Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Robust Use of External Information in Statistical Analysis

Dissertation zur Erlangung des Doktorgrades
Doctor philosophiae (Dr. phil.)

an der Universität Hamburg,
Fakultät für Psychologie und Bewegungswissenschaften,
Institut für Psychologie

Vorgelegt von

Martin Jann

Hamburg, 2025

Tag der Disputation: 16.09.2025

Promotionsprüfungsausschuss:

Vorsitzende/r: Prof. Dr. Alexander Redlich

1. Dissertationsgutachter: Prof. Dr. Martin Spieß

2. Dissertationsgutachter: Prof. Dr. Thomas Augustin

1. Disputationsgutachterin: Prof. Dr. Jenny Wagner

2. Disputationsgutachter: Prof Dr. Mike Wendt

Contents

1 Abstract

Scientific research involves different processes, including the accumulation, aggregation, and cumulation of knowledge. The latter is construed by using existing theories and empirical findings to obtain new results based on previous ones in further research. In psychology, the accumulation and aggregation of knowledge is employed in everyday research. This thesis sheds light on the cumulation of knowledge in psychology by focusing on statistical cumulation – the use of quantitative external information in statistical analyses. To prevent new results from being biased by misspecified external information, the uncertainties of the external information should be considered. This includes estimation and structural uncertainty, as external quantities are estimates and there are often structural differences between a new data set and external sources, such as different designs or populations. Furthermore, this thesis discusses a wide range of approaches for incorporating external information and considers how well they reflect present uncertainties. Previous approaches include generalized Bayesian analyses and inferential models that incorporate partial prior information about a parameter of interest. Careful consideration of previous approaches indicates that a frequentist approach had not been developed for this purpose in psychology. To address this issue, this thesis introduces an externally informed generalized method of moments approach, which was developed and outlined across the four attached papers. Within this novel approach, two uses of external information are possible: improving the statistical analysis of new data and testing the fit of external information and data to indicate structural differences between them. Furthermore, this approach can incorporate external information about variables in the form of statistical moment equations. Thus, it is a relevant addition to existing methods as generalized Bayesian and inferential model approaches have difficulty incorporating this type of external information. The main focus of the four attached papers was on the application of the externally informed generalized method of moments approach to multiple linear and repeated measures generalized linear models. Additionally, this PhD project provides software that allows applied researchers to use the developed approach with various models, such as multiple linear models, repeated measures generalized linear models, two-level mixed linear models, and structural equation models.

2 Introduction

The pursuit of knowledge has been a fundamental aspect of human history, driven by the efforts of numerous generations of researchers. Theories were developed and accumulated if they were not disproven or forgotten. The manner in which knowledge accumulates varies between scientific disciplines and is contingent upon the theoretical framework of each discipline.

In the field of mathematics, formal proofs of theorems are developed based on prerequisites that are meticulously defined. These proofs are designed to be applicable to any situation in which the prerequisites are satisfied. This fosters an optimal environment for the development of theories and a cumulative science, in the sense that existing theories are employed to derive new theories.

In the field of physics, theories are formulated using mathematical models derived from mathematical principles and sophisticated experiments. Once these models are established, they can be accumulated and used in the development of new theories. One of the most illustrative examples may be the development of electromagnetism, particularly the Maxwell equations, as described by Huray (2009): Coulomb's law was postulated in 1750 and confirmed in 1785. The invention of the battery by Volta in 1800 enabled experiments on electric currents. These experiments led Ampere to formulate his law relating circulating magnetic fields and currents in 1820. In 1831, Faraday discovered the law of electromagnetic induction through experimentation. All of these laws were then combined by Maxwell in 1864 into a unified theory through his equations. Each new theory in this example builds on older theories and experimental results. Furthermore, some theories are direct generalizations of older ones.

In the field of philosophy of science, cumulative science has been debated extensively. One major argument against the cumulative nature of science is the occurrence of paradigm shifts, revolutionary changes in scientific theories and in fundamental concepts, as discussed by Kuhn (1962). However, according to Kuhn (1987), there are both normal and revolutionary changes in science. The former is not made obsolete by the latter and represents the majority of scientific work in which knowledge accumulates within the current paradigm. Furthermore, Kuhn (1987) distinguishes between empirical laws

and theory. Although theories are holistic and prone to displacement during a paradigm shift, empirical laws can be tested directly using observations or experiments, adding to scientific knowledge. Though some difficulties remain in conception of empirical laws, from an idealized perspective, they seem to persist.

In the field of psychology, the concept of cumulative science may not be widespread, but the accumulation of theories and empirical results certainly is. One reason may be that modern psychology is a relatively young discipline. Its beginnings can be traced back to the mid-nineteenth century, when Gustav Theodor Fechner, Hermann von Helmholtz, Ernst Weber, and Wilhelm Wundt introduced the experimental method to psychological research (Schultz, 1981, Chapter 3).

As more and more theories and data emerged, the idea of cumulative science became increasingly attractive. In this context, Mischel (2009) criticized psychological researchers for treating theories like toothbrushes, using only their own. Mischel advocated for building a cumulative psychological science. Interestingly, he placed less emphasis on theoretical cumulation and more on the development of common tools, such as normed questionnaires and brain imaging techniques, as well as robust, replicable, and consequential findings.

New opportunities through online publications and open data facilitate the empirical realization of such a cumulative psychological science. However, the same adaptability that has been helping humanity develop and spread around the globe may be a fundamental reason for ongoing changes in psychological theory and empirical results. Nevertheless, stepping back from the idea of a global, persistent theoretical cumulation, there is still a chance for local, temporary cumulation, when it comes to empirical results.

2.1 Knowledge accumulation, aggregation and cumulation in psychology

This thesis defines knowledge accumulation as an increase in the number of theories and empirical findings maintained by researchers in a given field. In modern psychology, a vast number of articles are published each year. In June 2025, a search based on the term "psychology" in the PubMed database alone yielded 1,906,483 results (PubMed, 2025). Of these results, 85,165 were published in 2023, and 115,150 were published in 2024.

Due to the large number of studies, even in specific fields of psychological research, some aggregation of accumulated knowledge seems necessary. Knowledge aggregation refers to the development of a single source that encompasses the theories or empirical

findings of many sources, often with the goal of providing a concise summary. There are two common approaches to aggregating psychological knowledge from multiple studies.

On the one hand, narrative reviews provide a flexible, subjective method for analyzing a set of psychological studies. The possible goals of a narrative review include understanding a topic, developing theories, or evaluating and critiquing existing literature based on theoretical premises (Sukhera, 2022). These reviews are often qualitative, not quantitative, and do not use statistical or mathematical techniques.

On the other hand, there are systematic reviews and meta-analyses. The goal of a systematic review is to synthesize all available empirical evidence to answer a specific research question (Patole, 2021). Meta-analyses employ statistical methods to synthesize the results of a set of studies (Patole, 2021). There are criteria and checklists for conducting and reporting the results of systematic reviews and meta-analyses; see Higgins et al. (2019) and Page et al. (2021).

So far, the accumulation and aggregation of psychological knowledge has been discussed, but not its cumulation, i.e., how existing theories or empirical findings are used in subsequent research. Typically, authors cite previous studies in the introduction or theoretical background section of their paper, as required by the American Psychological Association (2020). The existence of a phenomenon is stated based on these citations. For instance, previous studies may indicate that two psychological concepts are (linearly) related. In some cases, the effect sizes of previous results are reported as small, medium, or large, according to the approach developed by Cohen (1962) for the categorization of statistical findings. Another convention is to indicate which models were used in previous studies or meta-analyses, as well as whether their parameters can be considered significantly different from zero. In this case, even though there is a mathematical model, the corresponding parameter estimates from previous studies are usually omitted.

It appears that the previous results are mainly reported so that the existence of a phenomenon can be established and new existence hypotheses can be derived. For instance, a researcher might hypothesize that the previously established correlation of psychological constructs also exists in a different scenario, or that other psychological variables influence this correlation. However, the estimated values of parameters or other statistics have been reported in previous empirical studies. The utilization of these quantities during the analysis of a novel data set, which will be referred to as *statistical cumulation* in the subsequent discussion, may prove advantageous for various reasons.

First, it may provide a theory-independent method of knowledge cumulation. Even if the models chosen in a previous paper were incorrect, the sample variance and mean

of an observed variable can still be valid estimates of its expected value and population variance. The reason is that the typical asymptotic properties of a single variable only depend on assumptions about that variable, not on its relationships with other variables (Casella & Berger, 2024). For example, even if theories about intelligence were falsified, the mean item score of an intelligence test in a specific sample remains the same estimate of the population item score.

Second, it may be helpful to derive a specific null hypothesis. The typical testing of whether a parameter is equal to zero, also known as the point-null hypothesis, was criticized early on for providing only weak support for a theory since it is almost always false (Meehl, 1967). As a solution, equivalence tests have been proposed for use in psychological research (Lakens et al., 2018). These consist of two one-sided tests in opposite directions, and the null hypotheses are constituted by the smallest effect size of interest (SESOI). An effect is only considered meaningful if its absolute value exceeds the SESOI. The specification of the SESOI may be guided by the results of previous empirical studies.

Third, statistical cumulation might help distinguish between different populations, detect selectivity, or analyze changes over time. If some relevant variable produces substantially different statistics in one sample compared to the previous results on the population of interest, then the findings and the theory may not be generalizable.

Fourth, it may improve the current statistical analyses. For example, if a value is known to be positive, a researcher may want to use a model that excludes negative values. Heuristically, if information is present, it should aid in inference or prediction much like a clue helps a detective solve a case by narrowing down the set of suspects.

Statistical cumulation involves using information that is external to the new data set in its statistical analysis. Throughout this thesis, *external information* will refer to quantitative information independent from the currently analyzed data. One way to develop a cumulative psychology is to develop methods for using external information in statistical analyses.

A note of caution should be made regarding the use of external information in multiple ways during statistical analysis. Using the same sources to define the SESOI and then to improve the statistical analysis may result in a dangerous conflation. The same external information may bias the statistical results and influence the stated hypotheses, raising the question of whether the new data sufficiently impact statistical inference.

Analyzing the effects of such a conflation would constitute its own research project and is outside of the scope of the current project. Thus, this thesis will focus on theory-independent cumulation, identifying differences between populations, and improving sta-

tistical analyses with external information, as outlined above. However, the thesis will not focus on externally informed hypotheses.

2.2 Goals of this dissertation

This dissertation aims to rigorously examine quantitative external information, its forms, uncertainties, and use in statistical analyses in psychology. It will discuss which types of uncertainties are present and how they can be reflected using an externally informed generalized method of moments approach based on F-probabilities. This approach is new to psychology and was developed as part of this PhD project. Additionally, it will provide an overview of existing approaches to incorporating external information into statistical analyses.

The focus of this work lies in techniques that can incorporate external information while representing its uncertainties. As will be discussed, there is extensive literature on using external information in a Bayesian framework. However, to the best of the authors' knowledge, no such framework existed for incorporating external information into frequentist analyses in psychology at the beginning of this research project in 2019. The main goal of the project has been to develop a novel framework for incorporating external information into frequentist analyses that reflects existing uncertainties.

It will be elaborated on how the developed approach incorporates external information. This will be followed by an explanation of how using external information can reduce the variance of estimates and thus increase statistical power. Techniques based on external information to assess whether the population behind the new data set is compatible with other populations will be discussed subsequently. Reflecting uncertainty provides a robust way of using external information, reducing or avoiding possible harm due to misspecified information, and indirectly making statistical inference more robust.

Following the exposition of the theoretical concepts, the contributions of each paper will be delineated. The first paper introduced a statistical test to detect differences between populations by assessing the fit of external information and data. The second paper analyzed how external information improves statistical analyses by reducing the variance of estimators, focusing on linear models. Building on the findings of the first paper, the third paper presents a general hypothesis-testing procedure for model parameters that incorporates external information and reflects its associated uncertainties. The fourth paper examined how external information improves statistical analyses of repeated measures generalized linear models. The contributions of each paper can be comprehended in the context of the presented, overarching theory.

3 External information and its uncertainties

For a precise definition of external information, some technical concepts are needed. There is a mathematical formulation behind statistical thinking, a foundation for estimation, statistical inference, and prediction. As a reference for the following discussion of mathematical foundations, consider Klenke (2020) or other introductory courses in probability theory or mathematical statistics. Let $(\Omega, \mathcal{A}, P)$ be an underlying probability space, where Ω denotes the sample space, $\mathcal{A}$ a σ-algebra (the set of measurable sets) and P a probability measure. As is common in statistics, this probability space is not specified. Here, random variables (measurable functions) on this space are of interest and the probability space just induces all relevant measures. The task of proving the existence of probability spaces for certain scenarios belongs to the domain of stochastics.

In applied statistics (and in this thesis), the existence of probability spaces is typically taken for granted. Since this thesis focuses on psychology, technical aspects such as measurability and the existence and uniqueness of distributional quantities (e.g., expected values) will not be discussed, but implicitly assumed to be true. Let $\mathbf{z}$ be a random variable on $(\Omega, \mathcal{A}, P)$, such that one unit of newly sampled data can be considered a realization of $\mathbf{z}$. Typically, multiple units are sampled from a vector-valued $\mathbf{z}$, inducing random variables $\mathbf{z}_i$ for $i = 1 \dots, n$, where n is the sample size. In this case the multiple units can be arranged as a matrix $\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1 & \cdots & \mathbf{z}_n \end{pmatrix}^T$, where the superscript T indicates that the matrix is transposed. In this thesis, bold uppercase letters represent matrices and bold lowercase letters represent vectors. Additionally, in regression models, y denotes the dependent variable and $\mathbf{x}$ denotes the independent variables.

Often, $P_{\mathbf{z}}$ and $P_{\mathbf{Z}}$ are linked by assuming that the $\mathbf{z}_i$ are independent random variables distributed identically like $\mathbf{z}$ (in short i.i.d.). In the following, $P_{\mathbf{z}}$ will be treated as a special case of $P_{\mathbf{Z}}$ for $n = 1$. Now, $\mathbf{Z}$ induces the pushforward measure $P_{\mathbf{Z}}$, called the distribution of $\mathbf{Z}$ or the *data generating process*. Typically, $P_{\mathbf{Z}}$ (or a conditional expectation based on it) is supposed to be specified by a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ denotes the parameter space, which is the set of possible parameter values. Throughout this thesis, the components of a parameter, such as its entries if the parameter is a vector or set, are referred to as "parameters."

A prominent example in psychology is the normal distribution $N(\mu, \sigma^2)$, which is fully specified by its mean, μ, and its variance, σ^2. Heuristically, if researchers knew the true values of both parameters, they would have all the possible knowledge about the phenomenon according to this statistical framework. The equivalent expressions $P_{\boldsymbol{\theta}}(\mathbf{Z})$ and $P(\mathbf{Z}|\boldsymbol{\theta})$ are used to denote the dependence of a probability distribution on a parameter. The true value of a parameter will be symbolized by a subscript of 0 to distinguish it from other parameter values.

In addition to the current data, there are often other sources about or related to the phenomena of interest. These will be called *external sources*. Technically, each external source provides a random variable, $\mathbf{Z}_{ex}$, which does not necessarily represent the same variables as $\mathbf{Z}$ and has its own data generating process, $P_{\mathbf{Z}_{ex}}$. The random variable $\mathbf{Z}_{ex}$ is called an *external sample*. Based on the realizations of $\mathbf{Z}_{ex}$, (external) empirical distributions of variables or models can be calculated. Therefore, if the full (realization of) $\mathbf{Z}_{ex}$ is given, then complete information regarding the external source is also given.

However, oftentimes only the reported results and quantities derived from the reported results are available. These quantities are called *external values* and are defined as $\mathbf{f}(\mathbf{Z}_{ex})$, where $\mathbf{f}$ is a function with values in $\mathbb{R}^k$ for some natural number k. For the sake of generality, the term "external values" will also refer to cases in which $\mathbf{Z}_{ex}$ is considered as a random variable, not just a realization. In this sense, expected values or probabilities are also external values. One example for this is expert knowledge, where an expert is consulted about population values, such as the probability of an event.

Further, external information is defined as any (quantitative) assertion about (an aspect of) $P_{\mathbf{Z}}$ that is based on a set of external values. This includes assertions about $\boldsymbol{\theta}$. Compared to external values, external information may be quite vague. For instance, one might assert that a construct is positively correlated with another, such as intelligence and the number of correct answers on a quiz. This assertion could be based on the fact that other studies and expert knowledge have provided positive external values of the correlation, despite the fact that these correlations differ in size.

It should be noted that the notion of external information links new data and external values. It is clear that external information may be incorrect. Therefore, the first step in using external information is understanding its uncertainties. Different types of uncertainty arise depending on the source of the external information. The following sections cover sources and types of external information, as well as uncertainties, and introduce a framework to represent these uncertainties. This is directly related to the quantitative representation of empirical knowledge in psychology, since external information is simply current quantitative knowledge.

3.1 Sources and types of external information

Given the proliferation of empirical studies in recent decades, the probability of identifying relevant external values has notably increased. An important property of an external source is the range of computable or reported quantities. Fortunately, many psychological papers follow the APA style, which requires reporting means and standard deviations, as well as test statistics and p-values (American Psychological Association, 2020). Some studies have data sets that are freely accessible online. This open data allows researchers to calculate any computable quantity of interest, making use of the full range of computable quantities.

Oftentimes there are multiple external sources containing information on the same population aspect and likely not agreeing on all results. One solution for this would be to aggregate all the available information using the methods discussed in the previous chapter. Although narrative reviews may provide more of a qualitative aggregation, systematic reviews with meta-analyses provide quantitative aggregation. A narrative review may conclude that a certain effect exists and has a direction. This can be translated into the assertion that a statistic, such as a correlation coefficient, is positive (or negative) in the population and, thus, in the data generating process of the new data. Meta-analyses, on the other hand, contain a variety of statistical results. Meta-analyses typically report effect estimates for individual studies, the overall effect estimate, and the respective confidence intervals (Higgins et al., 2019).

A variety of external information can be constructed based on a meta-analysis that varies in boldness. A (very) risky approach is to interpret the overall effect estimate as being equal to the population effect in $P_{\mathbf{Z}}$. A less risky approach would be to interpret the confidence interval of the overall effect as a confidence interval of the population effect in $P_{\mathbf{Z}}$. A conservative approach would be to create an external interval encompassing the full range of effect estimates from all the included studies. This yields the assertion that the population effect treated as an aspect of $P_{\mathbf{Z}}$ is bounded by the smallest and the largest estimates of the selected studies. Such external intervals could be calculated analogously using the lower and upper bounds of the reported confidence intervals of the individual studies. Another approach to formulating external information is acknowledging that the external values are estimates. For example, rather than claiming that the overall estimate equals the population effect, one could state that it is an unbiased or consistent estimate thereof.

If treated as an assessment of population values, external sources may be rated by the statistical quality of their results. The most reliable statistics on national populations are

likely provided by national administrative offices. These are known as official statistics.

As a reference for the following discussions of official statistics, see Wallgren and Wallgren (2014, Chapter 1). Three types of sources for official statistics can be distinguished: sample surveys, censuses, and (surveys based on) administrative registers. All three types aim to accurately describe aspects of the selected population. Sample surveys and censuses are based on collecting a data set using questionnaires developed for this purpose. The difference is that in a sample survey a random sample of the population is drawn (often based on address lists), whereas a census attempts to collect data from all members of the population.

Surveys based on administrative registers use existing data, which is often automatically collected for administrative purposes, not statistical ones. Examples include annual pay registers (collected to prevent tax evasion) and disease registers (collected by the healthcare system). For demographic variables (such as age) which are often assessed in psychological studies, there are official statistics, like the distribution of (a grouped version of) the variable in the population. One limitation to keep in mind is that a psychological study may not target the same population as existing official statistics. For example, a study of the cognitive abilities of depressed inpatients in Germany should not carelessly employ official statistics about the entire German population since it analyzes a subpopulation whose statistical characteristics may differ substantially.

In addition to published sources, there is subjective external information, such as expert opinions (Garthwaite et al., 2005; Kadane & Wolfson, 1998). For instance, a clinical expert may assert that at least half of untreated patients will experience worsening of a specific symptom within one year. The expert thus states that the median of the true distribution of symptom improvement for untreated patients is below zero. This may be the only available external source in new research areas where no published sources exist yet. One possible procedure in Bayesian statistics is to use experts to elicit a prior distribution of the parameter of interest, which will be discussed in Section 3.6. In this case the external information is that the subjective belief about the parameter $\boldsymbol{\theta}$ is reflected by the elicited distribution. Another approach would be to elicit only certain probability statements, such as "I am 90% sure that $\boldsymbol{\theta}$ lies in the interval $[1; 2]$".

External values, as well as the information derived from them, can be categorized based on statistical properties. In psychological studies, statistics such as means, variances, correlations, and parameter estimates are frequently reported. External means and (co)variances are examples of statistical moments. A statistical moment of a random variable v is defined as $E(v^k)$ for a natural number k (Casella & Berger, 2024). This notation includes sample moments when used with an empirical measure. If external

values can be considered as statistical moments of a random variable, then they are said to be of *moment-type*. External information is of moment-type if it is an assertion about statistical moments. This type of external information plays a crucial role in this thesis because it is often present in psychological studies. As will be discussed in Chapter 4, it challenges existing methods of incorporating external information into statistical analyses.

External information about the parameters of a statistical model will be called *model-specific*. For example, consider the assertion that the slope of a simple linear model of one variable on another is positive in the population. An important special case occurs when the external information is about the same model fitted to the new data. In this case, external information will be called *parameter-specific*. External information does not have to be model-specific. There is also *external information about variables*, such as the assertion that the expected value of a variable y lies within a specific external interval. This information is independent of any model linking y to other variables.

However, these two types of external information are not incompatible. External information about variables can be translated into model-specific information when a model is provided.

Example 1. (Jann, 2023) Suppose a simple linear regression model $y = \beta_1 + x\beta_2 + \epsilon$ is given and $E(y) = 100$ is known externally. Under the assumption $E(\epsilon) = 0$, the moment-type external information $E(y) = 100$ becomes a constraint on the parameters,

$$100 = E(y) = \beta_1 + E(x)\beta_2,$$

which is a linear constraint on intercept β_1 and slope β_2.

Similarly, under certain conditions, model-specific external information can be translated into information about variables. In a simple regression model, a positive slope is equivalent to a positive covariance between the dependent and independent variables. This assertion about a covariance is not model-specific – it is external information about variables. Nevertheless, these two types of external information should be distinguished because translating between them may be difficult for models that are more complex than linear ones. This translation would be necessary each time a new model is considered. Ideally, external information about variables would be used as is, without the need for translation.

The named external sources are characterized by a certain degree of uncertainty. That is to say, there is a need to ascertain whether the external information can be considered a correct assertion about aspects of $P_{\mathbf{Z}}$, the data generating process of the new data.

3.2 Estimation uncertainty

External information derived from other studies is based on an external sample $\mathbf{Z}_{ex}$, following its own data generating process $P_{\mathbf{Z}_{ex}}$. Assume that the external information about aspects of $P_{\mathbf{Z}}$ is correct if and only if it is also correct for the same aspects of $P_{\mathbf{Z}_{ex}}$. Informally, this means that the data generating processes agree on the aspects addressed by the external information. Even under this strict assumption, external information is still based on results reported in the external source. These results are only estimates of certain aspects of $P_{\mathbf{Z}_{ex}}$.

Therefore, such external information involves random variables and will most likely differ from sample to sample, even if the data generating process is the same. Due to the fact that population aspects are never exactly known but only estimated, this uncertainty was first called *estimation uncertainty* by Jann (2024) and will be referred to as such throughout this thesis. To reflect estimation uncertainty, a theory about the properties of the underlying estimators is necessary. Fortunately, estimation uncertainty is the most common type of uncertainty in statistics and is covered in both old and new introductory textbooks alike (Casella & Berger, 2024; Cox & Hinkley, 1974).

Estimation uncertainty can be addressed (asymptotically) by stating the (asymptotic) distribution of the estimator of which the external values are a realization. A typical frequentist approach to do this is to use laws of large numbers and central limit theorems to establish the consistency and asymptotic normality of estimators (Casella & Berger, 2024). These results are especially applicable to means or more general moment-type external information. If the estimator is asymptotically normally distributed and its variance has been estimated consistently, then the estimation uncertainty is asymptotically covered by stating its variance. Then, the asymptotic normal distribution is fully specified. To avoid imposing further assumptions, the variance should be calculated using the same external sample $\mathbf{Z}_{ex}$ and will be called *external variance*.

Even if the external sample is unavailable, variances are likely reported in the paper related to the external source due to APA style, as previously mentioned. However, if the external information is about a multidimensional aspect of $P_{\mathbf{Z}_{ex}}$, then a variance matrix, not just a variance, is needed for the asymptotic multivariate normal distribution to be fully specified. Covariances between variables may be more difficult to obtain than variances because they are not required to be reported. Fortunately, correlations between variables are often reported, so covariances can be calculated using correlations and variances.

Despite the asymptotic approach, another approach applicable to small samples is to

make direct assumptions about the form of the data generating process. Although this may be strict assumptions, it seems to be the only solution when the sample size is small and the data are not sufficiently representative for the entire population.

Estimation uncertainty is not directly applicable to subjective external information. Here, the goal is to accurately represent the beliefs or opinions of persons (e.g., experts), rather than objectively estimating population values (Garthwaite et al., 2005). However, it should be noted that experts and other professionals in a given field often base their knowledge on samples of the population, so estimation uncertainty may indirectly be at play. Section 3.6 will discuss the more prevalent uncertainty of subjective information in terms of the proper mathematical representation of subjective beliefs.

3.3 Structural uncertainty

External sources may not accurately reflect the scenario of the new data set. They may be based on a different population. Furthermore, sampling in psychology is often selective, which can appear as exclusion criteria or the "typical" student sample. This can result in biased estimates of population quantities (Spiess & Jordan, 2023). Previous experimental studies may have involved different stimuli, a greater or lesser number of trials, or varied procedures. Further, performance on a task of interest may vary across a day. This means that external information derived from studies in which participants were tested in the morning may not apply to a new study in which data was collected in the evening. In general, multiple developments occur over time, hence a data set that is several years old may not be compatible with the new data.

In many of these examples, it is questionable a priori whether or not the external information aligns with the new data. This discrepancy may be due to the external values being estimates or due to structural differences in the data generating processes of the external sample and the new data. Unlike estimation uncertainty, structural differences between external values and new data become more apparent and influential on statistical inference as sample sizes increase. For the case of selective sampling mechanisms Spiess and Jordan (2023) analyzed and reported this phenomenon. Structural differences may also be present when it comes to subjective information. This occurs if a subject specializes in a very specific area or works with particular populations. In these cases, the subject has only selective expertise.

Taken together, the assumption that the external information about aspects of $P_{\mathbf{Z}}$ is correct if and only if it is also correct for the same aspects of $P_{\mathbf{Z}_{ex}}$ seems to be too strict for psychological research. The uncertainty regarding this transferability of true assertions

is called *structural uncertainty*. This type of uncertainty is referred to by different names in the attached papers. For example, Jann (2023) used the term "epistemic uncertainty", whereas Jann (2024) used the term "qualitative uncertainty". The question remains of how to handle structural uncertainty. The following three approaches offer different ways to address this issue:

The first approach would be to either ignore structural uncertainty or identify theoretical obligations against its presence or relevance. This approach involves implicit statistical assumptions. For instance, one might assume that structural uncertainty cancels out on average.

If more than one study is available, then there are multiple external values and the question arises of how to combine them. If a sufficient number of studies are available, a meta-analysis is a sophisticated way to statistically aggregate sets of external values based on certain assumptions, as described in Section 2.1. However, the underlying assumptions may be incorrect. Meta-analyses form aggregated overall estimates across studies that may be structurally different, meaning their data generating processes differ in relevant ways. Even if the overall population mean is correctly estimated by a meta-analysis, each of the single studies as well as the new sample may differ from it structurally. When external information is used in statistical analyses, however, it should fit the respective population aspects of the new data set.

The second approach would be to directly analyze or even model structural uncertainty. Various psychological effects, even optical illusions such as the Müller-Lyer illusion, were shown to be different or even absent in some non-Western cultures (Henrich et al., 2010). Hence, external information from Western samples should be used with caution when adapting it to non-Western samples. With regard to performance variation over the course of a day, one could try to model the relationship between day time and performance. The resulting function could then be used to derive external information for a new data set that is sampled at a specific time of the day.

Although this seems to be the ideal scenario with structural uncertainty becoming structural knowledge, there are some drawbacks. This approach seems to be straightforward for metric variables like time, but it hardly applies to categorical variables like country of origin or to qualitative variables like stimuli or procedures. Technically, samples from all countries would be needed for a complete assessment of structural differences, since a country that is not sampled may differ substantially in its laws or culture. A prediction of the results in one country based on the results in another would require further assumptions regarding the underlying structure. Estimation uncertainty further complicates the analysis, as one would need a large enough sample size to infer that

differences are due to structural differences rather than random variation. Furthermore, it is impractical to model every aspect of structural uncertainty, since each new variable adds another dimension. For example, studies of the Müller-Lyer illusion may differ not only in population but also in procedure. Finally, an applied researcher typically only has access to limited external sources when conducting data analysis.

The third approach aims to address these problems by constructing external bounds to delimit the true value of the relevant aspects of the data generating process. This approach can be seen as a compromise between the first two. Rather than assuming that uncertainty is absent or averages out, it is assumed that the effect differs only within a certain range. No explicit model of the structural differences has to be specified, any shape within the bounds is valid.

To illustrate this approach, a thought experiment may be helpful. Consider how time of day affects task performance in a problem-solving experiment where the relative frequency of correct responses is of interest. Set aside estimation uncertainty and consider the following values to be the population values. Imagine that two previous studies had the same design but observed an effect mainly in the morning and evening, respectively. A new study may take place in the afternoon. Taking the range of values from both previous studies provides a safer estimate than taking a single aggregated value.

This is illustrated in Figure 3.1. Since the new value lies between the previous values, an interval that covers both also covers the new value. Although using an interval is safer than using a single value, some regions are not covered. At 9 p.m., the frequency of correct responses is not bounded by the previous values. In practice, achieving absolute safety is unrealistic since there are multiple possible sources of structural uncertainty. Furthermore, even if absolute safety were achievable, the resulting intervals might be uninformatively broad.

Even when ignoring structural uncertainty and relying on aggregated values, there may be multiple meta-analyses or reviews. Multiple official statistics may be available, such as administrative registers, which fosters the need for aggregation, often based on theoretical considerations (Wallgren & Wallgren, 2014, Chapter 1). One way to avoid developing higher-order aggregations is to apply the principle of structural uncertainty to these multiple sources of aggregated external information and construct an external interval based on their results.

Of the reasons discussed, the third approach was employed during this PhD project. Therefore, sets of external values were used instead of aggregated quantities. Note that sets of external values alone do not provide a reliable method for addressing structural uncertainty. When multiple external values are used without constructing an interval,

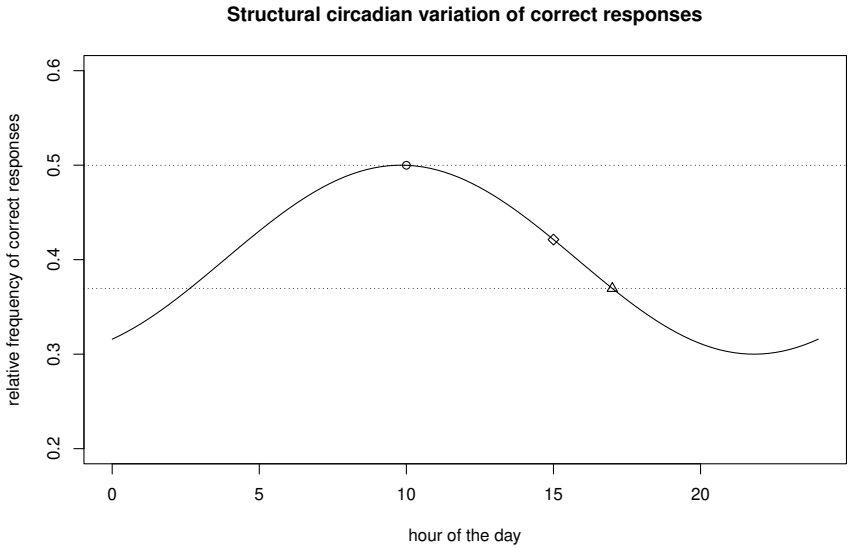


Figure 3.1: Hypothetical scenario of structural uncertainty. The curve depicts the true relative frequency of correct responses at a given time point. The circle and the triangle represent the frequencies reported in the previous studies in the morning and the evening, respectively. The square represents the frequency observed in a new data set. The dotted lines show the area that is covered by taking the interval from the lowest to the highest value of the previous studies, independent of time.

one of them is assumed to be correct, and even slight differences are not accounted for. Therefore, intervals are preferable because they cover deviations from the true values of other studies. Figure 3.1 shows that none of the old values (circle and triangle) is equal to the new value (square), but the square falls within the interval they define.

3.4 Combine structural and estimation uncertainty

In practice, structural and estimation uncertainty will occur simultaneously. The combined treatment of both, which is presented here, was developed by Jann (2024) and Jann and Spiess (2025). Figure 3.2 shows how the data generating process can be conceptualized when both types of uncertainty are present.

At the population level, there is not one true value, but rather a set of possible true values. Each new data set may be sampled based on a different true value. The mechanism behind selecting a true value can be deterministic, as shown in Figure 3.1. In practice, however, this selection process is unknown. Figure 3.2 only displays the range of possible true values to reflect this. After the true value is selected, the units of the data set are sampled according to a probability distribution. For example, it could be a

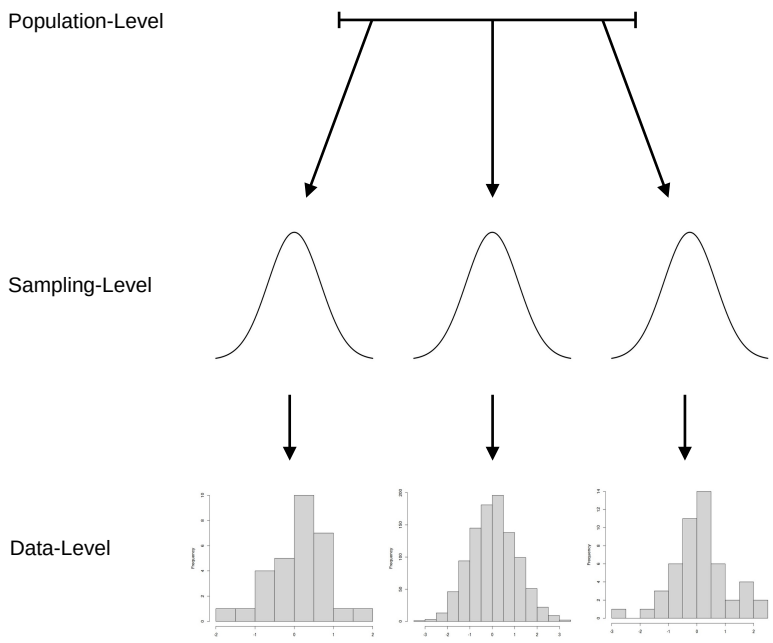


Figure 3.2: Scheme of the data generating process under structural and estimation uncertainty. The range of possible true values at the population level is displayed as a graphical interval.

normal distribution with the true value as the expected value. Note that this probability distribution may vary from data set to data set unless otherwise assumed. Thus, the observed data result from two nested mechanisms: a selection mechanism (which is not necessarily random) and a random draw from a probability distribution.

To reflect the uncertainties, the data generating process should be thought of as reversed. Estimates of a quantity of interest in a population are provided based on a number of studies or observed data sets. This set of estimates is denoted by $\mathbf{M}_{ex}$. For each estimate, the estimation uncertainty should be represented as described in Section 3.2.

Following the approach in Section 3.3, structural uncertainty can be represented by constructing a (multidimensional) interval covering all the estimates. The narrowest of such intervals can be constructed using the (elementwise) minimum and maximum estimates as the interval's boundaries. This interval is denoted by $\mathbf{I}_{ex}$. Note that this is an interval of estimates rather than an interval of population values. It is a random variable that would vary if the external samples were drawn again. Consequently, the

bounds only need to be unbiased (or at least consistent) estimates of the population values that encompass the true value of the new data.

The combination of both uncertainties is not fully realized at this stage. To illustrate this, consider means or moment-type external information. In this case the estimation uncertainty is represented by an asymptotic normal distribution and external variances as explained in Section 3.2. However, the whole interval $\mathbf{I}_{ex}$ is prone to estimation uncertainty. Even if many external sources are given, $\mathbf{I}_{ex}$ contains more values than there are estimates. This raises the question of how to consider estimation uncertainty for a non-estimate value.

In the one-dimensional case, each value in $\mathbf{I}_{ex}$ can be considered as a convex combination of the interval boundaries (Jann, 2024). As a consequence, the normal distribution applies to each value in $\mathbf{I}_{ex}$ and an external variance can be assigned by combining the external variances of the minimal and maximal estimates. This approach fails if there are multiple minimal or maximal estimates with different external variances. The process may also fail due to the non-uniqueness of convex combinations in the multidimensional case. Consider the midpoint of a square, for example. Because the two diagonals intersect at the midpoint, it can be described by two different convex combinations based on the two pairs of opposite corners of the square.

At a first glance, using the maximum external variance of all extremal estimates seems like it would solve the problem. This is considered an upper bound because it cannot be exceeded by any convex combination of the external variances of the extremal estimates. However, as will be explained in the following, this approach may be too pessimistic. There is a phenomenon called the *optimizer's curse* (J. E. Smith & Winkler, 2006). This "curse" is characterized by the fact that the maximum mean has a non-negative bias when estimating the maximum expected value. A similar statement immediately follows for the minimum.

Thus, if external means are considered as estimates and their minimum and maximum are used to construct an external interval, the expected mean interval will be at least as large as the population interval of expected values. In the context of utilizing external information, this effect is advantageous because it enhances the robustness of $\mathbf{I}_{ex}$ against misspecification. Regarding external variances, this effect reduces the need for an upper bound. As demonstrated in simulation studies, using the minimal external variance for inference purposes may suffice in some cases (Jann & Spiess, 2025).

It is debatable whether the proposed way of combining the two uncertainties may be a type of "double dipping" of uncertainty. On the one hand, different samples will almost always lead to different means even if sampled from the same distribution. Thus, the

same probabilistic variation affects both the width of $\mathbf{I}_{ex}$ and the external variances. However, when external sources have large sample sizes and consistent estimators are used, the external variances will likely be comparably small, reducing the "double dipping" effect as the sample size increases.

On the other hand, the number of external sources may be low, resulting in structural uncertainty not being covered for all possible true values. For example, consider Figure 3.1, where some areas are uncovered. A new study testing correct responses at night would not be covered by the current external interval. Therefore, a researcher should refrain from shortening $\mathbf{I}_{ex}$.

Overall, "double dipping" may be the price to pay for greater robustness against structural misspecification. The only reasonable way to reduce "double dipping" seems to be choosing a less conservative rule for calculating external variances for the values in $\mathbf{I}_{ex}$. Currently, only the concept of a data generating process is provided, in which multiple distributions are at play. To make the presented ideas usable for statistical inference, a mathematical formalism is needed.

3.5 F-probability as representation

The data generating process described in the previous section cannot be modeled by a single probability distribution because there are multiple distributions at play. A mathematical framework for addressing this issue is provided by imprecise probabilities (Augustin, Coolen, et al., 2014). In this framework, a set of distributions is called a *credal set*, and the idea is to treat this set as a mathematical object in its own right. As with probabilities, events can be evaluated based on credal sets. The following concept provides the theoretical basis for this:

Definition 1. (Augustin, 2002) Let Ω be a set and $\mathcal{A}$ be a σ-algebra on Ω. In addition, let $\mathcal{K}(\Omega, \mathcal{A})$ be the set of all probability measures on $(\Omega, \mathcal{A})$. Then a set-valued function F on $\mathcal{A}$ is called *Feasable-probability (F-probability)* with structure $\mathcal{M}$, if

1. there are functions $L, U : \mathcal{A} \to [0; 1]$ such that for every event $A \in \mathcal{A}$ it holds that $L(A) \leq U(A)$ and F has the form

$$\begin{aligned} F: \quad & \mathcal{A} \to \{[a; b] \,|\, a, b \in [0; 1] \text{ and } a \leq b\} \\ & A \mapsto F(A) \coloneqq [L(A); U(A)] \text{ for every event } A \in \mathcal{A}, \end{aligned}$$

2. the set $\mathcal{M} \coloneqq \{P \in \mathcal{K}(\Omega, \mathcal{A}) \,|\, L(A) \leq P(A) \leq U(A), \text{ for all } A \in \mathcal{A}\}$ is not empty,

3. if for all events $A \in \mathcal{A}$

$$\inf_{P \in \mathcal{M}} P(A) = L(A)$$
$$\sup_{P \in \mathcal{M}} P(A) = U(A).$$

The concept of F-probabilities was first introduced by Weichselberger (2001), presented in a detailed book (original in German). To refer to English-speaking literature, the subsequent discussion of F-probabilities is based on Augustin (2002). The functions L and U serve as the lower and upper bounds of possible probabilities.

For example, suppose that the probability of rain for tomorrow (coded as R) is reported differently by various sources, such as forecasting websites and apps. Thus, it can be concluded that the probability of R is between $L(R) = 0.1$ and $U(R) = 0.2$. The first requirement of Definition 1 ensures that these bounds are defined for all events and that they are reasonable. In terms of the example at hand, reasonable means $0 \leq L(R) \leq U(R) \leq 1$. The second requirement guarantees that there is at least one probability measure covered by L and U. In the current example, consider the event that it will not rain tomorrow, $\overline{R}$. For $L(\overline{R}) = 0.6$ and $U(\overline{R}) = 0.7$, the second requirement would not be met due to the complement rule of probability. In fact, the complement of $p \in [0.6; 0.7]$ is an element in $[0.3; 0.4]$, but this is disjoint from the interval $[0.1; 0.2]$ specified for R. Such cases, in which two assignments contradict the probability axioms, are ruled out. The third requirement ensures that the probability intervals are no wider than necessary. In the example, $F(R) = [0.1; 0.2]$, $P \in \mathcal{M}$ must satisfy $P(\overline{R}) \in [0.8; 0.9]$. Therefore, the infimum cannot be below 0.8, and the assignment $F(\overline{R}) = [0.5; 0.9]$ is not possible.

Interestingly, F-probabilities and credal sets are closely related. On the one hand, $\mathcal{M}$ is a credal set. On the other hand, given a credal set, one can use the third requirement of Definition 1 as a construction rule for an F-probability by using the credal set as $\mathcal{M}$. An important feature of constructing an F-probability based on a credal set is that the resulting $\mathcal{M}$ often contains more distributions than the credal set. All convex combinations of the elements of the credal set are included in $\mathcal{M}$. For a detailed discussion on this topic, see Jann and Spiess (2024). These convex combinations may differ substantially from the distributions in the credal set. For example, convex combinations of normal distributions can be skewed or multimodal.

To illustrate this property, consider the set of all normal distributions with variance equal to one and expected value in the interval $[0; 3]$. This set is denoted by the symbol

$\mathcal{N}[0;3]$. Figure 3.3 shows some of the convex combinations constructed based on $\mathcal{N}[0;3]$.

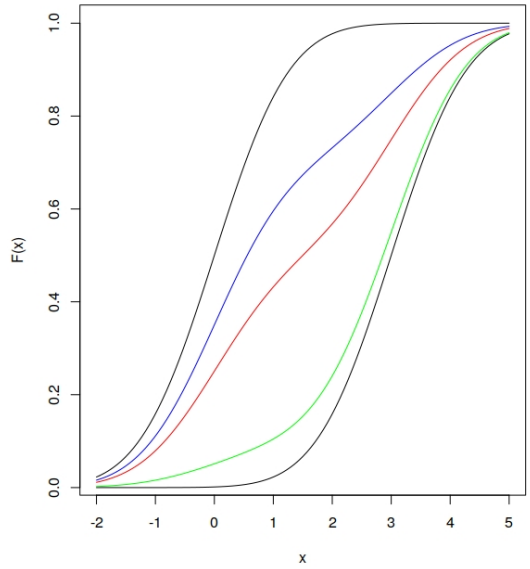


Figure 3.3: Displayed is the F-probability that was constructed based on the credal set $\mathcal{N}[0;3]$. Shown are the cumulative distribution functions of $N(0,1)$ (black, left) and $N(3,1)$(black,right), as well as convex combinations of the two (colored).

Now, all probability statements based on the constructed F-probability cover these convex combinations as well. The probability that they would assign to an event is always contained within the probability interval that F assigns to the same event.

This implies an increase in distributional robustness, since the true distribution is allowed to be non-normal, even if the credal set consists only of normal distributions. The approach is similar to the "neighborhoods" of distributions used in robust statistics to protect statistical analyses from small deviations from an assumed distribution (Huber & Ronchetti, 2009, p. 12). Despite violations of distributional assumptions, this effect may be helpful with small- to medium-sized samples, where the asymptotic distribution and the true distribution may differ to some extent.

Now, the data generating process from the previous section can be formalized. The population level is determined by a set of possible true values reflecting structural uncertainty. Each possible true value is mapped to a probability distribution reflecting its estimation uncertainty, and all resulting distributions are treated as a credal set. This credal set encompasses both the structural and the estimation uncertainty. Finally, an F-probability is derived based on the credal set. The F-probability generates an envelope

around the data generating process that includes more distributions than were used to create it. This increases distributional robustness. Thus, F-probabilities naturally arise when representing structural and estimation uncertainty.

3.6 Further uncertainties

Apart from structural and estimation uncertainty, further aspects should be considered before using external information in statistical analysis.

First, model-specific external information is only valid for a specific model. Assume there is external information about the slope of a simple linear regression model that states, "The slope is positive." for instance. Suppose a new study considers a multiple linear model that extends the simple linear model by an additional independent variable. One might be tempted to apply the external information about the slope of the independent variable in the simple linear model to the multiple linear model.

However, due to Simpson's paradox, the slope of the same independent variable may differ in a multiple linear model compared to a simple linear model (Simpson, 1951). The sign may even change, meaning the external information "The slope is positive" may not be directly transferable. A good discussion of Simpson's paradox in psychological research is provided by Kievit et al. (2013).

Second, even parameter-specific external information may not be correct for new data. This is because the employed model is likely incorrect, leading to model uncertainty. From a stochastic perspective, based on the data generating process, the estimator can still be seen as a random variable, even if the model is wrong. This random variable may still converge in probability to some aspect of the data generating process. This applies in the case of the typical i.i.d. assumption for the involved variables when treated as vectors, due to the asymptotic results discussed in Section 3.2. In a simple linear regression model, both the independent and dependent variables must align with the i.i.d. assumption, since the slope estimator involves both.

Suppose a simple linear model is chosen for statistical analysis, but the true relationship is $E(y|x) = e^x$. Further suppose that $y_i = e^{x_i} + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$, where $\sim$ denotes "(is) distributed like". In this case, the range of the observed values, $x_1, \ldots, x_n$, affects the slope of the fitted linear model. For instance, the result differs greatly depending on whether the range is $[-10; 0]$ or $[5; 10]$, because e^x increases exponentially with x. Therefore, researchers should investigate whether the distribution of the independent variable in the external sources is similar to that of the new data, before using external information regarding a simple linear regression slope. This is an example of structural

uncertainty, which shows that it may be better not to rely on a single external source.

A third important aspect is that many psychological constructs are latent variables that can not be observed directly. Instead, latent variables are measured indirectly based on manifest (observable) variables. This "indirect measurement" is achieved using a *latent variable model* (Loehlin & Beaujean, 2016). External information about a latent variable depends on the latent variable model. This can be considered a special case of model uncertainty.

However, a manifest score value is often reported and used in further analysis. This manifest score can be viewed independently of the latent variable model and represents an aspect of the data generating process. In practice, even if the latent variable model for a questionnaire changes, the mean scores (or other aspects of the manifest variables) in an external sample do not change.

Fourth, with regard to subjective external information, estimation uncertainty does not apply. Instead, uncertainty arises when representing the subjective knowledge of a person, such as an expert, in a precise mathematical form that can be used in statistical analysis. This issue is addressed by a procedure called *elicitation* (Kadane & Wolfson, 1998). The typical objective is to derive a distribution that mathematically represents subjective beliefs about a parameter and can serve as a Bayesian prior. See Section 4.1 for a description of the Bayesian approach. According to Garthwaite et al. (2005), medians and quantiles are easier to obtain than moments, and eliciting multivariate distributions poses additional difficulties. Elicitation is often done within a feedback loop, until the statistician and the person agree on the result.

Now, consider the following, admittedly provocative thought experiment. Suppose there is an agreed-upon final distribution. Then, calculate different aspects, such as the mean or a quantile, that differ from the elicited aspects. The results will have precise values that may seem odd to the person, such as a mean of 3.129, which may cause discomfort. A person may be unwilling to give a precise quantile or probability, but rather a statement such as "Thirty to fifty percent of patients are diagnosed with a major depression". Lastly, multiple persons may disagree with one another.

This could be explained by structural uncertainty. For example, one expert may base their experience on a structurally different phenomenon or group than another expert, even if the two seem similar. To allow for imprecision and circumvent "strange precision" or structural uncertainty, it may be better not to elicit a single distribution, but rather a credal set constituted by imprecise statements. The elicitation of credal sets (and imprecise probabilities in general) is discussed by Smithson (2014), who points out that it is a relatively new field of research.

Elicitation is certainly important from a non-subjective viewpoint as well. Based on current knowledge, it is the researcher's responsibility to determine which previous studies align with the new data set and which do not. Systematic reviews and meta-analyses address this issue by establishing inclusion criteria that align with existing guidelines. (Higgins et al., 2019). These guidelines can also help determine which external sources to include or exclude in a statistical analysis of new data.

However, some qualitative aspects of the new data may not be covered by external sources, particularly if there are few, if any, of them. In this case, it is important to avoid overly strict inclusion criteria. In order to construct an external interval that certainly captures the true effect, it may be beneficial to consider external sources that differ substantially from those recommended by guidelines. Including these additional sources will not result in misspecification. Rather, it will broaden the external interval $\mathbf{I}_{ex}$.

A fifth source of uncertainty pertains to research and publication practices in psychology. An essential feature of empirical science is that findings are reproducible. If researchers use the same or a similar study design, stimuli, and variables, they should find similar results. A large-scale attempt to replicate 100 psychological findings found that only 36% of the replicated results were statistically significant, whereas 97% of the original results were significant (Open Science Collaboration, 2015). Ensuing discussions on non-replicable findings in the field of psychology led to the establishment of the term *replication crisis*. Simmons et al. (2011) argue that p-hacking techniques, such as Hypothesis formulation After Results are Known (HARKing) or post hoc exclusion of participants, are a responsible factor for the replication crisis.

The presence of p-hacking can influence the reported statistics. Therefore, external information derived from this data may not be applicable to new data analyzed without p-hacking. Over the last decade, preregistration of studies has become a popular way to address the replication crisis. Preregistration involves publishing hypotheses, designs, and analysis plans online before sampling any data. Although preregistration may increase statistical power, its effectiveness in preventing p-hacking or HARKing is unclear (van den Akker et al., 2024).

To address issues such as p-hacking and HARKing, researchers should screen external sources. Researchers should check preregistration files to ensure transparency in the circumstances of the obtained results, such as exclusion criteria. This will help to rule out the most obvious cases of questionable research practices. However, more subtle cases, such as manual data manipulation, may go undetected when only a paper and its preregistration files are examined. If trustworthy, independent research groups manage

to replicate the findings, the external interval can still be valid. In this case, p-hacked results would only increase the width of an otherwise correctly specified external interval.

Apart from questionable research practices, there may be a *non-reporting bias* due to the fact that non-significant results are not published or are less visible to the research community (Higgins et al., 2019). To detect a possible non-reporting bias, meta-analyses often include (contour-enhanced) funnel plots. These plots display the effect sizes and their standard errors from all studies in the form of a scatter plot. Meta-analyses may also include sensitivity analyses based on these graphs (Higgins et al., 2019). One possible reason for the asymmetry of a funnel plot is a non-reporting bias. Sensitivity analyses, such as selection models, can then be used to determine how the results change under different assumptions about the origin of this asymmetry. The range of results from these sensitivity analyses constitutes another external set that can be used to construct an external interval.

Sixth, a technical aspect of external sources is how results are reported and the level of precision with which they are reported. The results reported in a paper are rounded to a specific number of digits. This means that the original result falls within a rounding interval of the reported value. For example, if the value 3.5 is reported, the original result falls within the interval $[3.45; 3.55]$. Although this may seem like a minor issue for large numbers, it is especially important for frequencies.

In some cases, the desired value is not reported, and instead, many other quantities are provided. An interesting approach to avoid discarding the external source is to use the presented results to approximate those of interest. However, the calculation is commonly restricted to the bounds on the value of interest. For example, Jann and Spiess (2025) calculated external intervals for frequencies of certain variable values based on means, variances, correlations, and other reported quantities while using rounding intervals for all quantities. This procedure yields sets of potential external values and, consequently, credal sets and F-probabilities. However, this approach has limitations. The calculated bounds may be very broad or uninformative, and the implementation may be both complex and case-dependent.

Seventh and finally, missing values in external samples can be a source of uncertainty. A statistical analysis based on fully observed cases only may be biased if some values are missing. Even if there is no structural uncertainty and the external sample size is large, the bias induced by missing data may persist. Thus, statistical techniques such as multiple imputations are needed to correct for this bias (Rubin, 1996). However, it is likely that not all previous studies employed valid techniques to handle missing data. It may not always be possible to perform multiple imputations on external samples. For

example, this method is inapplicable if the external sample lacks relevant variables for the intended imputation model or is unavailable.

One possible strategy to address these cases is the partial identification approach described by Manski (2003). This approach is particularly well-suited when the amount of missing data is small and the range of the variable or parameter is bounded. For example, consider a variable with values ranging from 1 to 4, such as a questionnaire score based on a four-point Likert scale. Suppose an external sample of size 100 has 10 missing values and a mean of 3.1, calculated on the basis of the observed values. The lowest possible mean would be reached if all missing values were 1. In this case, the mean would equal $0.1 \cdot 1 + 0.9 \cdot 3.1 = 2.89$. Similarly, the maximum possible value is $0.1 \cdot 4 + 0.9 \cdot 3.1 = 3.19$. Thus, the interval $[2.89; 3.19]$ encompasses all possible means with respect to the missing values.

Note that constructing this interval does not require any assumptions about the missing mechanism. However, this technique has limitations. The intervals broaden as the rate of missing values increases and the range of possible values widens. As a compromise, Manski (2003) provides a hierarchy of assumptions that enables researchers to balance the strength of additional assumptions with the width of the intervals.

The list of uncertainties discussed with regard to external information is not exhaustive. Other factors than those currently under consideration may become even more relevant in the future. In fact, it would be surprising to have certainty about all aspects of uncertainty. Although it may not be feasible to provide a comprehensive list of uncertainties, researchers should be considerate and transparent about the ones discussed here, even when deciding to disregard them in statistical analyses.

4 Existing statistical approaches that incorporate external information

This chapter examines statistical approaches that permit the use of external information. It discusses the extent to which these approaches can account for the uncertainties of that information. The approaches will vary in terms of the type of external information they are able to implement directly.

External information can be utilized for multiple purposes. Section 2.1 elucidated two primary applications of external information. First, it can be used to test the compatibility and fit of external information and data. Second, it can improve estimation, statistical inference, and prediction in new data sets. Note that this list is not exhaustive, and more uses may emerge in the future.

When asked about approaches to incorporate external information, many psychological researchers would likely name the Bayesian approach. It is arguably the most renowned approach for this purpose. From the author's perspective, it is debatable whether the standard Bayesian approach is appropriate for addressing structural uncertainty. Thus, the first section of this chapter is devoted to a generalized Bayesian approach that accounts for structural uncertainty and includes the classical Bayesian approach as a special case.

In contrast, there is no such general and renowned framework in frequentist statistics to incorporate external information. The second section of this chapter provides an overview of existing frequentist approaches. The mathematical details of a promising candidate for a frequentist framework, capable of representing structural and estimation uncertainty, will be presented in the next chapter. There are methods of statistical inference beyond the Bayesian and frequentist approaches that allow for the use of external information. One such approach will be discussed in Section 4.3.

4.1 Generalized Bayesian inference

Bayesian statistics is the most prominent method for incorporating external information. For a book-length treatment of the arguments presented here, see Bernardo and Smith

(1994). The foundation of Bayesian statistics is Bayes' theorem, also known as Bayes' rule, which is expressed by the equation

$$\underbrace{P(\boldsymbol{\theta}|\mathbf{Z})}_{\text{posterior}} = \underbrace{P(\mathbf{Z}|\boldsymbol{\theta})}_{\text{likelihood}} \cdot \underbrace{P(\boldsymbol{\theta})}_{\text{prior}} / \underbrace{P(\mathbf{Z})}_{\text{evidence}} .$$

The idea behind Bayesian statistics is to specify a prior distribution (in short prior) on the parameter space to reflect the prior knowledge and uncertainty of the parameter before seeing the data. Then, the prior is updated using Bayes' rule based on the observed data to obtain the posterior distribution (in short posterior). The prior is a genuine Bayesian concept that represents the degree of belief in certain parameter values as a probability distribution. Bayes' theorem can also be formulated with probability density functions, which is more convenient when dealing with continuous distributions. In its density form, the likelihood in Bayes's theorem is equal to the typical likelihood used in frequentist statistics and is often given by the density of the assumed sampling mechanism.

All statistical analyses are based on the posterior distribution. The shortest posterior interval with probability α, called the *highest density interval*, is the counterpart of a frequentist confidence interval. *Bayes factors* can be used to test relevant statistical hypotheses about the parameters. Predictions can be made by sampling values from the *posterior predictive distribution*.

The Bayesian approach is often accompanied by a subjective view of probability, interpreting it as a degree of belief. As stated in Section 3.6, subjective external information is elicited as a prior distribution, which can then be incorporated into a Bayesian analysis (Garthwaite et al., 2005). However, there are multiple approaches to Bayesian statistics. Objective Bayes for example, attempts to create "non-informative" default priors based on certain rules, such as Jeffrey's rule or reference priors. This approach avoids subjective elicitation but sacrifices the interpretability of the posterior (Martin & Chuanhai, 2015, p. 30).

Note that defining a prior is mandatory, not optional. However, specifying or eliciting a full prior distribution allows for many degrees of freedom and arbitrariness, particularly when the prior is continuous. This has often been used as an argument against the use of Bayesian statistics, prompting the development of methods to reduce arbitrariness, such as objective Bayes. Fortunately, as the sample size increases, the influence of the prior on the posterior decreases (given certain technical assumptions) (Bernardo & Smith, 1994, pp. 285 – 294). For small sample sizes, which are common in psychology, the prior can have a significant impact on the posterior.

Considering the different types of uncertainty in external information mentioned in Chapter 3 and its representation by F-probabilities, the Bayesian approach is inadequate because it only works with a single distribution. Model uncertainty plays an important role since the prior is specified for a parameter. Using a posterior distribution based on an external sample for a new data set requires employing the same model. Even for nested models, Simpson's paradox makes transferring external information risky.

Further, to reflect structural uncertainty, it may be better to rely on multiple sources and multiple previous posterior distributions that can be used as priors to form a credal set. In addition, it is uncommon for other studies to report complete posterior distributions. Often, only partial information is available, which is insufficient to identify a single prior, but rather a credal set. Moreover, as mentioned in Section 3.6, credal sets may be a way to address elicitation uncertainty.

Fortunately, it is straightforward to generalize Bayesian inference across the domain of credal sets and F-probabilities. The following elaboration is based on work by Augustin, Walter, and Coolen (2014). Instead of a single prior distribution, a credal set of prior distributions, denoted by $\mathcal{M}_{\boldsymbol{\theta}}$, is provided. This can be given through the set $\mathcal{M}$ of an F-probability representing external information. Then, each element of $\mathcal{M}_{\boldsymbol{\theta}}$ is updated according to Bayes' rule. The result is a credal set of posterior distributions, denoted by $\mathcal{M}_{\boldsymbol{\theta}|\mathbf{Z}}$. Note that this framework allows for a non-informative prior by defining $\mathcal{M}_{\boldsymbol{\theta}}$ as the set of all probability distributions on a given measurable space.

By constructing F-probabilities based on $\mathcal{M}_{\boldsymbol{\theta}}$ and $\mathcal{M}_{\boldsymbol{\theta}|\mathbf{Z}}$, this approach can be interpreted as updating an F-probability given the data. Interestingly, the set $\mathcal{M}_{\boldsymbol{\theta}}$ does not have to be equal to the set $\mathcal{M}$ of the prior F-probability for the update to produce the same posterior F-probability. It can be much smaller. According to the lower envelope theorem, if two prior credal sets induce the same F-probability, then their posterior credal sets will induce the same F-probability (Augustin, Walter, & Coolen, 2014, p. 154). For example, consider the F-probability shown in Figure 3.3 which is constructed based on the credal set $\mathcal{N}[0;3]$ as a prior. According to the lower envelope theorem, only the normal distributions in $\mathcal{N}[0;3]$ need to be updated, not their convex combinations. Updating the convex combinations would not alter the resulting posterior F-probability.

Statistical analyses using the generalized Bayesian approach are implemented by extending the corresponding methods from the classical Bayesian approach. Interval estimation can be performed by computing the highest posterior density interval for each element of the set $\mathcal{M}_{\boldsymbol{\theta}|\mathbf{Z}}$ and forming their union as sets (Walter & Augustin, 2009). Hypotheses can be tested using a generalized notion of Bayes factors based on a concept of practical relevance. This method was described in detail by Schwaferts (2022).

Thus far, the use of external information in Bayesian analysis has been discussed under the premise that the information is valid. However, if the validity of the external information is questionable, it may be better to contrast the data with external information to test whether they fit together. In Bayesian terms, a mismatch between data and external information is called a *prior-data conflict*.

In terms of the generalized Bayesian approach, Walter and Augustin (2009) addressed prior-data conflicts in generalized imprecise models with linearly updated conjugate prior knowledge, as well as potential solutions to these conflicts. As Jann (2024) argues, the prior-data conflict criterion developed by Walter and Augustin (2009) is liberal in that it incorrectly identifies many cases without conflict as having conflict. This is intentional, since the objective of Walter and Augustin (2009) was to rectify statistical inference for prior-data conflict. With this approach, missing a true conflict case is more problematic than generating a false positive. However, if the intention is to properly detect prior-data conflicts, then false positives matter.

To tackle this issue, Jann (2024) proposed an approach based on the work of Bickel (2015) for a conjugate normal model case. The approach is based on an assessment function, A, that evaluates the adequacy of a probability distribution given the data. Then, a credal set of probability distributions with an adequacy greater than a specified threshold is constructed. This credal set can then be compared to $\mathcal{M}_{\boldsymbol{\theta}}$. If there is no prior with an adequacy level that is at least as high as the lower bound, then that is considered a prior-data conflict. Since the data set is assumed to be observed and thus constant, it is omitted from the following notation.

Definition 2. (Bickel, 2015) Let $\mathcal{M}$ be a credal set. Let $A : \mathcal{M} \to \overline{\mathbb{R}}$ be a mapping to the extended real line. Then for $a \in \mathbb{R}$ a *set of a-adequate models* is

$$\mathcal{M}(a) \coloneqq \{P \in \mathcal{M} : A(P) \geq a\}.$$

As an assessment function, Bickel (2015) proposed the logarithmized Bayes factor or the integrated likelihood ratio. To derive a criterion for prior-data conflict, Jann (2024) proposed constructing the credal set $\mathcal{M}$ in Definition 2 using the general distributional form of the priors. For example, if all priors were normal distributions, then $\mathcal{M}$ would be the set of all normal distributions. The adequacy values of all $P \in \mathcal{M}$ are then determined based on A and the observed data.

Definition 3. Let $\mathcal{M}_{\boldsymbol{\theta}}$ be a credal set of prior distributions and $\mathcal{M}$ a credal set with the property $\mathcal{M}_{\boldsymbol{\theta}} \subset \mathcal{M}$. Then *prior-data conflict with threshold a (based on $\mathcal{M}$)* is said to occur if $\mathcal{M}(a) \cap \mathcal{M}_{\boldsymbol{\theta}} = \emptyset$.

Clearly, smaller values of a lead to a larger $\mathcal{M}(a)$ and thus the procedure less often decides in favor of prior-data conflict. Through simulation studies, Jann (2024) demonstrated that, for conjugate normal models with a variance of 1, values of a below 0 result in more conservative tests for prior-data conflict when using the integrated likelihood ratio assessment. Reasonable type I error rates were found for values of -0.1 or lower. However, power decreased as values of a decreased. Furthermore, when $a = 0$, it was shown that the prior-data conflict with threshold a performed identically to the criterion of Walter and Augustin (2009), so the latter can be considered a limit case in the scenarios examined. The limitations were that selecting and interpreting a can be challenging and that the results may be sensitive to slight variations in a (Bickel, 2015; Jann, 2024).

With respect to the type of external information, differences exist in how the information can be implemented in a Bayesian approach. In particular, external information about variables poses difficulties (Jann, 2023). It may be possible to use moment-type information about variables to restrict the set $\mathcal{M}_{\boldsymbol{\theta}}$. For example, if the true covariance of two variables is positive, then the slope of a simple linear model based on these two variables should also be positive. Therefore, a researcher could exclude prior distributions from $\mathcal{M}_{\boldsymbol{\theta}}$ for which the expected value of the slope is negative, or restrict the support of the prior distributions to positive values only.

To the best of the author's knowledge, no framework exists for directly including moment-type information about variables in generalized Bayesian inference thus far, though some investigations on this topic have been conducted. For example, Zellner (1996) developed a Bayesian method of moments approach. However, this approach has been criticized for not being properly Bayesian, but rather being based on the maximum entropy principle (Geisser, 1999). Additionally, Yin (2009) developed a Bayesian generalized method of moments approach. Unfortunately, this approach also used a full prior distribution for the parameter, which leads to the same issues when trying to incorporate moment-type information about variables.

4.2 Frequentist approaches

A variety of frequentist approaches allow for incorporating certain types of external information – some for a wide range of models. However, these approaches are not designed to reflect structural uncertainty. This chapter discusses the advantages and disadvantages of these approaches and concludes with an explanation of which approach could be extended to reflect structural uncertainty.

Some frequentist methods for incorporating external information are "domain spe-

cific". For example, suppose there are accurate estimates of the frequencies of values of a categorical variable within a population. These frequencies can then be used to adjust an estimate of a parameter in a new sample. By calculating a weighted estimate, the frequencies of the categorical variable in the new sample are adjusted to align with the previously estimated population values. This approach is called poststratification and can be used for various purposes, such as reducing the variance of an estimate or correcting for selective sampling (T. M. F. Smith, 1991). Since external frequencies should closely resemble population frequencies, they are often taken from census data. Suitable poststratification variables are demographic variables, such as age groups. As their name suggests, "domain-specific" methods for using external information are limited in the types of information they can incorporate. The presented form of poststratification is limited to incorporating the external values of the frequencies of categorical variables.

One general approach is the use of external information to establish constraints on the parameter space. For general constrained multiple linear and non-linear regression models with an additive error term, Knopov and Korkhin (2012) provide algorithms to calculate estimates and demonstrate the asymptotic properties of estimators, such as consistency and asymptotic normality. Their approach can be embedded into the broader field of stochastic optimization, which involves solving optimization problems where the objective function or the constraints include random variables (Rahimian & Mehrotra, 2022). This is achieved by approximating probability distributions and expected values with empirical distributions and means (Knopov & Korkhin, 2012, p. 73).

Stochastic optimization can be extended to distributionally robust optimization, for which only partial knowledge of the exact distribution is necessary, and which employs a set of possible distributions. For a discussion of the theoretical framework and recent developments in this field, see Rahimian and Mehrotra (2022). This approach can account for estimation uncertainty by treating parameter constraints as random and structural uncertainty by selecting an appropriate set of underlying distributions. However, distributionally robust optimization is not rooted in in frequentism. Therefore, establishing the necessary asymptotic theory for the approach when using common models in psychology might be better addressed in a research project in mathematical statistics. Nevertheless, such a project would be worthwhile because virtually all of the approaches presented in this thesis can be considered special cases of distributionally robust optimization.

Another issue with using external information as parameter constraints is that only parameter-specific information can be implemented directly. Therefore, the discussion about model uncertainty in Section 3.6 applies. Consider Example 1 again. Knowledge

of both $E(y)$ and $E(x)$ establishes an equality constraint on the parameters β_1 and β_2, which can be used for constraint optimization via the method outlined by Knopov and Korkhin (2012).

However, if only $E(y)$ is known, the constraint is partially specified and cannot be implemented directly within the framework of Knopov and Korkhin (2012). Using the empirical mean $\bar{x}$ instead of $E(x)$ makes the constraint random because $\bar{x}$ is a random variable. Therefore, stochastic optimization is necessary to implement this constraint. In addition to these technical issues, the parameter constraints imposed by moment-type information, such as the knowledge of $E(y)$, do not generalize well. These constraints must be derived anew for other models.

Another general approach, albeit a lesser-known one, is based on *confidence distributions*. A commendable reference to the prevailing theory in the one-dimensional case presented here is the review by Xie and Singh (2013). Consider a lower-sided $(100 \cdot \alpha)\%$ confidence interval defined by an upper bound and treat the confidence probability α as a variable rather than a fixed value. If the resulting function is continuous and increasing for each possible realized data set, then its inverse is a confidence distribution.

More generally, let $\mathcal{Z}$ be the sample space and Θ be the one-dimensional parameter space, then a function $H : \mathcal{Z} \times \Theta \to [0; 1]$ is a confidence distribution if two criteria are met. First, for each $z \in \mathcal{Z}$, the function $H(z, \cdot)$ must be a cumulative distribution function on Θ. Second, for the true value of the parameter, θ_0, the random variable $H(z, \theta_0)$ must be uniformly distributed on the unit interval $[0; 1]$. If these properties only hold asymptotically, H is called an *asymptotic confidence distribution*.

Multiple (asymptotic) confidence distributions, H_j, for $j = 1, ..., k$, on the same parameter space, with the same true value, θ_0, but possibly different sample spaces, can be combined into a single confidence distribution, H^c, using the combination rule

$$H^c = G_c(F_{de}^{-1}(H_1) + \cdots + F_{de}^{-1}(H_k)).$$

Here, F_{de} is the cumulative distribution function (CDF) of the standard double exponential distribution and G_c is the CDF of the random variable $F_{de}^{-1}(U_1) + \cdots + F_{de}^{-1}(U_k)$, where $U_1, \ldots, U_k$ are independent random variables with a uniform distribution on $[0; 1]$.

Using an equivalent formulation of this combination rule, Bickel (2012) developed a frequentist framework that incorporates external information regarding one-dimensional values. This is based on the idea that confidence distribution(s) of the new data set(s) can be combined with subjective confidence distributions derived from multiple independent persons (e.g., experts). This formalism allows for the incorporation of confidence

distributions from other studies, enabling the consideration of multiple sources of information, subjective or objective. Here, the underlying assumption is that the data generating processes of all studies possess the same true parameter value θ_0, although they are allowed to differ in other aspects.

This procedure can also be interpreted in a purely subjective manner, providing coherent inductive reasoning that is similar to a Bayesian approach (Bickel, 2012). Subjective confidence distributions resemble the prior distribution, and the combined H^c resembles the posterior distribution. Therefore, this approach is subject to the same critique regarding the elicitation of a single prior distribution (as discussed in Section 3.6).

An additional limitation to be considered is with respect to incorporating confidence distributions from multiple previous studies. Although this approach may cover estimation uncertainty, it does not cover structural uncertainty. If the true parameter value varies across studies, then H^c represents a mixture of these values, which may differ from the true parameter value in the new study. Finally, this approach does not easily generalize to multidimensional parameter problems. Although some results for multivariate confidence distributions are reported by Xie and Singh (2013), there seems to be no general solution to this problem to the best of the author's knowledge.

Lastly, there are two econometric approaches that provide frameworks with well-developed asymptotics and the ability to directly incorporate moment-type external information about variables. The first is the empirical likelihood approach (Owen, 1988), in short EL, and the second is the generalized method of moments (Hansen, 1982), in short GMM. The incorporation of moment-type external values of variables was demonstrated by Qin and Lawless (1994) for EL, and by Imbens and Lancaster (1994) as well as Hellerstein and Imbens (1999) for GMM. The EL approach will be outlined here. The GMM approach will be explained in detail in the next chapter.

Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be realizations of i.i.d. random variables, each distributed like $\mathbf{z}$. Subsequently, for a probability measure P on the same measure space as $\mathbf{z}$, the *empirical likelihood function (based on P)* is

$$L(P) = \Pi_{i=1}^{n} P(\mathbf{z}_i) = \Pi_{i=1}^{n} p_i.$$

For continuous distributions, it is imminent that $L(P)$ is zero. This is because they assign a probability of zero to each individual value, $\mathbf{z}_i$. Of interest are probability measures P that assign non-zero probability to each $\mathbf{z}_i$. This is the case for discrete distributions on the set of points $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, which may be possible empirical measures.

Assume that there is a function $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ with values in $\mathbb{R}^q$, that satisfies the moment

equations $E(\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)) = \mathbf{0}$. Statistical models are implemented by including the corresponding estimating equations in $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$. External information, such as $E(\mathbf{z}) = \mathbf{e}_0$, can be incorporated by adding its sample equivalent, $\mathbf{z} - \mathbf{e}_0$ to $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$, as an additional entry.

Note that only moment equations were specified to estimate the parameter, rather than specifying a likelihood function. The EL approach is named as such because it substitutes likelihood with empirical measures for realized data. Now, an estimate is derived by maximizing the empirical likelihood function subject to the constraints

$$p_i \geq 0 \text{ for } i = 1, \ldots, n, \sum_{i=1}^{n} p_i = 1 \text{ and } \sum_{i=1}^{n} p_i \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{0}.$$

This can be translated into a nested maximization procedure (Qin & Lawless, 1994). Given a parameter value, $\boldsymbol{\theta}$, the p_i values that maximize the empirical likelihood are searched for. As a result, the constrained optimization problem is nested within the unconstrained optimization problem of finding the $\boldsymbol{\theta}$ value that yields the highest overall empirical likelihood for maximal p_i. This maximum value of $\boldsymbol{\theta}$ constitutes the EL estimate and is consistent and asymptotically normally distributed under suitable regularity conditions (Qin & Lawless, 1994).

4.3 Approaches beyond frequentist and Bayesian

Taking a step back from the frequentist and Bayesian schools of thought, there are other approaches to statistical inference. Heuristically, statistical inference aims to "invert" the probabilistic data generation due to $P_{\boldsymbol{\theta}}(\mathbf{Z})$, meaning assertions about $P_{\boldsymbol{\theta}}(\mathbf{Z})$ are derived from the observed data set (Augustin, Walter, & Coolen, 2014). Expressing skepticism about the Bayesian approach, particularly the specification of a uniform prior distribution, Fisher (1935) endeavored to develop a prior-free inference method, which he called *fiducial inference*. In a manner analogous to the posterior distribution in Bayesian statistics, a fiducial distribution of the parameter based on the observed data is considered in this approach.

Fiducial inference has been the subject of extensive debates within the statistical community, particularly with regard to its interpretation and limitations in multidimensional scenarios. Some have even considered it to be Fisher's one great failure (Zabell, 1992). Although fiducial inference has historically garnered minimal attention beyond the domain of statistics, recent endeavors have emerged to develop inference methods that are (partly) motivated by fiducial arguments.

One such method is the approach of confidence distributions or confidence measures,

as presented in Section 4.2. There were disputes in the statistical community regarding the observation that confidence distributions appeared to be analogous to fiducial distributions. However, it should be noted that these distributions are based on frequentist considerations rather than the fiducial argument (Cox, 1958).

Another method motivated by fiducial arguments that has received attention in the field of imprecise probabilities is the inferential model approach described by Martin and Chuanhai (2015). The following is an introduction to inferential models based on this source. For the sake of simplicity, only the one-dimensional case will be considered.

In order to motivate the idea, one should consider what is given in a statistical inference problem. In most cases, a parametric sampling model is employed, which is a probability distribution that is intended to be equivalent to the data generating process. Suppose the sampling model is correctly specified, that is, assuming it is equal to $P_\theta(\mathbf{Z})$. Then, statistical inference will be conducted based on $P_\theta(\mathbf{Z})$. Therefore, the only unknown aspect remaining to be assessed is the true value of the parameter.

In a typical statistical inference problem, the true parameter value is considered a constant, as in frequentist inference. In addition, the data have been observed and are therefore also considered constant, as in a Bayesian posterior distribution. In this scenario, questions arise as to where the uncertainty comes from and how it should be modeled. Inferential models address these questions by introducing an unobserved auxiliary random variable, U, that introduces uncertainty and is primarily defined by the sampling model. Therefore, it is not necessary to specify a prior distribution for the parameter.

Consider, for example, the sampling mechanism $z \sim N(\theta_0, 1)$. This mechanism can be decomposed as $z = \theta_0 + \Phi^{-1}(U)$, where U is uniformly distributed on $(0; 1)$ and Φ^{-1} is the inverse of the standard normal CDF. If there is one observation z^* of z, then the equation becomes $z^* = \theta_0 + \Phi^{-1}(u^*)$ with a fixed value $u^* \in (0; 1)$. This indicates that a statistician could determine the constant θ_0 based on the known observed value z^* with knowledge of u^*. Thus, the decomposition separates the known observed value from the part that introduces uncertainty.

To formally describe inferential models, one must distinguish three steps: association, prediction, and combination. During the association step, an association function a is defined based on the assumed sampling mechanism, such that $z = a(U, \theta_0)$ with an auxiliary random variable U. By default, U is assumed to be uniformly distributed on $(0; 1)$. Looking at the realization $z^* = a(u^*, \theta_0)$, it is clear that possible values of θ_0 could be calculated by solving the equation if u^* were known.

The goal of the subsequent prediction step is to predict u^* based on the assumed

distribution of U. Randomly sampling a value of U is a good start, but using prediction sets can improve predictions. Therefore, random sets are used. For example, if $z \sim N(\theta_0, 1)$, then a valid random set for this purpose is

$$S(u) = \{u' \in (0;1) : |u' - 0.5| < |u - 0.5|\}.$$

The set $S(u)$ is larger for values of u that are further away from 0.5.

During the combination step, equation $z^* = a(u', \theta)$ is solved for each $u' \in S(u)$, and the results are combined. Let $\Theta_{z^*}(u')$ denote the set of solutions for $u' \in S(u)$. Then, the (random) set of possible parameter values is defined by

$$\Theta_{z^*}(S(u)) = \bigcup_{u' \in S(u)} \Theta_{z^*}(u').$$

Now, suppose that there is an assertion A about the parameter that can be viewed as a subset of the parameter space. For example for a real-valued parameter, the hypothesis $\theta_0 < 0$ corresponds to $A = (-\infty; 0)$. Let P_S be the pushforward measure of P_U under the random set. Assuming $\Theta_{z^*}(S(u)) \neq \emptyset$, the *belief function at A* is defined as

$$\text{bel}_{z^*}(A) = P_S\{\Theta_{z^*}(S(u)) \subseteq A\}.$$

Heuristically, the belief function can be interpreted as a "degree of belief" in the correctness of assertion A. The *plausibility function at A* is defined as $\text{pl}_{z^*}(A) = 1 - \text{bel}_{z^*}(\bar{A})$.

The belief and plausibility functions are both meaningful in a frequentist sense given certain requirements regarding the random set $S(u)$ (Martin & Chuanhai, 2015, pp. 60 – 62). Interestingly, $\text{pl}_{z^*}(A)$ can play a role similar to a p-value. Consider the procedure of rejecting the null hypothesis $\theta_0 \in A$, if $\text{pl}_{z^*}(A) \leq \alpha$ for some prespecified value $\alpha \in (0; 1)$, and maintaining the null hypothesis otherwise. Given the aforementioned requirements with regard to $S(u)$, this procedure controls the type I error rate at level α. Furthermore, the classic Bayesian update formula can be obtained using conditional inferential models.

Inferential models provide a framework for incorporating partial prior information (Martin, 2022, 2023a, 2023b). A detailed discussion of this framework would exceed the scope of this dissertation, therefore only a selection of aspects will be addressed. Martin (2022) considers the generalized Bayesian approach described in Section 4.1 to be a method of incorporating partial prior information in accordance with the inferential model framework. This indicates that inferential models can represent structural uncertainty.

To illustrate what partial prior information means, Martin (2023a) provides multiple examples. In one example, a person asserts that they are 90% sure that the true value of the parameter is less than or equal to 0.6. This assertion indirectly specifies that the 0.9 quantile of the prior distribution is 0.6. Another example is the external information that $E(\theta) = 0$ and $E(|\theta|) \leq K$, with a constant $K > 0$. In this case, the expected value and an upper bound on the expected absolute deviation from that value are both specified. In both examples, the prior distribution of θ is only partially specified. Therefore, a set of prior distributions is consistent with the partial information on the parameter.

One limitation of this approach is that the association function is usually based on a likelihood function, i.e., the assumed sampling model. It is unclear how to incorporate prior information when there is no likelihood. For example, if only estimating equations are provided for a model, this approach is not applicable (Martin, 2023b). Consequently, incorporating moment-type information about variables poses a challenge to the current stage of the inferential model framework.

4.4 Conclusion regarding existing methods

Generalized Bayesian and inferential model approaches are well-suited for representing structural uncertainty. In contrast, frequentist approaches require further development. The following comments seek to justify the development of a genuine frequentist method that incorporates external information. They are intended as a compromise between Bayesian and frequentist approaches.

In his work on multiple imputation (a technique based on the Bayesian approach that can also be applied to a frequentist analysis) Rubin (1987, p.67) argues that Bayesian and frequentist inferences will asymptotically lead to the same statistical inferences under some weak regularity conditions. Furthermore, he asserts that a Bayesian analysis that does not align asymptotically with a frequentist analysis is unreliable because it fails to approximate a relevant aspect of the population. Such an analysis would be a purely subjective speculation, which is not suited to the intersubjective endeavor of science. Therefore, in the asymptotic sense, frequentist inference is equally as justified as Bayesian inference. Regarding frequentist inference, Rubin (1987, p.67) urges that the results be interpretable in terms of knowledge about a parameter or quantity in a statistical model. The next chapter will address his calls by discussing a conditional interpretation of the proposed frequentist framework.

More generally, Martin (2022) proposed that the only way to resolve the conflict may be to embed typical frequentist techniques in a way that is compatible with the

Bayesian approach. Thus, the frequentist approach that will be presented in this thesis could contribute to this endeavor by offering a method for the incorporation of moment-type information about variables. As discussed in preceding sections, this has yet been difficult in the context of Bayesian and inferential model approaches.

Despite the increased use of Bayesian approaches, frequentist methods remain the most common way of performing statistical analyses in many areas of psychology. A text mining analysis of 57,909 articles on the PubMed Central database showed that the use of Bayesian statistics rose, but from 1% in 2010 to only 4% in 2021 (Böschen, 2023). The same analysis also showed that 85% of the articles reported frequentist p-values. In addition, a bibliometric analysis of articles published from 1962 to 2023 using the Scopus database revealed that no more than 3.1% of articles in any of the analyzed psychological subfields mentioned to use Bayesian methods (Jevremov & Pajić, 2024). In the context of cumulative science, it is imperative that applied researchers have the opportunity to incorporate external information, irrespective of their methodological approach, whether it be frequentist or Bayesian.

From the perspective of distributional robustness, as described by Huber and Ronchetti (2009), using a single likelihood function makes a strong assumption about the data generating process which is often violated in practice. GMM and EL only require the specification of moment equations based on certain aspects of the data generating process (Hansen, 1982; Owen, 1988). They do not require a full likelihood function. Furthermore, they can directly incorporate moment-type external information regarding variables, eliminating the need for prior translation in terms of parameter constraints for a specific model (Hellerstein & Imbens, 1999; Imbens & Lancaster, 1994; Qin & Lawless, 1994).

With regard to extensions that reflect structural uncertainty, EL appears less promising due to the computational complexity of the nested maximization used to calculate the EL estimate. Conversely, GMM enables reasonable computational procedures and analytical formulas for estimators, as demonstrated in the attached papers. The explanations in Section 3.3 show that an external interval should be used to reflect structural uncertainty. As will be discussed in Section 5.4, the evaluation of estimators or other statistics at an external interval is inherently difficult. Therefore, estimators that are less computationally intensive are preferable. Based on these arguments, GMM was chosen to be extended in order to reflect structural uncertainty.

5 Externally informed generalized method of moments

This chapter introduces the externally informed generalized method of moments approach. The fundamental technical concepts of GMM will be delineated. To demonstrate its applicability to psychological research, a discussion will be conducted on the range of relevant models that can be estimated using GMM. Subsequently, the text will elaborate on the incorporation of external information into the GMM framework and the representation of structural and estimation uncertainty in this process.

5.1 Basic technical concepts

The method of moments has a long history in statistics. Ronald A. Fisher and Karl Pearson likely coined this name during a dispute on maximum likelihood versus the use of sample moments (Pearson, 1936). In other terms, the idea can be expressed as follows:

Suppose the expected value of a function of a random variable, $E(f(\mathbf{z}))$, is to be estimated. If the data set can be considered a realization of random variables $\mathbf{z}_1, \ldots, \mathbf{z}_n$, i.i.d. like $\mathbf{z}$, then some law of large numbers and a central limit theorem may apply under minimal additional requirements (Casella & Berger, 2024). These theorems establish the consistency and normality of the estimator $\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i)$ of $E(f(\mathbf{z}))$. Thus, the rationale is to replace population moments with their corresponding sample moments or means.

When a statistical model is present, it is often possible to deduce sample moment equations that effectively serve as estimating equations. For example, employing a multiple linear regression model, $y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon$, the assumption $E(\epsilon|\mathbf{x}) = 0$ implies the population moment equations

$$\mathbf{0} = E(\epsilon\mathbf{x}) = E(\mathbf{x}(y - \mathbf{x}^T\boldsymbol{\beta})),$$

according to the law of iterated expectations (Cameron & Trivedi, 2005, p. 167). The corresponding sample moment equations are given by $\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta}))$. Solving for $\boldsymbol{\beta}$ yields the well-known ordinary least squares (in short OLS) estimator.

If there are more sample moment equations than parameters, then they are generally unsolvable because they pose an overidentified system of equations. For instance, this situation arises when external information is included, as will be demonstrated in Section 5.3. In his renowned work, Hansen (1982) developed the GMM and its asymptotics to resolve this situation. The underlying idea is to find an estimate that is "as close as possible" to a solution of the sample moment equations rather than seeking an exact solution, which probably does not exist.

To formalize this idea, let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ be a p-dimensional parameter, where p is a positive integer, and let $\boldsymbol{\theta}_0$ be its true value. Suppose the population moment equations are $E(\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)) = \mathbf{0}$, where $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ is an $\mathbb{R}^q$-valued function, and $q \geq p$. The corresponding sample moments are $\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$. To measure the distance of $\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ to $\mathbf{0}$, a mathematical norm on $\mathbb{R}^q$ is employed.

This raises the question of which norm should be used. For the Euclidean norm for example, the result is the sum of the squares of the entries of $\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$. Thus, all sample moment equations have equal influence on the result. However, heuristically speaking, one equation may be "more informative" than others. To address this issue, a weighting matrix $\hat{\mathbf{W}}$ can be used to create a norm in which the sample moment equations have different influences on the result.

Definition 4. (Newey & McFadden, 1994, p. 2116) Let $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ be a vector-valued function with values in $\mathbb{R}^q$, that meets the moment equations $E(\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)) = \mathbf{0}$. Let $\hat{\mathbf{W}} \in \mathbb{R}^{q,q}$ be a positive semi-definite, possibly random matrix. Then the *GMM estimator* $\hat{\boldsymbol{\theta}}_{ex}$ is defined as the value $\boldsymbol{\theta}$, that maximizes the following function:

$$\hat{Q}_n(\boldsymbol{\theta}) \coloneqq -(\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}))^T \hat{\mathbf{W}} (\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})) \ .$$

A set of assumptions is used to establish the statistical properties of the GMM estimator. These assumptions are called *regularity conditions*. This set is presented here to discuss some of the assumptions and give researchers the opportunity to determine whether their analysis scenario is suitable for GMM estimation (if this is not yet known).

Definition 5. (Newey & McFadden, 1994, pp. 2132, 2133, 2148) Let $\| \cdot \|$ denote the Euclidean norm of a matrix or a vector. Let $\nabla_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ denote the Jacobian matrix of $\mathbf{g}$ with respect to $\boldsymbol{\theta}$. The *regularity conditions for GMM estimation* are

1. The random variables $\mathbf{z}_i$ for $i = 1, \ldots, n$ are independent and identically distributed, or they follow a stationary and ergodic stochastic process.

2. $\hat{\mathbf{W}}$ converges in probability to a positive semi-definite matrix $\mathbf{W}$.

3. The equation $\mathbf{W}E(\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})) = \mathbf{0}$ only holds when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

4. The parameter space $\boldsymbol{\Theta}$ is a compact set.

5. The function $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ is almost surely continuous at each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

6. The expected value $E(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})\|)$ exists and is finite.

7. The true value of the parameter, $\boldsymbol{\theta}_0$, is contained in the interior of $\boldsymbol{\Theta}$.

8. The function $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ is almost surely continuously differentiable in a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$.

9. The expected value $E(\|\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)\|^2)$ exists and is finite.

10. The expected value $E(\sup_{\boldsymbol{\theta} \in \mathcal{N}} \|\nabla_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{z}, \boldsymbol{\theta})\|)$ exists and is finite.

11. Define $\mathbf{G} \coloneqq E(\nabla_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0))$, then $\mathbf{G}^T\mathbf{W}\mathbf{G}$ is non-singular.

The first regularity condition shows that the data set does not have to adhere to the i.i.d. assumption. Some dependencies are permitted. Although no formal definition of stationary ergodic processes will be provided here, the heuristics will be discussed below. For formal definitions and heuristic arguments, see Todorovic (1992). A process is (strictly) stationary if its marginal distributions remain unchanged under shifts in the time variable. An ergodic process is defined by the property that its time average converges to the ensemble average, i.e., the expected value of the process. In this case, observing a single unit multiple times and calculating the average over time asymptotically yields the same value as the mean of a sample of multiple units observed once.

A recent discussion about ergodicity in psychological data was presented by Hunter et al. (2024). According to the authors, ergodicity is implausible in many psychological scenarios, particularly with regard to its implied equality of within-person and between-person variance. However, stationary ergodic processes can be considered a generalization of i.i.d. processes, provided that the time and ensemble averages exist and are finite (Gray, 1988, pp. 210 – 212). Therefore, the fact that the i.i.d. assumption does not have to be exactly true still increases statistical robustness.

Note that the dependencies between entries in unit $\mathbf{z}_i$ are not restricted to any specific form. The term "unit" is not limited to participants. In educational research for example, a unit $\mathbf{z}_i$ could represent a school class, with its entries representing the students in that class. Students in the same class likely show correlated results.

The second regularity condition causes $\hat{\mathbf{W}}$, and thus its induced norm, to stabilize with larger sample size. In this sense, $\hat{\mathbf{W}}$ can be considered as a consistent estimate of the population weighting matrix $\mathbf{W}$. The third regularity condition, when combined with the second, ensures that the maximum of the objective function $\hat{Q}_n$ is asymptotically unique. If this condition is not met, then the estimator may be unidentifiable, or converge to an incorrect value. The fourth regularity condition is quite strict, as it requires the parameter space to be bounded. However, it can be replaced by a weaker version that will be stated below.

Under the regularity conditions 1 – 6, Theorem 2.6 of Newey and McFadden (1994, pp. 2132 – 2133) shows that $\hat{\boldsymbol{\theta}}_{ex}$ is a consistent estimator, in the sense that $\hat{\boldsymbol{\theta}}_{ex}$ converges in probability to $\boldsymbol{\theta}_0$. The asymptotic normality of $\hat{\boldsymbol{\theta}}_{ex}$ follows from Theorem 3.4 of Newey and McFadden (1994, p. 2148), if $\hat{\boldsymbol{\theta}}_{ex}$ is consistent and regularity conditions 7 – 11 are met. More precisely, define $\boldsymbol{\Omega} = E(\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)^T)$. Then, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{ex} - \boldsymbol{\theta}_0)$ converges in distribution to a normal distribution with a mean of $\mathbf{0}$ and the asymptotic variance of $\hat{\boldsymbol{\theta}}_{ex}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{ex}) = \frac{1}{n}(\mathbf{G}^T\mathbf{W}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G}(\mathbf{G}^T\mathbf{W}\mathbf{G})^{-1}.$$

So far, the selection of $\mathbf{W}$ and its estimator $\hat{\mathbf{W}}$ has not been elaborated on. Compared to other weighting matrices, the choice of $\mathbf{W} = \boldsymbol{\Omega}^{-1}$ leads to the most efficient estimator (Hansen, 1982). The resulting GMM estimator is called *optimal GMM estimator*. In this case, $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{ex})$ simplifies to $\frac{1}{n}(\mathbf{G}^T\mathbf{W}\mathbf{G})^{-1}$. A reasonable choice for the estimator of $\mathbf{W}$ is $\hat{\mathbf{W}} = \hat{\boldsymbol{\Omega}}^{-1}$, if $\hat{\boldsymbol{\Omega}}$ is a consistent estimator of $\boldsymbol{\Omega}$ and non-singular. For a singular matrix $\hat{\boldsymbol{\Omega}}$, it is possible to choose a generalized inverse, such as the Moore-Penrose inverse $\hat{\boldsymbol{\Omega}}^{+}$, as an estimator for $\mathbf{W}$, yielding the same asymptotic results (Xiao, 2020).

Note that the regularity conditions 1 through 6 could be replaced by any conditions implying consistency of $\hat{\boldsymbol{\theta}}_{ex}$. For example, the strict requirement that $\boldsymbol{\Theta}$ be compact could be changed to it being a convex set, while slightly tightening the other regularity conditions (see Theorem 2.7 of Newey and McFadden (1994)). Furthermore, Section 7 of Newey and McFadden (1994) shows that it is not necessary for $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ to be generally differentiable, only asymptotically differentiable. The regularity conditions for GMM estimation are sufficient but not necessary for the stated asymptotics. Therefore, similar asymptotics may hold under different conditions.

Following the establishment of asymptotic results, it is necessary to provide further information on estimation. The obvious sample moment estimators of $\mathbf{G}$ and $\boldsymbol{\Omega}$ are $\hat{\mathbf{G}} = \frac{1}{n}\sum_{i=1}^{n}(\nabla_{\boldsymbol{\theta}}\mathbf{g}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_{ex}))$ and $\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_{ex})\mathbf{g}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_{ex})^T$ (Cameron & Trivedi,

2005). The obvious variance estimator for the optimal GMM estimator is $\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_{ex}) = \frac{1}{n}(\hat{\mathbf{G}}^T\hat{\mathbf{W}}\hat{\mathbf{G}})^{-1}$. Unless otherwise stated, the obvious estimators and the optimal GMM estimator will always be chosen for the remainder of this thesis. Note that the obvious estimators are functions of the data and $\hat{\boldsymbol{\theta}}_{ex}$. In this case, however, $\hat{\mathbf{W}}$ is also a function of $\hat{\boldsymbol{\theta}}_{ex}$.

There are three asymptotically equivalent approaches to resolving the dependency of $\hat{\mathbf{W}}$ on $\hat{\boldsymbol{\theta}}_{ex}$ when calculating the GMM estimator (Hansen et al., 1996). They differ if small samples are given. The simplest approach is two-step GMM estimation. First, the GMM estimator is computed based on a simple, constant weighting matrix, such as the identity matrix $\mathbf{I}_q$. In this case, the second regularity condition is met, so the first-step estimator is consistent as long as no other relevant regularity condition is violated. The next step is to compute $\hat{\boldsymbol{\Omega}}$, based on the first-step estimator. Because the first-step estimator is consistent, the resulting matrix $\hat{\boldsymbol{\Omega}}$ is a consistent estimate of $\boldsymbol{\Omega}$. Then, the GMM estimator is recalculated using $\hat{\boldsymbol{\Omega}}$. The result is called a *two-step GMM estimator*.

An *iterative GMM estimator* is obtained by repeating this procedure until the estimates converge. If the optimization is performed on the function $\hat{Q}_n(\boldsymbol{\theta})$ while treating $\hat{\mathbf{W}}$ as a function of $\boldsymbol{\theta}$, the result is called a *continuously updating GMM estimator*. Although this optimization problem may be substantially harder than those in the other two approaches, it only requires one maximization. It should be noted that unless otherwise stated, the two-step GMM estimator will be used throughout the remainder of this thesis, as it has the least computational complexity. The first-order conditions for maximizing $\hat{Q}_n$ are $\hat{\mathbf{G}}^T\hat{\mathbf{W}}\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_{ex}) = \mathbf{0}$. The first-order conditions can be solved using root-finding algorithms, such as the Newton-Raphson method or the method of scoring (Cameron & Trivedi, 2005, pp. 341 – 348).

Statistical inference and prediction can be conducted based on asymptotic normality. The following delineation is based on Chapter 7 of Cameron and Trivedi (2005). In many application scenarios, only aspects of the parameter are of interest rather than the full parameter itself. Formally, let $\mathbf{f} : \boldsymbol{\Theta} \to \mathbb{R}^k$, where $k \leq p$, be a continuously differentiable function. The value of $\mathbf{f}(\boldsymbol{\theta}_0)$ may then be of interest. For example, the difference of two parameters could be represented by $f(\boldsymbol{\theta}) = \theta_1 - \theta_2$.

Now, plugging in the GMM estimator yields $\mathbf{f}(\hat{\boldsymbol{\theta}}_{ex})$. Let $\mathbf{R}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbf{f}(\boldsymbol{\theta})$ be the Jacobian matrix of $\mathbf{f}$, which is assumed to have full row rank, meaning its rows are not linearly dependent. Based on the asymptotic properties of the GMM estimator, an application of the delta method yields that $\mathbf{f}(\hat{\boldsymbol{\theta}}_{ex})$ is asymptotically normally distributed with a mean of $\mathbf{f}(\boldsymbol{\theta}_0)$ and an asymptotic variance of $\widehat{\text{Var}}(\mathbf{f}(\hat{\boldsymbol{\theta}}_{ex})) = \frac{1}{n}\mathbf{R}(\hat{\boldsymbol{\theta}}_{ex})(\hat{\mathbf{G}}^T\hat{\mathbf{W}}\hat{\mathbf{G}})^{-1}\mathbf{R}(\hat{\boldsymbol{\theta}}_{ex})^T$. These results can be used to construct significance tests and confidence intervals. When $\mathbf{f}$ is a

scalar-valued function, the asymptotic $(1-\alpha)\%$ confidence interval is given by

$$CI_{1-\alpha}(f(\boldsymbol{\theta}_0)) = \Big[f(\hat{\boldsymbol{\theta}}_{ex}) - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(f(\hat{\boldsymbol{\theta}}_{ex}))}; f(\hat{\boldsymbol{\theta}}_{ex}) + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(f(\hat{\boldsymbol{\theta}}_{ex}))}\Big],$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. The same procedure can be used to create a confidence interval for predicting the conditional mean. Suppose a model specifies the conditional mean of the dependent variable given the independent variables as $E(y|\mathbf{x}) = f(\mathbf{x}^T\boldsymbol{\theta})$. For a given value $\mathbf{x}_p$ of $\mathbf{x}$, the conditional mean is simply a function of the parameter, and the delta method can be applied as described above.

Similarly, significance tests can be constructed. Assuming that the null hyptohesis is $H_0 : f(\boldsymbol{\theta}_0) \leq 0$. Values different from 0 can be included by subtracting them from the current $f(\boldsymbol{\theta})$. For example, if the null hypothesis states that the difference between the two parameters is equal to or less than 3, then the result is $f(\boldsymbol{\theta}) = \theta_1 - \theta_2 - 3$. Based on the significance level, α, H_0 is rejected if

$$z_t = \frac{f(\hat{\boldsymbol{\theta}}_{ex})}{\sqrt{\widehat{\mathrm{Var}}(f(\hat{\boldsymbol{\theta}}_{ex}))}} > z_{1-\alpha}$$

holds. The other one-sided and two-sided tests work analogously. To account for estimated variances, it is reasonable to replace standard normal quantiles with the corresponding quantiles of a Student's t-distribution with $n-p$ degrees of freedom. More advanced significance tests are needed for multidimensional cases. Three such tests are typically used for hypothesis testing in the GMM framework.

Definition 6. (Cameron & Trivedi, 2005, p. 245) Suppose that the null hypothesis of interest is $H_0 : \mathbf{f}(\boldsymbol{\theta}_0) = \mathbf{0}$. Let $\hat{\boldsymbol{\theta}}_r$ denote the restricted GMM estimator using the unrestricted weighting matrix $\hat{\mathbf{W}}$, which is the maximum of $\hat{Q}_n(\boldsymbol{\theta})$ subject to the constraints $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$. Then, the *Wald test statistic* is

$$W \coloneqq \mathbf{f}(\hat{\boldsymbol{\theta}}_{ex})^T\big(\widehat{\mathrm{Var}}(\mathbf{f}(\hat{\boldsymbol{\theta}}_{ex}))\big)^{-1}\mathbf{f}(\hat{\boldsymbol{\theta}}_{ex}).$$

Let $\hat{\mathbf{G}}_r$ denote $\hat{\mathbf{G}}$ evaluated at $\hat{\boldsymbol{\theta}}_r$, not $\hat{\boldsymbol{\theta}}_{ex}$. The *Lagrange Multiplier test statistic* is

$$LM \coloneqq \big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_r)\big)^T\hat{\mathbf{W}}\hat{\mathbf{G}}_r(\hat{\mathbf{G}}_r^T\hat{\mathbf{W}}\hat{\mathbf{G}}_r)^{-1}\hat{\mathbf{G}}_r^T\hat{\mathbf{W}}\big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_r)\big)$$

and the *difference test statistic* is

$$D \coloneqq n(\hat{Q}(\hat{\boldsymbol{\theta}}_r) - \hat{Q}(\hat{\boldsymbol{\theta}}_{ex})).$$

Under the null hypothesis, all three test statistics are asymptotically χ^2_k-distributed with k degrees of freedom. As before, k represents the dimension of the values of $\mathbf{f}$. The null hypothesis is rejected with significance level α, if a test statistic exceeds the $1-\alpha$ quantile of the χ^2_k-distribution. The ideas behind the tests are quite different.

The Wald test statistic measures how far $\mathbf{f}(\hat{\boldsymbol{\theta}}_{ex})$ is from $\mathbf{0}$. The LM statistic measures how far the first-order conditions of the GMM are from $\mathbf{0}$, when evaluated at the restricted GMM estimate. The D statistic measures the difference between the maximum values of the GMM objective function for the unrestricted and the restricted estimation. Therefore, D appears to be similar to the classic likelihood ratio test. The likelihood ratio test is not discussed here since it does not generalize well to scenarios in which only estimating equations are given and no likelihood is provided (Cameron & Trivedi, 2005, p. 244).

Although these three tests are asymptotically equivalent, they may produce different results in small samples. In small samples, the Wald test seems to perform worse than the others (Bond & Windmeijer, 2005). Since the focus of frequentism is on asymptotic properties, the tests are sometimes defined differently, with these differences vanishing asymptotically. For example, Bond and Windmeijer (2005) calculated the variance matrix estimate in W based on a first-step estimator.

As Dufour et al. (2016) analyzed, the tests differ in terms of their invariance when it comes to equivalent hypothesis reformulations or reparametrizations. D is the only test that is invariant under equivalent hypothesis reformulations because the maximum values of the objective function, $\hat{Q}_n$, remain unchanged under such reformulations for both restricted and unrestricted optimization. Although none of these three tests are invariant under reparametrization, this property can be achieved by modifying D to use the continuously updating GMM estimator instead of the two-step GMM estimator.

Despite testing hypotheses about (aspects of) $\boldsymbol{\theta}_0$, the GMM framework offers a test for overidentifying restrictions when there are more moment equations than parameters.

Definition 7. (Hansen, 1982; Sargan, 1958) Suppose the dimension of $\boldsymbol{\theta}$ is lower than the number of moment equations, i.e., $q > p$. Assuming the prerequisites of Definition 4 are met, the **Sargan-Hansen test statistic** is defined as follows:

$$SH \coloneqq -n\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex}).$$

Under the null hypothesis $H_0 : E(\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)) = \mathbf{0}$, the test statistic SH has a χ^2_{q-p}-distribution. The null hypothesis is rejected at a significance level of α if SH exceeds the $1 - \alpha$ quantile of the χ^2_{q-p}-distribution. The Sargan-Hansen test can be applied to assess whether the data and the specified moment equations are compatible.

5.2 Scope of possible models and estimation methods

For the GMM approach to be useful in the field of psychology, it should encompass the majority of models employed in psychological research. This section aims to underline that this is the case indeed. As described in the previous section, multiple linear models can be used in the GMM via the sample moment equations $\mathbf{0} = \sum_{i=1}^{n}(\mathbf{x}_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta}))$, which solution is the OLS estimator. This already covers typical analysis of variance (ANOVA) models since they can be expressed as multiple linear models, as discussed in most introductory statistics courses, such as Rasch et al. (2011, pp. 401 – 402).

As a substantial extension, generalized linear models with repeated measures can be incorporated into the GMM using generalized estimating equations (GEE) (Cameron & Trivedi, 2005, p. 790). The GEE approach was developed by Liang and Zeger (1986). For an introduction to GEE, see Diggle et al. (2002). The idea behind GEE is to specify a link function, $\boldsymbol{\mu}$, that maps the linear combination of the regression parameters and independent variables, $\mathbf{X}\boldsymbol{\beta}$, to $E(\mathbf{y}|\mathbf{X})$. Common examples include the log-link for count data and the logit-link for logistic models. To model the dependencies among different measurements of the dependent variable, a working covariance matrix, denoted by the symbol $\boldsymbol{\Sigma}$, is used. This matrix is modeled as a function of multiple correlation parameters, $\boldsymbol{\alpha}$, and one scale parameter, ϕ. Now, the GEE are

$$\bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \mathbf{X}^T\mathbf{D}\boldsymbol{\Sigma}(\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})) = \mathbf{0},$$

where $\mathbf{D}$ is a diagonal matrix containing the derivatives of $\boldsymbol{\mu}$ with respect to $\mathbf{X}\boldsymbol{\beta}$.

Clearly, generalized linear models and repeated measures ANOVAs can be considered special cases. The GEE approach is a marginal method in the sense that the regression coefficients are of interest, while the remaining parameters of the covariance matrix are considered nuisance parameters. GEE can be applied not only to repeated measures, but also to correlated data in general. Notably, even in instances where the covariance matrix is misspecified, GEE can generate consistent estimates (Liang & Zeger, 1986). Although correlation parameters are usually considered nuisance parameters, GEE can be modified with additional estimating equations to consistently estimate covariance

parameters if they are of interest (Spiess, 2006).

In addition to repeated measures, there are other sources of dependency structures in psychological data. In educational, social, and organizational psychology, data sets are often hierarchical. For example, participants in a study may belong to groups such as classes, schools, or teams. Participants in the same group may share characteristics, such as having the same teacher or team leader. This creates statistical dependencies among participants' responses within the same group while keeping them independent of responses from other groups.

There are multiple approaches to analyzing these clustered data sets. Although most psychological researchers use multilevel analysis, clustered data can also be analyzed using GEE (McNeish et al., 2017). Considering the prevalence of multilevel analysis, it will also be briefly examined in the following. The presentation of multilevel analysis will align with standard textbooks on the subject (de Leeuw & Meijer, 2008). Multilevel analysis allows for both fixed and random regression coefficients, as well as different regression error variances.

Suppose the data set is structured into m groups, where $j = 1, \ldots, m$ denotes a specific group, and n_j its group size. Let $\mathbf{X}$ denote the design matrix of an underlying linear regression model corresponding to the variables with a fixed effect on the dependent variable $\mathbf{y}$. Due to the group structure, $\mathbf{X}$ can be split into parts $\mathbf{X}_j$ for $j = 1, \ldots, m$. Let $\mathbf{U}$ be another constant matrix containing the variables in the data set that have a random effect on $\mathbf{y}$. Again, let $\mathbf{U}_j$ denote the part of $\mathbf{U}$ for group j. Though not a requirement, the matrices $\mathbf{X}$ and $\mathbf{U}$ are allowed to have some variables in common. A *two-level mixed linear model* is given by

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{U}_j\boldsymbol{\delta}_j + \boldsymbol{\epsilon}_j$$

for $j = 1, \ldots, m$, where $\boldsymbol{\beta}$ represents the fixed regression coefficients and $\boldsymbol{\delta}_j$ as well as $\boldsymbol{\epsilon}_j$ are random variables. Suppose $\boldsymbol{\delta}_j$ is independent of $\boldsymbol{\delta}_k$ and $\boldsymbol{\epsilon}_l$ for all $k \neq j$ and $l = 1, \ldots, m$. In addition, assume that $\boldsymbol{\epsilon}_j$ is independent of $\boldsymbol{\epsilon}_k$ for all $k \neq j$. Furthermore, suppose $\boldsymbol{\epsilon}_j \sim N(\mathbf{0}, \sigma_j^2\mathbf{I}_{n_j})$ and $\boldsymbol{\delta}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Given these requirements, $\mathbf{y}_j \sim N(\mathbf{X}_j\boldsymbol{\beta}, \mathbf{V}_j)$, where $\mathbf{V}_j = \mathbf{U}_j\boldsymbol{\Sigma}\mathbf{U}_j^T + \sigma_j^2\mathbf{I}_{n_j}$. This result shows that including random effects introduces potential correlations between measures of the dependent variable.

Furthermore, the two-level mixed linear model model splits the variance matrix of $\mathbf{y}_j$ into two components. This is why some literature refers to these models as "variance components models". Special cases include two-level random coefficient models, random intercept models, and slopes-as-outcomes models. These models are frequently employed

in psychological research because they enable the analysis of within- and between-cluster variation through the respective variance components (McNeish et al., 2017).

Many psychological concepts of interest, such as intelligence and personality, are latent constructs that are not directly observable. *Structural equation models* can be used to model these latent variables and their relationships (Loehlin & Beaujean, 2016). These models can be specified using multiple linear regression models (Bollen, 1989, Chapter 8) or by defining the implied covariance or correlation matrix of the observed variables as a matrix product (McArdle & McDonald, 1984). The latter method will be presented here, as explained by Loehlin and Beaujean (2016).

Suppose the model consists of t variables in total, including m observable variables and $t-m$ latent variables. Let the t variables be in any order, so the implied order of the m observable variables can be determined by removing the latent variable ranks. The implied covariance structure can now be specified using three matrices. The order of the columns and rows in a matrix reflects the order of the variables that was previously determined.

First, there is an $m \times t$ filter matrix, $\mathbf{F}$, which indicates which variables are directly observable and which are not. For example, if the first variable is the first observable variable, then a 1 is placed in the top left cell of $\mathbf{F}$. Otherwise, a 0 is placed there. Next, a symmetric $t \times t$ matrix, $\mathbf{S}$, is defined that contains the assumed covariances and variances of all the variables. Then, an asymmetric $t \times t$ matrix, $\mathbf{A}$, is constructed to represent the assumed directed paths between the variables. These paths represent the relationships between latent variables, as well as the effects of latent variables on observable variables. In both $\mathbf{A}$ and $\mathbf{S}$, unknown elements are represented by symbols. These symbols correspond to parameters that need to be estimated. In conclusion, the covariance matrix of the observable variables implied by the three matrices is

$$\mathbf{C} = \mathbf{F}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{S}(\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{F}^T.$$

The implied covariance matrix can be compared to the sample covariance matrix, $\hat{\mathbf{S}}$, of the observable variables, which is not dependent on any particular model. The objective is to calculate parameter values that minimize the distance between the two matrices. These values are then used as estimates. If $\mathbf{C}$ contains more parameters than $\hat{\mathbf{S}}$ has entries, there may be multiple estimates that minimize the distance between the two matrices. To identify a single estimate, the variance of the latent variables is often fixed to 1, or the value of one path originating from each latent variable is set to 1.

Under the assumption of multivariate normality of the variables, both two-level mixed

linear models and structural equation models fully specify the distributions of the random variables involved. Therefore, the *maximum likelihood* method is often used to estimate their parameters (de Leeuw & Meijer, 2008; Loehlin & Beaujean, 2016). For a technical reference on maximum likelihood estimation and how it is a special case of GMM under the given regularity conditions, see Newey and McFadden (1994).

The idea behind maximum likelihood estimation is to maximize the density of the assumed sampling model $P_{\boldsymbol{\theta}}(Z)$for the parameter given the data. Thus, the roles of the parameter and the data have switched: The parameter is now treated as a variable, while the data are fixed. The resulting function is denoted by $L_{\mathbf{Z}}(\boldsymbol{\theta})$ and is called the *likelihood (function)*. Due to the multiplicative nature of the likelihood, it is often more convenient to work with the logarithmized likelihood, also known as the *log-likelihood*. Because it is additive, the log-likelihood often makes it easier to determine the first and second derivatives. Further, the log-likelihood can often be simplified, for example by removing constant terms, as they are irrelevant to maximization. To find the maximum likelihood estimate, the root of the first derivatives of the log-likelihood, also known as the score function, is approximated. This score function can be implemented directly within the GMM as sample moment equations (Newey & McFadden, 1994).

The log-likelihood of the two-level mixed linear model, ignoring the constant terms, is

$$\ln L_{\mathbf{y},\mathbf{X},\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \{\sigma_j^2, j = 1, \ldots, m\}) = -\sum_{j=1}^{m} \ln(\det(\mathbf{V}_j)) + (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta})^T \mathbf{V}_j^{-1} (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}),$$

where "ln" denotes the natural logarithm and "det" denotes the determinant of a matrix. Similarly, the log-likelihood for the restricted maximum likelihood (REML) estimation could be stated if an unbiased estimate of the variance components $\boldsymbol{\Sigma}$ and $\{\sigma_j^2, j = 1, \ldots, m\}$ is of interest (de Leeuw & Meijer, 2008, p. 22).

For the structural equation model, let $\boldsymbol{\theta}$ denote the vector of the parameters involved in $\mathbf{C} = \mathbf{C}(\boldsymbol{\theta})$. After removing the irrelevant constants, the log-likelihood is

$$\ln L_{\hat{\mathbf{S}}}(\boldsymbol{\theta}) = \ln(\det(\hat{\mathbf{S}})) + m - \ln(\det(\mathbf{C}(\boldsymbol{\theta}))) - \operatorname{tr}(\hat{\mathbf{S}}\mathbf{C}(\boldsymbol{\theta})^{-1}),$$

where in addition "tr" denotes the trace of a matrix (Loehlin & Beaujean, 2016, p. 55). Note that maximum likelihood is not the only method of formalizing the difference between the implied and observed covariance matrices. Other approaches, such as generalized least squares, can be used to generate different estimating equations for structural equation models (Loehlin & Beaujean, 2016). The score functions that can be incorporated into the GMM estimation process are constructed by stacking the first

derivatives of the stated log-likelihoods with respect to each of the parameters. The result is the gradient of the log-likelihood. As will be discussed in the next chapter, this can be done numerically without the need to analytically determine the gradients.

Finally, GMM permits the use of certain M-estimators. M-estimators generalize the maximum likelihood method and provide robust estimation techniques (Huber & Ronchetti, 2009). Let $\mathbf{z}_i$, for $i = 1, \ldots, n$, denote the observed data set (i.e., a realization, not a random variable) with the same sample space $\boldsymbol{\Omega}_\mathbf{z}$. Let ρ be a real-valued function on $\boldsymbol{\Omega}_\mathbf{z}$ and the parameter space, $\rho : \boldsymbol{\Theta} \times \boldsymbol{\Omega}_\mathbf{z} \to \mathbb{R}$. Then, an *M-estimator* is a solution of the minimization problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} \rho(\mathbf{z}_i, \boldsymbol{\theta}).$$

Using $\rho(\mathbf{z}, \boldsymbol{\theta}) = -\ln L_\mathbf{z}(\boldsymbol{\theta})$ yields the typical maximum likelihood estimator. If ρ is differentiable with respect to $\boldsymbol{\theta}$, then the first-order conditions that define the M-estimator are given by

$$\mathbf{0} = \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{z}_i, \boldsymbol{\theta}),$$

where $\boldsymbol{\psi}(\mathbf{z}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \rho(\mathbf{z}, \boldsymbol{\theta})$ is the gradient of ρ with respect to $\boldsymbol{\theta}$. These first-order conditions can be incorporated into GMM by dividing by n. Valid inference is only possible if the GMM regularity conditions or equivalent sufficient conditions that imply the same asymptotic results hold.

As a negative example, consider the function $\psi = \text{sign}(z - \theta)$, which leads to the sample median. This function is not differentiable in the classical sense, and the resulting estimator does not have a proper asymptotic distribution according to the asymptotic theory of M-estimation (Huber & Ronchetti, 2009, p.95). As a positive example, using the Huber loss as ρ does not conflict with the GMM regularity conditions and can therefore be implemented in GMM. See Corollary 3.5 of Huber and Ronchetti (2009) and the subsequent discussion for a more detailed elaboration on the underlying theory.

5.3 Incorporating external information

Imbens and Lancaster (1994) as well as Hellerstein and Imbens (1999) developed the idea of using moment equations to incorporate moment-type external values into the GMM. Suppose $E(y)$ is known to be 100. Then, this external value induces the population

moment equation $E(y) - 100 = 0$. As described in Example 1, it imposes constraints on the parameters of a linear model. This is despite the fact that it is information about a variable. The respective sample moment equation is $\bar{y} - 100 = 0$. This equation incorporates both a statistic derived from the new data set and the external value.

The fundamental idea is to add this external moment equation to the moment equations used to estimate the model, creating a combined set of moment equations. To fix the notation, let $\mathbf{h} : \mathbf{z} \to \mathbb{R}^{p_2}$ denote the external moment function with property $E(\mathbf{h}(\mathbf{z})) = \mathbf{0}$. Define $\overline{\mathbf{h}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\mathbf{z}_i)$. Then, the external (sample) moment equations are $\overline{\mathbf{h}} = \mathbf{0}$. Although $\mathbf{h}(\mathbf{z})$ contains the external values, they are constant and will thus be suppressed in the notation for now. This corresponds to the assumption of exact knowledge of the population value. First, the results for this case will be presented. As will be shown in the next section, these results can easily be extended to cases involving estimation and structural uncertainty.

It is assumed that $\mathbf{h}(\mathbf{z})$ does not depend on $\boldsymbol{\theta}$. Although this seems restrictive at first, note that only the specific parameter of the chosen model for the new data set is considered. If there is an external value of the regression coefficients of a linear model, $\boldsymbol{\beta}_{ex}$, then the expression $\frac{1}{n}\mathbf{X}^T(y - \mathbf{X}\boldsymbol{\beta}_{ex})$ is not a function of the parameter to be estimated for the new data set. Thus, it can be used as $\overline{\mathbf{h}}$.

Since $\mathbf{h}(\mathbf{z})$ is not a function of the parameter, the GMM regularity condition 8 is always satisfied, even if $\mathbf{h}(\mathbf{z})$ is not differentiable. For example, the external value of a median m_{ex} can be incorporated using the first-order condition for the M-estimator based on $\psi = \text{sign}(z - m_{ex})$. Although employing ψ poses problems when it is used to estimate a median, plugging in an external median turns it into a constant, which makes it differentiable with respect to the parameters.

There are multiple ways to incorporate parameter-specific external values. One approach is to formulate additional moment equations, as mentioned above. This may result in strong dependencies between moment equations. Another approach is to plug the external parameter values directly into the model's moment equations. Then, the values of the objective function can be derived, and a Sargan-Hansen test can be conducted to evaluate how well the previous parameter values fit the new data. An alternative approach to formulating moment equations is to express parameter-specific external information as constraints on the parameter space. However, constrained GMM estimation is beyond the scope of this thesis.

Let $\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})$ denote the $\mathbb{R}^{p_1}$-valued moment function for the statistical model of interest. This function induces $\overline{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{m}(\mathbf{z}_i, \boldsymbol{\theta})$, which constitutes the estimating equations of the model, $\overline{\mathbf{m}} = \mathbf{0}$. Then, the combined moment function is $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}) =$

$(\mathbf{m}(\mathbf{z},\boldsymbol{\theta})^T, \mathbf{h}(\mathbf{z})^T)^T$. The GMM procedure based on $\mathbf{g}(\mathbf{z},\boldsymbol{\theta})$ and the procedure's extensions in the next section are called an *externally informed GMM*. Note that the resulting system of moment equations is overidentified, regardless of whether $\overline{\mathbf{m}}$ is just-identified or overidentified. This justifies the use of GMM instead of the traditional method of moments. The remainder of this section discusses the analytical implications of using these combined moment equations, as they may lead to simplifications and new insights.

Lemma 1. [Separability](Jann, 2024) Let $\hat{\mathbf{W}} = \hat{\boldsymbol{\Omega}}^+$ be the Moore-Penrose inverse of the obvious estimator $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ and let $\mathbf{g}(\mathbf{z},\boldsymbol{\theta}) = (\mathbf{m}(\mathbf{z},\boldsymbol{\theta})^T, \mathbf{h}(\mathbf{z})^T)^T$. Then, $\hat{\boldsymbol{\Omega}}$ has the block form

$$\hat{\boldsymbol{\Omega}} = \begin{pmatrix} \hat{\boldsymbol{\Omega}}_m & \hat{\boldsymbol{\Omega}}_r^T \\ \hat{\boldsymbol{\Omega}}_r & \hat{\boldsymbol{\Omega}}_h \end{pmatrix},$$

with $\hat{\boldsymbol{\Omega}}_m \in \mathbb{R}^{p_1,p_1}$ and $\hat{\boldsymbol{\Omega}}_h \in \mathbb{R}^{p_2,p_2}$. If $\text{rank}(\hat{\boldsymbol{\Omega}}) = \text{rank}(\hat{\boldsymbol{\Omega}}_m)+\text{rank}(\hat{\boldsymbol{\Omega}}_h)$, it holds that

$$-\hat{Q}_n(\boldsymbol{\theta}) = (\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^+ \overline{\mathbf{h}})^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+ (\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^+ \overline{\mathbf{h}}) + \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^+ \overline{\mathbf{h}}.$$

Clearly, $\boldsymbol{\Omega}$ and $\hat{\boldsymbol{\Omega}}$ can both be represented by the same block form. The condition $\text{rank}(\hat{\boldsymbol{\Omega}}) = \text{rank}(\hat{\boldsymbol{\Omega}}_m)+\text{rank}(\hat{\boldsymbol{\Omega}}_h)$ is particularly satisfied if $\hat{\boldsymbol{\Omega}}$ is non-singular. The proof of Lemma 1 is a straightforward application of the Schur-inversion formula, as outlined by Puntanen et al. (2011, p. 294). This formula lays the foundation for the results in this chapter. Applying Lemma 1 to the first-order conditions of the externally informed GMM reduces them to

$$\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^+ \overline{\mathbf{h}} = \mathbf{0},$$

as Jann and Spiess (2025) demonstrated.

These simplified first-order conditions allow for a conditional interpretation that may be interesting from a Bayesian perspective. Assuming that $\mathbf{g}(\mathbf{z},\boldsymbol{\theta}_0)$ has a multivariate normal distribution and replacing $\hat{\boldsymbol{\Omega}}$ with $\boldsymbol{\Omega}$, the left side of the simplified first-order conditions becomes an estimate of the conditional expectation of $\mathbf{m}(\mathbf{z},\boldsymbol{\theta}_0)$ given $\mathbf{h}(\mathbf{z}) = \mathbf{0}$ (Puntanen et al., 2011, p. 197). Due to the GMM regularity conditions, the central limit theorem applies to $(\overline{\mathbf{m}}^T, \overline{\mathbf{h}}^T)^T$, allowing for the same interpretation in an asymptotic sense for $\overline{\mathbf{m}}$ and $\overline{\mathbf{h}}$. Therefore, the simplified first-order conditions will be referred to as *conditional moment equations* from this point forward.

According to Lemma 1, the GMM objective function can be separated into two parts: The expression $\overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^+ \overline{\mathbf{h}}$ measures the distance between the external values and the corre-

sponding values computed based on the new data. It is not a function of the parameter. The other part can be viewed as an objective function resulting from the conditional moment equations alone. Now, the heuristic of external information leading to a lower variance of the estimator can be formalized in the following.

Corollary 1. (Jann & Spiess, 2024) Assume that $\hat{\boldsymbol{\theta}}_M$ is the GMM estimator based on the model moment equations alone, ignoring the external moment equations. Suppose that $\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ are of the same dimension. Using the prerequisites of Lemma 1,

$$\begin{aligned}\operatorname{Var}(\hat{\boldsymbol{\theta}}_{ex}) &= \frac{1}{n}\big(E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0))^T\big)^{-1}\boldsymbol{\Omega}_m - \boldsymbol{\Omega}_r^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Omega}_r\big(E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0))\big)^{-1} \\ &= \operatorname{Var}(\hat{\boldsymbol{\theta}}_M) - \frac{1}{n}\big(E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0))^T\big)^{-1}\boldsymbol{\Omega}_r^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Omega}_r\big(E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0))\big)^{-1}.\end{aligned}$$

The corresponding variance estimator is

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}_{ex}) = \widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}_M) - \frac{1}{n}\big(\frac{1}{n}\sum_{i=1}^{n}(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_{ex}))^T\big)^{-1}\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^{+}\hat{\boldsymbol{\Omega}}_r\big(\frac{1}{n}\sum_{i=1}^{n}(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_{ex}))\big)^{-1}.$$

Assuming multivariate normality of $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)$, the expression $\boldsymbol{\Omega}_m - \boldsymbol{\Omega}_r^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Omega}_r$ in Corollary 1 is the conditional population variance matrix of $\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0)$ given $\mathbf{h}(\mathbf{z}) = \mathbf{0}$ (Puntanen et al., 2011, p. 197). Taken together, using additional moment equations independent of the parameter is consistent with using conditional moment equations. Heuristically, this reflects an update to the model's moment equations by conditioning them on external values under the assumption of multivariate normality.

The subtrahends in the variance formulas in Corollary 1 are positive semi-definite, so their diagonal elements are non-negative (Jann & Spiess, 2024). Therefore, the diagonal elements of $\operatorname{Var}(\hat{\boldsymbol{\theta}}_{ex})$ are equal to or less than the respective diagonal elements of $\operatorname{Var}(\hat{\boldsymbol{\theta}}_M)$. These diagonal elements represent the variances of the individual entries of $\hat{\boldsymbol{\theta}}_{ex}$ and $\hat{\boldsymbol{\theta}}_M$. Consequently, using an externally informed GMM will not increase the variances compared to an uninformed GMM.

Furthermore, proper variance reduction can only occur if the subtrahend is not the null matrix. Therefore, a necessary condition for reducing sample variance is $\hat{\boldsymbol{\Omega}}_r \neq \mathbf{0}$ (Jann & Spiess, 2024). Similarly, a necessary condition to reduce the expected variance is $\boldsymbol{\Omega}_r \neq \mathbf{0}$. This means that variance reduction only occurs if the external moment functions and the model moment functions are linearly dependent on each other.

An important counterexample involves external information about the independent variables in linear regression models. In this case, the error term, ϵ, and the independent variables, $\mathbf{x}$, are assumed to be independent. As discussed in Section 5.1, this allows for

the use of $E(\mathbf{x}\epsilon) = \mathbf{0}$ as population moment equations for the model. If only external information about $\mathbf{x}$ is available, the external moment function is just a function of $\mathbf{x}$, denoted by $\mathbf{h}(\mathbf{x})$. The independence of ϵ and $\mathbf{x}$ leads to $\boldsymbol{\Omega}_r = E(\mathbf{h}(\mathbf{x})\mathbf{x}^T\epsilon) = \mathbf{0}$, so no variance reduction can be expected (Jann & Spiess, 2024).

However, Jann and Spiess (2024) provide an example in which combining external information about independent variables with other external information reduced variance more than using the latter information alone. In contrast, using external information about the dependent variable of a linear regression model generally results in reduced variance because the dependent variable is linked to the error term. This includes covariances between the dependent and independent variables. According to simulation studies, using these covariances, or quantities that include them, in an externally informed GMM yields the lowest variance estimates in linear models, as opposed to other external values (Jann & Spiess, 2024).

The theoretical considerations made so far can be used to simplify two of the test statistics presented in Section 5.1.

Theorem 1. (Jann, 2024; Jann & Spiess, 2025) Suppose that the premises of Lemma 1 hold. Assume that $\hat{\boldsymbol{\Omega}}_h$ is not a function of $\boldsymbol{\theta}$. If a $\boldsymbol{\theta}_h \in \boldsymbol{\Theta}$ with the property $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}} = \mathbf{0}$ exists, then

$$-\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex}) = \overline{\mathbf{h}}^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}}.$$

As a consequence, the Sargan-Hansen and the difference test statistic simplify to

$$SH = n\overline{\mathbf{h}}^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}}$$

and

$$D = n(\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}})^T(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+(\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}}),$$

where D is evaluated at the restricted GMM estimator $\hat{\boldsymbol{\theta}}_r$.

As discussed in Section 5.1, the Sargan-Hansen test is a test of compatibility between the moment equations and the data. Incompatibility may arise from either the external or the model moment equations. However, Theorem 1 shows that the model moment equations cancel out, meaning the Sargan-Hansen test can be viewed as a test of the fit of external values and new data (Jann, 2024).

In case $\mathbf{h}(\mathbf{z})$ is linear in the external values, there are two choices for estimator $\hat{\boldsymbol{\Omega}}_h$ that behave differently for misspecified external values. The first is the obvious esti-

mator $\hat{\mathbf{\Sigma}}_h = \frac{1}{n}\Sigma_{i=1}^{n}\mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^T$. The second is the sample covariance matrix, $\hat{\mathbf{S}}_h = \frac{1}{n-1}\Sigma_{i=1}^{n}(\mathbf{h}(\mathbf{z}_i)-\overline{\mathbf{h}})(\mathbf{h}(\mathbf{z}_i)-\overline{\mathbf{h}})^T$. The external values cancel out in the latter, so it is not a function of the external values and is thus unaffected by misspecification. In contrast, the estimator $\hat{\mathbf{\Sigma}}_h$ is a function of the external values and will be affected by misspecification. One advantage of $\hat{\mathbf{\Sigma}}_h$ is that it can be used in the non-linear case, whereas $\hat{\mathbf{S}}_h$ is only valid in the case where $\mathbf{h}(\mathbf{z})$ is linear. Further details on the effects of using the different estimators of $\mathbf{\Omega}_h$ will be discussed in Chapter 6 and in the attached papers.

Now, only $\hat{\mathbf{\Omega}}_r$ remains to be specified. Two possible approaches exist. The first approach is to derive the asymptotic matrix $\mathbf{\Omega}_r$ analytically and then to use a consistent estimator for $\mathbf{\Omega}_r$. Jann and Spiess (2024) applied this approach to linear models. However, this approach is only feasible for simple models for which analytical results can be obtained.

The second approach is to use the obvious estimator, $\hat{\mathbf{\Omega}}_r = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}(\mathbf{z}_i)\mathbf{m}(\mathbf{z}_i,\hat{\boldsymbol{\theta}}_{ex})^T$. Unless otherwise stated, this approach will be employed for the remainder of this thesis. Using the obvious estimator $\hat{\mathbf{\Omega}}_r$, the conditional moment equations become

$$\mathbf{0} = \overline{\mathbf{m}} - \hat{\mathbf{\Omega}}_r^T\hat{\mathbf{\Omega}}_h^+\overline{\mathbf{h}} = \frac{1}{n}\sum_{i=1}^{n}(1-\mathbf{h}(\mathbf{z}_i)^T\hat{\mathbf{\Omega}}_h^+\overline{\mathbf{h}})\mathbf{m}(\mathbf{z}_i,\boldsymbol{\theta}). \tag{5.1}$$

This provides an alternative interpretation of the externally informed GMM as a weighting procedure, in which the model moment function for each unit is weighted by $(1-\mathbf{h}(\mathbf{z}_i)^T\hat{\mathbf{\Omega}}_h^+\overline{\mathbf{h}})$. This means that the model moment functions of units with values close to the external values are weighted differently than those of units with values far from the external values. The weights do not behave like frequencies. They sum up to $n-\sum_{i=1}^{n}\mathbf{h}(\mathbf{z}_i)^T\hat{\mathbf{\Omega}}_h^+\overline{\mathbf{h}}$ and not one. Furthermore, there may be negative weights. For example, consider the two-unit sample $h(z_1)=3$ and $h(z_2)=5$. In this case, the weights for the second unit are negative for both $\hat{\mathbf{\Sigma}}_h$ and $\hat{\mathbf{S}}_h$.

If $\mathbf{m}(\mathbf{z},\boldsymbol{\theta})$ is the gradient of a known objective function $M(\mathbf{z},\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, then Equation (5.1) is the first-order condition to optimize the objective function $\overline{M}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(1-\mathbf{h}(\mathbf{z}_i)^T\hat{\mathbf{\Omega}}_h^+\overline{\mathbf{h}})M(\mathbf{z}_i,\boldsymbol{\theta})$. For example, the log-likelihoods of structural equation models and two-level mixed linear models in Section 5.2 are objective functions. Their gradients can be extensive and must be calculated anew for different scenarios. In these cases, it may be easier to state $\overline{M}(\boldsymbol{\theta})$ than to state the first-order conditions based on equation (5.1). The objective function $\overline{M}(\boldsymbol{\theta})$ can be optimized using numerical derivatives, which eliminates the need to derive analytical first-order conditions.

In the field of psychology, the software environment R is commonly used to conduct statistical analyses (R Core Team, 2025). In R the function `optim` implements opti-

mization algorithms based on numerical derivatives. Therefore, it is possible to avoid cumbersome derivations when estimating the parameters of externally informed versions of structural equation models or two-level mixed linear models. Of course, this procedure's numerical stability should be evaluated to assess any potential divergence or error propagation in the results.

The objective of finding a root of the first-order conditions (5.1) can be achieved using the functions `fsolve` or `newtonsys`, which implement the Newton-Raphson algorithm and others, in the R package `pracma` (Borchers, 2023). Although both functions can calculate numerical derivatives, they can also be provided with analytical derivatives in the form of the Jacobian matrix. Since the weights are not a function of the parameter, the Jacobian matrix corresponding to the first-order conditions is $\frac{1}{n}\sum_{i=1}^{n}(1 - \mathbf{h}(\mathbf{z}_i)^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}})\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}_i, \boldsymbol{\theta})$. Thus, computing this Jacobian matrix does not require much knowledge beyond what is needed to compute the Jacobian matrix based on the model moment functions alone.

An R function that calculates the externally informed GMM estimator using the aforementioned numerical optimization and root-finding approaches has been developed during this PhD project and will soon be publicly available under the link https://github.com/MartinJann/exgmm.

5.4 Reflecting structural and estimation uncertainty

Thus far, the prevailing assumption has been that the external moment equations are correctly specified, i.e., that the external values are precisely equal to the corresponding population moments. However, in almost all cases, the external values are estimates. To reflect this estimation uncertainty, the external values will be treated as a random variable, denoted by $\mathbf{e}$, and not as a constant. Since $\mathbf{e}$ is derived from an external sample, it is assumed that it is independent of the new data set. Furthermore, it is assumed that the population variance matrix of $\mathbf{e}$ exists and is finite.

First, consider the linear case, in which $\overline{\mathbf{h}}(\mathbf{z}, \mathbf{e}) = \hat{\mathbf{h}}(\mathbf{z}) - \mathbf{e}$, where $\hat{\mathbf{h}}(\mathbf{z})$ is an unbiased estimate of the corresponding population moment $E(\mathbf{e})$ in the new data set. Important examples of this case include means, variances, covariances, and estimates of linear regression parameters. However, it excludes estimators that are consistent yet not unbiased, such as correlations and most estimates of non-linear model parameters. Now, inference is possible by relying on either the GMM regularity conditions or another set of sufficient conditions for consistency and asymptotic normality.

Given the independence of the external sample and the new data, the only change is

in $\hat{\boldsymbol{\Omega}}_h$. As first shown by (Jann, 2024), the new obvious estimator of $\boldsymbol{\Omega}_h$ is

$$\hat{\boldsymbol{\mathcal{S}}}_h = \hat{\boldsymbol{\Sigma}}_h + \widehat{\mathrm{Var}}(\mathbf{e}),$$

where $\widehat{\mathrm{Var}}(\mathbf{e})$ is a consistent estimate of the population variance matrix of $\mathbf{e}$. Thus, to fully reflect estimation uncertainty in the linear case, researchers only need to add a consistent variance estimate for the external values to the obvious estimator $\hat{\boldsymbol{\Sigma}}_h$.

However, $\widehat{\mathrm{Var}}(\mathbf{e})$ must be derived from the external sample. See Section 3.2 for a discussion of this issue. Note that since $\mathbf{e}$ is an unbiased estimator, its variance will asymptotically shrink to zero in most practical cases. Once again, consider the example $h(z) = \bar{z} - 100$, where 100 is now correctly considered a mean rather than the actual population value. If an external source reports a sample size of 144 and a variance of 81, then $\widehat{\mathrm{Var}}(\mathbf{e}) = 81/144 = 0.25$ is a reasonable estimate of the variance of $\mathbf{e}$.

Even for a random $\mathbf{e}$, GMM estimation requires that $E(\hat{\mathbf{h}}(\mathbf{z})) = E(\mathbf{e})$. Therefore, structural uncertainty is not yet covered. This can be achieved by defining k random variables $\mathbf{e}_1, \ldots, \mathbf{e}_k$, which reflect k external sources. It is reasonable to assume that these variables are independent of the new data set. Let $\mathbf{e}_0$ denote the corresponding population values of the new data set, i.e., the correctly specified external values.

The results from Section 3.4 now apply. Instead of using k external values pointwise, it is more reasonable to use the external interval $\mathbf{I}_{ex} = [\min_{j=1,\ldots,k} \mathbf{e}_j; \max_{j=1,\ldots,k} \mathbf{e}_j]$, where the minima and maxima are taken elementwise. The key assumption for the validity of the following arguments is that $\mathbf{e}_0 \in \mathbf{I}_0 = [\min_{j=1,\ldots,k} E(\mathbf{e}_j); \max_{j=1,\ldots,k} E(\mathbf{e}_j)]$. Due to the optimizer's curse, $\mathbf{I}_{ex}$ is a reasonable, conservative estimator of $\mathbf{I}_0$.

Furthermore, a combination rule for the external variance estimates must be chosen. The combined variance estimate is then used to compute $\hat{\boldsymbol{\mathcal{S}}}_h$. For example, the minimum variance estimate could be selected. Unfortunately, this is not possible with variance matrix estimates. Although the Löwner order can be used to compare matrices, comparability between matrices with respect to this order is not guaranteed (Puntanen et al., 2011, p. 13). If there are few cases, it may be reasonable to conduct the analysis using each incomparable matrix estimate and then aggregate the results. However, if there are many incomparable matrix estimates, it may be more appropriate to aggregate them differently. For example, since estimation uncertainty often depends on sample size, it is reasonable to choose the matrix estimate corresponding to the smallest external sample.

The fundamental idea to reflect structural and estimation uncertainty is that GMM estimation for each value in $\mathbf{I}_{ex}$ and for each selected (combined) external variance estimate should be performed. Due to the key assumption that $\mathbf{e}_0$ is bounded by the

expected values of $\mathbf{e}_1, \ldots, \mathbf{e}_k$, there is an unbiased point estimator of $\mathbf{e}_0$ in $\mathbf{I}_{ex}$. For example, it can be expressed as a convex combination of $\mathbf{e}_1, \ldots, \mathbf{e}_k$. The GMM estimator based on this unbiased estimator of $\mathbf{e}_0$ is correctly specified.

This is an example of "epistemic" uncertainty. Although there is a correctly specified estimator, it is unknown which one it is. Therefore, all values should be treated as equally plausible. The correctly specified GMM estimator is supposed to fulfill the GMM regularity conditions, making it consistent and asymptotically normally distributed. The other elements of $\mathbf{I}_{ex}$ correspond to misspecified GMM estimators if their expected values are not equal to $\mathbf{e}_0$. These estimators do not satisfy the GMM regularity conditions.

Ideally, misspecified estimators would form an "envelope" around the correctly specified estimator, similar to the classical neighborhood models that underlie robust statistics (Huber & Ronchetti, 2009). To rigorously demonstrate this, it is necessary to elaborate on the asymptotic distributions in the misspecified cases. Combined with the correctly specified estimator's asymptotic normal distribution, they form a credal set on the basis of which an F-probability can be constructed, enhancing distributional robustness. Typical estimation or inference concepts, such as confidence intervals or significance tests can then be extended via conservative principles or decision rules. This will be covered in the following subsections.

Now, some comments will be given on how to address estimation uncertainty when working with consistent yet biased estimators. Often, external variance matrices can be assumed to converge in probability to the null matrix. Conversely, if structural uncertainty is present, $\mathbf{I}_{ex}$ converges in probability to a constant interval. Therefore, estimation uncertainty with consistent estimators can be incorporated using the linear moment equations approach for unbiased estimators described above. This is possible provided that the external sample sizes are large enough so that the biases in the external values are negligible and $\mathbf{e}_0$ remains within $\mathbf{I}_0$. For an empirical indicator supporting this claim, see the simulation studies of Jann and Spiess (2024), which used an external interval of correlations without adding an external variance estimate. In the analyzed scenarios, the coverage rates were nominal even though the correlation estimator was consistent albeit not unbiased.

5.4.1 Estimation and prediction

All the results in this section are based on the conservative principle that each element of $\mathbf{I}_{ex}$ is potentially correctly specified, and no element is preferred over another. First, the generalization of point estimates and point predictions to the case of $\mathbf{I}_{ex}$ will be discussed. To define a GMM estimator based on $\mathbf{I}_{ex}$, the basic idea is to solve the

first-order conditions for each value in $\mathbf{I}_{ex}$ and for each selected external variance.

In practice, this computation may only be possible at certain grid points of $\mathbf{I}_{ex}$ because the first-order conditions can only be solved numerically. Once the results for each of the grid points are obtained, the elementwise minima and maxima over all results are taken to yield an interval. The resulting interval constitutes the GMM estimator based on $\mathbf{I}_{ex}$. Of course, the resulting interval is only an approximation, and the fewer grid points there are, the less precise it will be. For some models, analytical solutions exist for the first-order conditions, which substantially reduces the computational burden.

One of the main goals of the papers attached to this thesis was to address some of these cases and develop simplifications through analytical solutions. Even with analytical solutions, describing the resulting set of point estimates may be non-trivial. For example, determining the minimum or maximum of a statistic over $\mathbf{I}_{ex}$ is generally a non-convex optimization problem (Jann, 2023).If the variance or covariance estimates of the GMM estimator based on $\mathbf{I}_{ex}$ are of interest, the same procedure can be applied to the variance (matrix) estimator. The variance decomposition shown in Corollary 1 provides an analytical solution in any case. A similar procedure can be used to approximate the set of predictions of the conditional mean of a dependent variable based on a model given its independent variables. As described in Section 5.1, the model specifies $E(y|\mathbf{x}) = f(\mathbf{x}^T\boldsymbol{\theta})$, and the point prediction for a given $\mathbf{x}_p$ is $f(\mathbf{x}_p^T\hat{\boldsymbol{\theta}}_{ex})$. In summary, using $\mathbf{I}_{ex}$ instead of external values transforms point estimates and predictions into sets.

To determine the asymptotic properties of these set estimators, the asymptotic distributions have to be stated for each case. For the correctly specified estimate, the typical GMM asymptotics apply, as discussed in Section 5.1. Under weak additional assumptions, the remaining misspecified estimators asymptotically follow a normal distribution, albeit with bias and altered variance estimators (Hall & Inoue, 2003).

An important observation is that the bias not necessarily converges to zero as the sample size increases. This implies that using a misspecified estimator alone results in an externally informed estimator that is not a consistent estimator of the true parameter value. This can be seen in simulation studies such as those conducted by Jann and Spiess (2024). These studies demonstrate that for misspecified external values, the proportion of confidence intervals covering the true value of the parameters decreases with sample size. When considered as a whole, the set of estimators corresponds to an asymptotically valid credal set of normal distributions that induces an F-probability. In the misspecified case, the estimated variances should be modified to provide better estimates of the true variances, as discussed by Hall and Inoue (2003). However, to achieve the goal of providing a numerically feasible "envelope" around a consistent estimate, the approach

described here should be sufficient. Thus, the variance estimators derived by Hall and Inoue (2003) that are more computationally intensive are omitted, although they are of interest for further research. Nevertheless, using an F-probability provides greater robustness than using a single estimator and its resulting probability distribution, even if the single estimator is correctly specified.

The subsequent discussion will address the extension of confidence intervals to the presence of $\mathbf{I}_{ex}$. The fundamental concept is to use the set union of all confidence intervals derived from each element of $\mathbf{I}_{ex}$ and each selected external variance. This approach was utilized by Walter and Augustin (2009) in the context of Bayesian highest density intervals. Formally, let $CI_{1-\alpha}(\mathbf{e})$ be the $(1-\alpha)\%$ confidence interval for estimation or prediction based on the externally informed GMM estimator given the external values $\mathbf{e}$, then the $(1-\alpha)\%$ confidence union is $\bigcup_{\mathbf{e}\in\mathbf{I}_{ex}} CI_{1-\alpha}(\mathbf{e})$. See Section 5.1 for the formula for one-dimensional confidence intervals based on the delta method.

The $(1-\alpha)\%$ confidence union is valid in the sense that it covers the true parameter value with a probability of at least $1-\alpha$, provided that the GMM regularity conditions hold for the correctly specified estimator and the external values constituting the latter are included in $\mathbf{I}_{ex}$. The reason for this is that the $CI_{1-\alpha}$ for the correctly specified estimator is already valid in the aforementioned sense and is a subset of the confidence union. Again, the underlying asymptotic construct is the F-probability previously described, which may increase robustness. Therefore, the $(1-\alpha)\%$ confidence union also covers $(1-\alpha)\%$ confidence intervals based on distributions beyond the normal distribution, such as slightly skewed or multimodal distributions.

There is an interesting relationship between the reflection of uncertainties and the variance-reducing property of external information. In almost all cases, the confidence union is substantially broader than the confidence interval for the correctly specified externally informed GMM estimator. Thus, greater robustness comes at the cost of reduced precision. However, using an externally informed GMM estimator may result in lower variance than using an uninformed GMM estimator. Since both estimators have an asymptotic normal distribution, the confidence interval for the externally informed GMM estimator is narrower. Therefore, using external values improves precision.

When combined, the broadness of $\mathbf{I}_{ex}$ is counterbalanced by the precision gained through variance reduction. If the variance of the correctly specified estimator is low enough, the resulting confidence union may still be narrower than the uninformed confidence interval. This has been demonstrated by Jann and Spiess (2024) through simulation studies. Even if the confidence union is equal or slightly broader, there is a net gain in distributional robustness compared to the uninformed confidence interval.

5.4.2 Hypothesis testing

Significance tests are a common method of frequentist statistical inference in psychology to test hypotheses about parameters (Rasch et al., 2011). Therefore, it is important to provide an extension of significance tests to reflect the structural uncertainty of external information. A discussion of a naïve extension approach and its limitations will serve as the starting point for motivating more sophisticated developments. Consider a specific test from the GMM framework, such as the Sargan-Hansen or the Wald test. Sections 5.1 and 5.3 discuss how these tests can be performed when external values are present.

Following the conservative principle, these tests are extended to cases involving an external interval $\mathbf{I}_{ex}$ by calculating test statistics for each value in $\mathbf{I}_{ex}$ and each external variance selected. The result is a set of test statistics. The distribution of the test statistic under the null hypothesis is the same for each value in $\mathbf{I}_{ex}$, provided the value is correctly specified. Therefore, a single critical value can be compared to the set of test statistics. Using the distribution under the null hypothesis, the set of test statistics can be translated into a set of p-values and compared to the significance level, α.

There are two scenarios in which the test decision becomes seemingly obvious. If all p-values are below α, the null hypothesis will be rejected for each value in $\mathbf{I}_{ex}$. Thus, rejecting the null hypothesis overall is appropriate. Conversely, if all p-values are greater than α, the null hypothesis should be maintained because all possible tests would also maintain it.

As for the third case, in which some p-values are larger than α and some are smaller, it is not possible to make an overall decision without additional decision rules. However, this should be approached with caution. Suppose that the set of p-values forms an interval. Consider the rule of rejecting the null hypothesis if the center of the interval is below α. The rule encompasses the two obvious cases, reducing the entire test to a procedure solely based on the center of the p-value interval. This amounts to ignoring structural uncertainty, since any external interval that yields the same midpoint of the p-value interval would result in the same test decision, regardless of its breadth. However, the structural uncertainty reflected by a narrower or broader external interval differs.

Therefore, it may be better to neither reject nor maintain the null hypothesis in the third case. This third decision could therefore be described as "undecided". The idea is similar to the "decision withheld" scenario of generalized Bayes factors based on imprecise specifications (Schwaferts, 2022, p. 44). However, there is an important drawback to using this third decision in the frequentist setting. Consider the Sargan-Hansen test for the fit of external values and new data. If external information is provided by an external interval $\mathbf{I}_{ex}$, a reasonable extension would be to test the null hypothesis

that any value in $\mathbf{I}_{ex}$ fits the new data (Jann, 2023).

According to Theorem 1, for linear external moment functions, the Sargan-Hansen test for the fit of external values and new data is equivalent to a Wald test. The Wald test is consistent, with an asymptotic power of 1 for fixed alternatives (Cameron & Trivedi, 2005, p. 248). Thus, asymptotically, the Sargan-Hansen test will almost surely reject the fit of each misspecified value in $\mathbf{I}_{ex}$ to the new data set. However, it will maintain the fit for the correctly specified external values with probability α.

Taken together, the test decision for a naïvely extended Sargan-Hansen test will probably be "undecided" asymptotically if the correctly specified value falls within $\mathbf{I}_{ex}$. Consequently, this test is usually unable to uphold a true null hypothesis asymptotically. Any external interval that is not a singleton is susceptible to this effect.

One way to avoid this asymptotic indecisiveness is to establish reasonable decision rules for the third case. Note that structural uncertainty results in a scenario in which an F-probability is present. Huntley et al. (2014) provides an introduction to decision-making under imprecise probabilities, which underlies the following arguments. The arguments are informally presented to illustrate the basic idea. A more rigorous treatment, including proofs, is given by Jann (2024).

The procedure begins with the establishment of a credal set $\mathcal{M}_0$ of asymptotic distributions of the set of test statistics under the null hypothesis. As discussed in Section 5.1, the test statistics SH, W, LM, and D are asymptotically χ^2_k-distributed under the null hypothesis for correctly specified external values (for SH set $k = q - p$). However, for misspecified external values, the test statistics generally diverge to ∞ (Cameron & Trivedi, 2005, p. 248).

This can be reflected by including a probability distribution in the credal set that shifts all mass to infinity. Using the extended real line, its CDF can be denoted as 1_∞, the indicator function at ∞. In addition to these two extremes, local alternatives around $\mathbf{e}_0$, defined by $\mathbf{e}_n = \mathbf{e}_0 + \boldsymbol{\delta}/n$, exist. The test statistics for these local alternatives are asymptotically distributed according to a non-central χ^2_k-distribution with k degrees of freedom and a non-centrality parameter λ, denoted by $\chi^2_k(\lambda)$.

Therefore, the complete asymptotic credal set $\mathcal{M}_0$ consists of $\chi^2_k(\lambda)$ for $0 \leq \lambda < \infty$ and 1_∞. Based on this credal set, the six choice functions presented by Huntley et al. (2014) can be used to develop decision rules for the third case by selecting optimal test statistics. According to Proposition 2 by Jann (2024), the presence of 1_∞ causes four out of the six choice functions to identify all test statistics as optimal. This does not result in any changes to the "undecided" scenario. The remaining two functions select the lowest test statistic, thus collapsing to the same function. This collapsed choice

function is called Γ-*maximin*, and leads to the following significance test for a general null hypothesis $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$, where $\boldsymbol{\Theta}_0$ is a predefined set.

Definition 8. (Jann, 2024) Let $T(\boldsymbol{\theta}, \mathbf{e})$ be a test statistic that is a function of a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and the external values $\mathbf{e} \in \mathbf{I}_{ex}$. Let $\mathcal{T}$ be the set of observed test statistics based on $\boldsymbol{\Theta}_0$ and $\mathbf{I}_{ex}$, where $\underline{t}$ denotes its infimum. Let $\mathcal{M}$ be a credal set of possible distributions of the test statistics under the null hypothesis and $\underline{P}$ be the lower bound of the F-probability based on $\mathcal{M}$. For the null hypothesis $H_0 : \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$, a $\Gamma-$*maximin test with significance level* $\alpha \in (0; 1)$ is as follows:

$$\begin{aligned}&\text{If } \underline{P}(T > \underline{t}) < \alpha, \text{ then reject } H_0 : \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0,\\ &\text{else maintain } H_0 : \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0.\end{aligned}$$

The Γ-maximin test corresponds to the simple idea of rejecting the null hypothesis only if it would be rejected by all possible test statistics based on $\mathbf{I}_{ex}$, and maintaining it otherwise. The fact is that if the lower probability of $\{T > \underline{t}\}$ is below α, then the lower probability of each corresponding event for an observed test statistic is also below α, since $\underline{t}$ corresponds to the largest event. The following property establishes that the Γ-maximin test is a valid significance test that controls the probability of a type I error.

Definition 9. (Jann, 2024) Let $\mathbf{e}_0$ denote the correctly specified external values. A $\Gamma-$maximin test with significance level α has α-*level under the (asymptotic) distribution of* $T(\boldsymbol{\theta}_0, \mathbf{e}_0)$, if (asymptotically)

$$P_{T(\boldsymbol{\theta}_0, \mathbf{e}_0)}(\underline{T} > t_\alpha) \leq P_{T(\boldsymbol{\theta}_0, \mathbf{e}_0)}(T(\boldsymbol{\theta}_0, \mathbf{e}_0) > t_\alpha) \leq \alpha$$

holds, given the null hypothesis is true, where $\underline{T} = \inf_{\boldsymbol{\Theta}_0, \mathbf{I}_{ex}} T(\boldsymbol{\theta}, \mathbf{e})$ and t_α is the upper $1 - \alpha$ quantile of a distribution in $\mathcal{M}$ that constitutes the lower probability $\underline{P}$ at the event $\{T > \underline{t}\}$.

The right inequality in Definition 9 indicates that the test based on the correctly specified test statistic $T(\boldsymbol{\theta}_0, \mathbf{e}_0)$ is valid. The event $\{\underline{T} > t_\alpha\}$ is equal to the event that the Γ-maximin test rejects the null hypothesis. Given the null hypothesis and correctly specified external information, the probability of $\{\underline{T} > t_\alpha\}$ under the true distribution (i.e., the probability of type I errors) is bounded by α. There is a simple criterion under which a Γ-maximin test has significance level α under the (asymptotic) distribution of $T(\boldsymbol{\theta}_0, \mathbf{e}_0)$. This criterion generalizes Theorem 2 from Jann (2024) using the same proof.

Theorem 2. Suppose that the credal set $\mathcal{M}$ contains the (asymptotic) distribution of $T(\boldsymbol{\theta}_0, \mathbf{e}_0)$, and that this distribution yields the lower probability of the event $\{T(\boldsymbol{\theta}_0, \mathbf{e}_0) >$

$t_\alpha\}$, i.e., $\underline{P}(T(\boldsymbol{\theta}_0, \mathbf{e}_0) > t_\alpha) = P_{T(\boldsymbol{\theta}_0, \mathbf{e}_0)}(T(\boldsymbol{\theta}_0, \mathbf{e}_0) > t_\alpha)$. Then, a $\Gamma-$maximin test with significance level α has α-level under the (asymptotic) distribution of $T(\boldsymbol{\theta}_0, \mathbf{e}_0)$.

Now, consider the credal set $\mathcal{M}_0$ for the test statistics SH, W, LM and D. Since the $\chi^2_k(\lambda)$-distributions are stochastically ordered in λ, the lower probability of the credal set $\mathcal{M}_0$ at the events $\{T > t\}$ for all possible t is the central χ^2_k-distribution (Ghosh, 1973). Because this is also the asymptotic distribution of the correctly specified test statistic under the null hypothesis, $T(\boldsymbol{\theta}_0, \mathbf{e}_0)$, Theorem 2 can be applied. Therefore, significance tests based on the test statistics SH, W, LM and D can be extended into valid Γ-maximin tests. In practice, the minimum test statistic has to be computed and compared to a critical value from a χ^2_k-distribution, or its p-value has to be calculated based on this distribution. Note that for the Γ-maximin Sargan-Hansen test as derived by Jann (2024), the expected external values are the parameter, i.e., $\boldsymbol{\theta} = E(\mathbf{e})$. The Γ-maximin Sargan-Hansen test evaluates the null hypothesis that $\mathbf{I}_{ex}$ contains the correctly specified external values, resolving the aforementioned asymptotic indecisiveness.

At a first glance, the ability to freely specify $\boldsymbol{\Theta}_0$ seems general enough to include one-sided tests. However, the considered test statistics are quadratic in nature and inconclusive in terms of the sign of the deviations from the null hypothesis. A better approach would be constructing a test based on the asymptotic normality of the GMM estimator and extend it to a Γ-maximin test. The null hypothesis for a left-sided test is $f(\boldsymbol{\theta}_0) \geq 0$, and for a right-sided test $f(\boldsymbol{\theta}_0) \leq 0$. As discussed in Section 5.1, the corresponding test statistic z_t asymptotically follows a standard normal distribution under the null hypothesis. If the external values are misspecified, then $f(\hat{\boldsymbol{\theta}}_{ex})$ is asymptotically "distributed" as $1_{-\infty}$ or 1_{∞}. For local alternatives, the test statistic is distributed as $N(\lambda, 1)$ with $\lambda \in \mathbb{R}$, since it is the (signed) square root of the Wald test statistic, which is $\chi^2_1(\lambda^2)$-distributed. To apply Theorem 2, it is necessary to construct the credal set so that $N(0, 1)$ yields the lower probabilities for the events $\{z_l < z\}$ for each possible z. Consequently, the credal set should only include $N(\lambda, 1)$ for $\lambda \leq 0$ and $1_{-\infty}$.

Using $N(\lambda, 1)$ for $\lambda > 0$ would lead to the application of distributions that would reject the null hypothesis in cases when it is valid. For example, $N(3, 1)$ might reject a test statistic value of 0, however, this value is perfectly consistent with the nominal null hypothesis. To avoid this issue, only the $N(0, 1)$-distribution of the null hypothesis boundary should be included, rather than distributions consistent with the interior of the null hypothesis. Including distributions in a credal set ensures that they will be considered among the calculation of possible p-values. However, if they correspond to a different, more restrictive null hypothesis, the Γ-maximin test will test that hypothesis, which is easier to reject than the nominal null hypothesis.

6 Contributions

The externally informed GMM approach was successively developed over the course of this PhD project and the following four papers. Three of the papers have been published as of now (available via open access online, links can be found in the respective subsection), and one is currently under review (planned to be open access). Aside from theory development, the papers provide analytical formulas for estimators and test statistics in important use cases. They also present simulation studies that analyze the behavior of the externally informed GMM approach in small samples across various scenarios. The cases in question include the (Γ-maximin) Sargan-Hansen test, as well as linear models and generalized linear models with repeated measures. Generalized linear models with repeated measures are estimated using the GEE approach. All papers are provided in full text in appendices A – D (pp. 82 – 209). The individual contributions of the papers are summarized below.

6.1 Paper 1: Coherence of external information and data

Jann, M. (2023). Testing the coherence of data and external intervals via an imprecise Sargan-Hansen test. In E. Miranda, I. Montes, E. Quaeghebeur, & B. Vantaggi (Eds.), *Proceedings of the thirteenth international symposium on imprecise probability: Theories and applications* (pp. 249–258, Vol. 215). PMLR. https://proceedings.mlr.press/v215/jann23a.html

This paper established the Sargan-Hansen test as a test for the fit of external information and new data. It contains a proof of the separability lemma and the first part of Theorem 1 to show that the model moment equations cancel out. Further, the Sargan-Hansen test was extended to include an external interval to represent structural uncertainty. The obvious estimator $\hat{\mathbf{\Sigma}}_h$, the sample covariance matrix $\hat{\mathbf{S}}_h$, and a small sample version of the Sargan-Hansen test based on a normality assumption were developed. By and large, three versions of the Sargan-Hansen test were discussed. The employment of quadratic programming to compute the minimum test statistic for all

three versions of the test for linear external moment functions $\mathbf{h}(\mathbf{z})$ obviated the necessity of a grid search in these cases.

The paper provides the results of simulation studies for sample sizes of 30 and 50, as well as for various external moments. They considered normally distributed data and a linear model. When it comes to the type I error rate and power, the three tests only showed minor discrepancies. The observed power was sufficient and, in some cases, high for externally known expected values. The type I error rates were also below the nominal significance level. For an externally known variance of the dependent variable, $\text{Var}(y)$, the simulations indicated low power and increased type I error rates. The simulation studies also suggest that using multiple external moments does not necessarily lead to better test performance. In fact, combinations of $\text{Var}(y)$ with other external moments resulted in lower power than cases without $\text{Var}(y)$.

Taken together, the results indicate that the $\Gamma-$maximin Sargan-Hansen test is feasible and well suited for sample sizes commonly found in psychology. Thus, the first paper fulfilled the goal of providing a frequentist method of statistical cumulation that helps to distinguish between populations and detects structural uncertainty.

6.2 Paper 2: Using external information for more precise inferences

Jann, M., & Spiess, M. (2024). Using external information for more precise inferences in general regression models. *Psychometrika*, *89*(2), 439–460. https://doi.org/10.1007/s11336-024-09953-w

Statement of author's contribution:
I derived the analytical formulas, conducted the simulation studies, and wrote the initial draft of the paper. Then, Prof. Dr. Martin Spieß and I jointly discussed and finalized the draft for the initial submission, as well as for subsequent revisions.

This paper investigated the effects of using moment-type external information about variables to improve statistical analyses. Corollary 1 was proven in this paper. It showed the variance reduction property and the necessary condition $\mathbf{\Omega}_r \neq \mathbf{0}$ for variance reduction. The paper presents analytical formulas for the externally informed multiple linear regression model. The formulas sparked a discussion about which moments reduce the variance of which parameters. For instance, $E(y)$ only influenced the variance of the intercept. In addition, when used alone, covariate information had no variance-reducing effect. Knowledge of the covariances $\text{Cov}(y, x_j)$ reduced the variance of the intercept

and the respective slope β_j for x_j, but not the variances of the other slopes.

The coverage behavior of confidence interval unions was analyzed in simulation studies. These studies addressed structural uncertainty. However, estimation uncertainty was not reflected in order to test whether it was covered by using external intervals. The simulation studies used intervals of expected values, (co-) variances, correlations, simple linear regression slopes, and other moments as external information for sample sizes of 15, 30, 50, and 100. Two scenarios were considered: one with correctly specified external information and normally distributed regression error terms, and another with misspecified external information and χ^2_1-squared distributed regression error terms. As in the first study, the squared moments of the dependent variables, such as $\text{Var}(y)$, exhibited increased type I errors for sample sizes below 100.

Interestingly, in both simulation scenarios, the width of the confidence union was smaller than the width of the uninformed confidence intervals when covariances, correlations, or simple linear regression slopes were used. Furthermore, the confidence unions had valid type I error rates in the misspecified case, whereas the confidence intervals based on external values did not. These results suggest that using confidence unions increases distributional robustness. Finally, a real data set investigating the predictability of the premorbid intelligence of elderly individuals using lexical tasks was analyzed. This analysis used external information to demonstrate the variance reduction property of confidence unions in practice.

Overall, the results demonstrate the validity of confidence unions, even in misspecified cases and with sample sizes commonly found in psychological research. This paper introduced psychological researchers to a frequentist technique that enhances linear models using external information while accounting for structural uncertainty.

6.3 Paper 3: Fit of external information and data

Jann, M. (2024). Testing the fit of data and external sets via an imprecise Sargan-Hansen test. *International Journal of Approximate Reasoning*, *170*, 109214. https://doi.org/10.1016/j.ijar.2024.109214

This paper is an extended version of the first publication, as prompted by an invitation to contribute to the special issue of the International Journal of Approximate Reasoning on the Thirteenth International Symposium on Imprecise Probabilities: Theories and Applications. It expands upon the first paper's results in several ways. Rather than assuming that all matrices are regular, this paper generalized the results to generalized

inverses. The lemmas, corollaries, and theorems in this thesis are based on this extension. This paper introduced the use of an external variance estimate to represent estimation uncertainty. It was also the first to discuss and implement a combined representation of estimation and structural uncertainty. The Γ-maximin test was introduced through a rigorous discussion of decision-making under imprecise probabilities.

An algorithm was provided to compute the minimum Sargan-Hansen test statistic, thus eliminating the need for a grid approximation. Figure 1 was provided in the paper to illustrate the impact of various estimators of $\boldsymbol{\Omega}_h$ on the Sargan-Hansen test statistic. As a Bayesian competitor, this paper derived the criterion for prior-data conflict with threshold a, as discussed in Section 4.1.

Simulation studies were conducted to analyze the behavior of the proposed methods in small samples with directly simulated structural uncertainty. The results showed that the proposed methods had valid type I errors and sufficient power. For the Bayesian method, different values of the threshold a produced very different type I error rates. This may be a drawback, as the threshold a is more difficult to interpret and define a priori than the nominal type I error rate is in the frequentist case. Ultimately, the estimator of $\boldsymbol{\Omega}_h$ that produced the highest type I error rates was $\hat{\mathbf{S}}_h$. Therefore, the estimators $\hat{\boldsymbol{\Sigma}}_h$ or $\hat{\boldsymbol{\mathcal{S}}}_h$ should be preferred.

Taken together, the existing uncertainties can be addressed for a wide range of hypothesis testing problems using the proposed Γ-maximin test. Thus, this paper accomplished the goal of providing a robust frequentist method for using external information in statistical inference. It enables applied researchers to actually consider both estimation and structural uncertainty in their analyses and benefit from using external information.

6.4 Paper 4: Testing linear hypotheses using external information

Jann, M., & Spiess, M. (2025). *Testing linear hypotheses in repeated measures generalized linear models using external information* [Manuscript under review]

Statement of author's contribution:
I derived the analytical formulas, conducted the simulation studies, and wrote the initial draft of the paper. Then, Prof. Dr. Martin Spieß and I discussed and finalized the draft for the initial submission collaboratively.

This paper extended the tests statistics W, LM, and D to include an external interval by generalizing them to a Γ-maximin tests for general linear hypotheses. The test

statistics were analyzed for generalized linear models with repeated measures estimated with GEE. The special case of block invariant independent variables allowed for the derivation of analytical formulas for the GMM estimator, and a simplification of D. Constrained optimization methods were developed to calculate all three test statistics when external values are present. Then, these methods were extended to an external interval via grid approximation.

The paper provides the results of simulation studies analyzing the behavior of the three tests in small samples based on two real-data examples. The first example was error count data collected from 63 participants in a figure-ground segmentation experiment based on sequential distractor-response binding. The second example built upon an intervention study in sports psychology with 72 participants. It tested mood enhancement through exercise by comparing a control group with an exercise group measured three times. The dependent variable was categorized to employ and test cumulative logit models. Many simulation scenarios were tested to analyze the effects of estimation and structural uncertainty. To do so, two external samples based on different true values were generated.

Using an external interval rather than external values reduced type I errors. Nominal significance levels were achieved when the interval was correctly specified. However, misspecified intervals led to increased type I error rates. Therefore, caution is advised when constructing external intervals. Interestingly, the simulation studies showed that using external values that fully determine some of the model parameters caused the GMM estimator to set those parameters to that exact value and set the corresponding variance to zero. This was discussed as a strong argument for using external variances to reflect that the value is based on a sample. Otherwise, GMM estimation treats the external values as if they were population quantities. Finally, the two data sets were reanalyzed using external intervals. This resulted in reduced p-values, demonstrating the effect of the variance reduction property in practice.

Overall, externally informed generalized linear models with repeated measures showed good performance in small samples in scenarios relevant to psychological research. Thus, this paper achieved the goal of developing a frequentist method of statistical cumulation for generalized linear models with repeated measures that possibly reduces variance and increases power while being less susceptible to misspecification of external information.

7 Discussion

7.1 Conclusion and limitations

This thesis aimed to demonstrate the possibility of a cumulative psychological science in terms of statistical cumulation. The ideal has been that statistical cumulation (i.e., using external information) is theory-independent, allows for testing for differences between populations, and improves statistical analyses. The goals of the PhD project have been to rigorously examine external information and its uncertainties, explore existing approaches and develop a novel frequentist approach.

Chapter 3 provided a formal description of external information, identified various types of external information, such as moment-type external information about variables and discussed a variety of uncertainties related to external information. Estimation and structural uncertainty are particularly likely to be present when external information is derived from previous studies. Chapter 4 provided an overview of existing approaches and identified the generalized Bayesian and inferential model approaches as the only ones capable of reflecting structural uncertainty. To the best of the author's knowledge, no frequentist framework existed at the beginning of this PhD project that could incorporate external information while reflecting structural uncertainty.

The externally informed GMM approach presented in Chapter 5 is able to address this issue. It fills a gap in the scope of existing methods, as generalized Bayesian and inferential model approaches cannot incorporate moment-type external information about variables directly yet. Furthermore, moment-type external information about variables is not limited to one particular model. One piece of information can be used in numerous future studies with different models. Thus, the externally informed GMM approach provides a theory-independent method of statistical cumulation.

The Γ-maximin Sargan-Hansen test provides a method to test if two populations differ in certain aspects. Further, Section 5.3 discussed the variance reduction property of the externally informed GMM approach. This property improves statistical analyses by leading to higher power of significance tests and narrower confidence intervals. Even when these positive effects disappear when an external interval is used, distributional

robustness increases due to the F-probability induced by the external interval. One advantage of using external intervals instead of aggregated external values is that they are less susceptible to structural uncertainty. The attached papers include simulation studies indicating the validity of the externally informed GMM approach in small samples across various scenarios.

Overall, this PhD project demonstrated that the robust use of external information in statistical analysis and thus statistical cumulation is possible. The discussed approaches could serve as starting grounds for a cumulative psychological science that is able to use accumulated empirical findings, aggregated or not, in subsequent research. To enable applied researchers to use the externally informed GMM approach developed in this thesis, the analytical results from the four attached papers, as well as general approaches to solving the first-order conditions (5.1), have been implemented as R functions and will be made openly available via https://github.com/MartinJann/exgmm.

The most apparent limitation of the externally informed GMM approach is that the external interval has to be correctly specified to contain an unbiased estimator of the true moment value for the new data. This is related to the fact that structural uncertainty is rarely fully understood in practice, so some aspects of the new data set may not be covered by external sources. In this respect, the correct specification of the external interval will remain an assumption. Based on the available qualitative and quantitative information, it is up to the researcher to decide whether to rest their results on this assumption or not. This limitation can also be discussed in relation to other approaches of incorporating external information into statistical analyses of small samples, where the partial prior may substantially influence the results. This highlights that sensitivity to misspecification of external information is not unique to the current approach, but applies to all existing approaches. To enhance statistical analyses based on external information while maintaining their frequentist validity, accepting such assumptions may be the only option.

Moreover, if there are substantial doubts about the correctness of the external interval specification, the externally informed GMM approach enables researchers to test the fit of external information and new data using the Γ-Maximin Sargan-Hansen test. The fit can be evaluated for a new data set, and the external information can be applied to other data sets collected under the same conditions. Thus, the presented approach proves to be useful despite the risk of misspecified external intervals.

Another limitation of the externally informed GMM approach is that it is only feasible for low-dimensional external intervals. This is because the duration of a grid search increases exponentially with each additional dimension, assuming an equal number of

grid points in each dimension. However, adding some moments may not improve statistical analyses. For example, in the case of the Sargan-Hansen test, combining multiple moments did not increase the test's power if one of the moments was ineffective, as demonstrated in Paper 1. Rather than producing synergy, combining the moments produced a mixture of their behaviors. Therefore, one should refrain from using as many moments as possible and instead investigate their behavior.

7.2 Further research

Thus far, the externally informed GMM approach has been tested with multiple linear and repeated measures generalized linear models. However, many other models are employed in psychological research. Thus, more research is necessary to understand how the approach behaves with different models. For instance, further studies should examine two-level mixed linear models and structural equation models. Of particular interest are the performance of the externally informed GMM approach with small samples, its robustness under misspecification of underlying distributions, and the effectiveness of external information about specific statistical moments.

More research is also needed on parameter-specific external information. One possible approach would be to extend the externally informed GMM approach by using stochastic constraints on the parameters and applying the distributionally robust optimization framework. Additionally, further research is warranted to understand how to incorporate higher-dimensional external intervals. This requires computational methods other than grid approximation. Another approach could be to reduce the dimension of the external intervals based on efficiency. For instance, it ought to be investigated whether including multiple moments results in greater variance reduction than including a single, highly effective moment.

Applications of external information other than the ones discussed in this thesis may be explored in future endeavors. For instance, recognizing that misspecified external values generally result in a bias that does not vanish asymptotically could help to correct for sample selectivity. Constructing externally informed GMM estimators that mimic poststratification would be beneficial if there were reliable estimates of a variable's characteristics within the population of interest, such as official demographic statistics.

Finally, in the spirit of mutual completion, it is important to research and implement externally informed GMM, generalized Bayesian, and inferential model approaches alongside one another. This would allow applied researchers to access the full potential of statistical cumulation methods and to establish a cumulative psychological science.

Bibliography

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Augustin, T. (2002). Neyman–Pearson testing under interval probability by globally least favorable pairs: Reviewing Huber–Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference*, *105*(1), 149–173. https://doi.org/10.1016/S0378-3758(01)00208-7

Augustin, T., Coolen, F. P. A., De Cooman, G., & Troffaes, M. C. M. (2014). *Introduction to imprecise probabilities*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118763117

Augustin, T., Walter, G., & Coolen, F. P. A. (2014). Statistical inference. In T. Augustin, F. P. A. Coolen, G. De Cooman, & M. C. M. Troffaes (Eds.), *Introduction to imprecise probabilities* (pp. 135–189). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118763117.ch7

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316870

Bickel, D. R. (2012). A frequentist framework of inductive reasoning. *Sankhyā: The Indian Journal of Statistics, Series A*, *74*(2), 141–169. https://doi.org/10.1007/s13171-012-0020-x

Bickel, D. R. (2015). Inference after checking multiple Bayesian models for data conflict and applications to mitigating the influence of rejected priors. *International Journal of Approximate Reasoning*, *66*, 53–72. https://doi.org/10.1016/j.ijar.2015.07.012

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118619179

Bond, S., & Windmeijer, F. (2005). Reliable inference for GMM estimators? Finite sample properties of alternative test procedures in linear panel data models. *Econometric Reviews*, *24*(1), 1–37. https://doi.org/10.1081/ETC-200049126

Borchers, H. W. (2023). *Pracma: Practical numerical math functions* [R package version 2.4.4]. https://CRAN.R-project.org/package=pracma

Böschen, I. (2023). Changes in methodological study characteristics in psychology between 2010-2021. *PLOS ONE*, *18*(5), 1–25. https://doi.org/10.1371/journal.pone.0283353

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511811241

Casella, G., & Berger, R. (2024). *Statistical inference* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781003456285

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, *29*(2), 357–372. https://doi.org/10.1214/aoms/1177706618

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. Chapman; Hall/CRC. https://doi.org/10.1201/b14832

de Leeuw, J., & Meijer, E. (2008). Introduction to multilevel analysis. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 1–75). Springer New York. https://doi.org/10.1007/978-0-387-73186-5_1

Diggle, P. J., Heagerty, P. J., Liang, K.-y., & Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford University Press. https://doi.org/10.1093/oso/9780198524847.001.0001

Dufour, J.-M., Trognon, A., & Tuvaandorj, P. (2016). Invariant tests based on M-estimators, estimating functions, and the generalized method of moments. *Econometric Reviews*, *36*(1-3), 182–204. https://doi.org/10.1080/07474938.2015.1114285

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, *6*(4), 391–398. https://doi.org/10.1111/j.1469-1809.1935.tb02120.x

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–701. https://doi.org/10.1198/016214505000000105

Geisser, S. (1999). Remarks on the 'Bayesian' method of moments. *Journal of Applied Statistics*, *26*(1), 97–101. https://doi.org/10.1080/02664769922683

Ghosh, B. K. (1973). Some monotonicity theorems for χ^2, F and t distributions with applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, *35*(3), 480–492. https://doi.org/10.1111/j.2517-6161.1973.tb00976.x

Gray, R. M. (1988). *Probability, random processes, and ergodic properties*. Springer New York. https://doi.org/10.1007/978-1-4757-2024-2

Hall, A. R., & Inoue, A. (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, *114*(2), 361–394. https://doi.org/10.1016/S0304-4076(03)00089-7

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*(4), 1029–1054. https://doi.org/10.2307/1912775

Hansen, L. P., Heaton, J., & Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, *14*(3), 262–280. https://doi.org/10.1080/07350015.1996.10524656

Hellerstein, J. K., & Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *The Review of Economics and Statistics*, *81*(1), 1–14. https://doi.org/10.1162/003465399557860

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119536604

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9780470434697

Hunter, M. D., Fisher, Z. F., & Geier, C. F. (2024). What ergodicity means for you. *Developmental Cognitive Neuroscience*, *68*, 101406. https://doi.org/https://doi.org/10.1016/j.dcn.2024.101406

Huntley, N., Hable, R., & Troffaes, M. C. M. (2014). Decision making. In T. Augustin, F. P. A. Coolen, G. De Cooman, & M. C. M. Troffaes (Eds.), *Introduction to imprecise probabilities* (pp. 190–206). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118763117.ch8

Huray, P. G. (2009). *Maxwell's equations*. John Wiley & Sons. https://doi.org/10.1002/9780470549919

Imbens, G. W., & Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, *61*(4), 655–680. https://doi.org/10.2307/2297913

Jann, M. (2023). Testing the coherence of data and external intervals via an imprecise Sargan-Hansen test. In E. Miranda, I. Montes, E. Quaeghebeur, & B. Vantaggi (Eds.), *Proceedings of the thirteenth international symposium on imprecise*

probability: Theories and applications (pp. 249–258, Vol. 215). PMLR. https://proceedings.mlr.press/v215/jann23a.html

Jann, M. (2024). Testing the fit of data and external sets via an imprecise Sargan-Hansen test. *International Journal of Approximate Reasoning*, *170*, 109214. https://doi.org/10.1016/j.ijar.2024.109214

Jann, M., & Spiess, M. (2024). Using external information for more precise inferences in general regression models. *Psychometrika*, *89*(2), 439–460. https://doi.org/10.1007/s11336-024-09953-w

Jann, M., & Spiess, M. (2025). *Testing linear hypotheses in repeated measures generalized linear models using external information* [Manuscript under review].

Jevremov, T., & Pajić, D. (2024). Bayesian method in psychology: A bibliometric analysis. *Current Psychology*, *43*(10), 8644–8654. https://doi.org/10.1007/s12144-023-05003-3

Kadane, J. B., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *47*(1), 3–19. https://doi.org/10.1111/1467-9884.00113

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00513

Klenke, A. (2020). *Probability theory: A comprehensive course*. Springer International Publishing. https://doi.org/10.1007/978-3-030-56402-5

Knopov, P. S., & Korkhin, A. S. (2012). *Regression analysis under a priori parameter restrictions* (1st ed.). Springer New York. https://doi.org/10.1007/978-1-4614-0574-0

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago press.

Kuhn, T. S. (1987). What are scientific revolutions? In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution, vol. 1: Ideas in history*. The MIT Press.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.2307/2336267

Loehlin, J. C., & Beaujean, A. A. (2016). *Latent variable models: An introduction to factor, path, and structural equation analysis* (5th ed.). Routledge. https://doi.org/10.4324/9781315643199

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer New York. https://doi.org/10.1007/b97478

Martin, R. (2022). Valid and efficient imprecise-probabilistic inference with partial priors, I. First results. https://arxiv.org/abs/2203.06703

Martin, R. (2023a). Valid and efficient imprecise-probabilistic inference with partial priors, II. General framework. https://arxiv.org/abs/2211.14567

Martin, R. (2023b). Valid and efficient imprecise-probabilistic inference with partial priors, III. Marginalization. https://arxiv.org/abs/2309.13454

Martin, R., & Chuanhai, L. (2015). *Inferential models*. Chapman; Hall/CRC. https://doi.org/10.1201/b19269

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 234–251. https://doi.org/10.1111/j.2044-8317.1984.tb00802.x

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological methods*, *22*(1), 114. https://doi.org/10.1037/met0000078

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115. https://doi.org/10.1086/288135

Mischel, W. (2009). Becoming a cumulative science. *APS Observer*, *22*(1), 3. https://www.psychologicalscience.org/observer/becoming-a-cumulative-science

Newey, W. K., & McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. In *Handbook of econometrics* (pp. 2111–2245, Vol. 4). Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*(2), 237–249. https://doi.org/https://doi.org/10.2307/2336172

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Patole, S. (2021). Systematic reviews, meta-analysis, and evidence-based medicine. In S. Patole (Ed.), *Principles and practice of systematic reviews and meta-analysis*

(pp. 1–10). Springer International Publishing. https://doi.org/10.1007/978-3-030-71921-0_1

Pearson, K. (1936). Method of moments and method of maximum likelihood. *Biometrika*, *28*(1/2), 34–59. https://doi.org/10.2307/2334123

PubMed. (2025). *Search results based on the term "psychology"*. Retrieved June 13, 2025, from https://pubmed.ncbi.nlm.nih.gov/?term=psychology&sort=date&ac=yes&timeline=expanded

Puntanen, S., Styan, G. P. H., & Isotalo, J. (2011). *Matrix tricks for linear statistical models*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10473-2

Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*(1), 300–325. https://doi.org/10.1214/aos/1176325370

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rahimian, H., & Mehrotra, S. (2022). Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, *3*, Article 4, 1–85. https://doi.org/10.5802/ojmo.15

Rasch, D., Kubinger, K. D., & Yanagida, T. (2011). *Statistics in psychology using R and SPSS*. John Wiley & Sons. https://doi.org/10.1002/9781119979630

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316696

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. https://doi.org/10.1080/01621459.1996.10476908

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, *26*(3), 393–415. https://doi.org/10.2307/1907619

Schultz, D. (1981). *A history of modern psychology* (3rd ed.). Academic Press. https://doi.org/10.1016/C2013-0-11479-1

Schwaferts, P. (2022). *Improving practical relevance of Bayes factors*. Ludwig-Maximilians-Universität München. https://doi.org/10.5282/edoc.29449

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Smith, J. E., & Winkler, R. L. (2006). The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, *52*(3), 311–322. https://doi.org/10.1287/mnsc.1050.0451

Smith, T. M. F. (1991). Post-stratification. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *40*(3), 315–323. https://doi.org/10.2307/2348284

Smithson, M. (2014). Elicitation. In T. Augustin, F. P. A. Coolen, G. De Cooman, & M. C. M. Troffaes (Eds.), *Introduction to imprecise probabilities* (pp. 318–328). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118763117.ch15

Spiess, M. (2006). Estimation of a two-equation panel model with mixed continuous and ordered categorical outcomes and missing data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *55*(4), 525–538. https://doi.org/10.1111/j.1467-9876.2006.00551.x

Spiess, M., & Jordan, P. (2023). In models we trust: Preregistration, large samples, and replication may not suffice. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1266447

Sukhera, J. (2022). Narrative reviews: Flexible, rigorous, and practical. *Journal of Graduate Medical Education*, *14*(4), 414–417. https://doi.org/10.4300/JGME-D-22-00480.1

Todorovic, P. (1992). *An introduction to stochastic processes and their applications*. Springer New York. https://doi.org/10.1007/978-1-4613-9742-7

van den Akker, O. R., van Assen, M. A. L. M., Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2024). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. *Behavior Research Methods*, *56*(6), 5424–5433. https://doi.org/10.3758/s13428-023-02277-0

Wallgren, A., & Wallgren, B. (2014). *Register-based statistics: Statistical methods for administrative data* (2nd ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118855959

Walter, G., & Augustin, T. (2009). Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, *3*(1), 255–271. https://doi.org/10.1080/15598608.2009.10411924

Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept* [Elementary basic concepts of a more general probability theory I: Interval probability as a comprehensive concept]. Physica Heidelberg. https://doi.org/10.1007/978-3-642-57583-9

Xiao, Z. (2020). Efficient GMM estimation with singular system of moment conditions. *Statistical Theory and Related Fields*, *4*(2), 172–178. https://doi.org/10.1080/24754269.2019.1653159

Xie, M.-g., & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review / Revue Internationale de Statistique*, *81*(1), 3–39. https://doi.org/10.1111/insr.12000

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis*, *4*(2), 191–207. https://doi.org/10.1214/09-BA407

Zabell, S. L. (1992). R. A. Fisher and fiducial argument. *Statistical Science*, *7*(3), 369–387. https://doi.org/10.1214/ss/1177011233

Zellner, A. (1996). Bayesian method of moments (BMOM) analysis of mean and regression models. In J. C. Lee, W. O. Johnson, & A. Zellner (Eds.), *Modelling and prediction honoring Seymour Geisser* (pp. 61–72). Springer New York. https://doi.org/10.1007/978-1-4612-2414-3_4

Appendix A

Paper 1

Jann, M. (2023). Testing the coherence of data and external intervals via an imprecise Sargan-Hansen test. In E. Miranda, I. Montes, E. Quaeghebeur, & B. Vantaggi (Eds.), *Proceedings of the thirteenth international symposium on imprecise probability: Theories and applications* (pp. 249–258, Vol. 215). PMLR. https://proceedings.mlr.press/v215/jann23a.html

Proceedings of Machine Learning Research 215:249–258, 2023 ISIPTA 2023

Testing the Coherence of Data and External Intervals via an Imprecise Sargan-Hansen Test

Martin Jann MARTIN.JANN@UNI-HAMBURG.DE
Department of Research Methods and Statistics, Institute of Psychology, University of Hamburg, Germany

Abstract

When information about a population is sparse, it is difficult to test whether a data set originated from that population. In applied research, however, researchers often have access to external information in the form of (central) statistical moments such as mean or variance. To compensate for the uncertainty in the external point values, this paper uses external intervals instead to represent the information about moments. The Sargan-Hansen test from the generalized method of moments framework is used, which exploits point-valued external information about moments in the presence of a statistical model to test whether data and external information are in conflict. For the Sargan-Hansen test, a separability result is derived with respect to the model and the external information. This result leads to a simplification of the test in terms of its analytical form and the calculation of the test statistics. To allow the use of external intervals instead of point values, an imprecise version of the Sargan-Hansen test is created using the Gamma-maximin decision rule. Assuming that the variables are normally distributed, a small sample version of this imprecise Sargan-Hansen test is derived. The power and type I errors of the developed tests are analyzed and compared in a simulation study in different small sample scenarios.

Keywords: imprecise external information, information-data conflict, generalized method of moments, Sargan-Hansen test, credal set, robustness

1. Introduction

The use of (external) prior information on parameters has frequently been studied. Well-known techniques for incorporating external information into statistical analysis include informed prior distributions in Bayesian statistics [3] and constraints on the parameter space imposed by the external information, leading to constrained optimization (see, e.g. Knopov and Korkhin [11] for the case of multiple linear regression). However, in some research areas, there may not be enough information to determine the feasible region or a prior distribution. The following example is provided to support this assertion:

Example 1 *Suppose we have a simple linear regression model $y = \beta_1 + x\beta_2 + \epsilon$ under Gauss-Markov assumptions and only the expected value $E(y) = 100$ is known externally. Under the model assumptions, $E(y) = 100$ becomes a constraint on the parameter,*

$$100 = E(y) = \beta_1 + E(x)\beta_2, \tag{1}$$

which is a linear constraint on intercept β_1 and slope β_2. However, if $E(x)$ is not known, we cannot use Equation (1) directly as a constraint in the optimization. Equation (1) is also not sufficient to identify (the moments of) a prior distribution, since there are usually several different distributions that satisfy this condition.

The fact that the external information in Example 1 is in the form of a moment motivates another method of using external information. According to an idea proposed by Imbens and Lancaster [9], this type of external information implies moment conditions that can be combined with the moment conditions used to estimate a statistical model. In general, the resulting overidentified system of moment conditions does not have an exact solution, but the Generalized Method of Moments (GMM) [7] can be used to find estimators that are 'as close as possible' to a solution with respect to some norm. Imbens and Lancaster [9] showed for multiple linear models that the estimators found in this way generally have lower variances than the corresponding OLS estimators, provided that the external information is correct. This paper examines the opposite question: Given the combined moment conditions of the model and the external information, is the external information correct (for a given data set)? This concept is similar to the prior-data conflict in Bayesian statistics and will be referred to hereafter as *information-data conflict*. To answer this question in the GMM framework, the Sargan-Hansen test is typically used because it is a test for overidentifying restrictions [18, 7].

However, its role as a test for misspecification has been criticized in current research, especially with respect to models that use instrumental variables [14, 10]. Therefore, the results of this paper should be interpreted as a test of the coherence of external information and data rather than a test of misspecification of a model. This argument is supported by a small thought experiment. There are

two statements: *"The model assumptions are true."* and *"The expected value of the dependent variable is 100."* Both statements are logically independent, one is neither necessary nor sufficient for the other to be true. How might a model specification test benefit from this kind of external information? A mathematical formulation of this logical independence is proved in Section 2.

Most external information depends on population, time, and many other aspects, which makes the use of point values for the external information risky because the results of Imbens and Lancaster [9] depend on the correctness of these point values. To reduce the risk of potentially misspecified external information, this paper addresses the case where an interval is given that contains the true value of the external moments, but its exact position inside the interval is unknown. This epistemic uncertainty about the true value of the external moments leads directly to the use of imprecise probabilities in the form of credal sets, as we show in Section 2.

2. The Sargan-Hansen Test with External Information

2.1. The Point-Valued Case

We assume that the external information only consists of point values of the respective moments. The notation from Newey and McFadden [12] is adopted. In the following, italic lowercase letters are for (random) scalar values, bold lowercase letters are for (random) vectors, and bold uppercase letters are for (random) matrices, unless otherwise indicated. Now let $\mathbf{z}$ be a random variable over $\mathbb{R}^k$ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be $n > 1$ i.i.d. random variables distributed like $\mathbf{z}$.[1] Further let q be an integer and $\boldsymbol{\Theta} \subset \mathbb{R}^q$, then let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be a possible value for a (fixed) parameter of a statistical model, where $\boldsymbol{\theta}_0$ is the true value. Given a function $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ with the property $E[\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)] = \mathbf{0}$, one can try to estimate the parameter by the method of moments. Practically, this is done by formulating the equivalent sample moment conditions $\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{0}$ and solving for $\boldsymbol{\theta}$.

To explain this method, let's consider Example 1. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be an i.i.d. sample of random variables distributed like y, and let

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,q-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,q-1} \end{pmatrix}$$

be the design matrix containing the covariates assumed to be an i.i.d. sample of the random variable $\mathbf{x}$. The sample moment conditions for the OLS estimator can be derived by setting the mixed moment of the independent variables and the error term to zero, i.e. $E(\mathbf{g}(\mathbf{z}, \boldsymbol{\beta}_0)) = E(\mathbf{x}\epsilon) = \mathbf{0}$. The sample moment conditions are therefore $\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\beta}) = \frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ in this case denotes the parameter [4, p. 172].

However, sometimes the number of the moment conditions is larger than the dimension of the parameter. As a classic example from econometrics, we present estimation using instrumental variables, following the presentation of Cameron and Trivedi [4, p. 170]. As before, we assume a linear model. If some of the independent variables in $\mathbf{x}$ are correlated with the error term, then the Gauss-Markov assumptions are incorrect, and therefore OLS will not provide a consistent estimate of the regression parameter. A common idea to solve this problem is to find other variables that are correlated with $\mathbf{x}$ but uncorrelated with the error term. These variables are called instruments, and we represent their sample realizations by the $(n \times s)$ matrix $\mathbf{D}$. Similar to the OLS case, we can set the mixed moment of the instruments and the error term to zero. The corresponding sample moment conditions are $\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\beta}) = \frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. If the number of potential instruments is greater than the dimension of the parameter, the sample moment conditions are generally not solvable for $\boldsymbol{\beta}$, the system of equations is overidentified. Not using all the instruments would result in a loss of efficiency. Instead of solving the equations, the idea of the GMM is to find a value for $\boldsymbol{\beta}$ that makes $\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ as small as possible in terms of quadratic loss, i.e, by minimizing

$$(\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))^T \mathbf{W} (\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})),$$

where $\mathbf{W}$ is a chosen positive-definite weighting matrix. Note that this is a generalization of the case of solvable sample moment conditions, since a positive quadratic form in $\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is zero if and only if $\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$. In general, different $\mathbf{W}$ lead to different estimators. In the GMM approach, there is a way to choose the optimal weighting matrix with respect to the efficiency of the estimator. This optimality is achieved by $\mathbf{W} = \boldsymbol{\Omega}^{-1}$ with $\boldsymbol{\Omega} = E(\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})^T)$ [7]. This optimal $\mathbf{W}$ is almost always unknown and must be estimated by a random matrix $\hat{\mathbf{W}}$. Taken together, this leads to the following definition:

Definition 1 *[12, p. 2116] Let $p \geq q$ be an integer and $\boldsymbol{g}(\boldsymbol{z}, \boldsymbol{\theta})$ be a vector valued function with values in $\mathbb{R}^p$, that meets the moment conditions $E[\boldsymbol{g}(\boldsymbol{z}, \boldsymbol{\theta}_0)] = \boldsymbol{0}$. Further let $\hat{\boldsymbol{W}} \in \mathbb{R}^{p,p}$ be a positive semi-definite (and hence symmetric) random matrix such that $(\boldsymbol{r}^T \hat{\boldsymbol{W}} \boldsymbol{r})^{1/2}$ is almost surely a norm for all $\boldsymbol{r} \in \mathbb{R}^p$. Then a **GMM-estimator** $\hat{\boldsymbol{\theta}}_{ex}$ is defined as a $\boldsymbol{\theta}$, that maximizes the following objective function:*

$$\hat{Q}_n(\boldsymbol{\theta}) = -(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(z_i, \boldsymbol{\theta}))^T \hat{\boldsymbol{W}} (\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(z_i, \boldsymbol{\theta})) \,. \qquad (2)$$

[1] Some entries of $\mathbf{z}$ could possibly be fixed, as long as at least one entry is random.

Under mild regularity conditions, the GMM-estimator is point-identified, consistent, and asymptotically normally distributed [12, Theorem 3.4]. To emphasize the generality of the GMM, we give some examples. Special cases of GMM-estimators range from OLS estimators to maximum likelihood estimators (MLE) [4, p. 172] to estimators derived by generalized estimating equations [4, p. 790]. To see that GMM is an extension of MLE, note that maximizing the log-likelihood function implies setting the score function to zero. This corresponds to the first-order conditions for MLE and has exactly the form of sampling moment conditions. In addition, the regularity conditions of the MLE require that the expected value of the score function be zero at the true parameter value, which is exactly the requirement $E[\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)] = \mathbf{0}$ in Definition 1. This property of the score function is central to establishing the consistency and asymptotic normality of the MLE. For the mathematical details of incorporating the MLE into the GMM, see Cameron and Trivedi [4, p. 140]. Finally, there is also an important connection to robust statistics, since M-estimators with differentiable ρ (those of the $\psi-$type) are also derived by sample moment conditions and thus represent a special case of GMM estimators [4, p. 118].

Following Imbens and Lancaster [9], we include in $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ not only the moment conditions for the model, but also those for the external information, resulting in an overidentified system of moment conditions. Let $\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})$ denote the $p_1 \geq q$ moment conditions for the model and $\mathbf{h}(\mathbf{z})$ denote the p_2 moment conditions for the external information, which are assumed to be expressible as functions of the data alone, then $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}) = (\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})^T, \mathbf{h}(\mathbf{z})^T)^T$. For example, the condition for the external moment in Example 1 is $h(\mathbf{z}) = y - 100$. If one of the moment conditions for the external information depends only on the parameter, then the results derived here will not hold in general.

For the overidentified case $p > q$, under the null hypothesis that all moment conditions are correct, it holds that $-n\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex}) \xrightarrow{d} \chi^2_{p-q}$ if the regularity conditions hold and if $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W} = \boldsymbol{\Omega}^{-1}$. The χ^2-test that results from this distribution property is called the Sargan-Hansen test [18, 7]. For simplicity, in the remainder of this paper we assume that $\hat{\mathbf{W}}$ is invertible almost surely and therefore positive-definite almost surely by Definition 1. All the following results are derived for this almost sure case of invertible $\hat{\mathbf{W}}$ and thus hold almost surely. If $\hat{\mathbf{W}}$ is singular for certain data, one should first check whether the moment conditions are linearly dependent, and accordingly delete some conditions, so that the remaining ones are not linearly dependent. Otherwise, one could add random noise to $\hat{\mathbf{W}}$ to try to make it invertible, or use its Moore-Penrose inverse [20].

Let $\hat{\boldsymbol{\Omega}}$ be the inverse of $\hat{\mathbf{W}}$. For the sake of brevity we define $\overline{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^n \mathbf{m}(\mathbf{z}_i, \boldsymbol{\theta})$ and $\overline{\mathbf{h}} = \frac{1}{n}\sum_{i=1}^n \mathbf{h}(\mathbf{z}_i)$. By $\mathbf{A}/\mathbf{B}$ we denote the Schur complement of the block $\mathbf{B}$ of the matrix $\mathbf{A}$ and obtain

Lemma 2 *(Separability) From the premises of Definition 1 and $\mathbf{g}(z, \boldsymbol{\theta}) = (\boldsymbol{m}(z, \boldsymbol{\theta})^T, \boldsymbol{h}(z)^T)^T$ it follows that $\hat{\boldsymbol{\Omega}}$ has the block form*

$$\hat{\boldsymbol{\Omega}} = \begin{pmatrix} \hat{\boldsymbol{\Omega}}_m & \hat{\boldsymbol{\Omega}}_r^T \\ \hat{\boldsymbol{\Omega}}_r & \hat{\boldsymbol{\Omega}}_h \end{pmatrix},$$

where $\hat{\boldsymbol{\Omega}}_m \in \mathbb{R}^{p_1,p_1}$ and $\hat{\boldsymbol{\Omega}}_h \in \mathbb{R}^{p_2,p_2}$. Further,

$$-\hat{Q}_n(\boldsymbol{\theta}) = (\overline{\boldsymbol{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\boldsymbol{h}})^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}(\overline{\boldsymbol{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\boldsymbol{h}}) + \overline{\boldsymbol{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\boldsymbol{h}}.$$

Proof We take advantage of the fact that $\hat{\mathbf{W}}$ is symmetric, positive-definite, and can be written in block form

$$\hat{\mathbf{W}} = \begin{pmatrix} \hat{\mathbf{W}}_m & \hat{\mathbf{W}}_r^T \\ \hat{\mathbf{W}}_r & \hat{\mathbf{W}}_h \end{pmatrix},$$

where $\hat{\mathbf{W}}_m \in \mathbb{R}^{p_1,p_1}$ and $\hat{\mathbf{W}}_h \in \mathbb{R}^{p_2,p_2}$. The first statement follows from the fact that $\hat{\mathbf{W}} = \hat{\boldsymbol{\Omega}}^{-1}$ and the block form of $\hat{\mathbf{W}}$. For the second statement, note that $\hat{\mathbf{W}}$ is positive-definite and so is $\hat{\boldsymbol{\Omega}}$, so the Schur complement $\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h = \hat{\boldsymbol{\Omega}}_m - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1}\hat{\boldsymbol{\Omega}}_r$ is invertible. Now $\hat{\mathbf{W}}$ can be expressed by Schur complements:

$$\begin{aligned}
\hat{\mathbf{W}}_m &= (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}, \\
\hat{\mathbf{W}}_r &= -\hat{\boldsymbol{\Omega}}_h^{-1}\hat{\boldsymbol{\Omega}}_r(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}, \\
\hat{\mathbf{W}}_h &= \hat{\boldsymbol{\Omega}}_h^{-1} + \hat{\boldsymbol{\Omega}}_h^{-1}\hat{\boldsymbol{\Omega}}_r(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^{-1}.
\end{aligned}$$

It follows that

$$\begin{aligned}
-\hat{Q}_n(\boldsymbol{\theta}) &= (\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}))^T \hat{\mathbf{W}} (\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})) \\
&= \overline{\mathbf{m}}^T \hat{\mathbf{W}}_m \overline{\mathbf{m}} + 2\overline{\mathbf{m}}^T \hat{\mathbf{W}}_r^T \overline{\mathbf{h}} + \overline{\mathbf{h}}^T \hat{\mathbf{W}}_h \overline{\mathbf{h}} \\
&= \overline{\mathbf{m}}^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}\overline{\mathbf{m}} - 2\overline{\mathbf{m}}^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}} \\
&\quad + \overline{\mathbf{h}}^T (\hat{\boldsymbol{\Omega}}_h^{-1} + \hat{\boldsymbol{\Omega}}_h^{-1}\hat{\boldsymbol{\Omega}}_r(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^{-1})\overline{\mathbf{h}} \\
&= (\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}})^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}(\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}) \\
&\quad + \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}.
\end{aligned}$$

■

Lemma 2 can be interpreted as a separability result, since $\overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}$ is not a function of $\boldsymbol{\theta}$ if a suitable $\hat{\boldsymbol{\Omega}}_h$ is used, e.g, $\hat{\boldsymbol{\Sigma}}_h = \frac{1}{n}\sum_{i=1}^n \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^T$ or the sample covariance matrix $\hat{\mathbf{S}}_h = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}})(\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}})^T$. In these cases, $\hat{\boldsymbol{\Omega}}_h$ can be calculated from the data and external information

alone. Note that the matrix $\hat{\mathbf{S}}_h$ can be computed even without knowing the true external value. Both matrices are asymptotically identical if the null hypothesis of correctly specified external values holds, but different if it does not. The following important result holds for these examples.

Theorem 3 *Let the premises and notation of Lemma 2 be given. If $\hat{\boldsymbol{\Omega}}_h$ is not a function of $\boldsymbol{\theta}$ and if there is a $\boldsymbol{\theta}_h \in \boldsymbol{\Theta}$, for which $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\boldsymbol{h}} = \mathbf{0}$ holds, it follows that*

$$-\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex}) = \overline{\boldsymbol{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\boldsymbol{h}}.$$

Proof By Definition 1 we get

$$-\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex}) = -\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{Q}_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} -\hat{Q}_n(\boldsymbol{\theta}).$$

For $\boldsymbol{\theta}_h$ given in the premises, it follows from Lemma 2, that $-\hat{Q}_n(\boldsymbol{\theta}_h) = \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$. Since $\hat{\boldsymbol{\Omega}}$ is positive-definite, $(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}$ is also positive-definite. Therefore, $(\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}})^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1} (\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}})$ is a positive quadratic form and reaches its global minimum at 0, which is achieved by the given $\boldsymbol{\theta}_h$. Since $\overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$ is not a function of the parameter $\boldsymbol{\theta}$, the proof is complete. ■

Theorem 3 shows the reduction of the Sargan-Hansen test based on external information to a test of the fit of the external information and the data alone, without the model. Moreover, under the conditions of Theorem 3 the test statistic $-n\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex})$ has the form of a Wald statistic, and the Sargan-Hansen test is then equivalent to a Wald test of linear restrictions [4, p. 136]. The condition $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}} = \mathbf{0}$ is equivalent to the main separability result of Ahu and Schmidt [1], if the external information is interpreted as a parameter with only one possible value. Their result gives an indication of the meaning of $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}} = \mathbf{0}$, since they proved that it always holds when the first-order conditions for the GMM are satisfied. As an important special case, this result applies to OLS estimation in multiple linear models when the design matrix $\mathbf{X}$ has full rank, because the result then has the form $\frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}} = \mathbf{0}$, which can be directly resolved to $\boldsymbol{\beta}$. This is the mathematical form of logical independence mentioned in Section 1.

Finally, $\overline{\mathbf{h}}$, if the external information is correct, will in general almost surely be arbitrarily close to $\mathbf{0}$ for $n \to \infty$ as $E(\overline{\mathbf{h}}) = \mathbf{0}$, in which case the disturbance term $\hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$ vanishes. Overall, the case $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}} \neq \mathbf{0}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ seems to be rather pathological for models that are just-identified by their moment conditions, which is why it will not be treated in the rest of the paper and only the case of just-identified models, $q = p_1$, will be treated.

2.2. The Interval-Valued Case

The assumption of point-value external information is now weakened by the assumption that a (possibly multidimensional) closed interval $\mathbf{I}_{ex}$ is known, for which we want to test the null hypothesis that it contains the true value of the external moments. The nature of this external interval is that it is based on external data that is affected by random noise. Thus, it reflects the current state of knowledge about the (moments of the) variables. Now the regularity conditions of the GMM apply to this true value, but it is not known which value in $\mathbf{I}_{ex}$ it is. Therefore, $\mathbf{I}_{ex}$ can be interpreted as coarse data, and cautious data completion can be applied to the test statistic $n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$ to derive the set of possible test statistics without further assumptions [2, p. 182]. If $\mathbf{I}_{ex}$ is bounded and the test statistic is a continuous function of the external information, the result is a bounded and closed interval $[\underline{n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}}, \overline{n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}}]$, since in this case $\mathbf{I}_{ex}$ is compact and connected. The interval $[\underline{n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}}, \overline{n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}}]$ is denoted by $[\underline{\chi^2}, \overline{\chi^2}]$. However, if $\mathbf{I}_{ex}$ is unbounded, the cautious data completion may result in a right-unbounded interval $[\underline{\chi^2}, \infty)$, e.g, if $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$ is used. The test statistic interval cannot be left-unbounded because the test statistic $n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$ is a positive-definite quadratic form and therefore cannot be less than zero. In the following, we will focus on the case where the set of possible test statistics is an interval $[\underline{\chi^2}, \overline{\chi^2}]$.

One strategy for computing $[\underline{\chi^2}, \overline{\chi^2}]$ for a given data set is to use quadratic programming, as we will show now. To reflect the dependence of $\overline{\mathbf{h}}$ on the external value $\mathbf{e} \in \mathbf{I}_{ex}$, it is now written as a function $\overline{\mathbf{h}}(\mathbf{e})$. If $\hat{\boldsymbol{\Omega}}_h^{-1}$ is not a function of $\mathbf{e}$, e.g, $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$, the objective function $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e}) = \overline{\mathbf{h}}(\mathbf{e})^T (\hat{\boldsymbol{\Omega}}_h/n)^{-1} \overline{\mathbf{h}}(\mathbf{e})$ is already in quadratic form based on the variable $\overline{\mathbf{h}}(\mathbf{e})$. The feasible region becomes $\overline{\mathbf{h}}(\mathbf{I}_{ex})$, the image of $\mathbf{I}_{ex}$ under $\overline{\mathbf{h}}(\mathbf{e})$. If $\overline{\mathbf{h}}(\mathbf{e})$ can be written as $\overline{\mathbf{h}}(\mathbf{e}) = \hat{\mathbf{h}} - \mathbf{e}$, where $\hat{\mathbf{h}}$ represents the sample moment, then $\overline{\mathbf{h}}(\mathbf{I}_{ex}) = \hat{\mathbf{h}} - \mathbf{I}_{ex}$ holds (Again, $\hat{\mathbf{h}} - \mathbf{I}_{ex}$ denotes the image of $\hat{\mathbf{h}} - e$ on $\mathbf{I}_{ex}$.). In this case, the feasible region is an interval. Taken together, the optimization problem is now a quadratic programming problem.

If $\hat{\boldsymbol{\Omega}}_h$ depends on $\mathbf{e}$, for example $\hat{\boldsymbol{\Omega}}_h = \hat{\boldsymbol{\Sigma}}_h$, the optimization problem is more complex. Again, the dependence on $\mathbf{e}$ is denoted by the notation $\hat{\boldsymbol{\Omega}}_h(\mathbf{e})$. In this case, the problem is not necessarily convex, as Figure 1 shows. Another problem is that the matrix $\hat{\boldsymbol{\Omega}}_h(\mathbf{e})$ must be nonsingular for each $\mathbf{e}$ for the problem to be well-defined. In the case $\hat{\boldsymbol{\Omega}}_h(\mathbf{e}) = \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})$ both problems can be solved by

Theorem 4 *The matrix $\hat{\boldsymbol{\Sigma}}_h(\boldsymbol{e})$ is positive-definite for each $\boldsymbol{e} \in \boldsymbol{I}_{ex}$ if $\hat{\mathbf{S}}_h$ is positive-definite. Assuming that $\hat{\mathbf{S}}_h$ is*

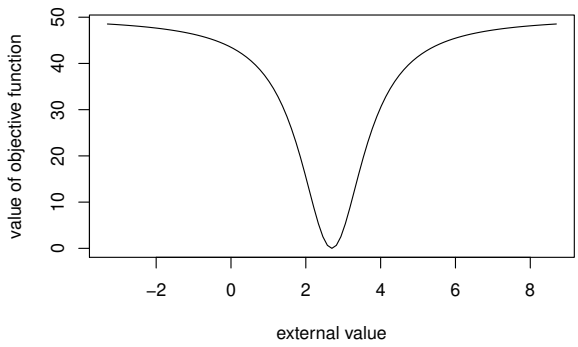


Figure 1: Graph of the objective function $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})^{-1}\overline{\mathbf{h}}(\mathbf{e})$ as a function of the external value $\mathbf{e}$ in the context of Example 1, i.e, $h(z) = y - e$, based on a sample of 50 i.i.d random variables $y_1, \ldots, y_{50}$ distributed like $N(3, 1)$

positive-definite, the objective function (2) *becomes*

$$n \cdot \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{\Sigma}}_h(\boldsymbol{e})^{-1}\overline{\boldsymbol{h}}(\boldsymbol{e}) = n \cdot \frac{\overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{S}}_h^{-1}\overline{\boldsymbol{h}}(\boldsymbol{e})}{\frac{n-1}{n} + \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{S}}_h^{-1}\overline{\boldsymbol{h}}(\boldsymbol{e})}$$

and reaches its minimum over $\boldsymbol{I}_{ex}$ *at the same point as the objective function* $n \cdot \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{S}}_h^{-1}\overline{\boldsymbol{h}}(\boldsymbol{e})$.

Proof For brevity, we will denote $\overline{\mathbf{h}}(\mathbf{e})$ by $\overline{\mathbf{h}}$ during the proof of the first statement. The first statement is clear by definition, since

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_h(\mathbf{e}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^T \\ &= \frac{1}{n}\sum_{i=1}^{n} (\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}} + \overline{\mathbf{h}})(\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}} + \overline{\mathbf{h}})^T \qquad (3) \\ &= \frac{n-1}{n}\hat{\mathbf{S}}_h + \overline{\mathbf{h}}\overline{\mathbf{h}}^T\end{aligned}$$

is a sum of the positive-definite matrix $\frac{n-1}{n}\hat{\mathbf{S}}_h$ and the positive semi-definite matrix $\overline{\mathbf{h}}\overline{\mathbf{h}}^T$, and hence positive-definite. Using (3) and applying the formula (13.72) in Puntanen et al. [16, p. 301] to $\frac{n-1}{n}\hat{\mathbf{S}}_h + \overline{\mathbf{h}}\overline{\mathbf{h}}^T$ now yields

$$\begin{aligned}n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})^{-1}\overline{\mathbf{h}} = n \cdot (&\overline{\mathbf{h}}^T (\frac{n-1}{n}\hat{\mathbf{S}}_h)^{-1}\overline{\mathbf{h}} \\ &- \frac{(\overline{\mathbf{h}}^T (\frac{n-1}{n}\hat{\mathbf{S}}_h)^{-1}\overline{\mathbf{h}})^2}{1 + \overline{\mathbf{h}}^T (\frac{n-1}{n}\hat{\mathbf{S}}_h)^{-1}\overline{\mathbf{h}}}).\end{aligned}$$

The second statement follows after a little algebra. The last statement follows from the fact, that the function $f(x) = \frac{x}{\frac{n-1}{n} + \frac{x}{n}}$ is strictly increasing for every $n > 1$ in $x \geq 0$. Thus, the extrema of quadratic form $x = n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1}\overline{\mathbf{h}}(\mathbf{e})$ over $\mathbf{I}_{ex}$ and the extrema of $f(x)$ over $\mathbf{I}_{ex}$ are attained by the same values in $\mathbf{I}_{ex}$. ■

Theorem 4 effectively reduces the case $\hat{\boldsymbol{\Omega}}_h = \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})$ to the case $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$, which is solvable by quadratic programming.

To extend the Sargan-Hansen test to the case of an external interval $\mathbf{I}_{ex}$, it is necessary to consider the distributional properties of the test statistic interval $[\underline{\chi^2}, \overline{\chi^2}]$. Each value $\mathbf{e} \in \mathbf{I}_{ex}$ can be specified correctly or incorrectly. If it is specified correctly, $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}(\mathbf{e}) \xrightarrow{d} \chi^2_{p_2}$, since the results of Section 2.1 apply. If it is not specified correctly, $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}(\mathbf{e}) \xrightarrow{d} \infty$ [4, p. 248], showing the inherent point value assumption. Only for values in a shrinking neighborhood around the true value, i.e, $\mathbf{e} = \mathbf{e}_0 + \boldsymbol{\delta}/n$, where $\mathbf{e}_0$ is the correctly specified value and $\boldsymbol{\delta}$ is a constant representing the bias, the asymptotic distribution of the test statistic is a noncentral $\chi^2_{p_2}-$distribution [4, p. 249]. The noncentral $\chi^2_{p_2}-$ distribution with the noncentrality parameter λ is denoted by $\chi^2_{p_2}(\lambda)$. The interval $\mathbf{I}_{ex}$ is assumed to be constant because it is constructed outside the data, so the problem of degenerate asymptotic distributions arises. To avoid this problem, the focus is on $\underline{\chi^2}$, the minimum value of the test statistic over $\mathbf{I}_{ex}$, using the heuristic that it should not go to ∞ if $\mathbf{e}_0 \in \mathbf{I}_{ex}$. To justify this decision and to develop a test based on $\underline{\chi^2}$, two arguments are given.

First, the task is to decide whether an external interval $\mathbf{I}_{ex}$ is coherent with the data, i.e. whether it contains a value that is 'close enough' to its sample equivalent. If a test decides that this is false for $\mathbf{I}_{ex}$, it should also decide that this is false for all intervals contained in $\mathbf{I}_{ex}$ as well. For example, if a test decides that the true value is negative, one should conclude that the test would also decide that it is not in $[0, 1]$. This requirement is satisfied when $\underline{\chi^2}$ is used as a single test statistic, because if $\underline{\chi^2}$ is greater than a critical value, then all values within $[\underline{\chi^2}, \overline{\chi^2}]$ are greater than it. Under the null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$, this critical value could be derived from the central χ^2-distribution to account for the fact that any value within $[\underline{\chi^2}, \overline{\chi^2}]$ could be the true one.

Second, this decision rule (reject the null hypothesis if $\underline{\chi^2}$ is greater than a critical value resulting from the central χ^2-distribution) amounts to a $\Gamma-$maximin decision rule [8, p. 193] for choosing the p-value. To recognize this, the corresponding set of gambles and the credal set must be specified. For an observed test statistic $\chi^2_{\mathbf{e}} \in [\underline{\chi^2}, \overline{\chi^2}]$, its

p-value is the probability of the event $\{\chi^2 > \chi^2_{\mathbf{e}}\}$ under the validity of the null hypothesis, where $\mathbf{e}$ is fixed. Therefore, the indicators of the events $\{\chi^2 > \chi^2_{\mathbf{e}}\}$ for all $\mathbf{e} \in \mathbf{I}_{ex}$ form the set of gambles. The possible asymptotic distributions for $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})$ under the null hypothesis are $\chi^2_{p_2}(\lambda)$ for $\lambda \in [0, \infty)$, so these distributions form the credal set. Now, the probabilities $P_{\chi^2_{p_2}(\lambda)}(\{\chi^2 > \chi^2_{\mathbf{e}}\})$ are increasing in λ if $\chi^2_{\mathbf{e}}$ is fixed [6], so the lower probability is reached at $\lambda = 0$. Note that this is equivalent to cumulative distribution functions that decrease pointwise in λ. But $\chi^2_{p_2}(0)$ is just the central $\chi^2_{p_2}-$distribution. Note that the above degenerate distributions at $n \to \infty$ are the limits for $\lambda \to \infty$ and thus the lower probability includes these 'distributions' as well. Finally, the lower probability $P_{\chi^2_{p_2}}(\{\chi^2 > \chi^2_{\mathbf{e}}\})$ is maximal at $\chi^2_{\mathbf{e}} = \underline{\chi^2}$, because

$$\{\chi^2 > \chi^2_{\mathbf{e}}\} \subset \{\chi^2 > \underline{\chi^2}\}$$

for all $\mathbf{e} \in \mathbf{I}_{ex}$.

Taken together, we calculate the maximum of the respective lower probabilities of the events $\{\chi^2 > \chi^2_{\mathbf{e}}\}$ for $\mathbf{e} \in \mathbf{I}_{ex}$ and compare it with the significance level α. Thus, the Sargan-Hansen test based on external intervals is

1. $P_{\chi^2_{p_2}}(\{\chi^2 > \underline{\chi^2}\}) \geq \alpha$
   $\Rightarrow$ maintain null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$
2. $P_{\chi^2_{p_2}}(\{\chi^2 > \underline{\chi^2}\}) < \alpha$
   $\Rightarrow$ reject null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$.

This test is conservative, but ensures that the asymptotic significance level is at most α, regardless of which value in $\mathbf{I}_{ex}$ is the true value under the null hypothesis.

The results obtained so far are asymptotic in nature. To derive a test for information-data conflict in small samples, distributional assumptions for $\overline{\mathbf{h}}(\mathbf{e})$ are required. So suppose that $\overline{\mathbf{h}}(\mathbf{e})$ is normally distributed for each $\mathbf{e} \in \mathbf{I}_{ex}$. If $\overline{\mathbf{h}}(\mathbf{e}) = \hat{\mathbf{h}} - \mathbf{e}$ holds, as assumed above for the application of quadratic programming, it is sufficient to assume that the sampling moment $\hat{\mathbf{h}}$ is normally distributed. Under this normality assumption, the test statistic $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})$ at a fixed $\mathbf{e} \in \mathbf{I}_{ex}$ has the scaled noncentral F-distribution $\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)$, where λ is again the noncentrality parameter [15, p. 889].[2] If the cumulative distribution functions of $\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)$ for $\lambda \in [0, \infty)$ are pointwise decreasing in λ, the same arguments used in the above construction of the Sargan-Hansen test based on external intervals can be applied. Now, the cumulative distribution function of $F_{p_2,n-p_2}(\lambda)$ is decreasing in λ [6]. This property carries over to $\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)$ since the scaling by $\frac{(n-1)p_2}{n-p_2}$ is a strictly increasing transformation and can be inverted using the definition of pushforward measures, i.e,

$$P_{\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)}(A) = P_{F_{p_2,n-p_2}(\lambda)}(\frac{n-p_2}{(n-1)p_2} \cdot A).$$

Taken together, the test for information-data conflict in small samples based on $\hat{\mathbf{S}}_h$ and the assumption of normality is

1. $P_{F_{p_2,n-p_2}}(\{\chi^2 > \frac{n-p_2}{(n-1)p_2}\underline{\chi^2}\}) \geq \alpha$
   $\Rightarrow$ maintain null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$
2. $P_{F_{p_2,n-p_2}}(\{\chi^2 > \frac{n-p_2}{(n-1)p_2}\underline{\chi^2}\}) < \alpha$
   $\Rightarrow$ reject null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$.

At first glance, one might think that the choice of $\hat{\boldsymbol{\Omega}}_h$ is always important when working with small samples. This is not necessarily the case, as we will show now. From the fact that the function $f(x)$ from the proof of Theorem 4 is strictly increasing for $x \geq 0$, it follows that for every $c \geq 0$ the inequality $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e}) > c$ is satisfied iff

$$n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})^{-1} \overline{\mathbf{h}}(\mathbf{e}) = f(n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})) > f(c)$$

holds. Both inequalities represent the same event in the common underlying probability space, and both are assigned the same probability. Therefore, c and $f(c)$ can be interpreted as quantiles with the same level α, and the small sample tests for information-data conflict based on $\hat{\mathbf{S}}_h$ and $\hat{\boldsymbol{\Sigma}}_h$, respectively, are the same.

The tests developed in this section are either asymptotic or assume a normal distribution. Therefore, it is important to check their properties in small samples when there is no normal distribution. Since a conservative $\Gamma-$Maximin decision rule was used, it would be interesting to compare the expected type I error rates with the significance level α. On the one hand, the use of lower probabilities could correct for the small sample bias of the asymptotic test or for the errors caused by deviations from the normal distribution. This is due to the fact that all distributions in the credal set and their convex combinations are undercut by the lower probability. On the other hand, the type I error rate could become very low if $\mathbf{I}_{ex}$ is very broad, possibly leading to low power of the tests for a fixed n. Regarding to the use of multiple external moments, the question is how this affects the type I error rate and the power of the tests. The inclusion of additional moments increases the degrees of freedom p_2, which may increase the critical values for a given significance level α. Thus, if the interval of the added moment includes or is close to the true value, the power may decrease. On the other hand, if the interval of the added moment is far from the true value, this could increase the power. We will analyze these issues through a short simulation study.

[2] The notation of Phillips [15] is very different from ours, so we explain it here: Their T is our n, their p is 1 in our case, and their q is our p_2.

3. A Simulation Study to Investigate Small Sample Properties

First, we choose sample sizes $n = 30$ and $n = 50$ so that each scenario occurs twice and the effect of increasing sample size can be analyzed. Based on Example 1, we use a simple linear regression model under Gauss-Markov assumptions and normally distributed errors. The slope is $\beta_2 = 1$ and the intercept is $\beta_1 = 16$. The sample values for the independent variable x and the dependent variable y are drawn i.i.d. as $x \sim N(4, 4)$ and $y = \beta_1 + \beta_2 x + \epsilon$ with $\epsilon \sim N(0, 60)$, where the second terms (4 and 60) are the variances. In these settings, the actual correlation between x and y is 0.25, which is low but quite typical for applied research, e.g, in psychology. The sample values are denoted by x_i and y_i for $i = 1, \ldots, n$. As external information, the moments $E(y)$, $E(x)$, and $\mathrm{Var}(y)$ are used individually or in combination of two or more of them, resulting in 7 moment scenarios. For $\mathrm{Var}(y)$, the moment function $h(\mathbf{z}) = \frac{n}{n-1}(y - \bar{y})^2 - e$ is used, where $\bar{y}$ is the sample mean of y and $\frac{n}{n-1}$ corrects for degrees of freedom. Note that using $\mathrm{Var}(y)$ leads to $\overline{h}(e) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 - e$, which is not normally distributed. Finally, two scenarios are chosen with respect to $\mathbf{I}_{ex}$ to investigate the type I error and the power, respectively. For the first scenario, $\mathbf{I}_{ex} = [0.95 \cdot \mathbf{e}_0, 1.05 \cdot \mathbf{e}_0]$ and for the second, $\mathbf{I}_{ex} = [1.2 \cdot \mathbf{e}_0, 1.3 \cdot \mathbf{e}_0]$.

To analyze the effect of the proximity of $\mathbf{I}_{ex}$ to the true value on the power of the tests, we use distributions for x and y that differ in terms of their standardized mean difference. To justify this, note that the square root of the test statistic can be simplified when using only one of the selected moments, as follows:

$$\sqrt{n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})} = \sqrt{n}\frac{|\hat{h} - e|}{\sqrt{\hat{\omega}_h}}, \tag{4}$$

where all expressions are not written in bold because they are now single-valued. Now, (4) resembles a t-test statistic and the typical effect size used for this test statistic is the standardized mean difference $d = \frac{|e_0 - e|}{\sqrt{\mathrm{Var}(h(\mathbf{z}))}}$ [5]. The value in $\mathbf{I}_{ex}$ that is closest to $\mathbf{e}_0$ is $1.2 \cdot \mathbf{e}_0$. For the Sargan-Hansen test based on external intervals using $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$, it holds for $E(x)$ that $d = \frac{|4 - 1.2 \cdot 4|}{2} = 0.4$, a small effect size, and for $E(y)$ that $d = \frac{|20 - 1.2 \cdot 20|}{8} = 0.5$, a medium effect size [5]. Thus, using $E(y)$ alone should result in higher power than using $E(x)$ alone. For $\mathrm{Var}(y)$ the calculation of d is a bit more complex. Under the above conditions, $\frac{n}{n-1}(y - \bar{y})^2$ has a scaled χ^2_1−distribution. However, d is scale invariant, so we can assume without loss of generality that $\frac{n}{n-1}(y - \bar{y})^2$ is χ^2_1− distributed. This leads to $d = 0.2\frac{1}{\sqrt{2}} = 0.1414$, which is below the threshold for small effects according to Cohen [5]. Note that the effect size for $\mathrm{Var}(y)$ does not depend on the value of any moment, which is a consequence of using a normally distributed y.

Taken together, these are 2 (sample sizes) × 7 (moment combinations) × 2 (choices of $\mathbf{I}_{ex}$) = 28 scenarios. For each scenario, the rejection rates of the null hypothesis are calculated for three tests, namely the Sargan-Hansen test based on external intervals using $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$ (abbreviated $\mathbf{SH}(\hat{\mathbf{S}}_h)$), the Sargan-Hansen test based on external intervals using $\hat{\boldsymbol{\Omega}}_h = \hat{\boldsymbol{\Sigma}}_h$ (abbreviated $\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$), and the small sample test for information-data conflict (**IDC**). The significance level is always set to $\alpha = 0.05$. For $\mathbf{SH}(\hat{\mathbf{S}}_h)$ and **IDC**, the test statistic $\underline{\chi^2}$ is computed using quadratic programming as described in Section 2.2, and for $\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$ it is computed using Theorem 4.

Simulations were performed in R, version 4.2.1 [17]. The R package *quadprog* [19] was used to perform quadratic programming. To calculate rejection rates, each simulation scenario was repeated 10000 times. The associated R script can be found in the electronic supplementary material. The results concerning type I error rates are presented in Table 1 and Table 2 and the results concerning power are presented in Table 3 and Table 4.

Table 1: Type I error rates for $n = 30$

Moments	$\mathbf{SH}(\hat{\mathbf{S}}_h)$	$\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$	**IDC**
$E(y)$	0.0085	0.0061	0.0065
$\mathrm{Var}(y)$	0.0752	0.0657	0.0674
$E(x)$	0.0162	0.0121	0.0124
$E(y), \mathrm{Var}(y)$	0.0573	0.0429	0.0460
$\mathrm{Var}(y), E(x)$	0.0578	0.0456	0.0482
$E(y), E(x)$	0.0115	0.0057	0.0069
$E(y), \mathrm{Var}(y), E(x)$	0.0487	0.0302	0.0345

Table 2: Type I error rates for $n = 50$

Moments	$\mathbf{SH}(\hat{\mathbf{S}}_h)$	$\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$	**IDC**
$E(y)$	0.0055	0.0043	0.0044
$\mathrm{Var}(y)$	0.0554	0.0502	0.0508
$E(x)$	0.0089	0.0070	0.0075
$E(y), \mathrm{Var}(y)$	0.0330	0.0293	0.0302
$\mathrm{Var}(y), E(x)$	0.0358	0.0298	0.0305
$E(y), E(x)$	0.0041	0.0029	0.0031
$E(y), \mathrm{Var}(y), E(x)$	0.0264	0.0195	0.0213

Table 3: Power for $n = 30$

Moments	SH($\hat{\mathbf{S}}_h$)	SH($\hat{\boldsymbol{\Sigma}}_h$)	IDC
$E(y)$	0.7811	0.7519	0.7564
$\mathrm{Var}(y)$	0.2530	0.2341	0.2364
$E(x)$	0.5994	0.5604	0.5680
$E(y), \mathrm{Var}(y)$	0.7421	0.6687	0.6852
$\mathrm{Var}(y), E(x)$	0.6027	0.5209	0.5409
$E(y), E(x)$	0.8116	0.7404	0.7576
$E(y), \mathrm{Var}(y), E(x)$	0.8076	0.6885	0.7189

Table 4: Power for $n = 50$

Moments	SH($\hat{\mathbf{S}}_h$)	SH($\hat{\boldsymbol{\Sigma}}_h$)	IDC
$E(y)$	0.9416	0.9339	0.9355
$\mathrm{Var}(y)$	0.2808	0.2677	0.2699
$E(x)$	0.8048	0.7877	0.7906
$E(y), \mathrm{Var}(y)$	0.9040	0.8807	0.8860
$\mathrm{Var}(y), E(x)$	0.7929	0.7526	0.7626
$E(y), E(x)$	0.9607	0.9478	0.9508
$E(y), \mathrm{Var}(y), E(x)$	0.9499	0.9219	0.9302

4. Discussion

4.1. Summary of the Simulation Results

All type I error rates were below the α significance level, except in the cases where $\mathrm{Var}(y)$ was used. When $\mathrm{Var}(y)$ was used alone, the type I error rates of all tests were above α, indicating that the tests could not compensate for deviations from the normal distribution. A possible explanation could be that $\mathbf{I}_{ex}$ was not large enough. In practice, however, $\mathbf{I}_{ex}$ is determined externally and should not be expanded carelessly, since a broader $\mathbf{I}_{ex}$ would result in lower power. Nevertheless, a larger sample size would be a possible solution, since in all our scenarios an increase in sample size resulted in lower Type I error rates and higher power. When $\mathrm{Var}(y)$ was used in combination with other moments, the type I error rates were below α at $n = 30$ for the tests **SH($\hat{\boldsymbol{\Sigma}}_h$)** as well as **IDC**, and at $n = 50$ for all tests, showing that combinations of normally and non-normally distributed sample moments can improve the type I error rate. When in doubt, a simulation of the practical scenario should be performed to analyze whether the significance level is exceeded. For the scenarios using $E(y)$ alone, the smallest type I error rates were 0.0061 for $n = 30$ and 0.0043 for $n = 50$, showing that the tests can be much more conservative than the significance level would suggest. This is the expected consequence of using the conservative $\Gamma-$maximin rule. In all moment scenarios, there was a clear order of the tests in terms of type I error rate. The test **SH($\hat{\mathbf{S}}_h$)** always had higher type I error rates than **IDC** and **IDC** always had higher error rates than **SH($\hat{\boldsymbol{\Sigma}}_h$)**.

As for the power of the tests, their order corresponds to the order of the type I error rate. In all moment scenarios, **SH($\hat{\mathbf{S}}_h$)** had the highest power, followed by **IDC** and **SH($\hat{\boldsymbol{\Sigma}}_h$)**. As expected, $E(y)$ yielded the highest power when used alone, followed by $E(x)$ and $\mathrm{Var}(y)$, clearly reflecting the effect size d calculated in Section 3. With powers ranging from 0.7519 to 0.7811 for $n = 30$ and from 0.9339 to 0.9416 for $n = 50$, the moment $E(y)$ shows that the use of an external interval does not erase all of the power of the tests in our simulation study. Even for the small effect size exerted by the moment $E(x)$, the power ranged from 0.7877 to 0.8048 for $n = 50$. However, using combinations of moments does not always result in higher power. Combinations with $\mathrm{Var}(y)$ resulted in lower power than the same combinations without $\mathrm{Var}(y)$, with the sole exception of $\mathrm{Var}(y)$ and $E(x)$ for the test **SH($\hat{\mathbf{S}}_h$)** in the case $n = 30$. The maximum power reduction due to the inclusion of $\mathrm{Var}(y)$ was 0.0832 for $n = 30$ and 0.0532 for $n = 50$, respectively, for the moment $E(y)$ for the test **SH($\hat{\boldsymbol{\Sigma}}_h$)**. This reduction property is explained by the very small effect size when using $\mathrm{Var}(y)$, which causes the increase in the critical value due to the higher degrees of freedom p_2 to exceed the expected increase in the test statistic due to the inclusion of $\mathrm{Var}(y)$. Only for the test **SH($\hat{\boldsymbol{\Sigma}}_h$)** and $n = 30$ did the combination of $E(x)$ and $E(y)$ result in lower power than using $E(y)$ alone. In all other cases, however, the combination of $E(x)$ and $E(y)$ led to an increase in power, although not as pronounced, since for $n = 30$ the power only increased by a maximum of 0.0305.

Despite the conservative $\Gamma-$maximin decision rule used to construct them, the tests had good power for small sample sizes at small and medium effect sizes in our simulation scenarios. However, when a moment is not normally distributed, one should be very careful with its use, as it may lead to too high a type I error rate when used alone and to a lower power when used in combination with other moments. The simulations suggest that in scenarios such as those used here, one should select the single normally distributed moment with the largest effect size rather than using multiple moments in combination. In particular, deviations from the normal distribution, which are likely to occur frequently in practice, need to be considered in further research.

4.2. Outlook

Most importantly, the robustness of the tests to deviations from the normal distribution should be further investigated. If only the variance of y is used as the external moment, one should correct for type I error rates by deriving the

specific distribution of the test statistic in this case, given normally distributed variables.

Since the $\Gamma-$maximin decision rule is conservative, one could analyze the effect of using other p-value decision rules on the tests developed here. There are some issues regarding this endeavor. First, the upper probabilities of the events in Section 2.2 would effectively be 1. This is because the credal set includes distributions with arbitrarily large noncentrality parameters. These distributions shift the probability mass to infinity, while the interval of test statistics for a fixed n is bounded almost surely. One way to deal with this problem would be to set an upper bound on the noncentrality parameter for a fixed n. Second, note that using a $\Gamma-$maximax decision rule would result in higher p-values and thus an even more conservative test. A more liberal procedure would be to minimize the lower probabilities. Since the external interval necessarily contains values that are not the true moment value, the p-values for these would asymptotically be 0, resulting in a test that always rejects the null hypothesis even if the interval contains the true moment value.

Another way to construct a more liberal test would be to use different significance levels, possibly increasing with n, since the actual type I error rates appear to be low even at $n = 50$. However, one should keep in mind that the type I error rate depends on where the true value lies within the external interval. Therefore, it would be interesting to analyze the type I errors for several locations of the true value to calculate the worst case type I error.

Although the Sargan-Hansen test has been reduced to a Wald test as shown here, there are still ways to use information about model parameters in the tests constructed in this paper, for example implementing the OLS estimator (a function based only on the data) and an external interval that represents the external information about the regression parameter. It would be interesting to study the properties of such 'indirect' model moment conditions. In addition, other tests or frameworks for using moment-type external information could be used and compared to the tests developed here, such as the Empirical Likelihood framework [13].

Finally, the results derived here may also be useful when working with interval-valued information about moments in other research areas, since the $\Gamma-$maximin decision rule for the p-values is based on the stochastic order of the underlying family of distributions of a test statistic. This is true for many econometric and psychometric procedures, such as the Wald test for general linear and nonlinear hypotheses, the likelihood ratio test, and the Langrange multiplier test, since their test statistics are asymptotically chi-squared distributed under the null hypothesis (see Cameron and Trivedi [4] for more details). The algebraic results could help to derive analytical formulas for the externally informed estimators of Imbens and Lancaster [9] and combine them with the use of external intervals. Since these estimators are more efficient than OLS estimators and since external intervals are a more realistic and robust representation of external information, there could be an interesting interaction between the two.

Acknowledgments

The author would like to thank the three ISIPTA reviewers for their thoughtful questions and comments. They helped to improve the didactic quality of this paper and to sharpen its statistical and theoretical arguments.

References

[1] Seung C. Ahu and Peter Schmidt. A separability result for gmm estimation, with applications to gls prediction and conditional moment tests. *Econometric Reviews*, 14(1):19–34, 1995. URL https://doi.org/10.1080/07474939508800301.

[2] Thomas Augustin, Gero Walter, and Frank P. A. Coolen. Statistical inference. In *Introduction to Imprecise Probabilities*, chapter 7, pages 135–189. John Wiley & Sons, Ltd, 2014. URL https://doi.org/10.1002/9781118763117.ch7.

[3] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Inc., 1994. URL https://doi.org/10.1002/9780470316870.

[4] Adrian Cameron and Pravin Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005. URL https://doi.org/10.1017/CBO9780511811241.

[5] Jacob Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992. URL https://doi.org/10.1037/0033-2909.112.1.155.

[6] Bhaskar K. Ghosh. Some monotonicity theorems for chi square, F and t distributions with applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(3):480–492, 1973. URL https://doi.org/10.1111/j.2517-6161.1973.tb00976.x.

[7] Lars P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. URL https://doi.org/10.2307/1912775.

[8] Nathan Huntley, Robert Hable, and Matthias C. M. Troffaes. Decision making. In *Introduction to Imprecise Probabilities*, chapter 8, pages 190–206. John Wiley & Sons, Ltd, 2014. URL https://doi.org/10.1002/9781118763117.ch8.

[9] Guido W. Imbens and Tony Lancaster. Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4):655–680, 1994. URL https://doi.org/10.2307/2297913.

[10] Jan F. Kiviet and Sebastian Kripfganz. Instrument approval by the Sargan test and its consequences for coefficient estimation. *Economics Letters*, 205, 2021. URL https://doi.org/10.1016/j.econlet.2021.109935.

[11] Pavel S. Knopov and Arnold S. Korkhin. *Regression Analysis Under A Priori Parameter Restrictions*, volume 54 of *Springer Optimization and Its Applications*. Springer Science & Business Media, New York, 2011. URL https://doi.org/10.1007/978-1-4614-0574-0.

[12] Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111–2245. Elsevier, 1994. URL https://doi.org/10.1016/S1573-4412(05)80005-4.

[13] Art B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988. URL https://doi.org/https://doi.org/10.2307/2336172.

[14] Paulo M. D. C. Parente and João M. C. Santos Silva. A cautionary note on tests of overidentifying restrictions. *Economics Letters*, 115(2):314–317, 2012. URL https://doi.org/10.1016/j.econlet.2011.12.047.

[15] Peter C. B. Phillips. The exact distribution of the Wald statistic. *Econometrica*, 54(4):881–895, 1986. URL https://doi.org/10.2307/1912841.

[16] Simo Puntanen, George P. H. Styan, and Jarkko Isotalo. *Matrix Tricks for Linear Statistical Models*. Springer Berlin Heidelberg, 2011. URL https://doi.org/10.1007%2F978-3-642-10473-2.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

[18] John D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26 (3):393–415, 1958. URL https://doi.org/10.2307/1907619.

[19] Berwin A. Turlach, Andreas Weingessel, and Cleve Moler. *quadprog: Functions to Solve Quadratic Programming Problems*, 2019. URL https://CRAN.R-project.org/package=quadprog. R package version 1.5-8.

[20] Zhiguo Xiao. Efficient gmm estimation with singular system of moment conditions. *Statistical Theory and Related Fields*, 4(2):172–178, 2020. URL https://doi.org/10.1080/24754269.2019.1653159.

Appendix B

Paper 2

Jann, M., & Spiess, M. (2024). Using external information for more precise inferences in general regression models. *Psychometrika*, *89*(2), 439–460. https://doi.org/10.1007/s11336-024-09953-w

PSYCHOMETRIKA—VOL. 89, NO. 2, 439–460
JUNE 2024
https://doi.org/10.1007/s11336-024-09953-w

USING EXTERNAL INFORMATION FOR MORE PRECISE INFERENCES IN GENERAL REGRESSION MODELS

MARTIN JANN
UNIVERSITY OF HAMBURG

MARTIN SPIESS
UNIVERSITY OF HAMBURG

Empirical research usually takes place in a space of available external information, like results from single studies, meta-analyses, official statistics or subjective (expert) knowledge. The available information ranges from simple means and proportions to known relations between a multitude of variables or estimated distributions. In psychological research, external information derived from the named sources may be used to build a theory and derive hypotheses. In addition, techniques do exist that use external information in the estimation process, for example prior distributions in Bayesian statistics. In this paper, we discuss the benefits of adopting generalized method of moments with external moments, as another example for such a technique. Analytical formulas for estimators and their variances in the multiple linear regression case are derived. An R function that implements these formulas is provided in the supplementary material for general applied use. The effects of various practically relevant moments are analyzed and tested in a simulation study. A new approach to robustify the estimators against misspecification of the external moments based on the concept of imprecise probabilities is introduced. Finally, the resulting externally informed model is applied to a dataset to investigate the predictability of the premorbid intelligence quotient based on lexical tasks, leading to a reduction of variances and thus to narrower confidence intervals.

Key words: external information, generalized method of moments, imprecise probabilities.

1. Introduction

When planning new empirical studies, researchers are confronted with a variety of information from previous studies, including statistical quantities such as means, variances or confidence intervals. However, this external information is mostly used qualitatively, i.e., to develop new theories, and rarely in a quantitative way, i.e., to estimate parameters. One advantage of using external information to estimate a parameter is that some parameter values can be excluded or considered less likely than without the external information, potentially leading to more efficient estimators. The usage of informed prior distributions, where the external information can be used to specify (certain aspects of) the prior distribution, is well known in Bayesian statistics (Bernardo & Smith, 1994). The underlying goal for its use must be clear. On the one hand, external information can facilitate the fitting or tuning of a model. On the other hand, it can make estimators more robust or efficient. This paper aims to achieve the latter of the two goals. Bayesian statistics refers to this as statistical elicitation (Kadane & Wolfson, 1998). The objective is to translate expert knowledge into a prior distribution. Therefore, many psychological biases, such as judgment by representativeness, availability, anchoring, adaptation, or hindsight bias and the intentional misleading by experts, must be considered. It should be noted that the aim is not

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-024-09953-w.

Correspondence should be made to Martin Jann, Department of Psychology, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany. Email: martin.jann@uni-hamburg.de

to achieve objectivity but to ensure a proper statistical representation of subjective knowledge (Garthwaite et al., 2005; Lele & Das, 2000). However, we believe that in applied psychological research, the researcher is usually the one who selects the external information, but is susceptible to the same psychological biases, e.g., in deciding which studies to include. Moreover, the difficulties in eliciting a (multivariate) prior distribution are well documented (Garthwaite et al., 2005, pp. 686–688). The method proposed in this paper allows a simplification of the elicitation compared to Bayesian statistics, since only moments need to be elicited. The elicitation of moments has been well studied for correlations, means, medians, or variances (Garthwaite et al., 2005). In Bayesian elicitation, there are several possible prior distributions for these externally given moments, e.g., with the same expected value or the same correlation, leading to different posterior distributions and thus potentially different results. This problem of prior sensitivity was addressed by Berger (1990) and led to work on robust Bayesian analysis (for an overview, see Insua & Ruggeri, 2000). However, it is somewhat arbitrary to choose the class of distributions for which one wants to make the analysis robust (Garthwaite et al., 2005, p. 695). In our framework, no restriction to a particular class of distributions is required, since it relies solely on moment information and a central limit theorem.

Another important point is that external information may not in general be precise and correct. As nearly all of the external quantities are estimates themselves, they are at least prone to sampling variation. If the external information is not correct (e.g., due to poor sampling or measurement protocols), its use can lead to biased conclusions that may even be worse than without external information. To address this problem, we suggest using an interval for the external information instead of point values, enabling researchers to incorporate any uncertainty about the external moments into the analysis. Inserting external intervals into estimators results in the imprecise probabilistic concept of *feasible probability (F-probability)* discussed in Sect. 4 (Augustin et al., 2014; Weichselberger, 2001). This approach provides an alternative way to enhance the robustness of elicitation compared to the classical Bayesian paradigm: Using intervals can reflect uncertainty about moments, and the resulting inference is still coherent if the interval contains the true value. However, researchers must be cautious of and avoid overconfidence bias when eliciting intervals; that is, the tendency to select intervals that are too narrow to represent current uncertainty (Winman et al., 2004). A test of the latter assumption is available, more specifically a test of the compatibility of the external interval and the data, which could serve as a pretest before applying the methods proposed here (Jann, 2023).

The insertion of intervals into estimators resembles creating fuzzy numbers (Kwakernaak, 1978; Zadeh, 1965), for which generalizations of traditional statistical methods already exist. This is particularly true for the special case of triangular numbers (Buckley, 2004). The possibility distributions induced by triangular numbers constitute special cases of imprecise probabilities and are constructed based on only one distribution (Augustin et al., 2014, pp. 84–87). This is the key difference between triangular numbers and F-probabilities, since the latter are constructed from a set of possible probability distributions, which can enhance the robustness of the outcomes compared to constructions based on only one distribution. Another difference lies in the fact that triangular numbers are constructed by varying the confidence probability of a confidence interval based on the estimator, while the external interval we use in this paper is fixed. Moreover, there is no probabilistic statement about the values within that interval.

In the present study, we analyze the frequentist properties of estimators if external information is used, that can be expressed as moment conditions and thus does not use complete distributions as prior information. To our knowledge, there is no general framework for robustly incorporating such quantitative external information into frequentist analysis. Since this would offer the advantage of improving upon classical inference procedures widely used in psychology, our goal is to present such a framework. The use of these external moment conditions in addition to the moment conditions used to estimate the model parameters leads to an overidentified system of moment

conditions. The main idea to find well performing estimators for such "externally" overidentified systems is the framework of the Generalized Method of Moments (GMM) (Hansen, 1982). This idea has already been used in the econometric literature, for example, by Imbens and Lancaster (1994) who combine micro- and macro-economic data and by Hellerstein and Imbens (1999) by constructing weights for regression models based on auxiliary data. A different yet related way to incorporate external moment-information is the empirical likelihood approach (Owen, 1988). This technique is quite frequently used in the literature, for example, in finite population estimation (Zhong & Rao, 2000) and for externally informed generalized linear models (Chaudhuri et al., 2008). Both approaches have in common that the use of external information may increase the efficiency of an estimator and/or reduce its bias.

Actually, in Sect. 3, we show that there will always be a variance reduction, if the external moment conditions and the ones for the model are correlated and if the covariance matrix of all moment conditions is positive definite. As the GMM allows the estimation of a large class of models, and many statistical measures like proportions, means, variances and covariances are statistical moments, the range of possible applications is large but far from being implemented in psychological research. For a multiple linear model, we derive the estimators analytically in Sect. 3. The use of imprecise probabilities will increase the overall variation of the estimator, and moreover, the effect of the variance reduction will decrease. As we will demonstrate, however, variance reduction will still be possible while increasing the robustness of the estimation. The proposed method and techniques allow more precise and robust inferences, which is particularly relevant in small samples. To illustrate the small sample performance of the externally informed models in multiple linear models, a simulation study is presented in Sect. 5. An application to a real data set analyzing the relation of premorbid (general) intelligence and performance in lexical tasks (Pluck & Ruales-Chieruzzi, 2021) is presented in Sect. 6.

2. Externally Informed Models

In a first step, we assume that precise external information is available, an assumption that will be relaxed in Sect. 4. Throughout, we assume that all variables will be considered as random variables if not given otherwise. For notational clarity, we will always write single-valued random variables in italic small letters. Vectors as well as vector-valued functions will be written in small bold letters and matrices in bold capital letters.

Although the basic concepts are presented in the following section, for the class of general regression models, we will consider the family of linear models for their illustration in a concrete class of models due to their frequent use. Note that, for example, ANOVA models are special cases of this model, however, with fixed factors instead of random covariates. Nevertheless, the results derived in this paper carry over to these models.

Let $\mathbf{z} = (z_1, \ldots, z_p)^T$ be a real-valued random vector and $\mathbf{z}_i$, $i = 1, \ldots, n$, be i.i.d. random vectors distributed like $\mathbf{z}$, representing the data. Suppose we want to fit a regression model to this data set with fixed parameter $\boldsymbol{\theta} \in \mathbb{R}^p$, where the adopted model reflects the interesting aspects of the true data-generating process and $\boldsymbol{\theta}_0$ is the true parameter value. In linear regression models, the parameter of scientific interest is usually the parameter of the mean structure denoted as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ with true value $\boldsymbol{\beta}_0$. The notation $\boldsymbol{\beta}$ will only be used for linear regression models, while we will use $\boldsymbol{\theta}$ to denote the regression coefficients in general regression models. The random vector $\mathbf{z}$ is given by $\mathbf{z} = (\mathbf{x}^T, y)^T$ with random explanatory variables $\mathbf{x} = (x_1, \ldots, x_p)^T$ and dependent variable y. Accordingly, the unit specific i.i.d. random vectors $\mathbf{z}$ are written as $\mathbf{z}_i = (\mathbf{x}_i^T, y_i)^T$ for $i = 1, \ldots, n$. Hence, the random $(n \times p)$-design matrix is $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, and we write $\mathbf{y} = (y_1, \ldots, y_n)^T$.

The multiple linear model can now be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$ with random error terms $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. As an illustration, suppose we want to investigate the effect of the explanatory variables *fluid intelligence* and *depression* on the dependent variable *mathematics skills*. We could design a study, in which fluid intelligence and math skills are measured via Cattell's fluid intelligence test, in short CFT 20-R, (x_2) and the number sequence test ZF-R (y), respectively (Weiss, 2006). Depression could be measured as a binary variable indicating if a person has a depression-related diagnosis (x_3). The model could be a linear multiple regression of the ZF-R score on the depression indicator and the CFT 20-R score for fluid intelligence. To include the intercept, x_1 is a degenerate variable with value 1.

In addition to the observed data and the assumptions justifying the model, we often have available external information like means, correlations or proportions, e.g., through official statistics, meta-analyses or already existing individual studies. In our applied example, there are various German norm groups for the CFT 20-R and the ZF-R, even for different ages (Weiss, 2006). Hence, we could always transform the results into scores with known expected value and variance, i.e. the CFT 20-R score can be transformed into an IQ-score based on a recent calibration sample from 2019, reported in the test manual (Weiss, 2019). Regarding the relation of fluid intelligence and math skills, a recent meta-analysis based on more than 370,000 participants in 680 studies from multiple countries suggests a correlation of $r = 0.41$ between the two variables (Peng et al., 2019). In addition, based on a study covering 87% of the German population aged at least 15 years, Steffen et al. (2020) report a prevalence of depression, defined as a F32, F33 or F34.1 diagnosis following the ICD-10-GM manual, of 15.7% in 2017.

Let us assume that these values can be interpreted as true population values, an assumption that will be relaxed later. Note that they have the form of statistical moments. For example, the observable depression prevalence is assumed to equal the expected value of the binary depression indicator (first moment), the mean (now considered as expected value) and variance of the test scores are set equal to the first moment and the second central moment, respectively, of the random variables CFT 20-R-score and ZF-R-score. Finally, the correlation is assumed to equal the mixed moment of the standardized CFT 20-R-score and ZF-R-score. Taking q to be the number of known external moments, we state

Definition 1. Let M be a statistical model. Further let $\mathbf{u}$ be a $(q \times 1)$-vector of statistical moment expressions and $\boldsymbol{\mu}_{\text{ex}}$ the corresponding $(q \times 1)$-vector of externally determined values for the statistical moments in $\mathbf{u}$. Then the model combining M and the conditions $\mathbf{u} = \boldsymbol{\mu}_{\text{ex}}$ is called *externally informed model*.

To illustrate the definition, we will use the applied example from above in which case the model M is a multiple linear regression model. Interpreting the norms for the dependent variable ZF-R from the calibration sample as population values, external knowledge about the corresponding moments, for example the means of ZF-R, is available. Let us assume that ZF-R is transformed into the IQ-scale. Then, if $\mathbf{u} = E(y)$ and $\boldsymbol{\mu}_{\text{ex}} = 100$, we get $E(\mathbf{y}) = 100 \times \mathbf{1}_n = E(\mathbf{X})\boldsymbol{\beta}_0$, where $\mathbf{1}_n$ is a $(n \times 1)$-vector of ones. Thus, $\mathbf{u} = \boldsymbol{\mu}_{\text{ex}}$ imposes conditions on $\boldsymbol{\beta}$.

3. Estimation and Properties of Externally Informed Models

3.1. Generalized Method of Moments with External Moments

The GMM approach (Hansen, 1982) allows to estimate (general) regression models and to incorporate external moments into the estimation (Imbens & Lancaster, 1994). To estimate the parameter of a general regression model, a "model moment function" $\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})$ must be given, which satisfies the conditions $E[\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})] = \mathbf{0}$ only for the true parameter value $\boldsymbol{\theta}_0$. The corresponding

"sample moment function" for $\mathbf{z}_i$ will be denoted as $\mathbf{m}(\mathbf{z}_i, \boldsymbol{\theta})$. In case of the linear regression model from Sect. 2, the model moment function corresponding to the method of Ordinary Least Squares (OLS) is $\mathbf{m}(\mathbf{z}, \boldsymbol{\beta}) = \mathbf{x}(y - \mathbf{x}^T\boldsymbol{\beta})$ (Cameron & Trivedi, 2005, p. 172). Given the model is correctly specified, for true parameter value $\boldsymbol{\beta}_0$, $E[\mathbf{m}(\mathbf{z}, \boldsymbol{\beta}_0)] = E[\mathbf{x}(y - \mathbf{x}^T\boldsymbol{\beta}_0)] = \mathbf{0}$ holds. Replacing these population model moment conditions by corresponding sample model moment conditions,

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}(\mathbf{z}_i, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right) = \frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \ ,$$

and solving these estimating equations for $\boldsymbol{\beta}$, leads to an estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$. The above conditions are identical to the estimating equations resulting from the least-squares or, if normality of the errors is assumed, the maximum likelihood method. Furthermore, the general classes of M- and Z-estimators can be written using estimating equations that have this moment form. This leads to broad applicability, since these classes, for example, include the median and quantiles (Vaart, 1998).

The possibly vector-valued "external moment function" will be denoted as $\mathbf{h}(\mathbf{z}) = \mathbf{u}(\mathbf{z}) - \boldsymbol{\mu}_{\text{ex}}$, where the functional form of $\mathbf{u}(\mathbf{z})$ depends on the external information included into the model. We assume that $\boldsymbol{\mu}_{\text{ex}} = E[\mathbf{u}(\mathbf{z})]$, so that $E[\mathbf{h}(\mathbf{z})] = \mathbf{0}$. If, for example, the expected value of y is known to be $E(y) = 100$, then $u(z) = y$, $\mu_{\text{ex}} = 100$ and $h(z) = y - 100$. The corresponding sample moment condition is $0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - 100)$ (Imbens & Lancaster, 1994).

To simplify the presentation, we define the combined moment function vector in general regression models as $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}) = [\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})^T, \mathbf{h}(\mathbf{z})^T]^T$ in what follows and assume that $E[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}_0)] = \mathbf{0}$ holds. Note that the number of moment conditions exceeds the number of parameters to be estimated, i.e. the externally informed model is overidentified. This means that there will in general be no estimator $\hat{\boldsymbol{\theta}}$ that solves the corresponding sample moment conditions $\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{0}$. To deal with the overidentification problem, we will use the GMM approach (Hansen, 1982), that finds an estimator as "close" as possible to a solution of the sample moment conditions. This is done by maximizing a quadratic form defined by a chosen symmetric, positive definite weighting matrix $\mathbf{W}$ in the moment functions of the sample. The efficiency of the estimator is affected by $\mathbf{W}$, and this can be chosen to maximize the asymptotic efficiency of the estimator in the class of all GMM-estimators based on the same sample moment conditions (Hansen, 1982). This optimal weighting matrix is $\mathbf{W} = \boldsymbol{\Omega}^{-1}$, where $\boldsymbol{\Omega} = E[\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)^T]$. However, this optimal $\mathbf{W}$ is unknown in practice and must be estimated by a consistent estimator $\hat{\mathbf{W}}$.

Definition 2. (Newey & McFadden, 1994, p. 2116) Let $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ be a vector-valued function with values in $\mathbb{R}^K$, that meets the moment conditions $E[\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)] = \mathbf{0}$. Further let $\hat{\mathbf{W}} \in \mathbb{R}^{K,K}$ be a positive-semidefinite, possibly random matrix, such that $(\mathbf{r}^T\hat{\mathbf{W}}\mathbf{r})^{1/2}$ is a measure of distance from $\mathbf{r}$ to $\mathbf{0}$ for all $\mathbf{r} \in \mathbb{R}^K$. Then, the **GMM-estimator** $\hat{\boldsymbol{\theta}}_{\text{ex}}$ is defined as the $\boldsymbol{\theta}$, which maximizes the following function:

$$\hat{Q}_n(\boldsymbol{\theta}) = -\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})\right]^T\hat{\mathbf{W}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})\right] \ .$$

The GMM approach provides consistent and normally distributed estimators under mild regularity conditions (Newey & McFadden, 1994, p. 2148) for a wide range of models, like linear or nonlinear, cross-sectional or longitudinal regression models. Note that we have not assumed that

$\hat{\mathbf{W}}$ is invertible because we will mainly derive asymptotic expressions based on $\mathbf{W}$ for which the invertibility of $\hat{\mathbf{W}}$ is not necessary. However, when deriving estimators, additional assumptions about invertibility must be made, which we explain in Sect. 3.2. Let $\mathbf{G} = E[\nabla_{\boldsymbol{\theta}}\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)]$ be a fixed matrix and $\mathbf{W}$ the optimal weighting matrix, then $\text{Var}(\hat{\boldsymbol{\theta}}_{\text{ex}}) = \frac{1}{n}(\mathbf{G}^T\mathbf{W}\boldsymbol{G})^{-1}$. This variance expression is not informative with respect to a possible efficiency gain of the GMM-estimator if external information is used. Hence, the following corollary explicitly shows the effect of the external information on the variance of $\hat{\boldsymbol{\theta}}_{\text{ex}}$.

Corollary 1. *Assume $\hat{\boldsymbol{\theta}}_M$ is the GMM-estimator based on the model estimating equations alone (ignoring the external moments), and that $\boldsymbol{m}(\boldsymbol{z}, \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ have the same dimension. Using the prerequisite $\boldsymbol{g}(\boldsymbol{z}, \boldsymbol{\theta}) = [\boldsymbol{m}(\boldsymbol{z}, \boldsymbol{\theta})^T, \boldsymbol{h}(\boldsymbol{z})^T]^T$ it follows, that $\boldsymbol{\Omega}$ has the block form*

$$\boldsymbol{\Omega} = \begin{pmatrix} E[\boldsymbol{m}(\boldsymbol{z}, \boldsymbol{\theta})\boldsymbol{m}(\boldsymbol{z}, \boldsymbol{\theta})^T] & E[\boldsymbol{m}(\boldsymbol{z}, \boldsymbol{\theta})\boldsymbol{h}(\boldsymbol{z})^T] \\ E[\boldsymbol{h}(\boldsymbol{z})\boldsymbol{m}(\boldsymbol{z}, \boldsymbol{\theta})^T] & E[\boldsymbol{h}(\boldsymbol{z})\boldsymbol{h}(\boldsymbol{z})^T] \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Omega}_M & \boldsymbol{\Omega}_R^T \\ \boldsymbol{\Omega}_R & \boldsymbol{\Omega}_h \end{pmatrix}$$

and that

$$\left\{E[\nabla_{\boldsymbol{\theta}}\boldsymbol{m}(\boldsymbol{z}, \boldsymbol{\theta}_0)]^T\right\}^{-1} \boldsymbol{\Omega}_R^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Omega}_R \{E[\nabla_{\boldsymbol{\theta}}\boldsymbol{m}(\boldsymbol{z}, \boldsymbol{\theta}_0)]\}^{-1} \tag{1}$$

A proof of Corollary 1 can be found in the supplementary materials online. Note that (1) shows that $\text{Var}(\hat{\boldsymbol{\theta}}_{\text{ex}})$ is equal to the conditional variance of $\hat{\boldsymbol{\theta}}_M$ under the external moment conditions, since the asymptotic distribution is normal. This equality shows why there is a reduction in the variance. Let the second term on the right-hand side of (1) be denoted by $\mathbf{D}$, then $\text{Var}(\hat{\boldsymbol{\theta}}_{\text{ex}})$ can be written as $\text{Var}(\hat{\boldsymbol{\theta}}_{\text{ex}}) = \text{Var}(\hat{\boldsymbol{\theta}}_M) - \mathbf{D}$. If $\mathbf{D}$ is nonnegative definite and not equal to $\mathbf{0}$, then including external moments leads to an expected efficiency gain in $\hat{\boldsymbol{\theta}}_{\text{ex}}$ as compared to $\hat{\boldsymbol{\theta}}_M$. That $\mathbf{D} \neq \mathbf{0}$ is nonnegative definite if $\boldsymbol{\Omega}_R \neq \mathbf{0}$ is easily seen by noting that $\boldsymbol{\Omega}_h^{-1}$ is positive definite and therefore can be written as $\boldsymbol{\Omega}_h^{-1} = \boldsymbol{\Omega}_h^{-1/2}\boldsymbol{\Omega}_h^{-1/2}$, where $\boldsymbol{\Omega}_h^{-1/2}$ is the positive definite square root of $\boldsymbol{\Omega}_h^{-1}$. Since $n\mathbf{D}$ can be written as the product of $\{E[\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0)]^T\}^{-1}\boldsymbol{\Omega}_R^T\boldsymbol{\Omega}_h^{-1/2}$ with its transpose, $\mathbf{D}$ is nonnegative definite. In summary, $\boldsymbol{\Omega}_R \neq \mathbf{0}$ is a necessary and sufficient condition for the presence of variance reduction based on Corollary 1. Finally, it should be noted that $\text{Var}(\hat{\boldsymbol{\theta}}_{\text{ex}})$ can consistently be estimated via the plug-in approach (e.g. Newey & McFadden, 1994, pp. 2171–2173) by replacing all unknown expected values by sample means.

3.2. The Externally Informed Multiple Linear Model

In linear models, $\hat{\boldsymbol{\theta}}_{\text{ex}}$ is denoted as $\hat{\boldsymbol{\beta}}_{\text{ex}}$. For analytical simplicity, in this section, we assume the Gauss–Markov assumptions hold, specifically $E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2, Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$ with $i, j = 1, \ldots, n$, and independence of the explanatory variables and the error terms $\boldsymbol{\epsilon}$. Furthermore, we assume the errors to be normally distributed in small samples. Analytical solutions to the estimating equations exist under these assumptions:

Theorem 1. *Let $\boldsymbol{H} = [\boldsymbol{h}(\boldsymbol{x}_1, y_1), \ldots, \boldsymbol{h}(\boldsymbol{x}_n, y_n)]^T$ be the $(n \times q)$ random matrix containing the externally informed sample moment functions and $\boldsymbol{1}_n$ a $(n \times 1)$-vector of ones. Further let $\hat{\boldsymbol{\Omega}}_h$ and $\hat{\boldsymbol{\Omega}}_R$ be consistent estimators of the corresponding matrices in Corollary 1. Then, the (consistent) externally informed OLS estimator is:*

$$\hat{\boldsymbol{\beta}}_{\text{ex}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\boldsymbol{\Omega}}_R^T\hat{\boldsymbol{\Omega}}_h^{-1}\boldsymbol{H}^T\boldsymbol{1}_n$$

and its variance is

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{ex}}) &= \mathrm{Var}(\hat{\boldsymbol{\beta}}) - \boldsymbol{D} \\ &= \frac{1}{n}\sigma^2\left[E\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right]^{-1} - \frac{1}{n}\left[E\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right]^{-1}\boldsymbol{\Omega}_R^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Omega}_R\left[E\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right]^{-1},\end{aligned}$$

where σ^2 is the variance of the error in the assumed linear model.

The proof of Theorem 1 is given in the supplementary materials online. Note that only the assumption of invertibility of $\hat{\boldsymbol{\Omega}}_h$ is made, which is weaker than the assumption that $\hat{\boldsymbol{\Omega}}$ is invertible. From Theorem 1, it is not immediately obvious which of several possibly available functions may lead to a variance reduction. Therefore, let us consider some external moment functions and their possible effects on the variance of $\hat{\boldsymbol{\beta}}_{\mathrm{ex}}$. Note that inclusion of external moment functions into the estimating equations may lead to expected efficiency gains only if $\boldsymbol{\Omega}_R^T = E[\mathbf{x}(y - \mathbf{x}^T\boldsymbol{\beta}_0)\mathbf{h}(\mathbf{x}, y)^T] = E[\mathbf{x}\epsilon\,\mathbf{h}(\mathbf{x}, y)^T] \neq \mathbf{0}$ holds.

Let the expressions σ_{x_j} and σ_y denote the population standard deviations of x_j and y, respectively, whereas $\sigma_{x_j,y}$ indicates the covariance of x_j and y. To denote the covariance vector $(\sigma_{x_1,x_j}, \ldots, \sigma_{x_p,x_j})^T$ of $\mathbf{x}$ and x_j the expression $\boldsymbol{\sigma}_{x_\cdot,x_j}$ is used, including $\sigma^2_{x_j}$ at the j-th position. Finally, $\rho_{x_j,y}$ is the population correlation of x_j and y.

First, consider some function $\mathbf{f}(\mathbf{x})$ of $\mathbf{x}$, i.e. $\mathbf{h}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - E[\mathbf{f}(\mathbf{x})]_{\mathrm{ex}}$, where $E[\mathbf{f}(\mathbf{x})]_{\mathrm{ex}}$ is the known expected value of $\mathbf{f}(\mathbf{x})$. If the assumptions underlying the linear model hold, then $\boldsymbol{\Omega}_R = E[\mathbf{x}\epsilon\,\mathbf{h}(\mathbf{x})^T] = \mathbf{0}$ because ϵ is independent of $\mathbf{f}(\mathbf{x})$ and $E(\epsilon) = 0$. Thus, according to the results of Sect. 3.1, there will be no variance reduction if the external moment function is a function of the explanatory variables only. In the example described in Sect. 2, there will be no efficiency gain if the 15.7%-prevalence of depression is used as external information to estimate the linear regression model.

On the other hand, if the external moment function is a function of ϵ, then generally, $E[\mathbf{x}\epsilon\,\mathbf{h}(\mathbf{x}, y)^T] \neq \mathbf{0}$. In the example, assume that the correlation between fluid intelligence and math skills reported in Peng et al. (2019) is taken as external information, in which case $\mathbf{h}(\mathbf{x}, y) = h(x_2, y) = [y - E(y)][x_2 - E(x_2)]/(\sigma_{x_2}\sigma_y) - \rho(x_2, y)_{\mathrm{ex}}$, where $\rho(x_2, y)_{\mathrm{ex}} = 0.41$. Then $E[\mathbf{x}\epsilon\,h(x_2, y)] = [\sigma^2/(\sigma_{x_2}\sigma_y)]\boldsymbol{\sigma}_{x_\cdot,x_2}$ will not in general be zero, and hence, there will, in general, be efficiency gains with respect to $\hat{\boldsymbol{\beta}}_{\mathrm{ex}}$. For more examples, see Table 1 and for the derivation of the results, see the supplementary materials online. It should be noted that if the distribution of the errors is not symmetrical, then $E(\mathbf{x})E(\epsilon^3)$ has to be added to the entries in column $\boldsymbol{\Omega}_R^T$ of Table 1 for the cases $E(y^2)$ and σ^2_y, see the supplementary materials online for further details.

Table 2 presents, in the second column, the absolute variance reduction for the parameters if the external information given in the first column is used to estimate the regression model. The third column in Table 2 shows which entries of the parameter $\boldsymbol{\beta}$ can be estimated more precisely if the external information is used. The results of Table 2 are derived in the supplementary materials online. Note that $\boldsymbol{\Omega}_h$ is written as ω_h here, as it is single-valued. It holds that $\omega_h = E[h(\mathbf{x}, y)^2]$, where $h(\mathbf{x}, y)$ is of the form given for various moments in Table 1. However, this expression often includes the terms $E(\epsilon)$ and $E(\epsilon^3)$, which are already set to zero in $\boldsymbol{\Omega}_R^T$ (see supplementary materials online). In order to avoid invalid estimates, $E(\epsilon)$ and $E(\epsilon^3)$ should be set to zero in ω_h. For example, if the correlation between fluid intelligence and math skills reported in Peng et al. (2019) would be used in the regression from math skills on fluid intelligence and depression, then the variance of the estimator weighting the variable fluid intelligence would be reduced by:

$$\frac{\sigma^4}{n\omega_h\sigma^2_y\sigma^2_{x_2}} = \frac{\sigma^4}{n\mathrm{Var}\{[x_2 - E(x_2)][y - E(y)]\}}.$$

TABLE 1.
Forms of $\mathbf{\Omega}_R^T$ for various single moments.

Moments	$h(\mathbf{x}, y)$	$\mathbf{\Omega}_R^T$
$E(y)$	$y - E(y)_{\text{ex}}$	$\sigma^2 E(\mathbf{x})$
$E(x_j y)$	$x_j y - E(x_j y)_{\text{ex}}$	$\sigma^2 E(x_j \cdot \mathbf{x})$
$E(y^2)$	$y^2 - E(y^2)_{\text{ex}}$	$2\sigma^2 E(\mathbf{x}\mathbf{x}^T)\boldsymbol{\beta}_0$
σ_y^2	$[y - E(y)]^2 - (\sigma_y^2)_{\text{ex}}$	$2\sigma^2[E(\mathbf{x}\mathbf{x}^T)\boldsymbol{\beta}_0 - E(y)E(\mathbf{x})]$
$\sigma_{x_j,y}$	$[y - E(y)][x_j - E(x_j)] - (\sigma_{x_j,y})_{\text{ex}}$	$\sigma^2 \boldsymbol{\sigma}_{x.,x_j}$
$\rho_{x_j,y}$	$\frac{[y-E(y)][x_j-E(x_j)]}{\sigma_{x_j}\sigma_y} - (\rho_{x_j,y})_{\text{ex}}$	$\frac{\sigma^2}{\sigma_{x_j}\sigma_y}\boldsymbol{\sigma}_{x.,x_j}$
$\beta_{x_j,y}$	$\frac{[y-E(y)][x_j-E(x_j)]}{\sigma_{x_j}^2} - (\beta_{x_j,y})_{\text{ex}}$	$\frac{\sigma^2}{\sigma_{x_j}^2}\boldsymbol{\sigma}_{x.,x_j}$

The subscript ex indicates externally determined values. In the last line, $\beta_{x_j,y}$ represents the expected value of the estimator of the slope from a simple linear regression model, which is identical to the true value of the slope only if x_j is independent of the other explanatory variables.

TABLE 2.
Effects of various single moments in terms of variance reduction.

Moments	$\mathbf{D}$	Effect on
$E(y)$	$\frac{\sigma^4}{n\omega_h}\mathbf{e}_1\mathbf{e}_1^T$	Only β_1
$E(x_j y)$	$\frac{\sigma^4}{n\omega_h}\mathbf{e}_j\mathbf{e}_j^T$	Only β_j
$E(y^2)$	$\frac{4\sigma^4}{n\omega_h}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^T$	All $\beta_j \neq 0$
σ_y^2	$\frac{4\sigma^4}{n\omega_h}[\boldsymbol{\beta}_0 - E(y)\mathbf{e}_1][\boldsymbol{\beta}_0 - E(y)\mathbf{e}_1]^T$	All $\beta_j \neq 0$ and β_1
$\sigma_{x_j,y}$	$\frac{\sigma^4}{n\omega_h}\tilde{\mathbf{e}}_j\tilde{\mathbf{e}}_j^T$	β_j and β_1
$\rho_{x_j,y}$	$\frac{\sigma^4}{n\omega_h\sigma_y^2\sigma_{x_j}^2}\tilde{\mathbf{e}}_j\tilde{\mathbf{e}}_j^T$	β_j and β_1
$\beta_{x_j,y}$	$\frac{\sigma^4}{n\omega_h\sigma_{x_j}^4}\tilde{\mathbf{e}}_j\tilde{\mathbf{e}}_j^T$	β_j and β_1

The expression $\mathbf{e}_j$ denotes the $(p \times 1)$-vector with 1 at the j-th position and zeros elsewhere. Further we set $\tilde{\mathbf{e}}_j := -E(x_j) \cdot \mathbf{e}_1 + \mathbf{e}_j$. In the last line, $\beta_{x_j,y}$ represents the expected value of the estimator of the slope from a simple linear regression model, which is identical to the true value of the slope only if x_j is independent of the other explanatory variables.

This means that there will be a variance reduction in all practically relevant cases, where $\sigma^2 \neq 0$ and $\text{Var}\{[x_2 - E(x_2)][y - E(y)]\} < \infty$ hold. For a comparison of the effects of the different external moments, the corresponding relative variance reductions may be of interest. These are obtained by dividing the j-th diagonal element of the absolute reductions in Table 2 by $\frac{1}{n}\sigma^2 E(\mathbf{x}\mathbf{x}^T)^{-1}_{(j,j)}$, where $E(\mathbf{x}\mathbf{x}^T)^{-1}_{(j,j)}$ denotes the element of the inverse of $E(\mathbf{x}\mathbf{x}^T)$ in the j-th row and the j-th column. For the resulting expressions it is clear that n factors out, as $\mathbf{D}$ also includes $\frac{1}{n}$ as the only factor depending on n, while the rest are fixed values. Hence, the relative efficiency gains do not vanish with increasing n, but are constant. In our example, the known correlation $\rho_{x_2,y} = .41$ exerts an expected relative variance reduction of

$$\frac{\sigma^2}{E(\mathbf{x}\mathbf{x}^T)^{-1}_{(2,2)}\text{Var}\{[x_2 - E(x_2)][y - E(y)]\}},$$

which is independent of n and does not vanish for large σ^2. Including more than one external moment is straightforward. In that case $\boldsymbol{\Omega}_h$ includes not only variances but also covariances of the external moments which may lead to additional variance reduction. To illustrate this effect, consider the example from Sect. 2 using the external moments $\rho(x_2, y)_{\text{ex}} = 0.41$ and $E(x_2)_{\text{ex}} = 100$. For the sake of simplicity and without loss of generality we assume x_2 and y to be centralized. In this example, the external moments $\rho_{x_2,y}$ and $E(x_2)$ are included in the externally informed multiple linear model, leading to $\boldsymbol{\Omega}_R^T = \left(\mathbf{0}\ \frac{\sigma^2}{\sigma_{x_2}\sigma_y}\boldsymbol{\sigma}_{x_\cdot,x_j}\right)$ according to Table 1, and

$$\boldsymbol{\Omega}_h = \begin{pmatrix} \text{Var}(x_2) & \frac{\text{Cov}(x_2^2,y)}{\sigma_{x_2}\sigma_y} \\ \frac{\text{Cov}(x_2^2,y)}{\sigma_{x_2}\sigma_y} & \frac{\text{Var}(x_2 y)}{\sigma_{x_2}^2\sigma_y^2} \end{pmatrix}$$

via definition, wherein $\text{Var}(x_2 y)$ is the scalar variance of x_2 times y. Using the notation of Table 2, the explicit inversion formula for (2×2)-matrices implies

$$\begin{aligned} \mathbf{D} &= \frac{1}{n}\left[E\left(\mathbf{x}\mathbf{x}^T\right)\right]^{-1}\boldsymbol{\Omega}_R^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Omega}_R\left[E\left(\mathbf{x}\mathbf{x}^T\right)\right]^{-1} \\ &= \frac{1}{n}\left[E\left(\mathbf{x}\mathbf{x}^T\right)\right]^{-1}\frac{\sigma^2}{\sigma_{x_2}\sigma_y}\boldsymbol{\sigma}_{x_\cdot,x_j}(\boldsymbol{\Omega}_h^{-1})_{(2,2)}\boldsymbol{\sigma}_{x_\cdot,x_j}^T\frac{\sigma^2}{\sigma_{x_2}\sigma_y}\left[E\left(\mathbf{x}\mathbf{x}^T\right)\right]^{-1} \\ &= \frac{\sigma^4(\boldsymbol{\Omega}_h)_{(1,1)}}{n\det(\boldsymbol{\Omega}_h)\sigma_{x_2}^2\sigma_y^2}\tilde{\mathbf{e}}_2\tilde{\mathbf{e}}_2^T = \frac{\sigma^4}{n\left[\text{Var}(x_2 y) - \frac{\text{Cov}(x_2^2,y)^2}{\sigma_{x_2}^2}\right]}\tilde{\mathbf{e}}_2\tilde{\mathbf{e}}_2^T, \end{aligned}$$

where $\det(\mathbf{A})$ denotes the determinant of matrix $\mathbf{A}$. Assuming both variances to be finite and positive and invoking the Cauchy–Schwartz inequality, the fraction $\text{Cov}(x_2^2, y)^2/\sigma_{x_2}^2$ will not exceed $\text{Var}(x_2 y)$ and hence $\mathbf{D}$ will be nonnegative. Further, if x_2^2 and y have a covariance different from 0, the variance will decrease even further, compared to the reduction due to $\rho_{x_2,y}$ alone. Hence, β_1 and β_2 can in general be estimated even more efficiently, if $E(x_2)$ is used in addition.

3.3. Additional Remarks

Using many moments, however, increases the risk of a near-singular $\boldsymbol{\Omega}$ matrix, especially if the moments are strongly mutually (linear) dependent. Calculation of the GMM-estimator with additional external moment functions often includes unknown population moments, like $E(\mathbf{x})$ or σ_y^2 (see Table 1), which may be replaced by the corresponding sample moments. However, $\boldsymbol{\Omega}_R$ and $\boldsymbol{\Omega}_h$ may in addition be functions of unknown σ^2 or $\boldsymbol{\beta}_0$, as can be seen in Table 1. Hence, the externally informed GMM-estimator is calculated iterating over the following steps until convergence: First estimate the model using the ordinary least squares approach without external moments to get $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$ and then estimate $\hat{\boldsymbol{\beta}}_{\text{ex}}$ based on the estimates from the former step.

Statistical inference with a GMM-estimator can be based on the Wald test, which simplifies to a t-test if single regression coefficients are tested and its approximative normality can be used to construct confidence intervals (Cameron & Trivedi, 2005). However, in small samples or when dealing with complex models, it is sometimes better to use a bootstrap method (Cameron & Trivedi, 2005; Spiess et al., 2019, p. 177).

As this approach combines data from different sources, one should take into account the issues arising in meta-analyses in general. The *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins et al., 2019) and the PRISMA statement (Page et al., 2021) should be considered

to select proper sources of external information, which are as up-to-date and as close as possible to the same population, method and design of the study one wishes to use the externally informed model in. This is important because a core regularity condition of the GMM is that the expected values of the moment functions are zero, which can be violated, if the external moment and the data were taken from different populations. As a possible approach to deal with this compatibility issue, the GMM framework incorporates the Sargan–Hansen test to test if the overidentification due to the additional moment conditions causes a $\hat{Q}_n(\hat{\boldsymbol{\theta}}_{\text{ex}})$ significantly larger than 0 (Hansen, 1982; Sargan, 1958). Another option to test for incompatibility especially in linear regression models is the Durbin–Wu–Hausman test (Hausman, 1978), as it compares two estimators of the same parameter. We will take a different approach here, as we will instead relax the assumption of correct external point-values to intervals containing the true value.

4. Robustness due to Interval Probability

External information is only an estimate itself and thus prone to uncertainty. A classical approach to analyze and prevent the issues of misspecification and thus misleading inferences is to use robust models (Huber, 1981). Hence, it is important to use techniques to robustify the estimation of the externally informed model. In this paper, we will adopt an approach based on the theory of imprecise probabilities due to Weichselberger (2001), that is capable of dealing with probabilistic and non-probabilistic uncertainty, not depending on a fully specified stochastic model. The advantage is that instead of distributional assumptions we only need bounds for the true external values. It would be possible to model the uncertainty in the external information within a probabilistic, e.g., a Bayesian, framework. However, this framework would replace uncertainty in the external information by assuming an additional parametric model of its estimation process in form of precise prior distributions. Moreover, it is not straightforward to represent only certain distributional aspects (moments) within a Baysian approach, e.g., the external information $100 = E(y) = E(\mathbf{x})^T\boldsymbol{\beta}_0$ presented in Sect. 2.

4.1. Externally Informed Models Based on Interval Information

Assume that I_{ex} is an interval containing the true value of an unknown external moment. Hence every value in the interval could be the true one. To illustrate a possible way to construct an I_{ex}, we use our earlier example. In our application example, we have a 95% confidence interval of [0.39, 0.44] for the correlation between fluid intelligence and mathematical skills (Peng et al., 2019). This is, of course, an interval that includes the true value only with a positive probability, but not with certainty. However, combining this confidence interval with the results of other studies on this or a similar correlation, and thus possibly widening the interval, the resulting interval serves as a subjective, rough approximation for I_{ex}. We illustrate the use of this technique in Sect. 6.
In this section, we discuss another way of constructing I_{ex}. Regarding the estimated depression prevalence of 0.157 in Steffen et al. (2020), we know that 87% of the population has been investigated. Thus, we can construct an interval by the technique proposed, e.g., in Manski (1993, 2003), Manski and Pepper (2013), Cassidy and Manski (2019). The two extreme cases that could occur are that no person of the 13% unobserved individuals has a depression and on the other extreme that all of these individuals have a depression. As 87% of 0.157 is 0.137, we get the interval [0.137, 0.267] for the prevalence. The advantage of such intervals is that they completely compensate for the missing values without any further assumptions. Having available an interval for the external information, one can adopt a technique denoted as cautious data completion proposed by Augustin et al. (2014, p. 182) to determine based on I_{ex} the sets of possible values for the estimator itself and its variance estimator. In our setting, this amounts to evaluating the

estimator for the externally informed linear model and its variance estimator from Theorem 1 traversing I_{ex}. This leads to a set $\mathcal{B}_{ex}$ of possible parameter estimates and a set $\mathcal{V}_{ex}$ of possible variance estimates. These sets of estimates are compact and connected in the strict mathematical sense, since both estimators are continuous functions on the external interval.

4.1.1. F-Probability Interval-based inferences can be justified by adopting the concept of F-probabilities (Augustin, 2002; Weichselberger, 2000).

Definition 3. (Augustin 2002) Let Ω be a set and $\mathcal{A}$ be a σ-algebra on Ω. Further, let $\mathcal{K}(\Omega, \mathcal{A})$ be the set of all probability measures on $(\Omega, \mathcal{A})$. Then, a set-valued function $F(\cdot)$ on $\mathcal{A}$ is called *F-probability* with structure $\mathcal{M}$, if

1. there are functions $L(\cdot), U(\cdot) : \mathcal{A} \to [0, 1]$ such that for every event $A \in \mathcal{A}$ it holds that $L(A) \leq U(A)$ and $F(\cdot)$ has the form

$$\begin{aligned} F(\cdot): \quad & \mathcal{A} \to \{[a, b] \mid a, b \in [0, 1] \text{ and } a \leq b\} \\ & A \mapsto F(A) := [L(A), U(A)] \text{ for every event } A \in \mathcal{A}, \end{aligned}$$

2. the set $\mathcal{M} := \{P(\cdot) \in \mathcal{K}(\Omega, \mathcal{A}) \mid L(A) \leq P(A) \leq U(A), \text{ for all } A \in \mathcal{A}\}$ is not empty,
3. for all events $A \in \mathcal{A}$ it holds that $\inf_{P(\cdot)\in\mathcal{M}} P(A) = L(A)$ and $\sup_{P(\cdot)\in\mathcal{M}} P(A) = U(A)$.

For most applications, it is sufficient to restrict to the case $\Omega = \mathbb{R}^d$ and let $\mathcal{A}$ be the corresponding Borel σ-algebra. F-probabilities are best understood as a representation of a "continuous" set of probability measures. For example, consider all normal distributions with a variance of 1 and a mean between -0.5 and 0.5. If we consider all these distributions as possible true distributions for a random variable X and evaluate an event in terms of its probability, we obtain a set of possible probability values. Consider the event $A = \{X \leq 0\}$, its possible probability ranges from 0.3085 (for mean 0.5) to 0.6915 (for mean -0.5) and thus $P(A) \in F(A) := [0.3085, 0.6915]$. If this procedure is performed for all $A \in \mathcal{A}$, the resulting $F(\cdot)$ is an F-probability. In general, given any nonempty set $\mathcal{P}$ of probability measures, one can construct the narrowest F-probability containing $\mathcal{P}$ by defining $F(A) := [\inf_{P\in\mathcal{P}} P(A), \sup_{P\in\mathcal{P}} P(A)]$ for each event $A \in \mathcal{A}$, cf. Remark 2.3. in Augustin (2002). If the intervals $F(A)$ consist of one element for all A, the F-probability simply corresponds to a single probability measure. Thus, it is a natural generalization of the conventional notion of probability, using simultaneously a range of probability measures between a lower bound and an upper bound. An important property of F-probabilities for ensuring robustness is that their structure $\mathcal{M}$ (all the probability measures covered by $F(\cdot)$, in the sense of condition 2 in Definition 3) is generally larger than the set $\mathcal{P}$ (called pre-structure) of probability measures used to construct them, since the structure is closed under convex combinations (Augustin, 2002). For two probability measures P and Q, this follows by the basic inequality that for all $0 \leq \epsilon \leq 1$ and $A \in \mathcal{A}$ it holds that

$$\min(P(A), Q(A)) \leq \epsilon P(A) + (1 - \epsilon)Q(A) \leq \max(P(A), Q(A)).$$

For example, convex combinations of normal distributions are not themselves normally distributed and include skewed and bimodal distributions. This illustrates that robustness with respect to distributional assumptions increases compared to using normal distributions alone. Unlike other concepts that reflect uncertainty about probability measures, such as triangular numbers (fuzzy numbers), there is no preference for one distribution over another caused by weighting functions or

possibility distributions. This agnosticism regarding the true distribution also covers deterministic ambiguity to some extent. For instance, in our example, a deterministic alteration of μ over time, where $\mu(t) \in [-0.5, 0.5]$ for all t, like $\mu(t) = 0.5\sin(t)$, would still be covered by the F-probability at any time t because the F-probability covers the range of $\mu(t)$. In applied research, the exact form of deterministic variation of μ is typically unknown, but if its bounds are known to lie within an interval, the F-probability based on this interval would account for it. Of course, these advantages come at the cost of greater conservatism than using a single probability distribution.

In our framework, the assumption of knowing the true moment value can be relaxed to assuming that an interval is known containing the unknown true moment value. As the GMM-estimator is asymptotically normally distributed for the true value of the external moment, we asymptotically get a pre-structure consisting of all normal distributions for estimator $\hat{\boldsymbol{\beta}}_{\text{ex}}$ with expected value inside $\mathcal{B}_{\text{ex}}$ and with variance inside $\mathcal{V}_{\text{ex}}$. This pre-structure is guaranteed to contain the normal distribution based on GMM asymptotics, since the true external moment value is assumed to be in I_{ex}. Therefore, for each event, the probability assigned to an event by this true normal distribution will lie between the lower and upper bounds assigned to that event by $F(\cdot)$, possibly leading to more conservative but valid statistical inference. Based on this pre-structure, we get an F-probability. Statistical inference based on F-probabilities is done by treating the probability intervals as a whole, e.g., by interval arithmetic. We demonstrate this principle by constructing an equivalent to confidence intervals in the context of F-probabilities in the next section.

4.1.2. Confidence Intervals for the Externally Informed Model Under F-Probabilities The construction of confidence intervals (point-CIs) is in general not possible in the framework of F-probabilities, because instead of a single probability value lower and upper bounds are assigned to an event. One possibility, however, is to use the union of all possible point-CIs traversing I_{ex}. The idea to calculate unions of intervals already has been investigated for Bayesian highest density intervals in an imprecise probability setting by Walter and Augustin (2009). Let $\hat{\boldsymbol{\theta}}_{e,j}$ be the j-th entry of the externally informed GMM-estimator $\hat{\boldsymbol{\theta}}_{\text{ex}}$ using external value e, we define the $(1-\alpha)\cdot 100\%$ confidence union for $\boldsymbol{\theta}_j$ to be

$$\bigcup \text{CI}_{1-\alpha} := \left[\inf_{e\in I_{\text{ex}}} [\hat{\boldsymbol{\theta}}_{e,j} - t_{1-\frac{\alpha}{2},n-p}\sqrt{\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_{e,j})}],\ \sup_{e\in I_{\text{ex}}} [\hat{\boldsymbol{\theta}}_{e,j} + t_{1-\frac{\alpha}{2},n-p}\sqrt{\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_{e,j})}] \right]$$

Because the true external moment value is in I_{ex}, the borders of the point-CI constructed via the true moment value lie between the infimum and the supremum of the lower and upper borders, respectively, of all point-CIs on I_{ex}. Therefore, $\bigcup \text{CI}_{1-\alpha}$ covers the point-CI constructed via the true moment value. The asymptotic normal distribution of $\hat{\boldsymbol{\beta}}_{\text{ex}}$ at the true value of the external moment implied by the asymptotic properties of GMM-estimators described in Sect. 2 ensures that the confidence union covers the true parameter asymptotically with probability at least $1-\alpha$. An approximation of the confidence union can be calculated using grid search traversing I_{ex}. If the point-CIs used to construct $\bigcup \text{CI}_{1-\alpha}$ differ, then the resulting interval is wider than every of these point-CIs. This demonstrates that the positive effect of the variance reduction (a shorter CI) can be reversed by the length of I_{ex}. The reason is that a broader I_{ex} increases the set over which infimum and supremum are taken, possibly expanding $\bigcup \text{CI}_{1-\alpha}$. However, we will show in a simulation study in Sect. 5 that in some cases it is possible to get a $\bigcup \text{CI}_{1-\alpha}$ shorter than the $(1-\alpha)$ confidence interval based on the OLS multiple linear regression. Hence, the variance reduction can compensate the broadening of $\bigcup \text{CI}_{1-\alpha}$ introduced by I_{ex}. Finally, using $\bigcup \text{CI}_{1-\alpha}$ strengthens the robustness through the F-probability, on which $\bigcup \text{CI}_{1-\alpha}$ is based, as it also includes, e.g., bimodal and skewed distributions.

5. A Simulation Study

5.1. Settings

To test the externally informed GMM approach for multiple linear models in small samples, we conducted two simulation studies. The first setting illustrates possible variance reduction if correctly specified external moments are used and shows that the usage of small external moment intervals can lead to confidence unions that may even be shorter than the OLS confidence interval. In the second setting, we focus on misspecified external information and non-normal errors. In this case, it is interesting to see, if inferences are still valid and whether the effects of the variance reduction illustrated in the first setting still occur. The simulation script was written and executed in R version 4.2.1 (R Core Team, 2022), the script can be found in the supplementary materials online. The function `interval_gmm()` implements the calculation of intervals of estimators and of their standard deviation, as well as confidence interval unions. In both settings, we used an intercept ($x_1 = 1$), a normally distributed variable $x_2 \sim N(2, 4)$ and a binary variable distributed according to Bernoulli distribution $x_3 \sim \text{Bernoulli}(0.4)$ as explanatory variables. The response variable was generated according to $y = x_1 + 0.5x_2 + 2x_3 + \epsilon$, where $\epsilon \sim N(0, 9)$ in the first setting. In the second setting, the errors were generated by affine transformation of a χ^2_1-distributed random sample, so that its mean is 0 and its variance is 9. The settings were selected, so that all required moments can easily be calculated, which is done before the simulations. The ratio of explained variance to total variance was $1 - 9/\text{Var}(y) = 1 - 9/10.96 = 0.178$ a value which is similar to often reported values in psychological research. This amounts to a relatively high error variance, a factor for possibly large variance reduction for some external moments (see Sect. 3).

Different moments have different scales, so a similar interval width of I_{ex} does not imply similar "sharpness" of the external information across scales. To create intervals for the external information, that are comparable across the different scales of the external moments, we have used external intervals where the ratio of half their width to their center is the same for all external moments in each setting. It should be noted that this technique is different from the design techniques discussed in Sect. 4.1. The reason for this difference is that the simulation study aims to compare the different moments in terms of their effectiveness and statistical validity in a context where the I_{ex} are comparable in magnitude and contain the true value. To motivate this, one could compare the given ratio to the coefficient of variation. For the standard IQ-scale, the coefficient of variation is $15/100 = 0.15$. For the first setting, we arbitrarily chose a ratio of 0.1 to represent somewhat more precise external information than one standard deviation in the IQ-scale around the center. For the second setting, we have chosen a ratio of 0.3 to represent a radius of two standard deviations in the IQ-scale and thus an approximate confidence interval width that takes the IQ-scale as a basis. In the first setting, we created intervals that were symmetrical around the true external value. Hence, if the true external value was e, then the interval was $I_{\text{ex}} = [0.9e, 1.1e]$. In the second setting, we first multiplied all true external moment values by 1.3. Since none of these true external values were equal to zero, this resulted in misspecified point values. These misspecified values were used as external point values during the simulation to test the sensitivity of the externally informed model based on point information. The constant 1.3 was again chosen arbitrarily and leads to a relative bias of 30%. Then, as in the first setting we generated a symmetric interval around the misspecified value. If e again denotes the true external value, $0.7 \cdot 1.3e = 0.91e$ was the lower limit and $1.3 \cdot 1.3e = 1.69e$ the upper limit of I_{ex}, i.e. $I_{\text{ex}} = [0.91e, 1.69e]$, which contains the true value e. As for sensitivity, tests with center width ratio and misspecification values similar to 0.1, 0.3 and 1.3 gave similar results.

Sample sizes n chosen are 15, 30, 50, 100. The moments used are those listed in Table 2 for both, x_2 and x_3. Given the results in Sect. 3, the expected relative variance reductions were calculated to check if these settings are capable of providing enough variance reduction. For every

moment condition in each setting we run 500 simulations. Only single moment conditions were used.

In a first step, all explanatory variables were generated and y was calculated as described above. In the second step, $\hat{\boldsymbol{\beta}}_{\text{ex}}$ and $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{ex}})$ were calculated according to the following two-step GMM algorithm:

1. Calculate $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ via the classical OLS method
2. Determine $\hat{\boldsymbol{\Omega}}_R$, $\hat{\omega}_h$ and $\hat{\boldsymbol{\beta}}_{\text{ex}}$ based on $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$
3. Recalculate $\hat{\sigma}^2$, $\hat{\boldsymbol{\Omega}}_R$ and $\hat{\omega}_h$ based on $\hat{\boldsymbol{\beta}}_{\text{ex}}$
4. Update $\hat{\boldsymbol{\beta}}_{\text{ex}}$ and calculate $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{ex}})$

Then, 95% confidence intervals were calculated based on $\hat{\boldsymbol{\beta}}_{\text{ex}}$ and its estimated variance, using a t-distribution with $n - 3$ degrees of freedom. Let $\hat{\beta}_{\text{ex}}$ be one element of $\hat{\boldsymbol{\beta}}_{\text{ex}}$, then it's 95% confidence interval is

$$\text{CI}_{0.95} = \left[\hat{\beta}_{\text{ex}} - t_{n-3,0.975}\sqrt{\widehat{\text{Var}}(\hat{\beta}_{\text{ex}})},\, \hat{\beta}_{\text{ex}} + t_{n-3,0.975}\sqrt{\widehat{\text{Var}}(\hat{\beta}_{\text{ex}})}\right].$$

To calculate $\bigcup \text{CI}_{0.95}$ a grid search algorithm was adopted. First we determined 101 equidistant points in the given I_{ex} (including the bounds of the interval). The number 101 was chosen after some preliminary tests of the algorithm as a compromise between precision and computing time. Then we traversed these grid points calculating $\hat{\beta}_{\text{ex}}$ and $\widehat{\text{Var}}(\hat{\beta}_{\text{ex}})$ using the two step procedure from above at each point. Comparing the bounds of the CIs sequentially, the minimal lower and maximal upper CI bounds on the grid points were determined and served as approximation for the bounds of $\bigcup \text{CI}_{0.95}$.

5.2. *Results*

As criteria to evaluate the statistical inferences, we calculated the mean $\bar{\hat{\boldsymbol{\beta}}}_{\text{ex}}$ of the estimates $\hat{\boldsymbol{\beta}}_{\text{ex}}$ and their variances $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{ex}})$ over 500 simulations. The latter will be compared to the corresponding means of estimated variances, $\overline{\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{ex}})}$. To evaluate possible variance reduction for β_j, the mean ratio of variance reduction to OLS-variance, $\hat{\Delta}_j := \overline{[\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{OLS}) - \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{ex}})]_{(j,j)}/[\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{OLS})]_{(j,j)}}$, will be considered. In addition, the actual coverage is calculated over simulations. For given $\alpha = 0.05$ and 500 simulations, the actual coverage should be between 0.93 and 0.97 for the point-valued moments (Spiess, 1998) and equal to or greater than 0.93 for the external moment intervals, as the confidence union is used to calculate the coverage in this case. Finally, $|\text{CI}| := \overline{\overline{\text{CI}}_{0.95} - \underline{\text{CI}}_{0.95}}$ and $|\bigcup \text{CI}| := \overline{\overline{\bigcup \text{CI}}_{0.95} - \underline{\bigcup \text{CI}}_{0.95}}$ were computed. They can be compared to the OLS-CI-length to evaluate the possible precision gains or losses.

5.2.1. Results for the Correctly Specified Setting The detailed results for sample size $n = 15$ are presented in Table 3, while the results for the other sample sizes are given in Tables 7 to 9 in the supplementary materials. Consistent with the theory in Sect. 3, the use of the moment $E(x_2)$ had no effect on the variances, neither for the correctly specified nor for the misspecified setting, and estimation results were equal to OLS estimation results. The corresponding results are presented for comparison. For all moments except $E(y^2)$ and σ_y^2 both the coverages for the point valued moments as well as the coverages for the external intervals exceeded 0.93. The coverages for σ_y^2 were in the valid range only for $n = 100$, while for $E(y^2)$ they were in the valid range already for $n = 50$. The undercoverage for sample sizes below $n = 100$ can be explained by the skewness

TABLE 3.
Results of the simulations with correctly specified external moments for sample size $n = 15$.

Moments	β_j	$\bar{\hat{\beta}}_{ex}$	$\text{Var}(\hat{\beta}_{ex})$	$\overline{\widehat{\text{Var}}(\hat{\beta}_{ex})}$	$\hat{\Delta}_j$	Cov	Cov_I	|CI|	$\lvert\bigcup \text{CI}\rvert$
$E(x_2)$	β_1	1.045	1.929	1.968	0	0.944	0.944	5.866	5.866
= OLS	β_2	0.499	0.205	0.211	0	0.950	0.950	1.909	1.909
	β_3	1.955	3.422	2.976	0	0.940	0.940	7.292	7.292
$E(y)$	β_1	1.048	1.453	1.593	0.217	0.942	0.964	5.215	5.639
$E(x_2 y)$	β_2	0.560	0.174	0.177	0.169	0.950	0.956	1.740	1.832
$\sigma_{x_2,y}$	β_1	0.802	1.478	1.456	0.239	0.954	0.954	5.054	5.222
	β_2	0.625	0.096	0.080	0.616	0.960	0.978	1.153	1.237
$\rho_{x_2,y}$	β_1	0.928	1.405	1.418	0.258	0.944	0.948	4.984	5.141
	β_2	0.562	0.069	0.074	0.639	0.976	0.984	1.110	1.188
$\beta_{x_2,y}$	β_1	0.967	1.263	1.414	0.255	0.960	0.964	4.982	5.131
	β_2	0.543	0.054	0.072	0.633	0.986	0.986	1.107	1.180
$E(x_3 y)$	β_3	2.150	2.745	2.380	0.192	0.936	0.946	6.525	6.837
$\sigma_{x_3,y}$	β_1	0.959	1.552	1.704	0.141	0.964	0.966	5.433	5.561
	β_3	2.202	0.783	0.949	0.689	0.976	0.976	3.952	4.273
$\rho_{x_3,y}$	β_1	0.977	1.659	1.694	0.148	0.960	0.960	5.415	5.541
	β_3	2.136	0.888	0.919	0.699	0.964	0.970	3.891	4.208
$\beta_{x_3,y}$	β_1	0.942	1.545	1.700	0.143	0.964	0.970	5.426	5.559
	β_3	2.182	0.692	0.908	0.698	0.976	0.980	3.888	4.213
$E(y^2)$	β_1	1.072	1.965	1.935	0.018	0.914	0.936	5.807	5.950
	β_2	0.507	0.207	0.204	0.031	0.942	0.946	1.875	1.929
	β_3	1.995	3.414	2.871	0.030	0.916	0.930	7.164	7.367
σ_y^2	β_1	0.905	1.951	1.708	0.142	0.920	0.928	5.434	5.667
	β_2	0.540	0.228	0.184	0.148	0.896	0.910	1.766	1.839
	β_3	2.092	3.646	2.584	0.151	0.858	0.878	6.725	7.005

The expressions $\bar{\hat{\beta}}_{ex}$, $\text{Var}(\hat{\beta}_{ex})$, $\overline{\widehat{\text{Var}}(\hat{\beta}_{ex})}$, $\hat{\Delta}_j$, |CI| and $\lvert\bigcup \text{CI}\rvert$ are defined in the beginning of Sect. 5.2. The results for the moment $E(x_2)$ are equivalent to the OLS results. Cov is the coverage for the external point value and Cov_I symbolizes the coverage for the confidence interval union based on the external interval. Only the affected coefficients are reported per moment. The true values are $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_3 = 2$.

of the distributions of their sample moment functions in small samples caused by the quadratic terms y^2, leading to higher sample size required for the asymptotic results to be applicable. Using confidence unions only reduced these required sample sizes to $n = 50$ and $n = 30$, respectively, showing that high skewness is also problematic for $\bigcup$ CI-based coverage in small samples.

For $\beta_{x_j,y}$ with $j = 2, 3$, the coverage for β_j was in many cases above 0.97 (up to 0.994) for all n. This was also the case, though not as pronounced, when the external information about the covariance between x_j and y was used. The reason for this is that the variances were mostly overestimated in these cases, as can be seen in Tables 3 and 4 as well as in Tables 7 to 12 in the supplementary materials by the fact that $\overline{\widehat{\text{Var}}(\hat{\beta}_{ex})}$ was larger than $\text{Var}(\hat{\beta}_{ex})$ for the respective β_j. Although variances are overestimated, the true and estimated variances nevertheless tend to be smaller than the variance of the OLS-estimators. Thus, inferences still tend to be more precise, suggesting a possible relationship with superefficiency (Bahadur, 1964).

As shown in Sect. 3, the relative variance reduction for each estimator of β_j, reported in column $\hat{\Delta}_j$ of Table 3 as well as Tables 7 to 9 in the supplementary materials, did not change significantly under the various conditions over the different sample sizes realized. The smallest relative variance reduction per β_j was attained by using the external information $E(y^2)$, ranging from 0.018 to 0.059, followed by σ_y^2 with a maximal relative variance reduction of 0.180. The

largest relative variance reduction was attained by using the covariance, the correlation and $\beta_{x_j,y}$ regarding β_j, ranging from 0.633 to 0.734 for $j = 2$ as well as from 0.698 to 0.857 for $j = 3$. For all other moments the values varied between 0.169 and 0.294, see Table 3 and Tables 7 to 9 in the supplementary materials.

These variance reductions translated for all moments directly into a reduction of the length of the confidence interval for the external point value. For the external interval, the length of the union of the confidence intervals is always greater than the one derived from a single external point. These differences increase with larger samples, as the variance estimator decreases with increasing sample sizes, while it can be seen from the formulas in Theorem 1 that the interval for $\hat{\beta}_{\text{ex}}$ is only affected by the difference between the estimators and the true values of $\mathbf{\Omega}_R$ and $\mathbf{\Omega}_h$, not directly by n. Finally, with regard to $|\bigcup \text{CI}|$ compared to $|\text{CI}|$ the results imply that at the sample sizes 15 and 30, using any moment except $E(y^2)$ resulted in a shorter confidence interval union than the OLS confidence interval. For $n = 50$, this was the case for all moments except $E(y^2)$, $E(y)$ and σ_y^2. Finally, for $n = 100$, only the moments $\sigma_{x_2,y}$, $\rho_{x_2,y}$, $\beta_{x_2,y}$, $\sigma_{x_3,y}$, $\rho_{x_3,y}$ and $\beta_{x_3,y}$ resulted in shorter confidence unions than point-CIs. This can be explained by the constancy of I_{ex} while n increases. There is always an interval inside $\bigcup \text{CI}$ which does not vanish for large n, while $|\text{CI}|$ converges to 0.

5.2.2. Results for the Misspecified Setting The detailed results for sample size $n = 50$ are presented in Table 4, while the results for the other sample sizes are given in Tables 10 to 12 in the supplementary materials. The coverage rates using the point-valued moments illustrate the expected sensitivity of the models due to misspecification. Even at $n = 15$ more than half of the coverage rates are below 0.93, although in most cases they are still above 0.9. The severeness increases with increasing n: For $n = 30$ only five coverage rates are in the acceptable range of at least 0.93. As seen in Table 4 for $n = 50$ the coverage is as low as 0.586 in the worst case for β_3 if $\sigma_{x_3,y}$ is used. Finally, for $n = 100$ all coverage rates are invalid, see Table 12 in the supplementary materials. Except for the moments $E(y^2)$ and σ_y^2, this is corrected by the union of confidence intervals based on the external interval, since all coverage rates in these cases are above 0.93, except the one for β_1 using $\sigma_{x_2,y}$ while $n = 15$. Like in the correctly specified setting, there are considerably larger coverage rates for the moments $\beta_{x_j,y}$ and lower coverage rates for σ_y^2 or $E(y^2)$ even in the cases $n = 30$ and $n = 15$. The explanations for these over- and undercoverages are the same as for the correctly specified case in Sect. 5.2.1. However, only the use of covariance, correlation or β for x_j and y for $j = 2, 3$ resulted in narrower confidence unions as compared to OLS confidence intervals, not the use of other moments. Regarding β_j for $j = 2, 3$ this is the case for every n, regarding β_1 this is only the case for $n = 15$. We conclude that the use of external intervals for covariances, correlations or β not only corrects low coverage rates due to misspecified point values for external moments, but can also lead to narrower (unions of) confidence intervals.

6. Application

To illustrate the possible benefits of using external information in a linear model, we reanalyze a dataset of Pluck and Ruales-Chieruzzi (2021), who investigated the estimation of premorbid intelligence based on lexical reading tasks in Ecuador. We will focus on their Study 2. Since the purpose of this analysis is to illustrate the proposed use of external information, we will only shortly sketch the theoretical background of the study. For a more detailed description, see Pluck and Ruales-Chieruzzi (2021). The dataset was downloaded from PsychArchives (Pluck, 2020a).

To quantify the cognitive impairment of patients, it is necessary to have an accurate baseline estimate observed in the premorbid state (Pluck & Ruales-Chieruzzi, 2021). As psychometric

TABLE 4.
Results of the simulations with misspecified external moments for sample size $n = 50$.

Moments	β_j	$\bar{\hat{\beta}}_{ex}$	$\mathrm{Var}(\hat{\beta}_{ex})$	$\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$	Cov	Cov_I	|CI|	$\lvert\bigcup \mathrm{CI}\rvert$
$E(x_2)$	β_1	1.012	0.440	0.509	0.936	0.936	2.778	2.778
= OLS	β_2	0.488	0.041	0.049	0.960	0.960	0.865	0.865
	β_3	2.042	0.751	0.803	0.954	0.954	3.494	3.494
$E(y)$	β_1	1.636	0.412	0.376	0.782	0.992	2.397	4.032
$E(x_2 y)$	β_2	0.615	0.031	0.040	0.924	0.984	0.779	1.062
$\sigma_{x_2,y}$	β_1	0.733	0.307	0.380	0.912	0.954	2.409	2.922
	β_2	0.627	0.016	0.016	0.864	0.984	0.496	0.747
$\rho_{x_2,y}$	β_1	0.756	0.243	0.380	0.952	0.968	2.401	2.914
	β_2	0.617	0.019	0.016	0.880	0.996	0.494	0.744
$\beta_{x_2,y}$	β_1	0.768	0.271	0.376	0.926	0.960	2.397	2.901
	β_2	0.610	0.008	0.015	0.948	0.998	0.488	0.735
$E(x_3 y)$	β_3	2.505	0.486	0.622	0.934	0.982	3.095	4.165
$\sigma_{x_3,y}$	β_1	0.827	0.341	0.413	0.930	0.966	2.502	3.026
	β_3	2.549	0.116	0.139	0.586	1.000	1.445	2.751
$\rho_{x_3,y}$	β_1	0.830	0.291	0.413	0.956	0.978	2.498	3.025
	β_3	2.525	0.341	0.138	0.642	0.978	1.434	2.739
$\beta_{x_3,y}$	β_1	0.822	0.346	0.412	0.922	0.966	2.499	3.030
	β_3	2.537	0.092	0.135	0.618	1.000	1.428	2.740
$E(y^2)$	β_1	1.136	0.512	0.513	0.914	0.944	2.802	3.151
	β_2	0.558	0.053	0.049	0.894	0.948	0.863	1.011
	β_3	2.318	0.846	0.790	0.894	0.958	3.477	4.093
σ_y^2	β_1	0.512	0.894	0.433	0.698	0.828	2.543	3.361
	β_2	0.626	0.067	0.044	0.754	0.896	0.806	1.029
	β_3	2.597	1.054	0.697	0.750	0.896	3.234	4.167

The expressions $\bar{\hat{\beta}}_{ex}$, $\mathrm{Var}(\hat{\beta}_{ex})$, $\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$,|CI| and $\lvert\bigcup \mathrm{CI}\rvert$ are defined in the beginning of Sect. 5.2. The results for the moment $E(x_2)$ are equivalent to the OLS results. Cov is the coverage for the external point value and Cov_I symbolizes the coverage for the confidence interval union based on the external interval. Only the affected coefficients are reported per moment. The true values are $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_3 = 2$.

intelligence tests can be too long or cumbersome for elderly people with emerging cognitive impairments, it is important to have short, yet reliable tests for general intelligence. It is argued in Pluck and Ruales-Chieruzzi (2021) that vocabulary has a high positive correlation with general intelligence, hence using short lexical tests could be helpful to estimate general intelligence. Following Cattell's classical theory, general intelligence can be divided into fluid and crystallized intelligence (Cattell, 1963). In this context, the variance reduction property of the externally informed linear model could provide an asymptotically unbiased estimate with higher precision than the estimates in Pluck and Ruales-Chieruzzi (2021), because external information about the correlation of general, fluid or crystallized intelligence and lexical tests is available. Although the different factors of intelligence are not identical, combining external information about them leads to a broader and thus more reliable external interval than using information about general intelligence alone, as the correlation between lexical tasks and fluid or crystallized intelligence may be lower or higher than for general intelligence.

In their Study 2 Pluck and Ruales-Chieruzzi (2021) used a Spanish, validated seven-subtest version of the Wechsler Adult Intelligence Scale in the 4th edition (WAIS-IV) (Meyers et al., 2013) to measure general intelligence, as well as three lexical tests, the Word Accentuation Test (WAT) in Spanish (Del Ser et al., 1997), the Stem Completion Implicit Reading Test (SCIRT) (Pluck,

2018) and the Spanish Lexical Decision Task (SpanLex) (Pluck, 2020b). The sample consists of 106 premorbid participants without neurological illness. As one participant has not completed the WAT, this person was excluded from the analysis regarding the WAT score. Simple linear regression models with the WAIS-IV as dependent and the lexical tests as independent variable, respectively, were conducted to determine the percentage of explained variance and to test the predictability of general intelligence through every single test. Therefore, the sample was randomly divided into two halves; hence, the net sample size for the linear regression models was 53 as the other half was used to test the prediction based on the regression models. We compared the widths of the 95% confidence intervals for the parameters of these regression models to the widths of the 95% confidence unions resulting from externally informed versions of the linear models. Because the OLS estimation does not account for heteroscedastic errors, which are common in practice, the standard errors are often too small (White, 1980). To correct for heteroscedasticity, we have computed robust standard errors of type HC3 using the package *sandwich* (Zeileis, 2004; Zeileis et al., 2020). Since the dependent variable is the WAIS-IV, an intelligence test with calibration sample, we calculated $E(y) = 100$. In the simulation study using the external information about ρ was found to lead to high variance reduction. Hence, by reviewing the literature, we identified the upper bound for the correlation between general intelligence and lexical tasks to be .85. This value was reported as correlation between WAT and the vocabulary scale of the Wechsler Adult Intelligence scale in Burin et al. (2000). A lower bound for the correlation between general intelligence and lexical tasks was found using the meta-analysis of Peng et al. (2019) or the study of Pluck (2018). Pluck (2018) argued based on a couple of studies that the correlation of general intelligence and lexical skills is typically higher than .70. In the meta-analysis of Peng et al. (2019), the reported 95% confidence interval for the correlation of fluid intelligence and reading is [0.36, 0.39]. To compare the results, both sources were used separately, leading to the lower bounds 0.4 and 0.7, where 0.4 is very conservative as it is derived from a correlation including a different variable (fluid intelligence). Together this amounts to the intervals [0.4, 0.85] and [0.7, 0.85], which are adopted for each of the three lexical tests. The confidence unions were calculated in the same way as in the simulations using grid search, but with 10001 grip points instead of 101 and $\hat{\mathbf{\Omega}}_h = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\mathbf{z})\mathbf{h}(\mathbf{z})^T$. The details of the analysis can be found in the R script in the online supplements to this article. The results for the interval [0.7, 0.85] are shown in Table 5 and the results for the interval [0.4, 0.85] are in Table 13 in the supplementary materials. First, the results of Pluck and Ruales-Chieruzzi (2021) were recalculated, showing no differences from the results reported in their Study 2. In addition, the corresponding OLS confidence intervals for the parameters were calculated based on the HC3 estimator (see column five of Table 5). Then, estimator and standard error intervals, as well as the unions of confidence intervals, were calculated for the externally informed model. For both [0.4, 0.85] and [0.7, 0.85], the maxima of all standard error intervals were below the respective standard errors calculated for the OLS models of Pluck and Ruales-Chieruzzi (2021). This clearly shows the variance reduction property of the externally informed model and was most pronounced for the SpanLex. For [0.4, 0.85] all estimation intervals included the OLS estimates and all confidence unions were larger than the corresponding OLS confidence intervals, indicating that [0.4, 0.85] is very conservative. For [0.7, 0.85], the estimation interval $[\hat{\beta}_j, \overline{\hat{\beta}_j}]$ included the OLS estimator only for the slope and intercept of the regression on SCIRT and the one based on the WAT. In this case, however, all confidence unions overlapped with the OLS-based confidence intervals. Using [0.7, 0.85], for every lexical test, the widths of the confidence unions from the externally informed model were smaller than the confidence intervals from the simple linear regression models, for both slopes and intercepts, except for the intercept of WAT. Since the prediction interval is calculated based on the distribution of parameter estimators, this would lead to shorter prediction intervals for a participant's general intelligence based on the externally informed model. In addition, the confidence union approach

TABLE 5.
Results using $\rho_{x,y} \in [0.7, 0.85]$ and $E(y) = 100$.

j	Test	Pluck and Ruales-Chieruzzi			Externally informed estimates		
		$\hat{\beta}_j$	$s(\hat{\beta}_j)$	$CI_{0.95}$	$[\underline{\hat{\beta}_j}, \overline{\hat{\beta}_j}]$	$[\underline{s(\hat{\beta}_j)}, \overline{s(\hat{\beta}_j)}]$	$\bigcup CI_{0.95}$
1	SpanLex	54.61	8.864	[37.06, 72.15]	[37.41, 47.15]	[2.373, 2.663]	[32.06, 51.91]
	WAT	62.81	4.701	[53.51, 72.12]	[60.02, 62.89]	[3.587, 3.612]	[52.77, 70.10]
	SCIRT	60.81	4.395	[52.11, 69.51]	[59.01, 61.28]	[3.910, 3.920]	[51.14, 69.13]
2	SpanLex	1.821	0.332	[1.163, 2.480]	[2.068, 2.430]	[0.124, 0.132]	[1.818, 2.696]
	WAT	2.083	0.240	[1.607, 2.559]	[2.041, 2.186]	[0.190, 0.191]	[1.659, 2.568]
	SCIRT	3.292	0.358	[2.583, 4.001]	[3.213, 3.393]	[0.309, 0.310]	[2.592, 4.015]

The third and fourth columns contain the recomputed results of Pluck and Ruales-Chieruzzi (2021) in terms of the OLS regression coefficients $\hat{\beta}_j$, where $\hat{\beta}_1$ is the intercept and $\hat{\beta}_2$ is the slope and the robust standard errors $s(\hat{\beta}_j)$ of the coefficients. The (robust) 95% confidence intervals $CI_{0.95}$ for the parameters were computed in addition. The estimator interval $[\underline{\hat{\beta}_j}, \overline{\hat{\beta}_j}]$, the standard error interval $[\underline{s(\hat{\beta}_j)}, \overline{s(\hat{\beta}_j)}]$ and the 95% confidence interval union $\bigcup CI_{0.95}$ are shown as results of the estimation of the externally informed model.

is more robust than OLS confidence intervals with respect to deviations from the assumed normal distribution. Taken together, this amounts to possibly more precise yet robust parameter estimation and prediction, if the external information is correct.

7. Discussion

In this paper, we show that incorporating external moments into the GMM framework by using intervals instead of point values can lead to more robust analyses, while a possible variance reduction can prevent the confidence unions from being too wide.

The results of the simulation study for point values show that the variance reduction can be considerable, over 70% using external information about covariances, correlations or $\beta_{x_j,y}$. However if the external moments deviate from the true values, the inferences will be biased, getting worse with increasing sample size. Instead, the use of external intervals often leads to correct inferences. However, the F-probability couldn't completely correct the undercoverages caused by using the moments σ_y^2 as well as $E(y^2)$, though it slightly improved them. The reason for these undercoverages is the skewed distribution induced by y^2, indicating a limitation of the distributional robustness in the presence of large deviations from the normal distribution. As these two moments also showed low variance reduction when used, one should thoughtfully decide on basis of their relative variance reduction if one wants to use them in small samples. However, bootstrap methods, like the bias-corrected accelerated bootstrap (Efron & Tibshirani, 1993), could be used instead to try to correct the undercoverage.

For small sample sizes, the use of covariances, correlations, and $\beta_{x_j,y}$, $j = 2, 3$, leads to variance reduction despite the use of external intervals. However, this was mostly the case for certain entries β_j of $\boldsymbol{\beta}$ in this setting, not for all elements in $\boldsymbol{\beta}$. Interestingly, the use of covariances and $\beta_{x_j,y}$, $j = 2, 3$, still resulted in overcoverage caused by overestimation of the variance. This means that inferences based on these moments would be more conservative than necessary, yet they had the highest variance reduction of all the moments tested, providing an interesting link to the concept of superefficiency (Bahadur, 1964). Further research on the variance estimator is needed to potentially correct for its overestimation.

Taken together, the simulation study showed promising results regarding very small sample sizes as $n = 15$, and however, one should still be cautious as the estimators are only proved to be consistent, not unbiased. To be sure that the inference will be valid in the sample at hand, a simulation to test the adopted scenario, i.e., model to be estimated and data set, is advised. In Sect. 6, we showed the applicability of the theoretical results to real data, where for the variable SpanLex the width of the confidence unions was significantly smaller than the width of the corresponding point-CI, if an appropriately small external interval is used. This shows the usefulness of adopting an externally informed model for applied problems.

A possible limitation of GMM is the assumption of the covariance matrix of the external moments being positive definite, which excludes distributions for which the required covariance matrix does not exist, e.g., the Cauchy distribution. Nevertheless, in many psychological applications the variables have a constrained range of values, so that at least the existence of the covariance matrix can be assumed. In general, the applicability of the method is not overtly limited by its assumptions. Another limitation is that the true value of the external moment must be within the external interval. However, this identifiability assumption, or an analogous assumption, exists in other approaches, and it is much weaker than point identifiability. Thus, a more robust use of external information is possible, up to using the full range of possible values, which would definitely lead to a valid, more robust, but also very conservative inference. The construction of the external moment interval in Sect. 6 was based on a rough, subjective approximation. The question of how to construct the external intervals requires further research. In particular, further links to existing techniques for eliciting intervals and preventing overconfidence bias would be important.

An application of the theory to generalized linear models or multi-level models is of inherent interest for psychological research, especially as Corollary 1 sets the foundation for research on more complex models. At first glance, the results appear to be in conceptional "conflict" with multi-level-models, since these often assume the random effects to be normally distributed and in this case there is no bounded interval, that includes the true parameter. However, even in these models there are fixed (hyper-)parameters one could know bounds for, and hence, it would be interesting for future research to analyze the behavior of these models in the external GMM framework. With respect to the limitation of robustness found in the simulation study, it would be interesting to investigate how robust the estimators are as a function of the length of the external interval. Finally, research on (the properties of) significance tests based on the use of an external intervals would be of great interest.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Data availability The dataset of Pluck and Ruales-Chieruzzi (2021) analyzed during the current study is available in the PsychArchive repository, https://doi.org/10.23668/psycharchives.2897. All data generated by the simulations in this study are included in this article and its supplementary information files.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Augustin, T. (2002). Neyman–Pearson testing under interval probability by globally least favorable pairs: Reviewing Huber–Strassen theory and extending it to general interval probability [Imprecise probability models and their applications]. *Journal of Statistical Planning and Inference, 105*(1), 149–173. https://doi.org/10.1016/S0378-3758(01)00208-7

Augustin, T., Coolen, F. P., De Cooman, G., & Troffaes, M. C. (2014). *Introduction to imprecise probabilities*. Hoboken: Wiley.

Bahadur, R. R. (1964). On Fisher's bound for asymptotic variances. *The Annals of Mathematical Statistics, 35*(4), 1545–1552.

Berger, J. O. (1990). Robust Bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference, 25*(3), 303–328. https://doi.org/10.1016/0378-3758(90)90079-A

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. Hoboken: Wiley.

Buckley, J. J. (2004). *Fuzzy statistics*. Berlin, Heidelberg: Springer.

Burin, D. I., Jorge, R. E., Arizaga, R. A., & Paulsen, J. S. (2000). Estimation of premorbid intelligence: The word accentuation test—Buenos Aires version. *Journal of Clinical and Experimental Neuropsychology, 22*(5), 677–685. https://doi.org/10.1076/1380-3395(200010)22:5;1-9;FT677

Cameron, A., & Trivedi, P. (2005). *Microeconometrics: Methods and applications*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511811241

Cassidy, R., & Manski, C. F. (2019). Tuberculosis diagnosis and treatment under uncertainty. *Proceedings of the National Academy of Sciences of the United States of America, 116*(46), 22990–22997.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1.

Chaudhuri, S., Handcock, M. S., & Rendall, M. S. (2008). Generalized linear models incorporating population level information: An empirical-likelihood based approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 70*(2), 311–328.

Del Ser, T., González-Montalvo, J.-I., Martinez-Espinosa, S., Delgado-Villapalos, C., & Bermejo, F. (1997). Estimation of premorbid intelligence in Spanish people with the word accentuation test and its application to the diagnosis of dementia. *Brain and Cognition, 33*(3), 343–356. https://doi.org/10.1006/brcg.1997.0877

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap* (Vol. 57). New York, NY: Chapman & Hall.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association, 100*(470), 680–701. https://doi.org/10.1198/016214505000000105

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50*(4), 1029–1054.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251–1271.

Hellerstein, J. K., & Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *The Review of Economics and Statistics, 81*(1), 1–14.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Hoboken: Wiley.

Huber, P. J. (1981). *Robust statistics*. Hoboken: Wiley. https://doi.org/10.1002/0471725250.ch7

Imbens, G. W., & Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies, 61*(4), 655–680.

Insua, D. R., & Ruggeri, F. (Eds.). (2000). *Robust Bayesian analysis*. New York: Springer. https://doi.org/10.1007/978-1-4612-1306-2

Jann, M. (2023). Testing the coherence of data and external intervals via an imprecise Sargan–Hansen test. In *International symposium on imprecise probability: Theories and applications* (pp. 249–258).

Kadane, J. B., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society. Series D (The Statistician), 47*(1), 3–19.

Kwakernaak, H. (1978). Fuzzy random variables—I. Definitions and theorems. *Information Sciences, 15*(1), 1–29. https://doi.org/10.1016/0020-0255(78)90019-1

Lele, S. R., & Das, A. (2000). Elicited data and incorporation of expert opinion for statistical inference in spatial studies. *Mathematical Geology, 32*, 465–487. https://doi.org/10.1023/A:1007525900030

Manski, C. F. (1993). Identification problems in the social sciences. *Sociological Methodology, 23*, 1–56.

Manski, C. F. (2003). *Partial identification of probability distributions*. Berlin: Springer.

Manski, C. F., & Pepper, J. V. (2013). Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology, 29*(1), 123–141.

Meyers, J. E., Zellinger, M. M., Kockler, T., Wagner, M., & Miller, R. M. (2013). A validated seven-subtest short form for the Wais-IV. *Applied Neuropsychology: Adult, 20*(4), 249–256. https://doi.org/10.1080/09084282.2012.710180

Newey, W. K., & McFadden, D. (1994). *Chapter 36 large sample estimation and hypothesis testing*. Amsterdam: Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika, 75*(2), 237–249.
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine, 18*(3), 1–15. https://doi.org/10.1371/journal.pmed.1003583
Peng, P., Wang, T., Wang, C., & Lin, X. (2019). A meta-analysis on the relation between fluid intelligence and reading/mathematics: Effects of tasks, age, and social economics status. *Psychological Bulletin, 145*(2), 189–236. https://doi.org/10.1037/bul0000182
Pluck, G. (2018). Lexical reading ability predicts academic achievement at university level. *Cognition, Brain, Behavior, 22*(3), 175–196.
Pluck, G. (2020a). *Datasets for: Estimation of premorbid intelligence and executive cognitive functions with lexical reading tasks*. https://doi.org/10.23668/psycharchives.2897
Pluck, G. (2020b). A lexical decision task to measure crystallized-verbal ability in spanish. *Revista Latinoamericana de Psicologia, 52*, 1–10.
Pluck, G., & Ruales-Chieruzzi, C. B. (2021). Estimation of premorbid intelligence and executive cognitive functions with lexical reading tasks. *Psychology and Neuroscience, 14*, 358.
R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.Rproject.org/
Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica, 26*(3), 393–415.
Spiess, M. (1998). A mixed approach for the estimation of probit models with correlated responses: Some finite sample results. *Journal of Statistical Computation and Simulation, 61*(1–2), 39–59. https://doi.org/10.1080/00949659808811901
Spiess, M., Jordan, P., & Wendt, M. (2019). Simplified estimation and testing in unbalanced repeated measures designs. *Psychometrika, 84*(1), 212–235. https://doi.org/10.1007/s11336-018-9620-2
Steffen, A., Thom, J., Jacobi, F., Holstiege, J., & Bätzing, J. (2020). Trends in prevalence of depression in Germany between 2009 and 2017 based on nationwide ambulatory claims data. *Journal of Affective Disorders, 271*, 239–247. https://doi.org/10.1016/j.jad.2020.03.082
Vaart, A. W. (1998). M–and z-estimators. In *Asymptotic statistics* (pp. 41– 84). Cambridge University Press. 10.1017/CBO9780511802256.006
Walter, G., & Augustin, T. (2009). Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice, 3*(1), 255–271. https://doi.org/10.1080/15598608.2009.10411924
Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning, 24*(2), 149–170. https://doi.org/10.1016/S0888-613X(00)00032-3
Weichselberger, K. (2001). Elementare grundbegriffe einer allgemeineren wahrschein-lichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes konzept (Vol. 1). Berlin Heidelberg: Springer.
Weiss, R. H. (2006). *Cft 20-r: Grundintelligenztest skala 2-revision*. Gottingen: Hogrefe.
Weiss, R. H. (2019). *Cft 20-r mit ws: Grundintelligenztest skala 2-revision (cft 20-r) mit wortschatztest und zahlenfolgentest-revision (ws/zf-r)* (2nd ed.). Gottingen: Hogrefe.
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817–838.
Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: How to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1167.
Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353.
Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software, 11*(10), 1–17. https://doi.org/10.18637/jss.v011.i10
Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object oriented implementation of clustered covariances in R. *Journal of Statistical Software, 95*(1), 1–36. https://doi.org/10.18637/jss.v095.i01
Zhong, B., & Rao, J. N. K. (2000). Empirical likelihood inference under stratified random sampling using auxiliary population information. *Biometrika, 87*(4), 929–938.

Manuscript Received: 21 DEC 2022
Published Online Date: 20 FEB 2024

Using External Information for More Precise Inferences in General Regression Models: Supplementary Material II - Proofs and Tables

1 Introduction

This document presents the proofs of Corollary 1, Theorem 1, and the expressions in Table 1 and Table 2 in Section 3, as well as the results of the simulation study and the application of the externally informed linear model discussed in Sections 5 and 6 of the main paper. Each heading includes the relevant section in the main paper that cites the results presented under that heading.

2 Proofs (Section 3)

2.1 Corollary 1 (Section 3.1)

We start with the proof of Corollary 1:

Corollary. Assume $\hat{\boldsymbol{\theta}}_M$ is the GMM-estimator based on the model estimating equations alone (ignoring the external moments), and that $\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ have the same dimension. Using the prerequisite $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}) = [\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})^T, \mathbf{h}(\mathbf{z})^T]^T$ it follows, that $\boldsymbol{\Omega}$ has the block form

$$\boldsymbol{\Omega} = \begin{pmatrix} E[\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})^T] & E[\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})\mathbf{h}(\mathbf{z})^T] \\ E[\mathbf{h}(\mathbf{z})\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})^T] & E[\mathbf{h}(\mathbf{z})\mathbf{h}(\mathbf{z})^T] \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Omega}_M & \boldsymbol{\Omega}_R^T \\ \boldsymbol{\Omega}_R & \boldsymbol{\Omega}_h \end{pmatrix}$$

and that

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{ex}) = \mathrm{Var}(\hat{\boldsymbol{\theta}}_M) - \frac{1}{n}\{E[\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0)]^T\}^{-1}\boldsymbol{\Omega}_R^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Omega}_R\{E[\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0)]\}^{-1} \tag{1}$$

Proof. The block form of $\boldsymbol{\Omega}$ follows directly. The variance is $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{ex}) = \frac{1}{n}(\mathbf{G}^T\mathbf{W}\boldsymbol{G})^{-1}$. Because $\mathbf{h}(\mathbf{z})$ does not depend on $\boldsymbol{\theta}$, we have $E(\nabla_{\boldsymbol{\theta}}\mathbf{h}(\mathbf{z})) = \mathbf{0}$, leading to $\mathbf{G} = E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0)^T, \mathbf{0})^T$. Using this form of $\mathbf{G}$ and partitioning $\mathbf{W}$ in the same way as $\boldsymbol{\Omega}$ leads to

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{ex}) = \frac{1}{n}[E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0))^T]^{-1}\mathbf{W}_M^{-1}[E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0))]^{-1}$$

as $E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0))^T$ is a square matrix and is non-singular because both $\mathbf{W}_M$ and $\mathbf{G}^T\mathbf{W}\boldsymbol{G}$ are non-singular. Applying results for inverse blocks of partitioned matrices based on Schur complements (Chamberlain, 1987, p. 329, Lemma A.1.)

to $\mathbf{W}$ and $\mathbf{\Omega}$, leads to $\mathbf{W}_M^{-1} = \mathbf{\Omega}_M - \mathbf{\Omega}_R^T\mathbf{\Omega}_h^{-1}\mathbf{\Omega}_R$. This completes the proof, since $\text{Var}(\hat{\boldsymbol{\theta}}_M) = \frac{1}{n}[E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0))^T]^{-1}\mathbf{\Omega}_M[E(\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{z}, \boldsymbol{\theta}_0))]^{-1}$. □

2.2 Theorem 1 (Section 3.2)

Now we continue with the proof of Theorem 1.

Theorem. Let $\mathbf{H} = [\mathbf{h}(\mathbf{x}_1, y_1), \ldots, \mathbf{h}(\mathbf{x}_n, y_n)]^T$ be the $(n \times q)$ random matrix containing the externally informed sample moment functions and $\mathbf{1}_n$ a $(n \times 1)$-vector of ones. Further let $\hat{\mathbf{\Omega}}_h$ and $\hat{\mathbf{\Omega}}_R$ be consistent estimators of the corresponding matrices in Corollary 1. Then the (consistent) externally informed OLS estimator is

$$\hat{\boldsymbol{\beta}}_{ex} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - (\mathbf{X}^T\mathbf{X})^{-1}\hat{\mathbf{\Omega}}_R^T\hat{\mathbf{\Omega}}_h^{-1}\mathbf{H}^T\mathbf{1}_n$$

and its variance is

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}_{ex}) &= \text{Var}(\hat{\boldsymbol{\beta}}) - \mathbf{D} \\ &= \frac{1}{n}\sigma^2[E(\mathbf{x}\mathbf{x}^T)]^{-1} - \frac{1}{n}[E(\mathbf{x}\mathbf{x}^T)]^{-1}\mathbf{\Omega}_R^T\mathbf{\Omega}_h^{-1}\mathbf{\Omega}_R[E(\mathbf{x}\mathbf{x}^T)]^{-1},\end{aligned}$$

where σ^2 is the variance of the error in the assumed linear model.

The variance of the estimator shown in Theorem 1 can be seen as a special case of the variance formula in Corollary 1 and it was also derived by Hellerstein and Imbens (1999), hence we will only derive $\hat{\boldsymbol{\beta}}_{ex}$ here:

Proof. Using the notation of Definition 2, the regularity conditions are fulfilled for the externally informed linear model. The first order conditions for the GMM-estimator are $\hat{\mathbf{G}}^T\hat{\mathbf{W}}[\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})] = \mathbf{0}$ (Newey & McFadden, 1994)[p.

2145], where $\hat{\mathbf{G}}$ is a consistent estimator for $\mathbf{G}$. In the multiple linear case $\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i,\boldsymbol{\theta})$ becomes $\begin{pmatrix}\frac{1}{n}\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ \frac{1}{n}\mathbf{H}^T\mathbf{1}\end{pmatrix}$ and it's $\hat{\mathbf{G}}$ is $\frac{1}{n}(\mathbf{X}^T\mathbf{X},\mathbf{0})^T$. Partitioning $\hat{\mathbf{W}}=\hat{\boldsymbol{\Omega}}^{-1}$ in the same manner as $\boldsymbol{\Omega}$ and solving for $\boldsymbol{\beta}$ we get

$$
\begin{aligned}
\mathbf{0} &= \hat{\mathbf{G}}^T\hat{\mathbf{W}}[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i,\boldsymbol{\theta})] = \frac{1}{n}(\mathbf{X}^T\mathbf{X},\mathbf{0})\hat{\mathbf{W}}\begin{pmatrix}\frac{1}{n}\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ \frac{1}{n}\mathbf{H}^T\mathbf{1}\end{pmatrix}\\
&= \mathbf{X}^T\mathbf{X}\begin{pmatrix}\hat{\mathbf{W}}_M & \hat{\mathbf{W}}_R^T\end{pmatrix}\begin{pmatrix}\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ \mathbf{H}^T\mathbf{1}\end{pmatrix} = \hat{\mathbf{W}}_M\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\hat{\mathbf{W}}_R^T\mathbf{H}^T\mathbf{1}\\
&\Rightarrow \hat{\mathbf{W}}_M\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \hat{\mathbf{W}}_M\mathbf{X}^T\mathbf{y}+\hat{\mathbf{W}}_R^T\mathbf{H}^T\mathbf{1} \qquad \text{(multiply by } \hat{\mathbf{W}}_M^{-1} \text{ and } (\mathbf{X}^T\mathbf{X})^{-1})\\
&\Rightarrow \hat{\boldsymbol{\beta}}_{ex} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}+(\mathbf{X}^T\mathbf{X})^{-1}\hat{\mathbf{W}}_M^{-1}\hat{\mathbf{W}}_R^T\mathbf{H}^T\mathbf{1}.
\end{aligned}
$$

The second order derivative is $-\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}_M\mathbf{X}^T\mathbf{X}$, which is negative definite if $\mathbf{X}$ has full column rank, which proves that $\hat{\boldsymbol{\beta}}_{ex}$ is indeed the searched maximum according to Definition 2. The structure of $\hat{\mathbf{W}}$ as a partitioned inverse provides the equality $\hat{\mathbf{W}}_M^{-1}\hat{\mathbf{W}}_R^T = -\hat{\boldsymbol{\Omega}}_R^T\hat{\boldsymbol{\Omega}}_h^{-1}$. This completes the proof. □

2.3 Expressions in Table 1 (Section 3.2)

We continue with the proof for the expressions in Table 1:

Forms of $\mathbf{\Omega}_R^T$ for various single moments

moments	$h(\mathbf{x}, y)$	$\mathbf{\Omega}_R^T$
$E(y)$	$y - E(y)_{ex}$	$\sigma^2 E(\mathbf{x})$
$E(x_j y)$	$x_j y - E(x_j y)_{ex}$	$\sigma^2 E(x_j \cdot \mathbf{x})$
$E(y^2)$	$y^2 - E(y^2)_{ex}$	$2\sigma^2 E(\mathbf{x}\mathbf{x}^T)\boldsymbol{\beta}_0$
σ_y^2	$[y - E(y)]^2 - (\sigma_y^2)_{ex}$	$2\sigma^2[E(\mathbf{x}\mathbf{x}^T)\boldsymbol{\beta}_0 - E(y)E(\mathbf{x})]$
$\sigma_{x_j,y}$	$[y - E(y)][x_j - E(x_j)] - (\sigma_{x_j,y})_{ex}$	$\sigma^2 \boldsymbol{\sigma}_{x\cdot,x_j}$
$\rho_{x_j,y}$	$\frac{[y-E(y)][x_j-E(x_j)]}{\sigma_{x_j}\sigma_y} - (\rho_{x_j,y})_{ex}$	$\frac{\sigma^2}{\sigma_{x_j}\sigma_y}\boldsymbol{\sigma}_{x\cdot,x_j}$
$\beta_{x_j,y}$	$\frac{[y-E(y)][x_j-E(x_j)]}{\sigma_{x_j}^2} - (\beta_{x_j,y})_{ex}$	$\frac{\sigma^2}{\sigma_{x_j}^2}\boldsymbol{\sigma}_{x\cdot,x_j}$

Note: The subscript *ex* indicates externally determined values. In the last line, $\beta_{x_j,y}$ represents the expected value of the estimator of the slope from a simple linear regression model, which is identical to the true value of the slope only if x_j is independent of the other explanatory variables.

Proof. We only have to prove the correctness of the third column (the one for $\mathbf{\Omega}_R^T$). First we note, that $\mathbf{\Omega}_R^T = E(\mathbf{x}(y - \mathbf{x}^T\boldsymbol{\beta}_0)h(\mathbf{x}, y)) = E(\mathbf{x}\epsilon h(\mathbf{x}, y))$. We can omit the exact values of the external moments, as they are constants and as ϵ has the expected value 0. For the first row we get

$$E(\mathbf{x}\epsilon y) = E(\mathbf{x}\epsilon^2 + \epsilon\mathbf{x}\mathbf{x}^T\boldsymbol{\beta}_0) = E(\mathbf{x}\epsilon^2) = \sigma^2 E(\mathbf{x})$$

by the Gauss-Markov-assumptions. The second row follows by the same argument

just with the additional factor x_j. For the second moment of y it follows that

$$\begin{aligned} E(\mathbf{x}\epsilon y^2) = E(\mathbf{x}\epsilon(\epsilon + \mathbf{x}^T\boldsymbol{\beta}_0)^2) &= E(\mathbf{x}\epsilon^3) + 2E(\epsilon^2\mathbf{x}\mathbf{x}^T\boldsymbol{\beta}_0) + E(\epsilon\mathbf{x}(\mathbf{x}^T\boldsymbol{\beta}_0)^2) \\ &= E(\mathbf{x})E(\epsilon^3) + 2\sigma^2 E(\mathbf{x}\mathbf{x}^T)\boldsymbol{\beta}_0. \end{aligned}$$

If the errors are assumed to be at least symmetrically distributed, the first summand vanishes, leaving the term written in the third row in Table 1. For the next row, we rewrite $(y - E(y))^2$ as $y^2 - 2yE(y) + E(y)^2$ and use the linearity of the expected value. Then the $\boldsymbol{\Omega}_R^T$ of the fourth row is just the one in the fourth row minus $2E(y)$ times the one in the second row. This is $2\sigma^2 E(\mathbf{x}\mathbf{x}^T)\boldsymbol{\beta}_0 - 2\sigma^2 E(\mathbf{x})E(Y)$, which is written in the fourth row. The expression in the fifth row is derived in the same manner as we can write

$$\mathbf{x}\epsilon(x_j - E(x_j))(y - E(y)) = \mathbf{x}\epsilon x_j y - \mathbf{x}\epsilon x_j E(y) - \mathbf{x}\epsilon y E(x_j) + \mathbf{x}\epsilon E(x_j)E(y).$$

The expected value of the second and the fourth term is zero, while the first term is equal to $\boldsymbol{\Omega}_R^T$ for the moment $E(x_j y)$ and the third term is equal to $\boldsymbol{\Omega}_R^T$ for the moment $E(y)$ times $E(x_j)$. The result is $\sigma^2 E(\mathbf{x}x_j) - \sigma^2 E(\mathbf{x})E(x_j)$, which is the vector of the covariances written in the fifth row. The last two rows follow from the fifth row, treating σ_{x_j} and σ_y as constants. □

2.4 Expressions in Table 2 (Section 3.2)

Effects of various single moments in terms of variance reduction.

moments	**D**	effect on
$E(y)$	$\frac{\sigma^4}{n\omega_h}\mathbf{e}_1\mathbf{e}_1^T$	only β_1
$E(x_j y)$	$\frac{\sigma^4}{n\omega_h}\mathbf{e}_j\mathbf{e}_j^T$	only β_j
$E(y^2)$	$\frac{4\sigma^4}{n\omega_h}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^T$	all $\beta_j \neq 0$
σ_y^2	$\frac{4\sigma^4}{n\omega_h}[\boldsymbol{\beta}_0 - E(y)\mathbf{e}_1][\boldsymbol{\beta}_0 - E(y)\mathbf{e}_1]^T$	all $\beta_j \neq 0$ and β_1
$\sigma_{x_j,y}$	$\frac{\sigma^4}{n\omega_h}\tilde{\mathbf{e}}_j\tilde{\mathbf{e}}_j^T$	β_j and β_1
$\rho_{x_j,y}$	$\frac{\sigma^4}{n\omega_h\sigma_y^2\sigma_{x_j}^2}\tilde{\mathbf{e}}_j\tilde{\mathbf{e}}_j^T$	β_j and β_1
$\beta_{x_j,y}$	$\frac{\sigma^4}{n\omega_h\sigma_{x_j}^4}\tilde{\mathbf{e}}_j\tilde{\mathbf{e}}_j^T$	β_j and β_1

Note: The expression $\mathbf{e}_j$ denotes the $(p \times 1)$-vector with 1 at the j-th position and zeros elsewhere. Further we set $\tilde{\mathbf{e}}_j := -E(x_j) \cdot \mathbf{e}_1 + \mathbf{e}_j$. In the last line, $\beta_{x_j,y}$ represents the expected value of the estimator of the slope from a simple linear regression model, which is identical to the true value of the slope only if x_j is independent of the other explanatory variables.

Proof. To prove the results in Table 2 it is sufficient to use Theorem 1. As ω_h is single valued, it holds that $\mathbf{D} = \frac{1}{n\omega_h}[E(\mathbf{x}\mathbf{x}^T)]^{-1}\boldsymbol{\Omega}_R^T\boldsymbol{\Omega}_R[E(\mathbf{x}\mathbf{x}^T)]^{-1}$. To derive $[E(\mathbf{x}\mathbf{x}^T)]^{-1}\boldsymbol{\Omega}_R^T$ the expressions of $\boldsymbol{\Omega}_R^T$ in Table 1 are used. The main idea is to factorize $E(\mathbf{x}\mathbf{x}^T)$ out of $\boldsymbol{\Omega}_R^T$. As $E(\mathbf{x} \cdot x_j) = E(\mathbf{x}\mathbf{x}^T)\mathbf{e}_j$ using the notation of Table 2 and noting that $x_1 = 1$, we get the results for $[E(\mathbf{x}\mathbf{x}^T)]^{-1}\boldsymbol{\Omega}_R^T$ in Table 6.

Table 6: Expressions for $[E(\mathbf{x}\mathbf{x}^T)]^{-1}\boldsymbol{\Omega}_R^T$ depending on the moment used.

moments	$E(y)$	$E(x_j y)$	$E(y^2)$	σ_y^2	$\sigma_{x_j,y}$	$\rho_{x_j,y}$	$\beta_{x_j,y}$
$[E(\mathbf{x}\mathbf{x}^T)]^{-1}\boldsymbol{\Omega}_R^T$	$\sigma^2\mathbf{e}_1$	$\sigma^2\mathbf{e}_j$	$2\sigma^2\boldsymbol{\beta}_0$	$2\sigma^2[\boldsymbol{\beta}_0 - E(y)\mathbf{e}_1]$	$\sigma^2\tilde{\mathbf{e}}_j$	$\frac{\sigma^2}{\sigma_y\sigma_{x_j}}\tilde{\mathbf{e}}_j$	$\frac{\sigma^2}{\sigma_{x_j}^2}\tilde{\mathbf{e}}_j$

This proves the results in Table 2. □

To illustrate how to determine ω_h, the case $E(y^2)$ is treated. Using $\epsilon \sim N(0, \sigma^2)$ and the Gauss-Markov-assumptions, we get

$$\begin{aligned}
\omega_h &= E\{[y^2 - E(y^2)]^2\} = E\{[\epsilon^2 + 2\epsilon \mathbf{x}^T\boldsymbol{\beta}_0 + (\mathbf{x}^T\boldsymbol{\beta}_0)^2 - E(y^2)]^2\} \\
&= E(\epsilon^4) + E[(2\epsilon \mathbf{x}^T\boldsymbol{\beta}_0)^2] + 2E\{\epsilon^2[(\mathbf{x}^T\boldsymbol{\beta}_0)^2 - E(y^2)]\} + E\{[(\mathbf{x}^T\boldsymbol{\beta}_0)^2 - E(y^2)]^2\} \\
&+ 2E\{2\epsilon \mathbf{x}^T\boldsymbol{\beta}_0[(\mathbf{x}^T\boldsymbol{\beta}_0)^2 - E(y^2)]\} + E(2\epsilon^3 \mathbf{x}^T\boldsymbol{\beta}_0) \\
&= 3\sigma^4 + 4\sigma^2 E[(\mathbf{x}^T\boldsymbol{\beta}_0)^2] + 2\sigma^2 E[(\mathbf{x}^T\boldsymbol{\beta}_0)^2 - E(y^2)] + E\{[(\mathbf{x}^T\boldsymbol{\beta}_0)^2 - E(y^2)]^2\}.
\end{aligned}$$

References

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, *34*(3), 305–334. https://doi.org/10.1016/0304-4076(87)90015-7

Hellerstein, J. K., & Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *The Review of Economics and Statistics*, *81*(1), 1–14. http://www.jstor.org/stable/2646780

Newey, W. K., & McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4

3 Detailed results of the simulations (Section 5)

3.1 Correctly specified external moments (Section 5.2.1)

Table 7: Results of the simulations with correctly specified external moments for sample size $n = 30$.

moments	β_j	$\bar{\hat{\beta}}_{ex}$	$\mathrm{Var}(\hat{\beta}_{ex})$	$\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$	$\hat{\Delta}_j$	Cov	Cov_I	$\lvert CI\rvert$	$\lvert\bigcup CI\rvert$
$E(x_2)$	β_1	0.979	0.880	0.866	0.000	0.950	0.950	3.757	3.757
=OLS	β_2	0.489	0.077	0.087	0.000	0.970	0.970	1.184	1.184
	β_3	2.077	1.383	1.326	0.000	0.934	0.934	4.675	4.675
$E(y)$	β_1	0.978	0.618	0.646	0.263	0.948	0.974	3.230	3.677
$E(x_2y)$	β_2	0.514	0.063	0.068	0.211	0.966	0.976	1.051	1.145
$\sigma_{x_2,y}$	β_1	0.860	0.663	0.628	0.261	0.934	0.944	3.207	3.364
	β_2	0.549	0.027	0.026	0.687	0.958	0.978	0.649	0.727
$\rho_{x_2,y}$	β_1	0.911	0.676	0.619	0.272	0.932	0.944	3.183	3.337
	β_2	0.523	0.021	0.025	0.700	0.970	0.982	0.634	0.710
$\beta_{x_2,y}$	β_1	0.928	0.615	0.618	0.270	0.952	0.954	3.183	3.332
	β_2	0.515	0.014	0.025	0.697	0.994	0.994	0.633	0.707
$E(x_3y)$	β_3	2.109	1.076	1.008	0.233	0.950	0.968	4.082	4.421
$\sigma_{x_3,y}$	β_1	0.994	0.678	0.703	0.188	0.952	0.956	3.379	3.518
	β_3	2.066	0.211	0.267	0.798	0.994	0.998	2.030	2.374
$\rho_{x_3,y}$	β_1	1.001	0.729	0.700	0.192	0.938	0.948	3.373	3.512
	β_3	2.033	0.256	0.260	0.803	0.970	0.986	2.007	2.348
$\beta_{x_3,y}$	β_1	0.983	0.677	0.702	0.190	0.948	0.956	3.377	3.519
	β_3	2.063	0.177	0.258	0.803	0.994	0.998	2.005	2.352
$E(y^2)$	β_1	0.991	0.887	0.848	0.020	0.932	0.952	3.717	3.843
	β_2	0.495	0.074	0.083	0.043	0.968	0.976	1.158	1.209
	β_3	2.103	1.374	1.260	0.050	0.922	0.938	4.554	4.770
σ_y^2	β_1	0.885	0.745	0.731	0.165	0.934	0.948	3.437	3.699
	β_2	0.514	0.072	0.076	0.141	0.928	0.960	1.100	1.171
	β_3	2.189	1.332	1.127	0.164	0.892	0.914	4.280	4.587

Note. The expressions $\bar{\hat{\beta}}_{ex}$, $\mathrm{Var}(\hat{\beta}_{ex})$, $\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$, $\hat{\Delta}_j$,$\lvert CI\rvert$ and $\lvert\bigcup CI\rvert$ are defined in the beginning of Section 5.2. The results for the moment $E(x_2)$ are equivalent to the OLS results. Cov is the coverage for the external point value and Cov_I symbolizes the coverage for the confidence interval union based on the external interval. Only the affected coefficients are reported per moment. The true values are $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_3 = 2$.

Table 8: Results of the simulations with correctly specified external moments for sample size $n = 50$.

moments	β_j	$\bar{\hat{\beta}}_{ex}$	$\mathrm{Var}(\hat{\beta}_{ex})$	$\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$	$\hat{\Delta}_j$	Cov	Cov_I	$\lvert CI\rvert$	$\lvert\bigcup CI\rvert$
$E(x_2)$	β_1	1.026	0.463	0.509	0.000	0.956	0.956	2.846	2.846
	β_2	0.491	0.050	0.049	0.000	0.936	0.936	0.886	0.886
	β_3	2.013	0.881	0.799	0.000	0.942	0.942	3.574	3.574
$E(y)$	β_1	1.024	0.355	0.366	0.285	0.952	0.976	2.408	2.865
$E(x_2y)$	β_2	0.506	0.038	0.038	0.227	0.946	0.968	0.778	0.873
$\sigma_{x_2,y}$	β_1	0.954	0.326	0.367	0.272	0.966	0.982	2.419	2.574
	β_2	0.528	0.015	0.014	0.714	0.960	0.982	0.466	0.543
$\rho_{x_2,y}$	β_1	0.975	0.332	0.365	0.277	0.964	0.978	2.411	2.564
	β_2	0.517	0.012	0.014	0.720	0.962	0.980	0.461	0.536
$\beta_{x_2,y}$	β_1	0.987	0.296	0.364	0.276	0.976	0.988	2.410	2.561
	β_2	0.511	0.008	0.013	0.719	0.994	0.996	0.460	0.534
$E(x_3y)$	β_3	2.071	0.649	0.594	0.252	0.948	0.968	3.085	3.437
$\sigma_{x_3,y}$	β_1	1.019	0.357	0.410	0.196	0.966	0.978	2.551	2.689
	β_3	2.049	0.137	0.133	0.834	0.982	0.996	1.412	1.764
$\rho_{x_3,y}$	β_1	1.020	0.383	0.409	0.198	0.956	0.966	2.548	2.687
	β_3	2.037	0.157	0.131	0.837	0.942	0.972	1.403	1.754
$\beta_{x_3,y}$	β_1	1.016	0.355	0.410	0.196	0.964	0.974	2.550	2.690
	β_3	2.036	0.103	0.130	0.837	0.986	0.996	1.400	1.751
$E(y^2)$	β_1	1.035	0.469	0.497	0.025	0.946	0.960	2.810	2.933
	β_2	0.491	0.048	0.047	0.052	0.938	0.950	0.863	0.917
	β_3	2.013	0.847	0.757	0.052	0.942	0.954	3.478	3.696
σ_y^2	β_1	0.989	0.382	0.428	0.166	0.956	0.982	2.602	2.871
	β_2	0.501	0.046	0.043	0.144	0.932	0.946	0.821	0.895
	β_3	2.055	0.810	0.685	0.150	0.912	0.940	3.296	3.604

Note. The expressions $\bar{\hat{\beta}}_{ex}$, $\mathrm{Var}(\hat{\beta}_{ex})$, $\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$, $\hat{\Delta}_j$,$\lvert CI\rvert$ and $\lvert\bigcup CI\rvert$ are defined in the beginning of Section 5.2. The results for the moment $E(x_2)$ are equivalent to the OLS results. Cov is the coverage for the external point value and Cov_I symbolizes the coverage for the confidence interval union based on the external interval. Only the affected coefficients are reported per moment. The true values are $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_3 = 2$.

Table 9: Results of the simulations with correctly specified external moments for sample size $n = 100$.

moments	β_j	$\bar{\hat{\beta}}_{ex}$	$\mathrm{Var}(\hat{\beta}_{ex})$	$\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$	$\hat{\Delta}_j$	Cov	Cov_I	$\lvert CI\rvert$	$\lvert\bigcup CI\rvert$
$E(x_2)$	β_1	0.968	0.252	0.246	0.000	0.952	0.952	1.962	1.962
	β_2	0.512	0.024	0.024	0.000	0.958	0.958	0.606	0.606
	β_3	2.020	0.377	0.383	0.000	0.944	0.944	2.449	2.449
$E(y)$	β_1	0.966	0.180	0.175	0.294	0.944	0.984	1.650	2.108
$E(x_2y)$	β_2	0.515	0.018	0.018	0.235	0.952	0.982	0.530	0.623
$\sigma_{x_2,y}$	β_1	0.954	0.174	0.178	0.276	0.954	0.976	1.666	1.819
	β_2	0.519	0.007	0.006	0.731	0.944	0.986	0.312	0.388
$\rho_{x_2,y}$	β_1	0.964	0.184	0.177	0.280	0.944	0.960	1.662	1.815
	β_2	0.513	0.006	0.006	0.735	0.966	0.996	0.309	0.385
$\beta_{x_2,y}$	β_1	0.972	0.164	0.177	0.279	0.960	0.978	1.662	1.813
	β_2	0.509	0.003	0.006	0.734	0.994	0.996	0.309	0.384
$E(x_3y)$	β_3	2.030	0.273	0.281	0.262	0.962	0.990	2.101	2.459
$\sigma_{x_3,y}$	β_1	0.972	0.194	0.195	0.209	0.964	0.972	1.744	1.884
	β_3	2.018	0.047	0.055	0.856	0.984	1.000	0.910	1.260
$\rho_{x_3,y}$	β_1	0.972	0.212	0.195	0.210	0.948	0.962	1.742	1.883
	β_3	2.013	0.061	0.055	0.857	0.934	0.996	0.906	1.257
$\beta_{x_3,y}$	β_1	0.969	0.194	0.195	0.209	0.956	0.972	1.744	1.885
	β_3	2.016	0.042	0.054	0.857	0.994	1.000	0.906	1.258
$E(y^2)$	β_1	0.973	0.260	0.241	0.022	0.938	0.958	1.940	2.054
	β_2	0.512	0.022	0.022	0.059	0.962	0.974	0.588	0.645
	β_3	2.020	0.348	0.361	0.055	0.942	0.960	2.379	2.604
σ_y^2	β_1	0.945	0.189	0.203	0.180	0.952	0.984	1.777	2.065
	β_2	0.518	0.019	0.020	0.154	0.954	0.980	0.557	0.638
	β_3	2.046	0.315	0.327	0.147	0.944	0.968	2.261	2.579

Note. The expressions $\bar{\hat{\beta}}_{ex}$, $\mathrm{Var}(\hat{\beta}_{ex})$, $\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$, $\hat{\Delta}_j$,$\lvert CI\rvert$ and $\lvert\bigcup CI\rvert$ are defined in the beginning of Section 5.2. The results for the moment $E(x_2)$ are equivalent to the OLS results. Cov is the coverage for the external point value and Cov_I symbolizes the coverage for the confidence interval union based on the external interval. Only the affected coefficients are reported per moment. The true values are $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_3 = 2$.

3.2 Misspecified external moments (5.2.2)

Table 10: Results of the simulations with misspecified external moments for sample size $n = 15$.

moments	β_j	$\bar{\hat{\beta}}_{ex}$	$\text{Var}(\hat{\beta}_{ex})$	$\overline{\widehat{\text{Var}}(\hat{\beta}_{ex})}$	Cov	Cov_I	|CI|	$\lvert\bigcup CI\rvert$
$E(x_2)$	β_1	0.982	2.210	2.096	0.926	0.926	5.676	5.676
	β_2	0.499	0.228	0.223	0.964	0.964	1.843	1.843
	β_3	2.128	3.110	3.148	0.966	0.966	7.051	7.051
$E(y)$	β_1	1.438	1.962	1.690	0.890	0.954	5.102	6.484
$E(x_2y)$	β_2	0.634	0.183	0.190	0.952	0.970	1.707	1.973
$\sigma_{x_2,y}$	β_1	0.547	1.581	1.612	0.896	0.924	5.012	5.546
	β_2	0.723	0.117	0.092	0.910	0.966	1.187	1.452
$\rho_{x_2,y}$	β_1	0.647	1.280	1.593	0.934	0.948	4.945	5.481
	β_2	0.672	0.102	0.088	0.958	0.978	1.154	1.417
$\beta_{x_2,y}$	β_1	0.711	1.337	1.560	0.914	0.932	4.931	5.424
	β_2	0.640	0.057	0.083	0.968	0.984	1.141	1.384
$E(x_3y)$	β_3	2.525	2.206	2.560	0.958	0.980	6.418	7.348
$\sigma_{x_3,y}$	β_1	0.794	1.711	1.821	0.922	0.936	5.279	5.751
	β_3	2.655	0.764	1.044	0.966	0.996	3.955	5.105
$\rho_{x_3,y}$	β_1	0.796	1.546	1.819	0.940	0.952	5.256	5.739
	β_3	2.616	1.400	1.032	0.926	0.980	3.897	5.065
$\beta_{x_3,y}$	β_1	0.771	1.712	1.815	0.918	0.936	5.268	5.759
	β_3	2.648	0.734	1.003	0.948	0.998	3.893	5.067
$E(y^2)$	β_1	1.046	2.124	2.108	0.896	0.914	5.726	6.081
	β_2	0.563	0.234	0.225	0.916	0.952	1.856	1.985
	β_3	2.343	2.964	3.151	0.928	0.954	7.085	7.585
σ_y^2	β_1	0.503	3.109	1.883	0.754	0.832	5.343	5.963
	β_2	0.636	0.280	0.204	0.804	0.876	1.746	1.932
	β_3	2.638	3.537	2.828	0.812	0.896	6.615	7.358

Note. The expressions $\bar{\hat{\beta}}_{ex}$, $\text{Var}(\hat{\beta}_{ex})$, $\overline{\widehat{\text{Var}}(\hat{\beta}_{ex})}$,$|CI|$ and $\lvert\bigcup CI\rvert$ are defined in the beginning of Section 5.2. The results for the moment $E(x_2)$ are equivalent to the OLS results. Cov is the coverage for the external point value and Cov_I symbolizes the coverage for the confidence interval union based on the external interval. Only the affected coefficients are reported per moment. The true values are $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_3 = 2$.

Table 11: Results of the simulations with misspecified external moments for sample size $n = 30$.

moments	β_j	$\bar{\hat{\beta}}_{ex}$	$\mathrm{Var}(\hat{\beta}_{ex})$	$\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$	Cov	Cov_I	|CI|	$\lvert\bigcup CI\rvert$
$E(x_2)$	β_1	1.009	0.871	0.853	0.920	0.920	3.586	3.586
	β_2	0.496	0.081	0.086	0.948	0.948	1.132	1.132
	β_3	1.984	1.386	1.341	0.950	0.950	4.486	4.486
$E(y)$	β_1	1.588	0.726	0.644	0.852	0.980	3.142	4.665
$E(x_2y)$	β_2	0.623	0.063	0.070	0.936	0.974	1.032	1.301
$\sigma_{x_2,y}$	β_1	0.689	0.634	0.636	0.890	0.932	3.126	3.643
	β_2	0.656	0.034	0.029	0.894	0.972	0.671	0.924
$\rho_{x_2,y}$	β_1	0.751	0.487	0.632	0.930	0.958	3.098	3.612
	β_2	0.626	0.037	0.028	0.928	0.992	0.657	0.909
$\beta_{x_2,y}$	β_1	0.764	0.564	0.624	0.912	0.942	3.094	3.595
	β_2	0.619	0.015	0.027	0.970	0.996	0.650	0.896
$E(x_3y)$	β_3	2.411	0.994	1.041	0.944	0.984	4.003	5.013
$\sigma_{x_3,y}$	β_1	0.795	0.670	0.700	0.910	0.934	3.250	3.767
	β_3	2.529	0.226	0.287	0.872	0.998	2.061	3.315
$\rho_{x_3,y}$	β_1	0.819	0.578	0.701	0.930	0.954	3.244	3.752
	β_3	2.462	0.564	0.287	0.822	0.986	2.043	3.278
$\beta_{x_3,y}$	β_1	0.779	0.671	0.699	0.904	0.934	3.249	3.778
	β_3	2.532	0.188	0.277	0.814	0.998	2.040	3.312
$E(y^2)$	β_1	1.127	0.951	0.866	0.888	0.914	3.635	3.962
	β_2	0.569	0.097	0.086	0.906	0.938	1.137	1.273
	β_3	2.287	1.524	1.344	0.900	0.938	4.505	5.031
σ_y^2	β_1	0.456	1.456	0.744	0.716	0.790	3.321	4.051
	β_2	0.645	0.117	0.077	0.772	0.880	1.058	1.268
	β_3	2.619	1.900	1.187	0.760	0.882	4.180	5.008

Note. The expressions $\bar{\hat{\beta}}_{ex}$, $\mathrm{Var}(\hat{\beta}_{ex})$, $\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$,$|CI|$ and $\lvert\bigcup CI\rvert$ are defined in the beginning of Section 5.2. The results for the moment $E(x_2)$ are equivalent to the OLS results. Cov is the coverage for the external point value and Cov_I symbolizes the coverage for the confidence interval union based on the external interval. Only the affected coefficients are reported per moment. The true values are $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_3 = 2$.

Table 12: Results of the simulations with misspecified external moments for sample size $n = 100$.

moments	β_j	$\bar{\hat{\beta}}_{ex}$	$\mathrm{Var}(\hat{\beta}_{ex})$	$\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$	Cov	Cov_I	|CI|	$\lvert\bigcup CI\rvert$
$E(x_2)$	β_1	1.012	0.257	0.252	0.928	0.928	1.959	1.959
	β_2	0.497	0.023	0.024	0.952	0.952	0.605	0.605
	β_3	2.004	0.388	0.392	0.956	0.956	2.446	2.446
$E(y)$	β_1	1.666	0.233	0.184	0.606	0.994	1.677	3.383
$E(x_2y)$	β_2	0.621	0.017	0.019	0.850	0.994	0.543	0.831
$\sigma_{x_2,y}$	β_1	0.767	0.175	0.187	0.900	0.960	1.692	2.205
	β_2	0.619	0.008	0.007	0.726	0.990	0.339	0.590
$\rho_{x_2,y}$	β_1	0.772	0.136	0.187	0.924	0.980	1.690	2.205
	β_2	0.616	0.010	0.008	0.744	0.988	0.339	0.591
$\beta_{x_2,y}$	β_1	0.787	0.157	0.186	0.914	0.966	1.686	2.194
	β_2	0.609	0.003	0.007	0.844	1.000	0.335	0.584
$E(x_3y)$	β_3	2.470	0.252	0.301	0.882	0.986	2.150	3.265
$\sigma_{x_3,y}$	β_1	0.805	0.192	0.201	0.902	0.966	1.749	2.289
	β_3	2.533	0.048	0.057	0.354	0.996	0.925	2.258
$\rho_{x_3,y}$	β_1	0.800	0.160	0.201	0.914	0.974	1.748	2.293
	β_3	2.538	0.157	0.057	0.432	0.970	0.923	2.264
$\beta_{x_3,y}$	β_1	0.801	0.194	0.201	0.896	0.968	1.748	2.292
	β_3	2.532	0.043	0.056	0.332	0.998	0.921	2.258
$E(y^2)$	β_1	1.141	0.294	0.253	0.908	0.954	1.970	2.334
	β_2	0.567	0.030	0.024	0.864	0.944	0.600	0.765
	β_3	2.286	0.486	0.384	0.878	0.954	2.424	3.090
σ_y^2	β_1	0.547	0.580	0.209	0.630	0.840	1.772	2.695
	β_2	0.625	0.039	0.021	0.708	0.914	0.558	0.813
	β_3	2.523	0.641	0.337	0.700	0.906	2.256	3.288

Note. The expressions $\bar{\hat{\beta}}_{ex}$, $\mathrm{Var}(\hat{\beta}_{ex})$, $\overline{\widehat{\mathrm{Var}}(\hat{\beta}_{ex})}$,$|CI|$ and $\lvert\bigcup CI\rvert$ are defined in the beginning of Section 5.2. The results for the moment $E(x_2)$ are equivalent to the OLS results. Cov is the coverage for the external point value and Cov_I symbolizes the coverage for the confidence interval union based on the external interval. Only the affected coefficients are reported per moment. The true values are $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_3 = 2$.

4 Results of the application of the externally informed model (Section 6)

Table 13: Results using $\rho_{x,y} \in [.4, .85]$ and $E(y) = 100$.

		Pluck & Ruales-Chieruzzi			externally informed estimates		
j	test	$\hat{\beta}_j$	$s(\hat{\beta}_j)$	$CI_{0.95}$	$[\underline{\hat{\beta}_j}, \overline{\hat{\beta}_j}]$	$[\underline{s(\hat{\beta}_j)}, \overline{s(\hat{\beta}_j)}]$	$\bigcup CI_{0.95}$
1	SpanLex	54.61	8.864	[37.06, 72.15]	[37.41, 66.90]	[2.336, 2.663]	[32.06, 71.90]
	WAT	62.81	4.701	[53.51, 72.12]	[60.02, 68.25]	[3.587, 3.689]	[52.77, 75.65]
	SCIRT	60.81	4.395	[52.11, 69.51]	[59.01, 65.48]	[3.910, 3.990]	[51.14, 73.50]
2	SpanLex	1.821	0.332	[1.163, 2.480]	[1.334, 2.430]	[0.124, 0.132]	[1.070, 2.696]
	WAT	2.083	0.240	[1.607, 2.559]	[1.773, 2.186]	[0.190, 0.196]	[1.379, 2.568]
	SCIRT	3.292	0.358	[2.583, 4.001]	[2.882, 3.393]	[0.309, 0.317]	[2.246, 4.015]

Note. Note: The third and fourth columns contain the recomputed results of in terms of Pluck & Ruales-Chieruzzi (2021) the OLS regression coefficients $\hat{\beta}_j$, where $\hat{\beta}_1$ is the intercept and $\hat{\beta}_2$ is the slope and the robust standard errors $s(\hat{\beta}_j)$ of the coefficients. The (robust) 95% confidence intervals $CI_{0.95}$ for the parameters were computed in addition. The estimator interval $[\underline{\hat{\beta}_j}, \overline{\hat{\beta}_j}]$, the standard error interval $[\underline{s(\hat{\beta}_j)}, \overline{s(\hat{\beta}_j)}]$ and the 95% confidence interval union $\bigcup CI_{0.95}$ are shown as results of the estimation of the externally informed model.

Appendix C

Paper 3

Jann, M. (2024). Testing the fit of data and external sets via an imprecise Sargan-Hansen test. *International Journal of Approximate Reasoning*, *170*, 109214. https://doi.org/10.1016/j.ijar.2024.109214

International Journal of Approximate Reasoning 170 (2024) 109214

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Testing the fit of data and external sets via an imprecise Sargan-Hansen test

Martin Jann

Department of Research Methods and Statistics, Institute of Psychology, University of Hamburg, Von-Melle-Park 5, Hamburg, 20146, Hamburg, Germany

A R T I C L E I N F O

Keywords:
Imprecise external information
Information-data conflict
Generalized method of moments
Sargan-Hansen test
Credal set
Stochastic order

A B S T R A C T

In empirical sciences such as psychology, the term cumulative science mostly refers to the integration of theories, while external (prior) information may also be used in statistical inference. This external information can be in the form of statistical moments and is subject to various types of uncertainty, e.g., because it is estimated, or because of qualitative uncertainty due to differences in study design or sampling. Before using it in statistical inference, it is therefore important to test whether the external information fits a new data set, taking into account its uncertainties. As a frequentist approach, the Sargan-Hansen test from the generalized method of moments framework is used in this paper. It tests, given a statistical model, whether data and point-wise external information are in conflict. A separability result is given that simplifies the Sargan-Hansen test statistic in most cases. The Sargan-Hansen test is then extended to the imprecise scenario with (estimated) external sets using stochastically ordered credal sets. Furthermore, an exact small sample version is derived for normally distributed variables. As a Bayesian approach, two prior-data conflict criteria are discussed as a test for the fit of external information to the data. Two simulation studies are performed to test and compare the power and type I error of the methods discussed. Different small sample scenarios are implemented, varying the moments used, the level of significance, and other aspects. The results show that both the Sargan-Hansen test and the Bayesian criteria control type I errors while having sufficient or even good power. To facilitate the use of the methods by applied scientists, easy-to-use R functions are provided in the R script in the supplementary materials.

1. Introduction

In statistical inference and inductive reasoning in general, the goal is to derive information about the population from a sample of data, often by “inverting” the probability laws that are assumed to have generated the data [1]. An important aspect of this data generation process is the sampling mechanism. In some applied fields, such as psychology, the influence of selective sampling is often neglected, ultimately leading to biased statistical inferences about the population [2]. If external (prior) information about the population or about other data drawn from the population is available, it is possible to compare this external information with quantities computed from the data at hand. Provided that other aspects of statistical inference are valid, a mismatch between external information and new data may indicate selective sampling. The aim of this paper is therefore to discuss and develop (robust)

E-mail address: martin.jann@uni-hamburg.de.

https://doi.org/10.1016/j.ijar.2024.109214
Received 17 January 2024; Received in revised form 4 May 2024; Accepted 15 May 2024

Available online 21 May 2024

approaches for testing the fit between the available data and existing external information. Regarding statistical methods that allow the use of prior information, there are two well-known possibilities.

First, Bayesian statistics allows external information to be incorporated via a prior probability distribution, which is then combined with a likelihood function by the Bayes rule to form a posterior probability distribution [3]. Second, external information can be used to formulate constraints on the parameter space. Since many statistical estimation techniques are optimization problems, such as ordinary least squares (OLS) or maximum likelihood estimation, constrained optimization serves as a method to incorporate these external constraints on the parameter. For example, the case of linear and nonlinear regression analysis under constraints is covered by Knopov and Korkhin [4]. However, both methods primarily incorporate the external information during estimation or updating and not per se to test the fit of external information and data. In order to achieve the goal of this paper, there are still challenges to overcome, which will motivate the further approach.

A first challenge is that external information comes in different forms, some of which are more convenient for constructing prior distributions or constraints, and some of which are more difficult to use. To justify this claim, consider the following example.

Example 1. Assume a linear regression model $y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon$ under Gauss-Markov assumptions. All that is known externally is the expected value $E(y) = 100$. Under the Gauss-Markov assumptions, it is true that

$$100 = E(y) = E(\mathbf{x})^T\boldsymbol{\beta}, \tag{1}$$

which is a linear constraint on the parameter $\boldsymbol{\beta}$. Since $E(\mathbf{x})$ is not known, (1) cannot be used directly as a constraint in the optimization. Furthermore, the condition (1) is satisfied by many distributions. Therefore, it is not sufficient to identify (the moments of) a prior distribution.

The consequence is not, that it is impossible to use moment-type information as in Example 1 in constrained optimization or Bayesian statistics, but that it may not be straightforward. Precisely for this external moment-type information, there is a third way to incorporate it into statistical inference. To motivate this idea, note that many statistical estimation techniques can be represented by solving moment conditions, such as OLS and maximum likelihood estimation (via solving the score function at zero) [5, p. 172]. According to Imbens and Lancaster [6], the moment conditions used to estimate model parameters can be viewed together with external moment conditions such as (1) as a system of equations. This system of equations contains more equations than parameters and will therefore generally be overidentified. To find an estimator anyway, the Generalized Method of Moments (GMM) can be used [7]. Instead of an exact solution of the system of equations (which generally does not exist), in the GMM it is sufficient to find a vector that comes "as close as possible" to a solution with respect to a (matrix-induced) norm in order to obtain a consistent and normally distributed estimator under mild regularity conditions. The GMM is thus a third technique for incorporating external information and will be the main focus of this paper.

A second challenge is that some method is needed to actually decide whether the external information and the data fit or not. For the GMM, the Sargan-Hansen test can serve this purpose, since it tests, for a given data set, whether the moment conditions for the model estimation and those representing the external information are close enough to zero and thus whether they fit the data [8,7]. It should be noted that the use of the Sargan-Hansen test to identify model misspecification has been questioned in the presence of instrumental variables [9,10]. In this paper, it is interpreted as a test of the fit of external information to the data and not as a test of misspecification. In (classical) Bayesian statistics, there are procedures to test for *prior-data conflict*, for example via Bayesian p-values computed for a sufficient statistic based on the predictive distribution [11] or via the data agreement criterion, which is a ratio of Kullback-Leibler divergences [12]. To the best of the author's knowledge, there are no sophisticated methods in constrained optimization to check whether the constraints are "correct", except for the naive approach of calculating the distance from the constrained to the unconstrained estimator with respect to some metric.

A third challenge is that the external information will be uncertain in several ways. There are always at least slight differences between data sets, even when they are sampled from the same population, because the external information depends on time, study design, and many other aspects. Under this "qualitative uncertainty", point values will never fit a new data set exactly and may be rejected by a test procedure even though the sample value of the data set is in the range of the external values. To reflect this, an appropriate representation could be an interval constructed by the external values, allowing all values in the range. Furthermore, external information is almost always estimated itself. Therefore, its variance should be reflected, especially if it comes from small samples, to avoid false rejections.

In constrained optimization, the use of an interval of estimated external information can be implemented using the distributionally robust optimization approach. For an overview of this approach, see Rahimian and Mehrotra [13]. External intervals can be used to construct ambiguity sets, sets of probability distributions over which optimization is performed. Moment-type information can be used to construct moment-based ambiguity sets, while reflecting qualitative uncertainty by using the bounds of external moment sets. Estimation uncertainty is then addressed by calibration. Another way to incorporate external information into distributionally robust optimization could be to formulate (imprecise) chance constraints. However, to the best of the authors' knowledge, the framework of distributionally robust optimization does not include a simple test of whether new data fits externally known (imprecise) chance constraints, or whether a given ambiguity set is correct for new data. The discussion of such methods is beyond the scope of this paper, but the methods derived in this paper could be seen as a version of such tests in the scenarios considered here.

In Bayesian statistics, the use of an interval of external values leads to a set of possible prior distributions. The generalized Bayesian rule can be used to compute a corresponding set of posterior distributions. Prior-data conflict in this generalized Bayesian

scenario has been studied by Walter and Augustin [14] for imprecise Linearly Updated Conjugate prior Knowledge (iLUCK) models, following the idea that prior-data conflict should lead to a larger set of posterior ambiguity, proposed by Walley [15]. Another approach for the generalized Bayes scenario was given by Bickel [16], using sets of a-adequate models. In the GMM approach, the Sargan-Hansen test uses only external point values, but it can be extended to sets of external values.

Section 2 discusses the mathematical background and properties of the Sargan-Hansen test when external point values are used. Section 3 extends the Sargan-Hansen test to external sets. Section 4 introduces the corresponding Bayesian methods. Finally, Section 5 evaluates and compares the false positive and false negative rates of the discussed methods in different small sample scenarios based on simulation studies.

2. External information and the Sargan-Hansen test

This section covers the development of a Sargan-Hansen test based on external information in the point-valued case. Throughout this section, the notation is based on Newey and McFadden [17]. Unless otherwise indicated, (random) scalar values are represented by italic lowercase letters, (random) vectors by bold lowercase letters, and (random) matrices by bold uppercase letters.

2.1. Introduction to the point-valued case

Assume that all external information given is point-valued. Let $\boldsymbol{\Theta} \subset \mathbb{R}^q$ be the parameter space of a (fixed) parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ in a statistical model. Let $\boldsymbol{\theta}_0$ denote the "true" value of the parameter for this model, which is induced by the data generating process. The data is then assumed to be a realization of $n > 1$ random variables $\mathbf{z}_1, \dots, \mathbf{z}_n$ that are i.i.d. like a random variable $\mathbf{z}$ over $\mathbb{R}^k$. In this setting, the basic idea of the traditional method of moments to estimate the parameter would be to find a function $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ that maps onto (a subset of) $\mathbb{R}^q$ and for which $E[\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)] = \mathbf{0}$ holds [5, p. 166]. Then, the corresponding sample moment conditions $\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{0}$ are solved for $\boldsymbol{\theta}$ and the result is the desired estimator for the parameter.

In Example 1, the method of moments can be used as follows: The design matrix based on an i.i.d. sample of the independent variable $\mathbf{x}$ is

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,q-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,q-1} \end{pmatrix},$$

and likewise $\mathbf{y} = (y_1, \dots, y_n)^T$ represents an i.i.d. sample of the dependent variable y. Under the Gauss-Markov assumptions, the mixed moment of the covariates $\mathbf{x}$ and the regression error term ϵ is zero, i.e., $\mathbf{0} = E(\mathbf{x}\epsilon) = E(\mathbf{x}(y - \mathbf{x}^T\boldsymbol{\beta}_0))$. Thus, the function $\mathbf{g}(\mathbf{x}, y, \boldsymbol{\beta}) = \mathbf{x}(y - \mathbf{x}^T\boldsymbol{\beta})$ is suitable for the method of moments. The corresponding sample moment conditions are $\mathbf{0} = \frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, which give the OLS estimator when solved for $\boldsymbol{\beta}$ [5, p. 172].

In general, moment conditions are only uniquely solvable if the dimension of the parameter is equal to the number of moment conditions. However, when external information is present, there are not only the moment conditions for estimating the model parameter, but also moment conditions representing external information. Thus, the traditional method of moments must be extended. A classic example to illustrate the idea behind this extension is instrumental variable estimation, a well-known technique in econometrics [5, p. 170]. In Example 1, assume that the linear model holds, but now there is a correlation between the error term and the covariates $\mathbf{x}$. In this case, the Gauss-Markov assumptions are violated and hence OLS will generally produce inconsistent estimates. Suppose the data set contains not only the covariates but also other variables that are known to be uncorrelated with the error term. Because these variables are not included in the model itself, but may be helpful in estimating the parameter, they are called instrumental variables. The mixed moment of the instrumental variables and the error term is again zero, allowing a procedure analogous to the OLS case. Let $\mathbf{V}$ be the $n \times s$ matrix containing the n realizations of the s instrumental variables, then the sample moment conditions are $\mathbf{0} = \frac{1}{n}\mathbf{V}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Often, however, the number of potential instrumental variables exceeds the dimension of $\boldsymbol{\beta}$. In this case, the sample moment conditions are an over-identified system of equations and therefore generally not solvable for the parameter. Deletion of some moment conditions would lead to a loss of efficiency, a consequence that (applied) researchers want to avoid. One way out of this dilemma is to use an estimate for $\boldsymbol{\beta}$ that is as close as possible to a solution with respect to quadratic loss. Specifically, let $\mathbf{W}$ be a chosen positive definite weighting matrix, then the $\boldsymbol{\beta}$ that minimizes

$$(\frac{1}{n}\mathbf{V}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))^T\mathbf{W}(\frac{1}{n}\mathbf{V}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))$$

is chosen as estimate.

This is the basic idea of the GMM. It is a generalization of the traditional method of moments, because a positive quadratic form in $\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ is zero (reaches the lowest possible minimum value of a positive quadratic form) if and only if $\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{0}$. In general the so derived estimator depends on $\mathbf{W}$. Therefore, the choice of $\mathbf{W}$ is important for the properties of the estimator. Since efficiency is of great importance (as it was in the instrumental variables example), the weighting matrix $\mathbf{W}$ is mainly chosen to maximize the efficiency of the estimator. Let $\boldsymbol{\Omega} := E(\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})^T)$, then the maximum efficiency is achieved at $\mathbf{W} = \boldsymbol{\Omega}^{-1}$ [7]. In practice, however, this optimal $\mathbf{W}$ is unknown and must be estimated. The corresponding estimator is denoted by $\hat{\mathbf{W}}$. All in all, the previous considerations motivate the following definition:

Definition 1. [17, p. 2116] Let $\mathbf{g}(\mathbf{z},\theta)$ be a function with values in $\mathbb{R}^p$, where $p \geq q$ and $E[\mathbf{g}(\mathbf{z},\theta_0)] = \mathbf{0}$ holds. Also let $\hat{\mathbf{W}} \in \mathbb{R}^{p,p}$ be a positive semidefinite (and thus symmetric) random matrix such that $||r|| = (\mathbf{r}^T \hat{\mathbf{W}} \mathbf{r})^{1/2}$ is almost surely a norm on $\mathbb{R}^p$. Then a **GMM estimator** $\hat{\theta}_{gmm}$ is defined as a maximizer of the objective function

$$\hat{Q}_n(\theta) = -(\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i,\theta))^T \hat{\mathbf{W}} (\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i,\theta)) \,. \tag{2}$$

The properties of a GMM estimator include point identification, consistency, and asymptotic normality under mild regularity conditions (including cases that are not i.i.d. but ergodic and stationary stochastic processes) [17, Theorem 3.4].

Special cases of GMM estimators are OLS estimators, maximum likelihood estimators (MLE) [5, p. 172] and estimators derived by generalized estimating equations [5, p. 790]. Generalized estimating equations are mainly used to model longitudinal data that are not normally distributed, especially discrete data [18]. In the case of MLE, its regularity conditions require that the expected value of the score function is zero when evaluated at the true parameter value. This implies the condition $E[\mathbf{g}(\mathbf{z},\theta_0)] = \mathbf{0}$ in Definition 1, with the score function serving as $\mathbf{g}$. Furthermore, the first-order conditions for maximizing the log-likelihood function are equivalent to setting the score function to zero, which provides the analog of the sample moment conditions. These properties of the score function are fundamental to establishing the consistency and asymptotic normality of the MLE (for mathematical details, see Cameron and Trivedi [5, p. 140]). Finally, some M-estimators are special cases of GMM estimators, providing a link to robust statistics. To see this, note that the equations based on a $\psi-$function that implicitly define M-estimators have the form of sample moment conditions [19, p. 46].

The inclusion of external information can now be realized by formulating additional moment conditions and combining them with the moment conditions used to estimate the model [6]. Let $\mathbf{m}(\mathbf{z},\theta)$ be a function with values in $\mathbb{R}^{p_1}$, where $p_1 \geq q$, which is used to estimate the model parameter. Let $\mathbf{h}(\mathbf{z})$ be a function that maps to $\mathbb{R}^{p_2}$, which represents the external information and is assumed to be a function of the data alone. The corresponding moment conditions are $E(\mathbf{h}(\mathbf{z})) = 0$. Results derived in this paper are generally not valid if $\mathbf{h}(\mathbf{z})$ is also dependent on the parameter. For example, a function representing the external information $E(y) = 100$ used in Example 1 is $h(\mathbf{z}) = y - 100$. Both functions are then combined to $\mathbf{g}(\mathbf{z},\theta) = (\mathbf{m}(\mathbf{z},\theta)^T, \mathbf{h}(\mathbf{z})^T)^T$, which has the dimension $p_1 + p_2 = p > q$.

Under the regularity conditions of the GMM, the maximum of the objective function plays an important role, since $-n\hat{Q}_n(\hat{\theta}_{gmm}) \xrightarrow{d} \chi^2_{p-q}$. The asymptotic χ^2 test that results from this property is called the Sargan-Hansen test [8,7]. The GMM regularity conditions can be found in Theorem 2.6 and Section 9.5 from Newey and McFadden [17], but will not be discussed in detail here. Only one regularity condition can be considered essential for the following discussions, namely $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W} = \mathbf{\Omega}^{-1}$, because it restricts the choice of $\hat{\mathbf{W}}$.

2.2. Choice of the weighting matrix $\hat{\boldsymbol{W}}$ and separability

Since $\mathbf{W} = \mathbf{\Omega}^{-1}$, it would suffice to find an estimator $\hat{\mathbf{\Omega}}$ that is nonsingular and has the property $\hat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$. Such an estimator satisfies the GMM regularity conditions using the continuous mapping theorem and the continuity of matrix inversion. By definition, $\mathbf{\Omega} = E(\mathbf{g}(\mathbf{z},\theta)\mathbf{g}(\mathbf{z},\theta)^T)$, so a natural choice would be a consistent estimator $\hat{\mathbf{\Omega}}$ that is symmetric and positive semidefinite, such as the sample analog $\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{g}(\mathbf{z}_i,\hat{\theta}_{gmm})\mathbf{g}(\mathbf{z}_i,\hat{\theta}_{gmm})^T)$, relying on a law of large numbers. Note that the dependence on $\hat{\theta}_{gmm}$ can be resolved by using an iterative estimation procedure based on a starting point for $\hat{\mathbf{\Omega}}$, such as the identity matrix. Here, two steps of iterative estimation are sufficient to provide estimators while satisfying the regularity conditions [17, p. 2171]. Now, the problem is that a (weak) law of large numbers does not guarantee the invertibility of $\hat{\mathbf{\Omega}}$.

Often, the invertibility can be assumed to hold almost surely, but in practice it is important to know what to do if a singular matrix occurs. A first approach is to check whether the entries of $\mathbf{g}(\mathbf{z},\theta)$ (viewed as random variables) are linearly dependent and, if so, to delete entries until the remaining ones are no longer linearly dependent. There are more sophisticated ways to solve the singularity problem [20]. The first is to add random noise to $\hat{\mathbf{\Omega}}$, which again assures the invertibility "only" almost surely and artificially increases the variance of the estimator. The second is to use generalized inverses, which always exist, even for singular matrices. For the Moore-Penrose inverse $\hat{\mathbf{\Omega}}^+$ of $\hat{\mathbf{\Omega}}$, it holds that $\hat{\mathbf{\Omega}}^+ \xrightarrow{p} \mathbf{\Omega}^+ = \mathbf{\Omega}^{-1}$, since we assumed $\mathbf{\Omega}$ to be invertible [20]. It is advisable to be cautious when using generalized inverses in practice, since the generalized inverse of a singular matrix can be very sensitive to small changes in the singular matrix [20]. However, since generalized inverses include the regular inverse as a special case, and since it is desirable that the reader should be able to decide which method to use, the following results are derived based on the Moore-Penrose inverse $\hat{\mathbf{W}} = \hat{\mathbf{\Omega}}^+$ where possible.

Accepting the arguments made so far, it is possible to split the objective function, and thus the test statistic of the Sargan-Hansen test, into two parts if a small additional assumption is made. To improve readability by using abbreviations, define $\overline{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}(\mathbf{z}_i,\theta)$, as well as $\overline{\mathbf{h}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}(\mathbf{z}_i)$, and denote the generalized Schur complement of a block $\mathbf{B}$ of a matrix $\mathbf{A}$ by $\mathbf{A}/\mathbf{B}$.

Lemma 1 *(Separability). Assume that the premises of Definition 1 are true. Let* $\boldsymbol{g}(\boldsymbol{z},\theta) = (\boldsymbol{m}(\boldsymbol{z},\theta)^T, \boldsymbol{h}(\boldsymbol{z})^T)^T$ *and* $\hat{\mathbf{\Omega}}$ *be symmetric and positive semidefinite. Then* $\hat{\mathbf{\Omega}}$ *has the block form*

$$\hat{\mathbf{\Omega}} = \begin{pmatrix} \hat{\mathbf{\Omega}}_m & \hat{\mathbf{\Omega}}_r^T \\ \hat{\mathbf{\Omega}}_r & \hat{\mathbf{\Omega}}_h \end{pmatrix},$$

with $\hat{\boldsymbol{\Omega}}_m \in \mathbb{R}^{p_1,p_1}$, $\hat{\boldsymbol{\Omega}}_h \in \mathbb{R}^{p_2,p_2}$. *If* $\text{rank}(\hat{\boldsymbol{\Omega}}) = \text{rank}(\hat{\boldsymbol{\Omega}}_m)+\text{rank}(\hat{\boldsymbol{\Omega}}_h)$, *it holds that*

$$\begin{aligned} -\hat{Q}_n(\theta) &= (\overline{\boldsymbol{m}}^T, \overline{\boldsymbol{h}}^T)\hat{\boldsymbol{\Omega}}^+(\overline{\boldsymbol{m}}^T, \overline{\boldsymbol{h}}^T)^T \\ &= (\overline{\boldsymbol{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\boldsymbol{h}})^T(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+(\overline{\boldsymbol{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\boldsymbol{h}}) + \overline{\boldsymbol{h}}^T\hat{\boldsymbol{\Omega}}_h^+\overline{\boldsymbol{h}}. \end{aligned} \tag{3}$$

Proof. The first statement follows from the symmetry of $\hat{\boldsymbol{\Omega}}$ and by partitioning it according to $\mathbf{g}(\mathbf{z},\theta)$, since $\mathbf{m}(\mathbf{z},\theta)$ is $\mathbb{R}^{p_1}-$valued and $\mathbf{h}(\mathbf{z})$ is $\mathbb{R}^{p_2}-$valued.

For the second statement, the first equality follows by Definition 1 with $\hat{\mathbf{W}} = \hat{\boldsymbol{\Omega}}^+$ and $\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i,\theta) = (\overline{\mathbf{m}}^T, \overline{\mathbf{h}}^T)^T$. Now, by Proposition 13.1 from Puntanen et al. [21, p. 294], the condition $\text{rank}(\hat{\boldsymbol{\Omega}}) = \text{rank}(\hat{\boldsymbol{\Omega}}_m)+\text{rank}(\hat{\boldsymbol{\Omega}}_h)$ and the block form of the symmetric positive semidefinite $\hat{\boldsymbol{\Omega}}$ from the first statement imply

$$\hat{\boldsymbol{\Omega}}^+ = \begin{pmatrix} (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+ & -(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+ \\ -\hat{\boldsymbol{\Omega}}_h^+\hat{\boldsymbol{\Omega}}_r(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+ & \hat{\boldsymbol{\Omega}}_h^+ + \hat{\boldsymbol{\Omega}}_h^+\hat{\boldsymbol{\Omega}}_r(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+ \end{pmatrix}. \tag{4}$$

It follows that

$$\begin{aligned} -\hat{Q}_n(\theta) &= (\overline{\mathbf{m}}^T, \overline{\mathbf{h}}^T)\hat{\boldsymbol{\Omega}}^+(\overline{\mathbf{m}}^T, \overline{\mathbf{h}}^T)^T \\ &= \overline{\mathbf{m}}^T(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+\overline{\mathbf{m}} - 2\overline{\mathbf{m}}^T(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}} \\ &\quad + \overline{\mathbf{h}}^T(\hat{\boldsymbol{\Omega}}_h^+ + \hat{\boldsymbol{\Omega}}_h^+\hat{\boldsymbol{\Omega}}_r(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+)\overline{\mathbf{h}} \\ &= (\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}})^T(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+(\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}}) + \overline{\mathbf{h}}^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}}. \quad \square \end{aligned}$$

Note, that the condition $\text{rank}(\hat{\boldsymbol{\Omega}}) = \text{rank}(\hat{\boldsymbol{\Omega}}_m)+\text{rank}(\hat{\boldsymbol{\Omega}}_h)$ will always hold if $\hat{\boldsymbol{\Omega}}$ is invertible. Furthermore, the condition will hold asymptotically, since for the consistent estimator $\hat{\boldsymbol{\Omega}}$ all involved matrices converge in probability to (invertible) submatrices of the invertible matrix $\boldsymbol{\Omega}$, which have full rank. Thus, the use of (3) could be justified by asymptotics. However, it is still advisable to check the rank condition to detect cases where the assumption that $\boldsymbol{\Omega}$ is invertible does not hold.

Now Lemma 1 separates the quadratic form $-\hat{Q}_n(\theta)$ into two quadratic forms in different variables. Since the Sargan-Hansen test is based on the maximum of $\hat{Q}_n(\theta)$, the quadratic form must be maximized. The following theorem shows that this optimization can be simplified if $\hat{\boldsymbol{\Omega}}_h$ is chosen not to be a function of θ, for example, by choosing $\hat{\boldsymbol{\Omega}}_h = \frac{1}{n}\sum_{i=1}^n \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^T$.

Theorem 1. *Suppose the premises of Lemma 1 hold. Suppose* $\hat{\boldsymbol{\Omega}}_h$ *is not a function of* θ. *If a* $\theta_h \in \boldsymbol{\Theta}$ *with the property* $\overline{\boldsymbol{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\boldsymbol{h}} = \boldsymbol{0}$ *exists, then*

$$-\hat{Q}_n(\hat{\theta}_{gmm}) = \overline{\boldsymbol{h}}^T\hat{\boldsymbol{\Omega}}_h^+\overline{\boldsymbol{h}}. \tag{5}$$

Proof. With Definition 1 the equations

$$-\hat{Q}_n(\hat{\theta}_{gmm}) = -\max_{\theta\in\boldsymbol{\Theta}} \hat{Q}_n(\theta) = \min_{\theta\in\boldsymbol{\Theta}} -\hat{Q}_n(\theta)$$

hold. The matrix $\hat{\boldsymbol{\Omega}}$ is positive semidefinite, so $\hat{\boldsymbol{\Omega}}^+ = \hat{\boldsymbol{\Omega}}^+\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Omega}}^+$ is also positive semidefinite. The matrix $\hat{\boldsymbol{\Omega}}_h$ is positive semidefinite because it is a principal submatrix of $\hat{\boldsymbol{\Omega}}$. By the same argument, $\hat{\boldsymbol{\Omega}}_h^+$ is also positive semidefinite. Applying (4) from the proof of Lemma 1, $(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+$ is a principal submatrix of $\hat{\boldsymbol{\Omega}}^+$ and thus also positive semidefinite. Now, using the separation based on Lemma 1, the objective function $-\hat{Q}_n(\theta)$ is a sum of two positive semidefinite quadratic forms, each with a possible global minimum of zero or higher. Using the θ_h defined in the premises, the first quadratic form $(\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}})^T(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^+(\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}})$ reaches the global minimum 0. Since $\hat{\boldsymbol{\Omega}}_h$ is not a function of θ, the second quadratic form $\overline{\mathbf{h}}^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}}$ is also not a function of the parameter θ. Thus, the second quadratic form is a constant in the optimization problem. Taken together, the global minimum is $\overline{\mathbf{h}}^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}}$. $\square$

Theorem 1 indicates that in the vast majority of scenarios, the model moment conditions simply cancel out. In these scenarios, the Sargan-Hansen test based on external information reduces to a goodness-of-fit test of the data and the external information alone. To illustrate this, consider again the context of linear regression.

Example 2. Based on Example 1, consider OLS estimation in multiple linear models when external information is known. When the design matrix $\mathbf{X}$ is of full rank, it holds that $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}} = \frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}} = \mathbf{0}$. This equation can be solved directly for $\boldsymbol{\beta}$, for all possible values of the external information.

To emphasize the importance of the condition $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^+\overline{\mathbf{h}} = \mathbf{0}$, note that it is equivalent to the main separability result of Ahu and Schmidt [22, see (5) on p. 21] (for invertible $\hat{\boldsymbol{\Omega}}_h$), if one interprets the external information as a parameter for which only

one value is possible. Furthermore, this condition follows from the first-order conditions for the GMM [22]. This implies, that if the condition fails, then the first-order conditions are not satisfiable, eventually violating the GMM regularity condition or at least making it very difficult to find the maximizer of the objective function in Definition 1.

Finally, even if the condition is not met, the parameter value is chosen for which the objective function is closest to the global minimum 0. The GMM attempts to compensate for the misspecified external information by shifting the parameter estimate. Therefore, the test statistic is generally not able to indicate the combined misspecification of the external information and the model, but instead eventually cushions overidentification or even misspecification of the model. Taken together, it is advisable to use the reduced test statistic (5) of Theorem 1 whenever possible.

To justify the use of (5) even if $\overline{\mathbf{m}} - \hat{\mathbf{\Omega}}_r^T \hat{\mathbf{\Omega}}_h^+ \overline{\mathbf{h}} \neq \mathbf{0}$, note that the reduced test statistic $-n\hat{Q}_n(\hat{\theta}_{gmm}) = n\overline{\mathbf{h}}^T \hat{\mathbf{\Omega}}_h^+ \overline{\mathbf{h}}$ has the form of a generalized Wald statistic [23,24]. This actually provides another way to prove the asymptotic validity of Lemma 1 for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{gmm}$ and hence of Theorem 1, since the Sargan-Hansen test can be reinterpreted as a generalized Wald Test [23, p. 348].[1] Theorem 2.1 from Hadi and Wells [24] or Theorem 1 from Andrews [23] establishes (under the respective regularity conditions) that the asymptotic χ^2-distribution of a generalized Wald test is invariant of the choice of the generalized inverse for $\hat{\mathbf{\Omega}}$. Choosing a generalized inverse in Banachiewicz-Schur form then allows to verify Lemma 1 and Theorem 1 with analogous proofs (see [21, p. 295] for the definition of the Banachiewicz-Schur form).

Besides consistency, the choice of $\hat{\mathbf{\Omega}}_h$ is also important to incorporate the external information in different ways. Consider the case where $\mathbf{h}(\mathbf{z})$ is linear in the external information, i.e., $\mathbf{h}(\mathbf{z}) = \hat{\mathbf{h}}(\mathbf{z}) - \mathbf{e}$, where $\mathbf{e}$ is the value of the external information and $\hat{\mathbf{h}}(\mathbf{z})$ is not a function of $\mathbf{e}$. This can be considered as a practically very important case (at least in psychology), since it covers externally given means, (co-)variances, and proportions, which are commonly reported in the majority of studies. The first possible consistent estimator is $\hat{\mathbf{\Sigma}}_h := \frac{1}{n}\sum_{i=1}^n \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^T = \frac{1}{n}\sum_{i=1}^n (\hat{\mathbf{h}}(\mathbf{z}_i) - \mathbf{e})(\hat{\mathbf{h}}(\mathbf{z}_i) - \mathbf{e})^T$, already defined above. A second possible estimator is the sample covariance matrix $\hat{\mathbf{S}}_h := \frac{1}{n-1}\sum_{i=1}^n (\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}})(\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}})^T = \frac{1}{n}\sum_{i=1}^n (\hat{\mathbf{h}}(\mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{h}}(\mathbf{z}_i))(\hat{\mathbf{h}}(\mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{h}}(\mathbf{z}_i))^T$. It is only a consistent estimator of $\hat{\mathbf{\Omega}}_h$ under the null hypothesis and the (GMM or Wald) regularity conditions, since then $\frac{1}{n}\sum_{i=1}^n \hat{\mathbf{h}}(\mathbf{z}_i) \xrightarrow{p} \mathbf{e}$. Both approaches are different because $\hat{\mathbf{S}}_h$ is not a function of the external information, a fact that will play an important role in the next section.

However, both estimators lead to test statistics that interpret the external information as fixed, and thus are unable to handle the fact that the external information is itself an estimate. To account for this fact, the first step is to consider $\mathbf{e}$ as a random variable. Under the null hypothesis, $\mathbf{0} = E(\hat{\mathbf{h}}(\mathbf{z}) - \mathbf{e})$, it follows that $E(\hat{\mathbf{h}}(\mathbf{z})) = E(\mathbf{e})$. Under these conditions, $\mathbf{\Omega}_h$ depends on the random variable $\mathbf{e}$. This is denoted by writing $\mathbf{\Omega}_{h,e}$ in this case. Now, $\mathbf{\Omega}_{h,e}$ can be derived under the null hypothesis as follows:

$$\begin{aligned}
\mathbf{\Omega}_{h,e} &= E(\mathbf{h}(\mathbf{z})\mathbf{h}(\mathbf{z})^T) \\
&= E((\hat{\mathbf{h}}(\mathbf{z}) - \mathbf{e} - E(\hat{\mathbf{h}}(\mathbf{z})) + E(\mathbf{e}))(\hat{\mathbf{h}}(\mathbf{z}) - \mathbf{e} - E(\hat{\mathbf{h}}(\mathbf{z})) + E(\mathbf{e}))^T) \\
&= \mathrm{Var}(\hat{\mathbf{h}}(\mathbf{z})) + \mathrm{Var}(\mathbf{e}) - \mathrm{Cov}(\hat{\mathbf{h}}(\mathbf{z}), \mathbf{e}) - \mathrm{Cov}(\mathbf{e}, \hat{\mathbf{h}}(\mathbf{z}))
\end{aligned}$$

Estimating $\mathbf{\Omega}_{h,e}$ requires more external knowledge than just knowing the value of $\mathbf{e}$. First, $\mathrm{Var}(\mathbf{e})$ is the entire covariance matrix for $\mathbf{e}$, so consistent estimates for all variances and covariances are needed. In many psychological papers, however, consistent variance estimates are reported or the complete data set is available as open data. For the rest of the paper, assume that consistent variance and covariance estimators exist and that their estimates are available from previous studies. Let $\widehat{\mathrm{Var}}(\mathbf{e})$ be the combined consistent covariance matrix estimator. Second, the terms $\mathrm{Cov}(\hat{\mathbf{h}}(\mathbf{z}), \mathbf{e})$ and $\mathrm{Cov}(\mathbf{e}, \hat{\mathbf{h}}(\mathbf{z}))$ need to be treated. They pose a bigger problem for the estimation of $\mathbf{\Omega}_{h,e}$, since $\mathbf{e}$ and $\hat{\mathbf{h}}(\mathbf{z})$ are based on different data sets. To solve this problem, assume that the external data and the new data are (at least) linearly independent, so that both terms vanish. Third, $\mathrm{Var}(\hat{\mathbf{h}}(\mathbf{z}))$ can be estimated consistently by $\hat{\mathbf{S}}_h$. Taken together, a consistent estimator for $\mathbf{\Omega}_{h,e}$ is $\hat{\mathbf{\Omega}}_{h,e} = \hat{\mathbf{S}}_h + \widehat{\mathrm{Var}}(\mathbf{e})$. Note that the derivations made here are based on the linearity of $\mathbf{h}(\mathbf{z})$ in $\mathbf{e}$, and thus linearity is assumed wherever $\hat{\mathbf{\Omega}}_{h,e}$ or $\hat{\mathbf{S}}_h$ are used as estimators for $\mathbf{\Omega}_h$.

3. Extension to external intervals via decision rules

In this section, the restrictive case of point-valued external information is extended to set-valued external information. To illustrate the practical relevance, consider the following example.

Example 3 *(Meta-analytical scenario).* In psychological research, and empirical research in general, there are often multiple studies on a particular effect or topic. Therefore, the same statistics are calculated for multiple data sets to see if the effect can be replicated, resulting in a set of available external information. This has led to the development of meta-analysis as a method to aggregate the variety of results (modern references for meta-analysis are [25] or [26]). Further replication never occurs under exactly the same conditions. There will be differences in study design, sampling, or confounding variables, leading to a qualitative uncertainty about the comparability of external information and new data.

[1] The theory of generalized Wald tests shows under very mild regularity conditions and the rank condition $P(\mathrm{rank}(\hat{\mathbf{\Omega}}) = \mathrm{rank}(\mathbf{\Omega})) \to 1$ if $n \to \infty$, that the asymptotic χ^2-distribution holds even if $\mathbf{\Omega}$ is singular. For an invertible $\mathbf{\Omega}$, this rank condition can be omitted, as was done in this paper. See Andrews [23], Hadi and Wells [24], Xiao [20] for the mathematical details.

Example 3 shows the presence of two different types of uncertainty, estimation uncertainty and (qualitative) "comparability" uncertainty. The first uncertainty can be addressed by using a variance term for the external information and thus using $\hat{\mathbf{\Omega}}_{h,e}$. The second uncertainty is addressed in current practice by using a random effects model when conducting a meta-analysis [25, pp. 75 – 76]. This study uses a different approach to avoid the pitfalls of assuming a single probability distribution as in a random effects model.

Let $\mathbf{M}_{ex}$ be a set representing the external information for which the fit to a data set is to be tested. In the meta-analytical scenario, the set $\mathbf{M}_{ex}$ would be a finite set of points, the observed estimates. In other scenarios, $\mathbf{M}_{ex}$ might be given by constraints. In general, $\mathbf{M}_{ex}$ is the raw external information. Sometimes this raw information does not have the structure that an applied researcher wants. If there is qualitative uncertainty due to design or sampling, it is unlikely that point estimates from other studies will be exactly correct in a new study. In this case, the researcher may want to include the intermediate values. Then the $\mathbf{M}_{ex}$ from the meta-analytical scenario could be extended to a convex set, since convexity is precisely the inclusion of the intermediate values. When the set $\mathbf{M}_{ex}$ is extended to a convex set, we will call the resulting set $\mathbf{I}_{ex}$. To simplify later optimization problems, we will assume that $\mathbf{I}_{ex}$ can be expressed by linear constraints. So if $\mathbf{M}_{ex}$ is already given by linear constraints, then $\mathbf{I}_{ex} = \mathbf{M}_{ex}$. If $\mathbf{M}_{ex}$ is given by a finite set, a different approach is preferable.

Definition 2. Let $\mathbf{M}_{ex} = \{\mathbf{e}_1, ..., \mathbf{e}_m\}$ be the set of external information with $\mathbf{e}_j = (e_{1,j}, \dots, e_{p_2,j})^T \in \mathbb{R}^{p_2}$ for $j = 1, ..., m$. Then the **interval hull of $\mathbf{M}_{ex}$** is

$$\langle \mathbf{M}_{ex} \rangle_I := \bigtimes_{i=1,\dots,p_2} [\min_{j=1,\dots,m} e_{i,j}, \max_{j=1,\dots,m} e_{i,j}]$$

For example, the points $(0,1), (0,4), (-1,3)$, and $(2,3)$ would result in the interval hull $[-1,2] \times [1,4]$ containing all observed points, but no observed point is a vertex.

While the convex hull would be sufficient to include values in between, the interval hull is used here for two reasons. First, its complexity scales exponentially in the dimension of the external values, not in the number of external values, which is helpful when a large meta-analysis is available. Second, it treats the dimensions separately in terms of maximum and minimum. In our example, the maximum of the first component is 2 and the maximum of the second is 4, but the point $(2,4)$ is only covered by the interval hull, not the convex hull. In general, the convex hull is a proper subset of the interval hull. The decision of which hull to use depends on whether a researcher wants to treat the dimensions of the external information separately, e.g., based on an independence assumption, in which case the interval hull is appropriate. We continue with $\mathbf{I}_{ex} = \langle \mathbf{M}_{ex} \rangle_I$, but the proofs in this paper also work with the convex hull.

Note that the decision to extend the set containing the external information to a convex set represents the idea that the true value fitting the data should lie within the extremes found so far. This idea is then tested as a hypothesis, i.e., the null hypothesis is $\mathbf{e}_0 \in \mathbf{I}_{ex}$, where $\mathbf{e}_0$ is the true value for the new data. However, the question of which studies to include as external information for a given effect (and which to exclude as measuring a different effect) is still qualitative, depends on the applied field, and should be answered a priori by inclusion and exclusion criteria [25, p. 5].

When $\mathbf{M}_{ex}$ is used instead of $\mathbf{I}_{ex}$, the null hypothesis is $\mathbf{e}_0 \in \mathbf{M}_{ex}$. The difference between using $\mathbf{M}_{ex}$ and $\mathbf{I}_{ex}$ when $\mathbf{M}_{ex}$ is a finite set of points can be summarized as follows: $\mathbf{I}_{ex}$ results in a more conservative test that accounts for the presence of qualitative uncertainty by including all values between the external values, while $\mathbf{M}_{ex}$ results in a more liberal test by testing only the observed external values point by point. This difference is generally more relevant the fewer points $\mathbf{M}_{ex}$ contains. Consider the case where only two previous studies are known. One study had a large estimate and the other had a small estimate, possibly due to a different design. When a new study is conducted, its estimate may be in the middle of these extremes. Using $\mathbf{M}_{ex}$ would likely reject the fit of the external information to the new data, while using $\mathbf{I}_{ex}$ would not. It is up to the researcher whether the external values should be used as a boundary for possible estimates in new studies, allowing for differences due to qualitative uncertainty, or whether the estimate should be close to one of the previous studies.

3.1. General results and asymptotics

For the rest of this paper, probability distributions are assumed to be σ-additive. The basis for extending the Sargan-Hansen test and the generalized Wald test from Section 2.2 is that the null hypotheses $\mathbf{e}_0 \in \mathbf{I}_{ex}$ and $\mathbf{e}_0 \in \mathbf{M}_{ex}$ include the point-value null hypothesis for the unknown value $\mathbf{e}_0$. Thus, there is a value in the set $\mathbf{M}_{ex}$ or in the set $\mathbf{I}_{ex}$ for which the results of Section 2.2 can be applied, but it is unknown for which value this is the case. According to the cautious data completion approach [1], all values in the external set must be considered, which leads to a set of test statistics by computing $n\overline{\mathbf{h}}^T \hat{\mathbf{\Omega}}_h^+ \overline{\mathbf{h}}$ for each value in the external set. Since $n\overline{\mathbf{h}}^T \hat{\mathbf{\Omega}}_h^+ \overline{\mathbf{h}}$ is a positive semidefinite quadratic form, this set of test statistics will always be bounded below by zero. However, there is no general upper bound, since $\mathbf{I}_{ex}$ can be arbitrarily large and it can be arbitrarily far away from the corresponding sample moment of the new data set. Regarding the distributional properties of each test statistic, multiple distributions are possible, so they form a credal set. To emphasize the dependence on $\mathbf{e}$, let $\chi^2(\mathbf{e})$ denote the test statistic based on $\mathbf{e}$. First, if $\mathbf{e}$ is the true value, then $\chi^2(\mathbf{e}) \xrightarrow{d} \chi^2_{p_2}$, since the results of Section 2.2 apply. Second, if $\mathbf{e}$ is not the true value, then $\chi^2(\mathbf{e}) \xrightarrow{d} \infty$, this is indicated by $\hat{\mathbf{\Omega}}_h^+ \xrightarrow{d} \mathbf{\Omega}_h^+ = \mathbf{\Omega}_h^{-1}$ and $\overline{\mathbf{h}} \xrightarrow{d} \mathbf{e}_0 - \mathbf{e} \neq \mathbf{0}$, which results in $\chi^2(\mathbf{e}) = n\overline{\mathbf{h}}^T \hat{\mathbf{\Omega}}_h^+ \overline{\mathbf{h}} \xrightarrow{d} n(\mathbf{e}_0 - \mathbf{e})^T \mathbf{\Omega}_h^{-1} (\mathbf{e}_0 - \mathbf{e}) \to \infty$ for $n \to \infty$, since all terms except n converge to constants [5, p. 248].

For $\mathbf{M}_{ex}$ these are all possible asymptotic distributions and thus the asymptotic credal set is $\mathcal{M}_m = \{\chi^2_{p_2}, 1_\infty\}$. Here 1_∞ denotes the indicator function at infinity, which represents the shift of the probability mass to infinity when misspecified external values are used (it can be interpreted as a probability measure using the extended real number line). For $\mathbf{I}_{ex}$ there are also external values in a shrinking neighborhood around the true value. These are defined as $\mathbf{e} = \mathbf{e}_0 + \delta/n$, where δ is a constant representing the bias. For these neighborhood values, the asymptotic distribution of the test statistic is a noncentral $\chi^2_{p_2}-$distribution under Wald regularity conditions, using Theorem 2.1 from [24]. Let $\chi^2_{p_2}(\lambda)$ denote the noncentral $\chi^2_{p_2}-$distribution with the noncentrality parameter λ. Taken together, the asymptotic credal set when using $\mathbf{I}_{ex}$ is $\mathcal{M}_I = \{\chi^2_{p_2}(\lambda) | \lambda \in [0, \infty)\} \cup \{1_\infty\}$. Note that $\chi^2_{p_2}(0)$ is the central $\chi^2_{p_2}-$distribution.

In order to derive a valid test for the fit of external information and data based on the set of test statistics and the credal set, it is useful to interpret the scenario as a decision making scenario. In the following, the presentation is based on Huntley et al. [27]. First, a set of gambles must be defined based on the set of test statistics. The test statistic $\chi^2(\mathbf{e})$ now symbolizes a realized value in a data set, not a random variable. For each $\mathbf{e} \in \mathbf{I}_{ex}$ (or $\mathbf{M}_{ex}$) consider the "p-value" event $\{\chi^2 > \chi^2(\mathbf{e})\}$, that for a fixed $\mathbf{e}$ a test statistic greater than the observed value $\chi^2(\mathbf{e})$ occurs. The indicator functions of these p-value events for all $\mathbf{e} \in \mathbf{I}_{ex}$ (or $\mathbf{M}_{ex}$) form the set of gambles.

Definition 3. Let $T(\theta)$ denote a one-dimensional test statistic dependent on a parameter θ and $t(\theta)$ denote a possible value of $T(\theta)$. Let the null hypothesis be $\theta_0 \in \mathbf{\Theta}_0$, where θ_0 is the true value of the parameter. Then the **set of p-value (event) gambles** is defined to be

$$\mathcal{K} = \{1_{\{T(\theta) > t(\theta)\}} | \theta \in \mathbf{\Theta}_0\}.$$

Since the gambles in $\mathcal{K}$ are indicator functions, linear previsions on $\mathcal{K}$ reduce to probabilities of the p-value events. Lower and upper previsions reduce to lower and upper probabilities, $\underline{P}$ and $\overline{P}$, of the p-value events, hence to lower and upper p-values. In the following, it is always assumed that $T(\theta)$ has the same measurable, closed set $A \subset \mathbb{R} \cup \infty$ as possible values for each $\theta \in \mathbf{\Theta}_0$, which allows to write $\{T > t(\theta)\}$ for the p-value events. Then $\mathcal{K}$ is a totally ordered set by using the set inclusion of the events $\{T > t(\theta)\}$ as order. The above credal sets have an important property which makes it easier to determine the upper and lower probabilities.

Definition 4. Let $\mathcal{P}_\lambda$ be a family of probability distributions on the real numbers, where λ is an one-dimensional real parameter. $\mathcal{P}_\lambda$ is said to be **stochastically ordered in** λ if for all $x \in \mathbb{R}$ and all $P_{\lambda_1}, P_{\lambda_2} \in \mathcal{P}_\lambda$ with $\lambda_1 \geq \lambda_2$ it holds that

$$P_{\lambda_1}(X > x) \geq P_{\lambda_2}(X > x).$$

While the stochastic ordering assumption may seem very restrictive, the results of this section can be applied to a variety of different cases. Many significance tests used in psychology or econometrics are based on families of distributions that are stochastically ordered in the noncentrality parameter, where the minimum distribution with respect to the stochastic order is the distribution under the null hypothesis. Examples are the noncentral χ^2- and F-distributions (later proved by references). Thus, the Wald test for general linear and nonlinear hypotheses, the likelihood ratio test, and the Langrange multiplier test are candidates for an application of the results derived here (see [5] for more details). These and other tests are used in a variety of scenarios of interest in hypothesis testing in applied research, goodness-of-fit testing (as in this paper), or model comparison [5,28].

In the proof of Corollary 1, it will be shown that the credal sets $\mathcal{M}_m$ and $\mathcal{M}_I$ are stochastically ordered in the noncentrality parameter λ. A trivial but important consequence of the stochastic ordering is that the lower and upper probabilities of the events $(X > x)$ are attained for the lowest and highest parameter values, respectively. To guarantee the existence of minimum and maximum values, it is assumed that the parameter space Λ for λ is a compact subset of the extended real line with respect to the order topology. This assumption is equivalent to the assertion that Λ is closed with respect to the order topology. With this simplification in hand, choice functions can be applied while using the implications of Figure 8.1 from Huntley et al. [27]. The choice functions treated here are $\Gamma-$maximax, $\Gamma-$maximin, E-admissible, maximal, Hurwicz and Interval dominance. In the following, $\underline{t}$ and $\overline{t}$ denote the infimum and supremum of the elements in a set of observed test statistics $\mathcal{T}_{\mathbf{\Theta}_0}$, based on $\mathbf{\Theta}_0$.

Proposition 1. *Let $\mathcal{T}_{\mathbf{\Theta}_0}$ be a set of observed test statistics, based on $\mathbf{\Theta}_0$. For each $\theta \in \mathbf{\Theta}_0$, let the credal set for the test statistic $T(\theta)$ be the same family of probability distributions $\mathcal{P}_\lambda$ stochastically ordered in λ. Then every set of optimal p-value gambles chosen by a choice function contains $1_{\{T > \underline{t}\}}$.*

Proof. Using the relations shown in Figure 8.1 from Huntley et al. [27, p. 196], it is sufficient to show that the choice functions $\Gamma-$maximax, $\Gamma-$maximin and Hurwicz choose $1_{\{T > \underline{t}\}}$ as an optimal gamble. Since $\mathcal{P}_\lambda$ is stochastically ordered in λ, the lower and upper probabilities for each event $\{T > t(\theta)\}$ are attained at $\underline{\lambda} = \min \lambda$ and $\overline{\lambda} = \max \lambda$, where minimum and maximum are taken over the possible values of λ. Since $\{T > t(\theta)\} \subset \{T > \underline{t}\}$ for every $t(\theta) \in \mathcal{T}_{\mathbf{\Theta}_0}$, it holds that

$$\underline{P}(T > \underline{t}) = P_{\underline{\lambda}}(T > \underline{t}) \geq P_{\underline{\lambda}}(T > t(\theta)) = \underline{P}(T > t(\theta)) \text{ and}$$
$$\overline{P}(T > \underline{t}) = P_{\overline{\lambda}}(T > \underline{t}) \geq P_{\overline{\lambda}}(T > t(\theta)) = \overline{P}(T > t(\theta)),$$

for every $t(\theta) \in \mathcal{T}_{\Theta_0}$. Since the events $\{T > t(\theta)\}$ correspond to the gambles $1_{\{T(\theta)>t(\theta)\}}$, this proves the optimality for $\Gamma-$maximin and $\Gamma-$maximax. The optimality of $1_{\{T>\underline{t}\}}$ for the Hurwicz choice function follows by noting that it is based on a convex combination of the two inequalities derived. □

Proposition 1 establishes that $1_{\{T>\underline{t}\}}$ can always be chosen as an optimal gamble, and the corresponding event forms the basis of p-value calculations. In the important case of asymptotic credal sets containing 1_∞, there are further simplifications.

Proposition 2. *Let $\mathcal{T}$ be a set of observed test statistics and $\mathcal{K}$ be the corresponding set of p-value gambles. Further, let the credal set contain 1_∞ and $A \cup \infty$ be the set of possible values of the test statistic. It follows that the choice functions $\Gamma-$maximax, E-admissible, maximal and Interval dominance choose all gambles in $\mathcal{K}$ as optimal. Furthermore $\Gamma-$maximin and Hurwicz choose the same set of gambles as optimal.*

Proof. The first statement only needs to be proved for $\Gamma-$maximax, the rest follows from Figure 8.1 from Huntley et al. [27, p. 196]. The measure 1_∞ assigns the value 1 to all gambles in $\mathcal{K}$, since the p-value events include all values greater than a value $t(\theta) \in \mathcal{T}$, including ∞ as a value.[2] Therefore, the upper probability is 1 for every event, and thus all gambles in $\mathcal{K}$ are optimal by the $\Gamma-$maximax rule. For the second statement, note that the upper probabilities are again equal to 1, and thus it follows for all $\beta \in [0,1]$ and all $f, g \in \mathcal{K}$ that

$$\beta \underline{P}(f) + (1-\beta)\overline{P}(f) \geq \beta \underline{P}(g) + (1-\beta)\overline{P}(g) \Leftrightarrow$$
$$\beta \underline{P}(f) + (1-\beta) \geq \beta \underline{P}(g) + (1-\beta) \Leftrightarrow \underline{P}(f) \geq \underline{P}(g). \quad \square$$

Now, Proposition 2 implies that there is asymptotically only one choice function, $\Gamma-$maximin, which effectively reduces the set of p-value gambles in the case of the credal sets $\mathcal{M}_m$ and $\mathcal{M}_I$. While one is tempted to attribute this reduction to infinity and question its validity for any finite sample, it is important to note, that for large samples the upper probabilities will generally be close to one, by convergence of the test statistic distribution to 1_∞ for misspecified values. There is an important relation to the consistency of a test statistic, i.e., the property that the power of the test converges to 1 for $n \to \infty$ [5, p. 248], since the upper probability for the misspecified cases is the maximum power based on the credal set. The consistency of a test is an important property in frequentist statistics, and tests are chosen to have the highest power, so upper probabilities will generally tend to be high when the basic frequentist scenario is assumed to hold, since then only one value is true and all other values are misspecified. To conclude, the test construction will be based on the $\Gamma-$maximin rule, i.e., choosing $\underline{P}(T > \underline{t})$ as the p-value and comparing it with the significance level α.

Definition 5. Let $T(\theta)$ be a test statistic that is a function of a parameter θ. Let $\mathcal{T}_{\Theta_0}$ be a set of observed test statistics, where $\underline{t}$ denotes its infimum. Let $\mathcal{M}$ be a credal set of possible distributions of the test statistics and $\underline{P}$ be the lower probability based on $\mathcal{M}$. Under the null hypothesis (H_0) $\theta_0 \in \Theta_0$, a **$\Gamma-$maximin test with significance level** $\alpha \in (0,1)$ is as follows:

If $\underline{P}(T > \underline{t}) < \alpha$, then reject $H_0 : \theta_0 \in \Theta_0$,

else maintain $\theta_0 \in \Theta_0$.

Note that only the lower probability $\underline{P}(T > \underline{t})$ is important for computing a $\Gamma-$maximin test, and thus it is uniquely defined, even if more events than $\{T > \underline{t}\}$ are optimal by the $\Gamma-$maximin choice function.

Aside from arguments from decision making, there are theoretical reasons to justify the use of a $\Gamma-$maximin test. Consider the above situation of testing whether an external set $\mathbf{M}_{ex}$ or its interval hull $\mathbf{I}_{ex}$ fits new data. One can argue in this situation that a rejection of the fit of the external set should also imply a rejection of all subsets of it. For example, if a test rejects the interval $[-1,1]$ as unfitting, i.e., not containing the true moment value for this data set, the test should also reject the interval $[-1,0]$ or the value 1 as unfitting. This argument translates to using $\{T > \underline{t}\}$, since this event will produce the highest p-value because it is the largest set based on the possible test statistics. If this p-value leads to a rejection of the null hypothesis, then all of the other p-values would as well. Regarding the use of the lower probability, consider the credal set $\mathcal{M}_m$, which contains only χ^2 and 1_∞, it is natural to rely on the lower probability in this case.

Furthermore, the external set is seen as a representation of the uncertainty about the true value. A straightforward strategy for constructing a test would be to apply only the distribution under the null hypothesis of the single true value. For the Sargan-Hansen test, this would be the χ^2-distribution. This motivates the following property to guarantee the validity of the test.

Definition 6. A $\Gamma-$maximin test with significance level α has **$\alpha-$level under the (asymptotic) distribution of** $T(\theta_0)$, if (asymptotically), given the null hypothesis is true,

$$P_{T(\theta_0)}(\underline{T} > t_\alpha) \leq P_{T(\theta_0)}(T(\theta_0) > t_\alpha) \leq \alpha$$

[2] If the reader is skeptical about using infinity as an element by extending the real line, it should be noted that another approach would be to use a sequence of indicator functions that shift mass to infinity and take the supremum to derive $\overline{P}$, i.e., the use of $\{1_n | n \in \mathbb{N}\}$ instead of 1_∞, which leads to the same conclusions.

holds, where $\underline{T} = \inf_{\boldsymbol{\Theta}_0} T(\boldsymbol{\theta})$ and t_α is the upper $1-\alpha-$quantile of a distribution in $\mathcal{M}$ that constitutes the lower probability $\underline{P}$ at the event $\{T > \underline{t}\}$.

The property of Definition 6 implies that the $\Gamma-$maximin test "contains" an $\alpha-$level significance tests for the (unknown) point null hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Sufficient conditions for this property can be stated as requirements for the credal set, as the following theorem shows.

Theorem 2. *Suppose that the credal set $\mathcal{M}$ is stochastically ordered, contains the (asymptotic) distribution of $T(\boldsymbol{\theta}_0)$, and that this distribution is a minimum with respect to the stochastic order of $\mathcal{M}$. Under these conditions, a $\Gamma-$maximin test with significance level $\alpha \in (0,1)$ has $\alpha-$level under the (asymptotic) distribution of $T(\boldsymbol{\theta}_0)$.*

Proof. Let $\underline{\lambda}$ denote the minimal λ. Using the stochastic order and the minimality of $\underline{\lambda}$, it follows that $P_{\underline{\lambda}}(T > t) = \underline{P}(T > t)$ for all $t \in \mathcal{T}_{\boldsymbol{\Theta}_0}$. Now, α can be mapped to the critical value t_α, defined as the upper $1-\alpha-$quantile of $P_{\underline{\lambda}}$. For the $\overline{\Gamma}-$maximin test, rejecting the null hypothesis is equivalent to the event $\underline{T} > t_\alpha$. It holds (asymptotically) that $P_{T(\boldsymbol{\theta}_0)}(T(\boldsymbol{\theta}_0) > t_\alpha) \leq \alpha$, since $P_{\underline{\lambda}}$ is the minimum with respect to the stochastic order and thus equal to $P_{T(\boldsymbol{\theta}_0)}$. Thus, under the null hypothesis $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$ it follows (asymptotically) that

$$P_{T(\boldsymbol{\theta}_0)}(\underline{T} > t_\alpha) \leq P_{T(\boldsymbol{\theta}_0)}(T(\boldsymbol{\theta}_0) > t_\alpha) \leq \alpha,$$

because $\underline{T} \leq T(\boldsymbol{\theta}_0)$ for all possible realizations. □

Now the Sargan-Hansen test (and the generalized Wald Test) can be extended to a $\Gamma-$maximin test.

Definition 7. Let $\chi^2(\mathbf{e}) = n\overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^+ \overline{\mathbf{h}}$ be the test statistic of the Sargan-Hansen test, which is a function of the external value $\mathbf{e}$. Let $\mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$) be a set of external values (representing the external information) and $\mathcal{T}_M$ be the set of observed test statistics, based on $\mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$). Suppose the credal set is $\mathcal{M}_m$ (or $\mathcal{M}_I$ in the case of $\mathbf{I}_{ex}$). The **Sargan-Hansen test for external sets with significance level** $\alpha \in (0,1)$ is defined to be the $\Gamma-$maximin test with significance level α for the null hypothesis $\mathbf{e}_0 \in \mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$) under the above conditions.

Finally, the results of this section and Section 2 can be combined to show the validity of the Sargan-Hansen test for external sets. Similar to $\underline{t}$, let $\underline{\chi^2}$ denote the minimum of the observed test statistics in the case of the Sargan-Hansen test for external sets.

Corollary 1. *Under the GMM or Wald regularity conditions of Section 2, the Sargan-Hansen test for external sets with significance level α has $\alpha-$level under the asymptotic distribution of $\chi^2(\boldsymbol{e}_0)$. It further reduces to the procedure*

If $P_{\chi^2_{p_2}}(T > \underline{\chi^2}) < \alpha$, then reject $H_0 : \boldsymbol{e}_0 \in \boldsymbol{M}_{ex}$(or $\boldsymbol{I}_{ex}$),
else maintain $\boldsymbol{e}_0 \in \boldsymbol{M}_{ex}$ (or $\boldsymbol{I}_{ex}$).

Proof. The GMM or Wald regularity conditions of Section 2 imply $\chi^2(\mathbf{e}_0) \xrightarrow{d} \chi^2_{p_2}$, which is contained in the credal sets $\mathcal{M}_m$ and $\mathcal{M}_I$. The credal set $\mathcal{M}_m$ is trivially stochastically ordered in λ by setting $\lambda = 0$ for $\chi^2_{p_2}$ and $\lambda = 1$ for 1_∞. The credal set $\mathcal{M}_I$ consists of the $\chi^2_{p_2}(\lambda)-$distributions for $\lambda \in [0,\infty)$ and 1_∞. The $\chi^2_{p_2}(\lambda)-$distributions are stochastically ordered in λ [29]. By setting $\lambda = \infty$ for 1_∞, it follows that $\mathcal{M}_I$ is stochastically ordered in λ. In both cases, the $\chi^2_{p_2}$-distribution is the minimum with respect to the stochastic order. Now, applying Theorem 2 proves the first statement. Since $\chi^2_{p_2}$ is the minimum with respect to the stochastic order, it follows that $\underline{P}(T > \underline{\chi^2}) = P_{\chi^2_{p_2}}(T > \underline{\chi^2})$. □

3.2. Small sample results

Since the Sargan-Hansen test for external sets is asymptotic in nature, it may not be an appropriate choice for small samples. To derive a test for the goodness of fit of the data and an external set in small samples, more assumptions are needed. First, assume that $\mathbf{h}(\mathbf{z})$ is linear, i.e., $\mathbf{h}(\mathbf{z}) = \hat{\mathbf{h}}(\mathbf{z}) - \mathbf{e}$. Suppose further that $\hat{\mathbf{h}}(\mathbf{z})$ is normally distributed, so that $\mathbf{h}(\mathbf{z})$ is normally distributed for each $\mathbf{e} \in \mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$). Then, by the results of Phillips [30, p. 889], the test statistic $n\overline{\mathbf{h}}\hat{\mathbf{S}}_h^{-1}\overline{\mathbf{h}}$ has the scaled noncentral F-distribution $\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)$ for a fixed $\mathbf{e}$, where λ is again the noncentrality parameter,[3] being only 0 for the true value $\mathbf{e}_0$. Since $\hat{\mathbf{S}}_h$ is a random matrix, its invertibility is not certain. However, replacing $\hat{\mathbf{S}}_h^{-1}$ with $\hat{\mathbf{S}}_h^+$ will eventually change the distribution of the test statistic. To avoid this change in distribution, it is assumed that $\hat{\mathbf{S}}_h$ is invertible almost surely.

[3] To bridge the different notation of Phillips [30] to the notation of this paper, note that his T is n here, his p is 1 here, so his $N = n-1$, and his q is p_2.

Definition 8. Let $T(\mathbf{e}) = n\overline{\mathbf{h}}\hat{\mathbf{S}}_h^+\overline{\mathbf{h}}$ be the test statistic, based on the external value $\mathbf{e}$. Let $\mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$) be a set of external values (representing the external information) and $\underline{t}$ be the infimum of the observed test statistics, based on $\mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$). Finally, the credal set $\mathcal{M}$ is set to be $\{\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda) | \lambda \in [0,\infty)\}$. Then the **small sample test for external sets with significance level** $\alpha \in (0,1)$ is defined to be the $\Gamma-$maximin test with significance level α for the null hypothesis $\mathbf{e}_0 \in \mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$) under the above conditions.

The validity of the small sample test for external sets follows directly from the results of the last section.

Theorem 3. *Suppose that $\hat{\mathbf{h}}(\mathbf{z})$ is normally distributed and $\hat{\mathbf{S}}_h$ is invertible almost surely. Then the small sample test for external sets with significance level α has $\alpha-$level under the distribution of $T(\mathbf{e}_0)$. Further, it can be reduced to the procedure*

If $P_{F_{p_2,n-p_2}}(T > \frac{n-p_2}{(n-1)p_2}\underline{t}) < \alpha$, *then reject* $H_0 : \boldsymbol{e}_0 \in \boldsymbol{M}_{ex}$ *(or* $\boldsymbol{I}_{ex}$*),*

else maintain $\boldsymbol{e}_0 \in \boldsymbol{M}_{ex}$ *(or* $\boldsymbol{I}_{ex}$*).*

Proof. From the above considerations, based on the results of Phillips [30], it follows that $T(\mathbf{e})$ has the distribution $\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)$, where $\lambda = 0$ for $\mathbf{e}_0$. Thus, the distribution of $T(\mathbf{e}_0)$ would be the minimum with respect to the stochastic order if $\mathcal{M}$ is ordered in λ. To prove the stochastic order of $\mathcal{M}$ in λ, first note that $F_{p_2,n-p_2}(\lambda)$ is stochastically ordered in λ [29]. The scaling by $\frac{(n-1)p_2}{n-p_2}$ is a strictly increasing transformation and can be inverted using the definition of pushforward measures, i.e.,

$$P_{\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)}(A) = P_{F_{p_2,n-p_2}(\lambda)}(\frac{n-p_2}{(n-1)p_2} \cdot A) \tag{6}$$

for all measurable sets A. Therefore, the stochastic order in λ carries over to $\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)$. Now, the first statement follows from Theorem 2. The reduced procedure can be derived by Definition 6, using $\underline{P}(T > \underline{t}) = P_{\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)}(T > \underline{t})$ and then using (6). □

Taken together, the small sample test for external sets is an exact small sample version of the Sargan-Hansen test for external sets for normally distributed sample moments.

3.3. Computation of the infimum of the test statistics

To apply the $\Gamma-$maximin tests developed in this paper, only the efficient computation of $\underline{t}$ is a problem, since the rest of the procedures simply consist of computing a p-value under a central χ^2- or F-distribution. To represent the dependence on the external value $\mathbf{e}$, $\overline{\mathbf{h}}$ is now written as $\overline{\mathbf{h}}(\mathbf{e})$ and $\hat{\mathbf{\Omega}}_h$ as $\hat{\mathbf{\Omega}}_h(\mathbf{e})$. If a relatively small discrete set $\mathbf{M}_{ex}$ is used, a simple way to find the infimum of the test statistic is to compute all possible test statistics and compare them. However, this method cannot be used for $\mathbf{I}_{ex}$. For $\mathbf{I}_{ex}$, in the case of a linear moment function $\mathbf{h}(\mathbf{z}) = \hat{\mathbf{h}}(\mathbf{z}) - \mathbf{e}$, efficient methods from quadratic programming are available. Linear moment functions cover the important cases where there is a closed-form estimator $\hat{\mathbf{h}}(\mathbf{z})$ for which the moment function $\mathbf{h}(\mathbf{z})$ satisfies the GMM regularity conditions. This covers means, variances, covariances, regression coefficients, and more. These are the most commonly used statistics in psychology and other disciplines. As such, they are the easiest to obtain external information about in these disciplines. It is important to note that the GMM regularity conditions are weaker than the typical i.i.d. assumption made in statistical inference. For a more detailed discussion of the importance of linear moment functions in psychology, see the work of Jann and Spiess [31].

Since $\mathbf{I}_{ex}$ is compact and $\overline{\mathbf{h}}(\mathbf{e})$ is linear, the objective function is $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{\Omega}}_h(\mathbf{e})^+ \overline{\mathbf{h}}(\mathbf{e})$ and attains its minimum on $\mathbf{I}_{ex}$, if $\hat{\mathbf{\Omega}}_h(\mathbf{e})^+$ is continuous on $\mathbf{I}_{ex}$. The simplest example is a $\hat{\mathbf{\Omega}}_h(\mathbf{e})$ that is not a function of $\mathbf{e}$ and therefore constant on $\mathbf{I}_{ex}$. This is achieved by $\hat{\mathbf{\Omega}}_h(\mathbf{e}) = \hat{\mathbf{S}}_h$. In this case, the objective function can be rewritten as $\overline{\mathbf{h}}(\mathbf{e})^T(\hat{\mathbf{S}}_h/n)^{-1}\overline{\mathbf{h}}(\mathbf{e})$ and is thus already in quadratic form on the basis of the variable $\overline{\mathbf{h}}(\mathbf{e})$. The corresponding feasible region is $\overline{\mathbf{h}}(\mathbf{I}_{ex}) = \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{h}}(\mathbf{z}_i) - \mathbf{I}_{ex}$, the image of $\mathbf{I}_{ex}$ under $\overline{\mathbf{h}}(\mathbf{e})$, which is again an interval. In summary, the optimization problem based on the use of $\hat{\mathbf{S}}_h$ is a simple quadratic program.

If $\hat{\mathbf{\Omega}}_h(\mathbf{e})$ is a continuous function on $\mathbf{e}$ that is not constant, the optimization problem can be much more complex. A simple example is the matrix $\hat{\mathbf{\Omega}}_h(\mathbf{e}) = \hat{\mathbf{\Sigma}}_h(\mathbf{e})$ from Section 2.2. For $\hat{\mathbf{\Sigma}}_h(\mathbf{e})$, the programming problem is not convex (see Fig. 1). However, there is a workaround for this case.

Theorem 4. *Suppose $\overline{\boldsymbol{h}}(\boldsymbol{e})$ is linear in $\boldsymbol{e}$. Suppose that $\overline{\boldsymbol{h}}(\boldsymbol{e})$ is in the column space of $\hat{\boldsymbol{S}}_h$, then*

$$n \cdot \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{\Sigma}}_h(\boldsymbol{e})^+ \overline{\boldsymbol{h}}(\boldsymbol{e}) = \frac{n \cdot \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{S}}_h^+ \overline{\boldsymbol{h}}(\boldsymbol{e})}{\frac{1}{n}(n - 1 + n \cdot \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{S}}_h^+ \overline{\boldsymbol{h}}(\boldsymbol{e}))},$$

which is a strictly increasing function in $n \cdot \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{S}}_h^+ \overline{\boldsymbol{h}}(\boldsymbol{e})$ for $n > 1$. If $\overline{\boldsymbol{h}}(\boldsymbol{e})$ is not in the column space of $\hat{\boldsymbol{S}}_h$, then $n \cdot \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{\Sigma}}_h(\boldsymbol{e})^+ \overline{\boldsymbol{h}}(\boldsymbol{e}) = n$.

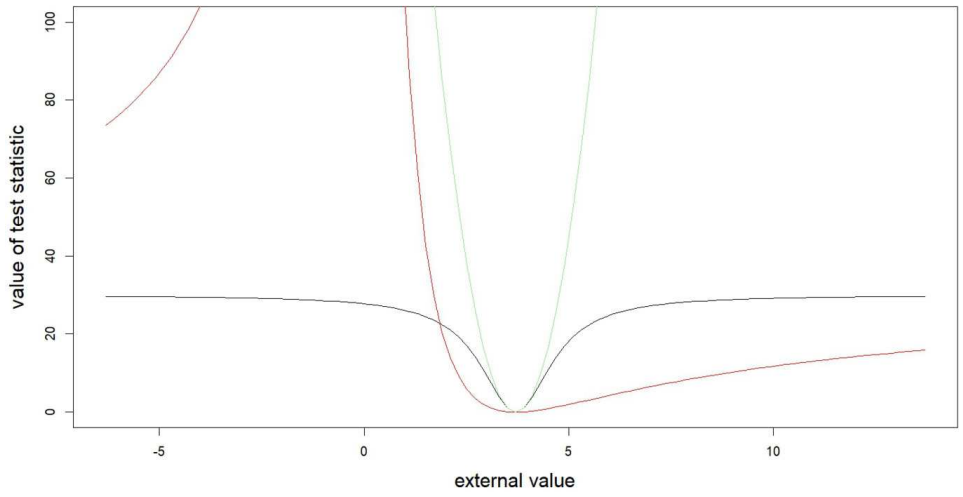


Fig. 1. Shown are plots of the test statistic $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{\Omega}}_h(\mathbf{e})^+ \overline{\mathbf{h}}(\mathbf{e})$ as a function of the external value $\mathbf{e}$ with the moment function $h(e) = y - e$, using $\hat{\mathbf{\Omega}}_h(\mathbf{e}) = \hat{\mathbf{S}}_h$ (green line), $\hat{\mathbf{\Omega}}_h(\mathbf{e}) = \hat{\mathbf{\Sigma}}_h(\mathbf{e})$ (black line), and $\hat{\mathbf{\Omega}}_h(\mathbf{e}) = \hat{\mathbf{\Omega}}_{h,e}$ (red line). In the latter case $\mathrm{Var}(e) = e^2$ was chosen. The plots are based on a sample of 30 random variables, which are i.i.d. like a normal distribution with mean 4 (true external value) and variance 1. See the R script in the supplementary materials for more details. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Proof. For the sake of brevity, $\overline{\mathbf{h}}(\mathbf{e})$ will be denoted by $\overline{\mathbf{h}}$ throughout this proof. First, note that

$$\begin{aligned}\hat{\mathbf{\Sigma}}_h(\mathbf{e}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^T \\ &= \frac{1}{n}\sum_{i=1}^{n} (\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}} + \overline{\mathbf{h}})(\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}} + \overline{\mathbf{h}})^T \\ &= \frac{n-1}{n}\hat{\mathbf{S}}_h + \overline{\mathbf{h}}\,\overline{\mathbf{h}}^T\end{aligned} \tag{7}$$

always holds, so the results of Meyer [32] can be applied to $\hat{\mathbf{\Sigma}}_h(\mathbf{e})^+$. If $\overline{\mathbf{h}}(\mathbf{e})$ is in the column space of $\hat{\mathbf{S}}_h$, the Corollary of Meyer [32, p. 320] can be applied, since $\overline{\mathbf{h}}^T(\frac{n-1}{n}\hat{\mathbf{S}}_h)^+\overline{\mathbf{h}}$ is a nonnegative quadratic form and thus $1 + \overline{\mathbf{h}}^T(\frac{n-1}{n}\hat{\mathbf{S}}_h)^+\overline{\mathbf{h}} > 0$, which yields

$$\begin{aligned}n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{\Sigma}}_h(\mathbf{e})^+ \overline{\mathbf{h}}(\mathbf{e}) &= n \cdot (\overline{\mathbf{h}}^T(\frac{n-1}{n}\hat{\mathbf{S}}_h)^+\overline{\mathbf{h}} - \frac{(\overline{\mathbf{h}}^T(\frac{n-1}{n}\hat{\mathbf{S}}_h)^+\overline{\mathbf{h}})^2}{1 + \overline{\mathbf{h}}^T(\frac{n-1}{n}\hat{\mathbf{S}}_h)^+\overline{\mathbf{h}}}) \\ &= \frac{n \cdot \overline{\mathbf{h}}^T \hat{\mathbf{S}}_h^+ \overline{\mathbf{h}}}{\frac{1}{n}(n - 1 + n \cdot \overline{\mathbf{h}}^T \hat{\mathbf{S}}_h^+ \overline{\mathbf{h}})}.\end{aligned}$$

Now, consider the function $f(x) = \frac{x}{\frac{1}{n}(n-1+x)}$, which maps $n \cdot \overline{\mathbf{h}}^T \hat{\mathbf{S}}_h^+ \overline{\mathbf{h}}$ to $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{\Sigma}}_h(\mathbf{e})^+ \overline{\mathbf{h}}(\mathbf{e})$. Since $f(x)$ is strictly increasing in x for $n > 1$ if $x \geq 0$, and since $n \cdot \overline{\mathbf{h}}^T \hat{\mathbf{S}}_h^+ \overline{\mathbf{h}} \geq 0$ holds, the first statement follows. If $\overline{\mathbf{h}}(\mathbf{e})$ is not in the column space of $\hat{\mathbf{S}}_h$, Theorem 1 of Meyer [32] leads to

$$\begin{aligned}\overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{\Sigma}}_h(\mathbf{e})^+ \overline{\mathbf{h}}(\mathbf{e}) &= \overline{\mathbf{h}}^T(\frac{n-1}{n}\hat{\mathbf{S}}_h)^+\overline{\mathbf{h}} - \frac{\overline{\mathbf{h}}^T(\frac{n-1}{n}\hat{\mathbf{S}}_h)^+\overline{\mathbf{h}}\,\overline{\mathbf{h}}^T(\mathbf{I}_{p_2} - \hat{\mathbf{S}}_h\hat{\mathbf{S}}_h^+)\overline{\mathbf{h}}}{\overline{\mathbf{h}}^T(\mathbf{I}_{p_2} - \hat{\mathbf{S}}_h\hat{\mathbf{S}}_h^+)\overline{\mathbf{h}}} \\ &\quad - \frac{\overline{\mathbf{h}}^T(\mathbf{I}_{p_2} - \hat{\mathbf{S}}_h^+\hat{\mathbf{S}}_h)\overline{\mathbf{h}}\,\overline{\mathbf{h}}^T(\frac{n-1}{n}\hat{\mathbf{S}}_h)^+\overline{\mathbf{h}}}{\overline{\mathbf{h}}^T(\mathbf{I}_{p_2} - \hat{\mathbf{S}}_h^+\hat{\mathbf{S}}_h)\overline{\mathbf{h}}} \\ &\quad + (1 + \overline{\mathbf{h}}^T(\frac{n-1}{n}\hat{\mathbf{S}}_h)^+\overline{\mathbf{h}})\frac{\overline{\mathbf{h}}^T(\mathbf{I}_{p_2} - \hat{\mathbf{S}}_h^+\hat{\mathbf{S}}_h)\overline{\mathbf{h}}\,\overline{\mathbf{h}}^T(\mathbf{I}_{p_2} - \hat{\mathbf{S}}_h^+\hat{\mathbf{S}}_h)\overline{\mathbf{h}}}{\overline{\mathbf{h}}^T(\mathbf{I}_{p_2} - \hat{\mathbf{S}}_h^+\hat{\mathbf{S}}_h)\overline{\mathbf{h}}\,\overline{\mathbf{h}}^T(\mathbf{I}_{p_2} - \hat{\mathbf{S}}_h^+\hat{\mathbf{S}}_h)\overline{\mathbf{h}}} \\ &= 1,\end{aligned}$$

where $\mathbf{I}_{p_2}$ is the corresponding identity matrix. □

Theorem 4 can be used to efficiently compute the minimum test statistic for $\hat{\mathbf{\Omega}}_h = \hat{\mathbf{\Sigma}}_h(\mathbf{e})$ by using quadratic programming based on $\hat{\mathbf{S}}_h$.

Algorithm 1. Suppose $\overline{\mathbf{h}}(\mathbf{e})$ is linear in $\mathbf{e}$. Let $\mathbf{A}$ be the matrix and $\mathbf{b}$ be the vector, so that $\overline{\mathbf{h}}(\mathbf{e}) \in \overline{\mathbf{h}}(\mathbf{I}_{ex})$ can be equivalently expressed as $\mathbf{A}\overline{\mathbf{h}}(\mathbf{e}) \leq \mathbf{b}$ (where $\leq$ is applied component-wise). Then the following algorithm computes $\underline{t} = \min_{\mathbf{e}\in\mathbf{I}_{ex}} n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{\Sigma}}_h(\mathbf{e})^+ \overline{\mathbf{h}}(\mathbf{e})$:

1. Minimize $n \cdot \mathbf{x}^T \hat{\mathbf{S}}_h \mathbf{x}$ in $\mathbf{x} \in \mathbb{R}^{p_2}$ subject to $\mathbf{A}\hat{\mathbf{S}}_h\mathbf{x} \leq \mathbf{b}$ and return the minimum value as $\underline{s}$.
2. Calculate $f(\underline{s}) = \frac{\underline{s}}{\frac{1}{n}(n-1+\underline{s})}$ and return $\underline{t} = f(\underline{s})$.

Proof. First, note that $\mathbf{A}$ and $\mathbf{b}$ always exist, since $\overline{\mathbf{h}}(\mathbf{e})$ is linear and hence $\overline{\mathbf{h}}(\mathbf{I}_{ex})$ is an interval. Theorem 4 distinguishes two cases, namely whether $\overline{\mathbf{h}}(\mathbf{e})$ is in the column space of $\hat{\mathbf{S}}_h$ or not. If it is, then $\overline{\mathbf{h}}(\mathbf{e}) = \hat{\mathbf{S}}_h\mathbf{x}$ for a $\mathbf{x} \in \mathbb{R}^{p_2}$. This leads to

$$n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^+ \overline{\mathbf{h}}(\mathbf{e}) = n \cdot \mathbf{x}^T \hat{\mathbf{S}}_h \hat{\mathbf{S}}_h^+ \hat{\mathbf{S}}_h \mathbf{x} = n \cdot \mathbf{x}^T \hat{\mathbf{S}}_h \mathbf{x},$$

and the feasible region transforms analogously. Furthermore, in this case it is valid to compute $\underline{s}$ and then transform it to $f(\underline{s}) = \frac{\underline{s}}{\frac{1}{n}(n-1+\underline{s})}$, since by Theorem 4 f is strictly increasing in $n \cdot \mathbf{x}^T \hat{\mathbf{S}}_h \mathbf{x}$ and thus preserves the minimum. Now, comparing the minimum for the first case with the minimum for the second case, which is n by Theorem 4, gives the minimum test statistic. Since $f(\underline{s})$ cannot be greater than n, the proof is complete. □

Besides its importance for computing $\min_{\mathbf{e}\in\mathbf{I}_{ex}} n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{\Sigma}}_h(\mathbf{e})^+ \overline{\mathbf{h}}(\mathbf{e})$ in the case of $\mathbf{I}_{ex}$, Theorem 4 can also be used to simplify computations in the case of a discrete set $\mathbf{M}_{ex}$. Naively, one would compute and compare all single test statistics over $\mathbf{M}_{ex}$, which would require computing a generalized inverse for each element in $\mathbf{M}_{ex}$. One could instead check which elements are in the column space of $\hat{\mathbf{S}}_h$, compute $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h \overline{\mathbf{h}}(\mathbf{e})$, and transform the minimum by f, using Theorem 4, similar to Algorithm 1.

Another interesting consequence of Theorem 4 is that a small sample test based on $\hat{\mathbf{\Sigma}}_h(\mathbf{e})$ would give the same results as the small sample test for external sets if $\hat{\mathbf{S}}_h$ is invertible. To see this, note that the column space of an invertible $\hat{\mathbf{S}}_h$ is the whole space, so the transformation f can always be used. Since f is strictly increasing, it is injective and so the event $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e}) > c$ is equivalent to $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{\Sigma}}_h(\mathbf{e})^{-1} \overline{\mathbf{h}}(\mathbf{e}) = f(n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})) > f(c)$. Thus, if c is a critical value, then $f(c)$ is the critical value for the same significance level by means of the pushforward measure, i.e., $P_{f(T)}(A) = P_T(f^{-1}(A))$. Taken together, the test decision would almost surely be unchanged if $\hat{\mathbf{\Sigma}}_h$ were used instead of $\hat{\mathbf{S}}_h$, since the small sample test for external sets assumes that $\hat{\mathbf{S}}_h$ is invertible almost surely. This results in a counterintuitive situation. The two asymptotic tests based on $\hat{\mathbf{\Sigma}}_h$ and $\hat{\mathbf{S}}_h$ are different (though asymptotically equivalent!) because the same χ^2-distribution is used to determine the p-value, while the observed test statistics are different. But their small sample versions are almost surely equivalent.

To conclude this section, the case $\hat{\mathbf{\Omega}}_{h,e} = \hat{\mathbf{S}}_h + \widehat{\mathrm{Var}}(\mathbf{e})$, which reflects the estimation uncertainty of the external information, is discussed. While for a discrete set $\mathbf{M}_{ex}$ one can just compute the test statistics and compare them, the case $\mathbf{I}_{ex}$ is more difficult. The estimator $\hat{\mathbf{\Omega}}_{h,e}$ was derived by treating $\mathbf{e}$ as a random variable. The consequence is that each $\mathbf{e} \in \mathbf{I}_{ex}$ must be treated as (the realization of) a random variable, and thus a variance estimate must be provided that is generally dependent on $\mathbf{e}$. In practice, only a finite number of variance estimates are available from external sources, e.g., from previous studies, possibly based on different sample sizes. Thus, further assumption have to be made about the variance structure of the $\mathbf{e} \in \mathbf{I}_{ex}$. The difficulties of simple approaches to solve this problem can be illustrated by the meta-analytical scenario.

Example 4. Suppose there are k independent sources of external information, indexed by $j = 1, \ldots, k$. The sources are each assumed to be based on an i.i.d. sample of size n_j. Suppose each source provides a mean $\bar{\mathbf{e}}_j$, which is assumed to be an unbiased estimate of the true moment value $E(\mathbf{e}_j)$ for that particular source, as well as a consistent covariance matrix estimate $\widehat{\mathrm{Var}}(\mathbf{e}_j)/n_j$. Since a new data set may have a slightly different true moment value, the interval hull $\mathbf{I}_{ex}$ of the $E(\mathbf{e}_j)$ is taken to counteract qualitative uncertainty. Then all $\mathbf{e} \in \mathbf{I}_{ex}$ can be written as $\mathbf{e} = \sum_{j=1}^k c_j E(\mathbf{e}_j)$ for some constants c_j. Although the true values are not known, an estimator $\hat{\mathbf{e}} = \sum_{j=1}^k c_j \bar{\mathbf{e}}_j$ can be used to estimate $\mathbf{e} \in \mathbf{I}_{ex}$ and plugged into the test statistic. To reflect the estimation uncertainty, $\widehat{\mathrm{Var}}(\hat{\mathbf{e}}) = \sum_{j=1}^k \frac{c_j^2}{n_j} \widehat{\mathrm{Var}}(\mathbf{e}_j)$ is a natural candidate for $\widehat{\mathrm{Var}}(\mathbf{e})$, since it is a consistent estimator for the variance of $\hat{\mathbf{e}}$. However, if there are different ways to write $\mathbf{e}$ as a linear or even convex combination of the $E(\mathbf{e}_j)$, then there are multiple, non-equivalent estimators for the variance. Then even the true variance of the estimator is not uniquely defined. A general solution is to write $\mathbf{e}$ as a linear combination of a fixed basis of the space spanned by the 2^{p_2} vertices of $\mathbf{I}_{ex}$, which is a unique representation based on the coefficients c_l. In the one-dimensional case, a convex combination of the two vertices is sufficient, since the interval is a line segment in this case.

In addition to the conceptual difficulties, modeling the variance of the estimator can lead to quite difficult optimization problems. In general, the objective function is not convex, see Fig. 1. In Example 4, the c_l are not known, but $\hat{\mathbf{e}}$ and $\widehat{\mathrm{Var}}(\hat{\mathbf{e}})$ depend on them. Because of the dependency, optimization should be in the c_l variables and not in $\hat{\mathbf{e}}$. However, the expected values that make up the vertices of $\mathbf{I}_{ex}$ are not known and must be estimated. Even worse, on the basis of the mean values alone, it is not possible to

know which external source is the source of the extremal expected values. This leads to the problem of estimating the minimum and maximum of a set of expected values based on given unbiased means.

For the one-dimensional case, attacking this problem using the maximum of the means is known as the optimizer's curse (see [33–35] for an overview). The name is inspired by the fact that the maximum mean $\max_j \bar{\mathbf{e}}_j$ has a nonnegative bias when estimating the maximum expected value, independent of any distributional assumption [33]. Using $\min_j \bar{\mathbf{e}}_j = -\max_j -\bar{\mathbf{e}}_j$, the minimum of the means has a nonpositive bias when estimating the minimum expected value.

The variance of $\max_j \bar{\mathbf{e}}_j$ is not known analytically, but has the upper bound $\sum_{j=1}^{k} \mathrm{Var}(\mathbf{e}_j)/n_j$ [34], which also holds for the minimum of the means, again by using $\min_j \bar{\mathbf{e}}_j = -\max_j -\bar{\mathbf{e}}_j$. Despite the existence of more sophisticated estimators for the maximum expected value, based on distributional assumptions [35], the simple maximum mean estimator is used, to test its bias in the simulation study. Based on the maximum mean estimator, a simple method to approximate an optimal value of the test statistic despite the nonconvexity of the problem is a grid search algorithm traversing the possible values of c_l. Note that since only the one-dimensional case is treated, there are 2 vertices and only one coefficient $c = c_l$, so a grid search is feasible with sufficient precision.

4. Bayesian methods

In order to have a common ground on which Bayesian methods and the tests developed in Section 3 can be compared in an easy way, Example 2 from Bickel [16] will be used and will be referred to as the given scenario in the following. Denote the normal distribution with mean μ and variance σ^2 by $N(\mu, \sigma^2)$. Let the possible sampling distributions be $N(\mu, 1)$ with $\mu \in \mathbb{R}$. For each $e \in \mathbb{R}$, the prior is chosen to be the conjugate prior, $N(e, \sigma_0^2)$, where σ_0^2 is known. For $i = 1, \dots, n$ let y_i be random variables identically distributed as $N(\mu, 1)$, where for all $i \neq j$ it is assumed that y_i is conditionally independent of y_j given μ. To test the fit of the external information to the data, the external information in the form of the set $\mathbf{I}_{ex}$ is used to constrain the prior parameter e. A prior-data conflict criterion can then be applied. The criterion from Walter and Augustin [14] is an example.

Definition 9. Let $\mathbf{I}_{ex}$ be an interval representing the external information and $\bar{y}$ be the observed sample mean. In the given scenario, the **degree of prior-data conflict** is

$$\Delta(\bar{y}; \mathbf{I}_{ex}) = \inf\{|\bar{y} - e| : e \in \mathbf{I}_{ex}\}$$

and prior-data conflict is defined by $\Delta(\bar{y}; \mathbf{I}_{ex}) > 0$.

This criterion for prior-data conflict is very similar to the Sargan-Hansen test for external sets in the case of linear moment functions discussed in Section 2 and 3. The only difference is that the metric used to measure the distance of the sample moment from the external value is the absolute value, rather than a metric induced by a matrix $\mathbf{\Omega}_h$. Moreover, this criterion can be much more liberal than any test considered so far, since the probability of $\bar{y} \notin \mathbf{I}_{ex}$ can be arbitrarily large for $\mathbf{I}_{ex}$ short enough. This is intended by Walter and Augustin [14], since they developed generalized iLUCK models that compensate for the prior-data conflict with imprecision. Therefore, their focus is on minimizing the number of false null hypotheses (not detecting true prior-data conflict), while false rejections only lead to slightly more imprecision. Note that this criterion would not work well when used with a discrete set $\mathbf{M}_{ex}$, since the probability of $\bar{y}$ being in a particular discrete set is 0 under the given scenario, so the criterion would always reject the fit of external information and data in this case. Now a second criterion is developed that is based on the work of Bickel [16].

Definition 10. In the given scenario, let $\bar{y}$ be the observed sample mean and let ϕ_{e,σ_0^2} denote the density function of $N(e, \sigma_0^2)$. Then for $a \in \mathbb{R}$ a **set of a-adequate models** is

$$\mathcal{M}(a) = \{e \in \mathbb{R} : \phi_{e,(\sigma_0^2 + \frac{1}{n})}(\bar{y}) \geq 2^a \phi_{\bar{y},(\sigma_0^2 + \frac{1}{n})}(\bar{y})\}$$

Sets of a-adequate models represent all models under the given scenario that are in adequate agreement with the data, that is, for which the likelihood exceeds a certain threshold. By simple analytic arguments, equality for the condition of $\mathcal{M}(a)$ is obtained for $e_{\min} = \bar{y} - \sqrt{-a(2\ln 2)(\sigma_0^2 + \frac{1}{n})}$ and $e_{\max} = \bar{y} + \sqrt{-a(2\ln 2)(\sigma_0^2 + \frac{1}{n})}$ if $a \leq 0$. Furthermore, in this case $\mathcal{M}(a)$ can be rewritten as $\{e \in \mathbb{R} : e_{\min} \leq e \leq e_{\max}\}$. The value a represents the strictness of the criterion. In the interpretation given by Bickel [16] based on Bayes factors, a value of -7 is a low threshold, allowing many e in $\mathcal{M}(-7)$, while -2 would be moderate and 0 would be a high threshold, allowing only $e = \bar{y}$ to be adequate. So far, $\mathcal{M}(a)$ is not linked to external information, this can be done by taking the intersection.

Definition 11. Let $\mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$) be a set representing the external information. Then **prior-data conflict with threshold a** is said to occur if $\mathcal{M}(a) \cap \mathbf{M}_{ex} = \emptyset$ (or $\mathcal{M}(a) \cap \mathbf{I}_{ex} = \emptyset$).

Note that prior-data conflict with threshold 0 is equivalent to testing whether the degree of prior-data conflict is greater than 0, i.e., for $a = 0$ both criteria are equivalent. The notion of sets of a-adequate models is more general than presented here (see [16]), but Definition 11 could easily be extended to this general notion, since only intersection is required.

5. Simulation studies

To compare the methods proposed in this paper and to investigate their small sample properties, two simulation studies were performed. The first treats the case where an interval $\mathbf{I}_{ex}$ is given but no variance for the external information, representing the case where an interval for the external information is elicited from experts. The second is based on the given scenario of Section 4 and represents the meta-analytical scenario of Example 3, so that Bayesian criteria and the Sargan-Hansen test for external sets can be compared. The simulations were performed in R, version 4.3.2 [36]. The R packages *quadprog* [37] and *MASS* [38] were used to solve the quadratic programs and to compute the Moore-Penrose inverses, respectively. All simulations are implemented in an R script, which can be found in the electronic supplementary material.

5.1. The expert opinion scenario

Based on Example 1, a linear regression model is used under Gauss-Markov assumptions and normally distributed errors. The slope is set to $\beta_2 = 1$ and the intercept to $\beta_1 = 16$. The sample sizes are $n = 30$ and $n = 50$. The samples x_i and y_i for $i = 1, \dots, n$ of the independent variable x and the dependent variable y are drawn i.i.d. as $x \sim N(4,4)$ and $y = \beta_1 + \beta_2 x + \epsilon$ with $\epsilon \sim N(0,60)$. The true R^2 is 0.0625, which is small but common in applied research such as psychology. The moments $E(y)$, $E(x)$, and $\text{Var}(y)$ and any combination of them are chosen as external information, resulting in 7 external moment scenarios. To correct for degrees of freedom, the moment function $h(z) = \frac{n}{n-1}(y - \bar{y})^2 - e$ is used for $\text{Var}(y)$. Note that $\bar{h}(e)$ is not normally distributed for $\text{Var}(y)$. To examine the type I error and the power, two scenarios are chosen regarding $\mathbf{I}_{ex}$. The first is $\mathbf{I}_{ex} = [0.95 \cdot \mathbf{e}_0, 1.05 \cdot \mathbf{e}_0]$ and the second is $\mathbf{I}_{ex} = [1.2 \cdot \mathbf{e}_0, 1.3 \cdot \mathbf{e}_0]$. The second scenario can be motivated by the proximity of $\mathbf{I}_{ex}$ to the true value $\mathbf{e}_0$ in terms of standardized mean differences. This is justified by noting that when using a single moment, the square root of the test statistic is simplified as follows:

$$\sqrt{n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})} = \sqrt{n} \frac{|\hat{h} - e|}{\sqrt{\hat{\omega}_h}} \tag{8}$$

Since (8) is similar to a t-test statistic, the standardized mean difference $d = \frac{|e_0 - e|}{\sqrt{\text{Var}(h(z))}}$ is the typical effect size [39]. To evaluate the effect sizes in the second scenario, the value $1.2 \cdot \mathbf{e}_0$ is used because it is closest to $\mathbf{e}_0$ in $\mathbf{I}_{ex}$. For $E(x)$ it is $d = \frac{|4 - 1.2 \cdot 4|}{2} = 0.4$, a small effect size, and for $E(y)$ it is $d = \frac{|20 - 1.2 \cdot 20|}{8} = 0.5$, a medium effect size [39]. Therefore, using $E(y)$ alone will asymptotically yield a higher power than using $E(x)$ alone. Under the chosen data generating process, $\frac{n}{n-1}(y - \bar{y})^2$ has a scaled χ_1^2-distribution. However, d is scale invariant, so without loss of generality, $\frac{n}{n-1}(y - \bar{y})^2$ can be assumed to have a χ_1^2-distribution with mean 1 and variance 2. Taken together, $d = 0.2 \frac{1}{\sqrt{2}} = 0.1414$ for $\text{Var}(y)$, which does not exceed the threshold for small effects [39]. Note that the effect size for $\text{Var}(y)$ is independent of any moment value, a consequence of y being normally distributed.

In total, there are 2 (sample sizes) $\times$7 (moment combinations) $\times$2 (choice of $\mathbf{I}_{ex}$) = 28 scenarios. The null hypothesis rejection rates are computed for three tests in each scenario, the Sargan-Hansen test for external sets using $\hat{\mathbf{S}}_h$ (abbreviated $\mathbf{SH}(\hat{\mathbf{S}}_h)$), the Sargan-Hansen test for external sets with $\hat{\boldsymbol{\Sigma}}_h$ (abbreviated $\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$), and the small sample test for external sets (**SES**). The significance level is $\alpha = 0.05$ for all scenarios. For $\mathbf{SH}(\hat{\mathbf{S}}_h)$ and **SES**, the test statistic $\underline{\chi^2}$ is computed using quadratic programming as described in Section 3.3, and for $\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$ it is computed using Algorithm 1.

For each simulation scenario the rejection rates were calculated twice with 100000 repetitions each to analyze the stability of the results. The type I error rate results are summarized in Table 1 and the test power results are summarized in Table 2.

5.2. The meta-analytical scenario

To implement the meta-analytical scenario, three sources of external information are sampled in each iteration of the simulation. The true expected values for the three sources are chosen to be $\mu_1 = 2$, $\mu_2 = 3$, and $\mu_3 = 4$. The sample sizes for the three sources are chosen to reflect the different estimation uncertainties present in applied fields, so that low ($n_1 = 20$), high ($n_2 = 1000$) and medium ($n_3 = 100$) sample sizes are chosen. For the Bayesian methods to be valid, the scenario given in Section 4 is used to generate the external samples, i.e., three i.i.d. samples are drawn based on $x_j \sim N(\mu_j, 1)$ for $j = 1, \dots, 3$. Then the external information is aggregated by calculating the mean and the variance for each source. After sampling the external information, a new data set is generated based on an i.i.d. sample of 50 values distributed like $N(\mu_0, 1)$, where $\mu_0 = 4$ (correctly specified scenario) or $\mu_0 = 4.5$ (misspecified scenario) is chosen. The value $\mu_0 = 4$ is the maximum of the external expected values and provides a harder test for correctness of type I errors than a value in the middle of the external expected values (as in Section 5.1). The value $\mu_0 = 4.5$ represents a medium effect size of $d = 0.5$. Then, asymptotic Sargan-Hansen tests for external sets based on $\hat{\mathbf{S}}_h$, $\hat{\boldsymbol{\Sigma}}_h$, and $\hat{\boldsymbol{\Omega}}_{h,e}$ are conducted.

To analyze the influence of different significance levels, the 13 values $0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99$ are used. It is also tested if the degree of prior-data conflict is greater than 0 and if there is a prior-data conflict with threshold a. To analyze the influence of a, the 13 values $0, -0.01, -0.10, -0.5, -1, -1.5, -2, -2.5, -3, -3.5, -4, -4.5, -5$ are used. For these 2 (correct or misspecified) times $(13 \times 3 + 13 + 1)$ (tests and criteria) = 106 scenarios, the null hypothesis / prior-data fit rejection rates were calculated based on 100000 simulation iterations each. The results are shown in Fig. 2.

Table 1
Type I error rates (correctly specified $\mathbf{I}_{ex}$).

Moments	$\mathbf{SH}(\hat{\mathbf{S}}_h)$	$\mathbf{SH}(\hat{\mathbf{\Sigma}}_h)$	**SES**
sample size $n = 30$			
$E(y)$	0.0106; 0.0110	0.0078; 0.0082	0.0081; 0.0086
$\mathrm{Var}(y)$	0.0762; 0.0764	0.0676; 0.0678	0.0690; 0.0693
$E(x)$	0.0159; 0.0155	0.0128; 0.0121	0.0132; 0.0126
$E(y), \mathrm{Var}(y)$	0.0590; 0.0574	0.0456; 0.0449	0.0488; 0.0475
$\mathrm{Var}(y), E(x)$	0.0602; 0.0593	0.0453; 0.0452	0.0485; 0.0482
$E(y), E(x)$	0.0101; 0.0099	0.0051; 0.0053	0.0063; 0.0061
$E(y), \mathrm{Var}(y), E(x)$	0.0508; 0.0492	0.0330; 0.0318	0.0371; 0.0357
sample size $n = 50$			
$E(y)$	0.0052; 0.0053	0.0044; 0.0044	0.0045; 0.0045
$\mathrm{Var}(y)$	0.0541; 0.0536	0.0501; 0.0492	0.0507; 0.0499
$E(x)$	0.0095; 0.0092	0.0080; 0.0078	0.0082; 0.0080
$E(y), \mathrm{Var}(y)$	0.0345; 0.0336	0.0287; 0.0279	0.0300; 0.0292
$\mathrm{Var}(y), E(x)$	0.0357; 0.0353	0.0294; 0.0286	0.0307; 0.0300
$E(y), E(x)$	0.0043; 0.0044	0.0029; 0.0028	0.0032; 0.0031
$E(y), \mathrm{Var}(y), E(x)$	0.0249; 0.0252	0.0185; 0.0185	0.0198; 0.0198

Note: The two numbers in each cell refer to the two runs of the simulations, where the left is from the first run and the right is from the second run.

Table 2
Power (misspecified $\mathbf{I}_{ex}$).

Moments	$\mathbf{SH}(\hat{\mathbf{S}}_h)$	$\mathbf{SH}(\hat{\mathbf{\Sigma}}_h)$	**SES**
sample size $n = 30$			
$E(y)$	0.7803; 0.7785	0.7495; 0.7474	0.7548; 0.7533
$\mathrm{Var}(y)$	0.2532; 0.2536	0.2353; 0.2351	0.2384; 0.2382
$E(x)$	0.5944; 0.5978	0.5549; 0.5588	0.5617; 0.5659
$E(y), \mathrm{Var}(y)$	0.7348; 0.7373	0.6623; 0.6649	0.6812; 0.6833
$\mathrm{Var}(y), E(x)$	0.6080; 0.6114	0.5289; 0.5311	0.5483; 0.5511
$E(y), E(x)$	0.8148; 0.8144	0.7443; 0.7440	0.7623; 0.7617
$E(y), \mathrm{Var}(y), E(x)$	0.8056; 0.8045	0.6959; 0.6936	0.7267; 0.7252
sample size $n = 50$			
$E(y)$	0.9405; 0.9410	0.9331; 0.9338	0.9345; 0.9350
$\mathrm{Var}(y)$	0.2755; 0.2753	0.2629; 0.2624	0.2650; 0.2645
$E(x)$	0.8060; 0.8078	0.7891; 0.7914	0.7918; 0.7943
$E(y), \mathrm{Var}(y)$	0.9091; 0.9090	0.8857; 0.8858	0.8914; 0.8914
$\mathrm{Var}(y), E(x)$	0.7936; 0.7934	0.7551; 0.7550	0.7638; 0.7639
$E(y), E(x)$	0.9614; 0.9607	0.9486; 0.9476	0.9516; 0.9508
$E(y), \mathrm{Var}(y), E(x)$	0.9491; 0.9494	0.9236; 0.9236	0.9305; 0.9304

Note: The two numbers in each cell refer to the two runs of the simulations, where the left is from the first run and the right is from the second run.

6. Discussion

6.1. Summary of the simulation studies

First, the results of the first simulation study are discussed, starting with the type I error rates, followed by the power of the tests. Except for the cases where $\mathrm{Var}(y)$ was used, all type I error rates were below the significance level. The use of $\mathrm{Var}(y)$ alone resulted in type I error rates of all tests above 0.05, but improving with sample size, indicating that in this case deviations from the normal distribution were not yet compensated by sample size. A possible compensation by increasing $\mathbf{I}_{ex}$ would result in lower power and violate the idea that $\mathbf{I}_{ex}$ is determined externally. Therefore, increasing the sample size seems to be the only viable solution, since in all scenarios increasing the sample size resulted in lower type I error rates and higher power. Using $\mathrm{Var}(y)$ in combination with other moments reduced type I error rates compared to using $\mathrm{Var}(y)$ alone. However, at $n = 30$ only the tests $\mathbf{SH}(\hat{\mathbf{\Sigma}}_h)$ and **SES** had correct type I error rates for moment combinations including $\mathrm{Var}(y)$, while at $n = 50$ all tests had correct type I error rates for moment combinations including $\mathrm{Var}(y)$. Using $E(y)$ alone resulted in the smallest type I error rates, as low as 0.0078 for $n = 30$ and 0.0044 for $n = 50$, illustrating that the tests are much more conservative than the significance level suggests. On the one hand, this can lead to a lower power, but on the other hand, it results in better small sample properties, since the chosen significance level is "reached more quickly" with respect to the sample size. In all scenarios, there was a clear order of tests in terms of type I error rate. The test $\mathbf{SH}(\hat{\mathbf{S}}_h)$ had higher type I error rates than **SES** and **SES** had higher rates than $\mathbf{SH}(\hat{\mathbf{\Sigma}}_h)$.

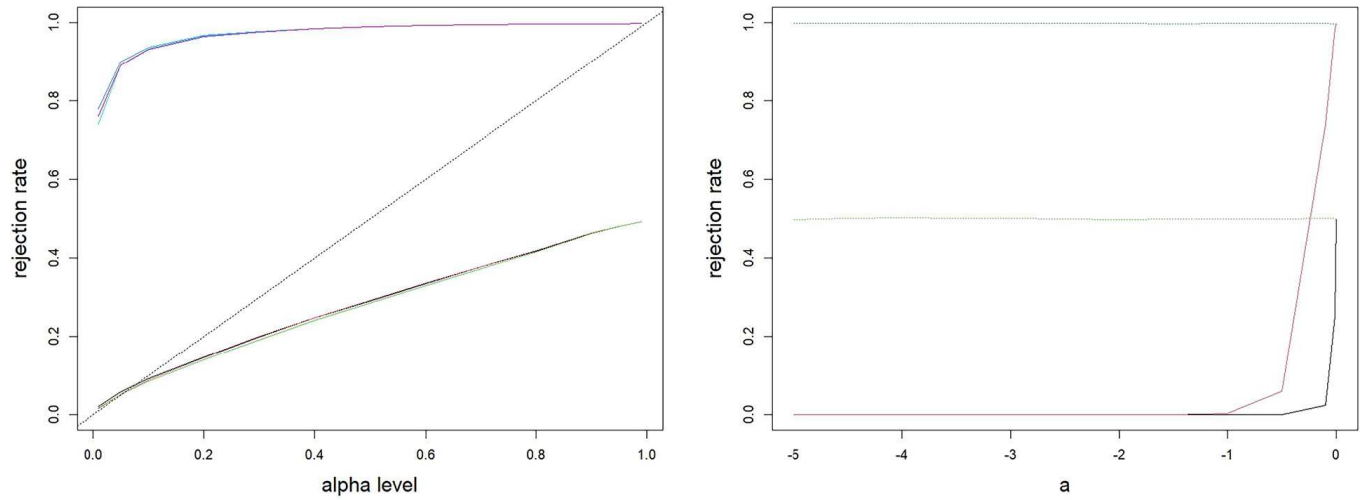


Fig. 2. Graphs of the rejection rates as a function of significance level α or threshold a. **Left Image:** Shows the rejection rates for the asymptotic Sargan-Hansen tests based on $\hat{\mathbf{S}}_h$, $\hat{\mathbf{\Sigma}}_h$, and $\hat{\mathbf{\Omega}}_{h,e}$ in the correctly specified case (the lower lines in black, red, and green) and in the misspecified case (the upper lines in blue, cyan and, magenta). The dotted line in the left image shows the identity function $y = x$ (rejection rates in the correctly specified case should be on or below the dotted line). **Right Image:** The solid lines show the proportion of prior-data conflicts with threshold a in the correct (black) and misspecified (red) cases. Also, the dotted lines show the proportion of degrees of prior-data conflict greater than 0 in the correct (green) and misspecified (blue) cases. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Regarding the power of the tests, their order was the same in all scenarios as in the type I error rate. The test **SH($\hat{\mathbf{S}}_h$)** had the highest power, followed by **SES** and **SH($\hat{\mathbf{\Sigma}}_h$)**. As derived from effect size considerations, the use of $E(y)$ alone resulted in the highest power, followed by $E(x)$ and $\text{Var}(y)$. The high power values for the moment $E(y)$ for $n = 30$ and $n = 50$ show that using external intervals instead of point values did not eliminate all the test power in the chosen scenarios. Even for the moment $E(x)$ with a small effect size, the power ranged from 0.7891 to 0.8078 for $n = 50$. Unfortunately, combinations of moments did not always result in higher power than single moments. For example, cases with $\text{Var}(y)$ resulted in lower power than the same cases without $\text{Var}(y)$, except for the combination of $\text{Var}(y)$ and $E(x)$ with **SH($\hat{\mathbf{S}}_h$)** and $n = 30$. This power reduction property can be explained by the very small effect size of $\text{Var}(y)$. The increase in the critical value due to a higher p_2 (degrees of freedom) may exceed the expected increase in the minimum test statistic by including $\text{Var}(y)$. However, combining $E(x)$ and $E(y)$ resulted in higher power than using either alone, except for $n = 30$ where **SH($\hat{\mathbf{\Sigma}}_h$)** is used. All the statements made so far are essentially true for both runs of the simulations.

Now, the results of the second simulation study are discussed. In both the correctly and the misspecified cases, the null hypothesis rejection rates of the three Sargan-Hansen tests were nearly identical, indicating that the estimation uncertainty did not play an important role here. Only for $\alpha = 0.01$ and $\alpha = 0.05$ were the type I error rates (correctly specified case) higher than 0.01 and 0.05, respectively. This indicates that the sample size was not large enough for the asymptotics to hold in this extreme case, where α is small and the true expected value of the new data is the maximum of the external expected values. Consistent with the first simulation study, the power was high even though an external interval is used. Furthermore, the power quickly went to one when α was increased. Regarding the Bayesian criteria, the use of the degree of prior-data conflict was, as expected, very liberal, with a rejection rate of about 0.5 in the correctly specified case. Thus, it was not ideal for confirming the fit between external information and data, but since it had a power of about 0.998, it was a viable detector of the slightest conflict between external information and data. However, the three Sargan-Hansen tests also had a power of about 0.998 when $\alpha = 0.99$ was used, while the type I error rate was about 0.49. The results for the prior-data conflict with threshold a are interesting. First, the power and type I error rates were above 0 only for a values between 0 and -1. However, in this range, the power and type I error rates were similar to the power and type I error rates of the Sargan-Hansen tests. For example, for $a = -0.01$, the type I error rate was 0.2514 with a power of 0.9843, while the Sargan-Hansen test based on $\hat{\mathbf{S}}_h$ with $\alpha = 0.4$ had a type I error rate of 0.2471 and a power of 0.9854. Another example is $a = -0.1$, which had a type I error rate of 0.0233 and a power of 0.7405, comparable to the Sargan-Hansen test based on $\hat{\mathbf{S}}_h$ with $\alpha = 0.01$, which had a type I error rate of 0.0202 and a power of 0.7798.

Taken together, the proposed methods can generally provide good power for low type I error rates, even for small sample sizes such as $n = 30$. However, if one wants to control for the type I error rate by α, as in the classical frequentist approach, there are a few things to consider. First, one should avoid moments with highly skewed distributions, such as $\text{Var}(y)$, even in combination with other moments. Second, one should prefer using $\hat{\mathbf{\Sigma}}_h$ or $\hat{\mathbf{\Omega}}_{h,e}$ to using $\hat{\mathbf{S}}_h$, since the latter resulted in the highest type I error rates, sometimes exceeding α, while the former did not. Third, it is better not to use too small a α when the sample size is small and there are no prior assumptions about the location of the true value in the external interval. In general, a good advice for practitioners is to run simulations of their own scenario to analyze whether the significance level is exceeded. Regarding the threshold a for the prior-data conflict, it is difficult to give an interpretation, since the case here deviates from the interpretation based on Bayes factors as given in Bickel [16]. Despite the use of external intervals, the Sargan-Hansen tests (and their small sample version) had good power for small sample sizes, even for the small and medium effect sizes used in the simulation studies.

6.2. Outlook

For the scenarios tested so far, more research is needed on the choice and interpretation of the threshold a. It is unclear whether the range of usable values depends on the scenario or whether values slightly below 0 can be recommended in general. It would also be helpful to analyze the relationship between a and the type I error rate and power, so that the choice of a can be guided by the power or type I error rate one wishes to achieve.

The scenarios tested in the two simulation studies are quite simple, so a natural question is how the proposed methods behave in more complex scenarios. While this is straightforward for the Sargan Hansen test, the extension of the Bayesian criteria seems to be more analytically challenging. The biggest problem is that the Sargan-Hansen test and the Bayesian criteria generally address slightly different questions. While the Sargan-Hansen test tests the fit of external information to the data for certain moment functions, the Bayesian criteria are designed to detect a conflict of the data with an entire (prior) distribution. Therefore, it is possible that the Bayesian criteria reject prior distributions for which the values of certain moment functions would fit the data (would be accepted by the Sargan-Hansen test). Another important issue is the robustness of the methods to small deviations from the normal distribution. For moments that are highly skewed in small samples, such as $\text{Var}(y)$, it may be helpful to derive the exact distribution of the test statistic under the assumption that y is normally distributed.

Although in most cases the model parameters cancel out in the Sargan-Hansen test, there are ways to use external information about the model parameters. For example, the OLS estimator can be interpreted as a function based only on the data. This function can be used as a moment function that incorporates external knowledge about parameter values. Such "indirect" model moment functions would be interesting to study. In addition, there are other tests or frameworks for using external moment-type information that could be investigated with respect to using external information about the parameter. An example is the Empirical Likelihood framework [40].

Another possible extension is externally informed tests for parameters. These can be realized by using the Cartesian product of the external set and the possible parameter values under the null hypothesis, since a Wald test would still have the same asymptotic credal set compared to the case where only the external set is used. Assuming the external information is correct, the joint null hypothesis reduces to the null hypothesis regarding only the parameter. An advantage could be a reduced variance of the parameter estimator (as shown by Imbens and Lancaster [6]) and thus higher power compared to the test without external information. Finally, it is crucial to highlight that the proofs of Proposition 1 and Theorem 2 of Section 3.1 primarily depend on the existence of a minimum and a maximum cumulative distribution function. These results can be directly extended from stochastically ordered credal sets to (σ-additive) probability boxes (p-boxes), which have lower and upper cumulative distribution functions by definition. This broadens the scope of potential applications.

CRediT authorship contribution statement

Martin Jann: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared my R script in the attach file step as supplementary material.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. This paper is an extended version of a paper presented at ISIPTA 2023 [41]. The discussions during the conference were helpful in thinking about Bayesian criteria and decision theory. The author would like to thank the guest editors Ignacio Montes, Enrique Miranda and Barbara Vantaggi for inviting him to contribute to this special issue, and the IJAR reviewers for their helpful comments and suggestions.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijar.2024.109214.

References

[1] T. Augustin, G. Walter, F.P.A. Coolen, Statistical inference, in: Introduction to Imprecise Probabilities, John Wiley & Sons, Ltd, 2014, pp. 135–189.
[2] M. Spiess, P. Jordan, In models we trust: preregistration, large samples, and replication may not suffice, Front. Psychol. 14 (2023), https://doi.org/10.3389/fpsyg.2023.1266447.

[3] J.M. Bernardo, A.F.M. Smith, Bayesian Theory, John Wiley & Sons, Inc., 1994.
[4] P.S. Knopov, A.S. Korkhin, Regression Analysis Under A Priori Parameter Restrictions, Springer Optimization and Its Applications, vol. 54, Springer Science & Business Media, New York, 2011.
[5] A. Cameron, P. Trivedi, Microeconometrics: Methods and Applications, Cambridge University Press, 2005.
[6] G.W. Imbens, T. Lancaster, Combining micro and macro data in microeconometric models, Rev. Econ. Stud. 61 (1994) 655–680, https://doi.org/10.2307/2297913.
[7] L.P. Hansen, Large sample properties of generalized method of moments estimators, Econometrica 50 (1982) 1029–1054, https://doi.org/10.2307/1912775.
[8] J.D. Sargan, The estimation of economic relationships using instrumental variables, Econometrica 26 (1958) 393–415, https://doi.org/10.2307/1907619.
[9] P.M.D.C. Parente, J.M.C. Santos Silva, A cautionary note on tests of overidentifying restrictions, Econ. Lett. 115 (2012) 314–317, https://doi.org/10.1016/j.econlet.2011.12.047.
[10] J.F. Kiviet, S. Kripfganz, Instrument approval by the Sargan test and its consequences for coefficient estimation, Econ. Lett. 205 (2021), https://doi.org/10.1016/j.econlet.2021.109935.
[11] M. Evans, H. Moshonov, Checking for prior-data conflict, Bayesian Anal. 1 (2006) 893–914, https://doi.org/10.1214/06-BA129.
[12] N. Bousquet, Diagnostics of prior-data agreement in applied Bayesian analysis, J. Appl. Stat. 35 (2008) 1011–1029, https://doi.org/10.1080/02664760802192981.
[13] H. Rahimian, S. Mehrotra, Frameworks and results in distributionally robust optimization, Open J. Math. Optim. 3 (2022) 1–85, https://doi.org/10.5802/ojmo.15.
[14] G. Walter, T. Augustin, Imprecision and prior-data conflict in generalized Bayesian inference, J. Stat. Theory Pract. 3 (2009) 255–271, https://doi.org/10.1080/15598608.2009.10411924.
[15] P. Walley, Statistical Reasoning with Imprecise Probabilities, Springer, 1991.
[16] D.R. Bickel, Inference after checking multiple Bayesian models for data conflict and applications to mitigating the influence of rejected priors, Int. J. Approx. Reason. 66 (2015) 53–72, https://doi.org/10.1016/j.ijar.2015.07.012.
[17] W.K. Newey, D. McFadden, Large sample estimation and hypothesis testing, in: Handbook of Econometrics, vol. 4, Elsevier, 1994, pp. 2111–2245.
[18] S.L. Zeger, K.-Y. Liang, Longitudinal data analysis for discrete and continuous outcomes, Biometrics 42 (1986) 121–130, https://doi.org/10.2307/2531248.
[19] P. Huber, E. Ronchetti, Robust Statistics, 2 ed., Wiley Series in Probability and Statistics, Wiley, 2009.
[20] Z. Xiao, Efficient GMM estimation with singular system of moment conditions, Stat. Theory Relat. Fields 4 (2020) 172–178, https://doi.org/10.1080/24754269.2019.1653159.
[21] S. Puntanen, G.P.H. Styan, J. Isotalo, Matrix Tricks for Linear Statistical Models, Springer Berlin Heidelberg, 2011.
[22] S.C. Ahu, P. Schmidt, A separability result for GMM estimation, with applications to GLS prediction and conditional moment tests, Econom. Rev. 14 (1995) 19–34, https://doi.org/10.1080/07474939508800301.
[23] D.W.K. Andrews, Asymptotic results for generalized Wald tests, Econom. Theory 3 (1987) 348–358, https://doi.org/10.1017/S0266466600010434.
[24] A.S. Hadi, M.T. Wells, A note on generalized Wald's method, Metrika 37 (1990) 309–315, https://doi.org/10.1007/BF02613538.
[25] S. Patole, Principles and Practice of Systematic Reviews and Meta-Analysis, vol. 1, Springer, Cham, 2021.
[26] J.T. Cleophas, H.A. Zwinderman, Modern Meta-Analysis: Review and Update of Methodologies, vol. 1, Springer, 2017.
[27] N. Huntley, R. Hable, M.C.M. Troffaes, Decision making, in: Introduction to Imprecise Probabilities, John Wiley & Sons, Ltd, 2014, pp. 190–206.
[28] D. Rasch, K.D. Kubinger, T. Yanagida, Statistics in Psychology Using R and SPSS, John Wiley & Sons, 2011.
[29] B.K. Ghosh, Some monotonicity theorems for chi square, F and t distributions with applications, J. R. Stat. Soc., Ser. B, Methodol. 35 (1973) 480–492, https://doi.org/10.1111/j.2517-6161.1973.tb00976.x.
[30] P.C.B. Phillips, The exact distribution of the Wald statistic, Econometrica 54 (1986) 881–895, https://doi.org/10.2307/1912841.
[31] M. Jann, M. Spiess, Using external information for more precise inferences in general regression models, Psychometrika (2024) 1–22, https://doi.org/10.1007/s11336-024-09953-w.
[32] C.D. Meyer Jr., Generalized inversion of modified matrices, SIAM J. Appl. Math. 24 (1973) 315–323, https://doi.org/10.1137/0124033.
[33] J.E. Smith, R.L. Winkler, The optimizer's curse: skepticism and postdecision surprise in decision analysis, Manag. Sci. 52 (2006) 311–322, https://doi.org/10.1287/mnsc.1050.0451.
[34] H. Van Hasselt, Estimating the maximum expected value: an analysis of (nested) cross validation and the maximum sample average, arXiv preprint, arXiv:1302.7175, 2013.
[35] C. D'Eramo, M. Restelli, A. Nuara, Estimating maximum expected value through Gaussian approximation, in: M.F. Balcan, K.Q. Weinberger (Eds.), Proceedings of the 33rd International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 48, PMLR, New York, New York, USA, 2016, pp. 1032–1040, https://proceedings.mlr.press/v48/deramo16.html.
[36] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2023, https://www.R-project.org/.
[37] B.A. Turlach, A. Weingessel, C. Moler, quadprog: Functions to Solve Quadratic Programming Problems, https://CRAN.R-project.org/package=quadprog, 2019, r package version 1.5-8.
[38] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, fourth ed., Springer, New York, ISBN 0-387-95457-0, 2002, https://www.stats.ox.ac.uk/pub/MASS4/.
[39] J. Cohen, A power primer, Psychol. Bull. 112 (1992) 155–159, https://doi.org/10.1037/0033-2909.112.1.155.
[40] A.B. Owen, Empirical likelihood ratio confidence intervals for a single functional, Biometrika 75 (1988) 237–249, https://doi.org/10.2307/2336172.
[41] M. Jann, Testing the coherence of data and external intervals via an imprecise Sargan-Hansen test, in: E. Miranda, I. Montes, E. Quaeghebeur, B. Vantaggi (Eds.), Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications, in: Proceedings of Machine Learning Research, vol. 215, PMLR, 2023, pp. 249–258, https://proceedings.mlr.press/v215/jann23a.html.

Appendix D

Paper 4

Jann, M., & Spiess, M. (2025). *Testing linear hypotheses in repeated measures generalized linear models using external information* [Manuscript under review]

The following page is a printout from the website https://mc.manuscriptcentral.com/psychometrika confirming that the fourth paper has been submitted to the journal *Psychometrika* and is currently under review.

Psychometrika
CAMBRIDGE UNIVERSITY PRESS

Submitted Manuscripts

STATUS	ID	TITLE	CREATED	SUBMITTED
Contact Journal ADM: Sage, Alison Awaiting Reviewer Scores	PSY-2025-0083	Testing linear hypotheses in repeated measures generalized linear models using external information View Submission Cover Letter	17-Apr-2025	17-Apr-2025

Testing linear hypotheses in repeated measures generalized linear models using external information

Abstract

In this paper we propose three tests for general linear hypotheses in generalized linear models with repeated measures using external moment information. To this end we embedded the approach of generalized estimating equations in the generalized method of moment framework. The developed method is capable of reflecting both uncertainty of external information due to estimation by including external variance estimates and uncertainty due to different designs, populations or procedures by using a set of external values. For block invariant designs we provide analytic expressions for estimators and some test statistics. Further, for these designs the dependence structure cancels out, so that our results are valid for every possible of these structures without the need to model it. The small sample validity and power of the three tests are investigated under a variety of conditions in two simulation studies based on real data scenarios. All three tests show nominal type I error rates when correct external information is used and in some cases the use of external information increases the power, even if uncertainties are reflected properly. Despite a gain in power, the use of external intervals may increase the robustness of validity.

Keywords: external information; general linear hypotheses; generalized estimating equations; generalized method of moments; imprecise probabilities

Introduction

The goal of this paper is to derive statistical tests for general linear hypotheses for parameters of externally informed models developed by Jann and Spiess (2024) using the Generalized Method of Moments (GMM) (Hansen, 1982). These models allow a researcher to incorporate knowledge of statistical moments into statistical analysis. External moment information often comes in the form of means or (co-)variances from other studies. The advantages of using external information are an increase in robustness (if the external information is properly represented) and possible efficiency gains in estimators and test statistics. It is therefore of interest to exploit this effect to increase power in parameter testing.

There is some literature on testing (general) linear hypotheses in a GMM framework, offering a variety of aspects to be represented in hypothesis testing. Starting with the seminal work of Newey and West (1987) comparing test statistics for the efficient GMM, there are extensions to the case where parameters are unidentified or weakly identified (Andrews & Guggenberger, 2017; Dufour, 2003) as well as robust extensions of tests (Ronchetti & Trojani, 2001). There are bootstrapping approaches (Brown & Newey, 2002) and tests that are invariant under hypothesis reformulations and reparameterizations (Dufour et al., 2017). Furthermore, the performance of tests for (general) linear hypotheses has been compared for a number of different data scenarios and designs, ranging from autoregressive panel data (Bond et al., 2001) to cluster-correlated data (Rotnitzky & Jewell, 1990) to linear panel data (Bond & Windmeijer, 2005).

In longitudinal designs, the use of repeated measures can increase estimation efficiency compared to cross-sectional studies (Diggle et al., 2002, pp.

24–25). In addition, repeated measures are common in clinical and general psychology, for example, suggesting their relevance in applied psychological research. The use of externally informed models for longitudinal data offers an opportunity for synergy in terms of efficiency gains. The focus will be on population means and thus marginal models, since they allow estimation via generalized estimating equations (GEE) (Liang & Zeger, 1986), which can be incorporated directly into GMM (Cameron & Trivedi, 2005, p. 790).

When using external information, one must represent its uncertainties in order to arrive at a valid statistical test. To properly represent the uncertainties, one must distinguish between uncertainty due to estimation and "qualitative uncertainty" due to different designs, sampling mechanisms, populations, and other aspects of the external data compared to the new data (Jann, 2024). While the former type of uncertainty can be represented by using variances of the estimates (the typical case in statistical analysis), the latter can be represented by using a range of possible values for the external estimates, for example in the form of external intervals, as was done by Jann and Spiess (2024). By traversing the external interval, one can construct intervals of possible estimates and variances or unions of confidence intervals. As demonstrated by Jann and Spiess (2024), this increases distributional robustness, leading to valid inference even when, contrary to the assumed normality, the data are not normally distributed.

However, the extension of these results to significance testing has only been done for the special case of the Sargan-Hansen test for overidentifying restrictions (Jann, 2023, 2024). For testing general (linear) hypotheses, there are three (asymptotically equal) types of tests typically used in the GMM framework, the Wald test, the Lagrange multiplier test, and criterion-based

tests (Bond & Windmeijer, 2005). These tests are presented in Section 2. Section 3 introduces an extension of these tests to the case where external sets are used. In Section 4, the tests with external sets are applied to generalized linear models with repeated measures, discussing designs where their computation is feasible. Section 5 compares the tests in terms of type I error rate and power in small samples for different scenarios through simulation studies. To demonstrate the applicability to real data, the tests are applied to two real psychological datasets.

Testing general linear hypotheses

To emphasize the application to regression analysis, the parameter of interest is denoted by $\boldsymbol{\beta}$ throughout the paper.

Definition 1. Let $\boldsymbol{\beta} \in \boldsymbol{\Theta} \subset \mathbb{R}^q$ be a parameter and $\boldsymbol{\beta}_0$ its true value. Let $\mathbf{R}$ be a (constant) $k \times q$ real-valued matrix, representing the linear constraints and $\mathbf{r}$ be a k-dimensional vector of known real constants. If $\text{rank}(\mathbf{R}) = k \leq q$, a **general linear hypothesis** is defined by the pair

$$H_0 : \mathbf{R}\boldsymbol{\beta}_0 = \mathbf{r} \qquad \text{and} \qquad H_1 : \mathbf{R}\boldsymbol{\beta}_0 \neq \mathbf{r}.$$

Important examples of a general linear hypothesis are main and interaction effects in ANOVAs. In effect coding, these hypotheses are of the form $\beta_1 = ... = \beta_k = 0$ and are thus equivalent to general linear hypotheses where $\mathbf{R}$ has the $(k \times k)$ identity matrix as a block at the corresponding position and all other elements are 0 while $\mathbf{r} = \mathbf{0}$. Other important examples are tests for single entries of the parameter where $\mathbf{R}$ has entry one at the corresponding position and zero entries elsewhere, and tests for differences of

parameter values where the only non-zero entries of $\mathbf{R}$ are 1 and -1 at the corresponding positions.

To fix the notation, the presentation from Bond and Windmeijer (2005) is adopted. Let $\mathbf{x}_i$ be a vector of observed variables for the i-th unit for $i = 1, \ldots, n$ and $\mathbf{g}(\mathbf{x}, \boldsymbol{\beta})$ be a function that assigns values in $\mathbb{R}^q$ to each $\mathbf{x}_i$ and $\boldsymbol{\beta} \in \boldsymbol{\Theta}$. The function $\mathbf{g}$ represents estimating equations used to estimate the true parameter value $\boldsymbol{\beta}_0$. Further define $\bar{\mathbf{g}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta})$ and $\mathbf{W}_n(\boldsymbol{\beta}) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta})\mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta})^T\right)^{-1}$. The function $\mathbf{W}_n$ represents a weighting matrix that balances the influences of the entries of the estimating equations. The choice of $\mathbf{W}_n$ is made to find the efficient GMM estimator (with respect to all possible weighting matrices) (Hansen, 1982). GMM estimation is typically carried out in two steps. First, minimize the quadratic form $\bar{\mathbf{g}}(\boldsymbol{\beta})^T \mathbf{W} \bar{\mathbf{g}}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ for some constant $(q \times q)$ matrix $\mathbf{W}$, for example $\mathbf{W} = \mathbf{I}$ the identity matrix. The resulting minimum value is denoted by $\hat{\boldsymbol{\beta}}_1$ and is called a one-step GMM estimator. Second, minimize the quadratic form $\bar{\mathbf{g}}(\boldsymbol{\beta})^T \mathbf{W}_n(\hat{\boldsymbol{\beta}}_1) \bar{\mathbf{g}}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Analogously the result is denoted by $\hat{\boldsymbol{\beta}}_2$ and is called a two-step GMM estimator. The corresponding variance estimator is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_2) = \frac{1}{n}\big((\nabla_{\boldsymbol{\beta}} \bar{\mathbf{g}}(\boldsymbol{\beta})|_{\hat{\boldsymbol{\beta}}_2})^T \mathbf{W}_n(\hat{\boldsymbol{\beta}}_2)(\nabla_{\boldsymbol{\beta}} \bar{\mathbf{g}}(\boldsymbol{\beta})|_{\hat{\boldsymbol{\beta}}_2})\big),$$

where $\nabla_{\boldsymbol{\beta}} \bar{\mathbf{g}}(\boldsymbol{\beta})|_{\hat{\boldsymbol{\beta}}_2}$ denotes the gradient operator with respect to $\boldsymbol{\beta}$ evaluated at $\hat{\boldsymbol{\beta}}_2$ (Cameron & Trivedi, 2005, p. 176).

Now, general linear hypotheses pose a significance testing problem that we want to address in the GMM framework in order to incorporate external moment information. The discussion of tests is based on the work of Bond and

Windmeijer (2005). All of the tests presented below have the same asymptotic distributions under mild regularity conditions. Under the null hypothesis, they have an asymptotic χ^2_k-distribution with k degrees of freedom (Bond & Windmeijer, 2005). Under local alternative hypotheses and using the efficient GMM estimator, they asymptotically follow a non-central $\chi^2_k(\lambda)$-distribution with k degrees of freedom and the non-centrality parameter λ (Cameron & Trivedi, 2005, p. 245). Thus, only the test statistic is left to specify the tests. The standard Wald test is given by the test statistic

$$T_W = (\mathbf{R}\hat{\boldsymbol{\beta}}_2 - r)^T(\mathbf{R}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_2)\mathbf{R}^T)^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_2 - r).$$

The Lagrange multiplier test uses the same test statistic, but with a different estimator. Let $\hat{\boldsymbol{\beta}}_r$ denote a two-step GMM estimator in which the weighting matrix $\mathbf{W}_n$ is computed in the second step with a one-step GMM estimator $\tilde{\boldsymbol{\beta}}_1$ that is restricted under the null hypothesis, i.e. where the first step quadratic form optimization is subject to the linear constraints $\mathbf{R}\boldsymbol{\beta} = r$. The resulting test statistic is

$$LM = (\mathbf{R}\hat{\boldsymbol{\beta}}_r - r)^T(\mathbf{R}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_r)\mathbf{R}^T)^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_r - r).$$

The criterion-based test uses the difference between the objective function evaluated at the two-step estimator constrained under the null hypothesis and at the unrestricted two-step estimator. Let $\tilde{\boldsymbol{\beta}}_2$ be the two-step GMM estimator obtained under the constraint $\mathbf{R}\boldsymbol{\beta} = r$, then the test statistic is

$$D_{RU} = n\big(\bar{\mathbf{g}}(\tilde{\boldsymbol{\beta}}_2)^T\mathbf{W}_n(\tilde{\boldsymbol{\beta}}_2)\bar{\mathbf{g}}(\tilde{\boldsymbol{\beta}}_2) - \bar{\mathbf{g}}(\hat{\boldsymbol{\beta}}_2)^T\mathbf{W}_n(\hat{\boldsymbol{\beta}}_2)\bar{\mathbf{g}}(\hat{\boldsymbol{\beta}}_2)\big).$$

Note that our definitions are slightly different from the one in Bond and Windmeijer (2005), since the weighting matrices are evaluated at the two-step estimators and not at the one-step estimators. However, the resulting weighting matrices have the same limits as those using one-step estimators, so the asymptotic results remain unaffected.

Regarding small sample performance for linear panel data models, the Langrange multiplier test and the criterion-based test have been found to be more reliable than a standard Wald test (Bond & Windmeijer, 2005). In addition, the standard Wald test appears to have poor small sample performance for within-subject designs (Spiess et al., 2019). We restrict our attention to these three tests because results for local alternative hypotheses are available for them, and the tests involving external sets developed in the next section depend on the results for local alternative hypotheses.

Hypothesis testing with external sets

The foundations of hypothesis testing when external moment information is known were developed by Jann (2024) to test whether data and external information fit. In this section, we extend the results to testing general linear hypotheses. The presentation here follows the arguments in Jann and Spiess (2024) and Jann (2024).

The basic idea of incorporating external moment information into GMM estimation by formulating additional moment conditions was given by Hellerstein and Imbens (1999) and Imbens and Lancaster (1994). For example, to include the information that the expected value of a random variable y is 3, one can use $\bar{\mathbf{h}}(y) = \bar{y} - 3$. Then one specifies $\bar{\mathbf{g}} = (\bar{\mathbf{m}}^T, \bar{\mathbf{h}}^T)^T$, where $\bar{\mathbf{m}}$ represents the estimating equations for the model parameters. Now, the three

tests presented in Section 2 can be applied if the GMM regularity conditions hold. However, the most critical regularity condition is $E(\bar{\mathbf{h}}) = 0$, i.e., the moment information must be correct. Since this is unlikely to be true, some modifications are needed to make this approach applicable. First, it is clear that the external information is an estimate. Suppose, as in the given example, that $\bar{\mathbf{h}}$ takes the form of a difference of some function of the data and the external value. In this case uncertainty due to estimation can be reflected by adding a corresponding variance (matrix) estimate to the lower right block of the inverse of the weighting matrix $\mathbf{W}_n$, using the independence of the external and the new data (Jann, 2024). More precisely, let $\mathbf{V}_{ex}$ be a variance (matrix) estimate for the external moments obtained from external sources, then the uncertainty due to estimation is reflected by using

$$\tilde{\mathbf{W}}_n = \left[\mathbf{W}_n^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{ex} \end{pmatrix} \right]^{-1}$$

instead of $\mathbf{W}_n$. See Jann (2024) for a derivation of this expression. This approach relaxes the assumption of knowing the true value to knowing an unbiased estimate of the moment. Second, there are often multiple sources of external information and thus a set $\mathbf{M}_{ex}$ of possible estimates. Due to differences in design, sampling or population, one may even doubt the assumption of unbiasedness of the estimates. This "qualitative" uncertainty is more difficult to quantify. The most conservative approach is to not aggregate the elements of $\mathbf{M}_{ex}$ and be agnostic about which element represents the true expected value. Even further, one could extend $\mathbf{M}_{ex}$ to a (multidimensional) interval by taking element-wise infima and suprema (element by element) of $\mathbf{M}_{ex}$, as was done by Jann and Spiess (2024). Such an interval extension of $\mathbf{M}_{ex}$

is denoted by $\mathbf{I}_{ex}$. As a result, the assumption is weakened to knowing bounds on the true expected value of the moment.

To reflect both uncertainties, the variance matrices of all estimates should be considered. Since computing the test statistic is a non-trivial optimization problem, this can be very time consuming, especially if there are many variance (matrix) estimates. A better solution would be to select a variance matrix with the lowest Löwner (partial) order (if it exists), separating the selection from the computation of the test statistic (see (Puntanen et al., 2011, p. 12) for a definition and technical details regarding the Löwner order). The rationale is to choose the smallest variance (matrix) estimate to compensate for the length of the external interval. Even if not all variance matrices are comparable, this provides a way to exclude at least some, namely all variance matrices with a Löwner order at least as high as one of the other variance matrices. While the implementation of uncertainty due to estimation is straightforward and does not change the tests, the same cannot be said for qualitative uncertainty since the external information is set-valued. Therefore, we will discuss its implementation below.

General hypothesis tests based on credal sets

Assume a (not necessarily linear) null hypothesis $H_0 : \boldsymbol{\beta}_0 \in \boldsymbol{\Theta}_0$ is given, and that a hypothesis test for the null hypothesis is known. Based on the assumption that $\mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$) contains the true value, the test statistics have a variety of possible distributions when calculated for different elements of $\mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$). For the true value it will be a central distribution, all other values it will be a non-central distribution. However, it is not known which value is the true one, so for each value only a set of possible distributions for the test

statistic can be determined. Such sets of probability measures are called credal sets. A general introduction to credal sets is given by Augustin et al. (2014, p. 19). In our case, a credal set is defined as a family of probability measures $\mathcal{M} = \{P_\theta \,|\, \theta \in \Theta\}$ (on the same measure space) indexed by a parameter θ, consisting of the possible distributions for the random variable of interest, here the test statistic. Let $\underline{P}$ be a set-valued function defined by assigning to each measurable set A the infimum probability with respect to $\mathcal{M}$, i.e. $\underline{P}(A) = \inf_{P \in \mathcal{M}} P(A)$. The function $\underline{P}$ is called **lower probability (based on** $\mathcal{M}$**)**. Based on credal sets and lower probability, there is the following concept for a hypothesis test:

Definition 2. (Jann (2024)) Let $T(\boldsymbol{\beta})$ be a test statistic that is a function of a parameter $\boldsymbol{\beta}$. Let $\mathcal{T}$ be a set of observed test statistics, where $\underline{t}$ denotes its infimum. Let $\mathcal{M}$ be a credal set of possible distributions of the test statistics and $\underline{P}$ be the lower probability based on $\mathcal{M}$. Under the null hypothesis $H_0 : \boldsymbol{\beta}_0 \in \boldsymbol{\Theta}_0$, a $\Gamma-$**maximin test with significance level** $\alpha \in (0, 1)$ is as follows:

$$\text{If } \underline{P}(T > \underline{t}) < \alpha, \text{ then reject } H_0 : \boldsymbol{\beta}_0 \in \boldsymbol{\Theta}_0, \text{else maintain } H_0.$$

A $\Gamma-$maximin test is based on the simple idea that the null hypothesis is rejected if it is rejected for every possible test statistic, where the lower probability of the credal set provides the p-value. By Theorem 2 of Jann (2024), a $\Gamma-$maximin test with significance level α "contains" a valid hypothesis test of level α, if the credal set $\mathcal{M}$ is stochastically ordered and its minimum is exactly the distribution of $T(\boldsymbol{\beta}_0)$. In other words, the $\Gamma-$maximin test will have a significance level of α or (usually) lower if the lower probability is equal to the

true distribution under the null hypothesis.

The most prominent examples of such credal sets are the families of non-central χ^2- and F-distributions. They are stochastically ordered in the non-centrality parameter (Ghosh, 1973), so their central versions (with the non-centrality parameter zero) are the lower probability. Under the null hypothesis, many test statistics are distributed like the central distribution in these families. Note that it is still important to show that test statistics follow the non-central distributions when the null hypothesis is violated. The reason is that the scale may also be affected. Thus not only non-central but also scaled distribution families have to be considered, which may not be stochastically ordered. However, all three tests of Section 2 satisfy this condition, since their (asymptotic) credal sets consist of the non-central χ^2-distributions for local alternatives and the degenerate distribution 1_∞ for fixed alternatives, which sets all masses to infinity and is therefore the maximum with respect to stochastic order.

Tests for general linear hypotheses using external sets

We will now extend the tests of Section 2 to the scenario where external sets are present. This is done by simply combining $\bar{\mathbf{g}}$ with the external moment function $\bar{\mathbf{h}}$, i.e., define $\bar{\mathbf{g}}_{ex} = (\bar{\mathbf{g}}^T, \bar{\mathbf{h}}^T)^T$. Let $\mathbf{e}_0$ be the true value of the considered moments in $\bar{\mathbf{h}}$ for the new dataset. At $\mathbf{e}_0$, the regularity conditions are satisfied, and the derivations and resulting test statistics are the same as in Section 2, only based on $\bar{\mathbf{g}}_{ex}$ instead of $\bar{\mathbf{g}}$. The set of possible test statistics $\mathcal{T}$ is given by computing the test statistic for each value in $\mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$). The credal set is the non-central χ^2-family, where the lower probability is reached by the central χ^2-distribution. The derived tests are summarized below.

Definition 3. Let $T_w(\mathbf{e}), LM(\mathbf{e})$ and $D_{RU}(\mathbf{e})$ be the test statistics for the standard Wald test, the Lagrange multiplier test and the criterion-based test in Section 2, based on $\bar{\mathbf{g}}_{ex}$ and $\mathbf{e} \in \mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$). The **externally informed Wald test, Lagrange multiplier test and criterion-based test** are defined by the test statistics

$$\underline{T_w} = \inf_{\mathbf{e}\in\mathbf{M}_{ex}(\mathbf{I}_{ex})} T_w(\mathbf{e}), \underline{LM} = \inf_{\mathbf{e}\in\mathbf{M}_{ex}(\mathbf{I}_{ex})} LM(\mathbf{e}) \text{ and } \underline{D_{RU}} = \inf_{\mathbf{e}\in\mathbf{M}_{ex}(\mathbf{I}_{ex})} D_{RU}(\mathbf{e})$$

respectively, while using critical values from a central χ^2-distribution.

Since the tests are constructed based on the minimal test statistic, they are valid if any element of $\mathbf{M}_{ex}$ (or $\mathbf{I}_{ex}$) represents the true moment value. Regarding robustness there are two different aspects to consider, robustness of validity and robustness of efficiency (Huber & Ronchetti, 2009). While the former means that the level of the test is not exceeded when the true distribution under the null hypothesis differs slightly from the assumed distribution, the latter means that a test has acceptable power under small deviations from (distributions under) specified alternatives. In general, the proposed tests can increase the robustness of validity because they use a (credal) set of possible distributions and a set of test statistics. The tests are performed using the minimum of all test statistics, which (as a random variable) is not expected to be greater than the test statistic at the true external value, given the null hypothesis and a correctly specified external variance. Applying the χ^2-distribution to the test statistic based on the true external value would asymptotically lead to a nominal type I error, so applying it to the minimum test statistic can only reduce the actual type I error or leave it unchanged. Thus, the inference will be valid for all distributions with

nominal type I errors at the minimal test statistic, for example distributions with (slightly) heavier tails.

The robustness of the power of the tests can be discussed on the basis of the initial "point" test. If the latter has more power, then it is reasonable to assume that the resulting interval tests may inherit it. The use of external moments can (strongly) reduce the variance of the estimator, as shown in Jann and Spiess (2024). However, the small sample performance may differ, and larger external sets are expected to lead to lower power (compared to smaller sets). We will explore these issues in the simulation studies in Section 5.

Generalized linear models with repeated measures

We will use these results in the context of marginal generalized linear models for longitudinal data, following the presentation of Diggle et al. (2002). The general idea is to specify the expectation of the dependent variable by a link function and to define additional nuisance parameters to model the correlation structure of the multiple measurements of the dependent variable on the same subject or unit. Count data can be modeled by the log link and binary data by the logit link. Logit models for binary data can be directly extended to the multinomial case. For example, cumulative logit models can be used for categorical ordinal dependent variables (Agresti, 2002). An important example of a correlation structure assumes equal correlations between variables, called the uniform correlation model or equicorrelation. In the case of a linear model, this uniform correlation model corresponds to a random intercept model (Diggle et al., 2002, p.54), which is widely used in applied psychological research. Another example is an exponential correlation model, where the correlation decays at a certain rate over time.

We adopt the notation of Spiess and Hamerle (1996). Let $\boldsymbol{\beta} \in \mathbb{R}^q$ denote the parameter. Let $i = 1, \ldots, n$ be the index of the units and $j = 1, \ldots, t$ be the measurement or time index. Note that we assume here that all units have been measured t times. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{it})^T$ be the response vector of participant i over the t measurements and $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)$ denote the vector of all nt responses. Similarly, $\mathbf{x}_{ij} \in \mathbb{R}^q$ denotes the values of the q covariates (including the constant 1 to represent the intercept, if one is specified) for the ij-th observation, and the aggregated matrices are

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1} & \cdots & \mathbf{x}_{it} \end{pmatrix}^T \text{ and } \mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T & \cdots & \mathbf{X}_n^T \end{pmatrix}^T,$$

where $\mathbf{X}$ is assumed to have full column rank. A subscript β indicates that an expression is a function of the parameter value. Now, define $\eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$ and $\boldsymbol{\eta}_\beta = (\eta_{11}, \ldots, \eta_{nt})^T$. To specify the marginal generalized linear model for longitudinal data, suppose that a link function $h(\cdot)$ is known such that $E(\mathbf{y}) = \boldsymbol{\mu}_\beta = (h(\eta_{11}), \ldots, h(\eta_{nt}))^T$ when evaluated at the true parameter value. Further, assume that $\text{Var}(y_{ij}) = \phi V(h(\eta_{ij}))$, where $V(\cdot)$ is a variance function, usually associated with the link function $h(\cdot)$, and ϕ is a scale parameter. The scale parameter ϕ allows the (co-)variances to be different from what would be expected based on the regression parameter $\boldsymbol{\beta}$ alone, allowing a certain type of underdispersion or overdispersion. This is especially relevant for modeling count data using the log link. Define the diagonal matrix $\mathbf{V}_\beta = \text{diag}\big(V(h(\eta_{11})), \ldots, V(h(\eta_{nt}))\big)$. We assume that the responses are uncorrelated between units, so we only need to specify the correlations within units. Suppose that the correlation matrix is the same for all units and that it can be viewed as a function of a (vector-valued) parameter $\boldsymbol{\alpha}$. We denote the

correlation matrix by $\mathbf{R}_\alpha$, where the subscript α again represents the fact that this expression can be seen as a function of a parameter. Under the present assumptions, the full "working" covariance matrix over all nt responses, denoted by $\mathbf{\Sigma}$, can be expressed as $\mathbf{\Sigma}_{\phi,\alpha,\beta} = \phi \mathbf{V}_\beta^{1/2}(\mathbf{I}_n \otimes \mathbf{R}_\alpha)\mathbf{V}_\beta^{1/2}$. Here $\mathbf{I}_n$ denotes the $(n \times n)$ identity matrix and $\otimes$ the Kronecker product, so the middle matrix is a block diagonal matrix with identical blocks equal to $\mathbf{R}_\alpha$. Additional care must be taken in the case of the cumulative logit model (Agresti, 2002, p.274–275). Let $\mathbf{x}_{ij}^T$ have a 1 as its first entry, representing the intercept, and suppose y can take values in one of J ordered categories. Then $\mathbf{X}$ must be expanded by replacing each $\mathbf{x}_{ij}^T$ by the $((J-1) \times (J-1)q)$ matrix $\mathbf{X}_{ij} = \mathbf{I}_{J-1} \otimes \mathbf{x}_{i1}^T$, where $\mathbf{x}_{i1}^T$ is repeated $J-1$ times. The idea of the cumulative logit model is to use separate logit models for the $J-1$ probabilities $P(y \leq k|\mathbf{x})$ of y being in category k or lower for all but the highest category. Therefore, each y_{ij} has to be replaced by a binary vector with $J-1$ entries, indicating into which category y_{ij} falls. These extensions introduce new parameters for each category except the highest, representing the effects of the covariates on the cumulative probabilities. Taken together, due to the assumptions made here, the model is fully specified by the parameters $\phi, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. To estimate the model parameter of interest, $\boldsymbol{\beta}$, Generalized Estimating Equations (GEE) can be used (Liang & Zeger, 1986). Instead of using the likelihood function, which can be difficult to specify, this approach specifies only moment-type conditions. For our repeated measures marginal generalized linear model, the estimating equations are

$$\bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{D}_\beta \mathbf{\Sigma}_{\phi,\alpha,\beta}^{-1}(\mathbf{y} - \boldsymbol{\mu}_\beta) = \mathbf{0},$$

where $\mathbf{D}_\beta = \partial(\boldsymbol{\mu}_\beta)/\partial(\boldsymbol{\eta}_\beta) = \text{diag}(\partial h(\eta_{11})/\partial\eta_{11}, \ldots, \partial h(\eta_{nt})/\partial\eta_{nt})$ is a diagonal

matrix containing the derivatives of $\boldsymbol{\mu}_\beta$ with respect to $\boldsymbol{\eta}_\beta$.

Adding external information

To estimate an externally informed version of this model, we interpret $\bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta})$ as moment conditions (see Cameron and Trivedi (2005, p. 790) for details) and combine them with the external moment conditions $\bar{\mathbf{h}}$ as mentioned in Section 3, resulting in $\bar{\mathbf{g}} = (\bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta})^T, \bar{\mathbf{h}}^T)^T$. Assuming the necessary inverses to exist and using the (Schur) inversion formula for block matrices, the corresponding weighting matrix has a block form (parameter indices are suppressed in the following):

$$\begin{aligned}\mathbf{W}_n &= n \begin{pmatrix} \mathbf{X}^T\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathrm{Var}(\mathbf{y})\boldsymbol{\Sigma}^{-1}\mathbf{D}\mathbf{X} & \mathbf{X}^T\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathrm{Cov}(y-\boldsymbol{\mu},\mathbf{h}) \\ \mathrm{Cov}(y-\boldsymbol{\mu},\mathbf{h})^T\boldsymbol{\Sigma}^{-1}\mathbf{D}\mathbf{X} & \mathrm{Var}(\mathbf{h}) \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \boldsymbol{\Omega}_m & \boldsymbol{\Omega}_r^T \\ \boldsymbol{\Omega}_r & \boldsymbol{\Omega}_h \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{W}_m & \mathbf{W}_r^T \\ \mathbf{W}_r & \mathbf{W}_h \end{pmatrix}.\end{aligned}$$

Based on the first-order conditions for minimizing the GMM objective function (Hansen, 1982) and the block form of $\mathbf{W}_n$, we obtain estimating equations:

$$\begin{aligned}(\nabla_\beta \bar{\mathbf{g}}(\boldsymbol{\beta}))^T\mathbf{W}_n\bar{\mathbf{g}}(\boldsymbol{\beta}) = \mathbf{0} \quad &\Leftrightarrow (\nabla_\beta \bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta}))^T\left(\mathbf{W}_m\bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta}) + \mathbf{W}_r^T\bar{\mathbf{h}}\right) = \mathbf{0} \\ &\Leftrightarrow \quad \bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta}) - \boldsymbol{\Omega}_r^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}} = \mathbf{0}.\end{aligned}$$

To compute the estimate based on the first-order conditions, an iterative multistep procedure is used, requiring consistent estimators $\hat{\phi}$ given $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\alpha}}$ given ϕ as well as $\boldsymbol{\beta}$ (Liang & Zeger, 1986). To avoid this computational complexity, we use the results of Spiess and Hamerle (1996), who showed that

GEE estimators are simplified if the covariates are all either invariant across blocks or block specific. The basic idea is that $\boldsymbol{\alpha}$ and ϕ cancel out (partially) in these cases. We will show that for the block-invariant case, this property carries over to the externally informed tests in the point-valued case, and thus to the interval case as well.

Covariates are said to be block invariant, if the design matrix can be written with identical blocks $\mathbf{X}_i = \mathbf{Z}$ for $i = 1, \ldots, n$. For block invariant covariates, Spiess and Hamerle (1996) showed that all appearing matrices have repeated blocks and can therefore be simplified by Kronecker products, i.e. $\mathbf{X} = \mathbf{1}_n \otimes \mathbf{Z}$, $\mathbf{D} = \mathbf{I}_n \otimes \mathbf{D}_t$ and $\boldsymbol{\Sigma} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}_t$, where $\mathbf{1}_n$ is a vector full of ones with dimension n. To derive simplifications, it is assumed that $\mathbf{D}_t, \boldsymbol{\Sigma}_t$ and $\mathbf{Z}$ are quadratic and nonsingular. It seems to be a strict requirement that the latter matrix is quadratic. Therefore, some examples are given to show that it still covers interesting models.

First, consider a study that is interested only in time effects, for example, how the effect of a particular intervention evolves over time. This scenario can be modeled by including an intercept in the model representing the first measurement and dummy variables for each subsequent measurement representing the difference from baseline at that measurement. Here $\mathbf{Z}$ is given by changing the first column of $\mathbf{I}_t$ to ones. Since the result is a lower triangular matrix with ones on the diagonal, it is not singular. Using a generalized linear hypothesis, any of the differences can be compared, and global tests can be constructed to test whether a difference is equal to zero. The latter indicates that saturated analysis of variance models are also possible application scenarios. Note that $\mathbf{X} = \mathbf{1}_n \otimes \mathbf{Z}$ has full column rank if and only if $\mathbf{Z}$ is nonsingular. Thus, all within-subject designs that use the same procedure (e.g.,

stimuli) for all subjects satisfy the above requirement if their design matrix has full column rank (or can be made to have full column rank by deleting columns).

Second, it's worth noting that the blocks don't have to be participants; they can be (equally sized) groups of participants. Consider a study where researchers want to examine how the effect of an intervention, measured at certain points in time, varies between groups. While time is a within factor and group is a between factor, we can still block format the design matrix under certain conditions. Assume that the number of participants, N, is equally divided into k groups, i.e., $n = N/k$, and that all are measured t times. Extending the regression model from the first example by k dummy variables for the k groups, and including the interactions of these group dummies with all time dummies, we end up with kt parameter values. Now, since N is evenly distributed across groups, we can form blocks of k participants, each from a different group, by rearranging $\mathbf{X}$ accordingly. The resulting $(kt \times kt)$ blocks are quadratic and not singular. Admittedly, such a balanced design is rare in practice, but it could easily be constructed by alternating assignments to groups or by sampling mechanisms. One reason for this is the following simplification for block invariant designs.

Proposition 1. Let the subscript i indicate that only the values for block i are considered for the indexed expression. If the covariates are block invariant, the first-order conditions for the externally informed GEE estimator $\hat{\boldsymbol{\beta}}_{ex}$ are

$$\frac{1}{n}(\mathbf{D}_t\mathbf{Z})^{-1}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_i - \mathrm{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}) = \mathbf{0}$$

and its variance (matrix) estimator is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{ex}) = \\ \frac{1}{n}(\mathbf{D}_t\mathbf{Z})^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\mathbf{y}_i) - \mathrm{Cov}(\mathbf{y}_i-\boldsymbol{\mu}_i,\mathbf{h})\boldsymbol{\Omega}_h^{-1}\mathrm{Cov}(\mathbf{y}_i-\boldsymbol{\mu}_i,\mathbf{h})^T\right)(\mathbf{Z}^T\mathbf{D}_t)^{-1}.$$

Both the first-order conditions and the variance estimator do not contain the parameters $\boldsymbol{\alpha}$ and ϕ.

Proof. See Appendix A. □

Proposition 1 allows for a discussion of what kind of external information would be useful. Both the first-order conditions and the variance estimate differ from those for (uninformed) GEE estimates only by the terms $\mathrm{Cov}(\mathbf{y}_i-\boldsymbol{\mu}_i,\mathbf{h}_i)\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}$ and $\mathrm{Cov}(\mathbf{y}_i-\boldsymbol{\mu}_i,\mathbf{h}_i)\boldsymbol{\Omega}_h^{-1}\mathrm{Cov}(\mathbf{y}_i-\boldsymbol{\mu}_i,\mathbf{h}_i)^T$. Thus, $\mathrm{Cov}(\mathbf{y}-\boldsymbol{\mu},\mathbf{h})$ is the important term, and if it is expected to be different from $\mathbf{0}$, then variance reduction is expected to occur. This is similar to the discussion in Jann and Spiess (2024). A linear dependence between covariates and errors is generally not expected (which is positive for the validity of the model), so information about covariates is not useful per se. Subject of interest is information about the mean or the variance of the dependent variable, as well as information about the correlations or covariances between the dependent variable and the covariates. One example is information about the expected value of the dependent variable at the baseline. From a technical point of view, the inclusion of baseline information is possible by defining the corresponding moment function h to be different from 0 only for $j = 1$. In this way, the expected value is 0 for $j \neq 1$, which satisfies regularity conditions. There will be examples of external moments and their implementation in the simulation

studies in Section 5.

The expression $\text{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)$ can be estimated by $(\mathbf{y}_i - \boldsymbol{\mu}_i)\mathbf{h}_i^T$. To compute $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{ex})$, the expression $\text{Var}(\mathbf{y}_i)$ can be replaced by $(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T$ (Spiess & Hamerle, 1996). Now, to compute an estimate $\hat{\boldsymbol{\beta}}_{ex}$, the first-order condition from Proposition 1 can be solved analytically. The solution varies depending on whether the estimation procedure is strictly two-step or not. In a strict two-step procedure, $\text{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)$ is computed based on the first-step estimator and thus is not a function of $\boldsymbol{\beta}$, but is constant in the second step. Hence, the solution is

$$\hat{\boldsymbol{\beta}}_{ex} = \mathbf{Z}^{-1}\mathbf{h}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - (\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_1))\mathbf{h}_i^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}})\right),$$

where $\mathbf{h}^{-1}$ is a vector-valued function containing only the inverse of the link function h^{-1} as entries. Note that in the absence of external information the estimator simplifies to $\hat{\boldsymbol{\beta}}_{ex} = \mathbf{Z}^{-1}\mathbf{h}^{-1}(\bar{\mathbf{y}})$, a generalization of the analytic solution of Spiess et al. (2019) in the linear case. If $\hat{\boldsymbol{\beta}}_1$ is computed without (information about) ϕ and $\boldsymbol{\alpha}$, then the estimate is not a function of them at all. This can always be achieved by choosing a weighting matrix for the first-step estimator that is not a function of ϕ and $\boldsymbol{\alpha}$. In addition to the two-step approach, the first-order conditions can also be solved in a single step. The derivation is explained in Appendix B, the result is the one-step estimator

$$\hat{\boldsymbol{\beta}}_{ex} = \mathbf{Z}^{-1}\mathbf{h}^{-1}\left(\frac{1}{1 - \bar{\mathbf{h}}^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}} \cdot \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i\mathbf{h}_i^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}})\right).$$

Taken together, if the covariates are block invariant, then the estimate for the parameter of the externally informed generalized linear regression model

for longitudinal data and its variance are independent of the working correlation matrix, and thus identical for any possible covariance structure. This property carries over directly to the test statistic T_W, since it is computed using only the GMM estimator based on the first-order conditions and the variance estimate derived from Proposition 1. In the following, we will always choose the one-step estimator to compute T_W, as it leads to the most computationally efficient test. The same holds for LM, using the first approach based on a strict two-step procedure, the only difference being that $\text{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)$ is computed based on a restricted first-step estimator.

Proposition 2. Let the covariates be block invariant and assume that the regularity conditions for the externally informed GEE estimator are fulfilled (especially that it is a solution to the first-order conditons), then the test statistic of the criterion-based test becomes

$$\begin{aligned} D_{RU} = & \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_i - \text{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}})^T \\ & \times \Big(\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T - \text{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)\boldsymbol{\Omega}_h^{-1}\text{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)^T\Big)^{-1} \\ & \times \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_i - \text{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}), \end{aligned}$$

where all expressions are evaluated at $\tilde{\boldsymbol{\beta}}_2$.

Proof. See Appendix C. □

Using Proposition 2, D_{RU} is not a function of ϕ and $\boldsymbol{\alpha}$ if the restricted estimator $\tilde{\boldsymbol{\beta}}_2$ can be computed without using ϕ or $\boldsymbol{\alpha}$. Details about the corresponding computation of restricted estimators are described in Appendix D. Note that $\mathbf{W}_n$ is fixed in each step and is recalculated after the new estimate

is found. In summary, with external point information and block invariant covariates, it is possible to compute test statistics for general linear hypothesis testing in generalized linear models with repeated measures without ϕ and $\boldsymbol{\alpha}$, i.e., without having to specify a working covariance matrix. Note that this applies to repeated-measures designs, such as crossover studies, as well as longitudinal studies, even when the interval between measurements of the dependent variable varies. In the latter case, however, one should still be cautious about the comparability of the slopes of the regression model. Furthermore, models involving between-subject variables with many values, e.g. age, generally do not allow the derived simplifications.

When using external sets, the minimum test statistic must be computed to perform a $\Gamma-$maximin test. Even if analytic formulas for estimators and test statistics are available, this is generally a non-convex optimization problem, which becomes a problem when an external interval is considered (Jann, 2024). In addition, calculating the restricted estimates $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$ is done by an iterative procedure. If the dimension of the external interval is small, grid searching on the external interval is a simple procedure to approximate a solution without further assumptions (Cameron & Trivedi, 2005, p. 337).

So far we have assumed that all matrices are invertible. However, there is one case where this assumption is certainly not true. If the external variance matrix $\mathbf{V}_{ex}$ is not regular, for example if $\mathbf{V}_{ex} = \mathbf{0}$, it may be that $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{ex})$ is singular. To illustrate this, consider the case where $\mathbf{h}$ encodes the true external information about the expected value of $\mathbf{y}$ in a particular group, say the k-th group. Then the k-th entry of $\mathrm{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h})$ is just the variance of $\mathbf{y}$ in group k, and the remaining entries are the covariances with the other groups. Since $\mathbf{V}_{ex} = 0$, $\boldsymbol{\Omega}_h$ is also just the variance of $\mathbf{y}$ in the k-th group. Thus, the k-th

column and the k-th row of $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{ex})$ computed based on Proposition 1 would be exactly $\mathbf{0}$, implying that it is singular. This can also apply to matrices calculated on the basis of $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{ex})$, and to the restricted estimation. Fortunately, the simple workaround is to use Moore-Penrose inverses, see (Lemma 1 and Theorem 1 in) Jann (2024).

Simulation studies and real data application

To analyze the validity and power of the proposed tests, we conducted simulation studies in two scenarios, both adapted from two published studies and thus from real data. An example of a within design for count data is analyzed. Further, we consider a mixed design with a categorical dependent variable using a cumulative logit model. We will describe the theory behind these papers very briefly, since we are only using them as examples. The simulations were performed in the statistical software R, version 4.3.2 (R Core Team, 2023). The implementation of test functions, simulation studies and data reanalysis can be found in the R scripts in the electronic supplementary material. Some technical details of the simulation studies have been omitted to keep the reading brief. These details can be found in Appendix E.

Count data scenario

The simulations in this section are based on the scenario of Experiment 3 of Schmalbrock and Frings (2022). The corresponding datasets for all three experiments are open source, see Schmalbrock (2022). Theoretically, the authors investigate the principle of figure-ground segmentation in an experiment using sequential distractor-response binding (DRB). Four within factors were manipulated: response relation (whether the response repeats or not), color

relationship (whether color repeats or not), prime layer (if color was presented as figure or background in the prime), and probe layer (if color was presented as figure or background in the probe). All participants had to complete 32 trials of each of the $2^4 = 16$ possible factor combinations. Repeated measures ANOVAs were then used to analyze changes in mean reaction times and mean error rates, including two-, three-, and four-way interaction terms. Similarly, we focus on the number of errors as a count variable. We restrict our attention to testing the basic DRB effect when both colors in the prime and probe are in the background, i.e., the two-way interaction between response and color relation.

As a source for external information we used experiments 1 and 2 by Schmalbrock and Frings (2022). They display a good source of external information because they are based on the same topic, conducted in the same lab, but with a different design, i.e., in Experiment 1 only the priming layer was varied and in Experiment 2 only the probe layer was varied. We used two approaches (a conservative as well as a liberal one, see Appendix E.1) to calculate external intervals for the number of errors for the condition where color and response are repeated and color is the background for both prime and probe, based on the data from Experiments 1 and 2. The resulting external intervals are $[1.1667, 2.5]$ for the conservative approach and $[1.3667, 1.8667]$ for the liberal approach. In both cases, the minimum variance of the conditions over which the range was built was 2.6023. We used these external intervals as well as the minimum variance for our simulation study.

We simulated the data in a scenario similar to Experiment 3, a $2 \times 2 \times 2 \times 2$ within-design with all interactions (resulting in 16 parameters). The sample size was set to 63. We calculated the mean number of errors for all 16 factor combinations and used these to generate data via Poisson

distributions. As a reference category, we used the condition where both response and color change, and color is the background in both prime and probe. To test validity, a null model where all parameters except the intercept were set to zero was used to generate the data. Therefore, we used the mean number of errors of the reference category, 1.7937, in all conditions. To test power, we used the full vector of the 16 calculated means as an alternative model. We used a simple third variable reduction approach described by Barbiero and Ferrari (2015) to generate correlated Poisson data with the defined means (see Appendix E.2).

To test the influence of external information on the three tests under different conditions, we created 11 cases: First, no external information was used as a baseline. Second, the true value based on the model was utilized as a point to evaluate the best case result. Then, the external intervals calculated above are applied to evaluate the influence of the interval width (both include the true value). To analyze the influence of misspecification, we used a false point value of 2.5 and a false interval of $[2.5, 4]$. All cases so far do not use external variance estimates. To test the influence of estimation uncertainty, we used all cases except the first and introduced estimation uncertainty by generating samples from a normal distribution in each simulation run. To accomplish this, the mean was set equal to the (minimum and maximum) external value of the previously discussed cases. The variance was set to the value calculated above, 2.6023, and the sample size was 30, as in experiments 1 and 2. The (minimum and maximum) means of these samples were used as external information and the (minimum) variance of the samples divided by 30 as external variance. During each simulation run, T_W (based on the one-step estimator), LM, and D_{RU} are computed, and it is evaluated whether the tests

reject the null hypothesis based on $\alpha = 0.05$. Finally, the null hypothesis rejection rate is computed over all simulations, and the number of divergences, i.e., when the Fisher scoring method for the restricted estimates fails to converge in less than 10000 iterations, is tracked. As a convergence criterion, we defined that two consecutive estimates have an Euclidean distance less than 10^{-12}. The number of simulation runs was set to 500. The simulation results in terms of type I error and power are shown in Table 1. There was no divergence of the Fisher scoring algorithm.

Table 1

Type I error rates and power resulting from the count data simulation.

	Type I errors			Power		
Scenario	T_W	LM	D_{RU}	T_W	LM	D_{RU}
without external information	0.050	0.050	0.022	0.862	0.862	0.804
external values (without V_{ex})						
true external value	0.072	0.066	0.066	0.928	0.920	0.912
$[1.1667, 2.5]$ (true)	0.002	0.002	0.002	0.430	0.270	0.250
$[1.3667, 1.8667]$ (true)	0.024	0.022	0.018	0.912	0.908	0.886
2.5 (false)	0.416	0.274	0.272	0.430	0.270	0.250
$[2.5, 4]$ (false)	0.416	0.168	0.256	0.058	0.002	0.012
external samples (with V_{ex})						
true external value	0.112	0.080	0.088	0.880	0.864	0.856
$[1.1667, 2.5]$ (true)	0.004	0.004	0.004	0.466	0.364	0.464
$[1.3667, 1.8667]$ (true)	0.058	0.048	0.042	0.860	0.834	0.830
2.5 (false)	0.382	0.238	0.358	0.500	0.366	0.476
$[2.5, 4]$ (false)	0.424	0.142	0.410	0.074	0.002	0.126

Note: The tests are symbolized by the respective test statistic. T_W stands for the Wald test, LM for the Lagrange multiplier test and D_{RU} for the criterion-based test. The significance level is $\alpha = 0.05$, so type I errors should be about 0.05 or less.

Categorical data scenario

The simulations in this section are based on the scenario of Zeibig et al. (2023). The corresponding dataset is open source, see Zeibig et al. (2022). The authors examined the positive long-term effect of exercise intervention on mental disorders in a sample of 74 outpatients awaiting psychotherapy.

Participants were assigned to either a passive control group or an exercise intervention group. Both groups were measured three times, at baseline, immediately after treatment, and one year after treatment. Thus, the design is mixed, including intervention group as a between factor and measurement time as a within factor. The between factor was originally unbalanced, with 38 participants in the exercise group, but two participants in this group had to be excluded due to misdiagnosis at baseline, resulting in a balanced between factor. Therefore, in principle, our approach can be applied as described in Section 4.1 by pairing participants from both groups. As outcome measure we only use the affect regulation subscale of the Physical Activity-related Health Competence Questionnaire (PAHCO). It is calculated as the mean of four items, each of which has a four-point Likert scale coded from 1 to 4. For our analysis, we formed three categories $[1, 2)$, $[2, 3)$, and $[3, 4]$ and named them category 1, 2, and 3. We applied a cumulative logit model to model the effect of the two factors on the probability of a participant being in category 1 and being in category 1 or 2 (abbreviated as $1 \cup 2$).

As a source of external information we used a validation study for the PAHCO based on 1028 patients in rehabilitation facilities at the beginning of their medical program, see Study A in Sudeck and Pfeifer (2016). Although they differed from the sample of Zeibig et al. (2023) in types of disorders (medical vs. mental) and type of upcoming treatment (rehabilitation vs. psychotherapy), both samples had a similar range of ages and had not started treatment but were waiting for it. In order to account for these differences, we planned to construct an interval of possible values for the relative frequencies of category 1 and $1 \cup 2$. We faced the problem that the dataset of Sudeck and Pfeifer (2016) is not available online and that the study does not report the

named frequencies. To approximate these frequencies, we used linear programming to calculate the possible minimum and maximum frequencies based on the given information, details are described in Appendix E.3. This provides another justification for the use of intervals: Not only to reflect the presence of multiple studies, but also in the case of uncertainty about the external information from a single study. The resulting intervals are $[0.238, 0.362]$ for category 1 and $[0.692, 0.806]$ for $1 \cup 2$.

We simulated the data according to the 3×2 mixed design with all interactions (resulting in 12 parameters, 6 for category 1 and 6 for $1 \cup 2$). As the reference category, we chose category 1 in the control group at baseline. The sample size was set to 72. To generate data according to the dataset of Zeibig et al. (2022), we wanted to calculate the frequency of category 1 as well as $1 \cup 2$ in all 6 factor combinations in the dataset. However, there was a significant amount of missing data for the PAHCO score. To restore the balance in the between factor without deleting data, we decided to use the multiple imputation method implemented by the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2011). Specifically, the PAHCO scores were imputed using predictive mean matching with 5 donors and 100 imputations. The rule of thumb that longitudinal data require at least as many imputations as the percentage of missing values accounts for this large number of imputations Wijesuriya et al. (2025). In each imputed dataset, the 12 frequencies were calculated and the mean of these frequencies over the 100 datasets were used as probabilities for the simulation study.

We then used a threshold model based on a joint normal distribution to induce dependencies across measurements (see Appendix E.4 for details). To test validity, a null model was applied to generate the data, where all

parameters except the intercepts for category 1 and for $1 \cup 2$ were set to zero. Thus, we utilized probabilities of 0.266 for category 1 and 0.482 for category 2. To test the power, we used the full vector of the 12 calculated frequencies as an alternative model.

We tested the same 11 external information conditions for both omnibus hypotheses as in the count data simulations. To analyze the influence of misspecification, we defined 0.35 and 0.5 as false values and $[0.35, 0.4]$ and $[0.4, 0.5]$ as false intervals. This time, however, we generated external samples before the simulation runs to test the influence of reusing the same external samples. Since external values represent probabilities, we used a Bernoulli distribution to generate the external samples, using the (minimum and maximum) external values of the above conditions as probabilities. The sample size for the external samples was set to 1028, the same as in Sudeck and Pfeifer (2016). Complete separation may occur because the sample size of 72 is relatively small. For a discussion of this phenomenon in the context of GEE and its remedy using a Firth-type penalty term, see Mondol and Rahman (2019). We restrict ourselves to excluding and counting the number of simulation runs where complete separation occurred. To maintain a sufficient number of simulation runs, we increased the number from 500 to 700 after having done some pilot simulations.

The resulting type I error rates and power of the simulation conditions are shown in Table 2 and Table 3. For the null model, no complete separation occurred, for the alternative model, 100 (category 1) and 101 ($1 \cup 2$) cases occurred, leaving over 500 runs in each of these cases. There were divergences for the simulations with $1 \cup 2$. For the alternative model and D_{RU} there was one divergence in the no external information condition, one for the external

Table 2
Type I error rates and power for category 1, categorical data.

	Type I errors			Power		
Scenario	T_W	LM	D_{RU}	T_W	LM	D_{RU}
without external information	0.063	0.063	0.046	0.633	0.633	0.622
external values (without V_{ex})						
true external value	0.091	0.077	0.067	0.772	0.788	0.943
$[0.238, 0.362]$ (true)	0.046	0.034	0.041	0.665	0.678	0.887
0.35 (false)	0.346	0.277	0.309	0.978	0.985	0.990
$[0.35, 0.4]$ (false)	0.343	0.277	0.307	0.978	0.985	0.990
external samples (with V_{ex})						
true external value	0.094	0.079	0.061	0.767	0.772	0.938
$[0.238, 0.362]$ (true)	0.046	0.036	0.046	0.675	0.688	0.898
0.35 (false)	0.281	0.221	0.259	0.970	0.978	0.990
$[0.35, 0.4]$ (false)	0.249	0.189	0.231	0.960	0.968	0.990

Note: The tests are symbolized by the respective test statistic. T_W stands for the Wald test, LM for the Lagrange multiplier test and D_{RU} for the criterion-based test. The significance level is $\alpha = 0.05$, so type I errors should be about 0.05 or less.

interval cases (with or without external variance) and 9 in the false external point condition, 53 for the false external interval, 9 for the false external point with external variance, and 51 for the false external interval with (minimal) external variance. For the alternative model and LM, two divergences occurred in the cases with false external points, respectively, as well as 5 for the false external interval without external variance and 4 with external variance. For the null model, only D_{RU} produced divergences, 2 for false external point, 19 for false external interval, 2 for false external point with external variance, and 17 for false external interval with (minimal) external variance.

Application to real data

We applied all three hypothesis tests to the real datasets from Sections 5.1 and 5.2, testing the same hypotheses as in the simulation studies using the specified external intervals and external variance, respectively. To adjust for missing values in the categorical data, we used the 100 imputations described

Table 3
Type I error rates and power for category $1 \cup 2$*, categorical data.*

	Type I errors			Power		
Scenario	T_W	LM	D_{RU}	T_W	LM	D_{RU}
without external information	0.086	0.086	0.054	0.806	0.806	0.711
external values (without V_{ex})						
true external value	0.109	0.096	0.074	0.995	0.995	0.967
$[0.692, 0.806]$ (true)	0.073	0.064	0.036	0.975	0.943	0.798
0.5 (false)	0.991	0.849	0.747	0.606	0.273	0.145
$[0.4, 0.5]$ (false)	0.991	0.843	0.711	0.593	0.229	0.097
external samples (with V_{ex})						
true external value	0.111	0.100	0.069	1.000	1.000	0.973
$[0.692, 0.806]$ (true)	0.071	0.061	0.034	0.972	0.940	0.788
0.5 (false)	0.993	0.853	0.747	0.608	0.272	0.145
$[0.4, 0.5]$ (false)	0.991	0.840	0.710	0.589	0.229	0.097

Note: The tests are symbolized by the respective test statistic. T_W stands for the Wald test, LM for the Lagrange multiplier test and D_{RU} for the criterion-based test. The significance level is $\alpha = 0.05$, so type I errors should be about 0.05 or less.

above and performed the tests based on the D_2 test statistic (Li et al., 1991). The results are shown in Table 4. For the count data case, using the wide interval resulted in an increase in the p-values, although they are still significant using 0.05 as the significance level, while using the narrow interval halved the p-values. For the categorical data, the p-values decreased, leading to significant results for the category 1 that were not significant before.

Discussion

Summary of simulation results and data reanalysis

In both simulation studies, the three tests led to valid type I errors when used without external information, except for category $1 \cup 2$, where T_W and LM had slightly increased type I errors, possibly due to the small sample size of 36 pairs. In all cases, using the true external point slightly increased type I errors, regardless of whether external variance was used or not. Combined with the fact that using external intervals (with or without external variance) reduced

Table 4

Resulting p-values from reanalysis for count and categorical data.

Test	Count	$(\text{Count})_{ex,1}$	$(\text{Count})_{ex,2}$	Cat 1	$(\text{Cat } 1)_{ex}$	Cat $1 \cup 2$	$(\text{Cat } 1 \cup 2)_{ex}$
T_W	0.0014	0.0132	0.0006	0.1411	0.0338	0.0262	0.0084
LM	0.0014	0.0385	0.0006	0.1411	0.0335	0.0262	0.0092
D_{RU}	0.0099	0.0352	0.0031	0.1363	0.0033	0.0380	0.0202

Note: The tests are symbolized by the respective test statistic. T_W stands for the Wald test, LM for the Lagrange multiplier test and D_{RU} for the criterion-based test. Count, Cat 1, and Cat $1 \cup 2$ symbolize the tests performed without external information for the hypotheses specified in the Sections 5.1 and 5.2. A subscript ex indicates that an external interval and an external variance are used. For the the count data, there were two intervals, so $ex, 1$ indicates the use of $[1.1667, 2.5]$ and $ex, 2$ indicates the use of $[1.3667, 1.8667]$.

type I errors compared to the case where no external information is used, this provides an argument for using intervals even when a reliable external value is known. This is supported by the observation that using false external points significantly increases type I errors, up to 0.993 for category $1 \cup 2$ in the categorical case using T_W. Furthermore, using false external intervals (with or without external variance) does not prevent increased type I errors. The fact that spurious external intervals lead to almost the same type I errors as spurious external points indicates that the width of the interval per se is not as important as the best case value within the interval. However, for the count data, the wide true external interval led to significantly lower type I errors than the narrower true external interval, reducing them from 0.058 to 0.004 in the case of T_W and using external variance. Thus, caution is required to construct external intervals that contain true values, or an unbiased or consistent estimate thereof, in order to benefit from the width of the interval in terms of validity. In both simulations, using external samples based on true external values or intervals did not greatly increase type I errors. For the categorical data, they were nearly equal to those using the true external value or interval, suggesting that the conditional use of external information (i.e., using the same external

sample each time) was not a problem in our case. However, this may be due to the large sample size of 1028. For the count data, type I errors doubled compared to the case without using external samples, but all were close to the nominal 0.05, suggesting that some caution is needed when interpreting external information unconditionally, e.g., repeatedly running multiple experiments and leveraging information from one experiment in the subsequent ones. This may be due to the small external sample size of 30.

Regarding power in the case of count data, using true external values without variance led to an increase in power of about 0.06 for T_W and LM and 0.11 for D_{RU}. This effect was reduced when external samples were used, to 0.02 for T_W, 0.002 for LM, and 0.05 for D_{RU}. Using the wide external true interval or false point values or intervals resulted in power of 0.5 or less, indicating a significant loss of power in these cases. Using the narrow true external interval without external variance reduced power by 0.03 at most compared to using the true external point, so there is still an increase in power for all tests. When the external variance is also considered, the power is lower when LM is used, the same when T_W is used, and higher when D_{RU} is used compared to the power without using external information. In terms of absolute power, not power reduction, T_W always had the highest power values, indicating that T_W is the most powerful test statistic here. Note that although externally informed T_W has lower power and similar type I errors compared to uninformed T_W, there may still be distributional robustness in validity, meaning that the type I errors may remain nominal in a certain neighborhood of the normal distribution due to the use of interval probability for test construction.

For the categorical data, the results are different for the categories. For category 1, the highest gain in power occurred for D_{RU}, which was 0.32. The

power gain was only slightly affected when external samples were used. Using a true external interval resulted in a maximum net power gain of only 0.04 for T_W and LM, while it remained above 0.26 for D_{RU}. Furthermore, in the presence of external information, the power was always highest for D_{RU}. Using false external values or intervals resulted in power values of 0.96 or higher. While this seems like a good effect, it is clearly offset by the increased type I errors, and it only occurred for category 1, so it is not reliable. For $1 \cup 2$ the results were partly the opposite. Again, D_{RU} led to the highest power gain and using external samples had no significant effect on power. However, using true external intervals resulted in a power gain of more than 0.199 for T_W, while it shrank to 0.08 for D_{RU}. Furthermore, T_W always led to the highest power. Finally, with false external information the power was lower than without external information. These results are reflected in the (reduced) p-values for the reanalysis of the datasets in Table 4. The test statistic T_W was adavantegous for count data and category $1 \cup 2$, while D_{RU} was adavantegous for category 1 of the ordinal data. While the simulation study for count data showed that there are no substantial power gains left when using external intervals with external variance, the reanalysis showed that the use of external information can still be useful for particular datasets.

Conclusion and further research

The summary of the results shows that no test is superior in all cases, and even for the same dataset it can vary which test is more powerful. Despite type I and type II errors, another important argument for test selection is computational complexity. Here, T_W is clearly advantageous because it can be calculated analytically, without iterative procedures, thus avoiding the need to

specify good initial values, which was important for D_{RU}. Further research could examine the behavior of the three test statistics proposed in this paper, and perhaps other possible test statistics to provide a more nuanced recommendation as to which test is best to use in different scenarios. It may be that no single test is best for the majority of scenarios, but that the choice of test is highly scenario dependent. It is therefore advisable to perform simulation studies before applying a test to another scenario.

We have shown that reflecting the uncertainty in the external information by using external intervals and external variances does not substantially reduce power. In the categorical case, the substantial increase in power was preserved, while in the count data case, power was not substantially reduced when a reasonably narrow external interval was used, so an increase in robustness of validity may be the net gain here. Further research is needed to assess whether and to what extent the robustness of validity increases for various violations of the distributional assumptions when external information is used. Since the use of false external information in many cases led to more type I and II errors, we strongly recommend using an external interval and external variance that are theoretically justified, i.e., constructed on the basis of studies that are as similar as possible to the present study, to avoid misspecification.

Finally, it would be interesting to extend our results on block invariant designs to unbalanced designs. The derived estimators can be applied to unbalanced designs as they are. This is because they are based only on means for each factor combination, and these are consistent estimates even when sample sizes vary across factor combinations. Unfortunately, the variance matrices do not generalize as easily, and additional assumptions may be needed to compute them for unbalanced designs, as is the case for unbalanced ANOVA.

References

Agresti, A. (2002). Logit models for multinomial responses. *Categorical data analysis* (pp. 267–313). John Wiley & Sons, Ltd. https://doi.org/10.1002/0471249688.ch7

Andrews, D. W. & Guggenberger, P. (2017). Asymptotic size of kleibergen's lm and conditional lr tests for moment condition models. *Econometric Theory*, *33*(5), 1046–1080. https://doi.org/10.2139/ssrn.2541746

Augustin, T., Coolen, F. P., De Cooman, G. & Troffaes, M. C. (2014). *Introduction to imprecise probabilities*. John Wiley & Sons. https://doi.org/10.1002/9781118763117

Barbiero, A. & Ferrari, P. A. (2015). Simulation of correlated poisson variables. *Applied Stochastic Models in Business and Industry*, *31*(5), 669–680. https://doi.org/10.1002/asmb.2072

Bond, S., Bowsher, C. & Windmeijer, F. (2001). Criterion-based inference for gmm in autoregressive panel data models. *Economics Letters*, *73*(3), 379–388. https://doi.org/10.2139/ssrn.265068

Bond, S. & Windmeijer, F. (2005). Reliable inference for gmm estimators? finite sample properties of alternative test procedures in linear panel data models. *Econometric Reviews*, *24*(1), 1–37. https://doi.org/10.1081/ETC-200049126

Brown, B. W. & Newey, W. K. (2002). Generalized method of moments, efficient bootstrapping, and improved inference. *Journal of Business & Economic Statistics*, *20*(4), 507–517. https://doi.org/10.1198/073500102288618649

Cameron, A. & Trivedi, P. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511811241

Diggle, P. J., Heagerty, P. J., Liang, K.-y. & Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford University Press. https://doi.org/10.1093/oso/9780198524847.001.0001

Dufour, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics/Revue canadienne d'économique*, *36*(4), 767–808. https://doi.org/10.1111/1540-5982.t01-3-00001

Dufour, J.-M., Trognon, A. & Tuvaandorj, P. (2017). Invariant tests based on m-estimators, estimating functions, and the generalized method of moments. *Econometric Reviews*, *36*(1-3), 182–204. https://doi.org/10.1080/07474938.2015.1114285

Ghosh, B. K. (1973). Some monotonicity theorems for chi square, F and t distributions with applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, *35*(3), 480–492. https://doi.org/10.1111/j.2517-6161.1973.tb00976.x

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*(4), 1029–1054. https://doi.org/10.2307/1912775

Hellerstein, J. K. & Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *The Review of Economics and Statistics*, *81*(1), 1–14. https://doi.org/10.1162/003465399557860

Huber, P. J. & Ronchetti, E. M. (2009). Robust tests. *Robust statistics* (pp. 297–305). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470434697.ch13

Imbens, G. W. & Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, *61*(4), 655–680. https://doi.org/10.2307/2297913

Jann, M. (2023). Testing the coherence of data and external intervals via an imprecise Sargan-Hansen test. In E. Miranda, I. Montes, E. Quaeghebeur & B. Vantaggi (Eds.), *Proceedings of the thirteenth international symposium on imprecise probability: Theories and applications* (pp. 249–258). PMLR. https://proceedings.mlr.press/v215/jann23a.html

Jann, M. (2024). Testing the fit of data and external sets via an imprecise sargan-hansen test. *International Journal of Approximate Reasoning*, 109214. https://doi.org/10.1016/j.ijar.2024.109214

Jann, M. & Spiess, M. (2024). Using external information for more precise inferences in general regression models. *Psychometrika*, 1–22. https://doi.org/10.1007/s11336-024-09953-w

Li, K.-H., Meng, X.-L., Raghunathan, T. E. & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, *1*(1), 65–92. Retrieved March 13, 2025, from http://www.jstor.org/stable/24303994

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.2307/2336267

Mondol, M. H. & Rahman, M. S. (2019). Bias-reduced and separation-proof gee with small or sparse longitudinal binary data. *Statistics in Medicine*, *38*(14), 2544–2560. https://doi.org/10.1002/sim.8126

Newey, W. K. & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, *28*(3), 777–787. Retrieved September 5, 2024, from https://doi.org/10.2307/2526578

Puntanen, S., Styan, G. P. H. & Isotalo, J. (2011). *Matrix tricks for linear statistical models*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10473-2

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ronchetti, E. & Trojani, F. (2001). Robust inference with gmm estimators. *Journal of econometrics*, *101*(1), 37–69. https://doi.org/10.1016/S0304-4076(00)00073-7

Rotnitzky, A. & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, *77*(3), 485–497. Retrieved September 3, 2024, from https://doi.org/10.2307/2336986

Schmalbrock, P. (2022). Dataset for: A mighty tool not only in perception: Figure-ground mechanisms control binding an retrieval alike. https://doi.org/10.23668/psycharchives.5619

Schmalbrock, P. & Frings, C. (2022). A mighty tool not only in perception: Figure-ground mechanisms control binding and retrieval alike. *Attention, Perception, & Psychophysics*, *84*(7), 2255–2270. https://doi.org/10.3758/s13414-022-02511-5

Spiess, M. & Hamerle, A. (1996). On the properties of gee estimators in the presence of invariant covariates. *Biometrical Journal*, *38*(8), 931–940. https://doi.org/10.1002/bimj.4710380805

Spiess, M., Jordan, P. & Wendt, M. (2019). Simplified estimation and testing in unbalanced repeated measures designs. *Psychometrika*, *84*(1), 212–235. https://doi.org/10.1007/s11336-018-9620-2

Sudeck, G. & Pfeifer, K. (2016). Physical activity-related health competence as an integrative objective in exercise therapy and health sports - conception and validation of a short questionnaire. *German Journal of Exercise and Sport Research*, *46*(2), 74–87. https://doi.org/10.1007/s12662-016-0405-4

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Wijesuriya, R., Moreno-Betancur, M., Carlin, J. B., White, I. R., Quartagno, M. & Lee, K. J. (2025). Multiple imputation for longitudinal data: A tutorial. *Statistics in Medicine*, *44*(3-4), e10274. https://doi.org/10.1002/sim.10274

Zeibig, J.-M., Seiffer, B., Frei, A. K., Takano, K., Sudeck, G., Rösel, I., Hautzinger, M. & Wolf, S. (2023). Long-term efficacy of exercise across diagnostically heterogenous mental disorders and the mediating role of affect regulation skills. *Psychology of Sport and Exercise*, *64*, 102340. https://doi.org/10.1016/j.psychsport.2022.102340

Zeibig, J.-M., Seiffer, B. A., Frei, A. K., Rösel, I., Hautzinger, M., Wolf, S. & Takano, K. (2022). Dataset and codebook for: Long-term efficacy of exercise across diagnostically heterogenous mental disorders and the mediating role of affect regulation skills. https://doi.org/10.23668/psycharchives.8254

Appendices to the paper

Testing linear hypotheses in repeated measures generalized linear models using external information

Contents

A: Proof of Proposition 1

Here we present the proof of Proposition 1:

Proof. By Liang and Zeger (1986, p. 21) $\nabla_{\boldsymbol{\beta}}\bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta})$ can asymptotically be expressed by the limit (in probability) $-\lim_{n\to\infty}\frac{1}{n}\mathbf{X}^T\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}\mathbf{X}$, so to derive consistent estimators we replace $\nabla_{\boldsymbol{\beta}}\bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta})$ by $\mathbf{G} := -\frac{1}{n}\mathbf{X}^T\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}\mathbf{X}$. $\mathbf{G}$ is nonsingular because $\mathbf{X}$ has full rank and $\mathbf{D}$ as well as $\boldsymbol{\Sigma}^{-1}$ are nonsingular because they have nonsingular blocks. Under block-invariant covariates, it holds that $\mathbf{G} = \frac{1}{n}(\mathbf{1}_n^T \otimes \mathbf{Z}^T)(\mathbf{I}_n \otimes \mathbf{D}_t)(\mathbf{I}_n \otimes \boldsymbol{\Sigma}_t^{-1})(\mathbf{I}_n \otimes \mathbf{D}_t)(\mathbf{1}_n \otimes \mathbf{Z})$. By the properties of the Kronecker product (see Puntanen et al. (2011, p. 52) for details) its inverse is $\mathbf{G}^{-1} = \frac{1}{n}(\mathbf{1}_n^T \otimes \mathbf{Z}^{-1})(\mathbf{I}_n \otimes \mathbf{D}_t^{-1})(\mathbf{I}_n \otimes \boldsymbol{\Sigma}_t)(\mathbf{I}_n \otimes \mathbf{D}_t^{-1})(\mathbf{1}_n \otimes (\mathbf{Z}^T)^{-1})$. Since $\mathbf{G}^{-1}$ is nonsingular, multiplying it to the first- order conditions does not change the solution, so the following equation is also a first-order condition:

$$
\begin{aligned}
\mathbf{0} &= \mathbf{G}^{-1}(\bar{\mathbf{m}}_{GEE}(\boldsymbol{\beta}) - \boldsymbol{\Omega}_r^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}) \\
&= \mathbf{G}^{-1}\mathbf{X}^T\mathbf{D}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu} - \mathrm{Cov}(y - \boldsymbol{\mu}, \mathbf{h})) \\
&= \mathbf{G}^{-1}(\mathbf{1}_n^T \otimes \mathbf{Z}^T)(\mathbf{I}_n \otimes \mathbf{D}_t)(\mathbf{I}_n \otimes \boldsymbol{\Sigma}_t^{-1})(\mathbf{y} - \boldsymbol{\mu} - \mathrm{Cov}(y - \boldsymbol{\mu}, \mathbf{h})\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}) \\
&= \frac{1}{n}(\mathbf{1}_n^T \otimes (\mathbf{D}_t\mathbf{Z})^{-1})(\mathbf{y} - \boldsymbol{\mu} - \mathrm{Cov}(\mathbf{y} - \boldsymbol{\mu}, \mathbf{h})\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}).
\end{aligned}
$$

Now the rest follows by evaluating the Kronecker product. Further, using (the proof of) Corollary 1 in Jann and Spiess (2024),

$$
\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{ex}) &= \frac{1}{n}(\mathbf{G}(\boldsymbol{\Omega}_m - \boldsymbol{\Omega}_r^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Omega}_r)^{-1}\mathbf{G})^{-1} = \frac{1}{n}\mathbf{G}^{-1}(\boldsymbol{\Omega}_m - \boldsymbol{\Omega}_r^T\boldsymbol{\Omega}_h^{-1}\boldsymbol{\Omega}_r)\mathbf{G}^{-1} \\
&= \frac{1}{n}\mathbf{G}^{-1}(\mathbf{1}_n^T \otimes \mathbf{Z}^T)(\mathbf{I}_n \otimes \mathbf{D}_t)(\mathbf{I}_n \otimes \boldsymbol{\Sigma}_t^{-1})\big(\mathrm{Cov}(\mathbf{y}) \\
&\quad - \mathrm{Cov}(\mathbf{y} - \boldsymbol{\mu}, \mathbf{h})\boldsymbol{\Omega}_h^{-1}\mathrm{Cov}(\mathbf{y} - \boldsymbol{\mu}, \mathbf{h})^T\big)(\mathbf{I}_n \otimes \boldsymbol{\Sigma}_t^{-1})(\mathbf{I}_n \otimes \mathbf{D}_t)(\mathbf{1}_n \otimes \mathbf{Z})\mathbf{G}^{-1},
\end{aligned}
$$

which leads to the desired expression of the variance estimator by the same arguments regarding the Kronecker products when multiplying $\mathbf{G}^{-1}$. Only $\boldsymbol{\Sigma}_t$ is a function of $\boldsymbol{\alpha}$ and ϕ, and it is cancelled out in both expressions. □

B: Derivation of the one-step estimator

Here we derive the one-step estimator described on page 20. Based on Proposition 1 the first-order conditions are

$$\frac{1}{n}(\mathbf{D}_t\mathbf{Z})^{-1}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_i - \mathrm{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}) = \mathbf{0}.$$

As described on page 19 we estimate $\mathrm{Cov}(\mathbf{y}_i - \boldsymbol{\mu}_i, \mathbf{h}_i)$ by $(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{h}_i)^T$. Further, by multiplying $n(\mathbf{D}_t\mathbf{Z})$ from the left we get

$$\begin{aligned}\mathbf{0} &= \sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_i - (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{h}_i)^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}) \\ &= \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i\mathbf{h}_i^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}) - n(1 - \bar{\mathbf{h}}^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}})\boldsymbol{\mu},\end{aligned}$$

because for block invariant designs the $\boldsymbol{\mu}_i$ are equal for all $i = 1, \ldots, n$, so $\boldsymbol{\mu}_i = \boldsymbol{\mu}$. Simple rearranging yields

$$\frac{1}{1 - \bar{\mathbf{h}}^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}} \cdot \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i\mathbf{h}_i^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}) = \boldsymbol{\mu} = \mathbf{h}(\mathbf{Z}\boldsymbol{\beta}).$$

By inverting $\mathbf{h}$ and then $\mathbf{Z}$ we finally derived the estimator

$$\hat{\boldsymbol{\beta}}_{ex} = \mathbf{Z}^{-1}\mathbf{h}^{-1}\left(\frac{1}{1 - \bar{\mathbf{h}}^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}}} \cdot \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i\mathbf{h}_i^T\boldsymbol{\Omega}_h^{-1}\bar{\mathbf{h}})\right),$$

reported on page 20. □

C: Proof of Proposition 2

Here we present the proof of Proposition 2:

Proof. Let $\mathbf{G}_r$ be the matrix $\mathbf{G}$ from the proof of Proposition 1 evaluated at $\tilde{\boldsymbol{\beta}}_2$. Since $\hat{\boldsymbol{\beta}}_2$ is assumed to be a solution of the first-order conditions, it follows that $\bar{\mathbf{m}}_{GEE}(\hat{\boldsymbol{\beta}}_2) - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \bar{\mathbf{h}} = \mathbf{0}$. Hence the conditions of Lemma 1 and Theorem 1 of Jann (2024) are fulfilled and by applying them and inserting $\mathbf{G}_r^{-1}\mathbf{G}_r$, we get

$$
\begin{aligned}
D_{RU} &= n\big(\bar{\mathbf{g}}(\tilde{\boldsymbol{\beta}}_2)^T \mathbf{W}_n(\tilde{\boldsymbol{\beta}}_2)\bar{\mathbf{g}}(\tilde{\boldsymbol{\beta}}_2) - \bar{\mathbf{g}}(\hat{\boldsymbol{\beta}}_2)^T \mathbf{W}_n(\hat{\boldsymbol{\beta}}_2)\bar{\mathbf{g}}(\hat{\boldsymbol{\beta}}_2)\big) \\
&= n(\bar{\mathbf{m}}_{GEE}(\tilde{\boldsymbol{\beta}}_2) - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \bar{\mathbf{h}})^T (\boldsymbol{\Omega}_m - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \boldsymbol{\Omega}_r)^{-1} (\bar{\mathbf{m}}_{GEE}(\tilde{\boldsymbol{\beta}}_2) - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \bar{\mathbf{h}}) \\
&= n(\bar{\mathbf{m}}_{GEE}(\tilde{\boldsymbol{\beta}}_2) - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \bar{\mathbf{h}})^T \mathbf{G}_r^{-1} \mathbf{G}_r (\boldsymbol{\Omega}_m - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \boldsymbol{\Omega}_r)^{-1} \mathbf{G}_r \mathbf{G}_r^{-1} \\
&\quad \times (\bar{\mathbf{m}}_{GEE}(\tilde{\boldsymbol{\beta}}_2) - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \bar{\mathbf{h}}) \\
&= (\mathbf{G}_r^{-1} \bar{\mathbf{m}}_{GEE}(\tilde{\boldsymbol{\beta}}_2) - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \bar{\mathbf{h}})^T (\frac{1}{n} \mathbf{G}_r^{-1} (\boldsymbol{\Omega}_m - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \boldsymbol{\Omega}_r) \mathbf{G}_r^{-1})^{-1} \\
&\quad \times (\mathbf{G}_r^{-1} \bar{\mathbf{m}}_{GEE}(\tilde{\boldsymbol{\beta}}_2) - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \bar{\mathbf{h}}) \\
&= (\mathbf{G}_r^{-1} \bar{\mathbf{m}}_{GEE}(\tilde{\boldsymbol{\beta}}_2) - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \bar{\mathbf{h}})^T (\widehat{\mathrm{Var}}(\tilde{\boldsymbol{\beta}}_2))^{-1} (\mathbf{G}_r^{-1} \bar{\mathbf{m}}_{GEE}(\tilde{\boldsymbol{\beta}}_2) - \boldsymbol{\Omega}_r^T \boldsymbol{\Omega}_h^{-1} \bar{\mathbf{h}}).
\end{aligned}
$$

The rest of the proof follows the same arguments as in the proof of Proposition 1 about Kronecker products. Note that $(\mathbf{D}_t\mathbf{Z})^{-1})$ then cancels out. Finally, the above expression is only a function of $\tilde{\boldsymbol{\beta}}_2$, not of ϕ and $\boldsymbol{\alpha}$. □

D: Nullspace method for constraint optimization

First, we follow the nullspace approach of Zörnig (2014, pp. 188–189). In our case, the linear constraints are $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. Using basic results from linear algebra, the feasible region can be easily described. Let $\mathbf{N}$ be a matrix whose columns are the basis of the nullspace, then the nullspace can be expressed by $\mathbf{N}\boldsymbol{\gamma}$ with $\gamma \in \mathbb{R}^{q-\text{rank}(R)}$. Thus, every feasible value can be expressed by $\boldsymbol{\beta} = \mathbf{N}\boldsymbol{\gamma} + \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ is a specific solution of $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, for example $\boldsymbol{\beta}^* = \mathbf{R}^T(\mathbf{R}\mathbf{R}^T)^{-1}\mathbf{r}$. Now we substitute this expression into the objective function of the GMM, which leads to an unconstrained optimization problem in $\boldsymbol{\gamma}$. After this reduction of variables, the (reduced) gradient and the Hessian matrix have to be derived again by applying the chain rule (Zörnig, 2014, p. 189). Consequently, the first-order conditions are

$$\mathbf{0} = \mathbf{N}^T(\nabla_\beta \bar{\mathbf{g}}(\boldsymbol{\beta}))^T \mathbf{W}_n \bar{\mathbf{g}}(\boldsymbol{\beta}) = \mathbf{N}^T(\nabla_\beta \bar{\mathbf{g}}(\mathbf{N}\boldsymbol{\gamma} + \boldsymbol{\beta}^*))^T \mathbf{W}_n \bar{\mathbf{g}}(\mathbf{N}\boldsymbol{\gamma} + \boldsymbol{\beta}^*)$$

and the Hessian matrix is $\mathbf{H} = \mathbf{N}^T(\nabla_\beta \bar{\mathbf{g}}(\mathbf{N}\boldsymbol{\gamma} + \boldsymbol{\beta}^*))^T \mathbf{W}_n (\nabla_\beta \bar{\mathbf{g}}(\mathbf{N}\boldsymbol{\gamma} + \boldsymbol{\beta}^*))\mathbf{N}$. Second, to find the restricted estimate, Fisher scoring can be used, as by Spiess and Hamerle (1996), which leads to the iterative procedure of choosing a starting value $\hat{\boldsymbol{\gamma}}_0$ and computing

$$\hat{\boldsymbol{\gamma}}_{j+1} = \hat{\boldsymbol{\gamma}}_j + \mathbf{H}^{-1}\mathbf{N}^T(\nabla_\beta \bar{\mathbf{g}}(\mathbf{N}\boldsymbol{\gamma} + \boldsymbol{\beta}^*))^T \mathbf{W}_n \bar{\mathbf{g}}(\mathbf{N}\boldsymbol{\gamma} + \boldsymbol{\beta}^*)$$

until a convergence criterion is met. Similar arguments as in (the proof of) Proposition 1 can be applied to this procedure, leading to a cancellation of ϕ and $\boldsymbol{\alpha}$. Finally, $\hat{\boldsymbol{\gamma}}_0$ can always be chosen independently of ϕ and $\boldsymbol{\alpha}$. To keep the number of iterations as small as possible, a starting value $\hat{\boldsymbol{\gamma}}_0$ is chosen, which is

heuristically close to the optimal value. To do this, we computed the unconstrained one-step estimator and projected it into the hyperplane defined by $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, resulting in $\hat{\boldsymbol{\gamma}}_0 = (\mathbf{N}^T\mathbf{N})^{-1}\mathbf{N}^T(\hat{\boldsymbol{\beta}}_{ex} - \boldsymbol{\beta}^*)$, so that $\mathbf{N}\hat{\boldsymbol{\gamma}}_0 + \boldsymbol{\beta}^*$ satisfies the constraints.

E: Further details of the simulation studies

E.1 Calculation of two I_{ex} for count data

As a first, conservative approach, we computed the range of means of the number of errors over all factor combinations where color was presented as background (8 in total). As a second, liberal approach, we assumed that the mean number of errors in Experiment 3 would be at least as high as in Experiment 1 or 2, so as a lower bound we used the maximum mean number of errors in the condition where color and response are repeated and color is the background across both experiments. As an upper bound, we used the minimum of the means of all conditions where exactly one of response or color is changed and color is the background. This assumption is consistent with the theory provided in the introduction by Schmalbrock and Frings (2022), that if only some but not all of the information is repeated, the entire event file is activated, causing a cognitive cost because the file encodes repetition but one aspect is not repeated.

While calculating the external intervals, we noticed that both datasets contained too many rows (they should contain only 14400 rows each). Further investigation showed that the extra rows were exact copies of the previous rows (which can be easily seen from the participant index), so we deleted them. As a result, we could not exactly replicate the results of Experiments 1 and 2, there were differences, but they were very small and did not change the conclusions.

E.2 Further details on the count data simulation

First, we describe our implementation of the approach of Barbiero and Ferrari (2015) to generate correlated poisson. The idea is to independently sample a

participant-specific poisson variable (similar to a random intercept) with parameter λ_0 and condition-specific poisson variables with parameter λ_k for $k = 1, ..., 16$ so that their means add up to the computed means or the null model means, respectively. Thus, the sum of each condition-specific variable with the person-specific variable is Poisson distributed with the pre-specified means, and there are positive correlations between conditions. For the null model, the correlations are all $\lambda_0/(\lambda_0 + \lambda_1)$, imposing an equicorrelation structure. We specified $\lambda_0 = 0.5381$ so that the correlations are 0.3, which is approximately the mean correlation in the data. For the alternative model, the same λ_0 was used, and the condition-specific means were derived by subtracting λ_0 from the calculated means.

Second, during the first test simulation, D_{RU} showed type I error rates of 1 when no external variance was provided. A closer look at the generalized matrix inversions that occur in the restricted estimation revealed that there seemed to be a numerical problem with the iterated matrix multiplication and the function `ginv` in the R package *MASS* (Venables & Ripley, 2002) in our case. This was indicated by the fact that analytically derived generalized inverses and the ones calculated with `ginv` differed substantially. We thus used our analytically derived generalized inverses.

E.3 Approximation of $\mathbf{I}_{ex}$ for ordinal data

Our approach to approximating an external interval for the frequencies of categories 1 and $1 \cup 2$ is inspired by Chapter 4.1 from Weichselberger (2001), and our solutions will be F-probabilities. Table 4 in Sudeck and Pfeifer (2016) reports means, standard deviations, skewnesses, excess kurtosis, and item-test correlations for the four items of the PAHCO Physical activity specific mood

regulation scale, as well as the scale mean. Cronbach's alpha is also reported, but we have omitted it because it only encodes the relationship between the item and scale mean variances, and these are all given. Unfortunately, no other measures, especially frequencies, are reported. Based on this information, we want to approximate the frequencies for our category 1, i.e. scale means in $[1, 2)$, as well as the frequencies for category $1 \cup 2$, i.e. scale means in $[1, 3)$. Since $1, 2, 3$ and 4 are all the possible values for the four items, the joint frequency distribution has $4^4 = 256$ cells / dimensions. The frequencies of certain scale mean values, as well as those of category 1 and category $1 \cup 2$, can be computed as sums of subsets of these cell frequencies. Furthermore, all the information given in Table 4 are linear functions in the cell frequencies, which leads to the following linear system for each item as well as for the mean score (substituting the cell means $\bar{x}_i$ for x_i, the grand mean $\bar{x}$ for the item specific means $\bar{x}_j$, as well as using the descriptive values for the scale mean value and omitting the last row)

$$\begin{pmatrix} x_1 & \dots & x_{256} \\ (x_1 - \bar{x}_j)^2 & \dots & (x_{256} - \bar{x}_j)^2 \\ (x_1 - \bar{x}_j)^3 & \dots & (x_{256} - \bar{x}_j)^3 \\ (x_1 - \bar{x}_j)^4 & \dots & (x_{256} - \bar{x}_j)^4 \\ (x_1 - \bar{x}_j)(\bar{x}_1 - \bar{x}) & \dots & (x_{256} - \bar{x}_j)(\bar{x}_{256} - \bar{x}) \end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_{256} \end{pmatrix} = \begin{pmatrix} \bar{x}_j \\ \frac{n-1}{n} s_j^2 \\ skew_j \\ kurt_j + 3 \\ \frac{n-1}{n} r_{jt} s_j^2 s^2 \end{pmatrix}, \tag{1}$$

where s^2 is the sample variance of the scale mean value, s_j^2 is the sample variance, $skew_j$ is the empirical skewness coefficient, $kurt_j$ is the empirical excess kurtosis, and r_{jt} is the item-test correlation of item j, respectively. We

have assumed, that the values reported in Table 4 in Sudeck and Pfeifer (2016) are equal to these statistics. If this is not the case (for example, when bias-corrected skewness was used), our approximation may be biased. However, sample size was 1028, so the error should not be too large. Combined with the typical frequency constraints $0 \leq f_i \leq 1$ and $\sum_{i=1}^{256} f_i = 1$, we have a set of linear constraints and thus the following type of linear optimization problems:

$$
\begin{aligned}
&\text{Find a vector } (f_1, \ldots, f_{256}) \\
&\text{that minimizes } \sum_{i \in \text{category 1}} f_i \\
&\text{subject to equation (1) for the mean score and for } j = 1, 2, 3, 4 \\
&\text{and } 0 \leq f_i \leq 1, \sum_{i=1}^{256} f_i = 1.
\end{aligned}
$$

This results in four programs by minimizing and maximizing for category 1 and for category $1 \cup 2$. Unfortunately, all reported values are rounded to two decimal places, so we have to account for rounding errors. Therefore, $\bar{x}_j, s_j, skew_j, kurt_j, r_{jt}$ must be thought of as intervals. For the lower bound we subtracted 0.006, since values like 1.2445 could be represented by the system as 1.245 and thus reported as 1.25, so 0.006 is a safe lower bound. For the upper bound, we added 0.005, since this is a value where one would round to the next number on the second digits. We also took into account the fact that the mean of the item scores should be equal to the mean of the scale scores. Using the lower and upper bounds of the item means, the upper bound of the mean of the scale means could be reduced to adding 0.0025 (no higher value is possible based on the upper bounds of the item means). Inserting the calculated intervals into equation (1) leads to the problem that both the left and right

sides of the matrix contain interval entries. If only the right-hand side contained intervals, one could interpret equation (1) as a set of inequalities, i.e., the left-hand side is less than or equal to the upper bound of the right-hand side and greater than or equal to the respective lower bound. To achieve this, we transformed equation (1) so that there are no item or grand means in the matrix on the left side, resulting in

$$\begin{pmatrix} x_1 & \dots & x_{256} \\ x_1^2 & \dots & x_{256}^2 \\ x_1^3 & \dots & x_{256}^3 \\ x_1^4 & \dots & x_{256}^4 \\ x_1\bar{x}_1 & \dots & x_{256}\bar{x}_{256} \end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_{256} \end{pmatrix} = \begin{pmatrix} \bar{x}_j \\ \frac{n-1}{n}s_j^2 + \bar{x}_j^2 \\ skew_j s_j^3 + 3\bar{x}_j s_j^2 + \bar{x}_j^3 \\ (kurt_j + 3)s_j^4 + 4skew_j s_j^3 \bar{x}_j + 6s_j^2 \bar{x}_j^2 + \bar{x}_j^4 \\ \frac{n-1}{n} r_{jt} s_j^2 s^2 + \bar{x}_j \bar{x} \end{pmatrix}. \tag{2}$$

Based on the rounding intervals, we calculated lower and upper bounds on the right side. We did this entry by entry, ignoring the fact that the same values occur in multiple entries. However, this conservative approach only leads to slightly wider intervals than necessary, since the expressions are mostly dominated by the powers of $\bar{x}_j$. We implemented the four linear programs in the statistical software R (R Core Team, 2023), using equation (2) instead of (1). The implementation can be found in the R scripts in the electronic supplementary material. To find the objective values of the four linear programs, we used the package *linprog* (Henningsen, 2022). Unfortunately, no feasible solutions were found for the programs. The reason for this may be incorrectly reported values or incorrect assumptions about the formulas used for the reported values (empirical vs. bias-corrected). To resolve this, we tested

whether relaxing a single inequality constraint (except those that are mandatory for frequencies) by adding or subtracting it by 1000 leads to feasible solutions. This was the case for the upper bounds of the item-test correlations (except the one for item 1) as well as the lower bound of the standard deviation of the scale mean. For these four constraints, we tested their effects more precisely by repeatedly adding/subtracting 0.01, relaxing them successively, and stopping the first time a feasible range could be found. For the upper bounds of the item-test correlations, the resulting intervals for category 1 frequency were quite wide, being $[0.124, 0.442]$, $[0.118, 0.446]$, and $[0.117, 0.446]$, while for the lower bound of the standard deviation, the result was $[0.238, 0.362]$. Since the latter was consistent with the other solutions, reasonably large, and the shortest interval, we used it as an approximation for the external interval. The same relaxed lower standard deviation constraint was used for the category $1 \cup 2$, resulting in the external interval $[0.692, 0.806]$. While reducing the constant 0.01 for successive subtraction would lead to narrower intervals, we decided to leave it at 0.01 to keep the intervals reasonably large, reflecting the uncertainty in our approximation approach.

E.4 Further details on the categorical data simulation

Here were describe the details of the threshold model based on a joint normal distribution that we used to generate correlated categorical data. We calculated the mean correlations of PAHCO scores between measurements over the 100 imputed datasets, separately for each group. The resulting 6 mean correlations were used as covariances of the joint normal, with variances equal to 1 and a mean of 0, so that the marginal distributions were all standard normal. The correlations between the groups were set to 0 to reflect the independence of the

two groups. The 12 mean frequencies computed earlier were used to compute quantiles of the standard normal distribution, 2 for each factor combination, so that the marginal standard normal distribution of that factor combination is divided into 3 regions with probability equal to the respective two frequencies of the factor combination, the lower region representing category 1, the middle region representing category 2, and the upper region representing category 3.

References

Barbiero, A. & Ferrari, P. A. (2015). Simulation of correlated poisson variables. *Applied Stochastic Models in Business and Industry*, *31*(5), 669–680. https://doi.org/10.1002/asmb.2072

Henningsen, A. (2022). *Linprog: Linear programming / optimization* [R package version 0.9-4]. https://CRAN.R-project.org/package=linprog

Jann, M. (2024). Testing the fit of data and external sets via an imprecise sargan-hansen test. *International Journal of Approximate Reasoning*, 109214. https://doi.org/10.1016/j.ijar.2024.109214

Jann, M. & Spiess, M. (2024). Using external information for more precise inferences in general regression models. *Psychometrika*, 1–22. https://doi.org/10.1007/s11336-024-09953-w

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.2307/2336267

Puntanen, S., Styan, G. P. H. & Isotalo, J. (2011). *Matrix tricks for linear statistical models*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10473-2

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Schmalbrock, P. & Frings, C. (2022). A mighty tool not only in perception: Figure-ground mechanisms control binding and retrieval alike. *Attention, Perception, & Psychophysics*, *84*(7), 2255–2270. https://doi.org/10.3758/s13414-022-02511-5

Spiess, M. & Hamerle, A. (1996). On the properties of gee estimators in the presence of invariant covariates. *Biometrical Journal*, *38*(8), 931–940. https://doi.org/10.1002/bimj.4710380805

Sudeck, G. & Pfeifer, K. (2016). Physical activity-related health competence as an integrative objective in exercise therapy and health sports - conception and validation of a short questionnaire. *German Journal of Exercise and Sport Research*, *46*(2), 74–87. https://doi.org/10.1007/s12662-016-0405-4

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth) [ISBN 0-387-95457-0]. Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Weichselberger, K. (2001). *Elementare grundbegriffe einer allgemeineren wahrscheinlichkeitsrechnung i: Intervallwahrscheinlichkeit als umfassendes konzept* (Vol. 1). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-57583-9

Zörnig, P. (2014). *Nonlinear programming*. De Gruyter. https://doi.org/10.1515/9783110315288

Danksagung

Ich danke meinem Doktorvater Prof. Dr. Martin Spieß zutiefst für sein Vertrauen, seine vielen guten Ideen, die vielen mir gestatteten Freiheiten und für die Zeit, die er für mich aufgewandt hat.

Ich danke Prof. Dr. Thomas Augustin, Prof. Dr. Alexander Redlich, Prof. Dr. Jenny Wagner und Prof. Dr. Mike Wendt für ihre Arbeit im Rahmen meines Promotionsprüfungsausschusses.

Ich danke meinen Eltern Norbert und Gisela Jann für all die Unterstützung über so lange Zeit, den Glauben an mich und dass sie mir ermöglicht haben meine akademischen Wege gehen zu können.

Ich danke meiner besten Freundin Nell Kindt dafür, dass sie mich durch ihre Reflektiertheit, Ehrlichkeit und Lösungsorientierung stets unterstützt hat und mir die richtigen Anstöße zur Bewältigung meiner persönlichen Krise während der Promotionszeit gab.

Ich danke Dr. Marlena Mayer und Leslie Förtsch dafür, dass sie mir ermöglicht haben, meinen Schreibstil und meine Schreibkompetenz in der englischen Sprache durch ihre Hilfe deutlich verbessern zu können.

Ich danke all meinen aktuellen und ehemaligen Kolleginnen und Kollegen aus beiden Methoden-Arbeitsbereichen, also Dr. Yasin Altinisik, Dr. Ingmar Böschen, Marcella Dudley, Luise Frappier, Prof. Dr. Simon Grund, Dr. Pascal Jordan, Mark Lustig, Christine Manor, Dennis Warnholtz, Dirk Werner und Helen Wright, für die schöne Zeit, die vielen netten, interessanten Gespräche und die tollen Weihnachtsfeiern.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR PSYCHOLOGIE UND
BEWEGUNGSWISSENSCHAFT
Institut für Bewegungswissenschaft
Institut für Psychologie

Erklärung gemäß *(bitte Zutreffendes ankreuzen)*

☐ **§ 4 (1c) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**

☒ **§ 5 (4d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich,

Martin, Jann (Vorname, Nachname),

dass ich mich an einer anderen Universität oder Fakultät noch keiner Doktorprüfung unterzogen oder mich um Zulassung zu einer Doktorprüfung bemüht habe.

Hamburg, 02.06.2025 — Ort, Datum

Unterschrift

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR PSYCHOLOGIE UND
BEWEGUNGSWISSENSCHAFT
Institut für Bewegungswissenschaft
Institut für Psychologie

Eidesstattliche Erklärung nach *(bitte Zutreffendes ankreuzen)*

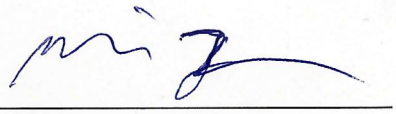
☐ **§ 7 (4) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**

☒ **§ 9 (1c und 1d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich an Eides statt,

1. dass die von mir vorgelegte Dissertation nicht Gegenstand eines anderen Prüfungsverfahrens gewesen oder in einem solchen Verfahren als ungenügend beurteilt worden ist.
2. dass ich die von mir vorgelegte Dissertation selbst verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die wörtlich oder inhaltlich übernommenen Stellen habe ich als solche kenntlich gemacht.

Hamburg, 02.06.2025
Ort, Datum


Unterschrift