Applications in High-Dimensional Econometrics

Cumulative Dissertation
to obtain the academic degree of a “doctor rerum oeconomicarum” (Dr. rer. oec.) according to doctoral degree regulations 2014

at the Faculty of Business Administration (Hamburg Business School)
Moorweidenstr. 18
20148 Hamburg (Germany)
of Universität Hamburg

submitted by: Philipp Simeon Bach
born on May 11, 1990 in Singen (Hohentwiel)

Hamburg, 2021

Thesis Committee:
Chairman: Prof. Dr. Knuth Haase
1st Examiner: Prof. Dr. Martin Spindler
2nd Examiner: Prof. Dr. Michael Merz
3rd Examiner: Prof. Dr. Anthony Strittmatter

Date of Disputation: May 17, 2021

Acknowledgement

I would never have been able to finish this doctoral thesis without the help of others.

First of all, I would like to say thank you to Martin Spindler who has been a great supervisor during the last four years. On the one hand, he always provided me the structure that I appreciated to pursue our joint research projects. On the other hand, he gave me all the freedom to do what I wanted and what I was interested in. It has been a lot of fun to work with such a creative and curious researcher who is permanently developing new ideas.

Second, I would like to thank my colleagues and coauthors Sven Klaaßen, Jannis Kück, Malte Kurz, and Zihao Yuan. I have learned a lot from them, in particular in terms of all the theoretical details in statistics and in terms of coding techniques. Working on joint projects has always been a lot of fun and I am really happy that all of us contributed to an atmosphere of collaboration and curiosity at the chair of statistics. Moreover, I am grateful to Cornelia Hartwig and Elke Thoma for helping me with the complicated administrative steps associated with a Ph.D. I am thankful for the support by the Hamburg Business School that set up an effective way to support young researchers like my colleagues and me. Thanks to this support, I was able to participate in many great workshops and conferences.

Furthermore, I would like to thank Victor Chernozhukov for the exciting and unique opportunity to collaborate on several research projects. It has been a great experience and an honor to get a little bit of an insight to his high standards of academic research. I'm grateful to Michael Merz and Anthony Strittmatter for agreeing to act as second and third examiner. Moreover, I would like to thank many more people in academia who had an impact on me during my studies at the University of Mannheim, the University of Kopenhagen, the LMU in Munich, the University of Hamburg as well as at the ZEW, the MEA, and the ifo institute.

I would like to thank my parents for their love and support and my siblings for their love, kindness and courage. Moreover, I am grateful to Herr Huber for his support during my studies, and to Crisi and Carlos for being such lovable persons. I would like to say thank you to all my friends who have always been supporting me in what I do although (or because) they never really wanted to know what it actually was that I've been doing during the last years.

Furthermore, I would like to thank my wonderful wife Desi for simply being the greatest person on earth. Ever since we have met, she has always been supporting me in everything I do. Every single day, Desi shows me the meaning of love and joy of life.

Contents

List of Figures . . . V
List of Tables . . . VII

1 General Introduction 1
1.1 Background and Intuition . . . 1
1.2 Conceptual Framework . . . 3
1.3 Outline . . . 5

2 DoubleML - An Object-Oriented Implementation of Double Machine Learning in R 7
2.1 Introduction . . . 7
2.2 Getting started . . . 8
2.2.1 Installation . . . 8
2.2.2 A Motivating Example: Basics of Double Machine Learning . . . 9
2.3 Key Causal Models . . . 10
2.3.1 Partially Linear Regression Model (PLR) . . . 11
2.3.2 Partially Linear Instrumental Variable Regression Model (PLIV) . . . 11
2.3.3 Interactive Regression Model (IRM) . . . 12
2.3.4 Interactive Instrumental Variable Model (IIVM) . . . 13
2.4 Basic Idea and Key Ingredients of Double Machine Learning . . . 14
2.4.1 Basic Idea behind Double Machine Learning for the PLR Model . . . 14
2.4.2 Key Ingredients of the Double Machine Learning Inference Approach . . . 15
2.5 The Double Machine Learning Inference Method . . . 18
2.5.1 Double Machine Learning for Estimation of a Causal Parameter . . . 18
2.5.2 Methods for Simultaneous Inference . . . 22
2.6 Implementation Details . . . 23
2.6.1 Class Structure . . . 23
2.6.2 Data-Backend and Causal Model . . . 24
2.6.3 Learners, Parameters and Tuning . . . 24
2.6.4 Modifications and Extensions . . . 25
2.7 Estimation of Causal Parameters with `DoubleML`: Real-Data and Simulated Examples. . . 25
2.7.1 Initialization of the Data-Backend . . . 25
2.7.2 Initialization of the Causal Model . . . 26
2.7.3 Estimation of the Causal Parameter in a PLR Model . . . 27
2.7.4 Estimation of the Causal Parameter in an IRM Model . . . 28
2.7.5 Simultaneous Inference in a Simulated Data Example . . . 29
2.7.6 Learners, Parameters and Tuning . . . 31
2.7.7 Specifications and Modifications of Double Machine Learning . . . 34
2.7.8 A Short Simulation Study . . . 37

2.8 Conclusion . . . 40
2.9 Appendix . . . 41
    2.9.1 Computation and Infrastructure . . . 41
    2.9.2 Suppressed Code Output . . . 41
    2.9.3 Additional Data Generating Processes, Simulation Study . . . 49

3 Valid Simultaneous Inference in High-Dimensional Settings with the hdm Package for R 51
3.1 Introduction . . . 51
3.2 Setting . . . 52
3.3 Simultaneous Post-Selection Inference in High Dimensions . . . 53
    3.3.1 Simultaneous Inference based on Double Selection . . . 53
    3.3.2 Simultaneous Inference based on Knockoffs . . . 54
    3.3.3 Summary: Simultaneous Post-Selection Inference . . . 56
3.4 Methods for Simultaneously Testing Multiple Hypotheses . . . 56
    3.4.1 Multiple Hypotheses Testing with Control of the Familywise Error Rate . . . 56
    3.4.2 Multiple Hypotheses Testing with Control of the False Discovery Rate . . . 58
    3.4.3 A Global Test for Joint Significance with Lasso Regression . . . 59
3.5 Simulation Study . . . 59
    3.5.1 Simulation Setting . . . 59
    3.5.2 Results . . . 60
3.6 Simultaneous Inference in a Real-Data Example with R . . . 63
    3.6.1 Implementation in the R Package `hdm` . . . 63
    3.6.2 A Real-Data Example for Simultaneous Inference in High Dimensions - The Gender Wage Gap Analysis . . . 64
3.7 Conclusion . . . 70
3.8 Appendix . . . 71
    3.8.1 Computation and Infrastructure . . . 71
    3.8.2 Additional Results, Simulation Study . . . 71
    3.8.3 Additional Results, Real-Data Example . . . 72

4 Uniform Inference in High-Dimensional Additive Models 76
4.1 Introduction . . . 76
    4.1.1 Organization of the Paper . . . 78
    4.1.2 Notation . . . 78
4.2 Setting . . . 79
    4.2.1 Motivation and Illustration . . . 79
    4.2.2 Formal Setting . . . 82
4.3 Estimation . . . 84
4.4 Main Results . . . 86
4.5 Simulation Results . . . 90
4.6 Illustration in a Real-Data Example . . . 96
4.7 Proofs . . . 98
4.8 Uniformly valid confidence bands . . . 117
4.9 Uniform nuisance function estimation . . . 122
    4.9.1 Uniform lasso estimation . . . 122
4.10 Computational Details . . . 131

4.10.1 Computation and Infrastructure . . . 131
4.10.2 Simulation Study: Smoothing Parameters in B-splines . . . 131
4.10.3 Empirical Application: Cross-Validation Procedure for Choice of Smoothing Parameter . . . 131
4.10.4 Empirical Application: Additional Plots for Explanatory Variables . . . 133

5 Heterogeneity in the U.S. Gender Wage Gap 134
5.1 Introduction . . . 134
5.2 Literature and Evidence on the Gender Wage Gap . . . 135
5.2.1 Empirical Methods in the Context of the Gender Wage Gap . . . 135
5.2.2 Literature Review: The Gender Wage Gap and Recent Developments . . . 136
5.3 An Econometric Model of a Heterogeneous Gender Wage Gap . . . 138
5.3.1 Valid Post-Selection Inference in High Dimensions . . . 138
5.4 Heterogeneity in the U.S. Gender Wage Gap . . . 140
5.4.1 Overview of the 2016 ACS data . . . 140
5.4.2 Descriptive Statistics . . . 141
5.5 Results . . . 146
5.5.1 Significant Variables for the U.S. Gender Wage Gap . . . 146
5.5.2 Heterogeneity in the Coefficient Function $\beta(x_i)$ . . . 150
5.5.3 A Robustness Check . . . 152
5.6 Conclusion . . . 152
5.7 Appendix . . . 154
5.7.1 Relation to Oaxaca-Blinder Decomposition . . . 154
5.7.2 Interpretation of $\beta(x_i)$ . . . 154
5.7.3 Additional Descriptive Statistics . . . 154
5.7.4 Additional Results . . . 162

6 Insights from Optimal Pandemic Shielding in a Multi-Group SEIR Framework 181
6.1 Introduction . . . 181
6.2 Multi-Group SEIR Model . . . 183
6.2.1 Model Assumptions . . . 184
6.2.2 Dynamics in the MG-SEIR Model . . . 187
6.2.3 Efficient Frontier . . . 188
6.3 Specification and Calibration . . . 188
6.3.1 Country-Specific Parameters . . . 189
6.3.2 Parameters for the SEIR Model . . . 191
6.3.3 Calibrating R_0 . . . 191
6.4 Results and Optimal Policies . . . 191
6.4.1 Efficient Frontier . . . 191
6.4.2 Optimal Policy . . . 191
6.4.3 Implementation of Optimal Policies . . . 209
6.5 Conclusion . . . 210
6.6 Appendix . . . 211

7 Conclusive Summary 218

Appendices **227**

A.1 Statement of Personal Contribution Pursuant to §6(4) PromO 228
A.2 Short Summary of Papers Pursuant to §6(6) PromO 230
A.3 List of Publications Pursuant to §6 (6) PromO 234

List of Figures

1.1 Naive variable selection based on t-statistics and ordinary least squares, simulation example. 2
1.2 Naive inference based on the lasso and random forests, simulation example. 3
1.3 Double machine learning with lasso and random forests, simulation example. 5

2.1 Causal diagram for PLR and IRM. 11
2.2 Causal diagram for PLIV and IIVM. 12
2.3 Performance of non-orthogonal and orthogonal estimators in simulated data example. . . 13
2.4 Performance of orthogonal estimators based on full sample and sample splitting in simulated data example. 18
2.5 Class structure of the DoubleML package for R. 24
2.6 Illustration of efficiency gains due to the use of cross-fitting. 38
2.7 Histogram of double machine learning estimators in key causal models. 39

3.1 Regression coefficients, simulation setting with $K = 60$. 60
3.2 Average wage gap in industries, ACS 2016. 65

4.1 Illustration of the estimator and confidence bands, simulation example. 81
4.2 Average confidence bands, $f_1(x_1)$, simulation study. 92
4.3 Average confidence bands, $f_2(x_2)$, simulation study. 93
4.4 Average confidence bands, $f_3(x_3)$, simulation study. 94
4.5 Average confidence bands, $f_4(x_4)$, simulation study. 95
4.6 Effects and confidence bands, Boston housing data example. 97
4.7 Additional results, Boston housing example. 133

5.1 Usual hours worked per week by gender. 143
5.2 Unconditional and conditional female-to-male mean (log) wage ratios. 144
5.3 Quantiles of effects with simultaneous confidence bands. 151
5.4 Share of women within occupational, industry and college major categories. 156
5.5 Distribution of women across occupational, industry and college major categories. 157
5.6 Income according to marital status, race and ethnicity. 158
5.7 Income according to English language ability, veteran status and region. 159
5.8 Income according to industry and occupation. 160
5.9 Income according to college major. 161
5.10 Selected results, high school degree subgroup. Marginal effects with joint 0.95-confidence bands. 163
5.11 Selected results, bachelor's degree subgroup. Marginal effects with joint 0.95-confidence bands. 164
5.12 Quantiles of effects with corresponding confidence bounds, $c = 0.5$. 180

6.1 Compartments in a two-group SEIR model. 184
6.2 Efficient frontier in SEIR model, baseline setting. 192
6.3 Optimal uniform shielding policy, baseline setting. 193
6.4 Share of uninfected and reproduction rate, baseline setting. 194
6.5 Policy frontiers with group distancing or reduced transmission rates. 195
6.6 Optimal policy with safety focus and group distancing, uniform shielding. 196
6.7 Share of uninfected and reproduction rate with group distancing, uniform shielding. . . . 197
6.8 Optimal policy with safety focus and group distancing, semi-targeted shielding. 198
6.9 Share of uninfected and reproduction rate group distancing, semi-targeted shielding. . . . 199
6.10 Policy frontiers with alternative contact rate adjustment. 200
6.11 Policy frontiers with improved testing and isolation. 202
6.12 Policy frontiers with improved conditions for working from home. 203
6.13 Policy frontiers and optimal policy with combination of policy measures. 204
6.14 Share of uninfected and reproduction rate with combination of policy measures 205
6.15 Policy frontier with improved medical treatment. 206
6.16 Optimal shielding policy with a vaccine arrival after one year. 207
6.17 Optimal shielding policy with a vaccine arriving after six months. 208
6.18 Policy frontiers, variation of the parameter λ and the ICU capacity constraint. 212
6.19 Optimal policy with combination of policy measures, uniform shielding. 213
6.20 Share of uninfected and reproduction rate with combination of policy measures, uniform shielding. 214
6.21 Policy frontiers, combination of improved medical treatment and group distancing. 215
6.22 Optimal policy, combination of improved medical treamtent and uniform group distancing policy. 216
6.23 Optimal policy, combination of improved medical treamtent and group distancing towards the vulnerable. 217

List of Tables

2.1 Family-wise error rate and average number of correct rejections in a simulation example. . 40

3.1 Simulation results, $R = 1000$ repetitions. 62
3.2 Number of rejected hypotheses, gender wage gap example. 68
3.3 Number of repetitions with selection of zero variables, Knockoffs. 71

4.1 Definitions of the data generating processes, simulation study. 91
4.2 Empirical coverage, simulation study. 91
4.3 List of variables, Boston housing data example. 96
4.4 Smoothing parameters, simulation study. 131
4.5 Smoothing parameters, Boston housing example. 132

5.1 Summary statistics. 142
5.2 List of variables. 145
5.3 Double lasso results. 147
5.4 Occupational effects, high school degree and bachelor's degree data, double lasso. 149
5.5 Complete double lasso results (1/3), high school degree data. 166
5.6 Complete double lasso results (2/3), high school degree data. 167
5.7 Complete double lasso results (3/3), high school degree data. 168
5.8 Complete double lasso results (1/4), bachelor's degree data. 169
5.9 Complete double lasso results (2/4), bachelor's degree data. 170
5.10 Complete double lasso results (3/4), bachelor's degree data. 171
5.11 Complete double lasso results (4/4), bachelor's degree data. 172
5.12 Complete double lasso results, lasso with $c = 0.5$ (1/3), high school degree data. 173
5.13 Complete double lasso results, lasso with $c = 0.5$ (2/3), high school degree data. 174
5.14 Complete double lasso results, lasso with $c = 0.5$ (3/3), high school degree data. 175
5.15 Complete double lasso results, lasso with $c = 0.5$ (1/4), bachelor's degree data. 176
5.16 Complete double lasso results, lasso with $c = 0.5$ (2/4), bachelor's degree data. 177
5.17 Complete double lasso results, lasso with $c = 0.5$ (3/4), bachelor's degree data. 178
5.18 Complete double lasso results, lasso with $c = 0.5$ (4/4), bachelor's degree data. 179

6.1 Parameters for the United States and Germany. 189
6.2 Hospital beds and ICU beds. 190

Chapter 1

General Introduction

1.1 Background and Intuition

One of the most significant developments to have occured in recent years has been the increasingly important role of data in the private and public sector: Leading high-tech companies base their business models on the collection and exploitation of big data sets. Furthermore, public organization and governments use empirical studies and methods to evaluate and optimize policy measures. In recent years, considerable progress has been made in terms of software and hardware that enable analysts and researchers to use data sets of an unprecedented volume, quality and structure. For example, the machine learning literature has developed methods to effectively process unstructured data like text or image data that can now be employed in statistical modeling.

Major advances in the machine learning literature have mainly been made in terms of prediction problems. For instance, the lasso estimator introduced by Tibshirani (1996) is a popular choice in high-dimensional linear models with excellent performance guarantees (Bickel et al., 2009, Bühlmann and van De Geer, 2011). In nonlinear models, regression trees and random forests are frequently used methods (Breiman, 2001). However, many important problems in social sciences and business are causal in nature: Policy programs, such as active labor market policies, should be cost-effective and efficient; managers in private companies have a great interest in finding optimal pricing or marketing strategies to maximize revenue and profits. In order to derive such an optimal policy rule, valid estimation of causal quantities is essential. For example, optimal pricing strategies require exact estimation of price elasticities. Thus, causal inference that exploits the powerful performance of state-of-the-art machine learning methods is greatly important. However, these modern estimation techniques cannot directly be used to estimate causal effects and, if they are not employed in a valid inference framework, may lead to substantially flawed conclusions and decisions.

In the past years, the statistical literature has resulted in several approaches for valid inference based on machine learning methods, for instance the double machine learning approach introduced in Belloni et al. (2014c) and Chernozhukov et al. (2018a) and the debiasing approach in Zhang and Zhang (2014) and van de Geer et al. (2014). Both approaches provide a framework to construct valid confidence intervals or test statistics based on machine learning methods. Moroever, recent work by Belloni et al. (2014a) and Belloni et al. (2018) that build on accompanying results and methods developed in Chernozhukov et al. (2013a) and Chernozhukov et al. (2014) make it possible to perform inference on high-dimensional vectors of unknown causal parameters. These results are highly useful, for example to perform inference on functionals or a possibly large number of causal effects as in the analysis of heterogeneous treatment effects.

In the following, a short example is introduced to shed light on the building blocks for valid inference

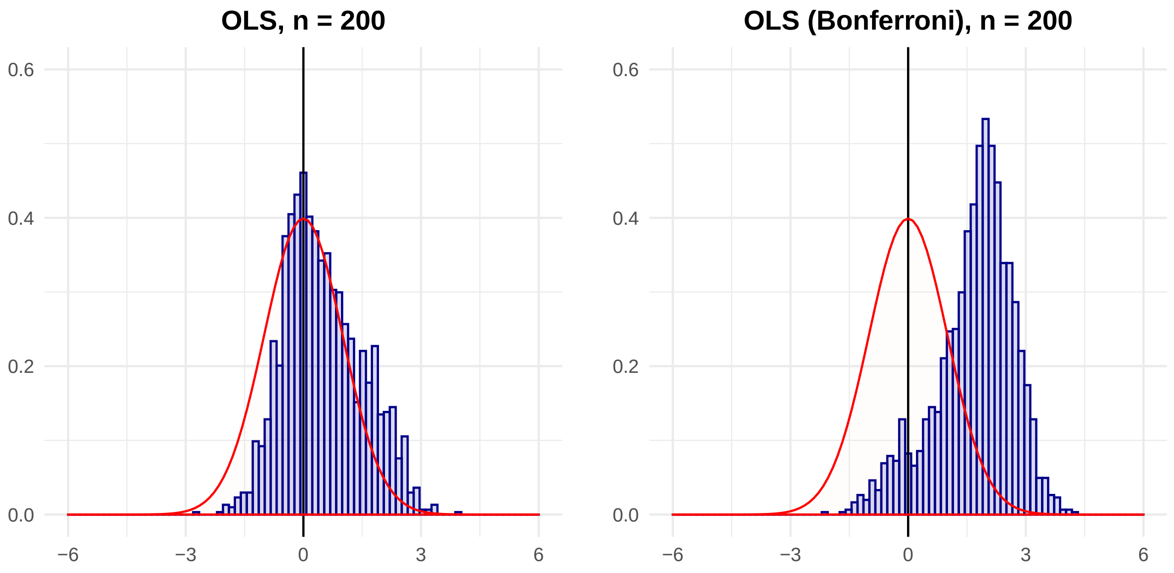


Figure 1.1: **Naive variable selection based on t-statistics and ordinary least squares, simulation example.**

The histograms illustrate the empirical distribution of studentized naive estimators that are based on ordinary least squares regression obtained in $R = 2000$ simulation replications. **Left panel:** Ordinary least squares regression after model selection based on t-statistics. Variables are excluded from the model if the null hypothesis $H_0 : \beta_j = 0$, with $j = 1, \ldots, p$, cannot be rejected at the 5% significance level. **Right panel:** Ordinary least squares regression after model selection based on t-statistics. Variables are excluded from the model if the null hypothesis $H_0 : \beta_j = 0$, with $j = 1, \ldots, p$, cannot be rejected at the 5% significance level with p-values being corrected according to the Bonferroni correction for multiple testing. In $R = 2000$ simulation repetitions, naively constructed confidence intervals for the estimators achieve an empirical coverage of 75% (left panel) and 30% (right panel).

using machine learning techniques. Suppose a researcher has access to a large number of covariates and is interested in estimation of the causal effect θ_0 in a partially linear regression model

$$Y = \theta_0 D + g_0(X) + \varepsilon, \tag{1.1}$$

with D being the treatment variable of interest and g_0 being an unknown function. In this model θ_0 measures the causal effect of the treatment variable D on the outcome Y, once it is controlled for the covariates X. We illustrate this regression example in a simulation with results shown in Figures 1.1 to 1.3. In the simulated example, there are $p = 120$ regressors X and the model is estimated on a sample with $n = 200$ observations.

In the following, we will shortly illustrate the invalidity of two different naive approaches. In the first naive procedure, a researcher performs an initial variable screening step that is based on t-tests as obtained from an ordinary least squares regression. Employing a linear specification, i.e., $g_0(X) = X'\beta$, all variables for which the null hypothesis $H_0 : \beta_j = 0$, with $j = 1, \ldots, p$, cannot be rejected at the 5% significance level are discarded from the model. After this selection step, an ordinary least squares regression is estimated and confidence intervals are constructed as if no variable selection was performed. Figure 1.1 illustrates that the corresponding estimator may not be asymptotically normally distributed and that the actually achieved empirical coverage of the confidence interval may substantially differ from the nominal coverage. The bias of the estimator is more pronounced with a stricter selection criterion being imposed in the first step, as for example by a Bonferroni correction (right panel of Figure 1.2).

A second naive procedure may involve the use of machine learning methods such as the lasso or random

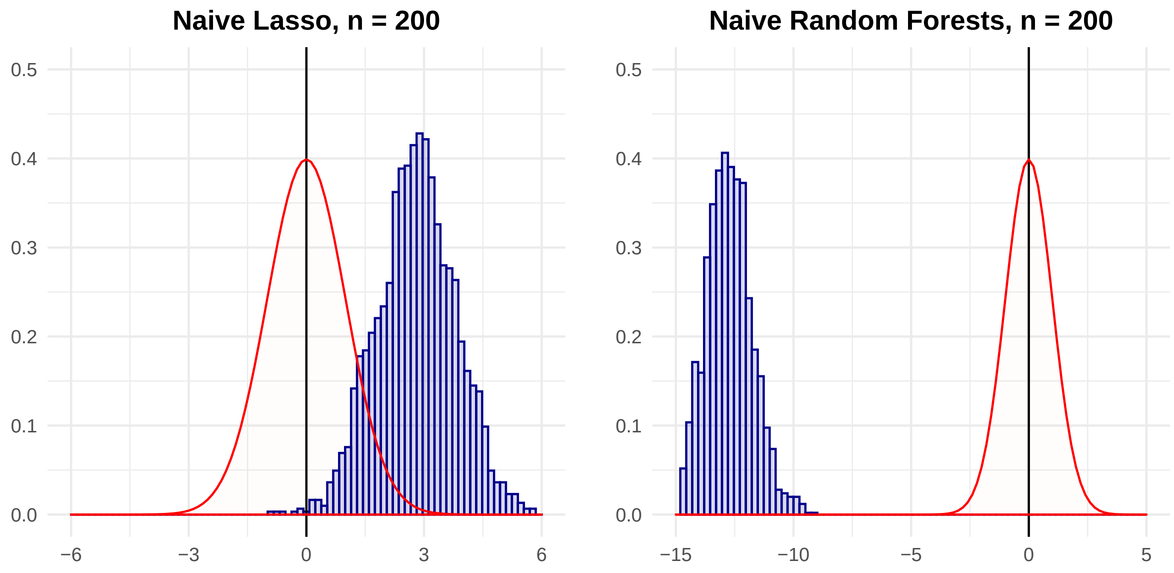


Figure 1.2: **Naive inference based on the lasso and random forests, simulation example.**

The histograms illustrate the empirical distribution of studentized naive estimators that are based on machine learning methods obtained in $R = 2000$ simulation replications. **Left panel:** A naive estimator that is based on a single selection step by lasso with cross-validated choice of the penalty. The variables that have been selected in this step are used in a linear regression model. **Right panel:** A naive estimator that is obtained if the unknown function $g_0(X)$ is estimated with a random forest learner. The empirical distribution is heavily biased and cannot be well-approximated by a normal distribution. In $R = 2000$ simulation repetitions, naively constructed confidence intervals for the estimators achieve an empirical coverage of 19% (left panel) and 0% (right panel).

forests. For example, the lasso is known to lead to sparse solutions and, hence, is attractive to perform variable selection. To perform inference, it may appear tempting to simply estimate ordinary least squares after the lasso selection to obtain confidence intervals and test statistics. The empirical distribution of such a naive estimator as obtained in the simulation example is illustrated in the left panel of Figure 1.2. The histogram shows that the estimator is severely biased implying that the empirical coverage is substantially lower than the nominal level.

Alternatively, a researcher may employ nonlinear machine learning methods such as random forests, for instance, if $g_0(X)$ is suspected to be nonlinear. Hence, the predictions from a random forest learner might simply be plugged in for $g_0(X)$. As can be observed in the right panel of Figure 1.2, the resulting estimator for θ_0 is heavily biased and any inferential statements that may be based on such an approach are likely to be invalid.

Contrarily to the previously presented approaches, estimation in the double machine learning framework, which is the basis for this dissertation, results in an asymptotically normally distributed estimator and, hence, makes it possible to construct valid confidence intervals and test statistics. Figure 1.3 illustrates the empirical distribution of double machine learning estimators as obtained for a lasso (left panel) and a random forest learner (right panel). In the following, we will review the key components of the double machine learning framework.

1.2 Conceptual Framework

In this section, the key components of the double machine learning framework are briefly presented. For the sake of brevity, we consider estimation of a scalar parameter θ_0 in the partially linear regression model in Equation (1.1) and will assume that technical assumptions, e.g., related to measurability, are satisfied.

A key component of the double machine learning framework is a score function

$$\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0,$$

with i.i.d. data $W = (Y, D, X)$, target parameter θ_0 and a nuisance term η_0. The double machine learning estimator is the solution to the empirical analog of the moment condition above. An essential property of the score is so-called Neyman orthogonality

$$\partial_\eta \, \mathbb{E}[\psi(W;\theta_0,\eta)]\big|_{\eta=\eta_0} = 0.$$

A Neyman-orthogonal score function allows the estimator being robust against biases in the estimation of η_0, which may be introduced by variable selection or machine learning methods. By violating the orthogonality property, this bias effectively translates into the behavior of the naive estimator for the causal parameter and eventually leads to a biased and non-normal distribution of this estimator as illustrated in Figures 1.1 and 1.2.

For instance, the naive model selection approaches create an omitted variable bias: Whereas these procedures are able to identify important predictors for the dependent variable Y, they may involve selection mistakes in terms of so-called confounding variables. These are variables that are correlated with Y *and* the treatment variable D. In the examples that are based on a naive use of machine learning techniques, these estimation methods effectively introduce regularization to the regression problems, for example by the l_1-norm in the case of lasso. Whereas regularization allows the estimators to effectively achieve a preferable predictive performance in high-dimensional or highly-complex settings, it may simultaneously introduce a substantial bias that, finally, invalidates the naive inference approach.

Being based on a Neyman-orthogonal score, the double machine learning estimator makes it possible to overcome the shortcomings of the naive procedures. Although the estimators presented in Figure 1.3 are based on exactly the same machine learning methods as those in Figure 1.2, Neyman-orthogonality leads to robustness against the generated regularization bias. In the partially linear regression example, it can be shown that orthogonality of the score can be achieved by including a second nuisance component in η. In other words, orthogonality can be obtained with the nuisance term being $\eta = (g, m)$ leading to the Neyman-orthogonal score

$$\psi(W;\theta_0,\eta_0) := (Y - g(X) - \theta(D - m(X)))(D - m(X)),$$

with g and m being functions that satisfy some regularity assumptions (Chernozhukov et al., 2018a). By only estimating the component $g(X)$ in Equation (1.1), the naive approaches only consider g as the nuisance part which violates the orthogonality property. In contrast, the double machine learning estimator includes estimation of a second nuisance component $m_0(X)$

$$D = m_0(X) + \nu.$$

Intuitively, $m_0(X)$ captures the relationship between the treatment variable D and the controls X. For example, in a setting with a binary treatment, $m_0(X)$ corresponds to the propensity score, which finally leads to the property of double robustness that is well-known in the treatment effect literature.

The double machine learning framework, which has been developed in a sequence of studies and recently generalized in Chernozhukov et al. (2018a), leads to an asymptotically normally distributed estimator if an orthogonal score function is used. Moreover, several additional conditions have to hold. Most importantly, the nuisance components have to be estimated consistently and the employed estimation methods need to exhibit good predictive performance in terms of the rate of convergence. Accordingly,

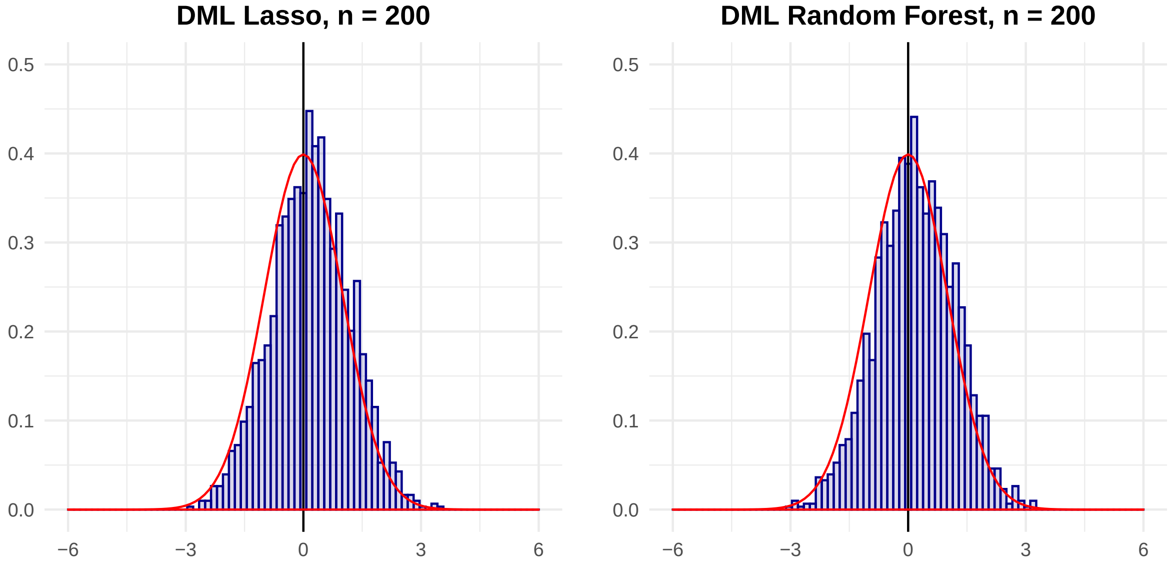


Figure 1.3: **Double machine learning with lasso and random forests, simulation example.**

The histograms illustrate the empirical distribution of a studentized double machine learning estimator that is based on an orthogonal moment condition and sample splitting obtained in $R = 2000$ simulation replications. **Left panel:** Double machine learning estimator based on the lasso with cross-validated choice of the penalty parameter. **Right panel:** Double machine learning estimator based on a random forest learner. Both estimators are asymptotically normally distributed and centered around the true value of the parameter θ_0. In $R = 2000$ simulation repetitions, constructing confidence intervals according to the double machine learning approach results in an empirical coverage of 94% (left panel) and 93% (right panel). In all simulation examples, the same data sets and specifications of the learners are used. A nominal coverage level of 95% is chosen for the confidence intervals in all cases.

the nuisance part η_0 needs to be estimated at a rate that is at least slightly faster than $n^{-1/4}$. Moreover, additional structural assumptions have to be satisfied. As an example, the lasso estimator can be used under the assumption that only a small set of the explanatory variables X have a non-zero influence on the outcome variable Y and the treatment variable D. This assumption is also called sparsity. Early work on valid inference using machine learning estimators has been based on the lasso as it allows for clarification of the corresponding technical requirements. In Chernozhukov et al. (2018a), a sample-splitting procedure is introduced that makes it possible to relax the involved structural assumptions and to use generic machine learning methods.

1.3 Outline

This doctoral thesis consists of a collection of research papers in the context of the double machine learning framework. Most of the papers provide empirical applications and implementations of the double machine learning framework. The first chapter provides a short summary and review of the double machine learning framework in Chernozhukov et al. (2018a). The overview is intended to provide guidance for users of the R and python packages `DoubleML` that have been developed as part of this doctoral thesis. The study presented in Chapter 3 provides a survey of methods for valid simultaneous inference in high-dimensional settings with a focus on the implementation in R. In settings where researchers or practitioners have to test a large number of components, it becomes greatly important to adjust the inferential procedure for multiple testing. Both of the first two papers present simulation examples to illustrate the use and validity of double machine learning in finite-sample settings. Chapter 4 provides an inferential procedure on a functional $f_1(x_1)$ in a high-dimensional additive model

$$Y = f_1(X_1) + \ldots + f_p(X_p) + \varepsilon,$$

with the number of components, p, possibly exceeding the number of observations in the data, n. The theoretical results that are based on the work in Belloni et al. (2018) are complemented by a simulation study and an illustration in an empirical application.

Chapter 5 provides an application of modern methods for valid simultaneous inference in an empirical application. The paper is concerned with a quantification of heterogeneity in the U.S. gender wage gap. The heterogeneity analysis is based on an interacted wage regression and multiple coefficients being tested simultaneously. The estimation framework is based on the double selection approach of Belloni et al. (2014a). This approach can be considered as a variant of the double machine learning framework developed for variable selection procedures such as the lasso estimator.

Finally, Chapter 6 is a study outside the double machine learning approach. In some sense, it can be considered as the other side of the same medal: Inferential procedures for causal inference result in estimation of causal parameters that are, in turn, highly important to design effective and efficient policy measures. Once these causal parameters are estimated, policy makers base their optimal policy decisions on an economic or structural framework. The COVID-19 pandemic creates a new challenge to policy makers who have to make difficult decisions facing the trade-off between protection of public health and mitigation of economic damage. Whereas some policy measures are highly effective in reducing the spread of the virus, these measures might be associated with severe economic consequences and vice versa. The study in Chapter 6 attempts to assess a variety of policy measures in the current pandemic. Finally, the conclusion presented in Chapter 7 presents potential extensions in future research.

Chapter 2

DoubleML - An Object-Oriented Implementation of Double Machine Learning in R

2.1 Introduction

Structural equation models provide a quintessential framework for conducting causal inference in statistics, econometrics, machine learning (ML), and other data sciences. The package `DoubleML` for R (R Core Team, 2020) implements partially linear and interactive structural equation and treatment effect models with high-dimensional confounding variables as considered in Chernozhukov et al. (2018a). Estimation and tuning of the machine learning models is based on the powerful functionalities provided by the `mlr3` package and the `mlr3` ecosystem (Lang et al., 2019). A key element of double machine learning (DML) models are score functions identifying the estimates for the target parameter. These functions play an essential role for valid inference with machine learning methods because they have to satisfy a property called Neyman orthogonality. With the score functions as key elements, `DoubleML` implements double machine learning in a very general way using object orientation based on the `R6` package (Chang, 2020). Currently, `DoubleML` implements the double / debiased machine learning framework as established in Chernozhukov et al. (2018a) for

- partially linear regression models (PLR),
- partially linear instrumental variable regression models (PLIV),
- interactive regression models (IRM), and
- interactive instrumental variable regression models (IIVM).

The object-oriented implementation of `DoubleML` is very flexible. The model classes `DoubleMLPLR`, `DoubleMLPLIV`, `DoubleMLIRM` and `DoubleIIVM` implement the estimation of the nuisance functions via machine learning methods and the computation of the Neyman-orthogonal score function. All other functionalities are implemented in the abstract base class `DoubleML`, including estimation of causal parameters, standard errors, t-tests, confidence intervals, as well as valid simultaneous inference through adjustments of p-values and estimation of joint confidence regions based on a multiplier bootstrap procedure. In combination with the estimation and tuning functionalities of `mlr3` and its ecosystem, this object-oriented implementation enables a high flexibility for the model specification in terms of

- the machine learning methods for estimation of the nuisance functions,

- the resampling schemes,
- the double machine learning algorithm, and
- the Neyman-orthogonal score functions.

It further can be readily extended regarding

- new model classes that come with Neyman-orthogonal score functions being linear in the target parameter,
- alternative score functions via callables, and
- customized resampling schemes.

Several other packages for estimation of causal effects based on machine learning methods exist for R. Probably the most popular packages are the `grf` package (Tibshirani et al., 2020), which implements generalized random forests (Athey et al., 2019), the package `hdm` (Chernozhukov et al., 2016a) for inference based on the lasso estimator and the `hdi` package (Dezeure et al., 2015) for inference in high-dimensional models. Previous implementations of the double machine learning (DML) framework of Chernozhukov et al. (2018a) have been provided by `postDoubleR` package (Szitas, 2019), the package `dmlmt` (Knaus, 2018) with a focus on lasso estimation, and `causalDML` (Knaus, 2020) for estimation of treatment effects under unconfoundedness. In python, `EconML` (Microsoft Research, 2019) offers an implementation of the double machine learning framework for heterogeneous effects. We would like to mention that the R package `DoubleML` was developed together with a Python twin (Bach et al., 2021) that is based on `scikit-learn` (Pedregosa et al., 2011). The python package is also available via GitHub, the Python Package Index (PyPI), and conda-forge.[1] Moreover, Kurz (2021) provides a serverless implementation of the python module `DoubleML`.

The rest of the paper is structured as follows: In Section 2.2, we briefly demonstrate how to install the `DoubleML` package and give a short motivating example to illustrate the major idea behind the double machine learning approach. Section 2.3 introduces the main causal model classes implemented in `DoubleML`. Section 2.4 shortly summarizes the main ideas behind the double machine learning approach and reviews the key ingredients required for valid inference based on machine learning methods. Section 2.5 presents the main steps and algorithms of the double machine learning procedure for inference on one or multiple target parameters. Section 2.6 provides more detailed insights on the implemented classes and methods of `DoubleML`. Section 2.7 contains real-data and simulation examples for estimation of causal parameters using the `DoubleML` package. Additionally, this section provides a brief simulation study that illustrates the validity of the implemented methods in finite samples. Section 5.6 concludes the paper. The code output that has been suppressed in the main text and further information regarding the simulations are presented in the Appendix. To make the code examples fully reproducible, the entire code is available online.

2.2 Getting started

2.2.1 Installation

The latest CRAN release of `DoubleML` can be installed using the command

```
install.packages("DoubleML")
```

---

[1] Resources for Python package: GitHub `https://github.com/DoubleML/doubleml-for-py`, PyPI: `https://pypi.org/project/DoubleML/`, conda-forge: `https://anaconda.org/conda-forge/doubleml`.

Alternatively, the development version can be downloaded and installed from the GitHub[2] repository using the command (previous installation of the `remotes` package is required)

```
remotes::install_github("DoubleML/doubleml-for-r")
```

Among others, `DoubleML` depends on the R package `R6` for object oriented implementation, `data.table` (Dowle and Srinivasan, 2020) for the underlying data structure, as well as the packages `mlr3` (Lang et al., 2019), `mlr3learners` (Lang et al., 2020a) and `mlr3tuning` (Becker et al., 2020) for estimation of machine learning methods, model tuning and parameter handling. Moreover, the underlying packages of the machine learning methods that are called in `mlr3` or `mlr3learners` must be installed, for example the packages `glmnet` for lasso estimation (**glmnet**) or `ranger` (Wright and Ziegler, 2017) for random forests.
Load the package after completed installation.

```
library(DoubleML)
```

2.2.2 A Motivating Example: Basics of Double Machine Learning

In the following, we provide a brief summary of and motivation to double machine learning methods and show how the corresponding methods provided by the `DoubleML` package can be applied. The data generating process (DGP) is based on the introductory example in Chernozhukov et al. (2018a). We consider a partially linear model: Our major interest is to estimate the causal parameter θ in the following regression equation

$$y_i = \theta d_i + g_0(x_i) + \zeta_i, \quad \zeta_i \sim \mathcal{N}(0,1),$$

with covariates $x_i \sim \mathcal{N}(0,\Sigma)$, where Σ is a matrix with entries $\Sigma_{kj} = 0.7^{|j-k|}$. In the following, the regression relationship between the treatment variable d_i and the covariates x_i will play an important role

$$d_i = m_0(x_i) + v_i, \quad v_i \sim \mathcal{N}(0,1).$$

The nuisance functions m_0 and g_0 are given by

$$\begin{aligned} m_0(x_i) &= x_{i,1} + \frac{1}{4}\frac{\exp(x_{i,3})}{1+\exp(x_{i,3})}, \\ g_0(x_i) &= \frac{\exp(x_{i,1})}{1+\exp(x_{i,1})} + \frac{1}{4}x_{i,3}. \end{aligned}$$

We construct a setting with $n = 500$ observations and $p = 20$ explanatory variables to demonstrate the use of the estimators provided in `DoubleML`. Moreover, we set the true value of the parameter θ to $\theta = 0.5$. The corresponding data generating process is implemented in the function `make_plr_CCDHNR2018()`. We start by generating a realization of a data set as a `data.table` object, which is subsequently used to create an instance of the data-backend of class `DoubleMLData`.

```
library(DoubleML)
alpha = 0.5
n_obs = 500
n_vars = 20
```

---

[2] GitHub repository for R package: `https://github.com/DoubleML/doubleml-for-r`.

```
set.seed(1234)
data_plr = make_plr_CCDDHNR2018(alpha = alpha, n_obs = n_obs, dim_x = n_vars,
                                return_type = "data.table")
```

The data-backend implements the causal model: We specify that we perform inference on the effect of the treatment variable d_i on the dependent variable y_i.

```
obj_dml_data = DoubleMLData$new(data_plr, y_col = "y", d_cols = "d")
```

In the next step, we choose the machine learning method as an object of class `Learner` from `mlr3`, `mlr3learners` (Lang et al., 2020a) or `mlr3extralearners` (Sonabend and Schratz, 2020). As we will point out later, we have to estimate two nuisance parts in order to perform valid inference in the partially linear regression model. Hence, we have to specify two learners. Moreover, we split the sample into two folds used for cross-fitting.

```
# Load mlr3 and mlr3learners package and suppress output during estimation
library(mlr3)
library(mlr3learners)
lgr::get_logger("mlr3")$set_threshold("warn")

# Initialize a random forests learner with specified parameters
ml_g = lrn("regr.ranger", num.trees = 100, mtry = n_vars, min.node.size = 2,
           max.depth = 5)
ml_m = lrn("regr.ranger", num.trees = 100, mtry = n_vars, min.node.size = 2,
           max.depth = 5)

doubleml_plr = DoubleMLPLR$new(obj_dml_data,
                               ml_g, ml_m,
                               n_folds = 2,
                               score = "IV-type")
```

To estimate the causal effect of variable d_i on y_i, we call the `fit()` method.

```
doubleml_plr$fit()
doubleml_plr$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##   Estimate. Std. Error t value Pr(>|t|)
## d   0.49398    0.04852   10.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows that the estimated coefficient is close to the true parameter $\theta = 0.5$. Moreover, we are able to reject the null hypotheses $H_0 : \theta = 0$ at all common significance levels.

2.3 Key Causal Models

`DoubleML` provides estimation of causal effects in four different models: Partially linear regression models (PLR), partially linear instrumental variable regression models (PLIV), interactive regression models (IRM) and interactive instrumental variable regression models (IIVM). We will shortly introduce these models.

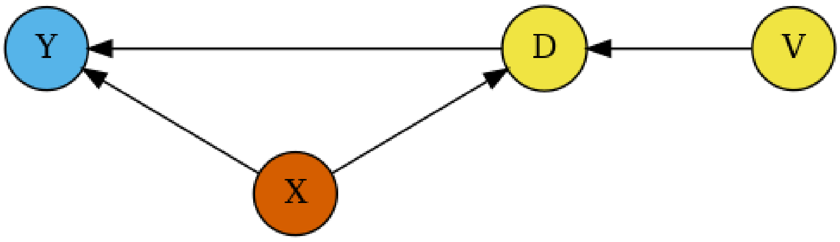


Figure 2.1: **Causal diagram for PLR and IRM.**

A causal diagram underlying Equation (2.1)-(2.2) and (2.5)-(2.6) under conditional exogeneity. Note that the causal link between D and Y is one-directional. Identification of the causal effect is confounded by X, and identification is achieved via V, which captures variation in D that is independent of X. Methods to estimate the causal effect of D must therefore approximately remove the effect of high-dimensional X on Y and D.

2.3.1 Partially Linear Regression Model (PLR)

Partially linear regression models (PLR), which encompass the standard linear regression model, play an important role in data analysis (Robinson, 1988). Partially linear regression models take the form

$$Y = D\theta_0 + g_0(X) + \zeta, \quad \mathbb{E}(\zeta|D,X) = 0, \tag{2.1}$$

$$D = m_0(X) + V, \quad \mathbb{E}(V|X) = 0, \tag{2.2}$$

where Y is the outcome variable and D is the policy variable of interest. The high-dimensional vector $X = (X_1, \ldots, X_p)$ consists of other confounding covariates, and ζ and V are stochastic errors. Equation (2.1) is the equation of interest, and θ_0 is the main regression coefficient that we would like to infer. If D is conditionally exogenous (randomly assigned conditional on X), θ_0 has the interpretation of a structural or causal parameter. The causal diagram supporting such interpretation is shown in Figure 2.1. The second equation keeps track of confounding, namely the dependence of D on covariates/controls. The characteristics X affect the policy variable D via the function $m_0(X)$ and the outcome variable via the function $g_0(X)$. The partially linear model generalizes both linear regression models, where functions g_0 and m_0 are linear with respect to a dictionary of basis functions with respect to X, and approximately linear models.

2.3.2 Partially Linear Instrumental Variable Regression Model (PLIV)

We next consider the partially linear instrumental variable regression model:

$$Y - D\theta_0 = g_0(X) + \zeta, \quad \mathbb{E}(\zeta|Z,X) = 0, \tag{2.3}$$

$$Z = m_0(X) + V, \quad \mathbb{E}(V|X) = 0. \tag{2.4}$$

Note that this model is not a regression model unless $Z = D$. Model (2.3)-(2.4) is a canonical model in causal inference, going back to Wright (1928), with the modern difference being that g_0 and m_0 are nonlinear, potentially complicated functions of high-dimensional X. The idea of this model is that there is a structural or causal relation between Y and D, captured by θ_0, and $g_0(X) + \zeta$ is the stochastic error, partly explained by covariates X. V and ζ are stochastic errors that are not explained by X. Since Y and D are jointly determined, we need an external factor, commonly referred to as an instrument, Z, to create exogenous variation in D. Note that Z should affect D. The X here serve again as confounding factors, so we can think of variation in Z as being exogenous only conditional on X.

A simple contextual example is from biostatistics (Permutt and Hebel, 1989), where Y is a health outcome and D is an indicator of smoking. Thus, θ_0 captures the effect of smoking on health. Health outcome Y

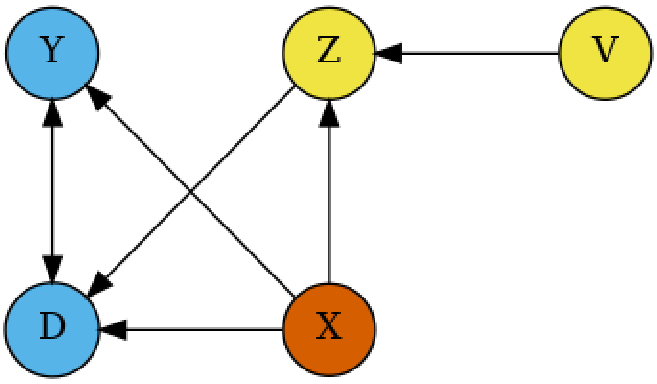


Figure 2.2: **Causal diagram for PLIV and IIVM.**

A causal diagram underlying Equation (2.3)-(2.4) and (2.7)-(2.8) under conditional exogeneity of Z. Note that the causal link between D and Y is bi-directional, so an instrument Z is needed for identification. Identification is achieved via V that captures variation in Z that is independent of X. Equations (2.3) and (2.4) do not model the dependence between D and X and Z, though a necessary condition for identification is that Z and D are related after conditioning on X. Methods to estimate the causal effect of D must approximately remove the effect of high-dimensional X on Y, D, and Z. Removing the confounding effect of X is done implicitly by the proposed procedure.

and smoking behavior D are treated as being jointly determined. X represents patient characteristics, and Z could be a doctor's advice not to smoke (or another behavioral treatment) that may affect the outcome Y only through shifting the behavior D, conditional on characteristics X.

2.3.3 Interactive Regression Model (IRM)

We consider estimation of average treatment effects when treatment effects are fully heterogeneous and the treatment variable is binary, $D \in \{0, 1\}$. We consider vectors (Y, D, X) such that

$$Y = g_0(D, X) + U, \quad \mathbb{E}(U|X, D) = 0, \tag{2.5}$$

$$D = m_0(X) + V, \quad \mathbb{E}(V|X) = 0. \tag{2.6}$$

Since D is not additively separable, this model is more general than the partially linear model for the case of binary D. A common target parameter of interest in this model is the average treatment effect (ATE),[3]

$$\theta_0 = \mathbb{E}[g_0(1, X) - g_0(0, X)].$$

Another common target parameter is the average treatment effect for the treated (ATTE),

$$\theta_0 = \mathbb{E}[g_0(1, X) - g_0(0, X)|D = 1].$$

In business applications, the ATTE is often the main interest, as it captures the treatment effect for those who have been affected by the treatment. A difference of the ATTE from the ATE might arise if the characteristics of the treated individuals differ from those of the general population.

The confounding factors X affect the policy variable via the propensity score $m_0(X)$ and the outcome variable via the function $g_0(X)$. Both of these functions are unknown and potentially complex, and we can employ ML methods to learn them.

---

[3] Without unconfoundedness/conditional exogeneity, these quantities measure association, and could be referred to as average predictive effects (APE) and average predictive effect for the exposed (APEX). Inferential results for these objects would follow immediately from Theorem 1.

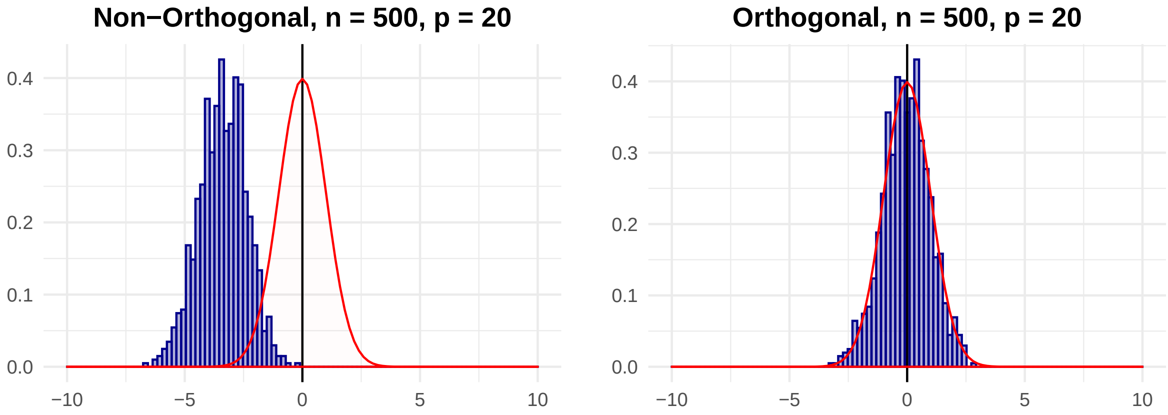


Figure 2.3: **Performance of non-orthogonal and orthogonal estimators in simulated data example.**

Left panel: Histogram of the studentized naive estimator $\hat{\theta}_0^{naive}$. $\hat{\theta}_0^{naive}$ is based on estimation of g_0 and m_0 with random forests and a non-orthogonal score function. Data sets are simulated according to the data generating process in Section 2.2.2. Data generation and estimation are repeated 1000 times. **Right panel:** Histogram of the studentized DML estimator $\tilde{\theta}_0$. $\tilde{\theta}_0$ is based on estimation of g_0 and m_0 with random forests and an orthogonal score function provided in Equation (2.17). Note that the simulated data sets and parameters of the random forest learners are identical to those underlying the left panel.

2.3.4 Interactive Instrumental Variable Model (IIVM)

We consider estimation of local average treatment effects (LATE) with a binary treatment variable $D \in \{0,1\}$, and a binary instrument, $Z \in \{0,1\}$. As before, Y denotes the outcome variable, and X is the vector of covariates. Here the structural equation model is:

$$Y = \ell_0(D,X) + \zeta, \quad \mathbb{E}(\zeta|Z,X) = 0, \tag{2.7}$$
$$Z = m_0(X) + V, \quad \mathbb{E}(V|X) = 0. \tag{2.8}$$

Consider the functions g_0, r_0, and m_0, where g_0 maps the support of (Z,X) to $\mathbb{R}$ and r_0 and m_0 map the support of (Z,X) and X to $(\epsilon, 1-\epsilon)$ for some $\epsilon \in (0,1/2)$, such that

$$Y = g_0(Z,X) + \nu, \quad \mathbb{E}(\nu|Z,X) = 0, \tag{2.9}$$
$$D = r_0(Z,X) + U, \quad \mathbb{E}(U|Z,X) = 0, \tag{2.10}$$
$$Z = m_0(X) + V, \quad \mathbb{E}(V|X) = 0. \tag{2.11}$$

We are interested in estimating

$$\theta_0 = \frac{\mathbb{E}[g_0(1,X)] - \mathbb{E}[g_0(0,X)]}{\mathbb{E}[r_0(1,X)] - \mathbb{E}[r_0(0,X)]}.$$

Under the well-known assumptions of Imbens and Angrist (1994), θ_0 is the LATE – the average treatment effect for compliers, in other words, those observations that would have $D = 1$ if Z were 1 and would have $D = 0$ if Z were 0.

2.4 Basic Idea and Key Ingredients of Double Machine Learning

2.4.1 Basic Idea behind Double Machine Learning for the PLR Model

Here we provide an intuitive discussion of how double machine learning works in the first model, the partially linear regression model. Naive application of machine learning methods directly to equations (2.1)-(2.2) may have a very high bias. Indeed, it can be shown that small biases in estimation of g_0, which are unavoidable in high-dimensional estimation, create a bias in the naive estimate of the main effect, $\hat{\theta_0}^{naive}$, which is sufficiently large to cause failure of conventional inference. The left panel in Figure 2.3 illustrates this phenomenon. The histogram presents the empirical distribution of the studentized estimator, $\hat{\theta}_0^{naive}$, as obtained in 1000 independent repetitions of the data generating process presented in Section 2.2.2. The functions g_0 and m_0 in the PLR model are estimated with random forest learners and corresponding predictions are then plugged into a non-orthogonal score function. The regularization performed by the random forest learner leads to a bias in estimation of g_0 and m_0. Due to non-orthogonality of the score, this translates into a considerable bias of the main estimator $\hat{\theta_0}^{naive}$: The distribution of the studentized estimator $\hat{\theta}_0^{naive}$ is shifted to the left of the origin and differs substantially from a normal distribution that would be obtained if the regularization bias was negligible as shown by the red curve. The PLR model above can be rewritten in the following residualized form:

$$W = V\theta_0 + \zeta, \quad \mathbb{E}(\zeta|D, X) = 0, \tag{2.12}$$

$$W = (Y - \ell_0(X)), \quad \ell_0(X) = \mathbb{E}[Y|X], \tag{2.13}$$

$$V = (D - m_0(X)), \quad m_0(X) = \mathbb{E}[D|X]. \tag{2.14}$$

The variables W and V represent original variables after taking out or *partialling out* the effect of X. Note that θ_0 is identified from this equation if V has a non-zero variance.

Given identification, double machine learning for a PLR proceeds as follows

(1) Estimate ℓ_0 and m_0 by $\hat{\ell}_0$ and $\hat{m}_0$, which amounts to solving the two problems of predicting Y and D using X, using any generic ML method, giving us estimated residuals

$$\hat{W} = Y - \hat{\ell}_0(X),$$

and

$$\hat{V} = D - \hat{m}_0(X).$$

The residuals should be of a cross-validated form, as explained below in Algorithm 1 or 2, to avoid biases from overfitting.

(2) Estimate θ_0 by regressing the residual $\hat{W}$ on $\hat{V}$. Use the conventional inference for this regression estimator, ignoring the estimation error in the residuals.

The reason we work with this residualized form is that it eliminates the bias arising from solving the prediction problems in stage (1). The estimates $\hat{\ell}_0$ and $\hat{m}_0$ carry a regularization bias due to having to solve prediction problems well in high-dimensions. However, the nature of the estimating equation for θ_0 are such that these biases are eliminated to the first order, as explained below. This results in a high-quality low-bias estimator $\tilde{\theta}_0$ of θ_0, as illustrated in the right panel of Figure 2.3.

The estimator is adaptive in the sense that the first stage estimation errors do not affect the second stage errors.

2.4.2 Key Ingredients of the Double Machine Learning Inference Approach

Our goal is to construct high-quality point and interval estimators for θ_0 when X is high-dimensional and we employ machine learning methods to estimate the nuisance functions such as g_0 and m_0. Example ML methods include lasso, random forests, boosted trees, deep neural networks, and ensembles or aggregated versions of these methods.

We shall use a method-of-moments estimator for θ_0 based upon the empirical analog of the moment condition

$$\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0, \tag{2.15}$$

where we call ψ the score function, $W = (Y, D, X, Z)$, θ_0 is the parameter of interest, and η denotes nuisance functions with population value η_0.

The first key input of the inference procedure is using a score function $\psi(W;\theta;\eta)$ that satisfies (2.15), with θ_0 being the unique solution, and that obeys the Neyman orthogonality condition

$$\partial_\eta \mathbb{E}[\psi(W;\theta_0,\eta)|_{\eta=\eta_0} = 0. \tag{2.16}$$

Neyman orthogonality (2.16) ensures that the moment condition (2.15) used to identify and estimate θ_0 is insensitive to small pertubations of the nuisance function η around η_0. The derivative ∂_η denotes the pathwise (Gateaux) derivative operator.

Using a Neyman-orthogonal score eliminates the first order biases arising from the replacement of η_0 with a ML estimator $\hat{\eta}_0$. Eliminating this bias is important because estimators $\hat{\eta}_0$ must be heavily regularized in high dimensional settings to be good estimators of η_0, and so these estimators will be biased in general. The Neyman orthogonality property is responsible for the adaptivity of these estimators – namely, their approximate distribution will not depend on the fact that the estimate $\hat{\eta}_0$ contains error, if the latter is mild.

The right panel of Figure 2.3 presents the empirical distribution of the studentized DML estimator $\tilde{\theta}_0$ that is based on an orthogonal score. Note that estimation is performed on the identical simulated data sets and with the same machine learning method as for the naive learner, which is displayed in the left panel. The histogram of the studentized estimator $\tilde{\theta}_0$ illustrates the favorable performance of the double machine learning estimator, which is based on an orthogonal score: The DML estimator is robust to the bias that is generated by regularization. The estimator is approximately unbiased, is concentrated around 0 and the distribution is well-approximated by the normal distribution.

- **PLR score:** In the PLR model, we can employ two alternative score functions. We will shortly indicate the option for initialization of a model object in `DoubleML` to clarify how each score can be implemented. Using the option `score = 'partialling out'` leads to estimation of the score function

$$\begin{aligned} &\psi(W;\theta,\eta) := (Y - \ell(X) - \theta(D - m(X)))\,(D - m(X))\,, \\ &\eta = (\ell, m), \quad \eta_0 = (\ell_0, m_0), \end{aligned} \tag{2.17}$$

where $W = (Y, D, X)$ and ℓ and m are P-square-integrable functions mapping the support of X to $\mathbb{R}$, whose true values are given by

$$\ell_0(X) = \mathbb{E}[Y|X], \quad m_0(X) = \mathbb{E}[D|X].$$

Alternatively, it is possible to use the following score function for the PLR via the option `score = 'IV-type'`

$$\psi(W; \theta, \eta) := (Y - D\theta - g(X))\,(D - m(X))\,, \quad \eta = (g, m), \quad \eta_0 = (g_0, m_0), \tag{2.18}$$

with g and m being P-square-integrable functions mapping the support of X to $\mathbb{R}$ with values given by

$$g_0 = \mathbb{E}[Y|X], \quad m_0(X) = \mathbb{E}[D|X].$$

The scores above are Neyman-orthogonal by elementary calculations. Now, it is possible to see the connections to the residualized system of equations presented in Section 2.4.1.

- **PLIV score:** In the PLIV model, we employ the score function (`score = 'partialling out'`)

  $$\begin{aligned}&\psi(W; \theta, \eta) := (Y - \ell(x) - \theta(D - r(X)))\,(Z - m(X))\,,\\ &\eta = (\ell, m, r), \quad \eta_0 = (\ell_0, m_0, r_0),\end{aligned} \tag{2.19}$$

  where $W = (Y, D, X, Z)$ and ℓ, m, and r are P-square integrable functions mapping the support of X to $\mathbb{R}$, whose true values are given by

  $$\ell_0(X) = \mathbb{E}[Y|X], \quad r_0(X) = \mathbb{E}[D|X], \quad m_0(X) = \mathbb{E}[Z|X].$$

- **IRM score:** For estimation of the ATE parameter of the IRM model, we employ the score (`score = 'ATE'`)

  $$\begin{aligned}&\psi(W; \theta, \eta) := (g(1, X) - g(0, X)) + \frac{D(Y - g(1, X))}{m(X)} - \frac{(1 - D)(Y - g(0, X))}{1 - m(X)} - \theta,\\ &\eta = (g, m), \quad \eta_0 = (g_0, m_0),\end{aligned} \tag{2.20}$$

  where $W = (Y, D, X)$ and g and m map the support of (D, X) to $\mathbb{R}$ and the support of X to $(\epsilon, 1 - \epsilon)$, respectively, for some $\epsilon \in (0, 1/2)$, whose true values are given by

  $$g_0(D, X) = \mathbb{E}[Y|D, X], \quad m_0(x) = \mathbb{P}[D = 1|X].$$

  This orthogonal score is based on the influence function for the mean for missing data from Robins and Rotnitzky (1995). For estimation of the ATTE parameter in the IRM, we use the score (`score = 'ATTE'`)

  $$\begin{aligned}&\psi(W; \theta, \eta) := \frac{D(Y - g(0, X))}{p} - \frac{m(X)(1 - D)(Y - g(0, X))}{p(1 - m(x))} - \frac{D}{p}\theta,\\ &\eta = (g, m, p), \quad \eta_0 = (g_0, m_0, p_0),\end{aligned} \tag{2.21}$$

  where $p_0 = \mathbb{P}(D = 1)$. Note that this score does not require estimating $g_0(1, X)$.

- **IIVM score:** To estimate the LATE paramter in the IIVM, we will use the score (`score = 'LATE'`)

$$\begin{aligned}\psi :=& g(1,X) - g(0,X) + \frac{Z(Y-g(1,X))}{m(X)} - \frac{(1-Z)(Y-g(0,X))}{1-m(X)} \\ &- \left(r(1,x) - r(0,X) + \frac{Z(D-r(1,x)}{m(X)} - \frac{(1-Z)(D-r(0,X)}{1-m(X)}\right) \times \theta, \\ \eta =& (g,m,r), \quad \eta_0 = (g_0,m_0,r_0),\end{aligned} \tag{2.22}$$

  where $W = (Y, D, X, Z)$ and the nuisance parameter $\eta = (g, m, r)$ consists of P-square integrable functions g, m, and r, with g mapping the support of (Z, X) to $\mathbb{R}$ and m and r, respectively, mapping the support of (Z, X) and X to $(\epsilon, 1-\epsilon)$ for some $\epsilon \in (0, 1/2)$.

The second key input is the use of high-quality machine learning estimators for the nuisance parameters.

For instance, in the PLR model, we need to have access to consistent estimators of g_0 and m_0 with respect to the $L^2(P)$ norm $\|\cdot\|_{P,2}$, such that

$$\|\hat{m}_0 - m_0\|_{P,2} + \|\hat{\ell}_0 - \ell_0\|_{P,2} \leq o(N^{-1/4}). \tag{2.23}$$

In the PLIV model, the sufficient condition is

$$\|\hat{r}_0 - r_0\|_{P,2} + \|\hat{m}_0 - m_0\|_{P,2} + \|\hat{\ell}_0 - \ell_0\|_{P,2} \leq o(N^{-1/4}). \tag{2.24}$$

These conditions are plausible for many ML methods. Different structured assumptions on η_0 lead to the use of different machine-learning tools for estimating η_0 as listed in Chernozhukov et al. (2018a, pp. 22-23):

1. The assumption of approximate or exact sparsity for η_0 with respect to some dictionary calls for the use of sparsity-based machine learning methods, for example the lasso estimator, post-lasso, l_2-boosting, or forward selection, among others.
2. The assumption of density of η_0 with respect to some dictionary calls for density-based estimators such as the ridge. Mixed structures based on sparsity and density suggest the use of elastic net or lava.
3. If η_0 can be well approximated by tree-based methods, regression trees and random forests are suitable.
4. If η_0 can be well approximated by sparse, shallow or deep neural networks, l_1-penalized neural networks, shallow neural networks or deep neural networks are attractive.

For most of these ML methods, performance guarantees are available that make it possible to satisfy the theoretical requirements. Moreover, if η_0 can be well approximated by at least one model mentioned in the list above, ensemble or aggregated methods can be used. Ensemble and aggregation methods ensure that the performance guarantee is approximately no worse than the performance of the best method.

The third key input is to use a form of sample splitting at the stage of producing the estimator of the main parameter θ_0, which allows us to avoid biases arising from overfitting.

Biases arising from overfitting could result from using highly complex fitting methods such as boosting, random forests, ensemble, and hybrid machine learning methods. We specifically use cross-fitted forms of

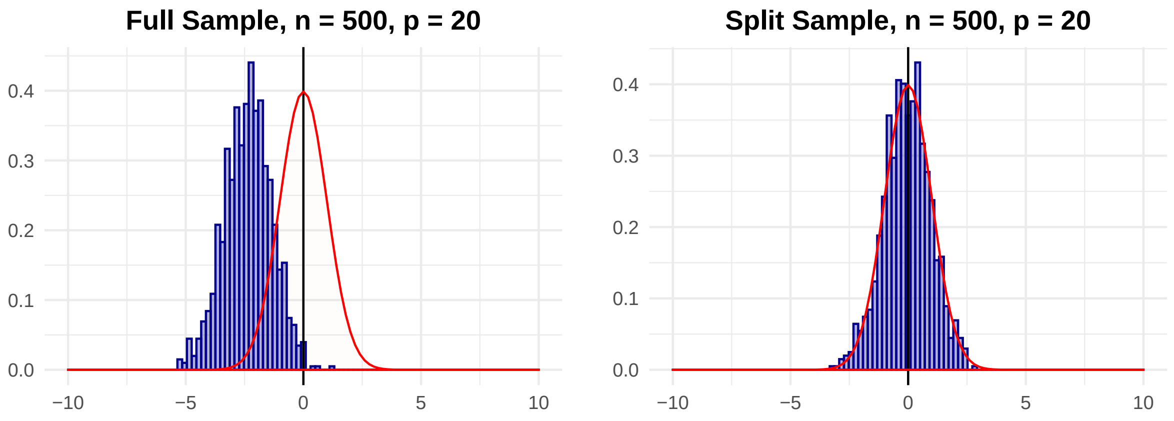


Figure 2.4: **Performance of orthogonal estimators based on full sample and sample splitting in simulated data example.**

Left panel: Histogram of the studentized estimator $\hat{\theta}_0^{nosplit}$. $\hat{\theta}_0^{nosplit}$ is based on estimation of g_0 and m_0 with random forests and a procedure without sample-splitting: The entire data set is used for learning the nuisance terms and estimation of the orthogonal score. Data sets are simulated according to the data generating process in Section 2.2.2. Data generation and estimation are repeated 1000 times. **Right panel:** Histogram of the studentized DML estimator $\tilde{\theta}_0$. $\tilde{\theta}_0$ is based on estimation of g_0 and m_0 with random forests and the cross-fitting described in Algorithm 2. Note that the simulated data sets and parameters of the random forest learners are identical to those underlying the left panel.

the empirical moments, as detailed below in Algorithms 1 and 2, in estimation of θ_0. If we do not perform sample splitting and the ML estimates overfit, we may end up with very large biases. This is illustrated in Figure 2.4. The left panel shows the histogram of a studentized estimator $\hat{\theta}_0^{nosplit}$ with $\hat{\theta}_0^{nosplit}$ being obtained from solving the orthogonal score of Equation (2.17) without sample splitting. All observations are used to learn functions g_0 and m_0 in the PLR model and to solve the score $\frac{1}{N}\sum_i^N \psi(W_i; \hat{\theta}_0^{nosplit}, \hat{\eta}_0)$. Consequently, this overfitting bias leads to a considerable shift of the empirical distribution to the left. The double machine learning estimator underlying the histogram in the right panel is obtained with cross-fitting according to Algorithm 2. The sample-splitting procedure makes it possible to completely eliminate the bias induced by overfitting.

2.5 The Double Machine Learning Inference Method

2.5.1 Double Machine Learning for Estimation of a Causal Parameter

We assume that we have a sample $(W_i)_{i1}^N$, modeled as i.i.d. copies of $W = (Y, D, Z, X)$, whose law is determined by the probability measure P. We assume that N is divisible by K in order to simplify the notation. Let $\mathbb{E}_N$ denote the empirical expectation

$$\mathbb{E}_N[g(W)] := \frac{1}{N}\sum_{i=1}^N g(W_i).$$

Algorithm 1: DML1. (Generic double machine learning with cross-fitting)

(1) **Inputs:** Choose a model (PLR, PLIV, IRM, IIVM), provide data $(W_i)_{i=1}^N$, a Neyman-orthogonal score function $\psi(W; \theta, \eta)$, which depends on the model being estimated, and specify machine learning methods for η.

(2) **Train ML predictors on folds:** Take a K-fold random partition $(I_k)_{k=1}^K$ of observation

indices $[N] = \{1, \ldots, N\}$ such that the size of each fold I_k is $n = N/K$. For each $k \in [K] = \{1, \ldots, K\}$, construct a high-quality machine learning estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_{0,k}\big((W_i)_{i \notin I_k}\big)$$

of η_0, where $x \mapsto \hat{\eta}_{0,k}(x)$ depends only on the subset of data $(W_i)_{i \notin I_k}$.

(3) For each $k \in [K]$, construct the estimator $\check{\theta}_{0,k}$ as the solution to the equation

$$\frac{1}{n} \sum_{i \in I_k} \psi(W_i; \check{\theta}_{0,k}, \hat{\eta}_{0,k}) = 0. \tag{2.25}$$

The estimate of the causal parameter is obtained via aggregation

$$\tilde{\theta}_0 = \frac{1}{K} \sum_{k=1}^{K} \check{\theta}_{0,k}.$$

(4) **Output:** The estimate of the causal parameter $\tilde{\theta}_0$ as well as the values of the evaluated score function are returned.

Algorithm 2: DML2. (Generic double machine learning with cross-fitting)

(1) **Inputs:** Choose a model (PLR, PLIV, IRM, IIVM), provide data $(W_i)_{i=1}^{N}$, a Neyman-orthogonal score function $\psi(W; \theta, \eta)$, which depends on the model being estimated, and specify machine learning methods for η.

(2) **Train ML predictors on folds:** Take a K-fold random partition $(I_k)_{k=1}^{K}$ of observation indices $[N] = \{1, \ldots, N\}$ such that the size of each fold I_k is $n = N/K$. For each $k \in [K] = \{1, \ldots, K\}$, construct a high-quality machine learning estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_{0,k}\big((W_i)_{i \notin I_k}\big)$$

of η_0, where $x \mapsto \hat{\eta}_{0,k}(x)$ depends only on the subset of data $(W_i)_{i \notin I_k}$.

(3) Construct the estimator for the causal parameter $\tilde{\theta}_0$ as the solution to the equation

$$\frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \tilde{\theta}_0, \hat{\eta}_{0,k}) = 0. \tag{2.26}$$

(4) **Output:** The estimate of the causal parameter $\tilde{\theta}_0$ as well as the values of the evaluated score function are returned.

Remark 1 (*Linear scores*) The score for the models PLR, PLIV, IRM and IIVM are linear in θ, having the form

$$\psi(W; \theta, \eta) = \psi_a(W; \eta)\theta + \psi_b(W; \eta),$$

hence the estimator $\tilde{\theta}_{0,k}$ for DML2 ($\check{\theta}_{0,k}$ for DML1) takes the form

$$\tilde{\theta}_0 = -\left(\mathbb{E}_N[\psi_a(W;\eta)]\right)^{-1}\mathbb{E}_N[\psi_b(W;\eta)]. \tag{2.27}$$

The linear score function representations of the PLR, PLIV, IRM and IIVM are

- **PLR** with `score = 'partialling out'`

$$\begin{aligned}\psi_a(W;\eta) &= -(D-m(X))(D-m(X)),\\ \psi_b(W;\eta) &= (Y-\ell(X))(D-m(X)).\end{aligned} \tag{2.28}$$

  PLR with `score = 'IV-type'`

$$\begin{aligned}\psi_a(W;\eta) &= -D(D-m(X)),\\ \psi_b(W;\eta) &= (Y-g(X))(D-m(X)).\end{aligned} \tag{2.29}$$

- **PLIV** with `score = 'partialling out'`

$$\begin{aligned}\psi_a(W;\eta) &= -(D-r(X))(Z-m(X)),\\ \psi_b(W;\eta) &= (Y-\ell(X))(Z-m(X)).\end{aligned} \tag{2.30}$$

- **IRM** with `score = 'ATE'`

$$\begin{aligned}\psi_a(W;\eta) &= -1,\\ \psi_b(W;\eta) &= g(1,X)-g(0,X)+\frac{D(Y-g(1,X))}{m(X)}-\frac{(1-D)(Y-g(0,X))}{1-m(x)}.\end{aligned} \tag{2.31}$$

  IRM with `score = 'ATTE'`

$$\begin{aligned}\psi_a(W;\theta,\eta) &= -\frac{D}{p}\\ \psi_b(W;\theta,\eta) &= \frac{D(Y-g(0,X))}{p}-\frac{m(X)(1-D)(Y-g(0,X))}{p(1-m(x))}\end{aligned} \tag{2.32}$$

- **IIVM** with `score = 'LATE'`

$$\begin{aligned}\psi_a(W;\eta) &= -\left(r(1,X)-r(0,X)+\frac{Z(D-r(1,X))}{m(X)}-\frac{(1-Z)(D-r(0,X))}{1-m(x)}\right),\\ \psi_b(W;\eta) &= g(1,X)-g(0,X)+\frac{Z(Y-g(1,X))}{m(X)}-\frac{(1-Z)(Y-g(0,X))}{1-m(x)}.\end{aligned} \tag{2.33}$$

Remark 2 (*Sample Splitting*) In Step (2) of the Algorithm DML1 and DML2, the estimator $\hat{\eta}_{0,k}$ can generally be an ensemble or aggregation of several estimators as long as we only use the data $(W_i)_{i\notin I_k}$ outside the k-th fold to construct the estimators.

Remark 3 (*Recommendation*) We have found that $K = 4$ or $K = 5$ to work better than $K = 2$ in a variety of empirical examples and in simulations. The default for the option `n_folds` that implements the value of K is `n_folds=5`. Moreover, we generally recommend to repeat the estimation procedure mutliple times and use the estimates and standard errors as aggregated over multiple repetitions as described in Chernozhukov et al. (2018a, pp. 30-31). This aggregation will be automatically executed if the number of repetitions `n_rep` is set to a value larger than 1.

The properties of the estimator are as follows.

Theorem 1. *There exist regularity conditions, such that the estimator* $\tilde{\theta}_0$ *concentrates in a* $1/\sqrt{N}$*-neighborhood of* θ_0 *and the sampling error* $\sqrt{N}(\tilde{\theta}_0 - \theta_0)$ *is approximately normal*

$$\sqrt{N}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0, \sigma^2),$$

with mean zero and variance given by

$$\begin{aligned} \sigma^2 &= J_0^{-2}\mathbb{E}(\psi^2(W; \theta_0, \eta_0)), \\ J_0 &= \mathbb{E}(\psi_a(W; \eta_0)). \end{aligned}$$

Algorithm 3: Variance Estimation and Confidence Intervals.

(1) **Inputs:** Use the inputs and outputs from Algorithm 1 (DML1) or Algorithm 2 (DML2).

(2) **Variance and confidence intervals:** Estimate the asymptotic variance of $\tilde{\theta}_0$ by

$$\begin{aligned} \hat{\sigma}^2 &= \hat{J}_0^{-2}\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\left[\psi(W_i; \tilde{\theta}_0, \hat{\eta}_{0,k})\right]^2, \\ \hat{J}_0 &= \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi_a(W_i; \hat{\eta}_{0,k}) \end{aligned}$$

and form an approximate $(1-\alpha)$ confidence interval as

$$[\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}].$$

(3) **Output:** Output variance estimator and the confidence interval.

Theorem 2. *Under the same regularity condition, this interval contains* θ_0 *for approximately* $(1-\alpha)\times 100$ *percent of data realizations*

$$\mathbb{P}\left(\theta_0 \in \left[\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}\right]\right) \to (1-\alpha).$$

Remark 4 (*Brief literature overview on double machine learning*) The presented double machine learning method was developed in Chernozhukov et al. (2018a). The idea of using property (16) to construct estimators and inference procedures that are robust to small mistakes in nuisance parameters can be traced back to Neyman (1959) and has been used explicitly or implicitly in the literature on debiased sparsity-based inference (Belloni et al., 2011; Belloni et al., 2014b; Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014; Chernozhukov et al., 2015b) as well as (implicitly) in the classical semiparametric learning theory with low-dimensional X (Bickel et al., 1993; Newey, 1994; Van der Vaart, 2000; Van der Laan and Rose, 2011). These references also explain that if we use scores ψ that are not Neyman-orthogonal in high dimensional settings, then the resulting estimators of θ_0 are not $1/\sqrt{N}$ consistent and are generally heavily biased.

Remark 5 (*Literature on sample splitting*). Sample splitting has been used in the traditional semiparametric estimation literature to establish good properties of semiparametric estimators under weak conditions (Schick, 1986; Van der Vaart, 2000). In sparse learning problems with high-dimensional X, sample splitting was employed in Belloni et al. (2012). There and here, the use of sample splitting results in weak conditions on the estimators of nuisance parameters, translating into weak assumptions on sparsity in the case of sparsity-based learning.

Remark 6 (*Debiased machine learning*). The presented approach builds upon and generalizes the approach of Belloni et al. (2011), Zhang and Zhang (2014), Javanmard and Montanari (2014), Javanmard and Montanari (2014), Javanmard and Montanari (2018), Belloni et al. (2014c), Belloni et al. (2014a), Bühlmann and van de Geer (2015), which considered estimation of the special case (2.1)-(2.2) using lasso without cross-fitting. This generalization, by relying upon cross-fitting, opens up the use of a much broader collection of machine learning methods and, in the case the lasso is used to estimate the nuisance functions, allows relaxation of sparsity conditions. All of these approaches can be seen as "debiasing" the estimation of the main parameter by constructing, implicitly or explicitly, score functions that satisfy the exact or approximate Neyman orthogonality.

2.5.2 Methods for Simultaneous Inference

In addition to estimation of target causal parameters, standard errors, and confidence intervals, the package `DoubleML` provides methods to perform valid simultaneous inference based on a multiplier bootstrap procedure introduced in Chernozhukov et al. (2013a) and Chernozhukov et al. (2014) and suggested in high-dimensional linear regression models in Belloni et al. (2014a). Accordingly, it is possible to (i) construct simultaneous confidence bands for a potentially large number of causal parameters and (ii) adjust p-values in a test of multiple hypotheses based on the inferential procedure introduced above.

We consider a causal PLR with p_1 causal parameters of interest $\theta_{0,1}, \ldots, \theta_{0,p_1}$ associated with the treatment variables $D_1, \ldots, D_{p_1}$. The parameter of interest $\theta_{0,j}$ with $j = 1, \ldots, p_1$ solves a corresponding moment condition

$$\mathbb{E}\left[\psi_j(W; \theta_{0,j}, \eta_{0,j})\right] = 0, \tag{2.34}$$

as for example considered in **zestim**. To perform inference in a setting with multiple target coefficients $\theta_{0,j}$, the double machine learning procedure implemented in `DoubleML` iterates over the target variables of interest. During estimation of the coefficient $\theta_{0,j}$, i.e., estimating the effect of treatment D_j on Y, the remaining treatment variables enter the nuisance terms by default.

Algorithm 4: Multiplier bootstrap.

(1) **Inputs:** Use the inputs and outputs from Algorithm 1 (DML1) or Algorithm 2 (DML2) and Algorithm 3 (Variance estimation) resulting in estimates $\tilde{\theta}_{0,1}, \ldots, \tilde{\theta}_{0,p_1}$, and standard errors $\hat{\sigma}_1, \ldots \hat{\sigma}_{p_1}$.

(2) **Multiplier bootstrap:** Generate random weights ξ_i^b for each bootstrap repetition $b = 1, \ldots, B$ according to a normal (Gaussian) bootstrap, wild bootstrap or exponential bootstrap. Based on the estimated standard errors given by $\hat{\sigma}_j$ and $\hat{J}_{0,j} = \mathbb{E}_N(\psi_{a,j}(W; \eta_{0,j}))$, we obtain bootstrapped versions of the coefficients $\tilde{\theta}_j^{*,b}$ and bootstrapped t-statistics $t_j^{*,b}$ for $j = 1, \ldots, p_1$

$$\begin{aligned}
\theta_j^{*,b} &= \frac{1}{\sqrt{N}\hat{J}_{0,j}} \sum_{k=1}^{K} \sum_{i \in I_k} \xi_i^b \cdot \psi_j(W_i; \tilde{\theta}_{0,j}, \hat{\eta}_{0,j;k}), \\
t_j^{*,b} &= \frac{1}{\sqrt{N}\hat{J}_{0,j}\hat{\sigma}_j} \sum_{k=1}^{K} \sum_{i \in I_k} \xi_i^b \cdot \psi_j(W_i; \tilde{\theta}_{0,j}, \hat{\eta}_{0,j;k}).
\end{aligned}$$

(3) **Output:** Output bootstrapped coefficients and test statistics.

> **Remark 7** (*Computational efficiency*) The multiplier bootstrap procedure of Chernozhukov et al. (2013a) and Chernozhukov et al. (2014) is computatioanally efficient because it does not require resampling and reestimation of the causal parameters. Instead, it is sufficient to introduce a random pertubation of the score ψ and solve for θ_0, accordingly.

To construct simultaneous $(1-\alpha)$-confidence bands, the multiplier bootstrap presented in Algorithm 4 can be used to obtain a constant $c_{1-\alpha}$ that will guarantee asymptotic $(1-\alpha)$ coverage

$$\left[\tilde{\theta}_{0,j} \pm c_{1-\alpha} \cdot \hat{\sigma}_j/\sqrt{N}\right]. \tag{2.35}$$

The constant $c_{1-\alpha}$ is obtained in two steps.

1. Calculate the maximum of the absolute values of the bootstrapped t-statistics, $t_j^{*,b}$ in every repetition b with $b = 1, \ldots, B$.
2. Use the $(1-\alpha)$-quantile of the B maxima statistics from Step 1 as $c_{1-\alpha}$ and construct simultaneous confidence bands according to Equation (2.35).

Moreover, it is possible to derive an adjustment method for p-values obtained from a test of multiple hypotheses, including classical adjustments such as the Bonferroni correction as well as the Romano-Wolf stepdown procedure (Romano and Wolf, 2005a; Romano and Wolf, 2005b). The latter is implemented according to the algorithm for adjustment of p-values as provided in Romano and Wolf (2016) and adapted to high-dimensional linear regression based on the lasso in Bach et al. (2018b).

2.6 Implementation Details

In this section, we briefly provide information on the implementation details such as the class structure, the data-backend and the use of machine learning methods. Section 2.7 provides a demonstration of `DoubleML` in real-data and simulation examples. More information on the implementation can be found in the DoubleML User Guide, that is available online[4]. All class methods are documented in the documentation of the corresponding class, which can be browsed online[5] or, for example, by using the commands `help(DoubleML)`, `help(DoubleMLPLR)`, or `help(DoubleMLData)` in R.

2.6.1 Class Structure

The implementation of `DoubleML` for R is based on object orientation as enabled by the the `R6` package (Chang, 2020). For an introduction to object orientation in R and the `R6` package, we refer to the vignettes of the `R6` package that are available online[6], Chapter 2.1 of Becker et al. (2021), and the chapters on object orientation in Wickham (2019). The structure of the classes are presented in Figure 2.5. The abstract class `DoubleML` provides all methods for estimation and inference, for example the methods `fit()`, `bootstrap()`, `confint()`. All key components associated with estimation and inference are implemented in `DoubleML`, for example the sample splitting, the implementation of Algorithm 1 (DML1) and Algorithm 2 (DML2), the estimation of the causal parameters, and the computation of the scores $\psi(W; \theta, \eta)$. Only the model-specific properties and methods are allocated at the classes `DoubleMLPLR`

---

[4] `https://docs.doubleml.org/stable/index.html`

[5] `https://docs.doubleml.org/r/stable/`

[6] `https://r6.r-lib.org/articles/`

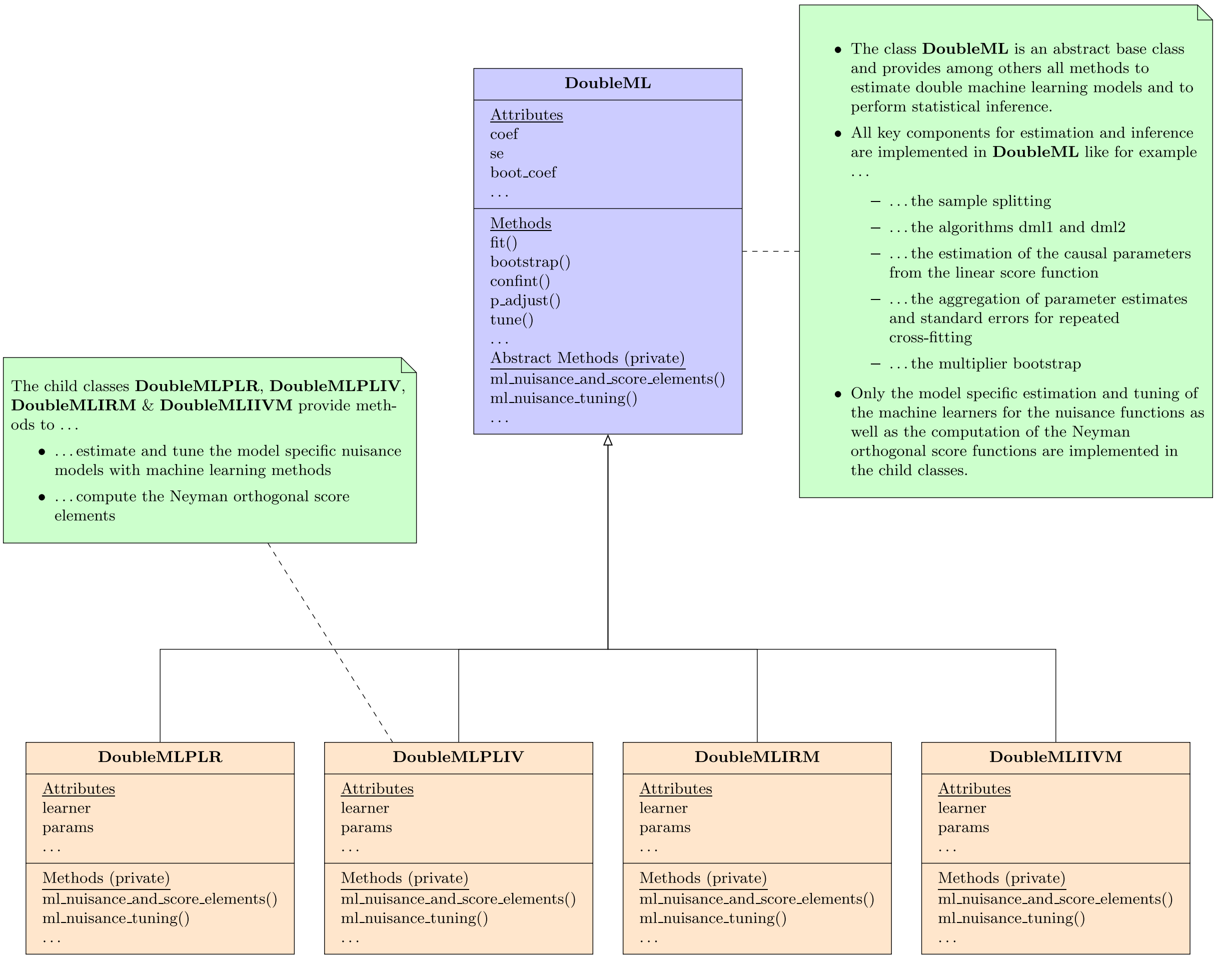

DoubleML
Attributes
coef
se
boot_coef
...
Methods
fit()
bootstrap()
confint()
p_adjust()
tune()
...
Abstract Methods (private)
ml_nuisance_and_score_elements()
ml_nuisance_tuning()
...
The class **DoubleML** is an abstract base class and provides among others all methods to estimate double machine learning models and to perform statistical inference.
All key components for estimation and inference are implemented in **DoubleML** like for example ...
...the sample splitting
...the algorithms dml1 and dml2
...the estimation of the causal parameters from the linear score function
...the aggregation of parameter estimates and standard errors for repeated cross-fitting
...the multiplier bootstrap
Only the model specific estimation and tuning of the machine learners for the nuisance functions as well as the computation of the Neyman orthogonal score functions are implemented in the child classes.
The child classes **DoubleMLPLR**, **DoubleMLPLIV**, **DoubleMLIRM** & **DoubleMLIIVM** provide methods to ...
...estimate and tune the model specific nuisance models with machine learning methods
...compute the Neyman orthogonal score elements
DoubleMLPLR
DoubleMLPLIV
DoubleMLIRM
DoubleMLIIVM
Attributes
learner
params
...
Methods (private)
ml_nuisance_and_score_elements()
ml_nuisance_tuning()
...

Figure 2.5: **Class structure of the DoubleML package for R.**

(implementing the PLR), `DoubleMLPLIV` (PLIV), `DoubleMLIRM` (IRM), and `DoubleMLIIVM` (IIVM). For example, each of the models has one or several Neyman-orthogonal score functions that are implemented for the specific child classes.

2.6.2 Data-Backend and Causal Model

The `DoubleMLData` class serves as the data-backend and implements the causal model of interest. The user is required to specify the roles of the variables in a data set at hand. Depending on the causal model considered, it is necessary to declare the dependent variable, the treatment variable(s), confounding variables(s), and, in the case of instrumental variable regression, one or multiple instruments. The data-backend can be initialized from a `data.table` (Dowle and Srinivasan, 2020). `DoubleML` provides wrappers to initialize from `data.frame` and `matrix` objects, as well.

2.6.3 Learners, Parameters and Tuning

Generally, all learners provided by the packages `mlr3`, `mlr3learners` and `mlr3extralearners` can be used for estimation of the nuisance functions of the structural models presented above. An interactive list of supported learners is available at the `mlr3extralearners` website.[7] The `mlr3extralearners` package makes it possible to add new learners, as well. The performance of the double machine learning estimator $\tilde{\theta}_0$ will depend on the predictive quality of the used estimation method. Machine learning

---

[7] `https://mlr3extralearners.mlr-org.com/articles/learners/list_learners.html`.

methods usually have several (hyper-)parameter that need to be adapted to a specific application. Tuning of model parameters can be either performed externally or internally. The latter is implemented in the method `tune()` and is further illustrated in an example in Section 2.7.6.2. Both cases build on the functionalities provided by the package `mlr3tuning`.

2.6.4 Modifications and Extensions

The flexible architecture of the `DoubleML` package allows users to modify the estimation procedure in many regards. Among others, users can provide customized sample splitting rules after initialization of the causal model via the method `set_sample_splitting()`. An example and the detailed requirements are provided in Section 2.7.7.1. Moreover, it is possible to adjust the Neyman-orthogonal score function by externally providing a customized function via the `score` option during initialization of the causal model object. A short example is presented in Section 2.7.7.2.

2.7 Estimation of Causal Parameters with `DoubleML`: Real-Data and Simulated Examples.

In this section, we will first demonstrate the use of `DoubleML` in a real-data example, which is based on data from the Pennsylvania Reemployment Bonus experiment (Bilias, 2000). This empirical example has been used in Chernozhukov et al. (2018a), as well. The goal in the empirical example is to estimate the causal parameter in a partially linear and an interactive regression model. We further provide a short example is given on how to perform simultaneous inference with `DoubleML`. Finally, we present results from a short simulation study as a brief assessment of the finite-sample performance of the implemented estimators.

2.7.1 Initialization of the Data-Backend

We begin our real-data example by downloading Pennsylvania Reemployment Bonus data set. To do so, we use the call (a connection to the internet is required).

```r
library(DoubleML)
# Load data as data.table
dt_bonus = fetch_bonus(return_type = "data.table")

# Output suppressed for the sake of brevity
dt_bonus
```

The data-backend `DoubleMLData` can be initialized from a `data.table` object by specifying the dependent variable Y via a character in `y_col`, the treatment variable(s) D in `d_cols`, and the confounders X via `x_cols`. Moreover, in IV models, an instrument can be specified via `z_cols`. In the next step, we assign the roles to the variables in the data set: `y_col = 'inuidur1'` serves as outcome variable Y, the column `d_cols = 'tg'` serves as treatment variable D and the columns `x_cols` specify the confounders.

```r
obj_dml_data_bonus = DoubleMLData$new(dt_bonus,
                            y_col = "inuidur1",
                            d_cols = "tg",
                            x_cols = c("female", "black", "othrace", "dep1", "dep2",
                                       "q2", "q3", "q4", "q5", "q6", "agelt35", "agegt54",
                                       "durable", "lusd", "husd"))
```

```
# Print data backend: Lists main attributes and methods of a DoubleMLData object
obj_dml_data_bonus
```

```
## <DoubleMLData>
##   Public:
##     all_variables: inuidur1 female black othrace dep1 dep2 q2 q3 q4 q5 q6 a ...
##     clone: function (deep = FALSE)
##     d_cols: tg
##     data: data.table, data.frame
##     data_model: data.table, data.frame
##     initialize: function (data = NULL, x_cols = NULL, y_col = NULL, d_cols = NULL,
##     n_instr: 0
##     n_obs: 5099
##     n_treat: 1
##     other_treat_cols: NULL
##     set_data_model: function (treatment_var)
##     treat_col: tg
##     use_other_treat_as_covariate: TRUE
##     x_cols: female black othrace dep1 dep2 q2 q3 q4 q5 q6 agelt35 ag ...
##     y_col: inuidur1
##     z_cols: NULL
```

```
# Print data set (output suppressed)
obj_dml_data_bonus$data
```

> **Remark 8** (*Interface for `data.frame` and `matrix`*) To initialize an instance of the class `DoubleMLData` from a `data.frame` or a collection of `matrix` objects, `DoubleML` provides the convenient wrappers `double_ml_data_from_data_frame()` and `double_ml_data_from_matrix()`. Examples can be found in the user guide and in the corresponding documentation.

2.7.2 Initialization of the Causal Model

To initialize a PLR model, we have to provide a learner for each nuisance part in the model in Equation (2.1)-(2.2). In R, this is done by providing learners to the arguments `ml_m` for nuisance part m and `ml_g` for nuisance part g. We can pass a learner as instantiated in `mlr3` and `mlr3learners`, for example a random forest as provided by the R package `ranger` (Wright and Ziegler, 2017). Previous installation of `ranger` is required. Moreover, we can specify the score (allowed choices for PLR are `'partialling out'` or `'IV-type'`) and the algorithm via the option `dml_procedure` (allowed choices `'dml1'` and `'dml2'`) . Optionally, it is possible to change the number of folds used for sample splitting through `n_folds` and the number of repetitions via `n_rep`, if the sample splitting and estimation procedure should be repeated.

```
set.seed(31415) # Required for reproducability of sample split
learner_g = lrn("regr.ranger", num.trees = 500, min.node.size = 2, max.depth = 5)
learner_m = lrn("regr.ranger", num.trees = 500, min.node.size = 2, max.depth = 5)
doubleml_bonus = DoubleMLPLR$new(obj_dml_data_bonus,
                                 ml_m = learner_m,
                                 ml_g = learner_g,
                                 score = "partialling out",
                                 dml_procedure = "dml1",
                                 n_folds = 5,
                                 n_rep = 1)
doubleml_bonus
```

```
## ================= DoubleMLPLR Object ==================
##
##
## ------------------ Data summary      ------------------
## Outcome variable: inuidur1
## Treatment variable(s): tg
## Covariates: female, black, othrace, dep1, dep2, q2, q3, q4, q5, q6, agelt35, agegt54, durable, lusd, husd
## Instrument(s):
## No. Observations: 5099
##
## ------------------ Score & algorithm ------------------
## Score function: partialling out
## DML algorithm: dml1
##
## ------------------ Machine learner   ------------------
## ml_g: regr.ranger
## ml_m: regr.ranger
##
## ------------------ Resampling        ------------------
## No. folds: 5
## No. repeated sample splits: 1
## Apply cross-fitting: TRUE
##
## ------------------ Fit summary       ------------------
##

## fit() not yet called.
```

2.7.3 Estimation of the Causal Parameter in a PLR Model

To perform estimation, call the `fit()` method. The output can be summarized using the method `summary()`.

```
doubleml_bonus$fit()
doubleml_bonus$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##    Estimate. Std. Error t value Pr(>|t|)
## tg  -0.07438    0.03543  -2.099   0.0358 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hence, we can reject the null hypothesis that $\theta_{0,tg} = 0$ at the 5% significance level. The estimated coefficient and standard errors can be accessed via the public fields `coef` and `se` of the object `doubleml_bonus`.

```
doubleml_bonus$coef
```

```
##          tg
## -0.07438411
```

```
doubleml_bonus$se
```

```
##         tg
## 0.03543316
```

After completed estimation, we can access the resulting score $\psi(W_i; \tilde{\theta}_0, \hat{\eta}_0)$ or the components $\psi_a(W_i; \hat{\eta}_0)$ and $\psi_b(W_i; \hat{\eta}_0)$. The estimated score for the first 5 observations can be obtained via.

```
# Array with dim = c(n_obs, n_rep, n_treat)
# n_obs: Number of observations in the data
# n_rep: Number of repetitions (sample splitting)
# n_treat: Number of treatment variables
doubleml_bonus$psi[1:5, 1, 1]
```

```
## [1] -0.2739454  0.7444154 -0.4509358  0.1813111 -0.3699474
```

Similarly, the components of the score $\psi_a(W_i; \hat{\eta}_0)$ and $\psi_b(W_i; \hat{\eta}_0)$ are available as public fields.

```
doubleml_bonus$psi_a[1:5, 1, 1]
```

```
## [1] -0.0981220 -0.1353987 -0.1276526 -0.4272341 -0.1126174
```

```
doubleml_bonus$psi_b[1:5, 1, 1]
```

```
## [1] -0.2812441  0.7343439 -0.4604311  0.1495317 -0.3783243
```

To construct a $(1 - \alpha)$ confidence interval, we use the `confint()` method.

```
doubleml_bonus$confint(level = 0.95)
```

```
##         2.5 %       97.5 %
## tg -0.1438318 -0.004936395
```

2.7.4 Estimation of the Causal Parameter in an IRM Model

The treatment variable D in the Pennsylvania Reemployment Bonus example is binary. Accordingly, it is possible to estimate an IRM model. Since the IRM requires estimation of the propensity score $\mathbb{P}(D|X)$, we have to specify a classifier for the nuisance part m_0.

```
# Classifier for propensity score
learner_classif_m = lrn("classif.ranger", num.trees = 500, min.node.size = 2, max.depth = 5)

doubleml_irm_bonus = DoubleMLIRM$new(obj_dml_data_bonus,
                                     ml_m = learner_classif_m,
                                     ml_g = learner_g,
                                     score = "ATE",
                                     dml_procedure = "dml1",
                                     n_folds = 5,
                                     n_rep = 1)
# Output suppressed
doubleml_irm_bonus
```

```
## fit() not yet called.
```

To perform estimation, call the `fit()` method. The output can be summarized using the method `summary()`.

```
doubleml_irm_bonus$fit()
doubleml_irm_bonus$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##    Estimate. Std. Error t value Pr(>|t|)
## tg  -0.07193    0.03554  -2.024    0.043 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated coefficient is very similar to the estimate of the PLR model and our conclusions remain unchanged.

2.7.5 Simultaneous Inference in a Simulated Data Example

We consider a simulated example of a PLR model to illustrate the use of methods for simultaneous inference. First, we will generate a sparse linear model with only three variables having a non-zero effect on the dependent variable.

```
set.seed(3141)
n_obs = 500
n_vars = 100
theta = rep(3, 3)

# Generate matrix-like objects and use the corresponding wrapper
X = matrix(stats::rnorm(n_obs * n_vars), nrow = n_obs, ncol = n_vars)
y = X[, 1:3, drop = FALSE] %*% theta  + stats::rnorm(n_obs)
df = data.frame(y, X)
```

We use the wrapper `double_ml_data_from_data_frame()` to specify a data-backend that assigns the first 10 columns of X as treatment variables and declares the remaining columns as confounders.

```
doubleml_data = double_ml_data_from_data_frame(df, y_col = "y",
                                               d_cols = c("X1", "X2", "X3",
                                                          "X4", "X5", "X6",
                                                          "X7", "X8", "X9",
                                                          "X10"))
```

```
## Set treatment variable d to X1.
```

```
# Output suppressed
doubleml_data
```

A sparse setting suggests the use of the lasso learner. Here, we use the lasso estimator with cross-validated choice of the penalty parameter λ as provided in the `glmnet` package for R (**glmnet**).

```
# Output messages during fitting are suppressed
ml_g = lrn("regr.cv_glmnet", s = "lambda.min")
ml_m  = lrn("regr.cv_glmnet", s = "lambda.min")
doubleml_plr = DoubleMLPLR$new(doubleml_data, ml_g, ml_m)
```

```
doubleml_plr$fit()
doubleml_plr$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##     Estimate. Std. Error t value Pr(>|t|)
## X1   3.017802   0.046180  65.348   <2e-16 ***
## X2   3.025812   0.042683  70.891   <2e-16 ***
## X3   3.000914   0.045849  65.452   <2e-16 ***
## X4  -0.034815   0.040955  -0.850   0.3953
## X5   0.035118   0.048132   0.730   0.4656
## X6   0.002171   0.044622   0.049   0.9612
## X7  -0.036129   0.046798  -0.772   0.4401
## X8   0.020361   0.044048   0.462   0.6439
## X9  -0.019439   0.043180  -0.450   0.6526
## X10  0.076180   0.043682   1.744   0.0812 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The multiplier bootstrap procedure can be executed using the `bootstrap()` method where the option `method` specifies the choice of the random pertubations and `n_rep_boot` the number of bootstrap repetitions.

```
doubleml_plr$bootstrap(method = "normal", n_rep_boot = 1000)
```

The resulting bootstrapped coefficients and t-statistics are available via the public fields `boot_coef` and `boot_t_stat`. To construct a simultaneous confidence interval, we set the option `joint = TRUE` when calling the `confint()` method.

```
doubleml_plr$confint(joint = TRUE)
```

```
##           2.5 %     97.5 %
## X1   2.88766757 3.14793595
## X2   2.90553386 3.14609021
## X3   2.87171334 3.13011430
## X4  -0.15022399 0.08059423
## X5  -0.10051468 0.17075155
## X6  -0.12357302 0.12791441
## X7  -0.16800517 0.09574654
## X8  -0.10376590 0.14448792
## X9  -0.14111984 0.10224143
## X10 -0.04691574 0.19927524
```

The correction of the p-values of a joint hypotheses test on the considered causal parameters is implemented in the method `p_adjust()`. By default, the adjustment procedure specified in the option `method` is the Romano-Wolf stepdown procedure.

```
doubleml_plr$p_adjust(method = "romano-wolf")
```

```
##        Estimate.  pval
## X1   3.017801759 0.000
## X2   3.025812035 0.000
## X3   3.000913821 0.000
## X4  -0.034814877 0.942
## X5   0.035118435 0.942
## X6   0.002170694 0.961
## X7  -0.036129317 0.942
## X8   0.020361010 0.951
## X9  -0.019439209 0.951
## X10  0.076179750 0.451
```

Alternatively, the correction methods provided in the `stats` function `p.adjust` can be applied, for example the Bonferroni, Bonferroni-Holm, or Benjamini-Hochberg correction. For example a Bonferroni correction could be performed by specifying `method = 'bonferroni'`.

```
doubleml_plr$p_adjust(method = "bonferroni")
```

```
##        Estimate.      pval
## X1   3.017801759 0.0000000
## X2   3.025812035 0.0000000
## X3   3.000913821 0.0000000
## X4  -0.034814877 1.0000000
## X5   0.035118435 1.0000000
## X6   0.002170694 1.0000000
## X7  -0.036129317 1.0000000
## X8   0.020361010 1.0000000
## X9  -0.019439209 1.0000000
## X10  0.076179750 0.8116808
```

2.7.6 Learners, Parameters and Tuning

The performance of the final double machine learning estimator depends on the predictive performance of the underlying ML method. First, we briefly show how externally tuned parameters can be passed to the learners in `DoubleML`. Second, it is demonstrated how the parameter tuning can be done internally by `DoubleML`.

2.7.6.1 External Tuning and Parameter Passing

Section 3 of the mlr3book (Becker et al., 2021) provides a step-by-step introduction to the powerful tuning functionalities of the `mlr3tuning` package. Accordingly, it is possible to manually reconstruct the `mlr3` regression and classification problems, which are internally handled in `DoubleML`, and to perform parameter tuning accordingly. One advantage of this procedure is that it allows users to fully exploit the powerful benchmarking and tuning tools of `mlr3` and `mlr3tuning`.

Consider the sparse regression example from above. We will briefly consider a setting where we explicitly set the parameter λ for a `glmnet` estimator rather than using the interal cross-validated choice with `cv_glmnet`.

Suppose for simplicity, some external tuning procedure resulted in an optimal value of $\lambda = 0.1$ for nuisance part m and $\lambda = 0.09$ for nuisance part g for the first treatment variable and $\lambda = 0.095$ and $\lambda = 0.085$ for the second variable, respectively. After initialization of the model object, we can set the parameter values using the method `set_ml_nuisance_params()`.

```
# Output messages during fitting are suppressed
ml_g = lrn("regr.glmnet")
ml_m  = lrn("regr.glmnet")
doubleml_plr = DoubleMLPLR$new(doubleml_data, ml_g, ml_m)
```

To set the values, we have to specify the treatment variable and the nuisance part. If no values are set, the default values are used.

```
# Note that variable names are overwritten by wrapper for matrix interface
doubleml_plr$set_ml_nuisance_params("ml_m", "X1", param = list("lambda" = 0.1))
doubleml_plr$set_ml_nuisance_params("ml_g", "X1", param = list("lambda" = 0.09))
doubleml_plr$set_ml_nuisance_params("ml_m", "X2", param = list("lambda" = 0.095))
doubleml_plr$set_ml_nuisance_params("ml_g", "X2", param = list("lambda" = 0.085))
```

All externally specified parameters are available at the public field `params`.

```
# Output omitted for the sake of brevity
str(doubleml_plr$params)
```

```
doubleml_plr$fit()
doubleml_plr$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##     Estimate. Std. Error t value Pr(>|t|)
## X1   3.041094   0.060030  50.660   <2e-16 ***
## X2   2.993916   0.054590  54.844   <2e-16 ***
## X3   2.993419   0.055144  54.283   <2e-16 ***
## X4  -0.035201   0.040637  -0.866    0.386
## X5   0.021541   0.047569   0.453    0.651
## X6  -0.006652   0.044715  -0.149    0.882
## X7  -0.039650   0.046823  -0.847    0.397
## X8   0.011146   0.044037   0.253    0.800
## X9  -0.021342   0.043237  -0.494    0.622
## X10  0.084426   0.043641   1.935    0.053 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.7.6.2 Internal Tuning and Parameter Passing

An alternative to external tuning and parameter provisioning is to perform the tuning internally. The advantage of this approach is that users do not have to specify the underlying prediction problems manually. Instead, `DoubleML` uses the underlying data-backend to ensure that the machine learning methods are tuned for the specific model under consideration and, hence, to possibly avoid mistakes. We initialize our structural model object with the learner. At this stage, we do not specify any parameters.

```
# Load required packages for tuning
library(paradox)
library(mlr3tuning)
# Set logger to omit messages during tuning and fitting
lgr::get_logger("mlr3")$set_threshold("warn")
lgr::get_logger("bbotk")$set_threshold("warn")
```

```
set.seed(1234)
ml_g = lrn("regr.glmnet")
ml_m = lrn("regr.glmnet")
doubleml_plr = DoubleMLPLR$new(doubleml_data, ml_g, ml_m)
```

To perform parameter tuning, we provide a grid of values used for evaluation for each of the nuisance parts. To set up a grid of values, we specify a named list with names corresponding to the learner names of the nuisance part (see method `learner_names()`). The elements in the list are objects of the class `ParamSet` of the `paradox` package (Lang et al., 2020b).

```
par_grids = list("ml_g" = ParamSet$new(list(
                                  ParamDbl$new("lambda", lower = 0.05, upper = 0.1))),
                 "ml_m" =  ParamSet$new(list(
                                  ParamDbl$new("lambda", lower = 0.05, upper = 0.1))))
```

The hyperparameter tuning is performed according to options passed through a named list `tune_settings`. The entries in the list specify options during parameter tuning with `mlr3tuning`:

- `terminator` is a `bbotk::Terminator` object passed to `mlr3tuning` that manages the budget to solve the tuning problem.
- `algorithm` is an object of class `mlr3tuning::Tuner` and specifies the tuning algorithm. Alternatively, algorithm can be a `character()` that is used as an argument in the wrapper `mlr3tuning` call `tnr(algorithm)`. The `Tuner` class in `mlr3tuning` supports grid search, random search, generalized simulated annealing and non-linear optimization.
- `rsmp_tune` is an object of class `resampling` object that specifies the resampling method for evaluation, for example `rsmp('cv', folds = 5)` implements 5-fold cross-validation. `rsmp('holdout', ratio = 0.8)` implements an evaluation based on a hold-out sample that contains 20 percent of the observations. By default, 5-fold cross-validation is performed.
- `measure` is a named list containing the measures used for tuning of the nuisance components. The names of the entries must match the learner names (see method `learner_names()`). The entries in the list must either be objects of class `Measure` or keys passed to `msr()`. If `measure` is not provided by the user, the mean squared error is used for regression models and the classification error for binary outcomes, by default.

In the next code chunk, the value of the parameter λ is tuned via grid search in the range 0.05 to 0.1 at a resolution of 11.[8] To evaluate the predictive performance in both nuisance parts, the cross-validated mean squared error is used.

```
# Provide tune settings
tune_settings = list(terminator = trm("evals", n_evals = 100),
                     algorithm = tnr("grid_search", resolution = 11),
                     rsmp_tune = rsmp("cv", folds = 5),
                     measure = list("ml_g" = msr("regr.mse"),
                                    "ml_m" = msr("regr.mse")))
```

With these parameters we can run the tuning by calling the `tune` method for `DoubleML` objects.

---

[8] The resulting grid has 11 equally spaced values ranging from a minimum value of 0.05 to a maximum value of 0.1. Type `generate_design_grid(par_grids$ml_g, resolution = 11)` to access the grid for nuisance part `ml_g`.

```
# Execution might take around 50 seconds
# Tune
doubleml_plr$tune(param_set = par_grids, tune_settings = tune_settings)

# Output omitted for the sake of brevity, available in the Appendix

# Acces tuning results for target variable "X1"
doubleml_plr$tuning_res$X1

# Tuned parameters
str(doubleml_plr$params)
```

```
# Estimate model and summary
doubleml_plr$fit()
doubleml_plr$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##      Estimate. Std. Error t value Pr(>|t|)
## X1   3.028980   0.059701  50.736   <2e-16 ***
## X2   3.008650   0.054301  55.407   <2e-16 ***
## X3   2.960571   0.053082  55.773   <2e-16 ***
## X4  -0.037859   0.040976  -0.924   0.3555
## X5   0.030018   0.047880   0.627   0.5307
## X6   0.003451   0.044419   0.078   0.9381
## X7  -0.025875   0.046936  -0.551   0.5814
## X8   0.022008   0.044172   0.498   0.6183
## X9  -0.014251   0.043765  -0.326   0.7447
## X10  0.088653   0.043691   2.029   0.0424 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By default, the parameter tuning is performed on the whole sample, for example in the case of K_{tune}-fold cross-validation, the entire sample is split into K_{tune} folds for evaluation of the cross-validated error. Alternatively, each of the K folds used in the cross-fitting procedure could be split up into K_{tune} subfolds that are then used for evaluation of the candidate models. As a result, the choice of the tuned parameters will be fold-specific. To perform fold-specific tuning, users can set the option `tune_on_folds = TRUE` when calling the method `tune()`.

2.7.7 Specifications and Modifications of Double Machine Learning

The flexible architecture of the `DoubleML` package allows users to modify the estimation procedure in many regards. We will shortly present two examples on how users can adjust the double machine learning framework to their needs in terms of the sample splitting procedure and the score function.

2.7.7.1 Sample Splitting

By default, `DoubleML` performs cross-fitting as presented in Algorithms 1 and 2. Alternatively, all implemented models allow a partition to be provided externally via the method `set_sample_splitting()`. Note that by setting `draw_sample_splitting = FALSE` one can prevent that a partition is drawn during initialization of the model object. The following calls are equivalent. In the first sample code, we use the standard interface and draw the sample-splitting with $K = 4$ folds during initialization of the `DoubleMLPLR` object.

```
# First generate some data, ml learners and a data-backend
learner = lrn("regr.ranger", num.trees = 100, mtry = 20, min.node.size = 2, max.depth = 5)
ml_g = learner
ml_m = learner
data = make_plr_CCDDHNR2018(alpha = 0.5, n_obs = 100, return_type = "data.table")
doubleml_data = DoubleMLData$new(data,
                                 y_col = "y",
                                 d_cols = "d")
```

```
set.seed(314)
doubleml_plr_internal = DoubleMLPLR$new(doubleml_data, ml_g, ml_m, n_folds = 4)
doubleml_plr_internal$fit()
doubleml_plr_internal$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##   Estimate. Std. Error t value Pr(>|t|)
## d    0.4892     0.1024   4.776 1.79e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the second sample code, we manually specify a sampling scheme using the `mlr3::Resampling` class. Alternatively, users can provide a nested list that has the following structure:

- The length of the outer list must match with the desired number of repetitions of the sample-splitting, i.e., `n_rep`.
- The inner list is a named list of length 2 specifying the `test_ids` and `train_ids`. The named entries `test_ids` and `train_ids` are lists of the same length.
  - `train_ids` is a list of length `n_folds` that specifies the indices of the observations used for model fitting in each fold.
  - `test_ids` is a list of length `n_folds` that specifies the indices of the observations used for calculation of the score in each fold.

```
doubleml_plr_external = DoubleMLPLR$new(doubleml_data, ml_g, ml_m,
                                        draw_sample_splitting = FALSE)

set.seed(314)
# Set up a task and cross-validation resampling scheme in mlr3
my_task = Task$new("help task", "regr", data)
my_sampling = rsmp("cv", folds = 4)$instantiate(my_task)

train_ids = lapply(1:4, function(x) my_sampling$train_set(x))
test_ids = lapply(1:4, function(x) my_sampling$test_set(x))
smpls = list(list(train_ids = train_ids, test_ids = test_ids))

# Structure of the specified sampling scheme
str(smpls)
```

```
## List of 1
##  $ :List of 2
##   ..$ train_ids:List of 4
##   .. ..$ : int [1:75] 1 7 11 18 19 20 21 31 32 37 ...
##   .. ..$ : int [1:75] 10 15 16 22 26 35 38 40 41 46 ...
##   .. ..$ : int [1:75] 10 15 16 22 26 35 38 40 41 46 ...
##   .. ..$ : int [1:75] 10 15 16 22 26 35 38 40 41 46 ...
##   ..$ test_ids :List of 4
##   .. ..$ : int [1:25] 10 15 16 22 26 35 38 40 41 46 ...
##   .. ..$ : int [1:25] 1 7 11 18 19 20 21 31 32 37 ...
##   .. ..$ : int [1:25] 3 5 6 8 17 24 25 28 29 34 ...
##   .. ..$ : int [1:25] 2 4 9 12 13 14 23 27 30 33 ...
```

```
# Fit model
doubleml_plr_external$set_sample_splitting(smpls)
doubleml_plr_external$fit()
doubleml_plr_external$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##   Estimate. Std. Error t value Pr(>|t|)
## d    0.4892     0.1024   4.776 1.79e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Setting the option `apply_cross_fitting = FALSE` at the instantiation of the causal model allows double machine learning being performed without cross-fitting. It results in randomly splitting the sample into two parts. The first half of the data is used for the estimation of the nuisance models with the machine learning methods and the second half for estimating the causal parameter, i.e., solution of the score. Note that cross-fitting performs well empirically and is recommended to remove bias induced by overfitting. Moreover, cross-fitting allows to exploit full efficiency: Every fold is used once for training the ML methods and once for estimation of the score (Chernozhukov et al., 2018a, pp. 6). A short example on the efficiency gains associated with cross-fitting is provided in Section 2.7.8.1.

2.7.7.2 Score Function

Users may want to adjust the score function $\psi(W;\theta_0,\eta_0)$, for example, to adjust the DML estimators in terms of a re-weighting. An alternative to the choices provided in `DoubleML` is to pass a function via the argument `score` during initialization of the model object. The following examples are equivalent. In the first example, we use the score option `'partialling out'` for the PLR model whereas in the second case, we explicitly provide a function that implements the same score. The arguments used in the function refer to the internal objects that implement the theoretical quantities in Equation (2.17).

```
# Use score "partialling out"
set.seed(314)
doubleml_plr_partout = DoubleMLPLR$new(doubleml_data, ml_g, ml_m, score = "partialling out")
doubleml_plr_partout$fit()
doubleml_plr_partout$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##   Estimate. Std. Error t value Pr(>|t|)
## d    0.5108     0.0959   5.326    1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We define the function that implements the same score and specify the argument `score` accordingly. The function must return a named list with entries `psi_a` and `psi_b` to pass values for computation of the score.

```
# Here:
# y: Dependent variable
# d: Treatment variable
# g_hat: Predicted values from regression of Y on X's
# m_hat: Predicted values from regression of D on X's
# smpls: Sample split under consideration, can be ignored in this example
score_manual = function(y, d, g_hat, m_hat, smpls) {
  resid_y = y - g_hat
  resid_d = d - m_hat

  psi_a = -1 * resid_d * resid_d
  psi_b = resid_d * resid_y
  psis = list(psi_a = psi_a, psi_b = psi_b)
  return(psis)
}
```

```
set.seed(314)
doubleml_plr_manual = DoubleMLPLR$new(doubleml_data, ml_g, ml_m, score = score_manual)
doubleml_plr_manual$fit()
doubleml_plr_manual$summary()
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##   Estimate. Std. Error t value Pr(>|t|)
## d    0.5108     0.0959   5.326    1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.7.8 A Short Simulation Study

To illustrate the validity of the implemented double machine learning estimators, we perform a brief simulation study.

2.7.8.1 The Role of Cross-Fitting

As mentioned in Section 2.7.7.1 the use of the cross-fitting Algorithms 1 (DML1) and 2 (DML2) makes it possible to use sample splitting and exploit full efficiency at the same time. To illustrate the superior performance due to cross-fitting, we compare the double machine learning estimator with and without a cross-fitting procedure in the simulation setting that was presented in 2.4.1. Figure 2.6 illustrates that efficiency gains can be achieved if the role of the random partitions is swapped in the estimation procedure. Using cross-fitting makes it possible to obtain smaller standard errors for the DML estimator: The empirical distribution of the double machine learning estimator that is based on the cross-fitting Algorithm 2 (DML2) exhibits a more pronounced concentration around zero.

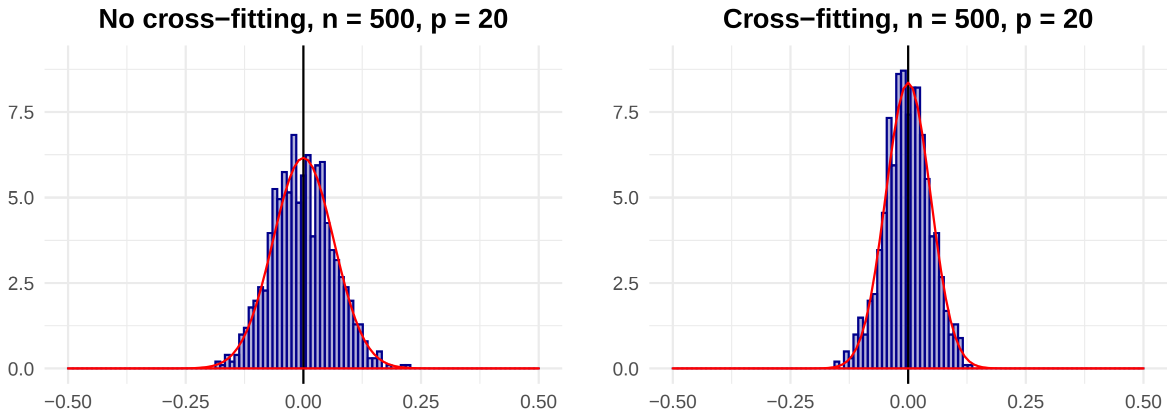


Figure 2.6: **Illustration of efficiency gains due to the use of cross-fitting.**

Left panel: Histogram of the centered dml estimator without cross-fitting, $\tilde{\theta}_0^{nocf} - \theta_0$. $\hat{\theta}_0^{nocf}$ is the double machine learning estimator obtained from a sample split into two folds. One fold is used for estimation of the nuisance parameters and the second fold is used for evaluation of the score function and estimation. The empirical distribution can be well-approximated by a normal distribution as indicated by the red curve. **Right panel:** Histogram of the centered dml estimator with cross-fitting, $\tilde{\theta}_0 - \theta_0$. The estimator is obtained from a split into two folds and application of Algorithm 2 (DML2). In both cases, the estimators are based on estimation of g_0 and m_0 with random forests and an orthogonal score function provided in Equation (2.17). Moreover, exactly the same data sets and exactly the same partitions are used for sample splitting. The empirical distribution of the estimator that is based on cross-fitting exhibits a more pronounced concentration around zero, which reflects the smaller standard errors.

2.7.8.2 Inference on a Structural Parameter in Key Causal Models

We provide simulation results for double machine learning estimators in the presented key causal models in Figure 2.7. In a replication of the simulation example in Section 2.4.1, we show that the confidence intervals for the DML estimator in the partially linear regression model achieves an empirical coverage close to the specified level of $1 - \alpha = 0.95$. The estimator is, again, based on a random forest learner. The corresponding results are presented in the top-left panel of Figure 2.7.

In a simulated example of a PLIV model, the DML confidence interval that is based on a lasso learner (`regr.cv_glmnet` of `mlr3`) achieves a coverage of 94.4%. The underlying data generating process is based on a setting considered in Chernozhukov et al. (2015a) with one instrumental variable. Moreover for simulations of the IRM model, we make use of a DGP of Belloni et al. (2017). The DGP for the IIVM is inspired by a simulation run in Farbmacher et al. (2020). We present the formal DGPs in the Appendix. To perform estimation of the nuisance parts in the interactive models, we employ the regression and classification predictors `regr.cv_glmnet` and `classif.cv_glmnet` as provided by the `mlr3` package. In all cases, we employ the cross-validated `lambda.min` choice of the penalty parameter with five folds, in other words, that λ value that minimizes the cross-validated mean squared error. Figure 2.7 shows that the empirical distribution of the centered estimators as obtained in finite sample settings is relatively well-approximated by a normal distribution. In all models the empirical coverage that is achieved by the constructed confidence bands is close to the nominal level.

2.7.8.3 Simultaneous Inference

To verify the finite-sample performance of the implemented methods for simultaneous inference, we perform a small simulation study in a regression setup which is similar as the one used in Bach et al. (2018b). We would like to perform valid simultaneous inference on the coefficients θ in the regression

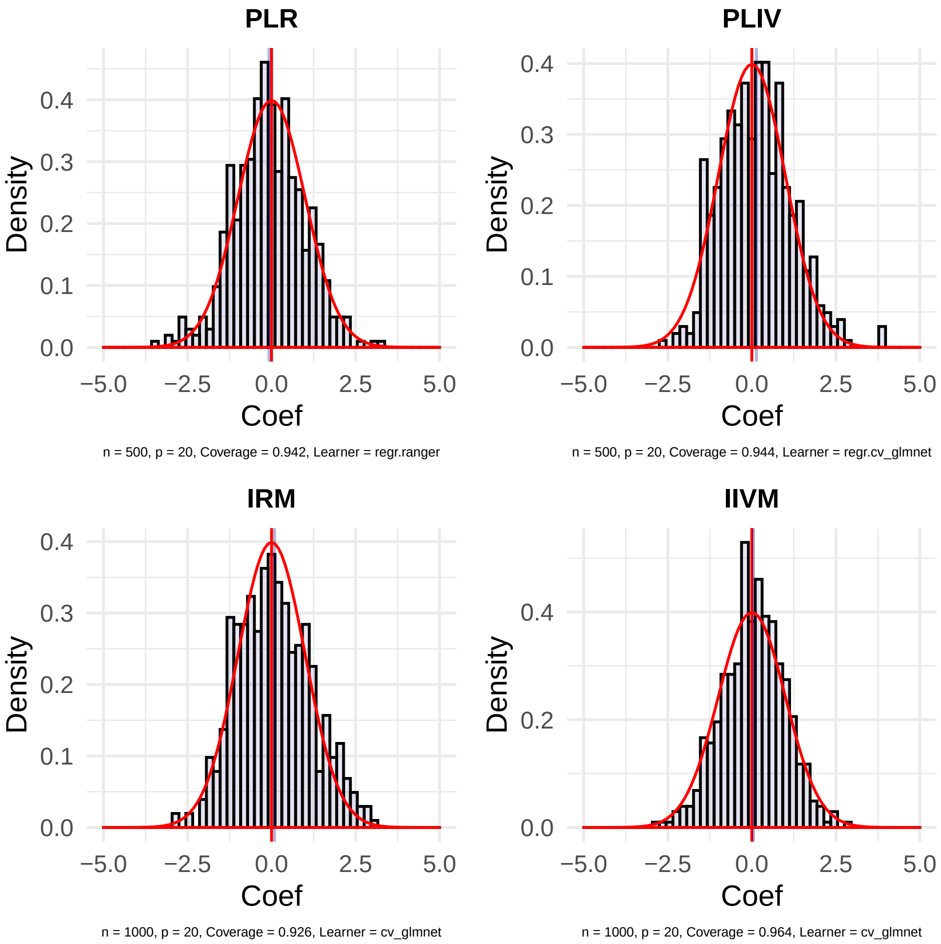


Figure 2.7: **Histogram of double machine learning estimators in key causal models.**

The figure shows the histograms of the realizations of the DML estimators in the PLR (top left), PLIV (top right), IRM (bottom left), and IIVM (bottom right) as obtained in $R = 500$ independent repetitions. Additional information on the data generating processes and implemented estimators are presented in the main text and the Appendix.

model

$$y_i = \beta_0 + d_i'\theta + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.36}$$

with $n = 1000$ and $p_1 = 42$ regressors. The errors ε_i are normally distributed with $\varepsilon_i \sim N(0, \sigma^2)$ and variance $\sigma^2 = 3$. The regressors d_i are generated by a joint normal distribution $d_i \sim N(\mu, \Sigma)$ with $\mu = \mathbf{0}$ and $\Sigma_{j,k} = 0.5^{|j-k|}$. The model is sparse in that only the first $s = 12$ regressors have a non-zero effect on outcome y_i. The p_1 coefficients $\theta_1, \ldots, \theta_{p_1}$ are generated as

$$\theta_j = \min\left\{\frac{\theta^{\max}}{j^a}, \theta^{\min}\right\},$$

for $j = 1, \ldots, s$ with $\theta^{\max} = 9$, $\theta^{min} = 0.75$, and $a = 0.99$. All other coefficients have values exactly equal to 0. Estimation of the nuisance components has been performed by using the lasso as provided by `regr.cv_glmnet` in `mlr3`.

We report the empirical coverage as achieved by a joint $(1 - \alpha)$-confidence interval for all $p_1 = 42$ coefficients and the realized family-wise error rate of the implemented p-value adjustments in $R = 500$ repetitions in Table 2.1. The finite sample performance of the Romano-Wolf stepdown procedure that is

	CI	RW	Bonf	Holm
FWER	0.09	0.11	0.09	0.10
Cor. Rejections	12.00	12.00	12.00	12.00

Table 2.1: **Family-wise error rate and average number of correct rejections in a simulation example.**

based on the multiplier bootstrap as well as the classical Bonferroni and Bonferroni-Holm correction are evaluated. Table 2.1 shows that all methods achieve an empirical FWER close to the specified level of $\alpha = 0.1$. In all cases, the double machine learning estimators reject all 12 false null hypotheses in every repetition.

2.8 Conclusion

In this paper, we provide an overview on the key ingredients and the major structure of the double/debiased machine learning framework as established in Chernozhukov et al. (2018a) together with an overview on a collection of structural models. Moreover, we introduce the R package `DoubleML` that serves as an implementation of the double machine learning approach. A brief simulation study provides insights on the finite sample performance of the double machine learning estimator in the key causal models.

The structure of `DoubleML` is intended to be flexible with regard to the implemented structural models, the resampling scheme, the machine learning methods and the underlying algorithm, as well as the Neyman-orthogonal scores considered. By providing the R package `DoubleML` together with its Python twin (Bach et al., 2021), we hope to make double machine learning more accessible to users in practice. Finally, we would like to encourage users to add new structural models, scores and functionalities to the package.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number 431701914.

2.9 Appendix

2.9.1 Computation and Infrastructure

The simulation study has been run on a x86_64-w64-mingw32/x64 (64-bit) (Windows 10 x64 (build 19041)) system using R version 3.6.3 (2020-02-29). The following packages have been used for estimation:

- `DoubleML`, version 0.1.2,
- `data.table`, version 1.13.2,
- `mlr3`, version 0.8.0,
- `mlr3tuning`, version 0.6.0,
- `mlr3learners`, version 0.4.2,
- `glmnet`, version 3.0.2,
- `ranger`, version 0.12.1,
- `paradox`, version 0.7.0
- `foreach`, version 1.5.1.

2.9.2 Suppressed Code Output

Pennsylvania Reemployment Data, Section 2.7

```r
library(DoubleML)
# Load data as data.table
dt_bonus = fetch_bonus(return_type = "data.table")
dt_bonus
```

```
##       inuidur1 female black othrace dep1 dep2 q2 q3 q4 q5 q6 agelt35 agegt54
##    1: 2.890372      0     0       0    0    1  0  0  0  1  0       0       0
##    2: 0.000000      0     0       0    0    0  0  0  0  1  0       0       0
##    3: 3.295837      0     0       0    0    0  0  0  1  0  0       0       0
##    4: 2.197225      0     0       0    0    0  0  1  0  0  0       1       0
##    5: 3.295837      0     0       0    1    0  0  0  0  1  0       0       1
##   ---
## 5095: 2.302585      0     0       0    0    0  0  1  0  0  0       1       0
## 5096: 1.386294      0     0       0    0    1  1  0  0  0  0       0       0
## 5097: 2.197225      0     0       0    0    1  1  0  0  0  0       1       0
## 5098: 1.386294      0     0       0    0    0  0  0  0  1  0       0       1
## 5099: 3.295837      0     0       0    0    0  0  0  1  0  0       0       1
##       durable lusd husd tg
##    1:       0    0    1  0
##    2:       0    1    0  0
##    3:       0    1    0  0
##    4:       0    0    0  1
##    5:       1    1    0  0
##   ---
## 5095:       0    0    0  1
## 5096:       0    0    0  1
## 5097:       0    1    0  0
## 5098:       0    0    0  1
## 5099:       1    1    0  0
```

```r
obj_dml_data_bonus = DoubleMLData$new(dt_bonus,
                                      y_col = "inuidur1",
                                      d_cols = "tg",
                                      x_cols = c("female", "black", "othrace", "dep1", "dep2",
                                                 "q2", "q3", "q4", "q5", "q6", "agelt35", "agegt54",
                                                 "durable", "lusd", "husd"))

# Print data backend: Lists main attributes and methods of a DoubleMLData object
obj_dml_data_bonus
```

```r
# Print data set (output suppressed)
obj_dml_data_bonus$data
```

```
##       inuidur1 female black othrace dep1 dep2 q2 q3 q4 q5 q6 agelt35 agegt54
##    1: 2.890372      0     0       0    0    1  0  0  0  1  0       0       0
##    2: 0.000000      0     0       0    0    0  0  0  0  1  0       0       0
##    3: 3.295837      0     0       0    0    0  0  0  1  0  0       0       0
##    4: 2.197225      0     0       0    0    0  0  1  0  0  0       1       0
##    5: 3.295837      0     0       0    1    0  0  0  0  1  0       0       1
##   ---
## 5095: 2.302585      0     0       0    0    0  0  1  0  0  0       1       0
## 5096: 1.386294      0     0       0    0    1  1  0  0  0  0       0       0
## 5097: 2.197225      0     0       0    0    1  1  0  0  0  0       1       0
## 5098: 1.386294      0     0       0    0    0  0  0  0  1  0       0       1
## 5099: 3.295837      0     0       0    0    0  0  0  1  0  0       0       1
##       durable lusd husd tg
##    1:       0    0    1  0
##    2:       0    1    0  0
##    3:       0    1    0  0
##    4:       0    0    0  1
##    5:       1    1    0  0
##   ---
## 5095:       0    0    0  1
## 5096:       0    0    0  1
## 5097:       0    1    0  0
## 5098:       0    0    0  1
## 5099:       1    1    0  0
```

```r
learner_classif_m = lrn("classif.ranger", num.trees = 500, min.node.size = 2, max.depth = 5)

doubleml_irm_bonus = DoubleMLIRM$new(obj_dml_data_bonus,
                                     ml_m = learner_classif_m,
                                     ml_g = learner_g,
                                     score = "ATE",
                                     dml_procedure = "dml1",
                                     n_folds = 5,
                                     n_rep = 1)
# Output suppressed
doubleml_irm_bonus
```

```
## ================= DoubleMLIRM Object ==================
##
##
## ------------------ Data summary      ------------------
## Outcome variable: inuidur1
## Treatment variable(s): tg
## Covariates: female, black, othrace, dep1, dep2, q2, q3, q4, q5, q6, agelt35, agegt54, durable, lusd, husd
## Instrument(s):
## No. Observations: 5099
##
## ------------------ Score & algorithm ------------------
## Score function: ATE
## DML algorithm: dml1
##
## ------------------ Machine learner   ------------------
## ml_g: regr.ranger
## ml_m: classif.ranger
##
## ------------------ Resampling        ------------------
## No. folds: 5
## No. repeated sample splits: 1
## Apply cross-fitting: TRUE
##
## ------------------ Fit summary       ------------------
##

## fit() not yet called.
```

Data-backend with multiple treatment variables, Section 2.7.5

```
doubleml_data = double_ml_data_from_data_frame(df, y_col = "y",
                                               d_cols = c("X1", "X2", "X3",
                                                          "X4", "X5", "X6",
                                                          "X7", "X8", "X9", "X10"))
```

```
## Set treatment variable d to X1.
```

```
# Output suppressed
doubleml_data
```

```
## <DoubleMLData>
##   Public:
##     all_variables: X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24  ...
##     clone: function (deep = FALSE)
##     d_cols: X1 X2 X3 X4 X5 X6 X7 X8 X9 X10
##     data: data.table, data.frame
##     data_model: data.table, data.frame
##     initialize: function (data = NULL, x_cols = NULL, y_col = NULL, d_cols = NULL,
##     n_instr: 0
##     n_obs: 500
##     n_treat: 10
##     other_treat_cols: X2 X3 X4 X5 X6 X7 X8 X9 X10
##     set_data_model: function (treatment_var)
##     treat_col: X1
##     use_other_treat_as_covariate: TRUE
##     x_cols: X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24  ...
##     y_col: y
##     z_cols: NULL
```

List of externally provided parameters, Section 2.7.6.1

```
# Output: Parameters after external tuning

# Tuned parameters
str(doubleml_plr$params)
```

```
## List of 2
##  $ ml_g:List of 10
##   ..$ X1 :List of 1
##   .. ..$ lambda: num 0.09
##   ..$ X2 :List of 1
##   .. ..$ lambda: num 0.085
##   ..$ X3 : NULL
##   ..$ X4 : NULL
##   ..$ X5 : NULL
##   ..$ X6 : NULL
##   ..$ X7 : NULL
##   ..$ X8 : NULL
##   ..$ X9 : NULL
##   ..$ X10: NULL
##  $ ml_m:List of 10
##   ..$ X1 :List of 1
##   .. ..$ lambda: num 0.1
##   ..$ X2 :List of 1
##   .. ..$ lambda: num 0.095
##   ..$ X3 : NULL
##   ..$ X4 : NULL
##   ..$ X5 : NULL
##   ..$ X6 : NULL
##   ..$ X7 : NULL
##   ..$ X8 : NULL
##   ..$ X9 : NULL
##   ..$ X10: NULL
```

List of internally tuned parameters, Section 2.7.6.2

```
# Output: parameters after internal tuning

# Access tuning results for target variable "X1"
doubleml_plr$tuning_res$X1
```

```
## $ml_g
## $ml_g[[1]]
## $ml_g[[1]]$tuning_result
## $ml_g[[1]]$tuning_result[[1]]
## $ml_g[[1]]$tuning_result[[1]]$tuning_result
##    lambda learner_param_vals  x_domain regr.mse
## 1:    0.1          <list[2]> <list[1]> 10.53451
##
## $ml_g[[1]]$tuning_result[[1]]$tuning_archive
##     lambda regr.mse                                uhash  x_domain
##  1:  0.100 10.53451 8c505081-e55c-41e8-88cd-75775716f9e8 <list[1]>
##  2:  0.095 10.60720 a0e8fb73-f402-4161-a620-ec59a40ac211 <list[1]>
##  3:  0.085 10.76577 9c32228b-eb8f-4363-aa21-bd8d1bae3531 <list[1]>
##  4:  0.055 11.32053 c7232bf6-0da0-426e-a208-815158954990 <list[1]>
##  5:  0.060 11.21736 103214b9-5fff-4649-b8b6-b05e22d13e40 <list[1]>
##  6:  0.050 11.42918 4925423b-a0f5-4641-bba8-7e4907039aff <list[1]>
##  7:  0.075 10.93077 6e0509a3-0701-44ed-9470-c2c8ff422fd1 <list[1]>
##  8:  0.065 11.11709 1b5c8d81-3045-4a65-908e-c422ff5c62d3 <list[1]>
##  9:  0.080 10.84518 53cc2cc5-9c74-4d9e-a2e5-27d0a2857cc5 <list[1]>
## 10:  0.070 11.02168 122e1304-9b36-4bee-ac09-d91aeb2d6f0b <list[1]>
## 11:  0.090 10.68576 525aafa7-44b9-49a6-a489-3e8a879c831d <list[1]>
##               timestamp batch_nr
##  1: 2021-04-14 15:08:24        1
##  2: 2021-04-14 15:08:24        2
##  3: 2021-04-14 15:08:24        3
##  4: 2021-04-14 15:08:25        4
##  5: 2021-04-14 15:08:25        5
##  6: 2021-04-14 15:08:25        6
##  7: 2021-04-14 15:08:25        7
##  8: 2021-04-14 15:08:25        8
##  9: 2021-04-14 15:08:26        9
## 10: 2021-04-14 15:08:26       10
## 11: 2021-04-14 15:08:26       11
##
## $ml_g[[1]]$tuning_result[[1]]$params
## NULL
##
##
##
## $ml_g[[1]]$params
## $ml_g[[1]]$params[[1]]
## $ml_g[[1]]$params[[1]]$family
## [1] "gaussian"
##
## $ml_g[[1]]$params[[1]]$lambda
## [1] 0.1
##
##
##
##
## $ml_g$params
## $ml_g$params[[1]]
## $ml_g$params[[1]]$family
## [1] "gaussian"
##
```

```
## $ml_g$params[[1]]$lambda
## [1] 0.1
##
##
##
##
## $ml_m
## $ml_m[[1]]
## $ml_m[[1]]$tuning_result
## $ml_m[[1]]$tuning_result[[1]]
## $ml_m[[1]]$tuning_result[[1]]$tuning_result
##    lambda learner_param_vals  x_domain  regr.mse
## 1:    0.1          <list[2]> <list[1]> 0.9794034
##
## $ml_m[[1]]$tuning_result[[1]]$tuning_archive
##     lambda  regr.mse                                uhash  x_domain
##  1:  0.090 0.9798230 11cb948a-0c49-4b6c-84dc-c319efb68de1 <list[1]>
##  2:  0.055 0.9971462 e9670de1-9d03-466b-86a9-dadbc1c072f9 <list[1]>
##  3:  0.075 0.9830963 2bd3b01e-dbd2-4784-9e0d-1ff673db1082 <list[1]>
##  4:  0.050 1.0045139 354a789a-bbe2-4186-b37c-54601f123e72 <list[1]>
##  5:  0.100 0.9794034 0d7f7c18-2d79-4c9a-8e62-4b3cebfd7387 <list[1]>
##  6:  0.060 0.9907519 2b710f09-82bf-4313-be56-abcef4fe3e8a <list[1]>
##  7:  0.065 0.9869171 89f7957b-4b73-4c82-a096-3d8939ac2dc3 <list[1]>
##  8:  0.095 0.9797396 456b1032-3c6a-4335-ab03-7872cb422455 <list[1]>
##  9:  0.085 0.9804282 1d496c32-cadf-4ea6-8866-c054b8af0aa5 <list[1]>
## 10:  0.070 0.9848766 a9b64f70-60d7-4c15-90f0-596bed74249b <list[1]>
## 11:  0.080 0.9813190 d7c2064c-dde6-4c95-8f9f-431945db2b0c <list[1]>
##               timestamp batch_nr
##  1: 2021-04-14 15:08:26        1
##  2: 2021-04-14 15:08:27        2
##  3: 2021-04-14 15:08:27        3
##  4: 2021-04-14 15:08:27        4
##  5: 2021-04-14 15:08:27        5
##  6: 2021-04-14 15:08:27        6
##  7: 2021-04-14 15:08:28        7
##  8: 2021-04-14 15:08:28        8
##  9: 2021-04-14 15:08:28        9
## 10: 2021-04-14 15:08:28       10
## 11: 2021-04-14 15:08:29       11
##
## $ml_m[[1]]$tuning_result[[1]]$params
## NULL
##
##
##
## $ml_m[[1]]$params
## $ml_m[[1]]$params[[1]]
## $ml_m[[1]]$params[[1]]$family
## [1] "gaussian"
##
## $ml_m[[1]]$params[[1]]$lambda
## [1] 0.1
##
##
##
##
## $ml_m$params
## $ml_m$params[[1]]
## $ml_m$params[[1]]$family
## [1] "gaussian"
##
## $ml_m$params[[1]]$lambda
## [1] 0.1
```

```
# Tuned parameters
str(doubleml_plr$params)
```

```
## List of 2
##  $ ml_g:List of 10
##   ..$ X1 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
##   ..$ X2 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
##   ..$ X3 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
##   ..$ X4 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.09
##   ..$ X5 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.07
##   ..$ X6 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.085
##   ..$ X7 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.085
##   ..$ X8 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.08
##   ..$ X9 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.09
##   ..$ X10:List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.075
##  $ ml_m:List of 10
##   ..$ X1 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
##   ..$ X2 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.095
##   ..$ X3 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.095
##   ..$ X4 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.095
##   ..$ X5 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
##   ..$ X6 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
##   ..$ X7 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
##   ..$ X8 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
##   ..$ X9 :List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
##   ..$ X10:List of 2
##   .. ..$ family: chr "gaussian"
##   .. ..$ lambda: num 0.1
```

2.9.3 Additional Data Generating Processes, Simulation Study

Data generating process for PLIV simulation
The DGP is based on Chernozhukov et al. (2015a) and defined as

$$\begin{aligned} z_i &= \Pi x_i + \zeta_i, \\ d_i &= x_i'\gamma + z_i'\delta + u_i, \\ y_i &= \alpha d_i + x_i'\beta + \varepsilon_i, \end{aligned} \tag{2.37}$$

with

$$\begin{pmatrix} \varepsilon_i \\ u_i \\ \zeta_i \\ x_i \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0.6 & 0 & 0 \\ 0.6 & 1 & 0 & 0 \\ 0 & 0 & 0.25 I_{p_n^z} & 0 \\ 0 & 0 & 0 & \Sigma \end{pmatrix}\right)$$

where Σ is a $p_n^x \times p_n^x$ matrix with entries $\Sigma_{kj} = 0.5^{|k-j|}$ and $I_{p_n^z}$ is an identity matrix with dimension $p_n^z \times p_n^z$. $\beta = \gamma$ is a p_n^x-vector with entries $\beta = \frac{1}{j^2}$ and $\Pi = (I_{p_n^z}, 0_{p_n^z \times (p_n^x - p_n^z)})$. In the simulation example, we have one instrument, i.e., $p_n^z = 1$ and $p_n^x = 20$ regressors x_i. In the simulation study, data sets with $n = 500$ observations are generated in $R = 500$ independent repetitions.

Data generating process for IRM simulation
The DGP is based on a simulation study in Belloni et al. (2017) and defined as

$$\begin{aligned} d_i &= 1\left\{\frac{\exp(c_d x_i'\beta)}{1 + \exp(c_d x_i'\beta)} > v_i\right\}, & v_i &\sim \mathcal{U}(0,1), \\ y_i &= \theta d_i + c_y x_i'\beta d_i + \zeta_i, & \zeta_i &\sim \mathcal{N}(0,1), \end{aligned} \tag{2.38}$$

with covariates $x_i \sim \mathcal{N}(0, \Sigma)$ where Σ is a matrix with entries $\Sigma_{kj} = 0.5^{|k-j|}$. β is a p_x-dimensional vector with entries $\beta_j = \frac{1}{j^2}$ and the constants c_y and c_d are determined as

$$c_y = \sqrt{\frac{R_y^2}{(1 - R_y^2)\beta'\Sigma\beta}}, \qquad c_d = \sqrt{\frac{(\pi^2/3)R_d^2}{(1 - R_d^2)\beta'\Sigma\beta}}.$$

We set the values of $R_y = 0.5$ and $R_d = 0.5$ and consider a setting with $n = 1000$ and $p = 20$. Data generation and estimation have been performed in $R = 500$ independent replications.

Data generating process for IIVM simulation
The DGP is defined as

$$\begin{aligned} d_i &= 1\left\{\alpha_x Z + v_i > 0\right\}, \\ y_i &= \theta d_i + x_i'\beta + u_i, \end{aligned} \tag{2.39}$$

with $Z \sim$ Bernoulli(0.5) and

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right).$$

The covariates are drawn from a multivariate normal distribution with $x_i \sim \mathcal{N}(0, \Sigma)$ with entries of the matrix Σ being $\Sigma_{kj} = 0.5^{|j-k|}$ and β being a p_x-dimensional vector with $\beta_j = \frac{1}{\beta^2}$. The data generating process is inspired by a process used in a simulation in Farbmacher et al. (2020). In the simulation study, data sets with $n = 1000$ observations and $p_x = 20$ confounding variables x_i have been generated in $R = 500$ independent repetitions.

Chapter 3

Valid Simultaneous Inference in High-Dimensional Settings with the hdm Package for R

3.1 Introduction

One of the most significant developments to have occurred during the course of digitalization has been the increased availability of individual information. Data sets have become richer in that more explanatory variables have become available to predict outcomes of interest. If researchers want to assess the significance of the relationship between a large number of regressors and the dependent variable, it is essential to correct for testing multiple hypotheses at the same time. Otherwise, the probability of incorrectly rejecting a true null hypothesis is likely to exceed the specified significance level α. Whereas valid inference for one or a small number of regression coefficients in cases with many explanatory variables (i.e., high-dimensional settings) has been an active research area in the past decade, and many of the developed methods have been incorporated into applied academic research, most empirical studies do not account for the risk associated with multiple testing. Suppose, for instance, a researcher wants to estimate a large number of regression coefficients and to assess which of these coefficients are significantly different from zero at a significance level α. It is well known that in such a situation an approach that simply ignores the fact that many hypotheses are tested at the same time will generally lead to flawed conclusions due to a large number of mistakenly rejected hypotheses.

The statistical literature has proposed various approaches to mitigate the consequences of testing multiple hypotheses at the same time. These methods can be grouped into two approaches according to the underlying criterion. The first approach, initiated by the famous Bonferroni correction, seeks to control the probability of at least one false rejection, which is called the *family-wise error rate* (FWER). Since the definition of the FWER refers to the probability of making at least one type I error, the FWER-criterion is appealing from an intuitive point of view. However, FWER control is often criticized to be conservative and, instead, the *false discovery rate* (FDR) control is frequently used as a criterion leading to the second major class of multiple testing correction methods, e.g., Benjamini and Hochberg (1995). The FDR refers to the expected share of falsely rejected null hypotheses and, hence, results from FDR-procedures differ from classical tests results in terms of interpretation.

Various approaches aim to maintain control of the FWER while reducing conservativeness at the same time by incorporating a stepwise procedure, for instance, the stepdown method of Holm (1979). Moreover, taking the dependence structure of test statistics into consideration allows for a reduction in the conservativeness of FWER-procedures, as in the stepdown procedure of Romano and Wolf (2005a) and Romano and Wolf (2005b) that is based on resampling methods.

In the following, we review methods to perform valid simultaneous inference in a high-dimensional regression setting, i.e., if the number of covariates exceeds the number of observations, and give examples how the presented approaches can be applied with the statistical software R. The literature review is not intended to provide a complete summary of the literature on multiple testing adjustment. Rather we hope to address the major risk associated with multiple testing and to outline the reasoning of potential solutions to the applied researcher in high-dimensional settings. Moreover, we emphasize methodological problems arising for classical linear regression in high-dimensional settings and illustrate how these can be handled using regularization methods. We also provide a simulation study in different high-dimensional settings to compare the methods and give some guidance.

It is well-known that classical regression methods, such as ordinary least squares, break down in high-dimensional settings. Instead, regularization methods, for example the lasso, can be used for estimation. However, post-selection inference is non-trivial and requires modification of the estimators. We provide a short overview on two major approaches to perform simultaneous inference using regularization methods, i.e., the double selection approach, which has been developed in Belloni et al. (2014c), and the knockoff framework of Barber and Candès (2015) and compare their performance in a simulation study. Moreover, the paper illustrates how valid simultaneous inference based on the double selection approach can be performed in a real-data example using the package `hdm` (Chernozhukov et al., 2016a) for R (R Core Team, 2020). `hdm` provides powerful tests for a large number of hypotheses that can be combined with various methods to adjust for multiple testing as well as the functionality to construct valid simultaneous confidence intervals that is based on a multiplier bootstrap procedure.

The remainder of the paper is organized as follows. First, the general setting is introduced and an overview on valid post-selection inference in high dimensions is provided. Second, a short and selective review on traditional and recent methods to adjust for multiple testing is presented. Third, we compare the performance of the previously presented methods in a simulation study. Fourth, the use of the functionalities provided by the R package `hdm` are illustrated in a replicable real-data example on heterogeneity in the gender wage gap. A conclusion is provided in the last section.

3.2 Setting

We are interested in testing a set of K hypotheses $H_1, \dots, H_K$ in a high-dimensional regression model, i.e., a regression where the number of covariates p is large, potentially much larger than the number of observations n, i.e., we have $p \gg n$. The ultimate objective in this setting is to perform inference on a set of regression coefficients, i.e., a vector of so-called *target* coefficients θ_k with $k = 1, \dots, K$ and possibly with $K > n$. For example, such a setting has been considered in Belloni et al. (2014c) and Belloni et al. (2018).

$$y_i = \beta_0 + d_i'\theta + x_i'\beta + \epsilon_i, \qquad i = 1, \dots, n, \tag{3.1}$$

where β_0 is an intercept and β denote the regression coefficients of the control variables x_i. Moreover, it is assumed that $\mathbb{E}_n[\epsilon_i x_i] = 0$, where $\mathbb{E}_n$ denotes the empirical expectation, $\mathbb{E}_n(x) = \frac{1}{n}\sum_{i=1}^n x_i$. In this setting, K hypotheses are tested for the coefficients that correspond to the effect of the "target" variables d_i on the outcome y_i

$$H_{0,k} : \theta_k = 0, \qquad k = 1, \dots, K. \tag{3.2}$$

For instance, such a high-dimensional regression setting arises in causal program evaluation studies, where a large number of regressors is included to approximate a potentially complicated, nonlinear population regression function using transformations with dictionaries, for example splines or polynomials. Alternatively, an analysis of heterogeneous treatment effects across possibly many subgroups as in the example in Section 3.6.2 might require a large number of interactions of the regressors.

Suppose, there is a test procedure for each of the hypotheses leading to test statistics $t_1, \dots, t_K$ and unadjusted p-values $p_1, \dots, p_K$. In the context of multiple testing, it is often helpful to sort the p-values in an increasing order (in other words, the "most significant" test result as the first in the row) and the hypotheses likewise, i.e., $p_{(1)}, \dots, p_{(K)}$ and $H_{(1)}, \dots, H_{(K)}$ with $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(K)}$. Also the test statistics are ordered by the same logic $|t_{(1)}| \geq |t_{(2)}| \geq \dots \geq |t_{(K)}|$. A researcher decides whether to accept or to reject a null hypothesis if the corresponding p-value p_k is above or below a prespecified significance level α. Generally, the significance level corresponds to the probability of erroneously rejecting a true null hypothesis. However, if the conclusions are based on a comparison of unadjusted p-values and the significance level, the probability of incorrectly rejecting at least one of the hypotheses will generally exceed the claimed level α. Hence, adjustment for multiple testing becomes necessary to draw appropriate inferential conclusions.

3.3 Simultaneous Post-Selection Inference in High Dimensions

3.3.1 Simultaneous Inference based on Double Selection

In high-dimensional settings, traditional regression methods such as ordinary least squares break down and testing the K hypotheses will severely suffer from the shortcomings of the underlying estimation method. Penalization methods, for instance the lasso or other machine learning techniques, provide an opportunity to overcome the failure of traditional least squares estimation as they regularize the regression problem in Equation (5.3) by introducing a penalization term. In the example of lasso, the ordinary least squares minimization problem is extended by a penalization of the regression coefficients using the l_1-norm. The lasso estimator is the solution to the maximization problem

$$\left(\hat{\theta}', \hat{\beta}'\right)' = \arg\min_{\theta,\beta} \mathbb{E}_n \left[(y_i - \beta_0 - d_i'\theta - x_i'\beta)^2\right] + \frac{\lambda}{n} \left\|\hat{\psi}\left(\theta', \beta'\right)'\right\|_1, \tag{3.3}$$

with $\| \bullet \|_1$ being the l_1-norm, λ is a penalization parameter and $\hat{\psi}$ denotes a diagonal matrix of penalty loadings. More details on the choice of λ and $\hat{\psi}$ as implemented in the `hdm` package can be found in the package vignette available at CRAN (Chernozhukov et al., 2016a). As a consequence of the l_1-penalization, some of the coefficients are shrunk towards zero and some of them are set exactly equal to zero. In general, inference after such a selection step is only valid if model selection by lasso is perfect - in other words, the lasso does only set those coefficients to zero that truly have no effect on y_i (Leeb and Pötscher, 2008). However, perfect model selection and the underlying assumptions are often considered unrealistic in real-world applications leading to a breakdown of the *naive* inferential framework and, thus, flawed inferential conclusions. Stated more explicitly, the regularization introduced by lasso penalization leads to imperfect model selection with regard to so-called confounders. These variables are correlated with the target variable of interest, d_i, and the dependent variable, y_i. Consequently, imperfect model selection might cause an omitted variable bias that leads to a bias of the final estimator θ.

In contrast to the naive procedure, the so-called double selection approach of Belloni et al. (2014a) tolerates imperfect model selection such that asymptotically valid confidence intervals and test procedures can be based on the lasso. The double selection method is based on orthogonal moment equations: The double selection estimator is insensitive to the bias that arises due to moderate selection mistakes by the lasso. To achieve orthogonality, an auxiliary (lasso) regression step is introduced for each of the target coefficients. Estimation of the double selection estimator proceeds as follows

(1) For each of the target variables in $d_{j,i}$, $j = 1, \ldots, K$, a lasso regression is estimated to identify the most important predictors among the covariates x_i and the remaining target variables $d_{-j,i}$.

(2) A lasso regression of the outcome variable y_i on all explanatory variables, except for $d_{j,i}$, is estimated to identify predictors of y_i. This step is executed for each of the target variables $d_{j,i}$ with $j = 1, \ldots, K$.

(3) Each of the target coefficients, θ_j, is estimated from a linear regression of the outcome on all target variables as well as all covariates that have been selected in either one of the corresponding lasso regressions in step (1) or (2).

As a consequence of the double selection procedure, the risk of an omitted variable bias that might arise due to imperfect variable selection is reduced. It can be shown that the double selection estimator $\hat{\theta}_k^{DS}$ is asymptotically normally distributed under a set of regularity assumptions. Probably, the most important of these assumptions is (approximate) sparsity. This assumption states that only a subset of the regressors suffice to describe the relationship of the outcome variable and the explanatory variables, and that all other regressors have no or only a negligible effect on the outcome. In general, valid post-selection inference is compatible with other tools from the machine learning literature, for instance elastic nets or tree-based methods such as boosting or random forests, as long as these methods satisfy some regularity conditions (Belloni et al., 2014a; Belloni et al., 2014c) or if they are used in combination with sample splitting (Chernozhukov et al., 2018a).

As the double selection approach provides an asymptotically normally distributed test statistic and p-value for each of the tested hypotheses, $H_{0,k}$, $k = 1, \ldots, K$, it is possible to adjust for multiple testing using correction methods that operate on the test statistics or on p-values. For example, Chernozhukov et al. (2013a) and Belloni et al. (2014a) show that a multiplier bootstrap version of the Romano-Wolf method can be used to construct a joint significance test in a high-dimensional setting such that asymptotic control of the FWER is obtained. In the original work by Belloni et al. (2014a), it is shown that a valid $(1 - \alpha)$ confidence interval can be constructed by using the multiplier bootstrap as established in Chernozhukov et al. (2013a) and Chernozhukov et al. (2014).

3.3.2 Simultaneous Inference based on Knockoffs

A second approach to perform simultaneous inference, the knockoff framework, has been suggested by Barber and Candès (2015). The knockoff framework has been designed as a variable selection procedure that guarantees control of the FDR in linear models. The idea of this framework is to generate variables artificially, so-called "knockoff variables", that have the same correlation structure as the original covariates. By the definition of their construction, it is known that the artificial variables do not have explanatory power for the dependent variable. The knowledge that the knockoff variables might be selected as false positives allows the procedure to base model selection on the FDR criterion. When model selection is performed by the lasso or some alternative variable selection procedure, the original and the knockoff variables are considered as candidate variables. The idea is to exploit the known distinction between the constructed knockoffs and original regressors. Intuitively, if a regressor has some explanatory power for the outcome variable, the lasso will likely select the original variable instead of its copy. However, if an explanatory variable has only a small or no effect on the dependent variable, the selection procedure has some difficulty to distinguish the original variable from the corresponding knockoff.

For lasso estimation, the order of entry of variables for decreasing penalty parameter λ is considered. Intuitively, large values of λ lead to a very sparse variable selection. By gradually lowering the value of

λ, more and more variables enter the model. Relevant variables should enter clearly before their knockoff counterpart and the corresponding λ value at entry should be large. For noise variables, i.e., variables with no effect on the outcome variable, the order might be reversed. Based on this observation, Barber and Candès (2015) derive test-statistics to control the false discovery rates at a certain level.

Framed in the model presented in Equation (5.3), we define the lasso estimators $\left(\hat{\theta}', \hat{\beta}'\right)'$ for a given penalty level λ as[1]

$$\left(\hat{\theta}', \hat{\beta}'\right)'(\lambda) = \arg\min_{(\theta,\beta)} \left\{ \frac{1}{2}\mathbb{E}_n\left[\left(y_i - \beta_0 - d_i'\theta - x_i'\beta\right)^2 \right] + \lambda \left\| (\theta', \beta')' \right\|_1 \right\}.$$

A variable enters the model for the first time at a value of λ given by

$$Z_j = \sup \lambda : \left(\hat{\theta}', \hat{\beta}'\right)'_j(\lambda) \neq 0.$$

Basically, for those variables that have strong explanatory power for the outcome, the value of Z_j is expected to be large, whereas it is expected to be small for the noise variables. If both the original variables and their knockoffs are considered as candidates in the lasso selection procedure, the statistics of the original variables Z_j and those of the corresponding knockoffs, $\tilde{Z}_j$, can be compared. Intuitively, one would expect that the difference between Z_j and $\tilde{Z}_j$ is largest for powerful predictors of the dependent variable, whereas a small difference is expected for the noise variables. Accordingly, the statistic W_j is based on the first entrance into the model

$$W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & , Z_j > \tilde{Z}_j \\ -1 & , Z_j < \tilde{Z}_j \end{cases}.$$

Given a target FDR of q, a data-dependent threshold T is defined

$$T = \min\left\{ t \in \mathcal{W} : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\},$$

with $\mathcal{W} = \{|W_j| : j = 1, \ldots, p\} \setminus \{0\}$. Selecting a model $\hat{S} = \{j : W_j \geq T\}$, Barber and Candès (2015) show that for any q, the knockoff method fulfills

$$\mathbb{E}\left[\frac{\#\{j : (\theta', \beta')'_j = 0 \text{ and } j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} q^{-1}} \right] \leq q.$$

For further results and a more in-depth discussion we refer to the original paper by Barber and Candès (2015).

Originally, knockoffs were designed as a model selection procedure that guarantees control of the FDR in linear models under homoskedasticity, fixed design X (hence, also denoted as "fixed knockoffs") and low-dimensional settings, i.e., $n \geq p$ (Barber and Candès, 2015). The approach has been extended to high-dimensional settings in Barber and Candès (2019) by introducing an initial screening procedure that imposes a dimension reduction on the model. Candès et al. (2018) introduce Model-X knockoffs, probabilistically constructed variables, allowing the dependent variable to be drawn from an arbitrary distribution. Moreover, the number of variables is possibly unrestricted and the considered model may be nonlinear. An extension of knockoffs to control k-FWER has been developed in Janson and Su (2016).

---

[1] Note that from a methodological point of view the double selection framework distinguishes target variables of interest, here d_i, from covariates, x_i. Inference is performed only for the target variable d_i. For example, in a causal model the confounders x_i must be included to achieve unconfoundedness or exogeneity conditional on x_i. This distinction is not made in the knockoff framework, and, thus, we reformulate the knockoff procedure according to Model (5.3) where all covariates, i.e., d_i and x_i, are subject to variable selection.

3.3.3 Summary: Simultaneous Post-Selection Inference

A methodological difference of the knockoff approach and double selection is that the knockoff framework incorporates the FDR-criterion as a guidance during the variable selection procedure and provides a set of selected variables that guarantees control of the FDR. In contrast, double selection first modifies the selection procedure by introducing a second lasso selection step, to obtain validity of the test statistics that are obtained for the target coefficients θ. Hence, it is possible to correct for multiple testing in an additional step after estimation has been performed by employing a variety of methods that operate either directly on the test statistics or the p-values.

3.4 Methods for Simultaneously Testing Multiple Hypotheses

In this section, we review classical and recently developed methods to adjust for simultaneously testing multiple hypotheses. The considered methods operate on p-values or test statistics and presume the exact or asymptotic validity of the corresponding inferential procedure. Hence, in a high-dimensional setting, we consider cases where double selection has been performed in a first step to obtain valid coefficient estimates, test statistics and p-values as summarized in Section 3.3.1. The following section is organized as follows: First, correction methods that control the FWER and FDR are presented in Sections 3.4.1 and 3.4.2, respectively. Section 3.4.3 presents a global test available for lasso regressions in high dimensions that is comparable to a F-test in a classical linear regression model.

3.4.1 Multiple Hypotheses Testing with Control of the Familywise Error Rate

The FWER is defined as the probability of falsely rejecting at least one hypothesis. The goal is to control the FWER and to secure that it does not exceed a prespecified level α. We assume that for the individual tests the significance level is set uniformly to α.

3.4.1.1 Bonferroni Correction

According to the Bonferroni correction the cutoff of the p-values is set to $\alpha^* = \alpha/K$ and all hypotheses with p-values below the adjusted level α^* are rejected. Boole's inequality then gives directly that the FWER is smaller or equal to α. Instead of adjusting the level of α to α^*, it is possible to adjust the p-values so that we reject a hypothesis H_k if $p_k^* = \min\{1, K \cdot p_k\} < \alpha$. A drawback of the procedure is that it is quite conservative, meaning that in many applications, in particular in high-dimensional settings when many hypotheses are tested simultaneously, often no or very few hypotheses are rejected, increasing the risk of accepting false null hypotheses (i.e., of a type II error).

3.4.1.2 Bonferroni-Holm Correction

We again assume that the p-values are ordered (from lowest to highest) $p_{(1)} \leq \ldots \leq p_{(K)}$ with corresponding hypotheses $H_{(1)}, \ldots, H_{(K)}$. Stepdown methods proceed in several rounds. In each round a decision is taken on the set of hypotheses being rejected. The algorithm continues until no further hypotheses are rejected. To illustrate the stepwise proceeding of the Bonferroni-Holm method, suppose that the Bonferroni correction as described in the previous section has been performed and it was possible to reject the first null hypotheses $H_{(1)}$. Then the Bonferroni correction could be applied a second time with respect to all hypotheses except for $H_{(1)}$, i.e., the significance level to test $H_{(2)}$ is adjusted to $\alpha^*_{(2)} = \frac{\alpha}{K-1}$. The Bonferroni-Holm correction proceeds in this sequential manner until no hypotheses can be rejected anymore. Formally, let k be the smallest index such that the corresponding p-value exceeds the adjusted cutoff α^*.

$$k = \min_j \{p_{(j)} > \underbrace{\frac{\alpha}{K-j+1}}_{\alpha^*}\},$$

The hypotheses $H_{(1)}, \ldots, H_{(k-1)}$ are then rejected and we accept $H_{(k)}, \ldots, H_{(K)}$. The Bonferroni-Holm procedure is considered as a general improvement over the Bonferroni correction that maintains control of the FWER and reduces the risk of a type II error at the same time. The adjusted p-value according to the Bonferroni-Holm correction are computed as $p^*_{(j)} = \max_{l \leq j} \min\{(K-j+1)p_{(j)}, 1\}$ with $l = 1, \ldots, j$.

3.4.1.3 Joint Confidence Region based on the Multiplier Bootstrap

Belloni et al. (2014a) derive valid $(1-\alpha)$ confidence regions for the vector of target coefficients, θ, in the high-dimensional regression setting in Equation (5.3) estimated with lasso. The confidence regions which are constructed with the multiplier bootstrap can be used equivalently to a joint significance test of the K hypotheses. Accordingly, the null hypotheses $H_{0,k} : \theta_k = 0$, $k = 1, \ldots, K$, would be rejected at the level α if the simultaneous $(1-\alpha)$ confidence region does not cover zero in dimension k.

The multiplier bootstrap procedure is based on random pertubations of the orthogonal score function. As mentioned above, the double selection estimator can be considered as the solution to the empirical analog of an orthogonal score function. The multiplier bootstrap procedure estimates bootstrapped coefficients $\hat{\theta}^{*,b}_j$ and test-statistics $t^{*,b}_j$, $b = 1, \ldots, B$, in B repetitions and for each of the coefficients of interest. The bootstrapped quantities are obtained from solving a pertubated version of the orthogonal score function, i.e., the score function being multiplied with independent random variables, for example independent draws from a standard normal distribution (Chernozhukov et al., 2013a). Relying on random pertubations, it is possible to avoid resampling and re-estimation of the double selection estimator, which might be computationally costly in high-dimensional settings.

3.4.1.4 Romano-Wolf Stepdown Procedure

The stepdown method of Romano and Wolf (2005a) and Romano and Wolf (2005b) is based on resampling methods. Thus, it is able to account for the dependence structure underlying the test statistics and to give less conservative results as compared to methods such as the Bonferroni and Holm correction. The idea of the Romano-Wolf procedure is to construct rectangular simultaneous confidence intervals in subsequent steps whereas in each step, the coverage probability is kept above a level of $(1-\alpha)$. If in step j, the confidence set does not contain zero in dimension k, the corresponding H_k is rejected. In step $j+1$, the algorithm proceeds analogously by constructing a rectangular joint confidence region for those coefficients for which the null hypotheses has not been rejected in step j or before. The algorithm stops if no hypothesis is rejected anymore. To take the dependence structure of the test hypotheses into account, the classical Romano-Wolf stepdown procedure uses resampling to compute the constant $c_{(1-\alpha)}$ that is needed to construct a rectangular confidence interval. This constant is estimated by the $(1-\alpha)$ quantile of the maxima of the bootstrapped test statistics in each step to guarantee the coverage probability of $(1-\alpha)$. The computational burden of the Romano-Wolf stepdown procedure can be reduced by using the multiplier bootstrap because it is only based on random permutations of the score function and does not require re-estimation of the double selection estimator based on bootstrap samples. We present a recent version of the Romano-Wolf method from Chernozhukov et al. (2013a) and Belloni et al. (2014a) who prove the validity of the procedure in combination with the multiplier bootstrap.

Algorithm 1: Romano-Wolf stepdown correction of p-values

1) Sort the test statistics in a decreasing order (in terms of their absolute values):

$$|t_{(1)}| \geq |t_{(2)}| \geq \ldots \geq |t_{(K)}|.$$

2) Draw B multiplier bootstrap versions for each of the test statistics $t^{*,b}_{(k)}$, $b = 1, \ldots, B$, and $k = 1, \ldots, K$,

3) For each b and k determine the maximum of the bootstrapped test statistics $m(t^{*,b}_{(k)}) = \max\left\{|t^{*,b}_{(k)}|, |t^{*,b}_{(k+1)}|, \ldots, |t^{*,b}_{(K)}|\right\}$.

4) Compute initial p-values, for $k = 1, \ldots, K$

$$p^{init}_{(k)} := \frac{\sum_{b=1}^{B} \mathbb{I}\{m(t^{*,b}_{(k)}) \geq |t_{(k)}|\}}{B},$$

with $\mathbb{I}\{\cdot\}$ being an indicator that is equal to 1, if the statement in curly brackets $\{\cdot\}$ is true.

5) Compute adjusted p-values by ensuring monotonicity

   a) if $k = 1$

   $$p^{*}_{(1)} := p^{init}_{(1)}.$$

   b) if $k = 2, \ldots, K$

   $$p^{*}_{(k)} := \max\{p^{init}_{(k)}, p^{*}_{(k-1)}\}.$$

The p-value adjustment algorithm parallels that in Romano and Wolf (2016) with the only difference that the bootstrap test statistics are computed efficiently with the multiplier bootstrap procedure instead of the classical bootstrap and that the test statistics are based on post-selection inference with the lasso. In Romano and Wolf (2005a) and Romano and Wolf (2016), a high number of bootstrap repetitions $B \geq 1000$ is recommended.

If the data stem from a randomized experiment, the method introduced in List et al. (2019) can be used. It is a variant of the Romano-Wolf procedure under unconfoundedness - an assumption that can be justified in an experimental setting if a treatment is assigned randomly conditional on observational characteristics. Moreover, it allows researchers to compare the effect of different treatments and several outcome variables simultaneously.

3.4.2 Multiple Hypotheses Testing with Control of the False Discovery Rate

The FWER is frequently considered a strict criterion, which often leads to very conservative conclusions. This means that in settings when thousands or hundred thousands of hypotheses are tested simultaneously, the FWER does often not detect useful signals. Hence, in large-scale settings frequently a less strict criterion, the so-called false discovery rate (FDR) is employed. The false discovery proportion (FDP) is defined as the ratio of the number of hypotheses, which are wrongly classified as significant (false positives) and the total number of positives. If the latter is zero, it is defined as zero. The FDR is defined as the expected value of the FDP : $FDR = \mathbb{E}(FDP)$. The FDR concept reflects the trade-off between false discoveries and true discoveries.

3.4.2.1 Benjamini-Hochberg Procedure

To control the FDR, the Benjamini-Hochberg (BH) procedure ranks the hypotheses according to the corresponding p-values and then chooses a cutoff along the ranking to control the FDR at a prespecified level of $\gamma \in (0,1)$. The BH procedure first uses a stepup comparison to find a cutoff p-value:

$$k = \max_j \{p_{(j)} \leq j \frac{\gamma}{K}\},$$

and then rejects all hypotheses $H_{(j)}, j = 1, \ldots, k$. In most applications, $\gamma = 0.1$ is chosen.

3.4.3 A Global Test for Joint Significance with Lasso Regression

A basic question frequently arising in empirical work is whether the lasso regression has explanatory power, comparable to a F-test for the classical linear regression model. The construction of a joint significance test follows Chernozhukov et al. (2013a, Appendix M) and has been presented earlier in Chernozhukov et al. (2016b). Based on the model $y_i = \beta_0 + d_i'\theta + x_i'\beta + \epsilon_i$ with intercept β_0, the null hypothesis of joint statistical non-significance is $H_0 : (\theta', \beta')' = \mathbf{0}$. The null hypothesis implies that

$$\mathbb{E}\left[(y_i - \beta_0)x_i\right] = 0,$$

and the restriction can be tested using the sup-score statistic:

$$S = \|\sqrt{n}\mathbb{E}_n\left[(y_i - \hat{\beta}_0)x_i\right]\|_\infty,$$

where $\hat{\beta}_0 = \mathbb{E}_n[y_i]$. The critical value for this statistic can be approximated by the multiplier bootstrap procedure, which simulates the statistic:

$$S^* = \|\sqrt{n}\mathbb{E}_n\left[(y_i - \hat{\beta}_0)x_i g_i\right]\|_\infty,$$

where g_i's are i.i.d. $N(0,1)$, conditional on the data. The $(1-\alpha)$ quantile of S^* serves as the critical value, $c_{(1-\alpha)}$. We reject the null if $S > c_{(1-\alpha)}$ in favor of statistical significance, and we keep the null of non-significance otherwise.

3.5 Simulation Study

The simulation study provides a finite-sample comparison of different multiple testing corrections in a high-dimensional setting, i.e., the Bonferroni method, the Bonferroni-Holm procedure, Benjamini-Hochberg adjustment and the Romano-Wolf stepdown method, as well as three different knockoff variants. In addition, the study illustrates the failure of an approach that ignores the problem of simultaneous hypotheses testing, i.e., without any correction of the significance level or p-values.

3.5.1 Simulation Setting

We consider a regression of a continuous outcome variable y_i on a set of regressors, d_i, in settings with $K \in \{60, 180, 200, 400\}$. To maintain comparability of the double selection and knockoff framework, we

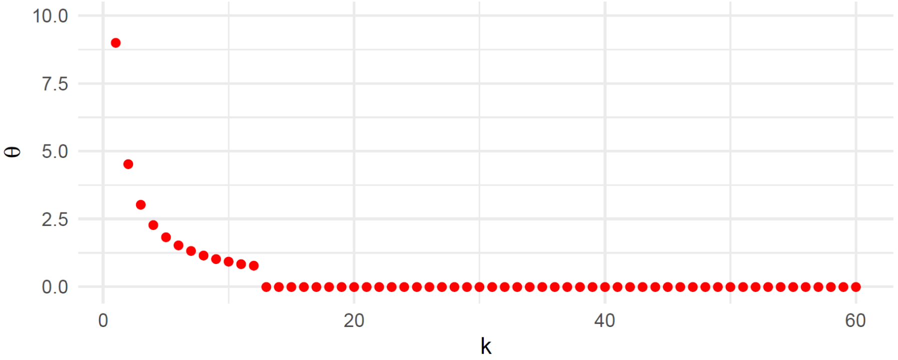


Figure 3.1: **Regression coefficients, simulation setting with $K = 60$.**

test the coefficients of all regressors, i.e., we specify $p = K$ throughout the simulation study.

$$y_i = \beta_0 + d_i'\theta + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3.4}$$

with a homoskedastic and normally distributed error $\varepsilon_i \sim N(0, \sigma^2)$ with variance $\sigma^2 = 3$. In our setting, the realizations of d_i are generated by a joint normal distribution $d_i \sim N(\mu, \Sigma)$ with $\mu = \mathbf{0}$ and $\Sigma_{j,k} = \rho^{|j-k|}$ with $\rho = 0.8$. We consider the case of an i.i.d. sample with $n = 200$ and $n = 500$ observations. The setting is sparse in that only $s = 12$ regressors are truly non-zero: The first $s = 12$ coefficients θ are generated according to the sparse model with decay

$$\theta_j = \min\left\{\frac{\theta^{\max}}{j^a}, \theta^{\min}\right\},$$

for $j = 1, \ldots, s$ with $\theta^{\max} = 9$, $\theta^{\min} = 0.75$, and $a = 0.99$. All other coefficients have values exactly equal to 0. Figure 3.1 presents the regression coefficients in the simulation study.

In the case of double selection, the regression in Equation (3.4) is estimated with post-lasso.[2] The K hypotheses are tested simultaneously

$$H_{0,k} : \theta_k = 0, \qquad k = 1, \ldots, K.$$

We implement three different specifications of the knockoff framework, i.e., (i) fixed-model knockoffs ("Fix-KO") (ii) second-order Gaussian model-X knockoffs ("Model-X"), (iii) Gaussian model-X knockoffs ("Model-X (Or.)"). In (ii) we specify that the joint distribution of the covariates is Gaussian with mean and covariance matrix being unknown, whereas in (iii) the oracle quantities μ and Σ are used to construct Gaussian knockoffs. It is worth to note that the assumptions required for fixed knockoffs are not satisfied in the high-dimensional simulation settings. Consequently, we expect an inferior performance or a breakdown of the method when $K > n$.

3.5.2 Results

The simulation results are summarized in Table 3.1. The reported results refer to averages from $R = 1000$ repetitions in terms of correct and incorrect rejections of null hypotheses at a specified level of $\alpha = 0.1$ for the FWER and $\gamma = 0.1$ for the FDR as well as the empirical FWER and FDR.

The results show that multiple testing adjustment is of great importance: Inferential statements without any multiple testing adjustment are likely invalid if the number of tested hypotheses is relatively large.

---

[2] More details on implementation of the simulation study are provided in the Appendix and the supplemental material available at `https://www.bwl.uni-hamburg.de/en/statistik/forschung/software-und-daten.html`.

The results in the column "None" correspond to such an approach. If each of the hypotheses is tested at a significance level of $\alpha = 0.1$ and no adjustment of the p-values is performed, the probability of at least one incorrect rejection is close or exactly equal to 1 in all settings considered.

Whereas a correction according to the Benjamini-Hochberg procedure (column "BH") already reduces the number of incorrect rejections: More than one hypothesis is rejected incorrectly in all settings considered, on average. At the same time, however, the empirical FDR is always close to the specified level of $\gamma = 0.1$ for the Benjamini-Hochberg adjustment whereas the performance benefits from larger sample size. Methods with asymptotic control of the FWER are much less likely to erroneously reject true null hypotheses. In most settings, the empirical FWER approaches the nominal level with the only exception being the high-dimensional setting with twice as many parameters as observations ($n = 200$, $p = 400$). The results improve in settings with larger n. Of these methods, the Bonferroni correction is the most conservative. Slight improvements in terms of power are achieved by the Holm procedure. The Romano-Wolf stepdown correction benefits from taking the dependence structure of the covariates into account and, thus, allows to reject slightly more correct hypotheses than the Holm method. However, this increase in power arises in parallel to an increased number of incorrect rejections.

In general, the FWER methods are more conservative than the FDR controlling approaches, in particular if the ratio n/p is small. However, in the settings with sample size $n = 500$ the FWER methods are able to reject almost all false null hypotheses while still controlling the number of incorrect rejections at a considerably lower level.

An interesting comparison is that of the double selection method with Benjamini-Hochberg correction with Model-X knockoffs as both approaches guarantee control of the FDR. In the simulation study, we made the observation that the performance of the knockoff algorithm depends on the method of knockoff generation. This instability of knockoffs in settings with a small number of non-zero coefficients has already been documented in a recent study by Gimenez and Zou (2019). Hence, we display the results for knockoffs in two ways. The results displayed in parentheses in Table 3.1 refer to all 1000 repetitions. The remaining results refer to only those simulation repetitions with a non-zero model selection, i.e., we excluded repetitions where no variables have been selected (corresponding to a threshold $T = \infty$). Table 3.3 in the Appendix presents the frequency of cases where the knockoff algorithm selected an empty set of variables. The degree of instability depends on the choice of the knockoff construction method and the relation of n and p. For example, the fixed knockoffs perform relatively well in the settings with $n = 500$ and $p = 60$. However, in the setting with $n = 200$ and $p = 180$ the procedure fails to deliver a reasonable model selection in almost every repetition.

Contrarily to the poor performance of the fixed knockoffs, Model-X knockoffs exhibit excellent performance and a high degree of stability in the high-dimensional setting with $n = 200$ and $p = 400$. However, the performance deteriorates when $n < p$. According to the documentation of the R package `knockoff` (Patterson and Sesia, 2018), regularization is performed during estimation of the covariance matrix in high-dimensional settings, which might explain the increased stability of Model-X in the setting with $n = 200$ and $p = 400$ (Candès et al., 2018).

n	K	None	BH	Bonf.	Holm	RW	Fix-KO	Model-X	Model-X (Or.)
Correct Rejections									
200	60	11.73	11.25	10.02	10.11	10.19	11.26	11.43	11.82
							(7.38)	*(6.71)*	*(5.74)*
200	180	11.76	10.76	9.51	9.53	9.62	8.00	10.88	11.80
							(0.05)	*(0.09)*	*(5.76)*
200	400	11.73	10.33	9.03	9.04	9.14	.	11.92	11.80
							(0.00)	*(11.92)*	*(5.77)*
500	60	12.00	11.98	11.87	11.88	11.90	11.81	11.96	11.99
							(10.23)	*(9.75)*	*(6.03)*
500	200	11.99	11.95	11.75	11.75	11.77	11.24	11.66	11.99
							(7.97)	*(6.81)*	*(6.04)*
Incorrect Rejections									
200	60	5.13	1.29	0.13	0.16	0.18	0.86	0.70	0.87
							(0.56)	*(0.41)*	*(0.42)*
200	180	18.17	1.69	0.19	0.20	0.21	2.83	0.50	0.94
							(0.02)	*(0.00)*	*(0.46)*
200	400	41.73	1.88	0.22	0.22	0.26	.	1.06	1.04
							(0.00)	*(1.06)*	*(0.51)*
500	60	4.85	1.27	0.10	0.13	0.14	1.11	0.91	0.93
							(0.96)	*(0.74)*	*(0.47)*
500	200	19.12	1.49	0.12	0.12	0.14	1.06	0.73	0.88
							(0.75)	*(0.42)*	*(0.44)*
Family-Wise Error Rate									
200	60	1.00	0.67	0.12	0.14	0.15	0.43	0.39	0.46
							(0.28)	*(0.23)*	*(0.22)*
200	180	1.00	0.74	0.16	0.17	0.18	1.00	0.25	0.45
							(0.01)	*(0.00)*	*(0.22)*
200	400	1.00	0.78	0.20	0.20	0.23	.	0.53	0.50
							(0.00)	*(0.53)*	*(0.24)*
500	60	0.98	0.65	0.09	0.11	0.12	0.51	0.47	0.50
							(0.44)	*(0.38)*	*(0.25)*
500	200	1.00	0.73	0.10	0.11	0.12	0.45	0.41	0.49
							(0.32)	*(0.24)*	*(0.25)*
False Discovery Rate									
200	60	0.29	0.09	0.01	0.01	0.02	0.06	0.05	0.06
							(0.04)	*(0.03)*	*(0.03)*
200	180	0.60	0.12	0.02	0.02	0.02	0.26	0.04	0.06
							(0.00)	*(0.00)*	*(0.03)*
200	400	0.78	0.14	0.02	0.02	0.03	.	0.07	0.07
							(0.00)	*(0.07)*	*(0.03)*
500	60	0.27	0.09	0.01	0.01	0.01	0.07	0.06	0.06
							(0.06)	*(0.05)*	*(0.03)*
500	200	0.61	0.10	0.01	0.01	0.01	0.07	0.05	0.06
							(0.05)	*(0.03)*	*(0.03)*

Table 3.1: **Simulation results, $R = 1000$ repetitions.**

A dot indicates that the procedure has resulted in zero-selection in all repetitions. Results in parentheses refer to all repetitions of the knockoff procedures, i.e., including those with zero variable selection.

Using the oracle quantities μ and Σ makes the performance of the knockoff procedure less dependent on the ratio of n and p. However, the procedure results in zero-selections in approximately every second repetition.

In terms of FWER control, the knockoff procedures never exceed the specified level of $\gamma = 0.1$.[3] Once we select only those repetitions with a positive number of selected variables, the knockoffs are able to exhibit excellent performance in most settings. Accordingly, the number of correct rejections exceeds considerably those of the Benjamini-Hochberg procedure in many cases. At the same time, the number of incorrect rejections is lower than for the FWER-controlling procedure based on double selection.

The results suggest that, once an appropriate choice of the knockoff procedure is made, the method can be a powerful tool for simultaneous inference in high-dimensional settings. Accordingly, the choice which version of knockoffs are used in an empirical application is likely to depend on the setting at hand. The performance of the double selection approach of Belloni et al. (2014a) is comparably more stable and allows to employ different criteria and methods for multiple testing adjustments.

3.6 Simultaneous Inference in a Real-Data Example with R

3.6.1 Implementation in the R Package `hdm`

Estimation of the high-dimensional regression model in Equation (5.3) and simultaneous inference on the target coefficients based on the double selection approach is implemented in the R package `hdm` available at CRAN (Chernozhukov et al., 2016a). `hdm` provides an implementation of the double selection approach of Belloni et al. (2014a) using the lasso as the variable selection device. The function `rlassoEffects()` does valid inference on a specified set of target parameters and returns an object of S3 class `rlassoEffects`. Correction for testing multiple hypotheses simultaneously is then performed on this output object as described in the following. More details on the `hdm` package and introductory examples are provided in the `hdm` vignette available at CRAN. The package `hdm` offers three ways to perform valid simultaneous inference in high-dimensional settings:

1. **Overall significance test**
   `hdm` provides a global significance test that is comparable to a F-test known from classical ordinary least squares regression. Based on Chernozhukov et al. (2013a, Appendix M) and Chernozhukov et al. (2016b), the null hypothesis that no covariate has explanatory power for the outcome y_i is tested, i.e.,

   $$H_0 : (\theta', \beta')' = \mathbf{0}.$$

   The test is performed automatically if `summary()` is called for an object of the S3 class `rlasso`. This object corresponds to the output of the function `rlasso()` which implements the lasso estimator using a theory-based rule for determining the penalization parameter.

2. **Joint confidence interval**
   Based on an object of the S3 class `rlassoEffects`, a valid joint confidence interval with coverage probability $(1 - \alpha)$ can be constructed for the specified target coefficients using the command `confint()` with the option `joint = TRUE`.

3. **Multiple testing adjustment of p-values**
   Starting with Version `0.3.0`, the `hdm` package offers the S3 method `p_adjust()` for objects inheriting

---

[3] The only case where the FDR is above γ is for the fixed knockoffs with $n = 200$ and $p = 180$. However, these results have to be interpreted with caution because of very many repetitions with a zero-selection.

from classes `rlassoEffects` and `lm`. By default, `p_adjust()` implements the Romano-Wolf step-down procedure using the computationally efficient multiplier bootstrap procedure (option `method = 'RW'`). Hence, `hdm` offers an implementation of the p-value adjustment that corresponds to a joint test in the sense of Romano-Wolf for post-selection inference that is based on double selection with the lasso as well as for ordinary least squares regression. Moreover, the `p_adjust()` call offers classical adjustment methods in a user-friendly way, i.e., the function can be executed directly on the output object returned from an estimation with `rlassoEffects()` or `lm()`. The hosted correction methods are the methods provided in the `p.adjust()` command of the basic `stats` package, i.e., Bonferroni, Bonferroni-Holm, and Benjamini-Hochberg among others. If an object of class `lm` is used, the user can provide an index of the coefficients to be tested simultaneously. By default, all coefficients are tested.

The `hdm` package can be installed from CRAN by the following command

```r
# To install the hdm from CRAN call
install.packages("hdm")
```

Once the package has been installed, it can be loaded via

```r
# Load the hdm package
library(hdm)
```

3.6.2 A Real-Data Example for Simultaneous Inference in High Dimensions - The Gender Wage Gap Analysis

The following section demonstrates the methods for valid simultaneous inference implemented in the package `hdm` and provides a comparison of the classical correction methods in a replicable real-data example. A simplistic although frequently encountered approach to assess wage gap heterogeneity is to compare the relative wage gap across female and male employees in subgroups defined in terms of a particular variable, for example by industry. It is obvious that this approach neglects the role of other variables relevant for the wage income, for example educational background, experience etc. As an exemplary illustration, the gender gap in average (mean) earnings in 12 industrial categories is presented in Figure 3.2, suggesting that the wage gap differs greatly across these subgroups. The category “Wholesale” is set as the baseline class as indicated by the gray box in Figure 3.1. If one would simply compare the wage gaps across categories in an approach such as this, one would conclude that there are several categories with higher and lower gender gaps, for example, in industry “Transportation”, “Agriculture” with the largest gap observed in “Finance, Insurance and Real Estate”.

In contrast to this simplistic approach, an extended wage equation including interaction terms of the gender indicator with observable characteristics is able to take the role of other labor market characteristics into account. Thus, this approach makes it possible to give insights on the potential drivers of the gender wage gap. The coefficients on the gender interactions can then be interpreted as changes of the wage gap as compared to the baseline category. As the regression approach leads to a large number of coefficients that are tested simultaneously, an appropriate multiple testing adjustment is required. The presented example is an illustration of the more extensive analysis of a heterogeneous gender wage gap in Bach et al. (2018a).

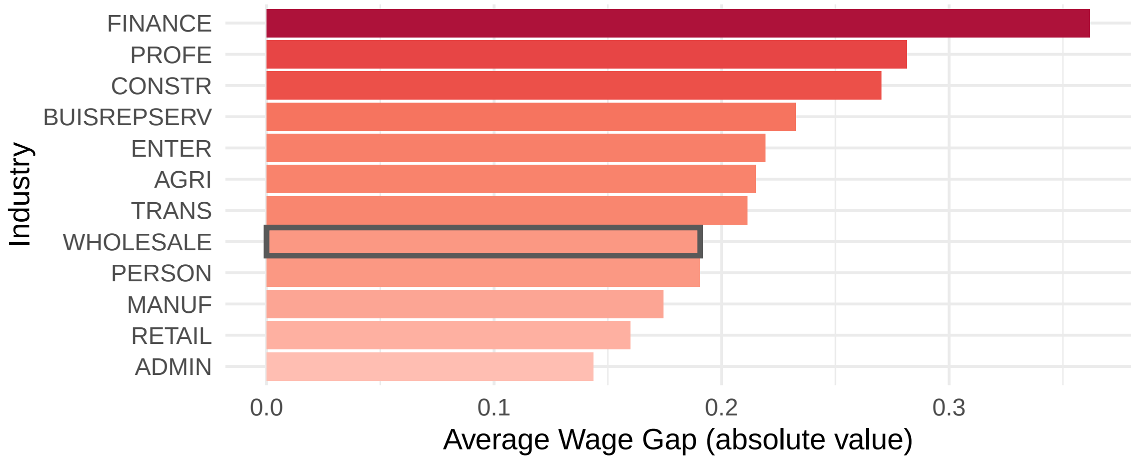


Figure 3.2: **Average wage gap in industries, ACS 2016.**

3.6.2.1 Data Preparation

The exemplary data is a subsample of the 2016 American Community Survey and can be replicated with the documentation “Appendix: Replicable Data Example” that is available online.[4]

The data provide information on civilian full-time working (35+ hours a week, 50+ weeks a year) White, non-Hispanic employees aged older than 25 and younger than 40 with earnings exceeding a the federal minimum level of earnings ($12,687.5 of yearly wage income). After preprocessing, the data set can be loaded.

```r
# load the ACS data (after preprocessing)
load("ACS2016_gender.rda")
```

3.6.2.2 Valid Simultaneous Inference on a Heterogeneous Gender Wage Gap

We would like to provide a valid answer to the question whether the gender wage gap differs according to the observable characteristics of female employees and whether this variation is significant. To do so, it is necessary to account for regressors that affect women’s job market environment. In the example, variables on marriage, the presence of own children, geographic variation, job characteristics (industry, occupation, hours worked), human capital variables (years of education, experience (squared)), and college major are considered. A wage regression is set up that includes all two-way interactions of the female dummy with the additional characteristics in addition to the baseline regressors, x_i.

$$\ln w_i = \beta_0 + \sum_{k=1}^{K} \theta_k \left(\text{female}_i \times x_{k,i}\right) + x_i'\beta + \varepsilon_i, \qquad i = 1, \dots, n, \tag{3.5}$$

The analysis begins with the construction of model matrix that implements the regression relationship of interest.

---

[4] `https://www.bwl.uni-hamburg.de/en/statistik/forschung/software-und-daten.html`

```
# Weekly log wages as outcome variable
y = data$lwwage

# Model Matrix containing 2-way interaction of female
# with relevant regressors + covariates (not interacted)
X = model.matrix( ~ 1 + fem + fem:(ind + occ + hw + deg + yos + exp + exp2 +
                  married + chld19 + region + msa ) +
                  married + chld19 + region + msa + ind + occ + hw + deg +
                  yos + exp + exp2,
                data = data)

# Exclude the constant variables
X = X[, which(apply(X, 2, var)!=0)]
dim(X)
```

```
## [1] 70473   123
```

```
# Replace column names for female indicator with shorter names
colnames(X) = gsub("femTRUE", "fem", colnames(X))
```

Accordingly, the regression model considered has $p = 123$ regressors in total and is estimated on the basis of $n = 70,473$ observations. The wage Equation (3.5)is estimated with the lasso with the theory-based choice of the penalty term as implemented in the function `rlasso`. To answer the question whether the included regressors have any explanatory power for the outcome variable, the global test of overall significance is run by calling `summary()` on the output object of the `rlasso()` function.

```
# run rlasso
lasso1 = rlasso(X,y)

# run global test
summary(lasso1, all = FALSE)
# Complete output provided in the Appendix
```

```
##
## Call:
## rlasso.default(x = X, y = y)
##
## Post-Lasso Estimation:  TRUE
##
## Total number of variables: 123
## Number of selected variables: 58
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.509477 -0.274888 -0.007866  0.255786  2.667454
##
##                                    Estimate
## (Intercept)                           4.696
## fem                                   0.008
## married                               0.116
## chld19                                0.088
## regionMiddle Atlantic Division        0.075
## regionEast North Central Div.        -0.037
##
## [...]
##
## fem:married                          -0.048
## fem:chld19                           -0.041
## fem:regionMountain Division          -0.007
##
## Residual standard error: 0.4633
## Multiple R-squared:  0.3906
## Adjusted R-squared:  0.3901
## Joint significance test:
##  the sup score statistic for joint significance test is 279.6 with a p-value of     0
```

The hypothesis that all coefficients in the model are zero can be rejected at all common significance levels. The main objective of the analysis is to estimate the effects associated with the gender interactions and to assess whether these effects are jointly significantly different from zero. The so-called "target" variables, in total 62 regressors, are specified in the `index` option of the `rlassoEffects()` function. Hence, it is necessary to indicate the columns of the model matrix that correspond to interactions with the female indicator.

```r
# Construct index for gender variable and interactions  (target parameters)
index.female = grep("fem", colnames(X))
K = length(index.female)

# Perform inference on target coefficients
# estimation might take some time (10 minutes)
effects = rlassoEffects(x = X, y = y, index = index.female, method = "double selection")

# here only present first rows of output; full output provided in the Appendix
summary(effects)
```

```
##                Estimate. Std. Error t value Pr(>|t|)
## fem              -0.0460     0.0660 -0.6976   0.4854
## fem:indAGRI      -0.1124     0.0542 -2.0752   0.0380
## fem:indCONSTR    -0.0525     0.0361 -1.4520   0.1465
## fem:indMANUF     -0.0110     0.0265 -0.4152   0.6780
## fem:indTRANS     -0.0418     0.0294 -1.4230   0.1547
## fem:indRETAIL     0.0284     0.0274  1.0353   0.3005
## fem:indFINANCE   -0.1349     0.0264 -5.1126   0.0000
```

The output presented in the Appendix shows the 123 estimated coefficients together with t-statistics and p-values that are not yet corrected for multiple testing. The next step is to adjust the p-values for multiple testing. Starting with Version `0.3.0`, the `hdm` offers the S3 method `p_adjust()` for objects inheriting from classes `rlassoEffects` and `lm`. It hosts the correction methods from the function `p.adjust()` of the `stats` package, for example the Bonferroni, Bonferroni-Holm, Benjamini-Hochberg as well as no correction at all. First, the naive approach without any correction is presented. Table 3.2 shows the number of rejections at significance levels $\alpha \in \{0.01, 0.05, 0.1\}$.

Method	Significance Level		
	0.01	0.05	0.10
None	14	24	25
Benjamini-Hochberg	10	14	21
Bonferroni	9	10	10
Holm	9	10	10
Joint Confidence Interval	9	10	10
Romano-Wolf	9	10	10

Table 3.2: **Number of rejected hypotheses, gender wage gap example.**

```
# Extract (unadjusted) p-values
pvals.unadj = p_adjust(effects, method = "none")

# Coefficients and p-values; show first rows of output only
head(pvals.unadj)

# Rejections at 1%, 5%, and 10% significance levels
#levels = list(0.01, 0.05, 0.1)
#lapply(levels, function(x) sum(pvals.unadj[,"pval"]< x))
```

```
##                Estimate.   pval
## fem              -0.0460 0.4854
## fem:indAGRI      -0.1124 0.0380
## fem:indCONSTR    -0.0525 0.1465
## fem:indMANUF     -0.0110 0.6780
## fem:indTRANS     -0.0418 0.1547
## fem:indRETAIL     0.0284 0.3005
## fem:indFINANCE   -0.1349 0.0000
```

Thus, without correction for multiple testing, 14, 24, and 25 hypotheses could be rejected given the significance levels of 1%, 5% and 10%, respectively. If one returns to the initial example on variation by industry, one would find significant variation of the wage gap by industry (as compared to the baseline category “Wholesale” in 3 categories, namely “Agriculture”, “Finance, Insurance, and Real Estate” and “Professional and Related Services” at a significance level of 0.1.

Second, classical correction methods like the Bonferroni, Bonferroni-Holm, and the Benjamini-Hochberg adjustments are used to account for testing the 62 hypotheses at the same time.

```
# Bonferroni
pvals.bonf = p_adjust(effects, method = "bonferroni")

# Holm
pvals.holm = p_adjust(effects, method = "holm")

head(pvals.bonf)
```

```
##                Estimate. pval
## fem              -0.0460    1
## fem:indAGRI      -0.1124    1
## fem:indCONSTR    -0.0525    1
## fem:indMANUF     -0.0110    1
## fem:indTRANS     -0.0418    1
## fem:indRETAIL     0.0284    1
## fem:indFINANCE   -0.1349    0
```

```
head(pvals.holm)
```

```
##                 Estimate. pval
## fem               -0.0460    1
## fem:indAGRI       -0.1124    1
## fem:indCONSTR     -0.0525    1
## fem:indMANUF      -0.0110    1
## fem:indTRANS      -0.0418    1
## fem:indRETAIL      0.0284    1
## fem:indFINANCE    -0.1349    0
```

As a general improvement, the Holm-corrected p-values are smaller or equal to those obtained from a Bonferroni adjustment. At significance levels 1%, 5% and 10%, it is possible to reject fewer hypotheses if p-values are corrected for multiple testing. As displayed in Table 3.2, nine hypotheses can be rejected according to the Bonferroni and the Holm procedure at a significance level of 1%. If the significance level of 5% and 10% are considered, 10 hypotheses can be rejected.
According to the Benjamini-Hochberg (BH) correction of p-values that achieves control of the FDR, it is possible to reject 10, 14 and 21 null hypotheses at specified values of the FDR, γ, at 0.01, 0.05 and 0.1.

```
pvals.BH = p_adjust(effects, method = "BH")
head(pvals.BH)
```

```
##                 Estimate.   pval
## fem               -0.0460 0.6404
## fem:indAGRI       -0.1124 0.1024
## fem:indCONSTR     -0.0525 0.2930
## fem:indMANUF      -0.0110 0.8084
## fem:indTRANS      -0.0418 0.2998
## fem:indRETAIL      0.0284 0.4903
## fem:indFINANCE    -0.1349 0.0000
```

Regarding variation by industry, the Bonferroni and Holm procedure find a significantly different wage gap (at the 10% significance level) only for industry “Finance, Insurance, and Real Estate”, whereas the Benjamini-Hochberg correction with $\gamma = 0.1$ finds a second significant effect for the industry “Professional and Related Services”.
The p-values can be adjusted according to the Romano-Wolf-stepdown algorithm by setting the option `method = ‘RW’` (default) of the `p_adjust()` call. The number of repetitions can be varied by specifying the option `B`, $B = 1000$ by default. Although the Romano-Wolf stepdown procedure leads to smaller p-values as compared to the Bonferroni and Holm correction, it is not possible to reject more hypotheses in the example considered.

```
set.seed(123)
pvals.RW = p_adjust(effects, method = "RW", B = 1000)

head(pvals.RW)
```

```
##                 Estimate.  pval
## fem               -0.0460 1.000
## fem:indAGRI       -0.1124 0.689
## fem:indCONSTR     -0.0525 0.980
## fem:indMANUF      -0.0110 1.000
## fem:indTRANS      -0.0418 0.980
## fem:indRETAIL      0.0284 1.000
## fem:indFINANCE    -0.1349 0.000
```

Finally, we can construct a simultaneous $(1-\alpha)$ confidence interval using the `confint()` command with specified option joint = TRUE. This is equivalent to performing a joint significance test at level α.

```
alpha = 0.1

set.seed(123)
CI = confint(effects, level = 1-alpha, joint = TRUE, B = 1000)
head(CI)
```

In line with the previous results, setting $\alpha = 0.1$ leads to 9 rejected hypotheses with a joint 0.9 confidence interval.

3.7 Conclusion

The previous sections provide a short overview on important methods for multiple testing adjustment in a high-dimensional regression setting. Throughout the paper, our intention was to present the concepts and the necessity of a multiple adjustment in a comprehensive way. Similarly, the tools for valid simultaneous inference in high-dimensional settings that are available in the R package `hdm` are intended to be easy to use in empirical applications. The demonstration of the methods in the real-data example are intended to motivate applied statisticians to (i) use modern statistical methods for high-dimensional regression, i.e., the lasso, and (ii) to appropriately adjust if multiple hypotheses are tested simultaneously. Since the `hdm` provides user-friendly adjustment methods for objects of the S3 class `lm`, we hope that applied researchers and data scientists will use the correction methods more frequently, even in classical least squares regression.

3.8 Appendix

3.8.1 Computation and Infrastructure

The simulation study has been run on a x86_64redhatlinux-gnu (64-bit) (CentOS Linux 7 (Core)) cluster using R version 3.5.3 (2019-03-11). All lasso estimations are performed using the R package `hdm`, version `0.3.1` (Chernozhukov, Hansen, and Spindler 2016a) which can be downloaded from CRAN. For the knockoff procedure, the R package `knockoff` (Patterson and Sesia, 2018) has been used. The package is available from CRAN.

3.8.2 Additional Results, Simulation Study

n	K	Fix-KO	Model-X	Model-X (Or.)
200	60	345	413	514
200	180	994	992	512
200	400	1000	0	511
500	60	134	185	497
500	200	291	416	496

Table 3.3: **Number of repetitions with selection of zero variables, Knockoffs.**

3.8.3 Additional Results, Real-Data Example

The complete summary outputs of the global significance test and the joint significance test for the target coefficients with lasso and double selection are omitted in the main text for the sake of brevity. For completeness, the output is presented in the following.

```
# run rlasso
lasso1 = rlasso(X,y)

# run global test
summary(lasso1, all = FALSE)
```

```
##
## Call:
## rlasso.default(x = X, y = y)
##
## Post-Lasso Estimation:  TRUE
##
## Total number of variables: 123
## Number of selected variables: 58
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.509477 -0.274888 -0.007866  0.255786  2.667454
##
##                                   Estimate
## (Intercept)                          4.696
## fem                                  0.008
## married                              0.116
## chld19                               0.088
## regionMiddle Atlantic Division       0.075
## regionEast North Central Div.       -0.037
## regionWest North Central Div.       -0.073
## regionEast South Central Div.       -0.120
## regionMountain Division             -0.059
## regionPacific Division               0.132
## msa                                  0.185
## indAGRI                             -0.217
## indMANUF                             0.076
## indTRANS                             0.092
## indRETAIL                           -0.136
## indFINANCE                           0.180
## indBUISREPSERV                       0.119
## indENTER                            -0.083
## indPROFE                            -0.057
## indADMIN                            -0.046
## occBus Operat Spec                  -0.043
## occComput/Math                       0.016
## occLife/Physical/Soc Sci.           -0.184
## occComm/Soc Serv                    -0.341
## occLegal                             0.109
## occEduc/Training/Libr               -0.350
## occArts/Design/Entert/Sports/Media   -0.168
## occHealthc Pract/Technic              0.105
## occProtect Serv                      -0.071
## occOffice/Administr Supp             -0.317
## occProd                              -0.324
## hw50to59                              0.215
## hw60to69                              0.284
## hw70plus                              0.252
## degComp/Inform Sci                    0.160
## degEngin                              0.207
## degEnglish/Lit/Compos                -0.061
## degLib Arts/Hum                      -0.060
## degBio/Life Sci                       0.040
## degMath/Stats                         0.156
## degPhys Fit/Parks/Recr/Leis          -0.077
## degPsych                             -0.041
## degCrim Just/Fire Prot               -0.041
## degPubl Aff/Policy/Soc Wo            -0.045
## degSoc Sci                            0.081
## degFine Arts                         -0.081
## degBus                                0.080
## degHist                              -0.063
## yos                                   0.111
## exp                                   0.033
## fem:indAGRI                          -0.073
## fem:indFINANCE                       -0.126
## fem:occArchit/Engin                   0.034
## fem:occOffice/Administr Supp         -0.023
## fem:degPubl Aff/Policy/Soc Wo        -0.007
## fem:exp2                             -0.017
## fem:married                          -0.048
## fem:chld19                           -0.041
## fem:regionMountain Division          -0.007
##
## Residual standard error: 0.4633
## Multiple R-squared:  0.3906
## Adjusted R-squared:  0.3901
## Joint significance test:
##  the sup score statistic for joint significance test is 279.6 with a p-value of     0
```

```
# Summary of significance test (no correction of p-values)
summary(effects)
```

```
##                                         Estimate. Std. Error t value Pr(>|t|)
## fem                                      -0.0460     0.0660 -0.6976   0.4854
## fem:indAGRI                              -0.1124     0.0542 -2.0752   0.0380
## fem:indCONSTR                            -0.0525     0.0361 -1.4520   0.1465
## fem:indMANUF                             -0.0110     0.0265 -0.4152   0.6780
## fem:indTRANS                             -0.0418     0.0294 -1.4230   0.1547
## fem:indRETAIL                             0.0284     0.0274  1.0353   0.3005
## fem:indFINANCE                           -0.1349     0.0264 -5.1126   0.0000
## fem:indBUISREPSERV                        -0.0356     0.0273 -1.3046   0.1920
## fem:indPERSON                             -0.0517     0.0420 -1.2305   0.2185
## fem:indENTER                              -0.0476     0.0458 -1.0394   0.2986
## fem:indPROFE                              -0.0548     0.0254 -2.1591   0.0308
## fem:indADMIN                              -0.0227     0.0277 -0.8185   0.4131
## fem:occBus Operat Spec                     0.0266     0.0162  1.6430   0.1004
## fem:occFinanc Spec                        -0.0506     0.0176 -2.8688   0.0041
## fem:occComput/Math                        -0.0030     0.0176 -0.1737   0.8621
## fem:occArchit/Engin                        0.0552     0.0220  2.5030   0.0123
## fem:occLife/Physical/Soc Sci.              0.0532     0.0230  2.3100   0.0209
## fem:occComm/Soc Serv                       0.1567     0.0194  8.0759   0.0000
## fem:occLegal                               0.0071     0.0256  0.2759   0.7827
## fem:occEduc/Training/Libr                  0.1123     0.0142  7.9243   0.0000
## fem:occArts/Design/Entert/Sports/Media     0.0432     0.0203  2.1277   0.0334
## fem:occHealthc Pract/Technic               0.0018     0.0212  0.0859   0.9315
## fem:occProtect Serv                        0.0210     0.0327  0.6417   0.5211
## fem:occSales                              -0.0169     0.0176 -0.9624   0.3358
## fem:occOffice/Administr Supp              -0.0094     0.0157 -0.5995   0.5488
## fem:occProd                                0.0135     0.0399  0.3376   0.7357
## fem:hw40to49                              -0.0548     0.0188 -2.9159   0.0035
## fem:hw50to59                              -0.0711     0.0204 -3.4914   0.0005
## fem:hw60to69                              -0.1290     0.0261 -4.9439   0.0000
## fem:hw70plus                              -0.2015     0.0414 -4.8652   0.0000
## fem:degAgri                                0.0137     0.0375  0.3668   0.7138
## fem:degComm                                0.0414     0.0199  2.0759   0.0379
## fem:degComp/Inform Sci                    -0.0650     0.0300 -2.1641   0.0305
## fem:degEngin                              -0.0019     0.0249 -0.0765   0.9390
## fem:degEnglish/Lit/Compos                  0.0308     0.0238  1.2972   0.1946
## fem:degLib Arts/Hum                        0.0602     0.0371  1.6241   0.1044
## fem:degBio/Life Sci                       -0.0330     0.0223 -1.4752   0.1402
## fem:degMath/Stats                         -0.0547     0.0341 -1.6036   0.1088
## fem:degPhys Fit/Parks/Recr/Leis           -0.0084     0.0273 -0.3080   0.7581
## fem:degPhys Sci                           -0.0591     0.0271 -2.1844   0.0289
## fem:degPsych                              -0.0114     0.0222 -0.5126   0.6082
## fem:degCrim Just/Fire Prot                -0.0745     0.0267 -2.7869   0.0053
## fem:degPubl Aff/Policy/Soc Wo             -0.0340     0.0440 -0.7717   0.4403
## fem:degSoc Sci                            -0.0515     0.0189 -2.7231   0.0065
## fem:degFine Arts                          -0.0194     0.0211 -0.9172   0.3590
## fem:degMed/Hlth Sci Serv                  -0.0301     0.0249 -1.2102   0.2262
## fem:degBus                                -0.0152     0.0164 -0.9259   0.3545
## fem:degHist                               -0.0613     0.0259 -2.3688   0.0178
## fem:yos                                    0.0069     0.0034  2.0285   0.0425
## fem:exp                                   -0.0019     0.0037 -0.5138   0.6074
## fem:exp2                                  -0.0078     0.0093 -0.8425   0.3995
## fem:married                               -0.0542     0.0089 -6.0739   0.0000
## fem:chld19                                -0.0507     0.0094 -5.4045   0.0000
## fem:regionMiddle Atlantic Division        -0.0239     0.0155 -1.5445   0.1225
## fem:regionEast North Central Div.         -0.0116     0.0148 -0.7837   0.4332
## fem:regionWest North Central Div.         -0.0146     0.0176 -0.8282   0.4076
## fem:regionSouth Atlantic Division         -0.0022     0.0149 -0.1478   0.8825
## fem:regionEast South Central Div.          0.0049     0.0197  0.2475   0.8045
## fem:regionWest South Central Div.        -0.0741     0.0174 -4.2652   0.0000
## fem:regionMountain Division              -0.0321     0.0180 -1.7827   0.0746
## fem:regionPacific Division               -0.0611     0.0161 -3.8013   0.0001
## fem:msa                                   0.0029     0.0134  0.2185   0.8270
```

Chapter 4

Uniform Inference in High-Dimensional Additive Models

4.1 Introduction

Nonparametric regression allows for estimation of the relationship f between a target variable Y and input variables $X = (X_1, \dots, X_p)^T$ without imposing restrictive functional assumptions

$$Y = f(X_1, \dots, X_p) + \varepsilon,$$

where ε denotes the random error term satisfying $\mathrm{E}[\varepsilon|X] = 0$. However, in settings with a large number of regressors p, possibly with p exceeding the number of observations n, the well-known curse of dimensionality makes it practically impossible to estimate the regression function $f(X_1, \dots, X_p)$. A very popular approach in statistics and econometrics to overcome this limitation of nonparametric estimation in practice is to impose an additive structure of the regression function leading to additive models

$$Y = \alpha + f_1(X_1) + \ldots + f_p(X_p) + \varepsilon, \tag{4.1}$$

where α is a constant and $f_j(\cdot)$, $j = 1, \dots, p$, are smooth univariate functions. The idea of additive models can be traced back to Friedman and Stuetzle (1981), Stone (1985) and Hastie and Tibshirani (1990). Estimation and inference in the low-dimensional setting with fixed p has been analyzed widely in the literature. For an introduction to additive models, we refer to the textbook treatments in Hastie and Tibshirani (1990) and Wood (2017). In recent years, considerable progress has been made in understanding and analyzing additive models in high-dimensional settings that allow the number of components to grow with the sample size. For example, the theoretical literature has provided results on the estimation rate in high-dimensional additive models, as in the work by Sardy and Tseng (2004), Lin and Zhang (2006) and many others (Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010; Koltchinskii and Yuan, 2010; Kato, 2012; Petersen et al., 2016; Lou et al., 2016). The derived theoretical guarantees in the high-dimensional setting rely on a sparsity assumption that requires that only a small number s of the components are non-zero. From an intuitive point of view, this assumption allows the model being endowed with additional structure if the number of covariates are allowed to grow with sample size. Despite the considerable efforts that have been made to gain theoretical insights in high-dimensional additive models, only few studies have been concerned with valid inference in this class of models, for example regarding the construction of valid hypothesis tests or confidence regions. Härdle (1989), Sun and Loader (1994), Fan and Zhang (2000), Claeskens and Keilegom (2003) and Zhang and

Peng (2010) provide approaches to construct confidence bands in the widely studied setting with fixed dimensions. Only recently new results on valid inference in additive models in a high-dimensional setting have been derived. We review these results in the following and highlight our contribution to the existing literature.

Our work contributes to the expanding literature on high-dimensional inference, especially to the debiased/double machine learning literature. The double machine learning approach (Belloni et al., 2014c) offers a general framework for uniformly valid inference in high-dimensional settings. Alternative approaches for valid confidence intervals for low-dimensional parameters in high-dimensional linear models were also derived in van de Geer et al. (2014) and Zhang and Zhang (2014). These studies are based on the so-called debiasing approach that provides an alternative framework for valid inference. The framework involves a one-step correction of the lasso estimator and, thus, gives rise to an asymptotically normally distributed estimator. For a survey on post-selection inference in high-dimensional settings and generalizations, we refer to Chernozhukov et al. (2015b).

In a recent contribution, which is related to our work, Kozbur (2015) proposes a so-called post-nonparametric double selection approach for a scalar functional of one component.We consider the same setting as Kozbur (2015), i.e., a more general additively separable model

$$Y = f_1(X_1) + f_{-1}(X_2, \dots, X_p) + \varepsilon,$$

that includes the additive model

$$Y = \alpha + f_1(X_1) + \ldots + f_p(X_p) + \varepsilon.$$

The focus in Kozbur (2015) is on inference on functionals of the form $\theta = a(f_1)$ leading to results on pointwise confidence intervals that are based on a penalized series estimator. Our framework allows to extend these results and to clarify the underlying assumptions. First, building on recent results on inference on high-dimensional target parameters by Belloni et al. (2018) and Belloni et al. (2014a), we are able to establish uniformly valid confidence bands for the whole function f_1. Second, Kozbur (2015) relies on two high level assumptions on lasso estimation and variable selection (see Assumptions 9 and 10 in Kozbur (2015)) that might be difficult to verify. Hence, we clarify the technical requirements and provide results on uniform lasso estimation that are necessary to establish valid inference.

In a recent study, which is based on the debiasing approach by Zhang and Zhang (2014) mentioned earlier, Gregory et al. (2016) propose an estimator for the first component f_1 in a high-dimensional additive model where the number of additive components p may increase with the sample size. The estimator is constructed in two steps. In the first step, an undersmoothed estimator based on near-orthogonal projections with a group lasso bias correction is constructed. A debiased version of the first step estimator is used to generate pseudo responses $\hat{Y}$. These pseudo responses are then used in the second step that involves a smoothing method being applied to a nonparametric regression problem with $\hat{Y}$ and covariates X_1. Under sparsity assumptions on the number of nonzero additive components, the so-called oracle property is shown. Accordingly, the proposed estimator in Gregory et al. (2016) is asymptotically equivalent to the oracle estimator that is based on the true functions $f_2, \dots, f_p$. The asymptotics of the oracle estimator are well understood and carry over to the proposed debiasing estimate including methodology to construct uniformly valid confidence intervals for f_1. Nevertheless, Gregory et al. (2016) do not explicitly focus on inference. In their analysis, much stronger assumptions are required to obtain the oracle property. For example, normally distributed errors that are independent to X are assumed as well as a bounded support of X. Similarly to our framework, a large set of basis functions is chosen, for example polynomials or splines, to approximate the components f_1 and f_{-1}. A feature that distinguishes our work from the work in Gregory et al. (2016) is that we allow the degree of approximating functions to grow to infinity with increasing sample size.

A procedure that explicitely addresses the construction of uniformly valid confidence bands for the components in high-dimensional additive models has been developed in Lu et al. (2020). The authors emphasize that unformly valid inference in these models is a challenging problem, as a direct generalization of the ideas for the fixed dimensional case is difficult. Whereas confidence bands are mostly built upon kernel methods in the low-dimensional case, the estimators for high-dimensional sparse additive models are typically sieve estimators based on dictionaries. To derive their results, Lu et al. (2020) combine both kernel and sieve methods to utilize the advantages of each method resulting in a kernel-sieves hybrid estimator. This leads to a two-step estimator with many tuning parameters, for example the bandwidth and penalization levels that need to be chosen by cross-validation. Due to the local structure of the hybrid estimator, the framework of Lu et al. (2020) differs from ours in that they consider an additive local approximation model with sparsity (ATLAS), in which they only need to impose a local sparsity structure.

The advantage of the estimator, which we propose in the following, is that we do not have to leave the sieves framework and are nevertheless able to establish the uniform validity of the resulting confidence bands. This is possible as we consider the problem as a high-dimensional Z-estimation problem utilizing recent results from Belloni et al. (2018). We also provide a theory driven choice of the penalization level involved in the lasso estimation steps that makes computationally intense cross-validation procedures obsolete. Similarly to Gregory et al. (2016), Lu et al. (2020) assume normally distributed errors that are independent to X. Our model framework allows us to refrain from the normality assumption and only requires sub-exponential tails of the distribution of the error term. Moreover, the main results are also compatible with a heteroskedastic error. Finally, we can overcome the requirement in Lu et al. (2020) that the number of non-zero components $s = O(1)$ is bounded. Instead, s may grow to infinity with increasing sample size.

The finite sample properties of our estimator are evaluated in a simulation study that is based on the data generating processes in Gregory et al. (2016). The results show that the suggested method is able to perform valid simultaneous inference even in small samples and high-dimensional settings. Finally, we include an empirical application to the Boston housing data and provide evidence on nonlinear effects of socio-economic factors on house prices.

4.1.1 Organization of the Paper

The paper is organized as follows. Section 4.2 introduces and motivates the main regression problem in a high-dimensional additive model. Section 4.3 provides an introduction to the estimation method. In Section 4.4, the main result is provided. A simulation study, highlighting the small sample properties and implementation of our proposed method, is presented in Section 4.5. Section 4.6 illustrates the use of the method in an empirical application to the Boston housing data. The proof of the main theorem is provided in Section 4.7. The Appendix includes additional technical material. In Appendix 4.8, a general result for uniform inference on a high-dimensional linear functional is presented. Appendix 4.9 provides results regarding uniform lasso estimation rates in high dimensions. Finally, computational details are presented in Appendix 4.10.

4.1.2 Notation

Throughout the paper, we consider a random element W from some common probability space $(\Omega, \mathcal{A}, P)$. We denote by $P \in \mathcal{P}_n$ a probability measure out of a large class of probability measures, which may vary with the sample size (since the model is allowed to change with n) and by $\mathbb{P}_n$ the empirical probability measure. Additionally, let $\mathbb{E}$ respectively $\mathbb{E}_n$ be the expectation with respect to P, respectively $\mathbb{P}_n$, and $\mathbb{G}_n(\cdot)$ denotes the empirical process

$$\mathbb{G}_n(f) := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(W_i) - \mathbb{E}[f(W_i)]\right)$$

for a class of suitably measurable functions $\mathcal{F} : \mathcal{W} \to \mathbb{R}$. $\|\cdot\|_{P,q}$ denotes the $L^q(P)$-norm. In the following, we write $\|\cdot\|_{\Psi_\rho}$ for the Orlicz-norm that is defined as

$$\|W\|_{\Psi_\rho} := \inf\left\{C > 0 : \mathbb{E}\left[\exp((|W|/C)^\rho) - 1\right] \le 1\right\}$$

for $\rho > 1$. Furthermore, $\|v\|_1 = \sum_{l=1}^{p} |v_l|$ denotes the ℓ_1-norm, $\|v\|_2 = \sqrt{v^T v}$ the ℓ_2-norm and $\|v\|_0$ equals the number of non-zero components of a vector $v \in \mathbb{R}^p$. We define $v_{-l} := (v_1, \dots, v_{l-1}, v_{l+1} \dots, v_p)^T \in \mathbb{R}^{p-1}$ for any $1 \le l \le p$. $\|v\|_\infty = \sup_{l=1,\dots,p} |v_l|$ denotes the sup-norm. Let c and C denote positive constants independent of n with values that may change at each appearance. The notation $a_n \lesssim b_n$ means $a_n \le C b_n$ for all n and some C. Furthermore, $a_n = o(1)$ denotes that there exists a sequence $(b_n)_{\ge 1}$ of positive numbers such that $|a_n| \le b_n$ for all n, where b_n is independent of $P \in \mathcal{P}_n$ for all n, and b_n converges to zero. Finally, $a_n = O_P(b_n)$ means that for any $\epsilon > 0$ there exists a C such that $P(a_n > C b_n) \le \epsilon$ for all n.

4.2 Setting

4.2.1 Motivation and Illustration

Before we introduce the formal setting in Section 4.2.2, we would like to motivate the basic ideas in a simplified example. We consider an additive model with two components

$$f(x) = f_1(x_1) + f_2(x_2) + \varepsilon. \tag{4.2}$$

Our goal is to perform valid inference on the object of interest f_1, in other words we would like to provide a uniform confidence band for f_1 as illustrated in an example in Figure 4.1. Hence, we consider f_2 in Equation (4.2) as a nuisance function. Next, we assume that it is possible to represent the two components by an approximately linear representation. For the first component, the representation is given by

$$f_1(X_1) = \theta_0^T g(X_1) + b_1(X_1).$$

Here, $g(X_1)$ is a basis (e.g., a spline basis, sieve terms or a polynomial series) consisting of d_1 terms, θ_0 is the corresponding coefficient vector and $b_1(X_1)$ is an approximation error. The existence of such a sparse linear approximation is a common assumption in high-dimensional statistics that states that only a subset of the coefficients in θ_0 have a coefficient that is different from zero. For the second component, we assume an analogous representation

$$f_2(X_2) = \beta_0^T h(X_2) + b_2(X_2), \tag{4.3}$$

where the basis $h(X_2)$ consists of d_2 terms. We allow the dimensions d_1 and d_2 to grow with the sample size in order to establish the asymptotic results in high-dimensional settings. Accordingly, the number of components p can grow with the sample size, as well.[1] To derive a uniformly valid confidence band,

---

[1] However, for the ease of notation we will later subsume them in the component $f_{-1}(X_2, \dots, X_p)$.

estimation and inference of

$$f_1(\cdot) \approx \theta_0^T g(\cdot) = \sum_{l=1}^{d_1} \theta_{0,l} g_l(\cdot) \tag{4.4}$$

is required. In a naive approach to estimate the first component $\theta_{0,1}$ of the vector θ_0, one could use lasso or other machine learning methods to select the relevant basis expansion terms in θ_0 and β_0 in the regression model (4.2). A possible second step would be to estimate a regression of the dependent variable on all components that have been selected in the lasso estimation step. The final estimator for the first component $\theta_{0,1}$ might be obtained from this regression, and the procedure could be repeated iteratively for all other components in $f_1(x_1)$. However, this approach is problematic because it fails to deliver valid results. In other words, the use of the lasso estimator makes it necessary to take the variable selection into account and, hence to deal with the non-standard problem of post-selection inference. Intuitively, the naive procedure that is based on only one lasso estimation step might involve selection mistakes that will invalidate the resulting confidence bands for $\theta_{0,1}$. The critical selection mistakes will not arise with regard to variables that have high predictive power for the dependent variable Y, but rather refer to variables that are potentially highly correlated with the basis expansion term $g_1()$. As a result, an omitted variable bias might arise that prevents the estimator from asymptotically converging to a normal distribution. To overcome this limitation of post-selection inference, the statistical literature has developed the Double Machine Learning and the Debiasing approach that we mentioned in the introduction.
To address the potential bias introduced by the lasso estimation, Belloni et al. (2014c) propose to include an auxiliary regression for the corresponding covariate of the target parameter. Here, we consider

$$g_1(X_1) = \gamma_0^T Z_{-1} + \nu \tag{4.5}$$

where ν is an error term and Z_{-1} is defined as

$$Z_{-1} = (g_2(X_1), \dots, g_{d_1}(X_1), h_1(X_{-1}), \dots, h_{d_2}(X_{-1}))^T.$$

Later, we will also allow for an approximation error in this equation. Belloni et al. (2014c) propose to include in the final regression not only the covariates selected in the first step of the naive approach but to augment this set of variables with Lasso-selected regressors from the auxiliary regression. This procedure is equivalent to constructing a so-called Neyman-orthogonal moment function with respect to the nuisance part. This is key for valid post-selection inference for the first component of the vector θ_0. In Section 4.2.2, we will provide more details about this property. Heuristically, the additional regression step in Equation (4.5) will lead to robustness against moderate selection mistakes. It can be shown, that this procedure implements an orthogonal moment equation

$$\mathbb{E}\left[\psi_1(W, \theta_{0,1}, \eta_{0,1})\right] = 0,$$

where the first component of θ_0 is our target parameter and all other involved parameters are considered as nuisance parameters. Belloni et al. (2014c) established an approach for valid inference for one parameter. In high-dimensional additive models, the major technical challenge is that we have to conduct inference for the potentially high-dimensional vector θ_0, in other words the number of elements in θ_0 for which we would like to construct a valid confidence region is allowed to grow with the sample size. Each component of θ_0, $\theta_{0,l}$ with $l = 1, \dots, d_1$, is determined by an orthogonal moment condition and we will show how uniformly valid confidence bands can be constructed by embedding the problem as a high-dimensional Z-estimation problem. Finally, we show how the estimation of θ_0 can be translated

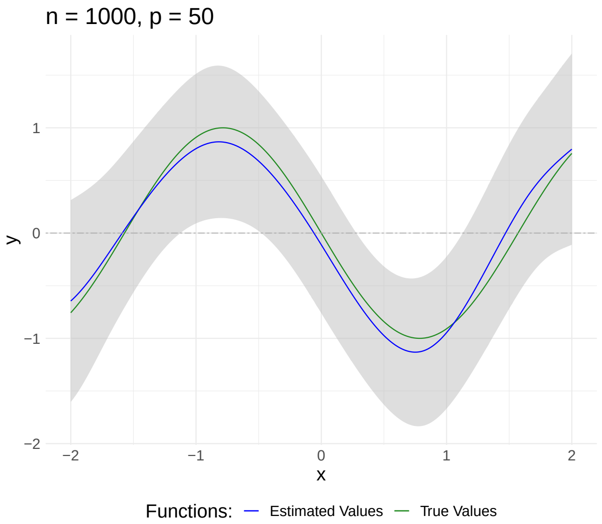


Figure 4.1: **Illustration of the estimator and confidence bands, simulation example.**

Predicted component $f_1(x_1)$ with 95%-confidence bands (gray shaded area) obtained by our proposed estimator for a component $f_1(x_1)$ with $f_1(x_1) = -\sin(2 \cdot x)$. The green curve corresponds to the true function $f_1(x_1)$. The predicted component $\hat{f}_1(x_1)$, which is obtained from our proposed estimator, is illustrated by the blue curve. The underlying data are generated according to the DGP in the simulation study in Section 4.5 with $n = 1000$ observations and $p = 50$ regressors. For more information on the DGP, we refer to the description of the DGP in Section 4.5.

to uniformly valid confidence intervals for the target function f_1 using a multiplier bootstrap procedure.

Let us now illustrate our estimation procedure in the motivating example in Equation (4.2) with two additive components f_1 and f_2 in a step-by-step explanation.

1. Perform valid inference on the lth component $\theta_{0,l}$ of θ_0 in Equation (4.4), where index l with $l = 1, \ldots, d_1$ indicates the target parameter under consideration. To obtain valid coefficient estimates and estimates of the variance covariance matrix, we
   1a. Estimate the potentially high-dimensional nuisance parameters by lasso regression. The nuisance terms include
      - The coefficient vector β_0 in the representation of f_2 in Equation (4.3),
      - All remaining components θ_k in θ_0 with $k \neq l$ for the representation of f_1 in Equation (4.4), and
      - The coefficient vector γ_0 in the auxiliary regression in Equation (4.5).

   1b. Plug in these estimates into the moment conditions $\mathbb{E}\left[\psi_l(W, \theta_{0,l}, \eta_{0,l})\right] = 0$ and solve these for the target parameter $\theta_{0,l}$.
2. This estimation method results in a de-biased estimator $\hat{\theta}_0$ of the target parameter that leads to the following estimator of the target component f_1:

$$\hat{f}_1(\cdot) \approx \hat{\theta}_0^T g(\cdot).$$

3. Using an appropriate multiplier bootstrap procedure allows us to construct uniformly valid confidence bands for $f_1(x)$ based on this estimator.

Figure 4.1 illustrates the use of our estimation by providing a preview of our simulation studies in Section 4.5. Our estimation methods provides an unbiased estimate for the target component f_1 with corresponding $(1-\alpha)$ confidence interval that is valid over a compact interval of X_1. In the next section, we consider a more general additively separable model and introduce a general formulation of the underlying problem.

4.2.2 Formal Setting

Consider the following nonparametric additively separable model

$$Y = f(X) + \varepsilon = f_1(X_1) + f_{-1}(X_{-1}) + \varepsilon$$

with $\mathbb{E}[\varepsilon|X] = 0$ and $\mathrm{Var}(\varepsilon|X) \geq c$. Let the scalar response Y and features $X = (X_1, \dots, X_p)$ take values in $\mathcal{Y}$, respectively in $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$. We assume to observe n i.i.d. copies $(W^{(i)})_{i=1}^n = (Y^{(i)}, X^{(i)})_{i=1}^n$ of $W = (Y, X)$, where the number of covariates p is allowed to grow with the sample size n. For identifiability, we assume $\mathbb{E}[f_{-1}(X_{-1})] = 0$. We aim to construct uniformly valid confidence regions for the first nonparametric component of the regression function, namely we want to find functions $\hat{l}(x)$ and $\hat{u}(x)$ converging to $f_1(x)$ with

$$P\left(\hat{l}(x) \leq f_1(x) \leq \hat{u}(x), \forall x \in I\right) \to 1 - \alpha.$$

Here, $I \subseteq \mathcal{X}_1$ is a bounded interval of interest where we want to conduct inference. We approximate f_1 and f_{-1} by a linear combination of approximating functions $g_1, \dots, g_{d_1}$ and $h_1, \dots, h_{d_2}$, respectively. Define

$$g(x_1) := (g_1(x_1), \dots, g_{d_1}(x_1))^T$$

for $x_1 \in \mathbb{R}$ and

$$h(x_{-1}) := (h_1(x_{-1}), \dots, h_{d_2}(x_{-1}))^T$$

for $x_{-1} \in \mathbb{R}^{p-1}$. It is important to note that we allow the number of approximating functions d_1 and d_2 to increase with sample size. Assume that the approximations are given by

$$f_1(X_1) = \theta_0^T g(X_1) + b_1(X_1), \tag{4.6}$$

where $\theta_{0,l} \in \Theta_l$ and analogously

$$f_{-1}(X_{-1}) := \beta_0^T h(X_{-1}) + b_2(X_{-1}), \tag{4.7}$$

where b_1 and b_2 denote the error terms. Additionally, it is convenient to define the combination

$$z(x) := (g_1(x_1), \dots, g_{d_1}(x_1), h_1(x_{-1}), \dots, h_{d_2}(x_{-1}))^T$$

for $x \in \mathbb{R}^p$, where we abbreviate

$$Z := z(X) = (g_1(X_1), \dots, g_{d_1}(X_1), h_1(X_{-1}), \dots, h_{d_2}(X_{-1}))^T.$$

For each element g_l of g, we consider

$$g_l(X_1) = (\gamma_0^{(l)})^T Z_{-l} + b_3^{(l)}(Z_{-l}) + \nu^{(l)} \tag{4.8}$$

with $\mathbb{E}[\nu^{(l)}|Z_{-l}] = 0$ and $\mathrm{Var}(\nu^{(l)}|Z_{-l}) \geq c$. This corresponds to

$$\mathbb{E}[g_l(X_1)|Z_{-l}] = (\gamma_0^{(l)})^T Z_{-l} + b_3^{(l)}(Z_{-l}),$$

with approximation error $b_3^{(l)}(Z_{-l})$. The second stage equation (4.8) is used to construct an orthogonal score function for valid inference in a high-dimensional setting as described in Section 4.2.1. Estimating

$$f_1(\cdot) \approx \theta_0^T g(\cdot)$$

can be recast into a general Z-estimation problem of the form

$$\mathbb{E}\left[\psi_l(W, \theta_{0,l}, \eta_{0,l})\right] = 0, \quad l \in 1, \ldots, d_1$$

with target parameter θ_0 where the score functions are defined by

$$\begin{aligned}\psi_l(W, \theta, \eta) = &\Big(Y - \theta g_l(X_1) - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X)\Big) \\ &\cdot \Big(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l})\Big).\end{aligned}$$

Here,

$$\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}, \eta^{(4)})^T$$

with $\eta^{(1)} \in \mathbb{R}^{d_1+d_2-1}, \eta^{(2)} \in \mathbb{R}^{d_1+d_2-1}$, $\eta^{(3)} \in \ell^\infty(\mathbb{R}^p)$ and $\eta^{(4)} \in \ell^\infty(\mathbb{R}^{d_1+d_2-1})$ are nuisance functions. The true nuisance parameter $\eta_{0,l}$ is given by

$$\begin{aligned}\eta_{0,l}^{(1)} &:= \beta_0^{(l)} \\ \eta_{0,l}^{(2)} &:= \gamma_0^{(l)} \\ \eta_{0,l}^{(3)}(X) &:= b_1(X_1) + b_2(X_{-1}) \\ \eta_{0,l}^{(4)}(Z_{-l}) &:= b_3^{(l)}(Z_{-l}),\end{aligned}$$

where $\beta_0^{(l)}$ is defined as

$$\beta_0^{(l)} := (\theta_{0,1}, \ldots, \theta_{0,l-1}, \theta_{0,l+1}, \ldots \theta_{0,d_1}, \beta_{0,1}, \ldots, \beta_{0,d_2})^T.$$

Essentially, the index l determines which coefficient is not contained in $\beta_0^{(l)}$. The third part of the nuisance functions captures the error made by the approximation of f_1 and f_{-1}, which is independent from l. Therefore, we sometimes omit l.

Comment 4.2.1. *The score ψ is linear in θ, meaning*

$$\psi_l(W, \theta, \eta) = \psi_l^a(X, \eta^{(2)}, \eta^{(4)})\theta + \psi_l^b(X, \eta)$$

with

$$\psi_l^a(X, \eta^{(2)}, \eta^{(4)}) = -g_l(X_1)(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l}))$$

and

$$\psi_l^b(X, \eta) = (Y - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X))(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l}))$$

for all $l = 1, \ldots, d_1$.

Comment 4.2.2. *The score function ψ satisfies the moment condition, namely*

$$\mathbb{E}\left[\psi_l(W, \theta_{0,l}, \eta_{0,l})\right] = 0$$

for all $l = 1, \ldots, d_1$, and, given further conditions mentioned in Section 4.4, the near Neyman orthogo-

nality condition

$$D_{l,0}[\eta,\eta_{0,l}] := \partial_t\big\{\mathbb{E}[\psi_l(W,\theta_{0,l},\eta_{0,l}+t(\eta-\eta_{0,l}))]\big\}\big|_{t=0} \lesssim \delta_n n^{-1/2},$$

where ∂_t *denotes the derivative with respect to* t *and* $(\delta_n)_{n\geq 1}$ *a sequence of positive constants converging to zero.*

4.3 Estimation

In this section, we describe our estimation method and how the uniform valid confidence bands are constructed. The nuisance functions are estimated by lasso regressions. Finally, they are plugged into the moment conditions and solved for the target parameters, which yield an estimate $\hat{f}_1$ for the first component in the additive regression model. The lower and upper curve of the confidence bands are finally based on the estimated covariance matrix and a critical value which is determined by a multiplier bootstrap procedure. The technical details for the estimation are given in this section.

Let

$$g(x) = (g_1(x),\dots,g_{d_1}(x))^T \in \mathbb{R}^{d_1\times 1},$$

and

$$\psi(W,\theta,\eta) = (\psi_1(W,\theta_1,\eta_1),\dots,\psi_{d_1}(W,\theta_{d_1},\eta_{d_1}))^T \in \mathbb{R}^{d_1\times 1}$$

for some vector

$$\theta = (\theta_1,\dots,\theta_{d_1})^T$$

and

$$\eta = (\eta_1,\dots,\eta_{d_1})^T.$$

For each $l = 1,\dots,d_1$, let $\hat{\eta}_l = \left(\hat{\eta}_l^{(1)},\hat{\eta}_l^{(2)},\hat{\eta}_l^{(3)},\hat{\eta}_l^{(4)}\right)$ be an estimator of the nuisance function. The estimator $\hat{\theta}_0$ of the target parameter

$$\theta_0 = (\theta_{0,1},\dots,\theta_{0,d_1})^T$$

is defined as the solution of

$$\sup_{l=1,\dots,d_1}\left\{\left|\mathbb{E}_n\left[\psi_l(W,\hat{\theta}_l,\hat{\eta}_l)\right]\right| - \inf_{\theta\in\Theta_l}\left|\mathbb{E}_n\left[\psi_l(W,\theta,\hat{\eta}_l)\right]\right|\right\} \leq \epsilon_n, \tag{4.9}$$

where $\epsilon_n = o\left(\delta_n n^{-1/2}\right)$ is the numerical tolerance. Finally, the target function $f_1(\cdot)$ can be estimated by

$$\hat{f}_1(\cdot) := \hat{\theta}_0^T g(\cdot). \tag{4.10}$$

Define the Jacobian matrix

$$J_0 := \frac{\partial}{\partial\theta}\mathbb{E}[\psi(W,\theta,\eta_0)]\bigg|_{\theta=\theta_0} = \operatorname{diag}(J_{0,1},\dots,J_{0,d_1}) \in \mathbb{R}^{d_1\times d_1}$$

with

$$J_{0,l} = E[\psi_l^a(W,\eta_{0,l}^{(2)},\eta_{0,l}^{(4)})]$$

$$
\begin{aligned}
&= -\mathbb{E}[((\gamma_0^{(l)})^T Z_{-l} + b_3^{(l)}(Z_{-l}) + \nu^{(l)})\nu^{(l)}] \\
&= -\mathbb{E}\Big[\big((\gamma_0^{(l)})^T Z_{-l} + b_3^{(l)}(Z_{-l})\big) \underbrace{\mathbb{E}[\nu^{(l)}|Z_{-l}]}_{=0}\Big] - \mathbb{E}[(\nu^{(l)})^2] \\
&= -\mathbb{E}[(\nu^{(l)})^2]
\end{aligned}
$$

for all $l = 1, \dots, d_1$. Observe that

$$\mathbb{E}\big[\psi(W,\theta_0,\eta_0)\psi(W,\theta_0,\eta_0)^T\big] =: \Sigma_{\varepsilon\nu}$$

is the covariance matrix of $\varepsilon\nu := (\varepsilon\nu^{(1)}, \dots, \varepsilon\nu^{(d_1)})$. Define the approximate covariance matrix

$$
\begin{aligned}
\Sigma_n :&= J_0^{-1}\mathbb{E}\big[\psi(W,\theta_0,\eta_0)\psi(W,\theta_0,\eta_0)^T\big](J_0^{-1})^T \\
&= J_0^{-1}\Sigma_{\varepsilon\nu}(J_0^{-1})^T \in \mathbb{R}^{d_1 \times d_1}
\end{aligned}
$$

with

$$
\Sigma_n := \begin{pmatrix}
\frac{\mathbb{E}[(\varepsilon\nu^{(1)})^2]}{\mathbb{E}[(\nu^{(1)})^2]^2} & \frac{\mathbb{E}\left[\varepsilon\nu^{(1)}\varepsilon\nu^{(2)}\right]}{\mathbb{E}[(\nu^{(1)})^2]\mathbb{E}[(\nu^{(2)})^2]} & \cdots & \frac{\mathbb{E}\left[\varepsilon\nu^{(1)}\varepsilon\nu^{(d_1)}\right]}{\mathbb{E}[(\nu^{(1)})^2]\mathbb{E}[(\nu^{(d_1)})^2]} \\
\frac{\mathbb{E}\left[\varepsilon\nu^{(2)}\varepsilon\nu^{(1)}\right]}{\mathbb{E}[(\nu^{(2)})^2]\mathbb{E}[(\nu^{(1)})^2]} & \frac{\mathbb{E}[(\varepsilon\nu^{(2)})^2]}{\mathbb{E}[(\nu^{(2)})^2]^2} & \cdots & \frac{\mathbb{E}\left[\varepsilon\nu^{(2)}\varepsilon\nu^{(d_1)}\right]}{\mathbb{E}[(\nu^{(2)})^2]\mathbb{E}[(\nu^{(d_1)})^2]} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\mathbb{E}\left[\varepsilon\nu^{(d_1)}\varepsilon\nu^{(1)}\right]}{\mathbb{E}[(\nu^{(d_1)})^2]\mathbb{E}[(\nu^{(1)})^2]} & \frac{\mathbb{E}\left[\varepsilon\nu^{(d_1)}\varepsilon\nu^{(2)}\right]}{\mathbb{E}[(\nu^{(d_1)})^2]\mathbb{E}[(\nu^{(1)})^2]} & \cdots & \frac{\mathbb{E}[(\varepsilon\nu^{(d_1)})^2]}{\mathbb{E}[(\nu^{(d_1)})^2]^2}
\end{pmatrix}.
$$

The approximate covariance matrix can be estimated by replacing every expectation by the empirical analog and plugging in the estimated parameters

$$
\begin{aligned}
\hat{\Sigma}_n :&= \hat{J}^{-1}\mathbb{E}_n\big[\psi(W,\hat{\theta},\hat{\eta})\psi(W,\hat{\theta},\hat{\eta})^T\big](\hat{J}^{-1})^T \\
&= \hat{J}^{-1}\hat{\Sigma}_{\varepsilon\nu}(\hat{J}^{-1})^T \\
&= \begin{pmatrix}
\frac{\mathbb{E}_n[(\hat{\varepsilon}\hat{\nu}^{(1)})^2]}{\mathbb{E}_n[(\hat{\nu}^{(1)})^2]^2} & \frac{\mathbb{E}_n\left[\hat{\varepsilon}\hat{\nu}^{(1)}\hat{\varepsilon}\hat{\nu}^{(2)}\right]}{\mathbb{E}_n[(\hat{\nu}^{(1)})^2]\mathbb{E}_n[(\hat{\nu}^{(2)})^2]} & \cdots & \frac{\mathbb{E}_n\left[\hat{\varepsilon}\hat{\nu}^{(1)}\hat{\varepsilon}\hat{\nu}^{(d_1)}\right]}{\mathbb{E}_n[(\hat{\nu}^{(1)})^2]\mathbb{E}_n[(\hat{\nu}^{(d_1)})^2]} \\
\frac{\mathbb{E}_n\left[\hat{\varepsilon}\hat{\nu}^{(2)}\hat{\varepsilon}\hat{\nu}^{(1)}\right]}{\mathbb{E}_n[(\hat{\nu}^{(2)})^2]\mathbb{E}_n[(\hat{\nu}^{(1)})^2]} & \frac{\mathbb{E}_n[(\hat{\varepsilon}\hat{\nu}^{(2)})^2]}{\mathbb{E}_n[(\hat{\nu}^{(2)})^2]^2} & \cdots & \frac{\mathbb{E}_n\left[\hat{\varepsilon}\hat{\nu}^{(2)}\hat{\varepsilon}\hat{\nu}^{(d_1)}\right]}{\mathbb{E}_n[(\hat{\nu}^{(2)})^2]\mathbb{E}_n[(\hat{\nu}^{(d_1)})^2]} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\mathbb{E}_n\left[\hat{\varepsilon}\hat{\nu}^{(d_1)}\hat{\varepsilon}\hat{\nu}^{(1)}\right]}{\mathbb{E}_n[(\hat{\nu}^{(d_1)})^2]\mathbb{E}_n[(\hat{\nu}^{(1)})^2]} & \frac{\mathbb{E}_n\left[\hat{\varepsilon}\hat{\nu}^{(d_1)}\hat{\varepsilon}\hat{\nu}^{(2)}\right]}{\mathbb{E}_n[(\hat{\nu}^{(d_1)})^2]\mathbb{E}_n[(\hat{\nu}^{(1)})^2]} & \cdots & \frac{\mathbb{E}_n[(\hat{\varepsilon}\hat{\nu}^{(d_1)})^2]}{\mathbb{E}_n[(\hat{\nu}^{(d_1)})^2]^2}
\end{pmatrix}.
\end{aligned}
$$

This estimated covariance matrix can be used to construct the confidence bands

$$
\begin{aligned}
\hat{u}(x) &:= \hat{f}_1(x) + \frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}c_\alpha}{\sqrt{n}} \\
\hat{l}(x) &:= \hat{f}_1(x) - \frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}c_\alpha}{\sqrt{n}},
\end{aligned}
$$

where c_α is a critical value determined by the following standard multiplier bootstrap method introduced in Chernozhukov et al. (2013a). Define

$$\hat{\psi}_x(\cdot) := (g(x)^T\hat{\Sigma}_n g(x))^{-1/2} g(x)^T \hat{J}^{-1}\psi(\cdot, \hat{\theta}_0, \hat{\eta}_0)$$

and let

$$\hat{\mathcal{G}} = \left(\hat{\mathcal{G}}_x\right)_{x\in I} = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi^{(i)}\hat{\psi}_x\left(W^{(i)}\right)\right)_{x\in I},$$

where $(\xi^{(i)})_{i=1}^n$ are independent standard normal random variables (especially independent from the data $(W^{(i)})_{i=1}^n$). The multiplier bootstrap critical value c_α is given by the $(1-\alpha)$-quantile of the conditional distribution of $\sup_{x\in I}|\hat{\mathcal{G}}_x|$ given $(W^{(i)})_{i=1}^n$. This estimation procedure can be summarized in the following algorithm:

Algorithm 1 HDAM

Input: n training examples of the form $W^{(i)} = (Y^{(i)}, X_1^{(i)}, X_{-1}^{(i)})$, where $Y^{(i)}$ is the response, $X_1^{(i)}$ the covariate of interest and $X_{-1}^{(i)}$ are additional covariates. Dictionaries of the approximating functions $g_1,\dots,g_{d_1}$ for f_1 and $h_1,\dots,h_{d_2}$ for f_{-1}, a significance level α, an interval I for inference and a number of bootstrap repetitions B.

```
Use the dictionary to construct the matrix $Z := (g_1(X_1),\dots,g_{d_1}(X_1),h_1(X_{-1}),\dots,h_{d_2}(X_{-1}))$.
Fit a Lasso/post-Lasso/sqrt-Lasso regression of the vector $Y$ onto $Z$ and save the estimated
  coefficients $(\tilde{\theta}_1,\dots,\tilde{\theta}_{d_1},\hat{\beta}_1,\dots,\hat{\beta}_{d_2})$ and the corresponding residuals $\hat{\varepsilon}$.
for $l = 1,\dots,d_1$ do
    Fit a Lasso/post-Lasso/sqrt-Lasso regression of the vector $g_l(X_1)$ onto $Z_{-l}$ and save the estimated
      coefficients $(\hat{\gamma}_1^{(l)},\dots\hat{\gamma}_{d_1+d_2-1}^{(l)})$ and the corresponding residuals $\hat{\nu}^{(l)}$.
    Plug in the estimated coefficients as nuisance parameters into the score function $\psi_l(W,\cdot,\hat{\eta}_l))$ to
      solve (4.9). Save the resulting estimate $\hat{\theta}_l$ and scores $\psi_l(W,\hat{\theta}_l,\hat{\eta}_l)$ into the corresponding vector
      $\hat{\theta}_0$ and matrix $\psi(W,\hat{\theta}_0,\hat{\eta}_0)$, respectively.
end for
Use the estimated residuals $\hat{\varepsilon}$ and $\hat{\nu}$ to construct the estimates $\hat{\Sigma}_n$ and $\hat{J}$.
for $x \in I$ do
    Calculate the vector $\hat{\psi}_x(W) := (g(x)^T\hat{\Sigma}_n g(x))^{-1/2}g(x)^T\hat{J}^{-1}\psi(W,\hat{\theta}_0,\hat{\eta}_0)$
    for $b = 1,\dots,B$ do
        Draw $(\xi_i^{(b)})_{i=1}^n$ independent standard normal random variables.
        Calculate $\hat{\mathcal{G}}_x^{(b)} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i^{(b)}\hat{\psi}_x(W^{(i)})$.
    end for
end for
Calculate the critical value $c_\alpha := (1-\alpha)$-quantile of $\sup_{x\in I}|\hat{\mathcal{G}}_x^{(b)}|$ with respect to the bootstrap
  repetitions.
for $x \in I$ do
    Construct the confidence band as $\hat{\theta}_0^T g(x) \pm \frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}c_\alpha}{\sqrt{n}}$
end for
```

4.4 Main Results

Now, we specify the conditions that are required to provide uniformly valid confidence bands by Algorithm 1. Since we would like to represent f_1 and f_{-1} by their approximations in (4.6) and (4.7), we need to choose an appropriate set of approximating functions $g = (g_1,\dots,g_{d_1})$ and $h = (h_1,\dots,h_{d_2})$, respectively. In this context, let $\bar{d}_n := \max(d_1, d_2, n, e)$ and C be a strictly positive constant independent of n and l, where e in $\bar{d}_n$ denotes the Euler's number. $(A_n)_{n\geq 1}$ denotes a sequence of positive constants, possibly going to infinity with $A_n \geq n$ for all n. For details, we refer to Appendix 4.8. The number of non-zero coefficients in Equation (4.6) and (4.8), respectively, is given by the sparsity index s. Additionally, we set $t_1 := \sup_{x \in I} \|g(x)\|_0 \leq d_1$. The definition of t_1 is helpful if the functions g_l, $l = 1, \ldots, d_1$, are local in the sense that for any point x in I there are at most $t_1 << d_1$ non-zero functions. Further, $g(I)$ denotes the image of the approximation functions with respect to the interval of interest I.

The following assumptions hold uniformly in $n \geq n_0$ and $P \in \mathcal{P}_n$:

Assumption A. 1.

(i) It holds

$$\inf_{x \in I} \|g(x)\|_2^2 \geq c > 0, \quad \sup_{x \in I} \sup_{l=1,\ldots,d_1} |g_l(x)| \leq C < \infty$$

and for all $\varepsilon > 0$

$$\log N(\varepsilon, g(I), \|\cdot\|_2) \leq C t_1 \log\left(\frac{A_n}{\varepsilon}\right).$$

(ii) There exists $1 \leq \rho \leq 2$ *such that*

$$\max_{l=1,\ldots,d_1} \|b_3^{(l)}(Z_{-l})\|_{\Psi_\rho} \leq C, \quad \|b_1(X_1) + b_2(X_{-1})\|_{\Psi_\rho} \leq C.$$

Additionally, the approximation errors obey

$$\mathbb{E}\Big[\big(b_1(X_1) + b_2(X_{-1})\big)^2\Big] \leq C s \log(\bar{d}_n)/n,$$
$$\max_{l=1,\ldots,d_1} \mathbb{E}\Big[\big(b_3^{(l)}(Z_{-l})\big)^2\Big] \leq C s \log(\bar{d}_n)/n$$

and

$$\mathbb{E}_n\Big[\big(b_1(X_1) + b_2(X_{-1})\big)^2\Big] - \mathbb{E}\Big[\big(b_1(X_1) + b_2(X_{-1})\big)^2\Big] \leq C s \log(\bar{d}_n)/n,$$
$$\max_{l=1,\ldots,d_1} \Big(\mathbb{E}_n\Big[\big(b_3^{(l)}(Z_{-l})\big)^2\Big] - \mathbb{E}\Big[\big(b_3^{(l)}(Z_{-l})\big)^2\Big]\Big) \leq C s \log(\bar{d}_n)/n$$

with probability $1 - o(1)$.

(iii) We have

$$\sup_{\|\xi\|_2 = 1} \mathbb{E}\left[(\xi^T Z)^2 \big(b_1(X_1) + b_2(X_{-1})\big)^2\right] \leq C \mathbb{E}\left[\big(b_1(X_1) + b_2(X_{-1})\big)^2\right]$$

and

$$\sup_{\|\xi\|_2 = 1} \mathbb{E}\left[(\xi^T Z)^2 \big(b_3^{(l)}(Z_{-l})\big)^2\right] \leq C \mathbb{E}\left[\big(b_3^{(l)}(Z_{-l})\big)^2\right]$$

for $l = 1, \ldots, d_1$.

(iv) It holds

$$\mathbb{E}\Big[\nu^{(l)}\big(b_1(X_1) + b_2(X_{-1})\big)\Big] \leq C \delta_n n^{-1/2}$$

with $\delta_n = o\big(t_1^{-\frac{3}{2}} \log^{-\frac{1}{2}}(A_n)\big)$.

Assumption A.1(i) contains regularity conditions on g. We assume that the infimum of the ℓ_2-norm of $g(x)$ is bounded away from zero, but the supremum is allowed to increase with sample size (affecting the growth conditions in A.2(v)). The lower bound on the infimum is not necessary and can be replaced by a decaying sequence at the cost of stricter growth rates. The Assumptions A.1(ii) and (iii) are tail and moment conditions on the approximation error. These assumptions are mild since the number of approximating functions may increase with sample size. Finally, Assumption A.1(iv) ensures that the violation of the exact Neyman Orthogonality due to the approximation errors is negligible. It is worth to notice that if $b_1(X_1)$ and $b_2(X_{-1})$ are measurable with respect to Z_{-l} (for example in the linear approximate sparse setting for the conditional expectation) the exact Neyman Orthogonality holds. Now, we go more into detail regarding the condition on the covering number of the image of g. Especially if $t_1 < d_1$, the complexity of the approximating functions is reduced significantly. One obtains

$$g(I) \subseteq \bigcup_{j=1}^{\binom{d_1}{t_1}} g^{(j)}(I),$$

where each $g^{(j)}(I)$ is only dependent on t_1 nonzero components. It is straightforward to see that for each $g^{(j)}(I)$ the covering numbers satisfy

$$N(\varepsilon, g^{(j)}(I), \|\cdot\|_2) \leq \left(\frac{6\sup_{x\in I}\|g(x)\|_2}{\varepsilon}\right)^{t_1}$$

(cf. Van der Vaart and Wellner (1996)), implying

$$\begin{aligned}
\log N(\varepsilon, g(I), \|\cdot\|_2) &\leq \log\left(\sum_{j=1}^{\binom{d_1}{t_1}} N(\varepsilon, g^{(j)}(I), \|\cdot\|_2)\right)\\
&\leq \log\left(\left(\frac{e\cdot d_1}{t_1}\right)^{t_1}\left(\frac{6\sup_{x\in I}\|g(x)\|_2}{\varepsilon}\right)^{t_1}\right)\\
&\leq t_1\log\left(\left(\frac{6ed_1\sup_{x\in I}\|g(x)\|_2}{t_1}\right)\frac{1}{\varepsilon}\right)\\
&\leq Ct_1\log\left(\frac{d_1}{\varepsilon}\right).
\end{aligned}$$

For specific classes of approximating functions the complexity can be further reduced.

Assumption A. 2.

(i) For all $l = 1,\dots,d_1$, Θ_l *contains a ball of radius*

$$\log(\log(n))n^{-1/2}\log^{1/2}(d_1 \vee e)\log(n)$$

centered at $\theta_{0,l}$ *with*

$$\sup_{l=1,\dots,d_1}\sup_{\theta_l\in\Theta_l}|\theta_l| \leq C.$$

(ii) It holds

$$\|\beta_0^{(l)}\|_0 \leq s, \quad \|\beta_0^{(l)}\|_2 \leq C$$

for all $l = 1,\dots,d_1$ *and*

$$\max_{l=1,\dots,d_1}\|\gamma_0^{(l)}\|_0 \leq s, \max_{l=1,\dots,d_1}\|\gamma_0^{(l)}\|_2 \leq C.$$

(iii) There exists $1 \leq \rho \leq 2$ such that

$$\max_{j=1,\dots,d_1+d_2} \|Z_j\|_{\Psi_\rho} \leq C, \quad \|\varepsilon\|_{\Psi_\rho} \leq C.$$

(iv) It holds

$$\inf_{\|\xi\|_2=1} \mathbb{E}[(\xi^T Z)^2] \geq c \textit{ and } \sup_{\|\xi\|_2=1} \mathbb{E}[(\xi^T Z)^4] \leq C,$$

and the eigenvalues of the covariance matrix $\Sigma_{\varepsilon\nu}$ are bounded from above and away from zero.

(v) There exists a fixed $\bar{q} \geq 4$ such that

(a) $n^{\frac{1}{\bar{q}}} \frac{s^2 t_1^3 \log^{2+\frac{4}{\rho}}(\bar{d}_n) \log(A_n)}{n} = o(1)$,

(b) $n^{\frac{1}{\bar{q}}} \frac{\sup_{x\in I} \|g(x)\|_2^6 s t_1^4 \log(\bar{d}_n) \log^2(A_n)}{n} \left(\log^{\frac{2}{\rho}}(d_1) \vee s\sqrt{\frac{s \log(\bar{d}_n)}{n}} \right) = o(1)$,

(c) $n^{\frac{1}{\bar{q}}} \frac{t_1^{13} \log^{\frac{6}{\rho}}(d_1) \log^7(A_n)}{n} = o(1)$.

Assumptions A.2(i) and (ii) are regularity and sparsity conditions, where the number of nonzero regression coefficients $s = s_n$ is allowed to grow to infinity with increasing sample size. A detailed comment on the sparsity condition is given in Comment 4.4.2. Assumption A.2(iii) contains tail conditions on the approximating functions (and therefore on the original variables) as well as for the error term. Assumption A.2(iv) is a standard eigenvalue condition, which restricts the correlation between the basis elements (and therefore between the original variables). For example, if the conditional variance of $\nu^{(l)}$ is uniformly bounded away from zero, the second inequality of A.2(iv) holds. Finally, Assumption A.2(v) provides the growth conditions. These are given in general terms and depend on the choice of the approximation functions. Choosing B-Splines simplifies the growth conditions significantly as we will discuss in Comment 4.4.1.

Theorem 3. *Under the Assumptions A.1 and A.2, it holds that*

$$P\left(\hat{l}(x) \leq f_1(x) \leq \hat{u}(x), \forall x \in I\right) \to 1 - \alpha$$

uniformly over $P \in \mathcal{P}_n$ where c_α is a critical value determined by the multiplier bootstrap method.

Comment 4.4.1. *[**B-Splines**] An appropriate and common choice in series estimation are B-Splines. B-Splines are positive and local in the sense that $g(x) \geq 0$ and $\sup_{x\in I} \|g(x)\|_0 \leq t_1$ for every x, where t_1 is the degree of the spline. The l_1-norm of B-Splines is equal to 1, meaning*

$$\|g(x)\|_1 = \sum_{j=1}^{d_1} g_j(x) = 1$$

for every x (partition of unity). Hence, Assumption A.1(i) is met with

$$\frac{1}{\sqrt{t_1}} \leq \inf_{x\in I} \|g(x)\|_2^2 \leq \sup_{x\in I} \|g(x)\|_2^2 \leq 1 \quad \textit{and} \quad \sup_{x\in I} \sup_{l=1,\dots,d_1} |g_l(x)| \leq 1.$$

The covering numbers of $g(I)$ is given by

$$\log N(\varepsilon, g(I), \|\cdot\|_2) \leq \log \left(\sum_{j=1}^{d_1} N(\varepsilon, g^{(j)}(I), \|\cdot\|_2) \right)$$

$$\leq t_1 \log\left(\left(\frac{6d_1^{\frac{1}{t_1}} \sup_{x\in I} \|g(x)\|_2}{\varepsilon}\right)\right)$$
$$\leq C \log\left(\frac{d_1}{\varepsilon}\right).$$

Choosing the degree of the B-Splines of order $t_1 = \log(n)$, the growth rates in Assumption A.2(v) simplify to

$$n^{\frac{1}{q}} \frac{s^2 \log^{2+\frac{4}{\rho}}(\bar{d}_n)\log(d_1)}{n} = o(1) \quad \text{and} \quad n^{\frac{1}{q}} \frac{\log^{7+\frac{6}{\rho}}(d_1)}{n} = o(1).$$

It is worth to notice that in the first growth condition

$$n^{\frac{1}{q}} \frac{s^2 \log^{2+\frac{4}{\rho}}(\bar{d}_n)\log(d_1)}{n} = o(1)$$

both the total number of approximating functions d_1 and d_2, and the number of relevant functions s may grow with the sample size in a balanced way. If s is bounded, the number of approximating functions can grow at an exponential rate with the sample size. This means that the set of approximating functions can be much larger than the sample size, only the number of relevant function s has to be smaller than the sample size. This situation is common for lasso based estimators. Our growth condition is in line with other results in the literature, e.g., Belloni et al. (2018), Belloni et al. (2014a) and many others. The second growth condition ensures that

$$n^{\frac{1}{q}} \frac{\log^{7+\frac{6}{\rho}}(d_1)}{n} = o(1)$$

and is in line with Chernozhukov et al. (2013a). It guarantees the validity of multiplier bootstrap in our setting and allows us to construct uniformly valid confidence regions.

Comment 4.4.2. *The sparsity condition in A.2(ii) restricts the number of nonzero regression coefficients $s = s_n$ in the Equations (4.6), (4.7) and (4.8). Through this, we especially assume that the regression function f can be approximated sufficiently well by only s relevant basis functions. Note that we do not directly control the number of relevant covariables, but the number of approximating functions in total. This sparsity condition is different from the one used in Gregory et al. (2016) and Lu et al. (2020) who restrict the number of relevant additive components in the model (4.1). Our model also includes the approximate sparse setting due to the error terms b_1 and b_2 in (4.6) and (4.7). This is more flexible and more realistic for many applications.*
Furthermore, we do not define $\theta_0^T g(X_1)$ as the best projection of $f_1(X_1)$ in (4.6) (and $\beta_0^T h(X_{-1})$ for $f_{-1}(X_{-1})$ in (4.7)) as it is done in Gregory et al. (2016). We only assume a sparse projection that is closeto the best projection where the distance is measured in terms of $\|\cdot\|_{P,2}$ as described in Assumption A.1(ii).

4.5 Simulation Results

To verify the theoretical guarantees of our estimator in practice, we perform a simulation study, which is based on the settings in Gregory et al. (2016) and Meier et al. (2009). We consider the finite sample performance of our estimator in a high-dimensional additive model of the form

$$y_i = \sum_{j=1}^{p} f_j(x_{i,j}) + \epsilon_{i,j},$$

Component	Function
1	$f_1(x_1) = -\sin(2 \cdot x)$
2	$f_2(x_2) = x^2 - \frac{25}{12}$
3	$f_3(x_3) = x$
4	$f_4(x_4) = \exp(-x) - \frac{2}{5} \cdot \sinh(\frac{5}{2})$
$5, \ldots, p$	$f_j(x_j) = 0.$

Table 4.1: **Definitions of the data generating processes, simulation study.**

Definitions of functions in data generating processes. Data generating processes are based on settings in Gregory et al. (2016) and Meier et al. (2009).

with $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The definitions of the functions $f_j(x_j)$, $j = 1, \ldots, j$, are presented in Table 4.1. We extend the initial setting in Gregory et al. (2016) to allow for heteroskedasticity by specifying an error term $\epsilon_j \sim N(0, \sigma_j(x_j))$ with $\sigma_j(x_j) = \underline{\sigma} \cdot (1 + |x_j|)$ and $\underline{\sigma} = \sqrt{\frac{12}{67}}$. This value of $\underline{\sigma}$ ensures a signal-to-noise ratio that is comparable to the settings in Gregory et al. (2016). Data sets are generated for scenarios with dimensions $n \in \{100, 1000\}$ and $p \in \{50, 150\}$. In all cases, sparsity is imposed by only allowing the first four components, $f_1, ..., f_4$, to be non-zero. The regressors X are marginally uniformly distributed on an interval, $I = [-2.5, 2.5]$ with correlation matrix Σ with $\Sigma_{k,l} = 0.5^{|k-l|}$, $1 \leq k, l \leq p$, which corresponds to the setting in Gregory et al. (2016) with the strongest correlation structure among the covariates.

In the simulation, we use the estimator and the multiplier bootstrap procedure we proposed in Section 4.3 to generate predictions $\hat{f}_j(x_j)$ for the function $f_j(x_j)$ and construct simultaneous confidence bands that are defined in terms of $\hat{l}_j(x_j)$ and $\hat{u}_j(x_j)$. The functions $f_j(x_j)$ in the additive model are approximated using cubic B-splines. Variable selection is performed using post-lasso with a theory-based choice of the penalty level as implemented in the R package `hdm` (Chernozhukov et al., 2016a). Further details related to the implementation and parametrization in the simulation study can be found in Appendix 4.10.

Table 4.2 presents the empirical coverage achieved by the estimated simultaneous 95%-confidence bands in $R = 2000$ repetitions which are constructed over the interval of values of x_j, $I = [-2, 2]$. A confidence band is considered to cover the function $f_j(x_j)$ if it entirely contains the true function, or, stated more formally, if for all values of $x_j \in I$ it holds that $\hat{l}_j(x_j) \leq f_j(x_j) \leq \hat{u}_j(x_j)$.

The results confirm the validity of our inference method in high-dimensional additive models. In all cases, the empirical coverage approaches 95% or is above the nominal level. This can be observed even in settings with more regressors than observations, i.e., with $n = 100$ and $p = 150$. For example, in this setting the overall dimensionality amounts to $d_1 + d_2 = 1500$ if the degrees of freedom of the B-splines are set to $k = 10$.

n	p	f_1	f_2	f_3	f_4
100	50	0.956	0.985	0.950	0.979
100	150	0.957	0.976	0.967	0.957
1000	50	0.985	0.987	0.952	0.987
1000	150	0.989	0.975	0.982	0.986

Table 4.2: **Empirical coverage, simulation study.**

Coverage achieved by simultaneous 0.95%-confidence bands in $R = 2000$ repetitions as generated over a range of values of x_j, $I = [-2, 2]$.

The presented results refer to one particular choice of the parameter k that specifies the degrees of freedom

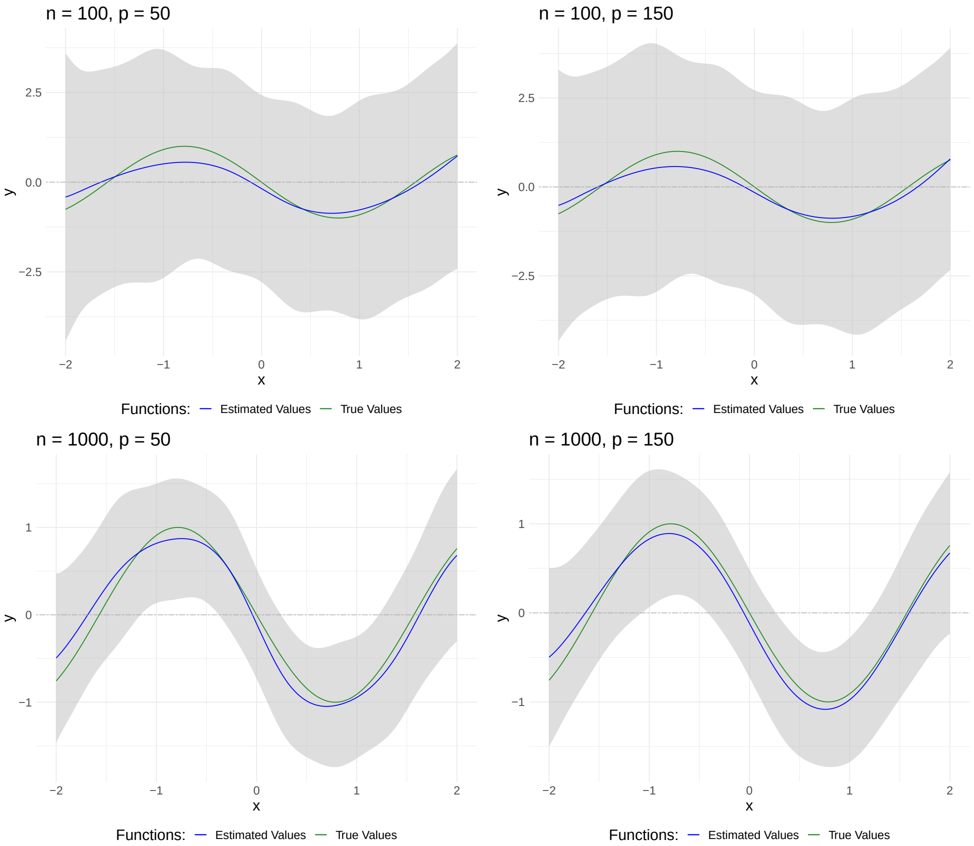


Figure 4.2: **Average confidence bands, $f_1(x_1)$, simulation study.**

Gray shaded areas illustrate averaged 95%-confidence bands obtained in $R = 2000$ repetitions for function $f_1(x_1)$. Blue curves correspond to the estimated functions $\hat{f}_j(x_j)$ and green curves to the true functions $f_j(x_j)$.

of the cubic B-splines as implemented in the R package `splines`. A table with the exact choice of k in all settings is presented in Appendix 4.10. In addition to the presented results, we experimented with the values of k and we conclude that the nominal coverage has been maintained in various parametrizations of the underlying spline components. The robustness with regard to the choice of the smoothing parameters provides additional support of the finite-sample validity of the proposed inferential procedure.

Figures 4.2 to 4.5 present the estimated confidence bands averaged over all $R = 2000$ repetitions. They illustrate that the estimation accuracy benefits from increasing sample size; the width of the confidence regions becomes smaller and the approximation of the true function improves in terms of accuracy. In several settings, we observe a slight bias emerging for values of x_j close to the boundary. Nonetheless, given the maintained coverage in all settings, the amount of this bias is tolerated by the estimator and the accompanying confidence bands.

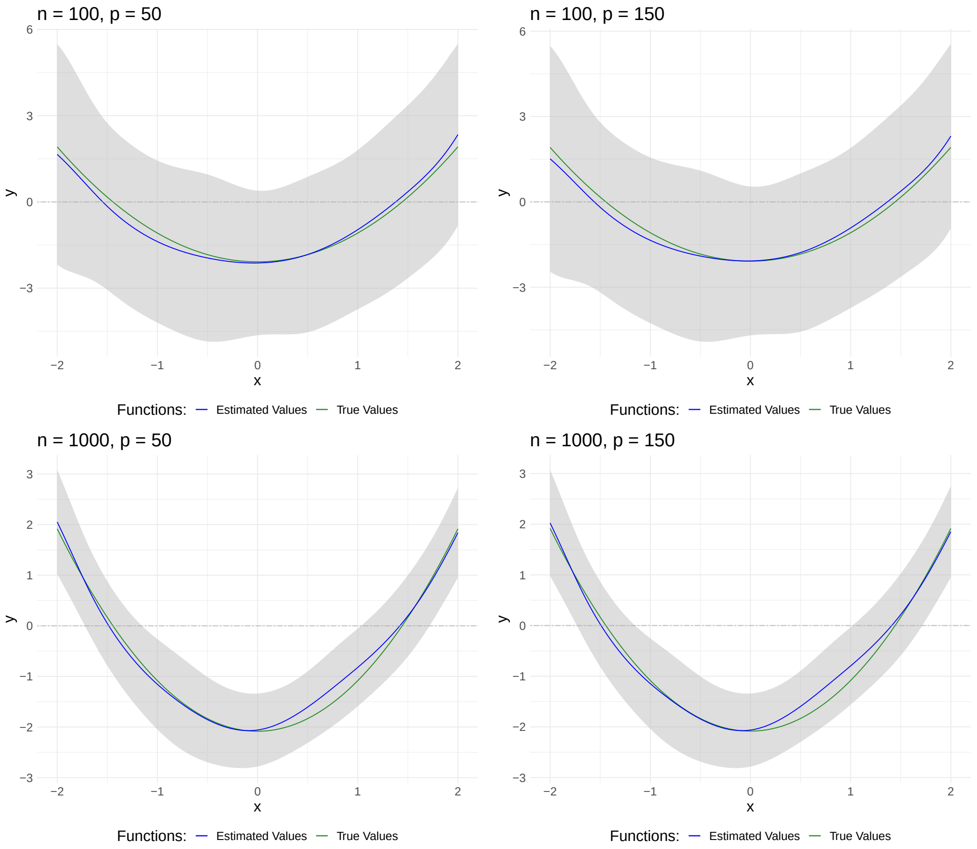


Figure 4.3: **Average confidence bands, $f_2(x_2)$, simulation study.**

Gray shaded areas illustrate averaged 95%-confidence bands obtained in $R = 2000$ repetitions for function $f_2(x_2)$. Blue curves correspond to the estimated functions $\hat{f}_j(x_j)$ and green curves to the true functions $f_j(x_j)$.

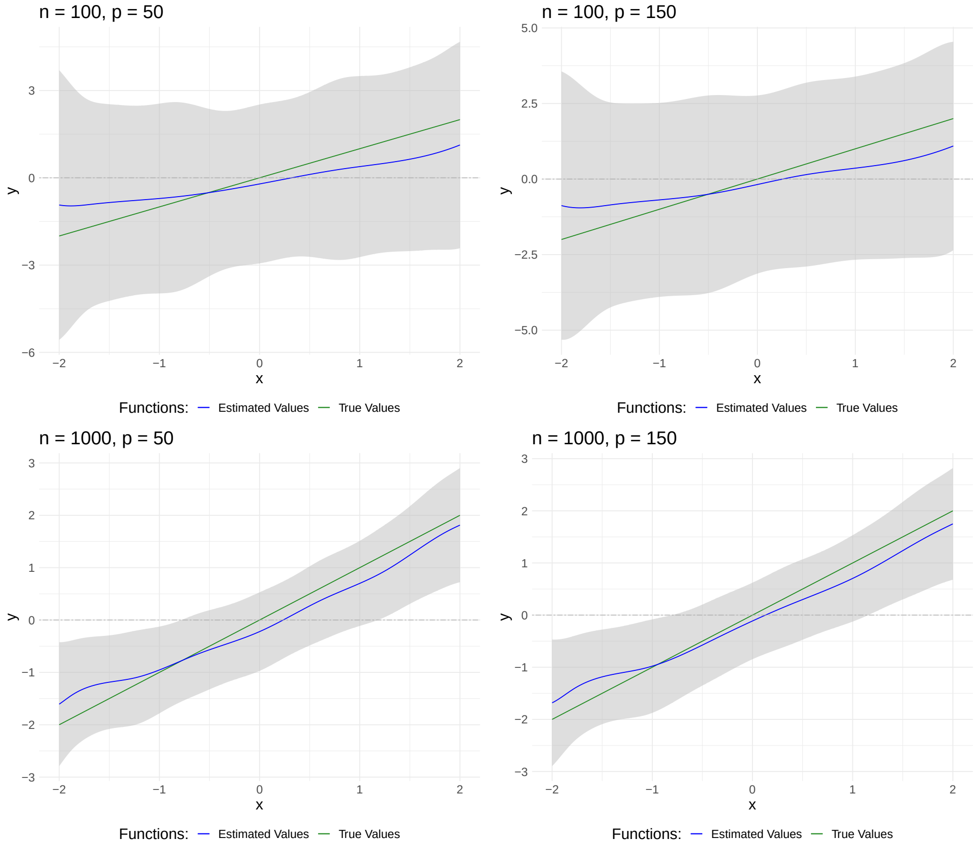


Figure 4.4: **Average confidence bands, $f_3(x_3)$, simulation study.**

Gray shaded areas illustrate averaged 95%-confidence bands obtained in $R = 2000$ repetitions for function $f_3(x_3)$. Blue curves correspond to the estimated functions $\hat{f}_j(x_j)$ and green curves to the true functions $f_j(x_j)$.

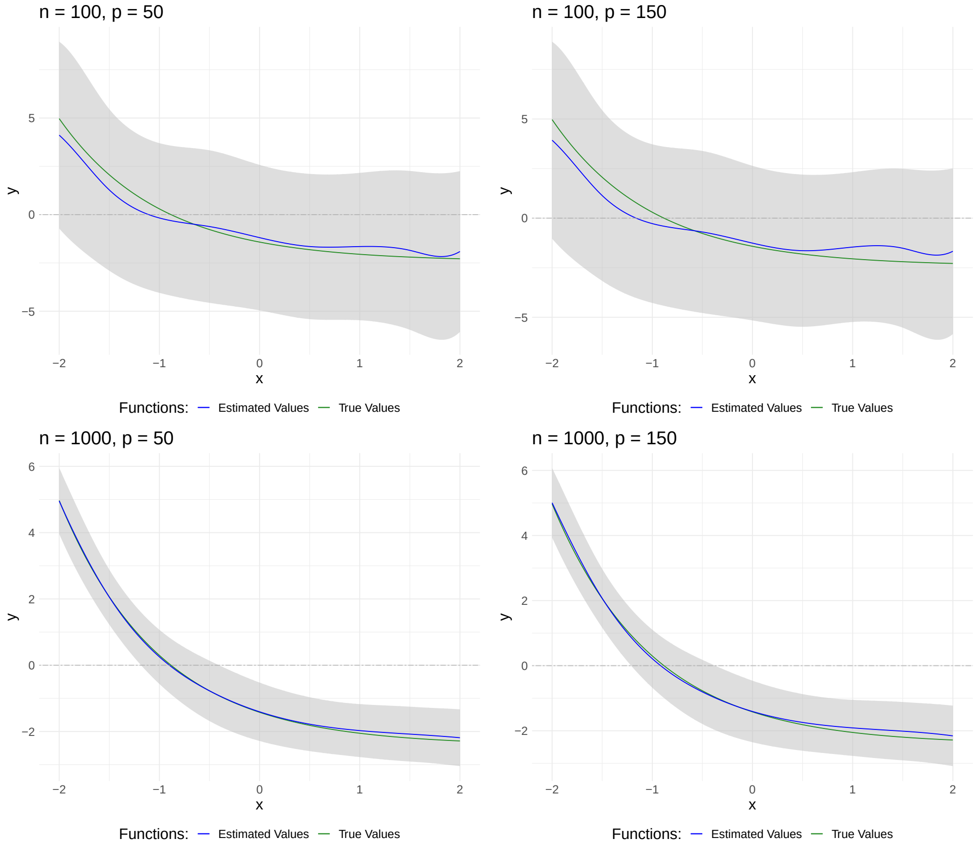


Figure 4.5: **Average confidence bands, $f_4(x_4)$, simulation study.**

Gray shaded areas illustrate averaged 95%-confidence bands obtained in $R = 2000$ repetitions for function $f_4(x_4)$. Blue curves correspond to the estimated functions $\hat{f}_j(x_j)$ and green curves to the true functions $f_j(x_j)$.

Variable	Description
$MEDV$	Median value of owner-occupied homes in USD 1000's
$LSTAT$	Percentage of lower status people of the population
$CRIM$	Per capita crime rate by town
NOX	Nitric oxides
TAX	Full-value property-tax rate per USD 10,000
AGE	Proportion of owner-occupied units built prior to 1940
$DIST$	Weighted distances to five Boston employment centres
RM	Average number of rooms per dwelling
$INDUS$	Proportion of non-retail business acres per town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft
$BLACK$	$1000(B-0.63)^2$ where B is the proportion of blacks by town
$PTRATIO$	Pupil-teacher ratio by town
$CHAS$	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

Table 4.3: **List of variables, Boston housing data example.**

4.6 Illustration in a Real-Data Example

As a real-data example, we apply our estimator to the Boston housing data that has been first used in Harrison Jr and Rubinfeld (1978) and later been reassessed in several studies, e.g., Kong and Xia (2012) and Doksum and Samarov (1995). The data set is available via the R package `mlbench` (Leisch and Dimitriadou, 2010; Newman et al., 1998). The data contain information on housing prices for $n = 506$ census tracts in Boston based on the 1970 census. We perform inference on the effect of 11 continuous variables on the dependent variable $MEDV$ which measures the median value of owner-occupied homes (in USD 1000's). A list of the explanatory variables is provided in Table 4.3.
The implemented model is given by

$$\begin{aligned}MEDV_i =& f_1(\mathrm{LSTAT}_i) + f_2(\mathrm{CRIM}_i) + f_3(\mathrm{NOX}_i) + f_4(\mathrm{TAX}_i) + \\ & f_5(\mathrm{AGE}_i) + f_6(\mathrm{DIST}_i) + f_7(\mathrm{RM}_i) + f_8(\mathrm{INDUS}_i) + \\ & f_9(\mathrm{ZN}_i) + f_{10}(\mathrm{BLACK}_i) + f_{11}(\mathrm{PTRATIO}_i) + \gamma \cdot \mathrm{CHAS} + \epsilon_i.\end{aligned}$$

Analogously to the simulation study, the functions $f_j(x_j)$ are approximated with cubic B-splines and variable selection is performed using post-lasso with theory-based choice of the penalty term. The smoothing parameters $k = \{k_j, k_{-j}\}$ have been determined according to a heuristic cross-validation rule that is outlined in Appendix 4.10. The results illustrated in Figure 4.6 suggest nonlinear and significant effects for the variables LSTAT and RM that are generally in line with economic intuition and the findings in Kong and Xia (2012) and Doksum and Samarov (1995). The variable LSTAT, the percentage of lower status people of the population, has a negative effect on the median home value. Whereas for small values of LSTAT, the estimated effect $\hat{f}_1(\mathrm{LSTAT})$ is sizable and positive, the effect decreases and eventually becomes negative for higher levels of the variable. The nonlinearities found for variable RM suggest that the average number of rooms per dwelling impacts housing prices strongly positively if the average number of rooms exceeds a value of 6.5. The results for the remaining regressors, which are presented in Appendix 4.10 also point at nonlinear effects that are not significant in most cases.

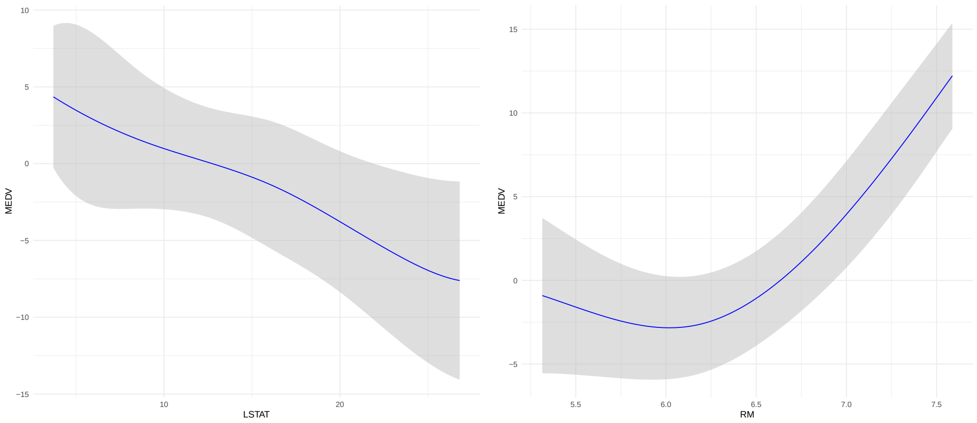


Figure 4.6: **Effects and confidence bands, Boston housing data example.**

Plots of $\hat{f}_1$(LSTAT) and $\hat{f}_7$(RM) with simultaneous 95%-confidence bands in the Boston housing data application.

4.7 Proofs

Proof of Theorem 3.
We will prove that the Assumptions A.1 and A.2 imply the Assumptions B.1-B.5 stated in Appendix 4.8 and then the claim follows by applying Theorem 4. Without loss of generality, we assume $\min(d_1, n) \geq e$ to simplify notation.

Assumption B.1
Both conditions (i) and (ii) are directly assumed in A.1(i). Due to A.1(ii) and A.2(iv) it holds

$$\begin{aligned}
\mathbb{E}\left[(\nu^{(l)})^2\right] &= \mathbb{E}\left[\left(g_l(X_1) - (\gamma_0^{(l)})^T Z_{-l} - b_3^{(l)}(Z_{-l})\right)^2\right] \\
&\leq C\left(\sup_{\|\xi\|_2=1} \mathbb{E}[(\xi^T Z)^2] + \mathbb{E}\left[\left(b_3^{(l)}(Z_{-l})\right)^2\right]\right) \\
&\lesssim C
\end{aligned}$$

where we used that $\|\gamma_0^{(l)}\|_2 \leq C$. It holds

$$\mathbb{E}\left[(\nu^{(l)})^2\right] \geq \operatorname{Var}(\nu^{(l)}|Z_{-l}) \geq c.$$

Since the eigenvalues of $\Sigma_{\varepsilon\nu}$ are bounded from above and away from zero,

$$\Sigma_n = J_0^{-1}\Sigma_{\varepsilon\nu}(J_0^{-1})^T \in \mathbb{R}^{d_1\times d_1}$$

directly implies B.1(iii).

Assumption B.2

For each $l = 1, \dots, d_1$, the moment condition holds

$$\begin{aligned}\mathbb{E}\left[\psi_l(W, \theta_{0,l}, \eta_{0,l})\right] &= \mathbb{E}\Big[\Big(Y - f(X)\Big)\Big(g_l(X_1) - (\gamma_0^{(l)})^T Z_{-l} - b_3^{(l)}(Z_{-l})\Big)\Big] \\ &= \mathbb{E}\big[\varepsilon \nu^{(l)}\big] \\ &= \mathbb{E}\big[\nu^{(l)} \underbrace{\mathbb{E}\left[\varepsilon | X\right]}_{=0}\big] \\ &= 0.\end{aligned}$$

For all $l = 1, \dots, d_1$, define the convex set

$$\begin{aligned}T_l := \Big\{\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}, \eta^{(4)})^T :& \eta^{(1)}, \eta^{(2)} \in \mathbb{R}^{d_1+d_2-1}, \\ & \eta^{(3)} \in \ell^\infty(\mathbb{R}^p), \eta^{(4)} \in \ell^\infty(\mathbb{R}^{d_1+d_2-1})\Big\}\end{aligned}$$

and endow ${}_l$ with the norm

$$\|\eta\|_e := \max\Big\{\|\eta^{(1)}\|_2, \|\eta^{(2)}\|_2, \|\eta^{(3)}(X)\|_{P,2}, \|\eta^{(4)}(Z_{-l})\|_{P,2}\Big\}.$$

Further, let $\tau_n := \sqrt{\frac{s\log(\bar{d}_n)}{n}}$ and define the corresponding nuisance realization set

$$\begin{aligned}\mathcal{T}_l := \Big\{\eta \in T_l :& \eta^{(3)} \equiv 0, \eta^{(4)} \equiv 0, \|\eta^{(1)}\|_0 \vee \|\eta^{(2)}\|_0 \le Cs, \\ & \|\eta^{(1)} - \beta_0^{(l)}\|_2 \vee \|\eta^{(2)} - \gamma_0^{(l)}\|_2 \le C\tau_n, \\ & \|\eta^{(1)} - \beta_0^{(l)}\|_1 \vee \|\eta^{(2)} - \gamma_0^{(l)}\|_1 \le C\sqrt{s}\tau_n\Big\} \cup \{\eta_{0,l}\}\end{aligned}$$

for a sufficiently large constant C. For arbitrary random variables X and Y, it holds

$$\begin{aligned}\|E[X|Y]\|_{\Psi_\rho} :&= \inf\{C > 0 : \mathbb{E}[\Psi_\rho(|\mathbb{E}[X|Y]|/C)] \le 1\} \\ &\le \inf\{C > 0 : \mathbb{E}[\mathbb{E}[\Psi_\rho(|X|/C)|Y]] \le 1\} \\ &= \|X\|_{\Psi_\rho}.\end{aligned}$$

Due to Assumption A.2(iii), this implies

$$\begin{aligned}\max_{l=1,\dots,d_1} \|\nu^{(l)}\|_{\Psi_\rho} &= \max_{l=1,\dots,d_1} \|g_l(X_1) - \mathbb{E}[g_l(X_1)|Z_{-l}]\|_{\Psi_\rho} \\ &\le \max_{l=1,\dots,d_1} \|g_l(X_1)\|_{\Psi_\rho} + \max_{l=1,\dots,d_1} \|\mathbb{E}[g_l(X_1)|Z_{-l}]\|_{\Psi_\rho} \\ &\lesssim C.\end{aligned}$$

Therefore, we are able to bound the q-th moments of the maxima by

$$\begin{aligned}\mathbb{E}\left[\max_{l=1,\dots,d_1} |\nu^{(l)}|^q\right]^{\frac{1}{q}} &= \|\max_{l=1,\dots,d_1} |\nu^{(l)}|\|_{P,q} \\ &\le q!\|\max_{l=1,\dots,d_1} |\nu^{(l)}|\|_{\Psi_1} \\ &\le q! \log^{\frac{1}{\rho}-1}(2) \|\max_{l=1,\dots,d_1} |\nu^{(l)}|\|_{\Psi_1} \\ &\le Cq! \log^{\frac{1}{\rho}-1}(2) \log^{\frac{1}{\rho}}(1 + d_1) \max_{l=1,\dots,d_1} \|\nu^{(l)}|\|_{\Psi_p}\end{aligned}$$

$$\leq C\log^{\frac{1}{\rho}}(d_1),$$

where C does depend on q and ρ but not on n. For $\mathcal{F} := \{\varepsilon\nu^{(l)} : l = 1, \dots, d_1\}$, it holds

$$\begin{aligned}\mathcal{S}_n :&= \mathbb{E}\left[\sup_{l=1,\dots,d_1}\left|\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right]\right|\right]\\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{G}_n(f)\right]\end{aligned}$$

and the envelope $\sup_{f\in\mathcal{F}}|f|$ satisfies

$$\begin{aligned}\|\max_{l=1,\dots,d_1}\varepsilon\nu^{(l)}\|_{P,q} &\leq \|\varepsilon\|_{P,2q}\|\max_{l=1,\dots,d_1}\nu^{(l)}\|_{P,2q}\\ &\leq C\log^{\frac{1}{\rho}}(d_1).\end{aligned}$$

We can apply Lemma P.2 from Belloni et al. (2018) with $|\mathcal{F}| = d_1$ to obtain

$$\mathcal{S}_n \leq C\log^{\frac{1}{2}}(d_1) + C\log^{\frac{1}{2}}(d_1)\left(n^{\frac{2}{q}}\frac{\log^{\frac{2}{\rho}+1}(d_1)}{n}\right)^{1/2} \lesssim \log^{\frac{1}{2}}(d_1),$$

due to A.2$(v)(a)$. Finally, Assumption A.2(i) implies B.2(i). Assumption B.2(ii) holds since for all $l = 1, \dots, d_1$, the map $(\theta_l, \eta_l) \mapsto \psi_l(X, \theta_l, \eta_l)$ is twice continuously Gateaux-differentiable on $\Theta_l \times \mathcal{T}_l$, which directly implies the differentiability of the map $(\theta_l, \eta_l) \mapsto \mathbb{E}[\psi_l(X, \theta_l, \eta_l)]$. Additionally, for every $\eta \in \mathcal{T}_l \setminus \{\eta_{0,l}\}$, we have

$$\begin{aligned}D_{l,0}[\eta,\eta_{0,l}] :&= \partial_t\big\{\mathbb{E}[\psi_l(W,\theta_{0,l},\eta_{0,l}+t(\eta-\eta_{0,l}))]\big\}\big|_{t=0}\\ &= \mathbb{E}\big[\partial_t\big\{\psi_l(W,\theta_{0,l},\eta_{0,l}+t(\eta-\eta_{0,l}))\big\}\big]\big|_{t=0}\\ &= \mathbb{E}\Bigg[\partial_t\bigg\{\Big(Y-\theta_{0,l}g_l(X_1)-\big(\eta_{0,l}^{(1)}+t(\eta^{(1)}-\eta_{0,l}^{(1)})\big)^T Z_{-l}\\ &\quad -\big(\eta_{0,l}^{(3)}(X)+t(\eta^{(3)}(X)-\eta_{0,l}^{(3)}(X))\big)\Big)\\ &\quad \Big(g_l(X_1)-\big(\eta_{0,l}^{(2)}+t(\eta^{(2)}-\eta_{0,l}^{(2)})\big)^T Z_{-l}\\ &\quad -\big(\eta_{0,l}^{(4)}(Z_{-l})+t(\eta^{(4)}(Z_{-l})-\eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\bigg\}\Bigg]\Bigg|_{t=0}\\ &= \mathbb{E}\left[\varepsilon(\eta_{0,l}^{(2)}-\eta^{(2)})^T Z_{-l}\right]+\mathbb{E}\left[\nu^{(l)}(\eta_{0,l}^{(1)}-\eta^{(1)})^T Z_{-l}\right]\\ &\quad +\mathbb{E}\left[\varepsilon\left(\eta_{0,l}^{(4)}(Z_{-l})-\eta^{(4)}(Z_{-l})\right)\right]+\mathbb{E}\left[\nu^{(l)}\left(\eta_{0,l}^{(3)}(X)-\eta^{(3)}(X)\right)\right]\end{aligned}$$

with

$$\mathbb{E}\left[\varepsilon(\eta_{0,l}^{(2)}-\eta^{(2)})^T Z_{-l}\right] = \mathbb{E}\left[((\eta_{0,l}^{(2)}-\eta^{(2)})^T Z_{-l}\mathbb{E}[\varepsilon|X]\right] = 0,$$

$$\mathbb{E}\left[\nu^{(l)}(\eta_{0,l}^{(1)}-\eta^{(1)})^T Z_{-l}\right] = \mathbb{E}\left[(\eta_{0,l}^{(1)}-\eta^{(1)})^T Z_{-l}\mathbb{E}[\nu^{(l)}|Z_{-l}]\right] = 0,$$

$$\mathbb{E}\left[\varepsilon\left(\eta_{0,l}^{(4)}(Z_{-l})-\eta^{(4)}(Z_{-l})\right)\right] = \mathbb{E}\left[\left(\eta_{0,l}^{(4)}(Z_{-l})-\eta^{(4)}(Z_{-l})\right)\mathbb{E}[\varepsilon|X]\right] = 0$$

and

$$\mathbb{E}\left[\nu^{(l)}\left(\eta_{0,l}^{(3)}(X)-\eta^{(3)}(X)\right)\right]=\mathbb{E}\left[\nu^{(l)}\big(b_1(X_1)+b_2(X_{-1})\big)\right]\leq C\delta_n n^{-1/2}$$

due to Assumption A.1 with $\delta_n = o\big(t_1^{-\frac{3}{2}}\log^{-\frac{1}{2}}(A_n)\big)$. Due to the linearity of the score and the moment condition, it holds

$$\mathbb{E}[\psi_l(W,\theta_l,\eta_{0,l})] = J_{0,l}(\theta_l-\theta_{0,l})$$

and due to

$$|J_{0,l}| = \mathbb{E}\left[(\nu^{(l)})^2\right]$$

Assumption B.2(iv) is satisfied.
For all $t\in[0,1)$, $l=1,\dots,d_1$, $\theta_l\in\Theta_l$ and $\eta_l\in\mathcal{T}_l\setminus\{\eta_{0,l}\}$, we have

$$\begin{aligned}
&\mathbb{E}\left[(\psi_l(W,\theta_l,\eta_l)-\psi_l(W,\theta_{0,l},\eta_{0,l}))^2\right]\\
&=\mathbb{E}\left[(\psi_l(W,\theta_l,\eta_l)-\psi_l(W,\theta_{0,l},\eta_l)+\psi_l(W,\theta_{0,l},\eta_l)-\psi_l(W,\theta_{0,l},\eta_{0,l}))^2\right]\\
&\leq C\bigg(\mathbb{E}\left[(\psi_l(W,\theta_l,\eta_l)-\psi_l(W,\theta_{0,l},\eta_l))^2\right]\\
&\qquad\vee\mathbb{E}\left[(\psi_l(W,\theta_{0,l},\eta_l)-\psi_l(W,\theta_{0,l},\eta_{0,l}))^2\right]\bigg)
\end{aligned}$$

with

$$\begin{aligned}
&\mathbb{E}\left[(\psi_l(W,\theta_l,\eta_l)-\psi_l(W,\theta_{0,l},\eta_l))^2\right]\\
&=|\theta_l-\theta_{0,l}|^2\mathbb{E}\left[\left(g_l(X_1)(g_l(X_1)-(\eta_l^{(2)})^TZ_{-l})-\eta_l^{(4)}(Z_{-l})\right)^2\right]\\
&\leq C|\theta_l-\theta_{0,l}|^2\left(\mathbb{E}\left[g_l(X_1)^4\right]\mathbb{E}\left[\left(g_l(X_1)-(\eta_l^{(2)})^TZ_{-l}-\eta_l^{(4)}(Z_{-l})\right)^4\right]\right)^{\frac{1}{2}}\\
&\leq C|\theta_l-\theta_{0,l}|^2
\end{aligned}$$

due to Assumption A.2(ii), (iv) and the definition of $\mathcal{T}_l$. With similar arguments, we obtain

$$\begin{aligned}
&\mathbb{E}\left[(\psi_l(W,\theta_{0,l},\eta_l)-\psi_l(W,\theta_{0,l},\eta_{0,l}))^2\right]\\
&=\mathbb{E}\Bigg[\bigg(\Big(Y-\theta_{0,l}g_l(X_1)-(\eta_l^{(1)})^TZ_{-l}-\eta_l^{(3)}(X)\Big)\Big(g_l(X_1)-(\eta_l^{(2)})^TZ_{-l}-\eta_l^{(4)}(Z_{-l})\Big)\\
&\qquad-\Big(Y-\theta_{0,l}g_l(X_1)-(\eta_{0,l}^{(1)})^TZ_{-l}-\eta_{0,l}^{(3)}(X)\Big)\Big(g_l(X_1)-(\eta_{0,l}^{(2)})^TZ_{-l}-\eta_{0,l}^{(4)}(Z_{-l})\Big)\bigg)^2\Bigg]\\
&=\mathbb{E}\Bigg[\bigg(\Big(Y-\theta_{0,l}g_l(X_1)-(\eta_l^{(1)})^TZ_{-l}-\eta_l^{(3)}(X)\Big)\\
&\qquad\cdot\Big((\eta_{0,l}^{(2)}-\eta_l^{(2)})^TZ_{-l}+\eta_{0,l}^{(4)}(Z_{-l})-\eta_l^{(4)}(Z_{-l})\Big)\\
&\qquad+\Big(g_l(X_1)-(\eta_{0,l}^{(2)})^TZ_{-l}-\eta_{0,l}^{(4)}(Z_{-l})\Big)\\
&\qquad\cdot\Big((\eta_{0,l}^{(1)}-\eta_l^{(1)})^TZ_{-l}+\eta_{0,l}^{(3)}(X)-\eta_l^{(3)}(X)\Big)\bigg)^2\Bigg]\\
&\leq C\bigg(\|\eta_{0,l}^{(2)}-\eta_l^{(2)}\|_2\vee\|\eta_{0,l}^{(1)}-\eta_l^{(1)}\|_2\vee\|\eta_{0,l}^{(3)}(X)\|_{P,2}\vee\|\eta_{0,l}^{(4)}(Z_{-l})\|_{P,2}\bigg)^2
\end{aligned}$$

$$= C\|\eta_{0,l} - \eta_l\|_e^2,$$

where we used the definition of $\mathcal{T}_l$, A.1(iii) and

$$\sup_{\|\xi\|_2=1} \mathbb{E}[(\xi^T Z)^4] \leq C.$$

Therefore, Assumption B.2$(v)(a)$ holds with $\omega = 2$ since it is straightforward to show Assumption B.2(v) for $\eta_l = \eta_{0,l}$. It holds

$$\begin{aligned}
&\left|\partial_t \mathbb{E}\Big[\psi_l(W, \theta_l, \eta_{0,l} + t(\eta_l - \eta_{0,l}))\Big]\right| \\
&= \bigg|\mathbb{E}\bigg[\partial_t\bigg\{\Big(Y - \theta_{0,l} g_l(X_1) - \big(\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)})\big)^T Z_{-l} \\
&\qquad - \big(\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X))\big)\Big) \\
&\qquad \cdot \Big(g_l(X_1) - \big(\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)})\big)^T Z_{-l} \\
&\qquad - \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\bigg\}\bigg]\bigg| \\
&= \bigg|\mathbb{E}\bigg[\Big(Y - \theta_{0,l} g_l(X_1) - (\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)}))^T Z_{-l} \\
&\qquad - (\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X)))\Big) \\
&\qquad \cdot \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)}))^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l}) - \eta_l^{(4)}(Z_{-l})\Big) \\
&\qquad + \Big(g_l(X_1) - (\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)}))^T Z_{-l} \\
&\qquad - \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big) \\
&\qquad \cdot \Big((\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X) - \eta_l^{(3)}(X)\Big)\bigg]\bigg| \\
&= |I_{1,1} + I_{1,2} + I_{1,3} + I_{1,4}|
\end{aligned}$$

with

$$\begin{aligned}
I_{1,1} &= \mathbb{E}\bigg[\Big(Y - \theta_{0,l} g_l(X_1) - (\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)}))^T Z_{-l} \\
&\qquad - (\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X)))\Big)\Big((\eta_{0,l}^{(2)} - \eta_l^{(2)}))^T Z_{-l}\Big)\bigg] \\
&\leq C\|\eta_{0,l}^{(2)} - \eta_l^{(2)}\|_2, \\
I_{1,2} &= \mathbb{E}\bigg[\Big(Y - \theta_{0,l} g_l(X_1) - (\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)}))^T Z_{-l} \\
&\qquad - (\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X)))\Big)\Big(\eta_{0,l}^{(4)}(Z_{-l})\Big)\bigg] \\
&\leq C\|\eta_{0,l}^{(4)}(X)\|_{P,2},
\end{aligned}$$

$$\begin{aligned}
I_{1,3} &= \mathbb{E}\bigg[\Big(g_l(X_1) - (\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)}))^T Z_{-l} \\
&\qquad - \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\Big((\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l}\Big)\bigg] \\
&\leq C\|\eta_{0,l}^{(1)} - \eta_l^{(1)}\|_2,
\end{aligned}$$

$$\begin{aligned}
I_{1,4} = \mathbb{E}\Big[&\Big(g_l(X_1) - (\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)}))^T Z_{-l} \\
&- \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\Big(\eta_{0,l}^{(3)}(X)\Big)\Big] \\
\leq{}& C\|\eta_{0,l}^{(3)}(X)\|_{P,2}.
\end{aligned}$$

This implies Assumption B.2$(v)(b)$ with $B_{1n} = C$. Finally, to obtain Assumption B.2$(v)(c)$ with $B_{2n} = C$, we note that

$$\begin{aligned}
&\partial_t^2 \mathbb{E}\left[\psi_l(W, \theta_{0,l} + t(\theta_l - \theta_{0,l}), \eta_{0,l} + t(\eta_l - \eta_{0,l}))\right] \\
={}& \partial_t \mathbb{E}\Big[\Big(Y - \big(\theta_{0,l} + t(\theta_l - \theta_{0,l})\big)g_l(X_1) - \big(\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)})\big)^T Z_{-l} \\
&- \big(\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X))\big)\Big) \\
&\cdot \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)}))^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l})\Big) \\
&+ \Big(g_l(X_1) - \big(\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)})\big)^T Z_{-l} \\
&- \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big) \\
&\cdot \Big((\theta_{0,l} - \theta_l))g_l(X_1) + (\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X)\Big)\Big] \\
={}& 2\mathbb{E}\Big[\Big((\theta_{0,l} - \theta_l)g_l(X_1) + (\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X)\Big) \\
&\cdot \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)}))^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l})\Big)\Big] \\
\leq{}& C\left(|\theta_{0,l} - \theta_l|^2 \vee \|\eta_{0,l} - \eta_l\|_e^2\right)
\end{aligned}$$

using the same arguments as above.

Assumption B.3

Note that the Assumptions B.3(ii) and (iii) both hold by the construction of $\mathcal{T}_l$ and the Assumptions A.1(ii) and A.2(ii). The main part to verify Assumption B.3 is to show that the estimates of the nuisance function are contained in the nuisance realization set with high probability. We will rely on uniform lasso estimation results stated in Appendix 4.9. Therefore, we have to check the Assumptions C.1(i) to (v). Due to Assumption A.2(iii), it holds

$$\max_{j=1,\dots,d_1+d_2} \|Z_j\|_{\Psi_\rho} \leq C \text{ and } \max_{l=1,\dots,d_1} \|\nu^{(l)}\|_{\Psi_\rho} \leq C,$$

which are the tail conditions in Assumption C.1(i) for the auxiliary regressions. Assumption C.1(ii) is directly implied by Assumption A.2(iv) and

$$\min_{l=1,\dots,d_1} \min_{j\neq l} \mathbb{E}\big[(\nu^{(l)})^2 Z_{-l,j}^2\big] = \min_{l=1,\dots,d_1} \min_{j\neq l} \mathbb{E}\big[Z_{-l,j}^2 \underbrace{\mathbb{E}[(\nu^{(l)})^2|Z_{-l}]}_{=\mathrm{Var}(\nu^{(l)}|Z_{-l})\geq c}\big] \geq c.$$

Additionally, the uniform sparsity condition in Assumption C.1(iii) holds by Assumption A.2(ii) and the growth condition in Assumption C.1(iv) by Assumption A.2$(v)(a)$. Finally, the condition on the approximation error in Assumption C.1(v) holds due to A.1(ii). Therefore,

$$\hat{\eta}_l^{(2)} \in \mathcal{T}_l \quad \text{for all } l = 1, \dots, d_1$$

with probability $1 - o(1)$. To estimate $\eta_{0,l}^{(1)}$, we run a lasso regression of Y on Z. With analogous arguments, it holds

$$\begin{aligned}
&\|\beta_0^{(l)} - \hat{\beta}^{(l)}\|_0 \leq \|\hat{\theta}\|_0 + \|\hat{\beta}\|_0 \leq Cs,\\
&\|\beta_0^{(l)} - \hat{\beta}^{(l)}\|_2 \leq \sqrt{\|\theta - \hat{\theta}\|_2^2 + \|\beta_0 - \hat{\beta}\|_2^2} \leq C\sqrt{\frac{s\log(\bar{d}_n)}{n}},\\
&\|\beta_0^{(l)} - \hat{\beta}^{(l)}\|_1 \leq \|\theta - \hat{\theta}\|_1 + \|\beta_0 - \hat{\beta}\|_1 \leq C\sqrt{\frac{s^2\log(\bar{d}_n)}{n}}
\end{aligned}$$

with probability $1 - o(1)$ using Assumptions A.1(ii), A.2(ii)-(v) and

$$\min_{l=1,\dots,d_1+d_2} \mathbb{E}\big[\varepsilon^2 Z_l^2\big] = \min_{l=1,\dots,d_1+d_2} \mathbb{E}\big[Z_l^2 \underbrace{\mathbb{E}[\varepsilon^2|X]}_{=\mathrm{Var}(\varepsilon|X)\geq c}\big] \geq c.$$

This directly implies that with probability $1 - o(1)$ the nuisance realization set $\mathcal{T}_l$ contains $\hat{\eta}_l^{(1)}$ for all $l = 1, \dots, d_1$.
Combining the results above with $\hat{\eta}^{(3)} \equiv 0$ and $\hat{\eta}^{(4)} \equiv 0$, we obtain Assumption B.3(i). Define

$$\mathcal{F}_1 := \big\{\psi_l(\cdot, \theta_l, \eta_l) : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l\big\}.$$

To bound the complexity of $\mathcal{F}_1$, we exclude the true nuisance function (the true nuisance function is the only element of $\mathcal{T}_l$ with a nonzero approximation error):

$$\mathcal{F}_{1,1} := \big\{\psi_l(\cdot, \theta_l, \eta_l) : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_0^{(l)}\}\big\} \subseteq \mathcal{F}_{1,1}^{(1)}\mathcal{F}_{1,1}^{(2)}$$

with

$$\begin{aligned}
\mathcal{F}_{1,1}^{(1)} &:= \big\{W \mapsto Y - \theta_l g_l(X_1) - (\eta_l^{(1)})^T Z_{-l} : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_0^{(l)}\}\big\}\\
\mathcal{F}_{1,1}^{(2)} &:= \big\{W \mapsto g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_0^{(l)}\}\big\}.
\end{aligned}$$

Note that the envelope $F_{1,1}^{(1)}$ of $\mathcal{F}_{1,1}^{(1)}$ satisfies

$$\begin{aligned}
\|F_{1,1}^{(1)}\|_{P,2q} &\leq \Bigg\| \sup_{l=1,\dots,d_1} \sup_{\theta_l\in\Theta_l, \|\eta_{0,l}^{(1)} - \eta_l^{(l)}\|_1 \leq C\sqrt{s}\tau_n} \Big(|\varepsilon| + |\eta_0^{(3)}(X)|\\
&\qquad + |(\theta_{0,l} - \theta_l)g_l(X_1)| + |(\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l}|\Big)\Bigg\|_{P,2q}\\
&\lesssim \|\varepsilon\|_{P,2q} + \|\eta_0^{(3)}(X)\|_{P,2q} + \|\sup_{l=1,\dots,d_1} g_l(X_1)\|_{P,2q}\\
&\qquad + \sqrt{s}\tau_n \|\sup_{j=1,\dots,d_1+d_2} Z_j\|_{P,2q}\\
&\lesssim C + \log^{\frac{1}{\rho}}(d_1) + \sqrt{s}\tau_n \log^{\frac{1}{\rho}}(d_1 + d_2)\\
&\lesssim \log^{\frac{1}{\rho}}(d_1)
\end{aligned}$$

due to A.1(ii), A.2(v) and analogously

$$\|F_{1,1}^{(2)}\|_{P,2q} \lesssim \log^{\frac{1}{\rho}}(d_1),$$

where we assumed $d_1 \geq 2$ without loss of generality. Next, note that due to Lemma 2.6.15 from Van der

Vaart and Wellner (1996) the set

$$\mathcal{G}_{1,1} := \left\{ Z \mapsto \xi^T Z : \xi \in \mathbb{R}^{d_1+d_2+1}, \|\xi\|_0 \leq Cs, \|\xi\|_2 \leq C \right\}$$

is a union over $\binom{d_1+d_2+1}{Cs}$ VC-subgraph classes $\mathcal{G}_{1,1,k}$ with VC indices less or equal to $Cs+2$. Therefore, $\mathcal{F}^{(1)}_{1,1}$ and $\mathcal{F}^{(2)}_{1,1}$ are unions over $\binom{d_1+d_2+1}{Cs}$ respectively $\binom{d_1+d_2}{Cs}$ VC-subgraph classes, which combined with Theorem 2.6.7 from Van der Vaart and Wellner (1996) implies

$$\sup_Q \log N(\varepsilon \|F^{(1)}_{1,1}\|_{Q,2}, \mathcal{F}^{(1)}_{1,1}, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1+d_2}{\varepsilon}\right)$$

and

$$\sup_Q \log N(\varepsilon \|F^{(2)}_{1,1}\|_{Q,2}, \mathcal{F}^{(2)}_{1,1}, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1+d_2}{\varepsilon}\right).$$

Using basic calculations, we obtain

$$\sup_Q \log N(\varepsilon \|F^{(1)}_{1,1}\mathcal{F}^{(2)}_{1,1}|_{Q,2}, \mathcal{F}_{1,1}, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1+d_2}{\varepsilon}\right),$$

where $F_{1,1} := F^{(1)}_{1,1}\mathcal{F}^{(2)}_{1,1}$ is an envelope for $\mathcal{F}_{1,1}$ with

$$\|F_{1,1}\|_{P,q} \leq \|F^{(1)}_{1,1}\|_{P,2q}\|F^{(2)}_{1,1}\|_{P,2q} \lesssim \log^{\frac{2}{\rho}}(d_1).$$

Define

$$\mathcal{F}_{1,2} := \left\{ \psi_l(\cdot, \theta_l, \eta_{0,l}) : l = 1, \ldots, d_1, \theta_l \in \Theta_l \right\}$$

and, with an analogous argument, we obtain

$$\sup_Q \log N(\varepsilon \|F_{1,2}\|_{Q,2}, \mathcal{F}_{1,2}, \|\cdot\|_{Q,2}) \lesssim \log\left(\frac{d_1}{\varepsilon}\right),$$

where the envelope $F_{1,2}$ of $\mathcal{F}_{1,2}$ obeys

$$\|F_{1,2}\|_{P,q} \lesssim \log^{\frac{2}{\rho}}(d_1).$$

Combining the results above, we obtain

$$\sup_Q \log N(\varepsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1+d_2}{\varepsilon}\right),$$

where the envelope $F_1 := F^{(1)}_{1,1}\mathcal{F}^{(2)}_{1,1} \vee F_{1,2}$ of $\mathcal{F}_1$ satisfies

$$\|F_1\|_{P,q} \lesssim \log^{\frac{2}{\rho}}(d_1).$$

Therefore, Assumption B.3(iv) holds with $\upsilon_n \lesssim s$, $a_n = d_1 \vee d_2$ and $K_n \lesssim \log^{\frac{2}{\rho}}(d_1)$. For all $f \in \mathcal{F}_1$, we have

$$\mathbb{E}[f^2]^{\frac{1}{2}} \lesssim \sup_{\|\xi\|_2=1} \mathbb{E}[(\xi^T Z)^4]^{\frac{1}{2}} \lesssim C$$

and for each $l = 1, \dots, d_1$

$$\begin{aligned}
&\mathbb{E}\left[\psi_l(W,\theta_l,\eta_l)^2\right]^{\frac{1}{2}}\\
&= \mathbb{E}\left[\big(Y-\theta_l g_l(X_1) - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X)\big)^2 \big(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l})\big)^2\right]^{\frac{1}{2}}\\
&= \mathbb{E}\Big[\big(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l})\big)^2\\
&\qquad \cdot \underbrace{\mathbb{E}\left[\big(Y-\theta_l g_l(X_1) - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X)\big)^2 | X\right]}_{\geq Var(\varepsilon|X)\geq c}\Big]^{\frac{1}{2}}\\
&\geq c
\end{aligned}$$

due to Assumption A.2(iv). This implies Assumption B.3(v). Assumption B.3(vi)(a) holds by the definition of τ_n and $\upsilon_n \lesssim s$. To verify the next growth condition, we note

$$\begin{aligned}
&(B_{1n}\tau_n + \mathcal{S}_n \log(n)/\sqrt{n})^{\omega/2}(\upsilon_n \log(a_n))^{1/2} + n^{-1/2+1/q}\upsilon_n K_n \log(a_n)\\
&\lesssim (\tau_n + \log^{\frac{1}{2}}(d_1)\log(n)/\sqrt{n})(s\log(a_n))^{1/2} + n^{-1/2+1/q} s \log^{\frac{2}{\rho}}(d_1)\log(a_n)\\
&\lesssim \left(n^{\frac{2}{q}} \frac{s^2 \log^{2+\frac{4}{\rho}}(\bar{d}_n)}{n}\right)^{\frac{1}{2}}\\
&\lesssim \delta_n
\end{aligned}$$

with $\delta_n = o\big(t_1^{-\frac{3}{2}} \log^{-\frac{1}{2}}(A_n)\big)$ due to Assumption A.2(v)(a) and analogously

$$n^{1/2}B_{1n}^2 B_{2n}^2 \tau_n^2 \lesssim n^{1/2}\tau_n^2 = \sqrt{\frac{s^2\log^2(\bar{d}_n)}{n}} \lesssim \delta_n,$$

since q can be chosen arbitrarily large.

Assumption B.4(i) $-$ (ii)
Define

$$\mathcal{F}_0 := \{\psi_x(\cdot) : x \in I\},$$

where $\psi_x(\cdot) := (g(x)^T \Sigma_n g(x))^{-1/2} g(x)^T J_0^{-1}\psi(\cdot,\theta_0,\eta_0)$. We note that for any $q > 0$ the envelope F_0 of $\mathcal{F}_0$ satisfies

$$\begin{aligned}
\|F_0\|_{P,q} &= \mathbb{E}\left[\sup_{x\in I}\left|(g(x)^T\Sigma_n g(x))^{-1/2} g(x)^T J_0^{-1}\psi(W,\theta_0,\eta_0)\right|^q\right]^{\frac{1}{q}}\\
&\lesssim \mathbb{E}\left[\sup_{x\in I}\left|g(x)^T J_0^{-1}\psi(W,\theta_0,\eta_0)\right|^q\right]^{\frac{1}{q}}\\
&= \mathbb{E}\left[\sup_{x\in I}\left|\sum_{l=1}^{d_1} g_l(x) J_{0,l}^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\right|^q\right]^{\frac{1}{q}}\\
&\lesssim \mathbb{E}\left[\sup_{x\in I}\left|\sum_{l=1}^{d_1} g_l(x)\varepsilon\nu^{(l)}\right|^q\right]^{\frac{1}{q}}\\
&\lesssim t_1 \mathbb{E}\left[\sup_{l=1,\dots,d_1}\left|\varepsilon\nu^{(l)}\right|^q\right]^{\frac{1}{q}}
\end{aligned}$$

$$\lesssim t_1 \log^{\frac{1}{\rho}}(d_1).$$

By using the same argument as above, we directly obtain B.4(ii) with

$$L_n \lesssim t_1^3 \log^{\frac{3}{\rho}}(d_1).$$

Therefore, we can find a larger envelope $\tilde{F}_0$ with

$$\|\tilde{F}_0\|_{P,q} \lesssim t_1^3 \log^{\frac{3}{\rho}}(d_1).$$

To bound the entropy of $\mathcal{F}_0$, we note that

$$\begin{aligned}
&\big\|\psi_x(W) - \psi_{\tilde{x}}(W)\big\|_{P,2} \\
&= \Bigg\|(g(x)^T\Sigma_n g(x))^{-1/2}\sum_{l=1}^{d_1} g_l(x)\mathbb{E}[(\nu^{(l)})^2]^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l}) \\
&\quad - (g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}\sum_{l=1}^{d_1} g_l(\tilde{x})\mathbb{E}[(\nu^{(l)})^2]^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\Bigg\|_{P,2} \\
&\le |(g(x)^T\Sigma_n g(x))^{-1/2} - (g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}| \\
&\quad \cdot \Bigg\|\sum_{l=1}^{d_1} g_l(x)\mathbb{E}[(\nu^{(l)})^2]^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\Bigg\|_{P,2} \\
&\quad + (g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}\Bigg\|\sum_{l=1}^{d_1}\big(g_l(x) - g_l(\tilde{x})\big)\mathbb{E}[(\nu^{(l)})^2]^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\Bigg\|_{P,2} \\
&= |(g(x)^T\Sigma_n g(x))^{-1/2} - (g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}|\Big\|g(x)^T J_0^{-1}\psi(W,\theta_{0,l},\eta_{0,l})\Big\|_{P,2} \\
&\quad + (g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}\Big\|\big(g(x) - g(\tilde{x})\big)^T J_0^{-1}\psi(W,\theta_{0,l},\eta_{0,l})\Big\|_{P,2} \\
&\lesssim |(g(x)^T\Sigma_n g(x))^{-1/2} - (g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}|\sup_{x\in I}\|g(x)\|_2 \\
&\quad + \|g(x) - g(\tilde{x})\|_2
\end{aligned}$$

due to the sub-multiplicativity of the spectral norm and the bounded eigenvalues.

Additionally, it holds

$$\begin{aligned}
&|(g(x)^T\Sigma_n g(x))^{-1/2} - (g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}| \\
&\lesssim \left|\left(\frac{g(\tilde{x})^T\Sigma_n g(\tilde{x})}{g(x)^T\Sigma_n g(x)}\right)^{1/2} - 1\right| \\
&\lesssim |g(\tilde{x})^T\Sigma_n g(\tilde{x}) - g(x)^T\Sigma_n g(x)| \\
&= |(g(x) - g(\tilde{x}))^T\Sigma_n(g(x) + g(\tilde{x}))| \\
&\le |\langle\Sigma_n(g(x) - g(\tilde{x})), (g(x) + g(\tilde{x}))\rangle| \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_x \|g(x)\|_2
\end{aligned}$$

which implies

$$\big\|\psi_x(W) - \psi_{\tilde{x}}(W)\big\|_{P,2} \lesssim \|g(x) - g(\tilde{x})\|_2 \sup_x \|g(x)\|_2^2.$$

Using the same argument as in Theorem 2.7.11 from Van der Vaart and Wellner (1996), we obtain

$$\begin{aligned}
&\sup_Q \log N(\varepsilon\|\tilde{F}_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2})\\
&\lesssim \sup_Q \log N\left(\left(\frac{\varepsilon t_1^3 \log^{\frac{3}{\rho}}(d_1)}{\sup_x \|g(x)\|_2^2}\right)\sup_x \|g(x)\|_2^2, \mathcal{F}_0, \|\cdot\|_{Q,2}\right)\\
&\leq \log N\left(\left(\frac{\varepsilon t_1^3 \log^{\frac{3}{\rho}}(d_1)}{\sup_x \|g(x)\|_2^2}\right), g(I), \|\cdot\|_2\right)\\
&\lesssim t_1 \log\left(\frac{A_n}{\varepsilon}\right).
\end{aligned}$$

Therefore, Assumption B.4(i) is satisfied with $\varrho_n = t_1$.

Assumption B.5

Next, we want to prove that with probability $1 - o(1)$ it holds

$$\sup_{l=1,\dots,d_1} |\hat{J}_l - J_{0,l}| = o(1),$$

where $\hat{J}_l = \mathbb{E}_n[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})]$. It holds

$$\begin{aligned}
|\hat{J}_l - J_{0,l}| \leq{}& |\hat{J}_l - \mathbb{E}[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})]|\\
&+ |\mathbb{E}[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})] + J_{0,l}|
\end{aligned}$$

with

$$\begin{aligned}
&|\mathbb{E}[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})] + J_{0,l}|\\
\leq& |\mathbb{E}[g_l(X_1)(\hat{\eta}_l^{(2)} - \eta_{0,l}^{(2)})^T Z_{-l})]| + |\mathbb{E}[g_l(X_1)\eta_{0,l}^{(4)}(Z_{-l})]|\\
\lesssim& \tau_n.
\end{aligned}$$

Let

$$\begin{aligned}
\tilde{\mathcal{G}}_1 := \Big\{& X \mapsto -g_l(X_1)(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l}) : l = 1, \dots, d_1, \|\eta_l^{(2)}\|_0 \leq Cs,\\
&\|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \leq C\tau_n, \|\eta^{(2)} - \eta_{0,l}^{(2)}\|_1 \leq C\sqrt{s}\tau_n \Big\}.
\end{aligned}$$

The envelope $\tilde{G}_1$ of $\tilde{\mathcal{G}}_1$ satisfies

$$\begin{aligned}
\mathbb{E}[\tilde{G}_1^q]^{\frac{1}{q}} \leq{}& \mathbb{E}\left[\sup_{l=1,\dots,d_1} \sup_{\eta^{(2)}:\|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \leq C\sqrt{s}\tau_n} |g_l(X_1)|^q |(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l})|^q\right]^{\frac{1}{q}}\\
\leq{}& \|\sup_{l=1,\dots,d_1} g_l(X_1)\|_{P,2q}\\
&\cdot \mathbb{E}\left[\sup_{l=1,\dots,d_1} \sup_{\eta^{(2)}:\|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \leq C\sqrt{s}\tau_n} |(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l})|^{2q}\right]^{\frac{1}{2q}}\\
\lesssim{}& \log^{\frac{1}{\rho}}(d_1)\left(\|\sup_{l=1,\dots,d_1} \nu^{(l)}\|_{P,2q} \vee \|\sup_{l=1,\dots,d_1} b_3^{(l)}(Z_{-l})\|_{P,2q}\right.
\end{aligned}$$

$$
\begin{aligned}
&\vee\, \mathbb{E}\left[\sup_{l=1,\dots,d_1}\ \sup_{\eta^{(2)}:\|\eta_l^{(2)}-\eta_{0,l}^{(2)}\|_2\le C\sqrt{s}\tau_n} (\eta_{0,l}^{(2)}-\eta_l^{(2)})^T Z_{-l})^{2q}\right]^{\frac{1}{2q}}\Bigg)\\
&\lesssim \log^{\frac{1}{\rho}}(d_1)\left(\log^{\frac{1}{\rho}}(d_1)\vee\sqrt{s}\tau_n\log^{\frac{1}{\rho}}(d_1+d_2)\right)\\
&\lesssim \log^{\frac{2}{\rho}}(d_1)
\end{aligned}
$$

and, with the same arguments as above, we obtain

$$\sup_Q \log N(\varepsilon\|\tilde{G}_1\|_{Q,2},\tilde{\mathcal{G}}_1,\|\cdot\|_{Q,2})\lesssim s\log\left(\frac{d_1+d_2}{\varepsilon}\right).$$

Therefore, by using Lemma P.2 from Belloni et al. (2018), it holds

$$
\begin{aligned}
\sup_{l=1,\dots,d_1}|\hat{J}_l-J_{0,l}| &\lesssim \sup_{f\in\tilde{\mathcal{G}}_1}|\mathbb{E}_n[f(X)]-\mathbb{E}[f(X)]|+\tau_n\\
&\lesssim K\left(\sqrt{\frac{s\log(\bar{d}_n)}{n}}+n^{\frac{1}{q}}\frac{s\log^{\frac{2}{\rho}}(d_1)\log(\bar{d}_n)}{n}\right)+\tau_n
\end{aligned}
$$

with probability $1-o(1)$. Next, we want to bound the restricted eigenvalues of $\hat{\Sigma}_{\varepsilon\nu}$ with high probability by showing

$$\sup_{\|v\|_2=1,\|v\|_0\le t_1}|v^T(\hat{\Sigma}_{\varepsilon\nu}-\Sigma_{\varepsilon\nu})v|\lesssim u_n \tag{4.11}$$

with

$$u_n\lesssim t_1\left(n^{\frac{1}{q}}\log^{\frac{2}{\rho}}(d_1)\tau_n^2\vee s\tau_n^3\right)^{\frac{1}{2}}$$

for a suitable $\tilde{q}>\bar{q}$. Define $\xi_i:=\varepsilon_i\nu_i$, $\hat{\xi}_i:=\hat{\varepsilon}_i\hat{\nu}_i$ and observe that

$$
\begin{aligned}
&\hat{\Sigma}_{\varepsilon\nu}-\Sigma_{\varepsilon\nu}\\
&=\frac{1}{n}\sum_{i=1}^n\hat{\xi}_i\hat{\xi}_i^T-\mathbb{E}[\xi_i\xi_i^T]\\
&=\frac{1}{n}\sum_{i=1}^n\xi_i\xi_i^T-\mathbb{E}[\xi_i\xi_i^T]\\
&\quad+\frac{1}{n}\sum_{i=1}^n\xi_i\big(\hat{\xi}_i-\xi_i\big)^T+\frac{1}{n}\sum_{i=1}^n\big(\hat{\xi}_i-\xi_i\big)\xi_i^T+\frac{1}{n}\sum_{i=1}^n\big(\hat{\xi}_i-\xi_i\big)\big(\hat{\xi}_i-\xi_i\big)^T.
\end{aligned}
$$

Using the Lemma Q.1 from Belloni et al. (2018), we can bound the first part.
Due to the tail conditions on ε and ν, we obtain

$$
\begin{aligned}
\left(\mathbb{E}\left[\max_{1\le i\le n}\|\varepsilon_i\nu_i\|_\infty^2\right]\right)^{1/2}&\le\left(\mathbb{E}\left[\max_{1\le i\le n}\|\varepsilon_i\|^4\right]\mathbb{E}\left[\max_{1\le i\le n}\|\nu_i\|_\infty^4\right]\right)^{1/4}\\
&\lesssim n^{\frac{2}{q}}\log^{\frac{1}{\rho}}(d_1)
\end{aligned}
$$

for an arbitrary but fixed $q\ge 4$. Then, Lemma Q.1 implies

$$\mathbb{E}\left[\sup_{\|v\|_2=1,\|v\|_0\le t_1}\left|v^T\Big(\frac{1}{n}\sum_{i=1}^n\xi_i\xi_i^T-\mathbb{E}[\xi_i\xi_i^T]\Big)v\right|\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sup_{\|v\|_2=1,\|v\|_0\le t_1}\left|\mathbb{E}_n\Big[\big(v^T\xi_i\big)^2-\mathbb{E}\big[\big(v^T\xi_i\big)^2\big]\Big]\right|\right]\\
&\lesssim \tilde{\delta}_n^2+\tilde{\delta}_n
\end{aligned}
$$

with

$$
\begin{aligned}
\tilde{\delta}_n &\lesssim \Big(n^{\frac{4}{q}}\log^{\frac{2}{\rho}}(d_1)t_1\log^2(t_1)\log(d_1)\log(n)n^{-1}\Big)^{\frac{1}{2}}\\
&\lesssim \left(n^{\frac{5}{q}}\frac{t_1\log^{1+\frac{2}{\rho}}(d_1)}{n}\right)^{\frac{1}{2}}
\end{aligned}
$$

and

$$
\frac{\tilde{\delta}_n^2}{u_n^2}\lesssim\Big(n^{\frac{1}{\tilde{q}}-\frac{5}{q}}t_1 s\Big)^{-1}=o(1)
$$

for $q>5\tilde{q}$. Using Markov's inequality, we directly obtain

$$
\sup_{\|v\|_2=1,\|v\|_0\le t_1}\left|v^T\Big(\frac{1}{n}\sum_{i=1}^n\xi_i\xi_i^T-\mathbb{E}[\xi_i\xi_i^T]\Big)v\right|\lesssim u_n
$$

with probability $1-o(1)$. Note that by applying the results on covariance estimation from Chen et al. (2012) instead would lead to comparable growth rates.

With probability $1-o(1)$, it holds

$$
\sup_{l=1,\dots,d_1}|\hat{\theta}_l-\theta_{0,l}|\lesssim\tau_n
$$

due to Appendix A from Belloni et al. (2018). Define

$$
\begin{aligned}
\tilde{\mathcal{G}}_2^2 := \big\{(\psi_l(\cdot,\theta_l,\eta_l)-\psi_l(\cdot,\theta_{0,l},\eta_{0,l}))^2 :& l=1,\dots,d_1, |\theta_l-\theta_{0,l}|\le C\tau_n\\
&\eta_l\in\mathcal{T}_l\setminus\{\eta_{0,l}\}\big\},
\end{aligned}
$$

with

$$
\sup_Q\log N(\varepsilon\|\tilde{G}_2^2\|_{Q,2},\tilde{\mathcal{G}}_2^2,\|\cdot\|_{Q,2})\lesssim s\log\left(\frac{d_1+d_2}{\varepsilon}\right).
$$

Here, $\tilde{G}_2^2$ is a measurable envelope of $\tilde{\mathcal{G}}_2^2$ with

$$
\tilde{G}_2^2=\sup_{l=1,\dots,d_1}\sup_{\theta_l:|\theta_l-\theta_{0,l}|\le C\tau_n,\eta_l\in\mathcal{T}_l}\big(\psi_l(W,\theta_l,\eta_l)-\psi_l(W,\theta_{0,l},\eta_{0,l})\big)^2
$$

and

$$
\begin{aligned}
&\|\tilde{G}_2^2\|_{P,q}\\
\lesssim&\Big\|\sup_{l,\theta_l,\eta_l^{(2)},\eta_l^{(4)}}\Big((\theta_{0,l}-\theta_l)g_l(X_1)\big(g_l(X_1)-(\eta_l^{(2)})^TZ_{-l}-\eta_l^{(4)}(Z_{-l})\big)\Big)^2\Big\|_{P,q}\\
&+\Big\|\sup_{l,\eta_l}\Big(\big(Y-\theta_{0,l}g_l(X_1)-(\eta_l^{(1)})^TZ_{-l}-\eta_l^{(3)}(X)\big)\\
&\quad\big((\eta_{0,l}^{(2)}-\eta_l^{(2)})^TZ_{-l}+\eta_{0,l}^{(4)}(Z_{-l})-\eta_l^{(4)}(Z_{-l})\big)\Big)^2\Big\|_{P,q}
\end{aligned}
$$

$$
\begin{aligned}
&+\Big\| \sup_{l,\eta_l^{(1)},\eta_l^{(3)}} \Big(\big(g_l(X_1) - (\eta_{0,l}^{(2)})^T Z_{-l} - \eta_{0,l}^{(4)}(Z_{-l})\big) \\
&\quad \big((\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X) - \eta_l^{(3)}(X)\big)\Big)^2 \Big\|_{P,q} \\
&=: T_1 + T_2 + T_3.
\end{aligned}
$$

It holds

$$
\begin{aligned}
T_1 &\lesssim \tau_n^2 \Big\| \sup_{l,\eta_l^{(2)},\eta_l^{(4)}} \Big(g_l(X_1)\big(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} - \eta_l^{(4)}(Z_{-l})\big)\Big)^2 \Big\|_{P,q} \\
&\leq \tau_n^2 \| \sup_l (g_l(X_1))^2 \|_{P,2q} \Big\| \sup_{l,\eta_l^{(2)},\eta_l^{(4)}} \Big(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} - \eta_l^{(4)}(Z_{-l})\Big)^2 \Big\|_{P,2q} \\
&\lesssim \tau_n^2 \log^{\frac{4}{\rho}}(d_1),
\end{aligned}
$$

$$
\begin{aligned}
T_2 &\leq \Big\| \sup_{l,\eta_l^{(1)},\eta_l^{(3)}} \Big(Y - \theta_{0,l} g_l(X_1) - (\eta_l^{(1)})^T Z_{-l} - \eta_l^{(3)}(X)\Big)^2 \Big\|_{P,2q} \\
&\quad \Big\| \sup_{l,\eta_l^{(2)},\eta_l^{(4)}} \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l}) - \eta_l^{(4)}(Z_{-l})\Big)^2 \Big\|_{P,2q} \\
&\lesssim s\tau_n^2 \Big\| \sup_l \|Z_{-l}\|_\infty^2 \Big\|_{P,2q} + \log^{\frac{2}{\rho}}(d_1) \\
&\lesssim s\tau_n^2 \log^{\frac{2}{\rho}}(d_1 + d_2) + \log^{\frac{2}{\rho}}(d_1)
\end{aligned}
$$

and

$$
\begin{aligned}
T_3 &\leq \| \sup_l (\nu^{(l)})^2 \|_{P,2q} \Big\| \sup_{l,\eta_l^{(1)},\eta_l^{(3)}} \Big(\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X) - \eta_l^{(3)}(X)\Big)^2 \Big\|_{P,2q} \\
&\lesssim \log^{\frac{2}{\rho}}(d_1) \Big(s\tau_n^2 \Big\| \sup_l \|Z_{-l}\|_\infty^2 \Big\|_{P,2q} + 1 \Big) \\
&\lesssim \log^{\frac{2}{\rho}}(d_1) \Big(s\tau_n^2 \log^{\frac{2}{\rho}}(d_1 + d_2) + 1 \Big).
\end{aligned}
$$

By using an analogous argument as above, we obtain

$$
\begin{aligned}
\tilde{\sigma} &:= \sup_{f \in \tilde{\mathcal{G}}_2^2} \mathbb{E}\left[f(X)^2\right]^{\frac{1}{2}} \\
&= \sup_{l=1,\dots,d_1} \sup_{\theta_l : |\theta_l - \theta_{0,l}| \leq C\tau_n, \eta_l \in \mathcal{T}_l} \mathbb{E}\Big[(\psi_l(W,\theta_l,\eta_l) - \psi_l(W,\theta_{0,l},\eta_{0,l}))^4 \Big]^{\frac{1}{2}} \\
&\lesssim \frac{s^2 \log(d_1 \vee d_2)}{n}.
\end{aligned}
$$

Again, we can apply Lemma P.2 from Belloni et al. (2018) to obtain

$$
\begin{aligned}
\sup_{f \in \tilde{\mathcal{G}}_2^2} |\mathbb{E}_n[f(X)] - \mathbb{E}[f(X)]| &\leq K \Bigg(\tilde{\sigma} \sqrt{\frac{s \log(\bar{d}_n)}{n}} + n^{\frac{1}{q}} \|\tilde{G}_2^2\|_{P,q} \frac{s \log(\bar{d}_n)}{n} \Bigg) \\
&\lesssim s\tau_n^3 \vee n^{\frac{1}{q}} \log^{\frac{2}{\rho}}(d_1) \tau_n^2
\end{aligned}
$$

with probability $1 - o(1)$. Note that we have already shown Assumption B.2$(v)(a)$ which implies

$$
\sup_{f \in \tilde{\mathcal{G}}_2^2} \mathbb{E}[f(X)] \leq C\left(|\theta_l - \theta_{0,l}|^2 \vee \|\eta_{0,l} - \eta_l\|_e^2 \right)
$$

$$\lesssim \tau_n^2.$$

Combined, this implies

$$\sup_{l=1,\dots,d_1} \mathbb{E}_n\left[\left(\hat{\varepsilon}_i\hat{\nu}_i^{(l)} - \varepsilon_i\nu_i^{(l)}\right)^2\right] \le \sup_{f\in\tilde{\mathcal{G}}_2^2} \mathbb{E}_n[f(X)] \lesssim n^{\frac{1}{q}}\log^{\frac{2}{\rho}}(d_1)\tau_n^2 \vee s\tau_n^3$$

and, with an analogous argument, we obtain

$$\sup_{l=1,\dots,d_1} \mathbb{E}_n\left[\left(\varepsilon_i\nu_i^{(l)}\right)^2\right] \lesssim 1.$$

Therefore, it holds

$$\begin{aligned}
&\sup_{\|v\|_2=1,\|v\|_0\le t_1} |v^T\frac{1}{n}\sum_{i=1}^n \xi_i\big(\hat{\xi}_i-\xi_i\big)^T v| \\
=&\sup_{\|v\|_2=1,\|v\|_0\le t_1} |\mathbb{E}_n\left[v^T\xi_i\big(\hat{\xi}_i-\xi_i\big)^T v\right]| \\
\le&\sup_{\|v\|_2=1,\|v\|_0\le t_1} \left|\left(\mathbb{E}_n\left[\left(v^T\xi_i\right)^2\right]\mathbb{E}_n\left[\left(v^T\big(\hat{\xi}_i-\xi_i\big)\right)^2\right]\right)^{\frac{1}{2}}\right| \\
\lesssim&\sup_{\|v\|_2=1,\|v\|_0\le t_1} \left|\left(\mathbb{E}_n\left[\left(v^T\big(\hat{\xi}_i-\xi_i\big)\right)^2\right]\right)^{\frac{1}{2}}\right| \\
=&\sup_{\|v\|_2=1,\|v\|_0\le t_1} \left(\sum_{k=1}^{d_1}\sum_{l=1}^{d_1} v_k v_l \mathbb{E}_n\left[(\hat{\varepsilon}_i\hat{\nu}_i^{(k)}-\varepsilon_i\nu_i^{(k)})(\hat{\varepsilon}_i\hat{\nu}_i^{(l)}-\varepsilon_i\nu_i^{(l)})\right]\right)^{\frac{1}{2}} \\
\lesssim& t_1 \sup_{l=1,\dots,d_1} \mathbb{E}_n\left[(\hat{\varepsilon}_i\hat{\nu}_i^{(l)}-\varepsilon_i\nu_i^{(l)})^2\right]^{\frac{1}{2}} \\
\lesssim& t_1\left(n^{\frac{1}{q}}\log^{\frac{2}{\rho}}(d_1)\tau_n^2 \vee s\tau_n^3\right)^{\frac{1}{2}}
\end{aligned}$$

and

$$\sup_{\|v\|_2=1,\|v\|_0\le t_1} |v^T\frac{1}{n}\sum_{i=1}^n \big(\hat{\xi}_i-\xi_i\big)\big(\hat{\xi}_i-\xi_i\big)^T v| \lesssim t_1^2\left(n^{\frac{1}{q}}\log^{\frac{2}{\rho}}(d_1)\tau_n^2 \vee s\tau_n^3\right)$$

with probability $1-o(1)$. Combining the steps above, implies (4.11) if $u_n = o(1)$ which is ensured by the growth conditions. Next, note that for every sparse vector $w\in\mathbb{R}^{d_1}$ ($\|w\|_0\le t_1$) there exists a corresponding matrix M_w

$$M_w\in\mathbb{R}^{d_1\times d_1} : (M_w)_{k,l} = \begin{cases} 1 \text{ if } w_k\neq 0 \wedge w_l\neq 0 \\ 0 \text{ else,}\end{cases}$$

such that

$$w^T(\hat{\Sigma}_{\varepsilon\nu}-\Sigma_{\varepsilon\nu})w = w^T\left(M_w\odot(\Sigma_n-\hat{\Sigma}_n)\right)w.$$

Due to (4.11), it holds

$$\sup_{\|w\|_0\le t_1}\sup_{\|v\|_2=1}\left|v^T\left(M_w\odot(\hat{\Sigma}_{\varepsilon\nu}-\Sigma_{\varepsilon\nu})\right)v\right| \le \sup_{\|v\|_2=1,\|v\|_0\le t_1}\left|v^T(\hat{\Sigma}_{\varepsilon\nu}-\Sigma_{\varepsilon\nu})v\right| \lesssim u_n,$$

which implies

$$\sup_{\|w\|_0\leq t_1} \|M_w \odot (\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu})\|_2 \lesssim u_n$$

and

$$\sup_{\|w\|_0\leq t_1} \|M_w \odot \hat{\Sigma}_{\varepsilon\nu}\|_2 \lesssim 1$$

due to Assumption A.2(iv). This can be used to show for $v \in \mathbb{R}^{d_1}$

$$\sup_{\|v\|_2=1,\|v\|_0\leq t_1} |v^T\big(\hat{\Sigma}_n - \Sigma_n\big)v| \lesssim u_n \tag{4.12}$$

with probability $1-o(1)$ which can be interpreted as an upper bound for the sparse eigenvalues of $\hat{\Sigma}_n-\Sigma_n$. It holds

$$\begin{aligned}
\hat{\Sigma}_n - \Sigma_n &= \hat{J}^{-1}\hat{\Sigma}_{\varepsilon\nu}(\hat{J}^{-1})^T - J_0^{-1}\Sigma_{\varepsilon\nu}(J_0^{-1})^T \\
&= \hat{J}^{-1}\hat{\Sigma}_{\varepsilon\nu}\big(\hat{J}^{-1} - J_0^{-1}\big)^T + \big(\hat{J}^{-1} - J_0^{-1}\big)\hat{\Sigma}_{\varepsilon\nu}(J_0^{-1})^T \\
&\quad + J_0^{-1}\big(\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu}\big)\big(J_0^{-1}\big)^T.
\end{aligned}$$

Note that

$$\begin{aligned}
&\sup_{\|v\|_2=1,\|v\|_0\leq t_1} |v^T\hat{J}^{-1}\hat{\Sigma}_{\varepsilon\nu}\big(\hat{J}^{-1} - J_0^{-1}\big)^T v| \\
=& \sup_{\|v\|_2=1,\|v\|_0\leq t_1} |v^T\hat{J}^{-1}\left(M_v \odot \hat{\Sigma}_{\varepsilon\nu}\right)\big(\hat{J}^{-1} - J_0^{-1}\big)^T v| \\
\leq& \left\|\hat{J}^{-1}\right\|_2 \sup_{\|w\|_0\leq t_1} \left\|\left(M_w \odot \hat{\Sigma}_{\varepsilon\nu}\right)\right\|_2 \left\|\big(\hat{J}^{-1} - J_0^{-1}\big)^T\right\|_2 \\
\lesssim& \, n^{\frac{1}{q}}\frac{s\log^{\frac{2}{\rho}}(d_1)\log(\bar{d}_n)}{n} + \tau_n
\end{aligned}$$

due to the sub-multiplicative spectral norm and an analogous argument holds for the second term. The third term can be bounded by

$$\sup_{\|v\|_2=1,\|v\|_0\leq t_1} |v^T J_0^{-1}\big(\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu}\big)\big(J_0^{-1}\big)^T v| \lesssim u_n.$$

This implies (4.12). We finally obtain

$$\begin{aligned}
\sup_{x\in I}\left|\frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}}{(g(x)^T\Sigma_n g(x))^{1/2}} - 1\right| &\lesssim \sup_{x\in I}\left|g(x)^T\big(\hat{\Sigma}_n - \Sigma_n\big)g(x)\right| \\
&\leq \sup_{x\in I}\|g(x)\|_2^2 \sup_{\|v\|_2=1,\|v\|_0\leq t_1} |v^T\big(\hat{\Sigma}_n - \Sigma_n\big)v| \\
&\lesssim \sup_{x\in I}\|g(x)\|_2^2 u_n
\end{aligned}$$

with probability $1-o(1)$ and $\epsilon_n \lesssim \sup_{x\in I}\|g(x)\|_2^2 u_n$ which is the first part of Assumption B.5.

Assumption B.4$(iii) - (iv)$

Define

$$\begin{aligned}\sigma_x &:= (g(x)^T\Sigma_n g(x))^{1/2},\\ \hat{\sigma}_x &:= (g(x)^T\hat{\Sigma}_n g(x))^{1/2}\end{aligned}$$

and

$$\hat{\mathcal{F}}_0 := \{\psi_x(\cdot) - \hat{\psi}_x(\cdot) : x \in I\}$$

with $\hat{\psi}_x(\cdot) := \hat{\sigma}_x^{-1} g(x)^T \hat{J}_0^{-1}\psi(\cdot, \hat{\theta}, \hat{\eta})$. For every x and $\tilde{x}$, it holds

$$\begin{aligned}
&\|\psi_x(W) - \hat{\psi}_x(W) - (\psi_{\tilde{x}}(W) - \hat{\psi}_{\tilde{x}}(W))\|_{\mathbb{P}_n,2}\\
=&\Big\|\sigma_x^{-1} g(x)^T J_0^{-1}\psi(W,\theta_0,\eta_0) - \sigma_{\tilde{x}}^{-1} g(\tilde{x})^T J_0^{-1}\psi(W,\theta_0,\eta_0)\\
&\quad - \big(\hat{\sigma}_x^{-1} g(x)^T \hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta}) - \hat{\sigma}_{\tilde{x}}^{-1} g(\tilde{x})^T \hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta})\big)\Big\|_{\mathbb{P}_n,2}\\
=&\Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x}))J_{0,l}^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\\
&\quad - \sum_{l=1}^{d_1}(\hat{\sigma}_x^{-1} g_l(x) - \hat{\sigma}_{\tilde{x}}^{-1} g_l(\tilde{x}))\hat{J}_l^{-1}\psi_l(W,\hat{\theta}_l,\hat{\eta}_l)\Big\|_{\mathbb{P}_n,2}\\
\le&\Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x}))\Big(J_{0,l}^{-1} - \hat{J}_l^{-1}\Big)\psi_l(W,\theta_{0,l},\eta_{0,l})\Big\|_{\mathbb{P}_n,2}\\
&\quad + \Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x}))\hat{J}_l^{-1}\Big(\psi_l(W,\theta_{0,l},\eta_{0,l}) - \psi_l(W,\hat{\theta}_l,\hat{\eta}_l)\Big)\Big\|_{\mathbb{P}_n,2}\\
&\quad + \Big\|\sum_{l=1}^{d_1}\Big((\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x})) - (\hat{\sigma}_x^{-1} g_l(x) - \hat{\sigma}_{\tilde{x}}^{-1} g_l(\tilde{x}))\Big)\hat{J}_l^{-1}\psi_l(W,\hat{\theta}_l,\hat{\eta}_l)\Big\|_{\mathbb{P}_n,2}\\
=:&I_{4,1} + I_{4,2} + I_{4,3}.
\end{aligned}$$

We obtain

$$\begin{aligned}
I_{4,1} &= \Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x}))\Big(J_{0,l}^{-1} - \hat{J}_l^{-1}\Big)\psi_l(W,\theta_{0,l},\eta_{0,l})\Big\|_{\mathbb{P}_n,2}\\
&\le \sigma_x^{-1}\Big\|(g(x) - g(\tilde{x}))^T\Big(J_0^{-1} - \hat{J}^{-1}\Big)\psi(W,\theta_0,\eta_0)\Big\|_{\mathbb{P}_n,2}\\
&\quad + |\sigma_x^{-1} - \sigma_{\tilde{x}}^{-1}|\Big\|g(\tilde{x})^T\Big(J_0^{-1} - \hat{J}^{-1}\Big)\psi(W,\theta_0,\eta_0)\Big\|_{\mathbb{P}_n,2}\\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{\|v\|_2=1,\|v\|_0\le 2t_1}\Big\|v^T\Big(J_0^{-1} - \hat{J}^{-1}\Big)\psi(W,\theta_0,\eta_0)\Big\|_{\mathbb{P}_n,2}\\
&\quad + \|g(x) - g(\tilde{x})\|_2 \sup_{x\in I}\|g(x)\|_2^2 \sup_{\|v\|_2=1,\|v\|_0\le t_1}\Big\|v^T\Big(J_0^{-1} - \hat{J}^{-1}\Big)\psi(W,\theta_0,\eta_0)\Big\|_{\mathbb{P}_n,2}\\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{x\in I}\|g(x)\|_2^2 u_n,
\end{aligned}$$

where we used that

$$\begin{aligned}
&\sup_{\|v\|_2=1,\|v\|_0\le t_1}\Big\|v^T\Big(J_0^{-1} - \hat{J}^{-1}\Big)\psi(W,\theta_0,\eta_0)\Big\|_{\mathbb{P}_n,2}^2\\
=&\sup_{\|v\|_2=1,\|v\|_0\le t_1}\Big|v^T\Big(J_0^{-1} - \hat{J}^{-1}\Big)\frac{1}{n}\sum_{i=1}^{n}\xi_i\xi_i^T\Big(J_0^{-1} - \hat{J}^{-1}\Big)^T v\Big|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\|J_0^{-1} - \hat{J}^{-1}\right\|_2^2 \sup_{\|v\|_0 \leq t_1} \left\|M_v \odot \left(\frac{1}{n}\sum_{i=1}^n \xi_i \xi_i^T\right)\right\|_2^2 \\
&\lesssim u_n^2.
\end{aligned}
$$

Analogously, we obtain

$$
\begin{aligned}
I_{4,2} &= \Big\| \sum_{l=1}^{d_1} (\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x})) \hat{J}_l^{-1} \Big(\psi_l(W, \theta_{0,l}, \eta_{0,l}) - \psi_l(W, \hat{\theta}_l, \hat{\eta}_l)\Big)\Big\|_{\mathbb{P}_n,2} \\
&\leq \sigma_x^{-1} \Big\| (g(x) - g(\tilde{x}))^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta})\Big)\Big\|_{\mathbb{P}_n,2} \\
&\quad + |\sigma_x^{-1} - \sigma_{\tilde{x}}^{-1}| \Big\| g(\tilde{x})^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta})\Big)\Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{\|v\|_2 = 1, \|v\|_0 \leq 2t_1} \Big\| v^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta})\Big)\Big\|_{\mathbb{P}_n,2} \\
&\quad + \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 \sup_{\|v\|_2 = 1, \|v\|_0 \leq t_1} \Big\| v^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta})\Big)\Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 u_n.
\end{aligned}
$$

It holds

$$
\begin{aligned}
I_{4,3} &= \Big\| \sum_{l=1}^{d_1} \Big((\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x})) - (\hat{\sigma}_x^{-1} g_l(x) - \hat{\sigma}_{\tilde{x}}^{-1} g_l(\tilde{x}))\Big) \hat{J}_l^{-1} \psi_l(W, \hat{\theta}_l, \hat{\eta}_l)\Big\|_{\mathbb{P}_n,2} \\
&\leq \big|\sigma_x^{-1} - \hat{\sigma}_x^{-1}\big| \Big\| (g(x) - g(\tilde{x}))^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta})\Big\|_{\mathbb{P}_n,2} \\
&\quad + \big|(\sigma_x^{-1} - \hat{\sigma}_x^{-1}) - (\sigma_{\tilde{x}}^{-1} - \hat{\sigma}_{\tilde{x}}^{-1})\big| \Big\| g(\tilde{x})^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta})\Big\|_{\mathbb{P}_n,2}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\big|(\sigma_x^{-1} - \hat{\sigma}_x^{-1}) - (\sigma_{\tilde{x}}^{-1} - \hat{\sigma}_{\tilde{x}}^{-1})\big| \\
&= \Big| \frac{1}{\sigma_x \sigma_{\tilde{x}}} (\sigma_{\tilde{x}} - \sigma_x) - \frac{1}{\hat{\sigma}_x \hat{\sigma}_{\tilde{x}}} (\hat{\sigma}_{\tilde{x}} - \hat{\sigma}_x)\Big| \\
&= \frac{1}{\hat{\sigma}_x \hat{\sigma}_{\tilde{x}}} \Big| \frac{\hat{\sigma}_x \hat{\sigma}_{\tilde{x}}}{\sigma_x \sigma_{\tilde{x}}} (\sigma_{\tilde{x}} - \sigma_x) - (\hat{\sigma}_{\tilde{x}} - \hat{\sigma}_x)\Big| \\
&\lesssim \big|(\sigma_{\tilde{x}} - \sigma_x) - (\hat{\sigma}_{\tilde{x}} - \hat{\sigma}_x)\big| + \Big|\frac{\hat{\sigma}_x \hat{\sigma}_{\tilde{x}}}{\sigma_x \sigma_{\tilde{x}}} - 1\Big| \big|\sigma_{\tilde{x}} - \sigma_x\big|
\end{aligned}
$$

with

$$
\begin{aligned}
\Big|\frac{\hat{\sigma}_x \hat{\sigma}_{\tilde{x}}}{\sigma_x \sigma_{\tilde{x}}} - 1\Big| \big|\sigma_{\tilde{x}} - \sigma_x\big| &\leq \Big(\Big|\frac{\hat{\sigma}_x}{\sigma_x} - 1\Big| \frac{\hat{\sigma}_{\tilde{x}}}{\sigma_{\tilde{x}}} + \Big|\frac{\hat{\sigma}_{\tilde{x}}}{\sigma_{\tilde{x}}} - 1\Big|\Big) |\sigma_{\tilde{x}} - \sigma_x| \\
&\lesssim \epsilon_n \frac{1}{\sigma_x} \big|\sigma_{\tilde{x}}^2 - \sigma_x^2\big| \\
&\lesssim \epsilon_n \|g(x) - g(\tilde{x})\|_2 \sup_x \|g(x)\|_2
\end{aligned}
$$

uniformly over $x \in I$ with probability $1 - o(1)$ and

$$
\begin{aligned}
&\big|(\sigma_{\tilde{x}} - \sigma_x) - (\hat{\sigma}_{\tilde{x}} - \hat{\sigma}_x)\big| \\
&\leq \frac{1}{(\hat{\sigma}_{\tilde{x}} + \hat{\sigma}_x)} \big|(\sigma_{\tilde{x}}^2 - \sigma_x^2) - (\hat{\sigma}_{\tilde{x}}^2 - \hat{\sigma}_x^2)\big| + \Big|\Big(\frac{1}{(\sigma_{\tilde{x}} + \sigma_x)} - \frac{1}{(\hat{\sigma}_{\tilde{x}} + \hat{\sigma}_x)}\Big)(\sigma_{\tilde{x}}^2 - \sigma_x^2)\Big| \\
&\lesssim \big|(\sigma_{\tilde{x}}^2 - \sigma_x^2) - (\hat{\sigma}_{\tilde{x}}^2 - \hat{\sigma}_x^2)\big| + \Big|\frac{(\hat{\sigma}_{\tilde{x}} + \hat{\sigma}_x)}{(\sigma_{\tilde{x}} + \sigma_x)} - 1\Big| \big|\sigma_{\tilde{x}}^2 - \sigma_x^2\big|.
\end{aligned}
$$

Using an analogous argument as in the verification of Assumption B.5, we obtain

$$\begin{aligned}|(\sigma_x^2-\hat{\sigma}_x^2)-(\sigma_{\tilde{x}}^2-\hat{\sigma}_{\tilde{x}}^2)| &= |(g(x)-g(\tilde{x}))^T(\Sigma_n-\hat{\Sigma}_n)(g(x)+g(\tilde{x}))|\\ &\le \|(\Sigma_n-\hat{\Sigma}_n)(g(x)-g(\tilde{x}))\|_2\sup_{x\in I}\|g(x)\|_2\\ &\lesssim \|g(x)-g(\tilde{x})\|_2 u_n\sup_{x\in I}\|g(x)\|_2\end{aligned}$$

with probability $1-o(1)$ where the last inequality holds due the order of the sparse eigenvalues in (4.12). Additionally,

$$\begin{aligned}\Big|\frac{(\hat{\sigma}_{\tilde{x}}+\hat{\sigma}_x)}{(\sigma_{\tilde{x}}+\sigma_x)}-1\Big|\big|\sigma_{\tilde{x}}^2-\sigma_x^2\big| &\le \sup_{x\in I}\Big|\frac{\hat{\sigma}_x}{\sigma_x}-1\Big|\big|\sigma_{\tilde{x}}^2-\sigma_x^2\big|\\ &\lesssim \epsilon_n\|g(x)-g(\tilde{x})\|_2\sup_{x\in I}\|g(x)\|_2\end{aligned}$$

with probability $1-o(1)$. Therefore, we obtain

$$\begin{aligned}I_{4,3} &\lesssim \epsilon_n\|g(x)-g(\tilde{x})\|_2\sup_{\|v\|_2=1,\|v\|_0\le 2t_1}\Big\|v^T\hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta})\Big\|_{\mathbb{P}_n,2}\\ &\quad+(\epsilon_n\vee u_n)\|g(x)-g(\tilde{x})\|_2\sup_{x\in I}\|g(x)\|_2^2\sup_{\|v\|_2=1,\|v\|_0\le t_1}\Big\|v^T\hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta})\Big\|_{\mathbb{P}_n,2}\\ &\lesssim \|g(x)-g(\tilde{x})\|_2\epsilon_n\sup_{x\in I}\|g(x)\|_2^2.\end{aligned}$$

Combining the steps above, we obtain

$$\|\psi_x(W)-\hat{\psi}_x(W)-(\psi_{\tilde{x}}(W)-\hat{\psi}_{\tilde{x}}(W))\|_{\mathbb{P}_n,2}\le\|g(x)-g(\tilde{x})\|_2\|\hat{F}_0\|_{\mathbb{P}_n,2}$$

with

$$\|\hat{F}_0\|_{\mathbb{P}_n,2}\lesssim\epsilon_n\sup_{x\in I}\|g(x)\|_2^2=o(1)$$

due to the growth condition in Assumption A.2$(v)(b)$ as shown below. Using the same argument as Theorem 2.7.11 from Van der Vaart and Wellner (1996), we obtain with probability $1-o(1)$

$$\begin{aligned}\log N(\varepsilon,\hat{\mathcal{F}}_0,\|\cdot\|_{\mathbb{P}_n,2}) &\le \log N(\varepsilon\|\hat{F}_0\|_{\mathbb{P}_n,2},\hat{\mathcal{F}}_0,\|\cdot\|_{\mathbb{P}_n,2})\\ &\le \log N(\varepsilon,g(I),\|\cdot\|_2)\\ &\le \bar{\varrho}_n\log\left(\frac{\bar{A}_n}{\varepsilon}\right)\end{aligned}$$

with $\bar{\varrho}_n=t_1$ and $\bar{A}_n\lesssim A_n$. Additionally, it holds

$$\begin{aligned}&\|\psi_x(W)-\hat{\psi}_x(W)\|_{\mathbb{P}_n,2}\\ &=\Big\|\sigma_x^{-1}g(x)^TJ_0^{-1}\psi(W,\theta_0,\eta_0)-\hat{\sigma}_x^{-1}g(x)^T\hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta})\Big\|_{\mathbb{P}_n,2}\\ &\le\sigma_x^{-1}\Big\|g(x)^T\Big(J_0^{-1}-\hat{J}^{-1}\Big)\psi(W,\theta_0,\eta_0)\Big\|_{\mathbb{P}_n,2}\\ &\quad+\sigma_x^{-1}\Big\|g(x)^T\hat{J}^{-1}\Big(\psi(W,\theta_0,\eta_0)-\psi(W,\hat{\theta},\hat{\eta})\Big)\Big\|_{\mathbb{P}_n,2}\\ &\quad+|\sigma_x^{-1}-\hat{\sigma}_x^{-1}|\Big\|g(x)^T\hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta})\Big\|_{\mathbb{P}_n,2}\\ &\lesssim\sup_{x\in I}\|g(x)\|_2(u_n\vee\epsilon_n)\end{aligned}$$

$$\lesssim \sup_{x\in I}\|g(x)\|_2\epsilon_n$$

with an analogous argument as above. Therefore, B.4(iii) holds with

$$\bar{\delta}_n \lesssim \sup_{x\in I}\|g(x)\|_2\epsilon_n.$$

To complete the proof, we verify all growth conditions from Assumptions B.4 and B.5. As shown in the verification of B.3(vi), it holds

$$t_1^2\delta_n^2\varrho_n\log(A_n)=\delta_n^2t_1^3\log(A_n)=o(1).$$

Additionally,

$$n^{-\frac{1}{7}}L_n^{\frac{2}{7}}\varrho_n\log(A_n)=\frac{t_1^{\frac{13}{7}}\log^{\frac{6}{7\rho}}(d_1)\log(A_n)}{n^{\frac{1}{7}}}=o(1)$$

and

$$n^{\frac{2}{3q}-\frac{1}{3}}L_n^{\frac{2}{3}}\varrho_n\log(A_n)=n^{\frac{2}{3q}}\frac{t_1^3\log^{\frac{2}{\rho}}(d_1)\log(A_n)}{n^{\frac{1}{3}}}=o(1)$$

for q large enough due to growth condition in Assumption A.2$(v)(c)$. Note that

$$\varepsilon_n\varrho_n\log(A_n)=\varepsilon_nt_1\log(A_n)\lesssim\bar{\delta}_nt_1\log(A_n).$$

Hence, we need to show that

$$\bar{\delta}_n^2\bar{\varrho}_n\varrho_n\log(\bar{A}_n)\log(A_n)=\bar{\delta}_n^2t_1^2\log^2(A_n)=o(1).$$

It holds

$$\begin{aligned}\bar{\delta}_n^2t_1^2\log^2(A_n)&\lesssim u_n^2\sup_{x\in I}\|g(x)\|_2^6t_1^2\log^2(A_n)\\&\lesssim\left(n^{\frac{1}{q}}\log^{\frac{2}{\rho}}(d_1)\tau_n^2\vee s\tau_n^3\right)\sup_{x\in I}\|g(x)\|_2^6t_1^4\log^2(A_n)\\&=o(1)\end{aligned}$$

due to Assumption A.2$(v)(b)$.

□

4.8 Uniformly valid confidence bands

As in Belloni et al. (2018), we consider the problem of estimating the set of parameters $\theta_{0,l}$ for $l=1,\dots,d_1$ in the moment condition model,

$$\mathbb{E}[\psi_l(W,\theta_{0,l},\eta_{0,l})]=0,\qquad l=1,\dots,d_1,\tag{4.13}$$

where W is a random variable, ψ_l a known score function, $\theta_{0,l}\in\Theta_l$ a scalar of interest, and $\eta_{0,l}\in T_l$ a high-dimensional nuisance parameter where T_l is a convex set in a normed space equipped with a norm $\|\cdot\|_e$. Let $\mathcal{T}_l$ be some subset of T_l, which contains the nuisance estimate $\hat{\eta}_l$ with high probability. Belloni et al. (2018) provide an appropriate estimator $\hat{\theta}_l$ and are able to construct simultaneous confidence bands for $(\theta_{0,l})_{l=1,\dots,d_1}$ where d_1 may increase with sample size n. In this section, we are particularly interested in the linear functional

$$G(x) = \sum_{l=1}^{d_1} \theta_{0,l} g_l(x),$$

where $(g_l)_{l=1,\dots,d_1}$ is a given set of functions with

$$g_l : I \subseteq \mathbb{R} \to \mathbb{R}, \qquad l = 1, \dots, d_1.$$

We assume that the score functions ψ_l are constructed to satisfy the near-orthogonality condition, namely

$$D_{l,0}[\eta, \eta_{0,l}] := \partial_t \big\{ \mathbb{E}[\psi_l(W, \theta_{0,l}, \eta_{0,l} + t(\eta - \eta_{0,l}))] \big\} \big|_{t=0} \lesssim \delta_n n^{-1/2}, \tag{4.14}$$

where ∂_t denotes the derivative with respect to t and $(\delta_n)_{n\geq 1}$ a sequence of positive constants converging to zero. We aim to construct uniform valid confidence bands for the target function $G(x)$, namely

$$P(\hat{l}(x) \leq G(x) \leq \hat{u}(x), \forall x \in I) \to 1 - \alpha.$$

Let $\hat{\eta}_l = \left(\hat{\eta}_l^{(1)}, \hat{\eta}_l^{(2)}\right)$ be an estimator of the nuisance function. The estimator $\hat{\theta}_0$ of the target parameter

$$\theta_0 = (\theta_{0,1}, \dots, \theta_{0,d_1})^T$$

is defined as the solution of

$$\sup_{l=1,\dots,d_1} \left\{ \left| \mathbb{E}_n \left[\psi_l\big(W, \hat{\theta}_l, \hat{\eta}_l\big) \right] \right| - \inf_{\theta \in \Theta_l} \left| \mathbb{E}_n \left[\psi_l\big(W, \theta, \hat{\eta}_l\big) \right] \right| \right\} \leq \epsilon_n, \tag{4.15}$$

where $\epsilon_n = o\left(\delta_n n^{-1/2}\right)$ is the numerical tolerance and $(\delta_n)_{n\geq 1}$ a sequence of positive constants converging to zero. Let

$$g(x) = (g_1(x), \dots, g_{d_1}(x))^T \in \mathbb{R}^{d_1 \times 1}$$

and

$$\psi(W, \theta, \eta) = (\psi_1(W, \theta, \eta), \dots, \psi_{d_1}(W, \theta, \eta))^T \in \mathbb{R}^{d_1 \times 1}.$$

Define the Jacobian matrix

$$J_0 := \frac{\partial}{\partial \theta} \mathbb{E}[\psi(W, \theta, \eta_0)] \bigg|_{\theta = \theta_0} = \operatorname{diag}(J_{0,1}, \dots, J_{0,d_1}) \in \mathbb{R}^{d_1 \times d_1}$$

and the approximate covariance matrix

$$\Sigma_n : = J_0^{-1} \mathbb{E}\big[\psi(W, \theta_0, \eta_0) \psi(W, \theta_0, \eta_0)^T\big] (J_0^{-1})^T \in \mathbb{R}^{d_1 \times d_1}.$$

Additionally, define

$$\mathcal{S}_n := \mathbb{E}\left[\sup_{l=1,\dots,d_1} \left| \sqrt{n} \mathbb{E}_n \left[\psi_l(W, \theta_{0,l}, \eta_{0,l}) \right] \right| \right]$$

and

$$t_1 := \sup_{x\in I} \|g(x)\|_0.$$

The definition of t_1 is helpful if the functions g_l, $l = 1,\dots,d_1$ are local in the sense that for any point x in I there are at most $t_1 \ll d_1$ non-zero functions. We state the conditions needed for the uniformly valid confidence bands.

Assumption B. 1. *It holds*

(i) $\inf_{x\in I} \|g(x)\|_2^2 \geq c > 0$

(ii) $\sup_{x\in I} \sup_{l=1,\dots,d_1} |g_l(x)| \leq C < \infty$

(iii) The eigenvalues from Σ_n are uniformly bounded from above and away from zero.

Since the proof of our main result in this section relies on the techniques in Belloni et al. (2018), we try formulate the following conditions as similar as possible to make the use of their methodology transparent.

Assumption B. 2. *For all $n \geq n_0, P \in \mathcal{P}_n$ and $l \in \{1,\dots,d_1\}$, the following conditions hold:*

(i) The true parameter value $\theta_{0,l}$ obeys (4.13), and Θ_l contains a ball of radius $C_0 n^{-1/2}\mathcal{S}_n \log(n)$ centered at $\theta_{0,l}$.

(ii) The map $(\theta_l, \eta_l) \mapsto \mathbb{E}[\psi_l(W,\theta_l,\eta_l)]$ is twice continuously Gateaux-differentiable on $\Theta_l \times \mathcal{T}_l$.

(iii) The score function ψ_l obeys the near orthogonality condition (4.14) for the set $\mathcal{T}_l \subset T_l$.

(iv) For all $\theta_l \in \Theta_l$, $|\mathbb{E}[\psi_l(W,\theta_l,\eta_{0,l})]| \geq 2^{-1}|J_{0,l}(\theta_l-\theta_{0,l})| \wedge c_0$, where $J_{0,l}$ satisfies $c_0 \leq |J_{0,l}| \leq C_0$.

(v) For all $r \in [0,1)$, $\theta_l \in \Theta_l$ and $\eta_l \in \mathcal{T}_l$

(a) $\mathbb{E}[(\psi_l(W,\theta_l,\eta_l) - \psi_l(W,\theta_{0,l},\eta_{0,l}))^2] \leq C_0(|\theta_l-\theta_{0,l}| \vee \|\eta_l-\eta_{0,l}\|_e)^\omega$

(b) $|\partial_r \mathbb{E}[\psi_l(W,\theta_l,\eta_{0,l}+r(\eta_l-\eta_{0,l}))]| \leq B_{1n}\|\eta_l-\eta_{0,l}\|_e$

(c) $|\partial_r^2 \mathbb{E}[\psi_l(W,\theta_{0,l}+r(\theta_l-\theta_{0,l}),\eta_{0,l}+r(\eta_l-\eta_{0,l}))]| \leq B_{2n}(|\theta_l-\theta_{0,l}|^2 \vee \|\eta_l-\eta_{0,l}\|_e^2)$.

Note that the notation $\mathbb{E}$ abbreviates $\mathbb{E}_P$. For a detailed discussion about the ideas and intuitions of these and the following assumptions, see Belloni et al. (2018).

Let $(\Delta_n)_{n\geq 1}$ and $(\tau_n)_{n\geq 1}$ be some sequences of positive constants converging to zero. Also, let $(a_n)_{n\geq 1}$, $(\upsilon_n)_{n\geq 1}$, and $(K_n)_{n\geq 1}$ be some sequences of positive constants, possibly growing to infinity where $a_n \geq n \vee K_n$ and $\upsilon \geq 1$ for all $n \geq 1$. Finally, let $q \geq 2$ be some constant.

Assumption B. 3. *For all $n \geq n_0$ and $P \in \mathcal{P}_n$, the following conditions hold:*

(i) With probability at least $1-\Delta_n$, we have $\hat{\eta}_l \in \mathcal{T}_l$ for all $l = 1,\dots,d_1$.

(ii) For all $l = 1,\dots,d_1$ and $\eta_l \in \mathcal{T}_l$, it holds $\|\eta_l-\eta_{0,l}\|_e \leq \tau_n$.

(iii) For all $l = 1,\dots,d_1$, we have $\eta_{0,l} \in \mathcal{T}_l$.

(iv) The function class $\mathcal{F}_1 = \{\psi_l(\cdot,\theta_l,\eta_l) : l = 1,\dots,d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l\}$ is suitably measurable and its uniform entropy numbers obey

$$\sup_Q \log N(\epsilon\|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \leq \upsilon_n \log(a_n/\epsilon), \quad \textit{for all } 0 < \epsilon \leq 1,$$

where F_1 is a measurable envelope for $\mathcal{F}_1$ that satisfies $\|F_1\|_{P,q} \leq K_n$.

(v) For all $f \in \mathcal{F}_1$*, we have* $c_0 \leq \|f\|_{P,2} \leq C_0$.

(vi) The complexity characteristics a_n *and* υ_n *satisfy*

  (a) $(\upsilon_n \log(a_n)/n)^{1/2} \leq C_0\tau_n$,

  (b) $(B_{1n}\tau_n + \mathcal{S}_n \log(n)/\sqrt{n})^{\omega/2}(\upsilon_n \log(a_n))^{1/2} + n^{-1/2+1/q}\upsilon_n K_n \log(a_n) \leq C_0\delta_n$,

  (c) $n^{1/2}B_{1n}^2 B_{2n}^2 \tau_n^2 \leq C_0\delta_n$.

Whereas the Assumptions B.2 and B.3 are identical to the Assumptions 2.1 and 2.2 from Belloni et al. (2018), the analogs to their Assumptions 2.3 and 2.4 need modifications to fit our setting constructing a uniformly valid confidence band for the linear functional $G(x)$. In this context, define

$$\psi_x(\cdot) := (g(x)^T\Sigma_n g(x))^{-1/2} g(x)^T J_0^{-1}\psi(\cdot, \theta_0, \eta_0)$$

and the corresponding plug-in estimator

$$\hat{\psi}_x(\cdot) := (g(x)^T\hat{\Sigma}_n g(x))^{-1/2} g(x)^T \hat{J}_0^{-1}\psi(\cdot, \hat{\theta}_0, \hat{\eta}_0).$$

Let $(\bar{\delta}_n)_{n\geq 1}$ be a sequence of positive constants converging to zero. Also, let $(\varrho_n)_{n\geq 1}$, $(\bar{\varrho}_n)_{n\geq 1}$, $(A_n)_{n\geq 1}$, $(\bar{A}_n)_{n\geq 1}$, and $(L_n)_{n\geq 1}$ be some sequences of positive constants, possibly growing to infinity where $\varrho \geq 1$, $A_n \geq n$, and $\bar{A}_n \geq n$ for all $n \geq 1$. In addition, assume that $q > 4$.

Assumption B. 4. *For all* $n \geq n_0$ *and* $P \in \mathcal{P}_n$*, the following conditions hold:*

(i) The function class $\mathcal{F}_0 = \{\psi_x(\cdot) : x \in I\}$ *is suitably measurable and its uniform entropy numbers obey*

$$\sup_Q \log N(\varepsilon\|F_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}) \leq \varrho_n \log(A_n/\varepsilon), \quad \textit{for all } 0 < \epsilon \leq 1,$$

where F_0 *is a measurable envelope for* $\mathcal{F}_0$ *that satisfies* $\|F_0\|_{P,q} \leq L_n$.

(ii) For all $f \in \mathcal{F}_0$ *and* $k = 3, 4$*, we have* $\mathbb{E}[|f(W)|^k] \leq C_0 L_n^{k-2}$.

(iii) The function class $\hat{\mathcal{F}}_0 = \{\psi_x(\cdot) - \hat{\psi}_x(\cdot) : x \in I\}$ *satisfies with probability* $1 - \Delta_n$*:*

$$\log N(\varepsilon, \hat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_n,2}) \leq \bar{\varrho}_n \log(\bar{A}_n/\varepsilon), \quad \textit{for all } 0 < \epsilon \leq 1,$$

and $\|f\|_{\mathbb{P}_n,2} \leq \bar{\delta}_n$ *for all* $f \in \hat{\mathcal{F}}_0$.

(iv) $t_1^2\delta_n^2\varrho_n \log(A_n) = o(1)$*,* $L_n^{2/7}\varrho_n \log(A_n) = o(n^{1/7})$ *and* $L_n^{2/3}\varrho_n \log(A_n) = o(n^{1/3-2/(3q)})$.

Additionally, we need to be able to estimate the variance of the linear functional sufficiently well. Let $\hat{\Sigma}_n$ be an estimator of Σ_n.

Assumption B. 5. *For all* $n \geq n_0$ *and* $P \in \mathcal{P}_n$*, it holds*

$$P\left(\sup_{x\in I}\left|\frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}}{(g(x)^T\Sigma_n g(x))^{1/2}} - 1\right| > \varepsilon_n\right) \leq \Delta_n,$$

where $\varepsilon_n\varrho_n \log(A_n) = o(1)$ *and* $\bar{\delta}_n^2\bar{\varrho}_n\varrho_n \log(\bar{A}_n)\log(A_n) = o(1)$.

As in Chernozhukov et al. (2013a), we employ the Gaussian multiplier bootstrap method to estimate the relevant quantiles. Let

$$\hat{\mathcal{G}} = \left(\hat{\mathcal{G}}_x\right)_{x\in I} = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\hat{\psi}_x(W_i)\right)_{x\in I},$$

where $(\xi_i)_{i=1}^n$ are independent standard normal random variables (especially independent from $(W_i)_{i=1}^n$). Define the multiplier bootstrap critical value c_α as the $(1-\alpha)$ quantile of the conditional distribution of $\sup_{x\in I}|\hat{\mathcal{G}}_x|$ given $(W_i)_{i=1}^n$.

Theorem 4. *Define*

$$\hat{u}(x) := \hat{G}(x) + \frac{(g(x)'\hat{\Sigma}_n g(x))^{1/2}c_\alpha}{\sqrt{n}}$$
$$\hat{l}(x) := \hat{G}(x) - \frac{(g(x)'\hat{\Sigma}_n g(x))^{1/2}c_\alpha}{\sqrt{n}}$$

with $\hat{G}(x) = g(x)^T\hat{\theta}_0$. Under the Assumptions B.1 - B.5, it holds

$$P\left(\hat{l}(x) \le G(x) \le \hat{u}(x), \forall x \in I\right) \to 1-\alpha$$

uniformly over $P \in \mathcal{P}_n$.

Proof. Since Theorem 2.1 in Belloni et al. (2018) is not directly applicable to our problem, we have to modify the proof to obtain a uniform Bahadur representation. We want to prove that

$$\sup_{x\in I}\left|\sqrt{n}(g(x)^T\Sigma_n g(x))^{-1/2}g(x)^T\left(\hat{\theta}-\theta_0\right)\right| = \sup_{x\in I}\left|\mathbb{G}_n(\psi_x)\right| + O_P(t_1\delta_n). \tag{4.16}$$

Assumptions B.2 and B.3 contain Assumptions 2.1 and 2.2 from Belloni et al. (2018) which enables us to use parts of their results. Therefore, it holds

$$\sup_{l=1,\dots,d_1}\left|J_{0,l}^{-1}\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right] + \sqrt{n}\left(\hat{\theta}_l-\theta_{0,l}\right)\right| = O_P(\delta_n).$$

Using Assumption B.1, this implies

$$\begin{aligned}
&\sup_{x\in I}\left|\sqrt{n}\mathbb{E}_n\left[g(x)^TJ_0^{-1}\psi(W,\theta_0,\eta_0)\right] + \sqrt{n}g(x)^T\left(\hat{\theta}-\theta_0\right)\right| \\
=\ &\sup_{x\in I}\left|\sum_{j=1}^{d_1} g_l(x)\left(J_{0,l}^{-1}\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right] + \sqrt{n}(\hat{\theta}_l-\theta_{0,l})\right)\right| \\
\le\ & t_1\underbrace{\sup_{x\in I}\sup_{l=1,\dots,d_1}|g_l(x)|}_{\le C}\sup_{l=1,\dots,d_1}\left|J_{0,l}^{-1}\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right] + \sqrt{n}\left(\hat{\theta}_l-\theta_{0,l}\right)\right| \\
=\ & O_p(t_1\delta_n).
\end{aligned}$$

Since the minimal eigenvalue of Σ_n is uniformly bounded away from zero, it follows that $g(x)^T\Sigma_n g(x)$ is uniformly bounded away from zero as long as $\|g(x)\|_2^2$ is uniformly bounded away from zero due to Assumption B.1. This implies (4.16).

Due to Assumption B.5, it holds

$$P\left(\sup_{x\in I}\left|\frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}}{(g(x)^T\Sigma_n g(x))^{1/2}}-1\right|>\varepsilon_n\right)\leq\Delta_n,$$

with $\Delta_n=o(1)$, which is an analogous version of the Assumption 2.4 from Belloni et al. (2018). Therefore, given the Assumptions B.2 - B.5, the proofs of Corollary 2.1 and 2.2 from Belloni et al. (2018) can be applied implying the stated theorem. □

4.9 Uniform nuisance function estimation

To establish uniform estimation properties of the nuisance function, we rely on uniform estimation results from Klaassen et al. (2018). Consider the following linear regression model

$$Y_r=\sum_{j=1}^{p}\beta_{r,j}X_{r,j}+a_r(X_r)+\varepsilon_r=\beta_r X_r+a_r(X_r)+\varepsilon_r$$

with centered regressors and $a_r(X_r)$ accounts for an approximation error. The errors ε_r are assumed to satisfy $\mathbb{E}[\varepsilon_r|X_r]=0$ for each $r=1,\dots,d$.
The true parameter obeys

$$\beta_r\in\arg\min_{\beta}\mathbb{E}[(Y_r-\beta X_r-a_r(X_r))^2].$$

We show that the lasso and post-lasso lasso estimators have sufficiently fast uniform estimation rates if the vector β_r is sparse for all $r=1,\dots,d$. Due to the approximation error $a_r(X_r)$, the sparsity assumption is quite mild and contains an approximate sparse setting. In this setting, $d=d_n$ is explicitly allowed to grow with n. In the following analysis, the regressors and errors need to have at least subexponential tails. In this context, we define the Orlicz norm $\|X\|_{\Psi_\rho}$ as

$$\|X\|_{\Psi_\rho}=\inf\{C>0:\mathbb{E}[\Psi_\rho(|X|/C)]\leq 1\}$$

with $\Psi_\rho(x)=\exp(x^\rho)-1$.

4.9.1 Uniform lasso estimation

Define the weighted lasso estimator

$$\hat{\beta}_r\in\arg\min_{\beta}\left(\frac{1}{2}\mathbb{E}_n\left[(Y_r-\beta X_r)^2\right]+\frac{\lambda}{n}\|\hat{\Psi}_{r,m}\beta\|_1\right)$$

with the penalty level

$$\lambda=c_\lambda\sqrt{n}\Phi^{-1}\left(1-\frac{\gamma}{2pd}\right)$$

for a suitable $c_\lambda>1$, $\gamma\in[1/n,1/\log(n)]$ and a fix $m\geq 0$. Define the post-regularized weighted least squares estimator as

$$\tilde{\beta}_r\in\arg\min_{\beta}\left(\frac{1}{2}\mathbb{E}_n\left[(Y_r-\beta X_r)^2\right]\right):\quad \operatorname{supp}(\beta)\subseteq\operatorname{supp}(\hat{\beta}_r).$$

The penalty loadings $\hat{\Psi}_{r,m} = \operatorname{diag}(\{\hat{l}_{r,j,m}, j = 1, \ldots, p\})$ are defined by

$$\hat{l}_{r,j,0} = \max_{1 \leq i \leq n} ||X_r^{(i)}||_\infty$$

for $m = 0$ and for all $m \geq 1$ by the following algorithm:

Algorithm 2 Penalty loadings

```
1. Set $\bar{m} = 0$. Compute $\hat{\beta}_r$ based on $\hat{\Psi}_{r,\bar{m}}$.
2. Set $\hat{l}_{r,j,\bar{m}+1} = \mathbb{E}_n\left[\left(\left(Y_r - \hat{\beta}_r X_r\right) X_{r,j}\right)^2\right]^{1/2}$.
3. If $\bar{m} = m$ stop and report the current value of $\hat{\Psi}_{r,m}$, otherwise set $\bar{m} = \bar{m} + 1$.
```

Let $a_n := \max(p, n, d, e)$. In order to establish uniform convergence rates, the following assumptions are required to hold uniformly in $n \geq n_0$ and $P \in \mathcal{P}_n$:

Assumption C. 1.

(i) There exists $1 \leq \rho \leq 2$ *such that*

$$\max_{r=1,\ldots,d} \max_{j=1,\ldots,p} \|X_{r,j}\|_{\Psi_\rho} \leq C \textit{ and } \max_{r=1,\ldots,d} \|\varepsilon_r\|_{\Psi_\rho} \leq C.$$

(ii) For all $r = 1, \ldots, d_n$, *it holds*

$$\inf_{\|\xi\|_2 = 1} \mathbb{E}\left[(\xi X_r)^2\right] \geq c, \sup_{\|\xi\|_2 = 1} \mathbb{E}\left[(\xi X_r)^2\right] \leq C$$

and

$$\min_{j=1,\ldots,p} \mathbb{E}[\epsilon_r^2 X_{r,j}^2] \geq c > 0.$$

(iii) The coefficients obey

$$\max_{r=1,\ldots,d} \|\beta_r\|_0 \leq s.$$

(iv) There exists a positive number $\tilde{q} > 0$ *such that the following growth condition is fulfilled:*

$$n^{\frac{1}{\tilde{q}}} \frac{s \log^{1+\frac{4}{\rho}}(a_n)}{n} = o(1).$$

(v) The approximation error obeys

$$\max_{r=1,\ldots,d} \|a_r(X_r)\|_{P,2} \leq C \sqrt{\frac{s \log(a_n)}{n}}$$

and

$$\max_{r=1,\ldots,d} (\mathbb{E}_n[(a_r(X_r))^2] - E[(a_r(X_r))^2]) \leq C \frac{s \log(a_n)}{n}$$

with probability $1 - o(1)$.

Theorem 5. *Under the Assumption C.1, the lasso estimator* $\hat{\beta}_r$ *obeys uniformly over all* $P \in \mathcal{P}_n$ *with* *probability* $1 - o(1)$

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r\|_2 \leq C\sqrt{\frac{s\log(a_n)}{n}}, \tag{4.17}$$

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r\|_1 \leq C\sqrt{\frac{s^2\log(a_n)}{n}} \tag{4.18}$$

and

$$\max_{r=1,\dots,d} \|\hat{\beta}_r\|_0 \leq Cs. \tag{4.19}$$

Additionally, the post-lasso estimator $\tilde{\beta}_r$ *obeys uniformly over all* $P \in \mathcal{P}_n$ *with probability* $1 - o(1)$

$$\max_{r=1,\dots,d} \|\tilde{\beta}_r - \beta_r\|_2 \leq C\sqrt{\frac{s\log(a_n)}{n}}, \tag{4.20}$$

$$\max_{r=1,\dots,d} \|\tilde{\beta}_r - \beta_r\|_1 \leq C\sqrt{\frac{s^2\log(a_n)}{n}}. \tag{4.21}$$

Proof of Theorem 5.

In the following, we use C for a strictly positive constant, independent of n, which may have a different value in each appearance. The notation $a_n \lesssim b_n$ stands for $a_n \leq Cb_n$ for all n for some fixed C. Additionally, $a_n = o(1)$ stands for uniform convergence towards zero meaning that there exists a sequence $(b_n)_{n\geq 1}$ with $|a_n| \leq b_n$, where b_n is independent of $P \in \mathcal{P}_n$ for all n and $b_n \to 0$. Finally, the notation $a_n \lesssim_P b_n$ means that for any $\epsilon > 0$, there exists C such that uniformly over all n we have $P_P(a_n > Cb_n) \leq \epsilon$.

Due to Assumption C.1(i), we can bound the q-th moments of the maxima of the regressors uniformly by

$$\begin{aligned}
\mathbb{E}\left[\max_{r=1,\dots,d} \|X_r\|_\infty^q\right]^{\frac{1}{q}} &= \|\max_{r=1,\dots,d}\max_{j=1,\dots,p} |X_{r,j}|\|_{P,q} \\
&\leq q!\|\max_{r=1,\dots,d}\max_{j=1,\dots,p} |X_{r,j}|\|_{\psi_1} \\
&\leq q!\log^{\frac{1}{\rho}-1}(2)\|\max_{r=1,\dots,d}\max_{j=1,\dots,p} |X_{r,j}|\|_{\psi_\rho} \\
&\leq q!\log^{\frac{1}{\rho}-1}(2)K\log^{\frac{1}{\rho}}(1+dp)\max_{r=1,\dots,d}\max_{j=1,\dots,p} \|X_{r,j}\|_{\psi_\rho} \\
&\leq C\log^{\frac{1}{\rho}}(a_n),
\end{aligned}$$

where C does depend on q and ρ but not on n. For the norm inequalities, we refer to Van der Vaart and Wellner (1996). Now, we essentially modify the proof from Theorem 4.2 from Belloni et al. (2018) to fit our setting and keep the notation as similar as possible. Let $\mathcal{U} = \{1, \dots, d\}$ and

$$\beta_r \in \arg\min_{\beta\in\mathbb{R}^p} \mathbb{E}\Big[\underbrace{\frac{1}{2}\left(Y_r - \beta X_r - a_r(X_r)\right)^2}_{:=M_r(Y_r,X_r,\beta,a_r)}\Big]$$

for all $r = 1, \dots, d$. The approximation error $a_r(X_r)$ is estimated with $\hat{a}_r \equiv 0$. Define

$$M_r(Y_r, X_r, \beta) := M_r(Y_r, X_r, \beta, \hat{a}_r) = \frac{1}{2}\left(Y_r - \beta X_r\right)^2.$$

Then, we have

$$\hat{\beta}_r \in \arg\min_{\beta \in \mathbb{R}^p} \left(\mathbb{E}_n\left[M_r(Y_r, X_r, \beta)\right] + \frac{\lambda}{n}\|\hat{\Psi}_r\beta\|_1\right)$$

and

$$\tilde{\beta}_r \in \arg\min_{\beta \in \mathbb{R}^p} \left(\mathbb{E}_n\left[M_r(Y_r, X_r, \beta)\right]\right): \quad \operatorname{supp}(\beta) \subseteq \operatorname{supp}(\hat{\beta}_r).$$

First, we verify the Condition WL from Belloni et al. (2018). Since $N_n = d$, we have $N(\varepsilon, \mathcal{U}, d_{\mathcal{U}}) \leq N_n$ for all $\varepsilon \in (0, 1)$ with

$$d_{\mathcal{U}}(i, j) = \begin{cases} 0 & \text{for } i = j \\ 1 & \text{for } i \neq j. \end{cases}$$

To prove WL(i), we note that

$$S_r = \partial_\beta M_r(Y_r, X_r, \beta, a_r)\big|_{\beta = \beta_r^{(1)}} = -\varepsilon_r X_r.$$

Since $\Phi^{-1}(1-t) \lesssim \sqrt{\log(1/t)}$, uniformly over $t \in (0, 1/2)$, it holds

$$\begin{aligned} \|S_{r,j}\|_{P,3}\Phi^{-1}(1-\gamma/2pd) &= \|\varepsilon_r X_{r,j}\|_{P,3}\Phi^{-1}(1-\gamma/2pd) \\ &\leq \left(\|\varepsilon_r\|_{P,6}\|X_{r,j}\|_{P,6}\right)^{1/2}\Phi^{-1}(1-\gamma/2pd) \\ &\leq C\log^{\frac{1}{2}}(a_n) \lesssim \varphi_n n^{\frac{1}{6}} \end{aligned}$$

with

$$\varphi_n = O\left(\frac{\log^{\frac{1}{2}}(a_n)}{n^{\frac{1}{6}}}\right) = o(1)$$

uniformly over all $j = 1, \dots, p$ and $r = 1, \dots, d$ by Assumption C.1(i) and C.1(iv). Further, it holds

$$\begin{aligned} \mathbb{E}\left[S_{r,j}^2\right] &= \mathbb{E}\left[\varepsilon_r^2 X_{r,j}^2\right] \\ &\leq \left(\mathbb{E}\left[\varepsilon_r^4\right]\mathbb{E}\left[X_{r,j}^4\right]\right)^{1/2} \\ &\leq C \end{aligned}$$

for all $j = 1, \dots, p$ and $r = 1, \dots, d$ by Assumption C.1(i) and

$$\mathbb{E}\left[S_{r,j}^2\right] = \mathbb{E}\left[\varepsilon_r^2 X_{r,j}^2\right] \geq c$$

by Assumption C.1(ii), which implies Condition $WL(ii)$. Note that Condition $WL(iii)$ simplifies to

$$\max_{r=1,\dots,d}\max_{j=1,\dots,p}|(\mathbb{E}_n - \mathbb{E})[S_{r,j}^2]| \leq \varphi_n$$

with probability $1 - \Delta_n$. Now, we use the Maximal Inequality, see Lemma $P.2$ in Belloni et al. (2018). Let $\mathcal{W} = (\mathcal{Y}, \mathcal{X})$ with $Y = (Y_1, \dots, Y_d) \in \mathcal{Y}$ and $X = (X_1, \dots, X_d) \in \mathcal{X}$. Define

$$\mathcal{F} := \{f_{r,j}^2 | r = 1, \dots, d, j = 1, \dots, p\}$$

with

$$\begin{aligned} f_{r,j} :& \mathcal{W} = (\mathcal{Y}, \mathcal{X}) \to \mathbb{R} \\ & W = (Y, X) \mapsto -\left(Y_r - \beta_r X_r - a_r(X_r)\right) X_{r,j} = -\varepsilon_r X_{r,j} = S_{r,j}. \end{aligned}$$

Note that

$$\begin{aligned} \| \sup_{f\in\mathcal{F}} |f| \|_{P,q} &= \| \max_{r=1,\dots,d} \max_{j=1,\dots,p} |f^2_{r,j}| \|_{P,q} \\ &= \mathbb{E}\left[\max_{r=1,\dots,d} \max_{j=1,\dots,p} \varepsilon_r^{2q} X_{r,j}^{2q} \right]^{1/q} \\ &\leq \mathbb{E}\left[\max_{r=1,\dots,d} \varepsilon_r^{2q} \max_{r=1,\dots,d} \max_{j=1,\dots,p} X_{r,j}^{2q} \right]^{1/q} \\ &\leq \left(\mathbb{E}\left[\max_{r=1,\dots,d} \varepsilon_r^{4q} \right]^{1/4q} \mathbb{E}\left[\max_{r=1,\dots,d} \max_{j=1,\dots,p} X_{r,j}^{4q} \right]^{1/4q} \right)^2 \\ &\leq C \log^{\frac{4}{\rho}}(a_n). \end{aligned}$$

Since

$$\sup_{f\in\mathcal{F}} \|f\|^2_{P,2} = \max_{r=1,\dots,d} \max_{j=1,\dots,p} \mathbb{E}\left[S^4_{r,j}\right] \leq \max_{r=1,\dots,d} \max_{j=1,\dots,p} \mathbb{E}\left[\varepsilon_r^8\right]^{1/2} \mathbb{E}\left[X^8_{r,j}\right]^{1/2} \leq C,$$

we can choose a constant with

$$\sup_{f\in\mathcal{F}} \|f\|^2_{P,2} \leq C \leq \| \sup_{f\in\mathcal{F}} |f| \|^2_{P,2}.$$

Additionally, it holds $|\mathcal{F}| = dp$ which implies

$$\log \sup_{Q} N(\epsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leq \log(dp) \lesssim \log(a_n/\epsilon), \quad 0 < \epsilon \leq 1.$$

Using Lemma $P.2$ from Belloni et al. (2018), we obtain with probability not less than $1 - o(1)$

$$\begin{aligned} \max_{r=1,\dots,d} \max_{j=1,\dots,p} |(\mathbb{E}_n - \mathbb{E})[S^2_{r,j}]| &= n^{-1/2} \sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)| \\ &\leq n^{-1/2} C \left(\sqrt{\log(a_n)} + n^{-1/2+1/q} \log^{1+\frac{4}{\rho}}(a_n) \right) \\ &= C \left(\sqrt{\frac{\log(a_n)}{n}} + \frac{\log^{1+\frac{4}{\rho}}(a_n)}{n^{1-1/q}} \right) \\ &\leq \varphi_n = o(1) \end{aligned}$$

by the growth condition in Assumption C.1(iv). We proceed by verifying Assumption $M.1$ in Belloni et al. (2018). The function $\beta \mapsto M_r(Y_r, X_r, \beta)$ is convex, which is the first requirement of Assumption $M.1$. We now proceed with a simplified version of proof of $K.1$ from Belloni et al. (2018). To show Assumption $M.1$ (a), note that for all $\delta \in \mathbb{R}^p$

$$\begin{aligned} & \left| \mathbb{E}_n \left[\partial_\beta M_r(Y_r, X_r, \beta_r) - \partial_\beta M_r(Y_r, X_r, \beta_r, a_r) \right]^T \delta \right| \\ = & \left| \mathbb{E}_n \left[X_r(a_r(X_r)) \right]^T \delta \right| \leq ||a_r(X_r)||_{\mathbb{P}_n,2} ||X_r^T \delta||_{\mathbb{P}_n,2} \\ \lesssim_P & \sqrt{\frac{s\log(a_n)}{n}} ||X_r^T \delta||_{\mathbb{P}_n,2} \end{aligned}$$

for all $r = 1, \dots, d$ due to C.1(v). Further, we have

$$\begin{aligned} &\mathbb{E}_n\left[\frac{1}{2}\left(Y_r - (\beta_r + \delta^T)X_r\right)^2\right] - \mathbb{E}_n\left[\frac{1}{2}\left(Y_r - \beta_r X_r\right)^2\right] \\ &= -\mathbb{E}_n\left[\left(Y_r - \beta_r X_r\right)\delta^T X_r\right] + \frac{1}{2}\mathbb{E}_n\left[(\delta^T X_r)^2\right], \end{aligned}$$

where

$$-\mathbb{E}_n\left[\left(Y_r - \beta_r X_r\right)\delta^T X_r\right] = \mathbb{E}_n\left[\partial_\beta M_r(Y_r, X_r, \beta_r)\right]^T \delta$$

and

$$\frac{1}{2}\mathbb{E}_n\left[(\delta^T X_r)^2\right] = ||\sqrt{w_r}\delta^T X_r||^2_{\mathbb{P}_n,2}$$

with $\sqrt{w_r} = 1/4$. This gives us Assumption $M.1$ (c) with $\Delta_n = 0$ and $\bar{q}_{A_r} = \infty$. Since Condition $WL(ii)$ and $WL(iii)$ hold, we have with probability $1 - o(1)$

$$1 \lesssim l_{r,j} = \left(\mathbb{E}_n[S^2_{r,j}]\right)^{1/2} \lesssim 1$$

uniformly over all $r = 1, \dots, d$ and $j = 1, \dots, p$ which directly implies

$$1 \lesssim \|\hat{\Psi}^{(0)}_r\|_\infty := \max_{j=1,\dots,p} |l_{r,j}| \lesssim 1$$

and additionally

$$1 \lesssim \|(\hat{\Psi}^{(0)}_r)^{-1}\|_\infty := \max_{j=1,\dots,p} |l^{-1}_{r,j}| \lesssim 1.$$

For now, we suppose that $m = 0$ in Algorithm 2. Uniformly over $r = 1, \dots, d$ and $j = 1, \dots, p$, we have

$$\hat{l}_{r,j,0} = \left(\mathbb{E}_n[\max_{1\le i\le n} \|X^{(i)}_r\|^2_\infty]\right)^{1/2} \ge \left(\mathbb{E}_n[\|X_r\|^2_\infty]\right)^{1/2} \gtrsim_P 1,$$

where the last inequality holds due to Assumption C.1(ii) and an application of the Maximal Inequality. Also uniformly over $r = 1, \dots, d$, $j = 1, \dots, p$ and for an arbitrary $q > 0$, it holds

$$\begin{aligned} \hat{l}_{r,j,0} &= \max_{1\le i\le n} \|X^{(i)}_r\|_\infty \\ &\le n^{1/q}\left(\frac{1}{n}\sum_{i=1}^n \|X^{(i)}_r\|^q_\infty\right)^{1/q} \\ &= n^{1/q}\left(\mathbb{E}_n[\|X_r\|^q_\infty]\right)^{1/q} \end{aligned}$$

with

$$\mathbb{E}[\|X_r\|^q_\infty]^{1/q} \lesssim \log^{\frac{1}{\rho}}(a_n).$$

By Maximal Inequality, we obtain with probability $1 - o(1)$ for a sufficiently large $q' > 0$

$$\begin{aligned} &\max_r |\mathbb{E}_n[\|X_r\|^q_\infty] - \mathbb{E}[\|X_r\|^q_\infty]| \\ &\lesssim C\left(\sqrt{\frac{\log^{\frac{2q}{\rho}+1}(a_n)}{n}} + n^{1/q'-1}\log^{\frac{q}{\rho}+1}(a_n)\right) \\ &\lesssim \log^{\frac{q}{\rho}}(a_n) \end{aligned}$$

since

$$\mathbb{E}[\max_r \|X_r\|_\infty^{qq'}]^{1/q'} \lesssim \log^{\frac{q}{\rho}}(a_n) \text{ and } \max_r \mathbb{E}[\|X_r\|_\infty^{q2}]^{1/2} \lesssim \log^{\frac{q}{\rho}}(a_n).$$

We conclude

$$\begin{aligned}
\hat{l}_{r,j,0} &\leq n^{1/q} \left(\mathbb{E}_n[\|X_r\|_\infty^q]\right)^{1/q} \\
&\leq n^{1/q} \left(|\mathbb{E}_n[\|X_r\|_\infty^q] - \mathbb{E}[\|X_r\|_\infty^q]| + \mathbb{E}[\|X_r\|_\infty^q]\right)^{1/q} \\
&\lesssim_P n^{1/q} \log^{\frac{1}{\rho}}(a_n)
\end{aligned}$$

uniformly over r. Therefore, Assumption $M.1(b)$ holds for some $\Delta_n = o(1)$, $L \lesssim n^{1/q} \log^{\frac{1}{\rho}}(a_n)$ and $l \gtrsim 1$. Hence, we can find a c_l with $l > 1/c_l$. Setting $c_\lambda > c_l$ and $\gamma = \gamma_n \in [1/n, 1/\log(n)]$ in the choice of λ, we obtain

$$P\left(\frac{\lambda}{n} \geq c_l \max_{r=1,\dots,d} \|(\hat{\Psi}_r^{(0)})^{-1}\mathbb{E}_n[S_r]\|_\infty\right) \geq 1 - \gamma - o(\gamma) - \Delta_n = 1 - o(1)$$

due to Lemma $M.4$ in Belloni et al. (2018). Now, we uniformly bound the sparse eigenvalues. Set

$$l_n = \log^{\frac{2}{\rho}}(a_n) n^{2/\bar{q}}$$

for a $\bar{q} > 5\tilde{q}$ with $\tilde{q}$ in C.1(iv). We apply Lemma $Q.1$ in Belloni et al. (2018) with $K \lesssim n^{1/\bar{q}} \log^{\frac{1}{\rho}}(a_n)$ and

$$\begin{aligned}
\delta_n &\lesssim K\sqrt{sl_n} n^{-1/2} \log(sl_n) \log^{\frac{1}{2}}(a_n) \log^{\frac{1}{2}}(n) \\
&\lesssim \sqrt{n^{\frac{4}{\bar{q}}} \log(n) \log^2(sl_n) \frac{s \log^{1+\frac{4}{\rho}}(a_n)}{n}} \\
&\lesssim \sqrt{n^{\frac{5}{\bar{q}}} \frac{s \log^{1+\frac{4}{\rho}}(a_n)}{n}}
\end{aligned}$$

for n large enough. Hence, by the growth condition in Assumption C.1(iv), it holds

$$\delta_n = o(1)$$

which implies

$$1 \lesssim \min_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}^2}{\|\delta\|_2^2} \leq \max_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}^2}{\|\delta\|_2^2} \lesssim 1$$

with probability $1 - o(1)$ uniformly over $r = 1, \dots, d$.
Define $T_r := \operatorname{supp}(\beta_r^{(1)})$ and

$$\tilde{c} := \frac{Lc_l + 1}{lc_l - 1} \max_{r=1,\dots,d} \|\hat{\Psi}_r^{(0)}\|_\infty \|(\hat{\Psi}_r^{(0)})^{-1}\|_\infty \lesssim L.$$

Let the restricted eigenvalues be defined as

$$\bar{\kappa}_{2\tilde{c}} := \min_{r=1,\dots,d} \inf_{\delta \in \Delta_{2\tilde{c},r}} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}}{\|\delta_{T_r}\|_2},$$

where $\Delta_{2\tilde{c},r} := \{\delta : \|\delta_{T_r}^c\|_1 \leq 2\tilde{c}\|\delta_{T_r}\|_1\}$. By the argument given in Bickel et al. (2009), it holds

$$\bar{\kappa}_{2\tilde{c}} \geq \left(\min_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}^2}{\|\delta\|_2^2}\right)^{1/2} - 2\tilde{c}\left(\max_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}^2}{\|\delta\|_2^2}\right)^{1/2} \left(\frac{s}{sl_n}\right)^{1/2}$$

$$\begin{aligned}
&\gtrsim \left(\min_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2}\right)^{1/2} - 2n^{\frac{1}{q}-\frac{1}{\bar{q}}} \left(\max_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2}\right)^{1/2} \\
&\gtrsim 1
\end{aligned}$$

with probability $1 - o(1)$ for a suitable choice of q with $q > \bar{q}$. Since

$$\frac{\lambda}{n} \lesssim n^{-1/2}\Phi^{-1}\left(1 - \gamma/(2dp)\right) \lesssim n^{-1/2}\sqrt{\log(2dp/\gamma)} \lesssim n^{-1/2}\log^{\frac{1}{2}}(a_n)$$

and the uniformly bounded penalty loading from above and away from zero, we obtain

$$\max_{r=1,\dots,d} \|(\hat{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2} \lesssim_P L\sqrt{\frac{s\log(a_n)}{n}}$$

by Lemma $M.1$ from Belloni et al. (2018). To show Assumption $M.1(b)$ for $m \geq 1$, we proceed by induction. Assume that the assumption holds for $\hat{\Psi}_{r,m-1}$ with some $\Delta_n = o(1)$, $l \gtrsim 1$ and $L \lesssim n^{1/q}\log^{\frac{1}{\rho}}(a_n)$. We have shown that the estimator based on $\hat{\Psi}_{r,m-1}$ obeys

$$\max_{r=1,\dots,d} \|(\hat{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2} \lesssim L\sqrt{\frac{s\log(a_n)}{n}}$$

with probability $1 - o(1)$. This implies

$$\begin{aligned}
|\hat{l}_{r,j,m} - l_{r,j}| &= \left|\mathbb{E}_n\left[\left(\left(Y_r - \hat{\beta}_r X_r\right)X_{r,j}\right)^2\right]^{1/2} - \mathbb{E}_n\left[\left(\left(Y_r - \beta_r X_r\right)X_{r,j}\right)^2\right]^{1/2}\right| \\
&\leq \left|\mathbb{E}_n\left[\left(\left((\hat{\beta}_r - \beta_r)X_r\right)X_{r,j}\right)^2\right]^{1/2}\right| \\
&\lesssim \|(\hat{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2} \max_{1\leq i\leq n}\max_{r=1,\dots,d}\|X_r^{(i)}\|_\infty \\
&\lesssim_P L\sqrt{\frac{s\log(a_n)}{n}}n^{1/q}\log^{\frac{1}{\rho}}(a_n) \\
&\lesssim \sqrt{n^{4/q}\frac{s\log^{1+\frac{4}{\rho}}(a_n)}{n}} = o(1)
\end{aligned}$$

uniformly over $r = 1, \dots, d$ and $j = 1, \dots, p$. Therefore, Assumption $M.1(b)$ holds for $\hat{\Psi}_{r,m}$ for some $\Delta_n = o(1)$, $l \gtrsim 1$ and $L \lesssim 1$. Consequently, we obtain

$$\max_{r=1,\dots,d} \|(\hat{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2} \lesssim \sqrt{\frac{s\log(a_n)}{n}}.$$

and

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r\|_1 \lesssim \sqrt{\frac{s^2\log(a_n)}{n}}$$

with probability $1 - o(1)$ due to Lemma $M.1$ in Belloni et al. (2018). Uniformly over all $r = 1, \dots, d$, it holds

$$\begin{aligned}
&\left|\left(\mathbb{E}_n\left[\partial_\beta M_r(Y_r, X_r, \hat{\beta}_r) - \partial_\beta M_r(Y_r, X_r, \beta_r)\right]\right)^T \delta\right| \\
&= \left|\left(\mathbb{E}_n\left[(\hat{\beta}_r - \beta_r)X_r X_r^T\right]\right)^T \delta\right|
\end{aligned}$$

$$\leq \|(\hat{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2}\|\delta X_r\|_{\mathbb{P}_n,2} \leq L_n\|\delta X_r\|_{\mathbb{P}_n,2}$$

with probability $1 - o(1)$ where $L_n \lesssim (s\log(a_n)/n)^{1/2}$. Since the maximal sparse eigenvalues

$$\phi_{max}(l_n s, r) := \max_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2}$$

are uniformly bounded from above, Lemma $M.2$ from Belloni et al. (2018) implies

$$\max_{r=1,\dots,d} \|\hat{\beta}_r\|_0 \lesssim s$$

with probability $1 - o(1)$. Combining this result with the uniform restrictions on the sparse eigenvalues from above, we obtain

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r\|_2 \lesssim \max_{r=1,\dots,d} \|(\hat{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2} \lesssim \sqrt{\frac{s\log(a_n)}{n}}$$

with probability $1 - o(1)$. We now proceed by using Lemma $M.3$ in Belloni et al. (2018). We obtain uniformly over all $r = 1, \dots, d$

$$\begin{aligned}\mathbb{E}_n[M_r(Y_r, X_r, \tilde{\beta}_r)] - \mathbb{E}_n[M_r(Y_r, X_r, \beta_r)] &\leq \frac{\lambda L}{n}\|\hat{\beta}_r - \beta_r\|_1 \max_{r=1,\dots,d}\|\hat{\Psi}_r^{(0)}\|_\infty \\ &\lesssim \frac{\lambda}{n}\|\hat{\beta}_r - \beta_r\|_1 \\ &\lesssim \frac{s\log(a_n)}{n}\end{aligned}$$

with probability $1 - o(1)$, where we used $L \lesssim 1$ and $\max_{r=1,\dots,d}\|\hat{\Psi}_r^{(0)}\|_\infty \lesssim 1$. Since

$$\max_{r=1,\dots,d}\|\mathbb{E}_n[S_r]\|_\infty \leq \max_{r=1,\dots,d}\|\hat{\Psi}_r^{(0)}\|_\infty \|\big(\hat{\Psi}_r^{(0)}\big)^{-1}\mathbb{E}_n[S_r]\|_\infty \lesssim \frac{\lambda}{n} \lesssim n^{-1/2}\log^{\frac{1}{2}}(a_n)$$

with probability $1 - o(1)$, we obtain

$$\max_{r=1,\dots,d}\|(\tilde{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2} \lesssim \sqrt{\frac{s\log(a_n)}{n}}$$

with probability $1 - o(1)$, where we used

$$\max_{r=1,\dots,d}\|\hat{\beta}_r\|_0 \lesssim s,\ C_n \lesssim (s\log(a_n)/n)^{1/2}$$

and that the minimum sparse eigenvalues are uniformly bounded away from zero. With the same argument as above, we obtain

$$\max_{r=1,\dots,d}\|\tilde{\beta}_r - \beta_r\|_2 \lesssim \max_{r=1,\dots,d}\|(\tilde{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2} \lesssim \sqrt{\frac{s\log(a_n)}{n}}.$$

This completes the proof. □

4.10 Computational Details

4.10.1 Computation and Infrastructure

The simulation study has been run on a x86_64_redhat_linux-gnu (64-bit) (CentOS Linux 7 (Core)) cluster using R version `3.6.1 (2019-07-05)`. All lasso estimations are performed using the R package `hdm`, version `0.3.1` by Chernozhukov et al. (2016a) which can be downloaded from CRAN. Construction of B-splines is based on the R package `splines`. R code is available upon request.

4.10.2 Simulation Study: Smoothing Parameters in B-splines

Table 4.4 presents the corresponding smoothing parameters $k = \{k_j, k_{-j}\}$ of the cubic B-splines that are used in the simulation study. k_j denotes the degrees of freedom chosen to approximate the function $f_j(x_j)$ and k_{-j} is chosen for all other functions. In our revised simulation, we consider settings with $k_j = k_{-j}$.

n	p	f_1	f_2	f_3	f_4
100	50	10	7	8	7
100	150	9	7	7	7
1000	50	11	8	10	7
1000	150	9	8	9	7

Table 4.4: **Smoothing parameters, simulation study.**

Smoothing parameters $k = \{k_j, k_{-j}\}$ corresponding to simulation results in Table 4.2.

4.10.3 Empirical Application: Cross-Validation Procedure for Choice of Smoothing Parameter

The choice of the degrees of freedom parameter k for construction of B-splines in the empirical application is based on a heuristic cross-validation which exploits the additive structure of the model. Let $k = \{k_j, k_{-j}\}$ be the degrees of freedom with k_j specifying the smoothing parameters for $f_j(x_j)$ and k_{-j} denoting the parameter for all other functions $f_{-j}(x_{-j})$. To explicitly address the dependence of the fitted function on the chosen degrees of freedom parameter, we use a notation $\hat{f}_j(x_j, k_j)$ which leads to the model

$$y_i = f_j(x_{i,j}, k_j) + f_{-j}(x_{i,-j}, k_{-j}) + \epsilon_i,$$

Then, the heuristic rule for choosing k proceeds as

- For $j = 1, ..., p$,
    1. Set up a grid of values for k_{-j},
    2. Perform a 5-fold cross-validated search for an optimal k_j over a grid of values $\underline{k}_j, ..., \overline{k}_j$, i.e., fit the regression
    $$y_i = f_j(x_{i,j}, k_j) + f_{-j}(x_{i,-j}, k_{-j}) + \epsilon_i$$
    and compute $MSE_{CV}(k_j, k_{-j})$, where $MSE_{CV}(k_j, k_{-j})$ is the cross-validated mean squared error in prediction provided values k_j and $k_{0,-j}$.
    3. Find the optimal value of k_j^* which minimizes MSE_{CV} over all values of k_{-j}.

We experimented with different settings and repeated the procedure multiple times. The resulting parameters are listed in Table 4.5.

Variable	k
NOX	11
$CRIM$	6
ZN	3
$INDUS$	6
RM	6
AGE	5
$DIST$	9
TAX	5
$PTRATIO$	11
$BLACK$	5
$LSTAT$	7

Table 4.5: **Smoothing parameters, Boston housing example.**

4.10.4 Empirical Application: Additional Plots for Explanatory Variables

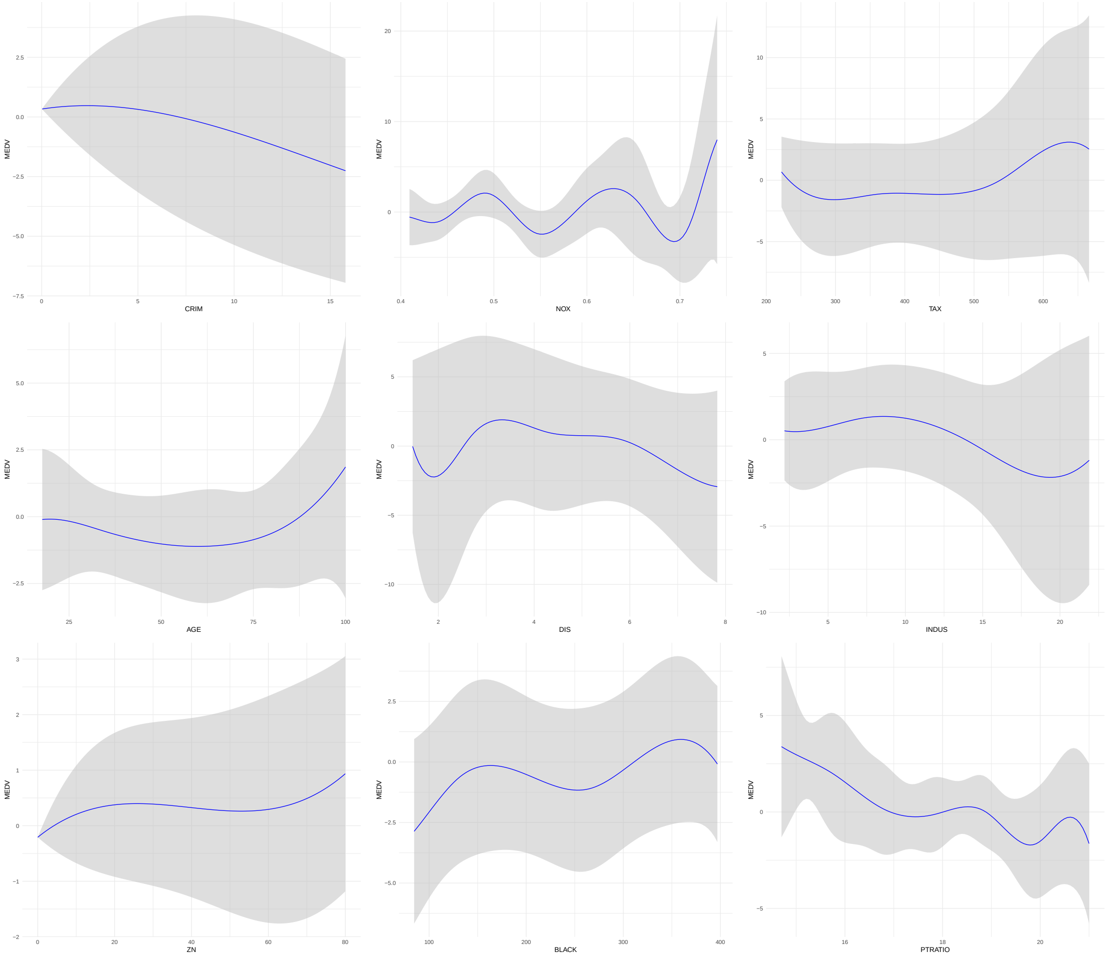


Figure 4.7: **Additional results, Boston housing example.**

Additional plots of the effect of the explanatory variables on the dependent variable $MEDV$ with simultaneous 95%-confidence bands in the Boston housing data application.

Chapter 5

Heterogeneity in the U.S. Gender Wage Gap

5.1 Introduction

Empirical studies on the gender wage gap have come to play an important role in the public and academic debate. Most studies that have attempted to quantify gender inequality in earnings to date employ decomposition methods (Oaxaca, 1973; Blinder, 1973) and, as a key result, report an unadjusted and an adjusted estimate of the gender wage gap. Using recent data from the 2016 American Community Survey (ACS), the unadjusted gender wage gap amounts to around 24% for full-time, year-round employed women with a high school degree or lower level of education. For employees with higher educational attainment, the unadjusted gap is approximately 33%. This unconditional measure of gender based inequality in earnings is often used in political discussion but does not account for differences in observational variables between men and women. For example, men often choose industries and college majors that lead to higher wages on average. Hence, controlling for variables is key and adjusting for differences in observable characteristics changes the pay gap to approximately 17% (for women with high school degree or lower) and 14% (for women with at least a college degree) A classical Oaxaca-Blinder decomposition focuses on average differences in wages and observable characteristics. The goal of this paper, however, is to analyze heterogeneity in the gender wage gap and to contribute to a more comprehensive understanding of gender inequality in income. This is particularly important to design efficient policies to establish equal pay for discriminated groups.

The extent to which the gender wage gap differs across women has attracted public attention and the interest of policy makers. While numerous policy reports and media articles have attempted to quantify heterogeneity in the wage gap, they have generally taken a simplistic approach based on comparing descriptive statistics across subgroups of people. In these studies, the subgroups are usually defined in terms of one characteristic only, such as region, age, race, ethnicity, or occupation (Baxter, 2015; Overberg and Adamy, 2016; The American Association of University Women, 2018; Vasel, 2017; U.S. Bureau of Labor Statistics, 2017; Nelson, 2016). Approaches such as this are likely to lead to flawed conclusions, however, because they neglect heterogeneity due to other observable variables.

We would like to shed light on potential drivers of the gender wage gap. Therefore, we suggest a model that allows for estimation of *ceteris paribus* changes of the gender wage gap according to observable characteristics - in other words, to estimate marginal effects keeping all other variables at fixed levels.

In a first step of our analysis, we estimate heterogeneity in the coefficient of the gender dummy variable in a Mincer Equation like wage regression. In a simple Mincerian wage equation, the gender coefficient

(β) usually measures the average difference in wage between men and women holding all other covariates constant (*ceteris paribus*). We allow the coefficient to vary with socio-economic variables, x, giving rise to a regressor-dependent coefficient $\beta(x)$; in other words, the gender coefficient in the wage equation is determined for different subgroups. With the set of potential covariates being relatively large, we use recent machine learning methods, namely double lasso, for estimation of the effects and inference. In a second step, we illustrate the heterogeneity in the coefficients $\beta(x)$ in quantile plots and provide simultaneous confidence bands. The analysis reveals interesting patterns and shows that the gender wage gap is quite heterogeneous across different groups.

Our study contributes to the literature as it allows to estimate and quantify heterogeneity in the U.S. gender wage gap. Using data from the 2016 ACS, it provides an empirical assessment of a rich set of potential determinants of the gender pay gap apart from classical Oaxaca-Blinder decompositions. We aim to provide a more accurate picture of the gender wage gap than previous studies by allowing the gap to vary with a large number of individual characteristics. In doing so, we consider family and household related demographic variables (i.e., marital status or having biological, adopted or stepchildren at home), race, ethnicity (i.e., Hispanic origin), English language ability, geographic information (i.e., U.S., census region and metropolitan statistical area), veteran status, labor market characteristics (i.e., industry, occupations and hours worked), and the classic human capital variables (i.e., labor market experience and years of education). For people with a bachelor's degree, we also include information on their college major. Considering a large number of explanatory variables creates a challenge in terms of statistical inference. To address this challenge, we employ up-to-date statistical methods that even allow for the number of variables to exceed the number of observations in the data set. In our analysis, statistical inference is performed using the lasso estimator in the double selection approach by Belloni et al. (2014c). In the following, we will refer to this estimator as "double lasso". Furthermore, we contribute to the literature on gender inequality in earnings by estimating the gender wage gap for each woman in the data set and illustrating the resulting distribution of wage gaps for full-time employed women in the United States. We find that, in 2016, the U.S. gender wage gap was highly heterogeneous and, thus, differed considerably from woman to woman depending on individual socio-economic characteristics. Whereas a substantial share of full-time employed women experienced a gap that exceeded the usually reported estimates, the estimated gender wage gap was non-significant for a considerable fraction of female full-time employees.

The paper is structured as follows. In Section 5.2.1, we give a short review on empirical methods in the context on the gender wage gap and recent developments in the literature. Section 5.3 presents the heterogeneous wage gap model together with the inferential framework. Section 5.4 introduces the data used in the empirical analysis, i.e., the American Community Survey (ACS) from 2016. In Section 6.4, we present results on heterogeneity of the U.S. gender wage gap. In Section 5.6, we conclude and present an outlook for future research of the gender wage gap. Additional descriptive statics and results are provided in the Appendix.

5.2 Literature and Evidence on the Gender Wage Gap

5.2.1 Empirical Methods in the Context of the Gender Wage Gap

Traditionally, decomposition methods as initially introduced in Oaxaca (1973) and Blinder (1973) are used in empirical studies that assess the gender wage gap. A detailed and comprehensive overview on decomposition methods and recent extensions thereof is provided in Fortin et al. (2011). The objective of the Oaxaca-Blinder decomposition is to distinguish whether the overall wage difference between men and women arises due to gender differences in observable characteristics or due to a different valuation of these characteristics in the labor market, sometimes referred to as a "wage structure effect" (Fortin et al., 2011). An example for the first effect is a situation with higher labor market experience, on average, for men than for women. Hence, if returns to labor market experience are positive, average earnings are higher for male employees than for female employees. The second effect emerges from the difference of the regression coefficients from two wage regressions that are separately estimated for male and female observations in the data. An example for the structural effect is a situation with higher returns to labor market experience for men than for women. In such a situation, women who have the same level of labor market experience earn on average less than men, provided larger returns to experience for men than for women. Such a gender difference in valuations of labor market characteristics is often considered as an indicator of discrimination, although it might also reflect non-discriminatory effects, for instance unobserved productivity effects (Blau and Kahn, 2017). Recently, the econometric literature has developed innovative methodological extensions of the basic Oaxaca-Blinder decomposition that base upon quantile regression, for instance Chernozhukov et al. (2013b). These methods are able to detect heterogeneous patterns of the gender gap at different points of the income distribution. For instance, the gender wage gap was found to be more pronounced at the top of the income distribution than in the middle or at the bottom Blau and Kahn (2017).

Goldin (2014) provides a recent study of the gender wage gap that can be related to our approach. In the empirical analysis of Goldin (2014) that is based on ACS data, an ordinary least squares regression of an extended wage equation is estimated that included interactions of gender with a large number (i.e., 469) of occupation dummies. Being based on a theoretical argument, the gender wage gap is allowed to vary across occupational categories, and, hence, the focus of the heterogeneity analysis is on variation by occupation. The results of Goldin (2014) illustrate the variation of the gender wage gap in an appealing way. Unfortunately, the significance of the effects is not reported. Under statistical considerations, however, the question of joint significance of heterogeneous effects is of great importance: If the number of tested hypotheses is large, adjustments for simultaneously testing multiple hypotheses are necessary in order to draw valid conclusions.

An approach that is related to our econometric framework has been recently developed by Chernozhukov et al. (2018b). Similar to the quantile plots, which we present in Section 6.4, the so-called sorted effects methods provides estimates and confidence bands for an ordered sequence of partial effects that quantify heterogeneity in terms of observational characteristics. Indeed, the quantile plots in Section 6.4 coincide with the sorted effects if ordinary least squares regression is employed and an appropriate structural regression model is chosen in both approaches. Whereas the interpretation of our quantile plots and the sorted effects is similar, the approach to analyze heterogeneity as a variation of the partial effects in terms of observed variables in Chernozhukov et al. (2018b) differs from our analysis. The so-called classification analysis in Chernozhukov et al. (2018b) provides an inferential framework for testing differences in observational characteristics of individuals in the most and least affected subgroup. In contrast, the focus in our study is on the variation of the gender wage gap estimate according to observational characteristics, in other words variation of $\beta(x)$ according to differences in x. Moreover, we base estimation of the regression equation on the lasso estimator and the double selection framework of Belloni et al. (2014c).

5.2.2 Literature Review: The Gender Wage Gap and Recent Developments

A great number of empirical studies have focused on the gender wage gap, its determinants and its development over time and the life cycle. Due to the richness of the gender gap literature, we restrict attention to the literature on the gender gap in earnings and its determinants. Blau and Kahn (2017) provide an extensive and detailed review of various explanations of the gender wage gap together with an empirical reassessment of many theories.

The second half of the 20th century was characterized by a substantial convergence of the gender wage gap paralleled by a considerable convergence of men and women in terms of education, labor market experience and participation, and occupational choices, among others (Goldin, 2014; Blau and Kahn, 2017). A large part of the reduction of the gender wage gap that began in the 1980s and still continues until today, although in a less steady and slower manner, is attributed to the convergence in traditional human capital factors. Today, women achieve higher levels of education than men and almost the same levels of actual experience, on average. In a recent analysis, Blau and Kahn (2017) provide evidence that gender differences in observable characteristics such as experience, occupation and industry variables, explained two thirds of the total gender gap in 2010. As gender differences in terms of traditional human capital characteristics have diminished over time, these factors have become less important in explaining the gender wage gap. For instance, in the decomposition of Blau and Kahn, 2017, Table 4, differences in human capital characteristics could only explain 13% of the total gender wage gap in 2010 compared to 25% in 1980.

Consequently, alternative explanations have been developed in the labor economics literature. A recently proposed reasoning by Goldin (2014) focuses on the structure of jobs. Temporal flexibility, referring to factors like the total number of hours worked and the time when they are provided, translates into a convex relationship of working hours and the salary. Since women typically value flexibility more than men because of a greater involvement in child rearing, gender inequality in earnings is expected to be more pronounced in inflexible occupations. Goldin (2014) presents evidence that the wage gap was larger and increased over the life cycle in inflexible occupations, for example in the area of business or law, compared to more flexible occupations like pharmacy, science or technology. Moreover, in less flexible occupations, the gender wage gap was found to increase with the number of hours worked due to a more convex hours-earnings relationship. The explanatory power of occupations for the gender gap, together with industries, was also empirically confirmed in the analysis of Blau and Kahn (2017).

The argumentation of Goldin (2014) and other studies is related to the fact that women are more likely to interrupt their work life because of having children and a greater responsibility in child rearing. Using data on actual labor market experience, Blau and Kahn (2017) emphasize the role of work life interruptions for the wage gap. In general, the effect of interruptions is relatively difficult to assess in empirical studies due to limited availability of actual labor market experience in many data sets. Moreover, explanations in favor of a “family” or “motherhood penalty” (Waldfogel, 1998; Sigle-Rushton and Waldfogel, 2007) have been proposed and confirmed empirically implying that mothers tend to experience larger wage gaps than women without children. Recent studies, which mainly use administrative data from Scandinavian countries, assess the dynamics of the motherhood penalty over women’s working history (Kleven et al., 2019; Angelov et al., 2016; Albrecht et al., 2018). A recently published study by Bütikofer et al. (2018) focuses on the motherhood penalty in high-paying jobs in Norway and assesses differences across four occupational categories. The analysis is based on the flexibility argumentation of Goldin (2014) and finds an association of greater motherhood penalties and occupations with lower flexibility.

Furthermore, behavioral explanations suggest that psychological attributes and norms, for example weak preferences for competition and negotiations, cause gender differences in wages (Mueller and Plug, 2006; Manning and Swaffield, 2008). However, Blau and Kahn (2017) conclude that these explanations cannot explain a large fraction of the gender wage gap and that further empirical non-laboratory evidence with stronger external validity is required to assess the importance of these theories.

Finally, taste-based or statistical discrimination is a potential source of the gender wage gap. The adjusted gender wage gap from an Oaxaca-Blinder decomposition is frequently taken as a measure of discrimination. However, the unexplained gap might as well be the result of unobserved factors related to productivity. Hence, there is no unambiguous empirical evidence of discrimination that is based on observational data. Real-world experiments point at a discrimination against women and mothers, for example Neumark et al. (1996) and Correll et al. (2007). Blau and Kahn (2017) conclude that a part of the convergence of the wage gap in the 20th century might be explained by reduced discrimination against women in the labor market.

5.3 An Econometric Model of a Heterogeneous Gender Wage Gap

To motivate our approach, we start with a basic log wage regression where the coefficient β measures the relative difference in pay that arises between men and women if one controls for the effects of observable characteristics. In the following, we use a gender variable that is 1 if a person is female and 0 if male.

$$\ln w_i = \alpha + \beta \cdot \text{gender}_i + x_i'\gamma + \varepsilon_i, \tag{5.1}$$

By construction of the wage equation, estimation of β in Equation (5.1) results in an average gender wage gap that is of the same magnitude for all women - even if they differ in terms of their observable characteristics. Hence, the resulting estimator will not be helpful in determining the driving forces of the wage gap nor to reveal heterogeneity in the wage gap. In order to model heterogeneity, we extend the basic wage equation in (5.1) and let the gender coefficient $\beta = \beta(x_i)$ be a function of individual characteristics.

$$\ln w_i = \alpha + \beta(x_i) \cdot \text{gender}_i + z_i'\delta + \varepsilon_i. \tag{5.2}$$

The $\beta(x_i)$ coefficient can be a linear or a more complicated function of the p_1 observable characteristics x_i, for example using transformations with splines or polynomials of higher order to approximate complex relationships of gender and the other explanatory variables. The covariates z_i in wage Equation (5.2) are natural or constructed regressors, for instance it is possible to apply a so-called dictionary $p(x_i)$ to the initial regressors, x_i, to approximate the relationship of $\ln w_i$ and the observable characteristics. In our empirical application, we approximate $\beta(x_i)$ with a linear function of the regressors, i.e. $\beta(x_i) = \sum_{j=1}^{p_1} \beta_j \cdot x_{i,j}$ and the variables in z_i comprise all two-way interactions of the initial covariates x_i, including a constant. With this specification, the model corresponded to

$$\ln w_i = \alpha + \sum_{j=1}^{p_1} (\beta_j \cdot x_{i,j}) \cdot \text{gender}_i + z_i'\delta + \varepsilon_i. \tag{5.3}$$

We consider p_1 initial characteristics x_i with corresponding coefficients β_j, $j = 1, \ldots, p_1$, that enter $\beta(x_i)$. Together with the dimension p_2 of z_i and the corresponding vector of coefficients δ, the overall dimension of the model is $p = p_1 + p_2 + 1$. A negative β_j, $\beta_j < 0$, is interpreted as an increase of the absolute value of the wage gap. Hence, by default and in line with the presented empirical evidence in the literature, the "gender wage gap" is interpreted as lower earnings for women, although the opposite might be observed in the data. More information on the interpretation is provided in Section 5.7.2 of the Appendix together with a short note on the relation to the Oaxaca-Blinder decomposition.

5.3.1 Valid Post-Selection Inference in High Dimensions

Studying the heterogeneity of the gender wage gap in the presented model requires a rich set of observable characteristics and, hence, modern statistical methods to deal with high-dimensional data. We estimate the wage Equation (5.2) with the lasso and base inference on the double selection approach of Belloni et al. (2014c) that, in combination with the work by Belloni et al. (2014a), provides a uniformly valid inference framework for a vector of "target" coefficients after model selection. In our example, the interactions with gender correspond to the target variables. Hence, under a set of assumptions including sparsity, it is possible to perform valid post-selection inference even in cases where the number of regressors (p) exceeds the number of observations (n). In the context of the heterogeneous wage gap regression, basic lasso estimators $(\widehat{\alpha}, \widehat{\beta}(x_i), \widehat{\delta})$ are defined as the solutions to

$$\begin{aligned}&\left(\widehat{\alpha}, \widehat{\beta}(x_i), \widehat{\delta}\right) \in &(5.4)\\ &\arg \min_{\alpha,\beta(x_i),\delta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\ln w_i - (\alpha + \beta(x_i) \cdot \text{gender}_i + z_i'\delta)\right)^2 + \frac{\lambda}{n} \parallel \Psi \left(\beta(x_i), \delta'\right)' \parallel_1 \right\},\end{aligned}$$

with Ψ being a diagonal matrix with data-dependent weights and $\beta(x_i)$ being specified as $\beta(x_i) = \sum_{j=1}^{p_1} \beta_j \cdot x_{i,j}$ in our empirical application. The lasso as initially developed in Tibshirani (1996) introduces a penalization by the l_1-norm of the coefficients to the least squares problem. This penalization results in a shrinkage being applied to the regression coefficients towards zero. Finally, some of the coefficients are shrunk to a value exactly equal to zero such that the lasso provides a selection device for the set of regressors. Under the assumptions that sparsity holds in the underlying data generating process, solutions obtained with the lasso are sparse, in other words only relatively few, say s, of the p candidate regressors have explanatory power for the outcome variable. Sparsity avoids overfitting that is likely to arise in ordinary least squares regression with many regressors. Estimation of (5.4) requires a choice of the penalty λ. Frequently, λ is set by (k-fold) cross-validation. However, since cross-validation is not backed by theoretical arguments in a high-dimensional setting and computationally expensive, we determined λ by the theory-based rule of Belloni et al. (2012) which is also applicable to the case of heteroskedasticity. An intuitive introduction to the lasso and the reasoning of the penalty choice can be found in Belloni and Chernozhukov (2011). To be more exact, we estimated a post-selection version of the lasso, the so-called *post-lasso* (Belloni and Chernozhukov, 2013): The variables that have been selected by the lasso are used to set up a re-estimation step that is estimated by ordinary least squares regression, which produces the final coefficient estimates. The shrinkage performed the lasso causes a bias of the coefficient estimates, which can be alleviated, in part, by using post-lasso.

Post-selection inference, in other words inference on coefficients after a model selection stage, has been an active research area in the statistics literature in the last years. In general, simply conducting ordinary least squares inference after estimation of Equation (5.4) with the lasso as if there was no variable selection does not result in valid inference unless perfect model selection is achieved. However, the latter is only guaranteed under perfect model selection. However this property is achieved only under relatively restrictive assumptions, for example a so-called *beta-min* assumption that requires that the non-zero coefficients of the true model are well-distinguishable from zero.

The challenge for valid inference after a model selection step with the lasso or other machine learning methods is to avoid selection mistakes for variables that are both correlated with the outcome and the target variables of interest, in other words incorrectly excluding *confounders* from the model. The failure of inference validity due to that omitted variable bias is illustrated in an intuitive example in Belloni et al. (2014c). However, the double selection approach of Belloni et al. (2014c), and more generally, estimation based on orthogonalized moment conditions as in Belloni et al. (2014a), offer an opportunity to overcome the problems of inference after a model selection stage. The idea of the method is to introduce an auxiliary lasso regression for every target coefficient to ensure that only moderate selection mistakes might occur. Double-selection proceeds as follows:

1. For each of the p_1 target variables, estimate a lasso regression of the dependent variable in Equation (5.2), $\ln w_i$, on regressors z_i and the remaining targets. The target variables correspond to the

interactions with gender in Equation (5.2).

2. Estimate an auxiliary lasso regression of each of the p_1 target variables on all remaining independent variables as regressors, i.e., the regressors z_i as well as the remaining targets.
3. Equation (5.2) is re-estimated with ordinary least squares regression with all variables being included that have been selected in either the first or the auxiliary regression steps.

For more details on the double selection approach, we refer to Belloni et al. (2014c). Following Belloni et al. (2014c), Belloni et al. (2014a) and Chernozhukov et al. (2018+) and from asymptotic normality of the double selection estimators, it is possible to show that under sparsity, $\hat{\beta}(x_i)$ as estimated by the double selection approach asymptotically follows a normal distribution

$$\sqrt{n}\left(\hat{\beta}(x_i) - \beta_0(x_i)\right) \rightsquigarrow^d N\left(0, x_i'\Omega x_i\right), \tag{5.5}$$

with variance-covariance matrix Ω of the $\hat{\beta}_j$, $j = 1, \ldots, p_1$, in $\hat{\beta}(x_i)$, which can be estimated according to Belloni et al. (2014c).

As the number of target parameter in the heterogeneous gender wage model is large, it is necessary to adjust for multiple testing. We implement the multiplier bootstrap procedure developed in Chernozhukov et al. (2013a) and Chernozhukov et al. (2014) to construct uniformly valid confidence intervals for $\beta(x_i)$ and perform a valid joint test for the marginal effect targets β_j, $j = 1, ..., p_1$, as suggested in Belloni et al. (2014a). Moreover, to adjust p-values in the joint hypothesis test, we apply the stepdown procedure of Romano and Wolf (2005a), Romano and Wolf (2005b), and Romano and Wolf (2016), as recently established in Chernozhukov et al. (2013a) and Belloni et al. (2014a). For a more detailed presentation of the Romano-Wolf stepdown procedure and the underlying algorithm to construct p-values, we refer to Bach et al. (2018b). More technical details on the inferential framework for the regressor-dependent coefficient $\beta(x)$ is provided in the forthcoming work by Chernozhukov et al. (2018+).

5.4 Heterogeneity in the U.S. Gender Wage Gap

5.4.1 Overview of the 2016 ACS data

In the empirical study, we use data from the 2016 American Community Survey (ACS) as provided by Ruggles et al. (2020) and extracted from the IPUMS-USA website[1]. The ACS provides a representative 1%-sample of the U.S. population. Participation in the survey is mandatory. A large number of socio-economic characteristics at the individual and household level are available, for example referring to education, industry, and occupation. We restrict attention to employed individuals working full time (35+ hours) and year-round, i.e., at least 50 weeks a year, to compare men and women with a similarly strong attachment to the labor force. Weekly earnings are computed as annual earnings divided by 52 (weeks). We focus on individuals aged 25 to 65 and discard persons with income below the mandated federal minimum level of wages corresponding to an hourly wage of \$7.25 or - in terms of annual wage income - to \$12,687.50 according to our sample composition. As the federal minimum level has not been adjusted since 2009, we consider our exclusion rule as not restrictive. However, the rule is sufficient to exclude unrealistic weekly wages, for instance wages corresponding to less than \$1 per hour. The final data set comprises 642,229 individual observations and is stratified into two subgroups according to individuals' highest educational degree. The "bachelor's degree data" comprises 288,095 individuals with at least a bachelor's degree and the "high school degree data" consists of 354,134 observations with at most graduation from high school, GED or equivalent.

---

[1] `https://usa.ipums.org/usa/`

5.4.2 Descriptive Statistics

Table 5.1 provides summary statistics for a selection of variables available in the 2016 ACS data. The descriptive statistics illustrate that wages are substantially higher for college graduates on average. As expected, the individuals holding at least a bachelor degree are in education for a longer time and have less labor market experience, on average. The shares of Hispanics and Blacks are lower in the bachelor's degree subgroup, whereas the share of Chinese is higher. College graduates tend to live in metropolitan statistical areas more frequently and to work longer hours, on average. Also the shares of persons who live with their biological, adopted or stepchildren aged 18 or younger are higher in the bachelor's degree data. Similarly, the share of persons who reside with their biological, adopted or stepchildren aged four or younger is higher in the sample of college graduates. The patterns in terms of marital status differ across the educational attainment subgroups. The share of married (with spouse present) persons is higher in the bachlor's degree data.

In both samples, average earnings of men exceed those of women by far, both in terms of the mean (around 32% for the high school and 49% for the bachelor's degree sample) and the median weekly wage (33% and 42%). An interesting descriptive finding can be observed with regard to the human capital characteristics years of education and experience. The summary statistics for the high school degree data reflect the frequently mentioned reversal of the gender gap in terms of labor market characteristics (Blau and Kahn, 2017). However, we cannot confirm this observation for the sample of college graduates, probably due to selection into full-time employment. The gender gap in terms of years of education is virtually zero. Moreover, we observe that the gender gap in terms of hours worked is still considerable with men working for about 2.4 (bachelor's degree) and 3 (high school) hours each week longer than their female counterparts, on average. Figure 5.1 illustrates the fact that the share of men in the group of employees who regularly work overtime is disproportionately large. Figures 5.4 to 5.9 in the Appendix provide further insights on the distribution of the observable characteristics given gender and educational attainment.

Variable	High school degree data		Bachelor's degree data	
	Men	Women	Men	Women
Weekly wage (mean)	1,098.95	833.83	2,244.29	1,508.86
	(863.34)	(618.87)	(1,996.49)	(1,240.98)
Weekly wage (median)	923.08	692.31	1692.31	1192.31
Single/never married	0.21	0.19	0.19	0.24
	(0.16)	(0.16)	(0.15)	(0.18)
Married, spouse present	0.63	0.54	0.72	0.60
	(0.23)	(0.25)	(0.20)	(0.24)
Child age ≤ 4	0.12	0.08	0.16	0.13
	(0.11)	(0.07)	(0.14)	(0.11)
Child age ≤ 18	0.37	0.33	0.44	0.38
	(0.23)	(0.22)	(0.25)	(0.24)
White	0.86	0.81	0.84	0.81
	(0.12)	(0.16)	(0.14)	(0.16)
Black	0.10	0.15	0.05	0.09
	(0.09)	(0.12)	(0.05)	(0.08)
Chinese	0.01	0.01	0.03	0.03
	(0.01)	(0.01)	(0.03)	(0.03)
Hispanic	0.14	0.12	0.06	0.06
	(0.12)	(0.11)	(0.05)	(0.06)
Experience (years)	27.03	28.48	21.61	20.38
	(11.23)	(11.08)	(10.99)	(11.19)
Years of education	12.43	12.66	16.95	16.97
	(1.17)	(1.06)	(1.28)	(1.23)
Hours worked (mean)	44.93	41.90	46.20	43.79
	(8.56)	(6.34)	(8.65)	(7.34)
Hours worked (median)	40	40	43	40
Veteran status	0.11	0.01	0.07	0.02
	(0.31)	(0.12)	(0.25)	(0.12)
MSA	0.85	0.86	0.94	0.93
	(0.36)	(0.35)	(0.23)	(0.26)
No. of observations	207,549	146,585	154,833	133,262
(%)	58.61	41.39	53.74	46.26

Table 5.1: **Summary statistics.**

Mean and median values for selected observable characteristics. Standard deviation in parantheses. *Data*: Subsample of the American Community Survey 2016 according to sample composition explained in the text.

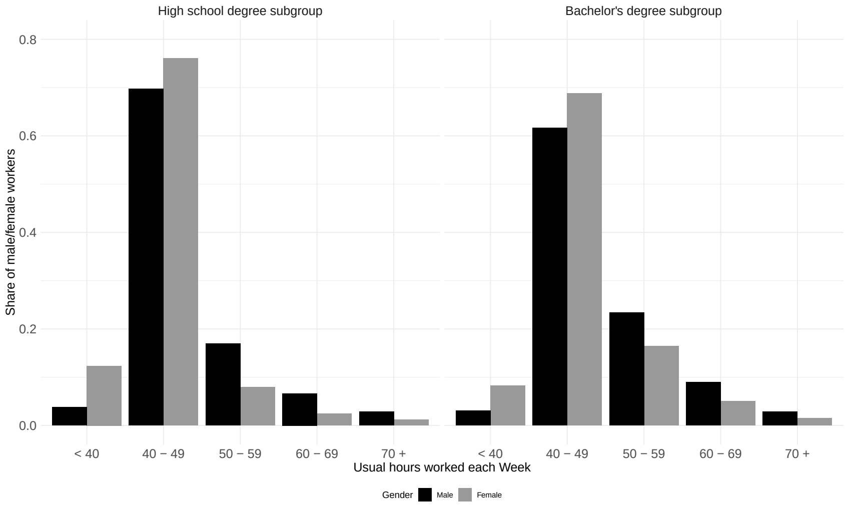


Figure 5.1: **Usual hours worked per week by gender.**

The bar plot in Figure 5.1 illustrates the distribution of male (black bars) and female (gray bars) employees across categories of usual hours worked each week, separately for the two educational attainment subgroups.

The above mentioned decrease of the gender wage gap in terms of human capital characteristics in the high school degree subgroup are in concordance with the estimated female-to-male wage ratios shown in Figure 5.2. We base the log wage ratio analysis on that in Blau and Kahn, 2017, Chapter 2.1. Accordingly, the female-to-male wage ratio is slightly smaller if we conditioned on human capital factors. The resulting wage ratios are 77% if we control for human capital characteristics and 78% if we consider the unconditional gap. If we condition on additional individual characteristics including occupation, industry and hours worked, among others, average wages of female employees are around 17% lower than wages of the male employees.[2] The patterns observed for the sample of academics reveal that conditioning on human capital factors lifts the female-to-male wage ratio from a level of 72% to 76%. Including additional individual characteristics lead to an estimated wage ratio of 87% corresponding to a residual wage gap of approximately 14%.

In the empirical analysis, we use a set of 16 initial regressors to model heterogeneity in the gender wage gap. The variables are listed in Table 5.2 together with information on the baseline categories. The dependent variable in the wage regression is log weekly wage implying that wage gap estimates are reported in log scale throughout the paper. We model parenthood by including two binary variables. The first of these variables indicates that a person resides with one or more biological, adopted or stepchildren of age 18 or younger. The second variable takes on value one if a person lives in the same household with a biological, adopted or stepchild aged 4 or younger. We include both variables to analyze heterogeneity in the motherhood penalty in terms of the age of the child. We use the 14 major groups of the 1990 Census Bureau industry classification scheme available in the ACS (3-digit). Similarly, the Census Bureau provides a 2010 ACS classification of occupations (4-digit) that are clustered into 26 major categories in the ACS. The variable on hours worked is a categorical variable indicating the number of hours usually worked each week in the last 12 months. For the bachelor's degree subgroup, we additionally include the variable college major to account for individuals' educational background in more detail. The exact coding of the categories and definition of the variables can be found on the corresponding documentation website of IPUMs.

To model heterogeneity in the wage gap, we construct all two-way interactions of the initial regressors

---

[2] The 17% wage loss corresponds to the "average female residual from the male wage equation" in Blau and Kahn (2017, p. 800).

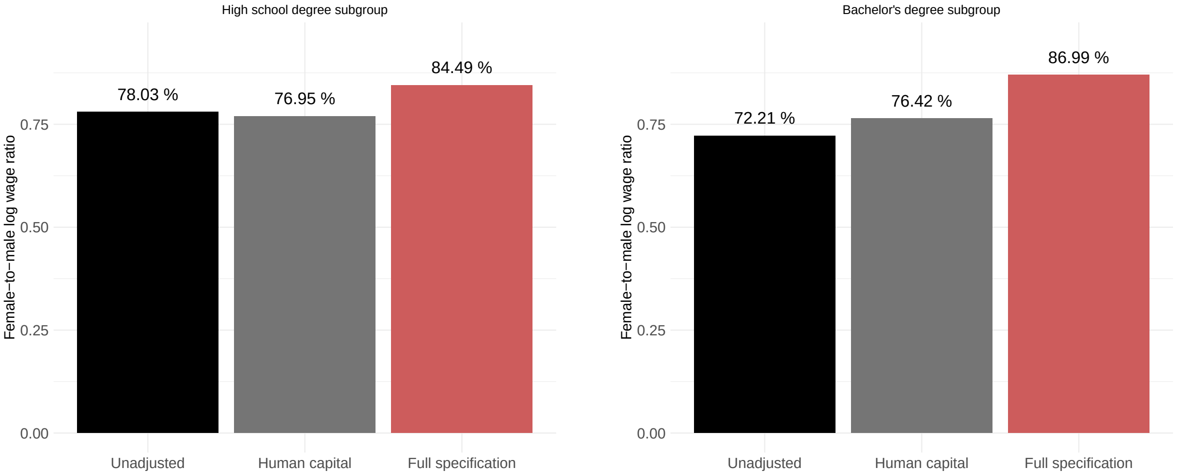


Figure 5.2: **Unconditional and conditional female-to-male mean (log) wage ratios.**

The bar plots indicate the female-to-male log wage ratios, i.e. the quotient $\exp(\overline{X}_f\gamma_f)/\exp(\overline{X}_f\gamma_m)$, with regression coefficients γ_f and γ_m from a regression separately performed for the female and male observations. The unadjusted wage ratio corresponds to $\exp(\overline{\log(w_f)})/\exp(\overline{\log(w_m)})$. The human capital specification includes the regressors years of education, experience and experience squared. In the full regression specification, variables on industry, occupation, the number of hours worked each week and, for the bachelor's degree data only, the field of college major are included. In the regressions, we additionally control for English language ability; veteran status; U.S. census region; race; Hispanic origin; binary variables indicating if a person lives in the same household with a biological, adopted or stepchild children of age 4 or younger and, respectively, of age 18 or younger; metropolitan statistical area; and marital status.

and end up with a high-dimensional setting with a value of p that corresponds in total 2,068 (high school degree subgroup) and 4,382 (bachelor's degree subgroup) regressors (categorical variables are transformed to level-wise dummies and variables with zero-variation are dropped). Of these regressors, 71 (high school) and 106 (bachelor) refer to the initial set of characteristics x_i and dimensions p_1. and 1,997 (high school) and 4,276 (bachelor) to the interacted regressors z_i and dimension p_2 in regression Equation (5.2) in Section 5.3.

In the following, we will selectively report the main findings. Table that provide the coefficient estimates and p-values for all coefficients are presented in the Appendix together with an illustration of the joint confidence bands.

Variable	Type	Baseline Category
Dependent Variable		
Log weekly wage	continuous	
Independent Variables		
Female	binary	
Marital status	6 categories	never married/single
Child age ≤ 4 (One or more biological, adopted or stepchildren at home aged 4 or younger)	binary	
Child age ≤ 18 (One or more biological, adopted or stepchildren at home aged 18 or younger)	binary	
Race	4 categories	White
Hispanic	binary	
English language ability	5 categories	speaks only English
Experience (years)	continuous	
Experience squared	continuous	
Years of education	continuous	
Veteran status	binary	
Industry	14 categories	wholesale trade
Occupation	26 catergories	management, business, science, and arts
Hours worked each week (usually worked in last 12 months)	5 categories	35 to 40 hours
College major (Bachelor's degree data only)	37 categories	education administration and teaching
Region (U.S. census)	9 categories	New England division
MSA (metropolitan statistical area)	binary	

Table 5.2: **List of variables.**

5.5 Results

5.5.1 Significant Variables for the U.S. Gender Wage Gap

In a first step, we focus on the most important drivers of the gap emerging between men's and women's weekly wages. Tables 5.3 and 5.4 present selected β_j-estimates from Equation 5.3 that are jointly significant at the 5%-level. The p-values are obtained from a joint significance test using the multiplier bootstrap procedure developed in Chernozhukov et al. (2013a) and Chernozhukov et al. (2014) and suggested for valid simultaneous inference after model selection in Belloni et al. (2014a) in combination with the stepdown procedure of Romano and Wolf (2005a) as recently established in Chernozhukov et al. (2013a) and Belloni et al. (2014a). The estimated coefficients of discrete regressors indicate changes of the gender wage gap as compared to the gap in the baseline group.[3] For instance, the results obtained for the high school degree subgroup show that married women (spouse present) experience a wage gap that is *ceteris paribus* about 10 to 11 percentage points (pp) larger (in absolute terms) than that of never married women, on average.

Overall, we find that the gender gap varied substantially with individual characteristics in both educational attainment subgroups. The gender wage gap is found to change heavily with family and household-related characteristics (i.e., marriage and motherhood), race, and labor market conditions (i.e., industries and occupations). Patterns with respect to the organization of the household and family as well as race exhibit similarities in the two subsamples. The effects associated with job-related variables like industry, occupation and hours worked, differ in sign and magnitude across the two educational attainment subgroups.

Family and Household-Related Characteristics

We find that, in 2016, the gender wage gap of married women is *ceteris paribus* around nine to 11 percentage points larger than that of women who have never been married. The magnitude of the marriage effect is relatively large compared to the motherhood penalty. According to the results, mothers, who are defined as women living in the same household with at least one biological, adopted or stepchildren aged 18 or younger, have a gender gap that is approximately five to six percentage points larger than that of women who do not reside with a child aged 18 or younger. In both samples, the wage gap for mothers of young children - defined as women who live in the same household with at least one biological, adopted or stepchild aged 4 or younger - is smaller than for those with older children. This effect is particularly sizable in the bachelor's degree subgroup.

These women experience an overall reduction of the wage gap by more than two percentage points as compared to women who do not reside with a biological, adopted or stepchild (of age 18 or younger). This result might serve as evidence of a time-persistent motherhood penalty, at least for mothers with a strong labor force attachment and high levels of education, instead of an immediate wage reduction due to child birth.

---

[3] The gender wage gap in the baseline group is indicated by the constant β_0, i.e. the average gender gap experienced by women with characteristics never married,White, wholesale trade industry, ... All baseline definitions are listed in Table 5.2.

	High school degree subgroup		Bachelor's degree subgroup	
Variable	Estimate	p-value	Estimate	p-value
Constant	-0.0463	0.9070	0.0428	1.0000
Marital status				
Married, spouse pres.	-0.1096	0.0000	-0.0973	0.0000
Married, spouse abs.	-0.0737	0.0010	-0.0535	0.2630
Separated	-0.0575	0.0030	-0.1205	0.0000
Divorced	-0.0571	0.0000	-0.0548	0.0000
Widowed	-0.0536	0.0700	-0.1152	0.0110
Child				
Age 18 or younger	-0.0507	0.0000	-0.0531	0.0000
Age 4 or younger	0.0289	0.0180	0.0809	0.0000
Race				
Black/African American	0.0789	0.0000	0.0679	0.0000
Chinese	0.0819	0.0100	0.0589	0.0020
Other Asian or Pacific Isl.	0.0716	0.0000	0.0437	0.0010
Veteran status				
Veteran	0.0429	0.0140	0.0204	0.9930
Experience				
Exp	-0.0040	0.0000	-0.0024	0.2770
Industry				
TRANS	-0.0535	0.0030	0.0217	1.0000
RETAIL	-0.0444	0.0150	-0.0216	1.0000
FINANCE	-0.0493	0.0180	-0.0799	0.0000
BUISREPSERV	-0.0433	0.0640	-0.0557	0.0450
PROFE	-0.0742	0.0000	-0.0668	0.0000
ADMIN	-0.0527	0.0140	-0.0091	1.0000
Usual hours worked				
40 to 49	-0.0456	0.0000	-0.0104	1.0000
50 to 59	-0.0374	0.0150	-0.0048	1.0000
60 to 69	-0.0534	0.0150	-0.0207	0.9980
>70	-0.1186	0.0000	-0.0623	0.2150
College major				
Comp/Inform Sci	.	.	-0.0666	0.0000
Engin	.	.	-0.0545	0.0000
Bio/Life Sci	.	.	-0.0496	0.0040
Math/Stats	.	.	-0.0683	0.0110
Phys Sci	.	.	-0.0570	0.0040
Psych	.	.	-0.0705	0.0000
Crim Just/Fire Prot	.	.	-0.0788	0.0000
Soc Sci	.	.	-0.0613	0.0000
Bus	.	.	-0.0621	0.0000
Hist	.	.	-0.0561	0.0440

Table 5.3: **Double lasso results.**

Selected results from post-lasso estimation using double selection (double lasso). Coefficients printed in black are significant at the 5% level. p-values are obtained from a joint test of all β_j coefficients in $\beta(x_i)$ from Equation (5.2) using the multiplier bootstrap procedure suggested in Belloni et al. (2014a) with 1000 repetitions in combination with the stepdown procedure of Romano and Wolf (2005a). Results on occupation are presented in Table 5.4.

Race and Ethnicity

The results suggest that there is substantial heterogeneity in the gender gap according to race. Women in race categories other than White experience a significantly smaller gender wage gap. The race-based gender gap differentials are more pronounced in the group with lower educational attainment where the wage gap for non-Whites is seven to eight percentage points smaller than for Whites. Hence, the frequently reported variation of the gender gap according to race is robust to controlling for a large set of characteristics such as years of education, experience, occupation, industry, and level of English language ability, even in the bachelor's degree subgroup. Controlling for observable characteristics render the gender gap differential non-significant only for individuals in the group of Hispanics.

Education and Experience

As we will point out in Section 5.5.2, the heterogeneity patterns of the gender gap for full-time and year-round employed women vary substantially across the two subgroups. Within the two samples, however, classic human capital variables such as years of education and labor market experience only have a minor impact on the magnitude of the gender wage gap, if any. In the wage regression, we additionally include the level of English language ability and do not find a significant effect on the magnitude of the gender gap. The effect of labor market experience is small and only significant for the high school degree subgroup, indicating that for this group the gender gap increases slightly over the employment history. However, the non-significant coefficient on experience squared points at a weak linear relationship of labor market experience and the magnitude of the wage gap. Whereas the coefficient on years of education is non-significant, the gender gap of individuals with post-secondary education is found to vary by several college categories. We found that the gender gap is significantly larger than that in the baseline category "Education Administration and Teaching" in 10 out of the 36 college majors, among others, in natural science disciplines such as "Biology and Life Science", "Physical Sciences", as well as in the categories "Social Sciences" and "Business".

We find distinct patterns of job-related effects in both educational attainment subgroups. As pointed out in Goldin (2014), the lack of temporal flexibility in a job might place women at a disadvantage in terms of earnings as compared to men and, thus, lead to a larger wage gap. Our results on the number of hours usually worked each week are in line with this argumentation as we find that longer working hours are associated with a larger gender wage gap in the high school degree subgroup. The effects are relatively sizable and tend to increase steeply with the number of hours worked. However, for the college graduates the effects are smaller and non-significant.

Industries, Occupation, and Hours Worked

Previous studies provided evidence that the gender wage gap varied according to job-related characteristics associated with specific industries and occupations (Goldin, 2014; Blau and Kahn, 2017).

High school degree subgroup			Bachelor's degree subgroup		
Variable	Coefficient	p-value	Variable	Coefficient	p-value
Educ/Training/Libr	-0.1836	0.0000	Healthc Supp	-0.1022	0.3950
Extract	-0.1448	0.9600	Milit Specific	-0.0799	0.9930
Prod	-0.0974	0.0000	Farm/Fish/Forestry	-0.0498	1.0000
Arts/Design/Entert/ Sports/Media	-0.0304	0.9200	Office/Administr Supp	-0.0465	0.0000
			Healthc Pract/Technic	-0.0407	0.0260
Sales	-0.0187	0.7210	Financ Spec	-0.0348	0.0690
Financ Spec	-0.0127	0.9760	Pers Care/Serv	-0.0287	1.0000
Build/Grounds Clean/ Mainten	-0.0108	0.9710	Build/Grounds Clean/ Mainten	-0.0248	1.0000
Transp	-0.0085	0.9760	Sales	-0.0162	0.9980
			Food Prepar/Serving	-0.0011	1.0000
Archit/Engin	0.0189	0.9760	Prod	0.0065	1.0000
Pers Care/Serv	0.0200	0.9200	Transp	0.0228	1.0000
Comput/Math	0.0246	0.7830	Comput/Math	0.0372	0.0020
Farm/Fish/Forestry	0.0286	0.9760	Bus Operat Spec	0.0377	0.0110
Food Prepar/Serving	0.0290	0.2360	Arts/Design/Entert/ Sports/Media	0.0469	0.0330
Milit Specific	0.0391	0.9760			
Technic	0.0419	0.9200	Legal	0.0495	0.0810
Protect Serv	0.0479	0.0510	Educ/Training/Libr	0.0606	0.0000
Healthc Supp	0.0530	0.0420	Archit/Engin	0.0620	0.0000
Bus Operat Spec	0.0571	0.0030	Protect Serv	0.0666	0.0110
Office/Administr Supp	0.0635	0.0000	Life/Physical/Soc Sci.	0.0719	0.0000
Life/Physical/Soc Sci.	0.0703	0.5360	Technic	0.1126	0.5270
Install/Mainten/Rep	0.0742	0.0070	Constr	0.1469	0.0810
Constr	0.0895	0.0150	Install/Mainten/Rep	0.1496	0.0020
Legal	0.1042	0.3400	Comm/Soc Serv	0.1702	0.0000
Healthc Pract/Technic	0.1075	0.0000			
Comm/Soc Serv	0.1205	0.0000			

Table 5.4: **Occupational effects, high school degree and bachelor's degree data, double lasso.**

The table presents occupational effects for the high school degree subgroup (left) and the bachelor's degree subgroup (right) in increasing order to provide a comparison of the occupational patterns observed for both subgroups. p-values are obtained from a joint test of all β_j coefficients in $\beta(x_i)$ from Equation (5.2) using the multiplier bootstrap procedure suggested in Belloni et al. (2014a) with 1000 repetitions in combination with the stepdown procedure of Romano and Wolf (2005a). Significant (printed black) and non-significant (printed gray) coefficients at a 5% significance level are presented.

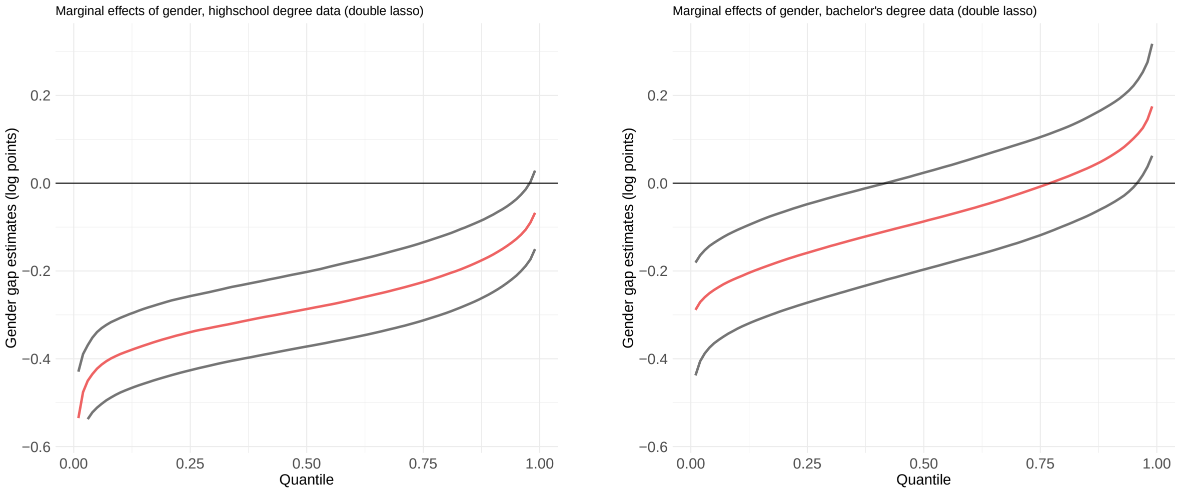
Variation by industry is found to be more pronounced for the high school degree subgroup. For this subgroup the variation of the gap is significant in five industries as compared to only three significant coefficients observed for persons with a college degree. The patterns for the industries “Finance, Insurance, and Real Estate” and “Professional and related Services” are similar in both subgroups corresponding to a larger wage gap by five to eight percentage points than for the baseline industry. For people with at most a high school degree, we observe larger wage gaps in the industries “Transportation, Communications, and other Public Utilities”, “Retail Trade”, and “Public Administration” compared to the baseline category “Wholesale Trade”. Heterogeneity in the wage gap by occupation tends to be more pronounced than heterogeneity by industry which is generally in line with the argumentation and results in Goldin (2014). The magnitude of occupational effects is larger than that of industry effects and the patterns tend to be different in the two educational attainment subgroups. Table 5.4 presents the estimates of the occupational effects in an increasing order. Apparently, the difference in wages between male and female employees are particularly small in occupations in the category “Community and Social Services” (approximately 12pp smaller gender gap for the high school degree subgroup and 17pp for the bachelor’s degree data), as well as in “Healthcare Practitioners and Technical”, “Construction” and “Installation, Maintenance, and Repair” for persons with at most a high school degree. The patterns observed for college graduates are different. Jobs in the category “Healthcare Practitioners and Technical” are associated with a relatively large gender wage gap. Highly educated employees in “Education, Training, and Library” occupations experience a small wage gap as compared to other occupations. In contrast, this occupational category is associated with the largest occupational effect on the gender wage gap for the high school degree subgroup. Moreover, the wage gap for employees with a bachelor’s degree working in a scientific or technical occupation (e.g. “Architecture and Engineering” or “Life, Physical, and Social Science”) experience a wage gap that is relatively small as compared to the baseline category.

In general, the different patterns of job-related effects in the two subsamples can be justified with an extension of the flexibility-based argumentation of Goldin (2014). Basically, the effect of higher education on flexibility can go either way: higher educational achievement might be associated either with an increase or a decrease in job flexibility. On the one hand, higher education, in other words, the acquisition of more abstract skills, might make it easier for women to work in a more flexible work environment given an occupational category, for example in education or science. Whereas on the other hand, higher education might allow to enter jobs, that are in general less flexible than jobs for less-educated in the same occupational category. For example, compared to a nurse, a surgeon as a healthcare practitioner might be less able to work fewer hours or decide when to provide these hours. Recognizing heterogeneity in the gender wage gap is highly relevant for policy makers as it might help them to choose efficient and effective policy measures. In order to contribute to a more comprehensive understanding of gender inequality in earnings, we estimate the gender gap for every woman in the data set. The $\beta(x_i)$ coefficient from Equation (5.2) summarizes all effects attributable to the observed characteristics, and hence, gives an estimate on the gender wage gap of a woman with characteristics x_i. Moreover, the inferential framework presented in Section 5.3.1 allows us to construct confidence bands for $\beta(x_i)$. Thus, we are able to judge whether a women’s wage gap is meaningful from a statistical point of view. We present quantile plots of the individual gender gap estimates together with confidence bands as obtained for all women in the two educational attainment subgroups in Figure 5.3. We additionally report the ordinary least squares results to allow for comparison.

5.5.2 Heterogeneity in the Coefficient Function $\beta(x_i)$

The quantile plot illustrates that the U.S. gender wage gap for full-time and year-round employed women was highly heterogeneous in 2016. Rather than affecting all women to the same extent, gender inequality

Panel A: Quantiles of effects with corresponding confidence bounds, double lasso.



Panel B: Quantiles of effects with corresponding confidence bounds, OLS.

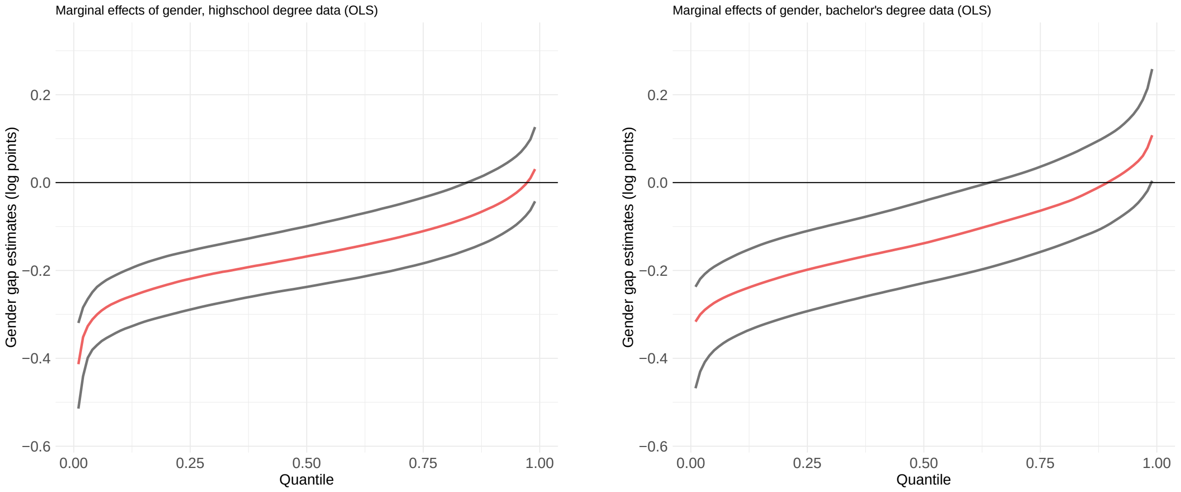


Figure 5.3: **Quantiles of effects with simultaneous confidence bands.**

The plots show the quantiles of the individual gender wage gap estimates as computed for all women in the educational attainment subgroups of the ACS 2016 data together with simultaneous 0.95 confidence bands (gray lines) obtained from the multiplier bootstrap procedure with 500 repetitions. Estimates in Panel (a) are obtained from a high-dimensional wage regression using the double lasso estimator, with log weekly wages as the dependent variable. Plots on the left refer to the high school degree subgroup and plots on the right to the bachelor's degree subgroup. In addition, ordinary least squares results are provided in Panel (b) for reasons of comparison.

in wages consists of a range of wage penalties that differ greatly from woman to woman. For most women, the estimated gap deviates from the above mentioned estimates, derived from traditional analysis. In Section 5.4.2 we reported an unadjusted gap of approximately 24% (high school degree subgroup) and 33% (bachelor's degree subgroup) and adjusted wage gaps of 17% and 14%, respectively.

Patterns of heterogeneity varied substantially across the two samples, with gender wage inequality being more prevalent and more severe among women with lower educational attainment. Whereas more than 90% of female employees with a high school degree or lower earn significantly less than their male counterparts, only 40% of female employees who hold at least a bachelor's degree experience a significant wage penalty according to the double lasso results in Figure 5.3 Panel A. Moreover, at any given quantile, the wage gap is larger for women who do not have a college degree. The median of the estimated wage gaps is around 9% (non-significant) for women with post-secondary education. In contrast, half of the women with at most a high school degree experience a wage gap of at least 29%. Interestingly, there is evidence of a reversal of the gender wage gap for a small share of women with a college degree, i.e., 4% of the full-time and full-year employees with post-secondary education earn significantly more than comparable men according to our double lasso results.

5.5.3 A Robustness Check

We performed a robustness check with regard to the degree of penalization in the lasso estimation steps. The lasso estimator that is based on a theoretical choice of the penalty term λ involves a constant c. A lower value of this parameter is associated with a less severe penalization. The corresponding quantile plots and result tables are presented in the Appendix. The major conclusions drawn in the previous section continue to hold in the setting with less severe penalization. Moreover, we compare the results in Figure 5.3 Panel A to ordinary least squares estimation presented in Figure 5.3 Panel B. However, we observed difficulties of the ordinary least squares estimator in estimation of the covariance matrix. The covariance matrix is used in the multiplier bootstrap procedure used to construct the simultaneous confidence bands. Hence, the corresponding results must be interpreted with caution.

5.6 Conclusion

We started with the objective to analyze heterogeneity in the gender wage gap. Our attempt to answer this question provided detailed insights to the determinants of the gender wage gap and their consequences in the aggregate. Thereby we considered characteristics related to the organization of the family and household, race and ethnicity as well as job-related information.

In summary, our empirical analysis reveals that in 2016, most full-time employed women in the U.S. experienced a substantial wage penalty compared to otherwise identical men. However, the extent to which women were affected by gender inequality in earnings differed greatly according to individual characteristics, including educational attainment, marital status, having children at home, race, and job-related characteristics such as occupation and industry. The commonly used average estimates of the gender wage gap can therefore be seen as a poor approximation of the wage penalty that is experienced by most women. By illustrating and quantifying heterogeneity in the wage gap, we hope to contribute to both the public and the academic discussion, and to provide information that policy makers can use to design more effective policies.

Finally, we would like to point at potential directions for future research. First of all, we consider our analysis as a first step to develop a more precise understanding of gender inequality in earnings. Future work might reproduce and reassess the reported heterogeneity patterns in other data sets, for example the Panel Study of Income Dynamics (PSID) or the Current Population Survey (CPS) and, of course, in other countries. A point that we adopt from the recommendations of Blau and Kahn (2017) is that improved data quality on actual labor market experience, family interrelationships and information at the firm level would allow to get a better impression of the gender wage gap. Moreover, our study was restricted to full-time and year-round working employees. Thus, the extent to which women in our data set adjusted their labor market supply was restricted to the intensive margin, in other words how much hours they work each week in a full-time position (at least 35 hours). Future studies might analyze the gender gap heterogeneity in a yet broader sample including part-time working female workers, as well. Second, we would like to encourage additional work on causal mechanisms that drive the gender wage gap. Whereas our study is a first attempt describing heterogeneous patterns and, hence, reports association of variables and variation in the wage gap, future work on the underlying causal channels is of great relevance. Third, our methodology might be used in future to elaborate optimal equal pay policies. For instance, one could generate hypothetical quantile plots of the gender wage gap for different policy measures targeted at different subgroups and choose the policy that is optimal given a certain objective function and budget constraints.

5.7 Appendix

5.7.1 Relation to Oaxaca-Blinder Decomposition

Comment 5.7.1. *[**Relation to the Oaxaca-Blinder Decomposition**] In the case of a linear function $\beta(\cdot)$ and the covariate vector z_i comprising all two-way interactions of the initial covariates x_i, the heterogeneous gender gap model can be related to the Oaxaca-Blinder decomposition. Suppose, one estimates the wage regression*

$$\ln w_i = \alpha + \beta(z_i) \cdot gender_i + z_i'\delta + \varepsilon_i.$$

Then, the mean of $\beta(z_i)$ corresponds to the negative of the total unexplained gender gap ("the structural effect") from an Oaxaca-Blinder decomposition, in other words the part of the gender wage gap that emerges due to different valuations of labor market characteristics for men and women

$$\overline{\beta(z_i)} = \frac{1}{n_f} \sum_{i=1}^{n_f} \beta(z_i) = -\overline{z}_f' \left(\gamma_m - \gamma_f\right),$$

with (γ_m, γ_f) being the coefficients with regard to z_i that are obtained from the regressions being performed separately for the subset of men (m) and women (f). $n_f = \sum_i^n gender_i$ is the number of female observations, and $\overline{z}_f$ being the matrix collecting the mean values of the interacted initial observable characteristics of women, x_i.

5.7.2 Interpretation of $\beta(x_i)$

The proposed model captures the heterogeneity of the gender gap in the function $\beta(x_i)$. We interpret a negative $\beta(x_i)$ as the approximate gender wage gap experienced by a woman with characteristics x_i on average. Hence, a woman in the subgroup of individuals with characteristics x_i earns approximately $\beta(x_i) \cdot 100\%$ less than a male employee in the same subgroup, in other words a man with the same educational attainment, working in the same industry and occupation, and so on.

We are not only interested in estimating the gender pay gap for every woman in the sample, but also in assessing the determinants of the wage gap. In a linear specification including a constant, $\beta(x_i) = x_i'\beta$, the jth component of β, β_j, indicates the marginal change of the wage gap for a woman differing only with regard to this variable. In case a regressor is continuous, the gender gap change due to a marginal change in variable x_j is ceteris paribus

$$\frac{\partial \beta(x_i)}{\partial x_j} = \beta_j. \tag{5.6}$$

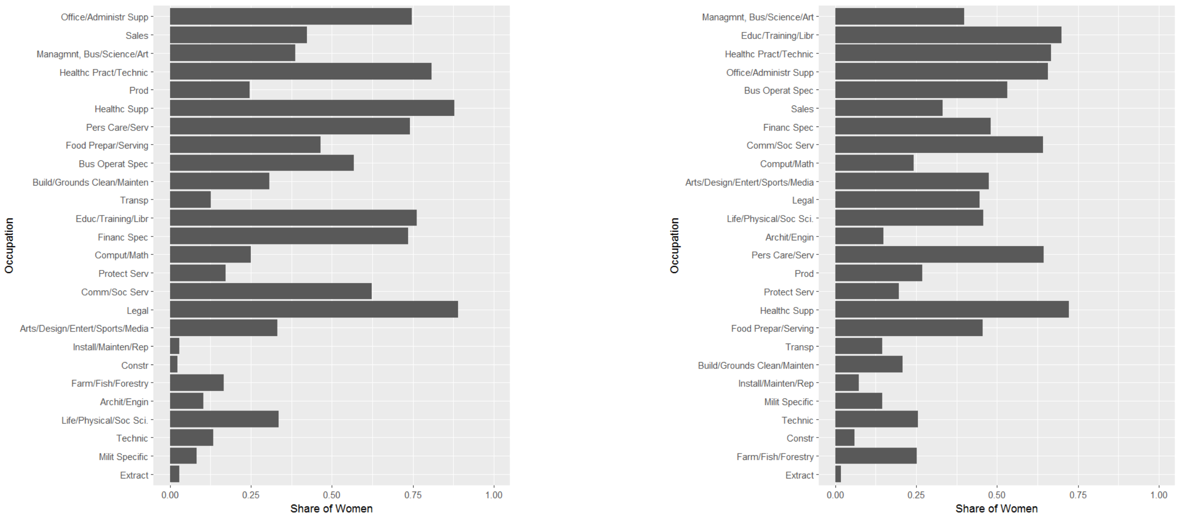
5.7.3 Additional Descriptive Statistics

In the following, summary statistics are generated to illustrate

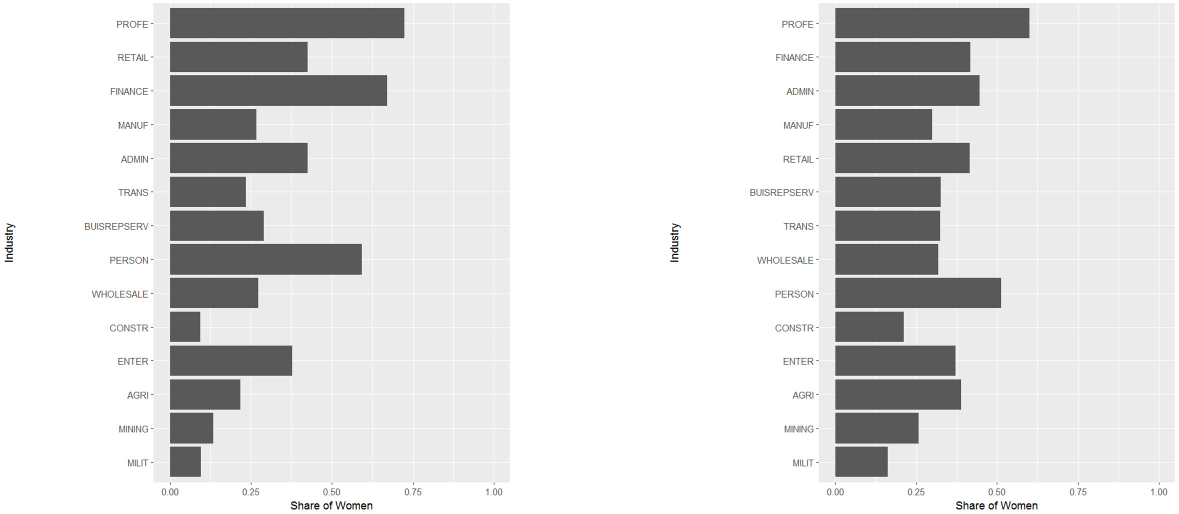
- the share of women within certain subgroups, e.g. within an occupational category or an industry,
- the share of women across certain subgroups in order to illustrate how women distribute in terms of occupations, industries and so on,
- income distribution separately illustrated for male and female full-time employees in order to provide a comparison to the simplistic approach which is mentioned in the introduction of the paper, i.e. a comparison of wages or wage gaps that solely conditions on one characteristic.

Figures are generated separateley for the high school degree subgroup (plots on the left) and the bachelor's degree subgroup (plots on the right).

Panel A: Share of women within occupations.



Panel B: Share of women within industries.



Panel C: Share of women within fields of degree.

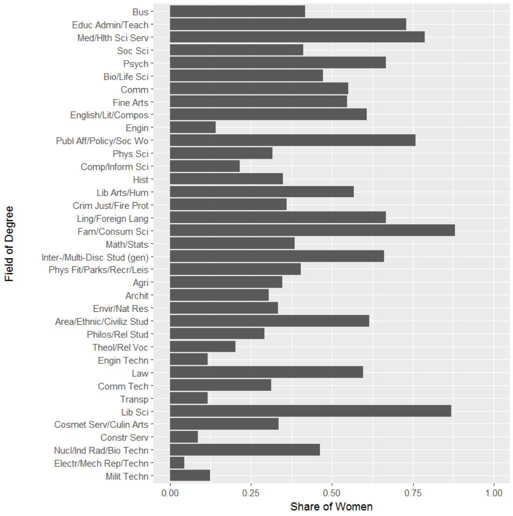
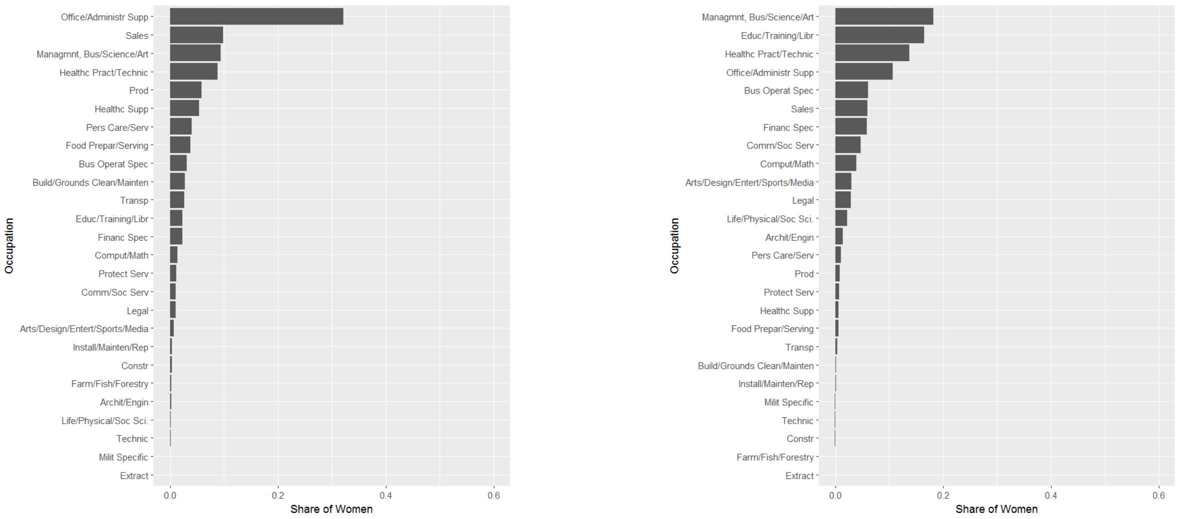


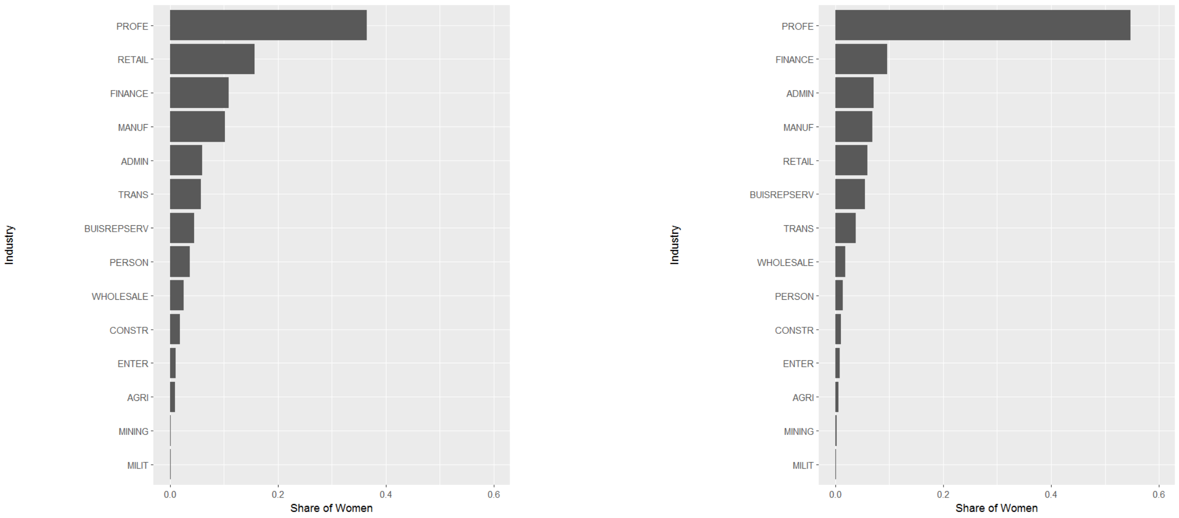
Figure 5.4: **Share of women within occupational, industry and college major categories.**

Plots on the left side refer to the high school degree subgroup, plots on the right side to the bachelor's degree subgroup.

Panel A: Share of women across occupations.



Panel B: Share of women across industries.



Panel C: Share of women across fields of degree.

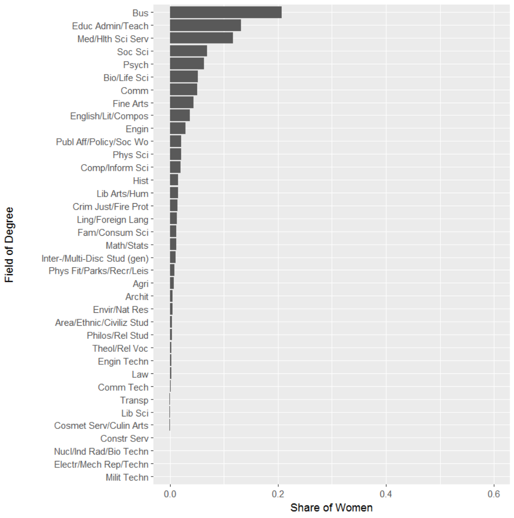
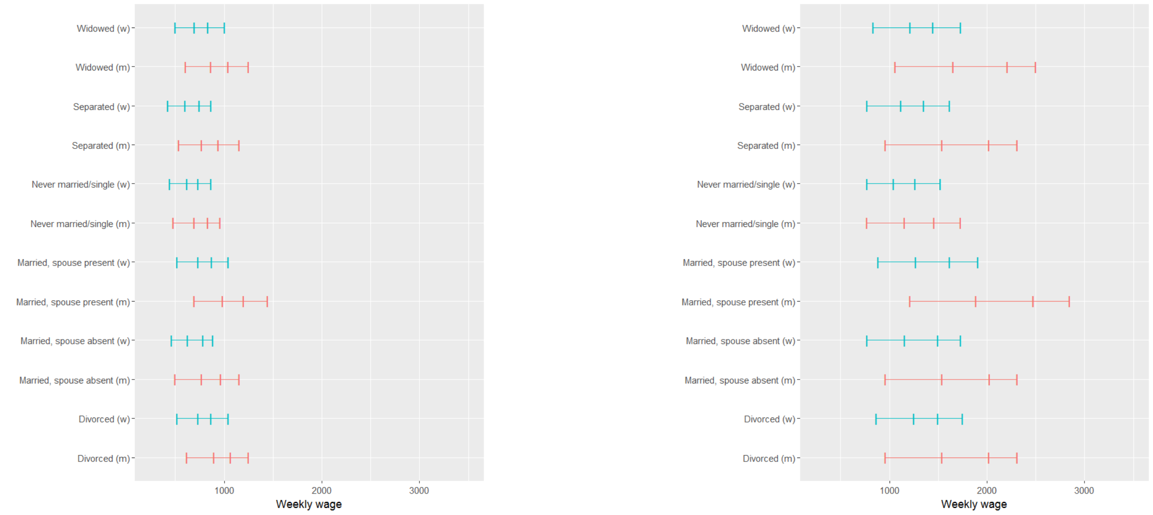


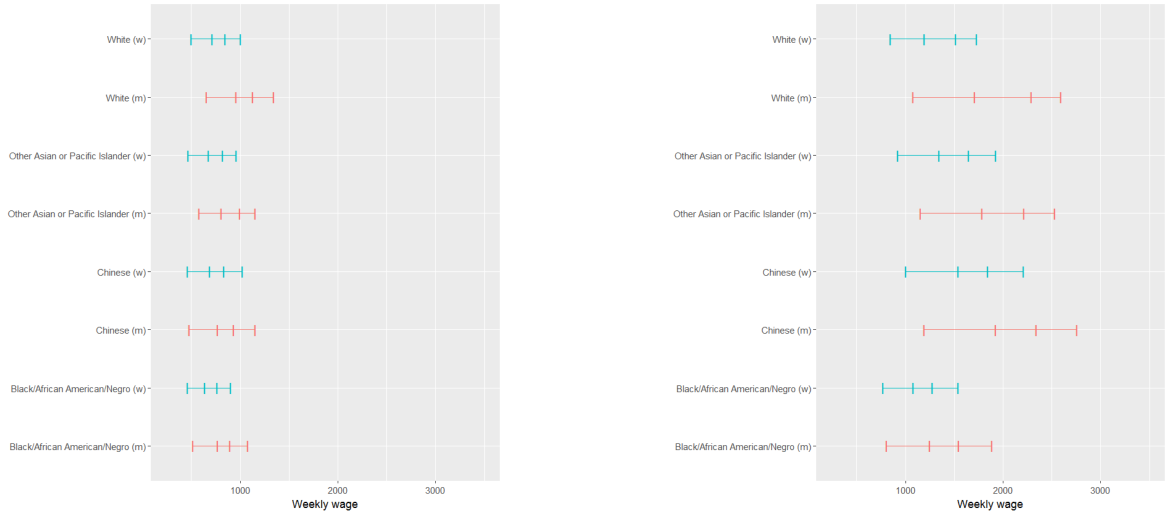
Figure 5.5: **Distribution of women across occupational, industry and college major categories.**

Plots on the left side refer to the high school degree subgroup, plots on the right side to the bachelor's degree subgroup.

Panel A: Marital status



Panel B: Race



Panel C: Hispanic

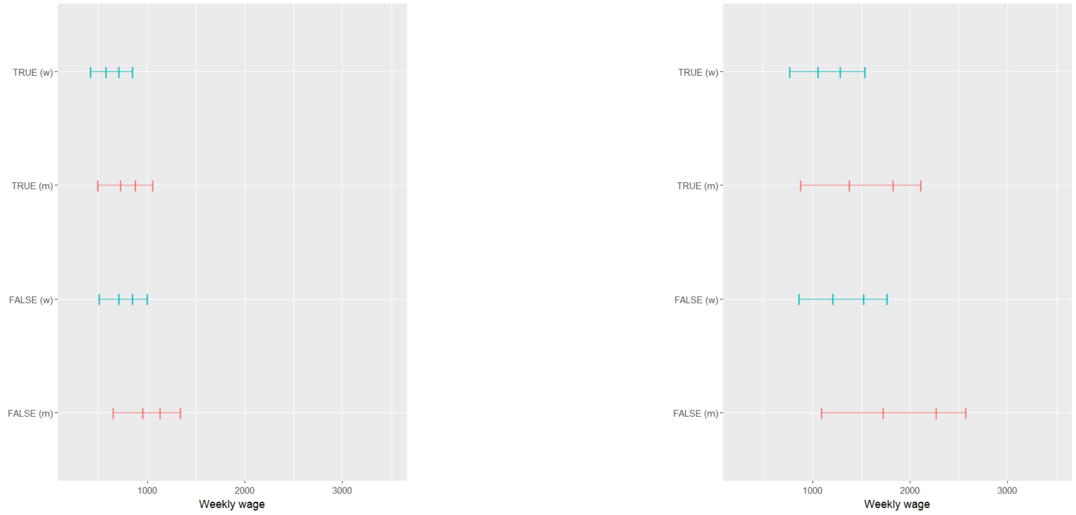
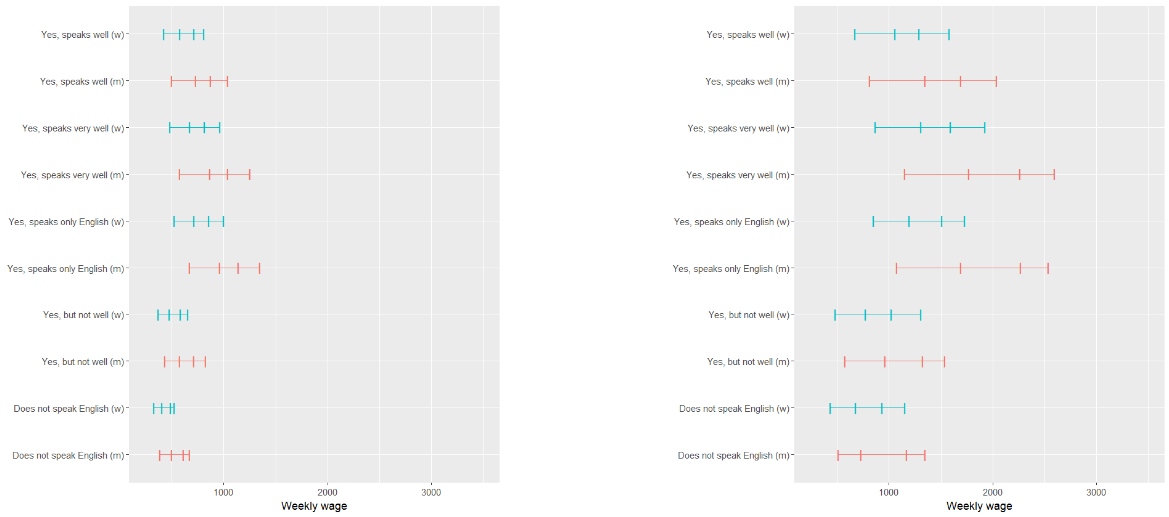


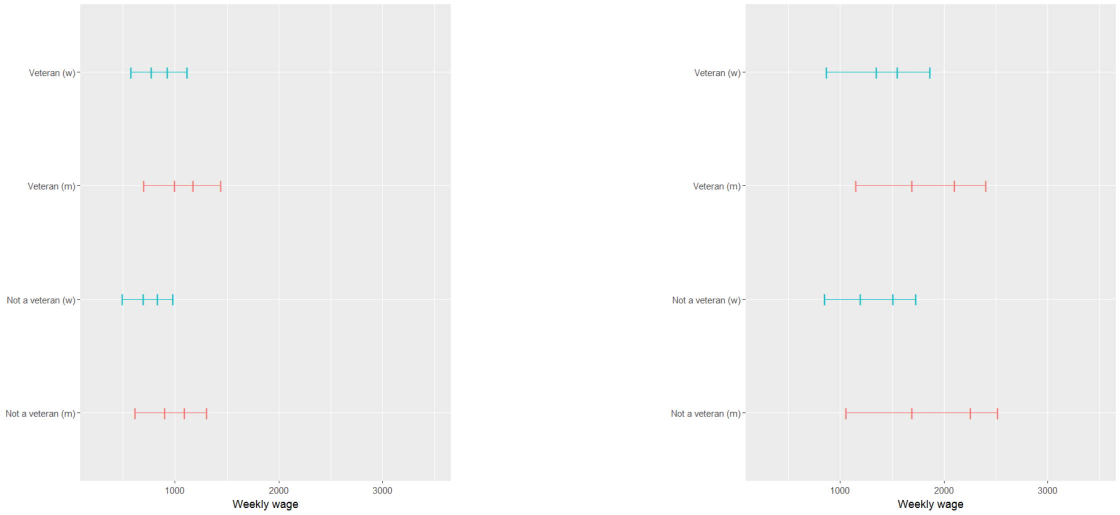
Figure 5.6: **Income according to marital status, race and ethnicity.**

Plots on the left side refer to the high school degree subgroup, plots on the right side to the bachelor's degree subgroup. The lines in a plot indicate the 0.25-quantile, the median, the mean and the 0.75-quantile of the income distribution in a category.

Panel A: English language ability



Panel B: Veteran status



Panel C: Region

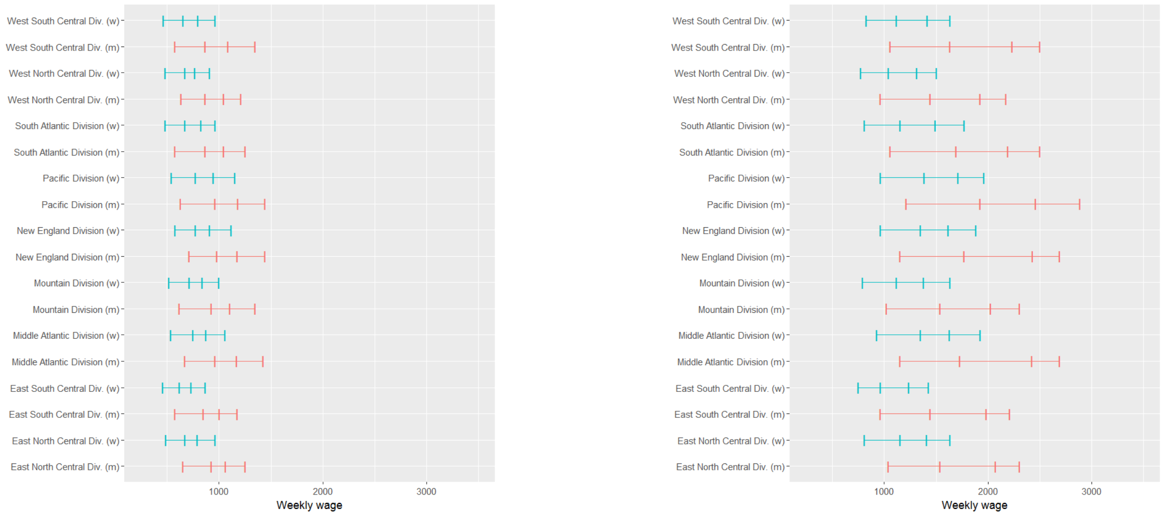
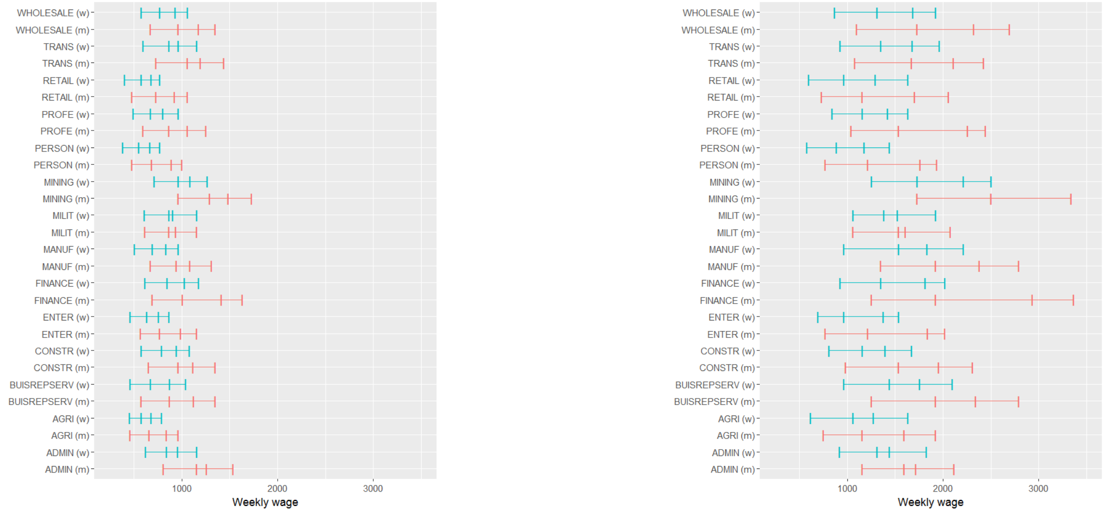


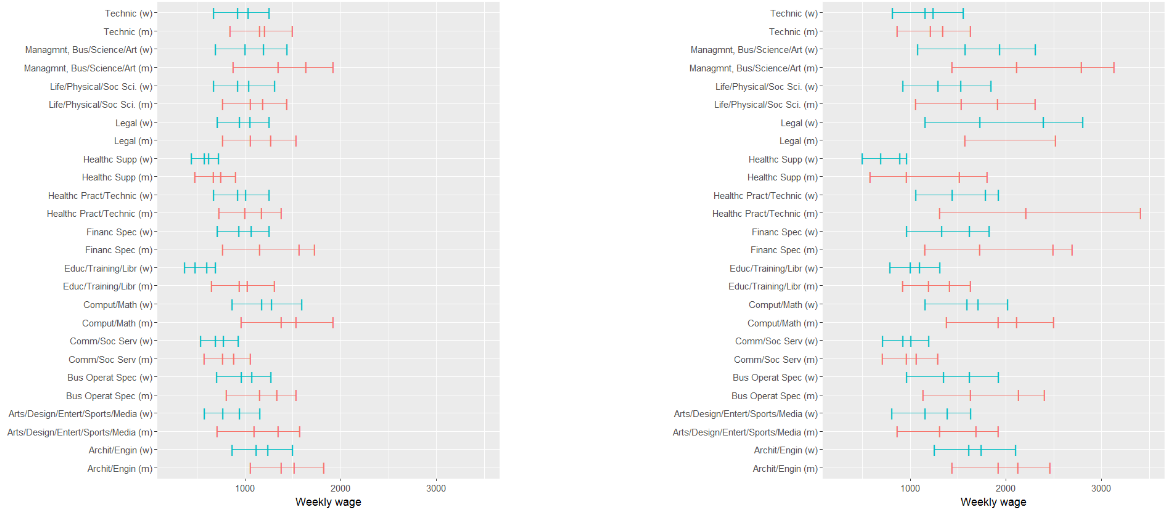
Figure 5.7: **Income according to English language ability, veteran status and region.**

Plots on the left side refer to the high school degree subgroup, plots on the right side to the bachelor's degree subgroup. The lines in a plot indicate the 0.25-quantile, the median, the mean and the 0.75-quantile of the income distribution in a category.

Panel A: Industry



Panel B: Occupation (1/2)



Panel C: Occupation (2/2)

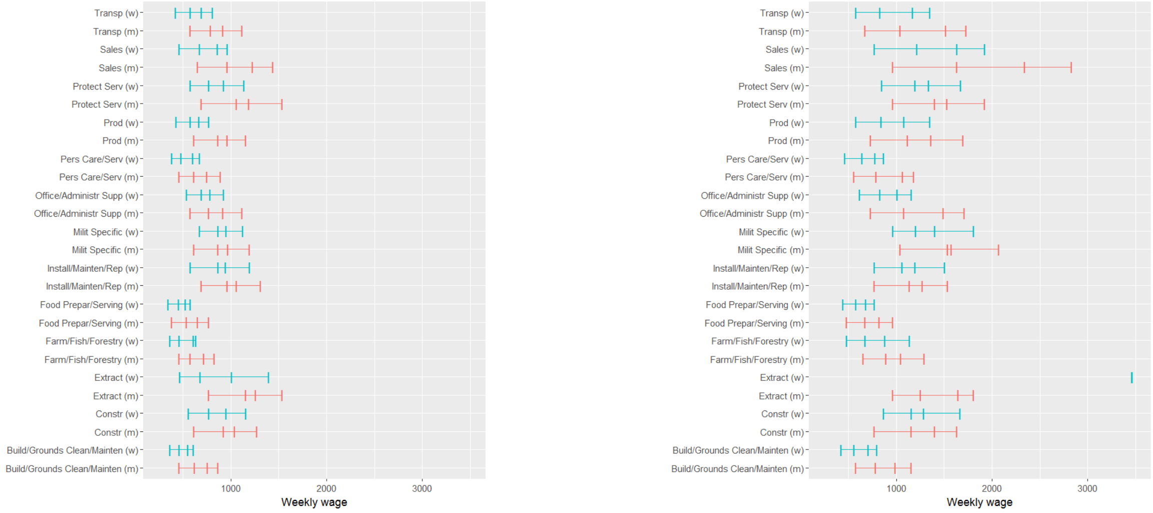
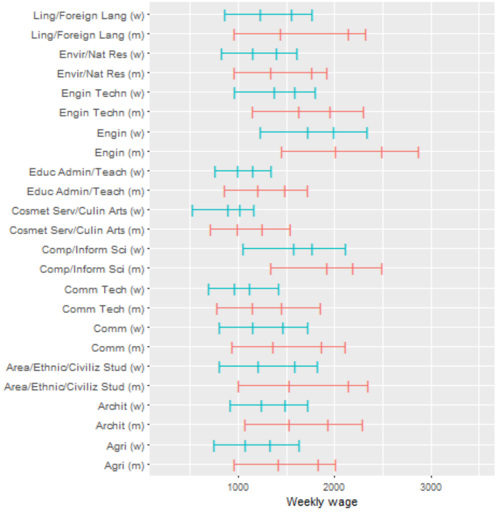


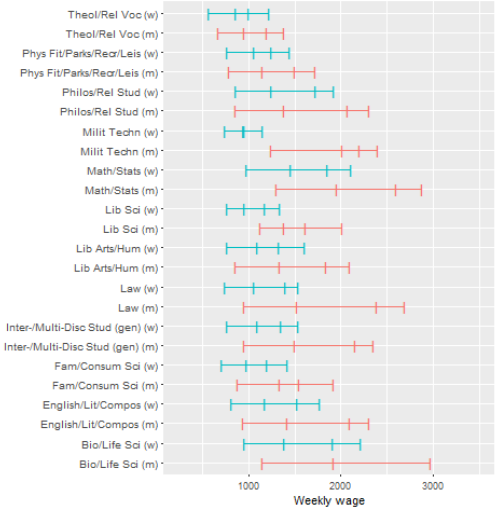
Figure 5.8: **Income according to industry and occupation.**

Plots on the left side refer to the high school degree subgroup, plots on the right side to the bachelor's degree subgroup. The lines in a plot indicate the 0.25-quantile, the median, the mean and the 0.75-quantile of the income distribution in a category.

Panel A: College major (1/3)



Panel B: College major (2/3)



Panel C: College major (3/3)

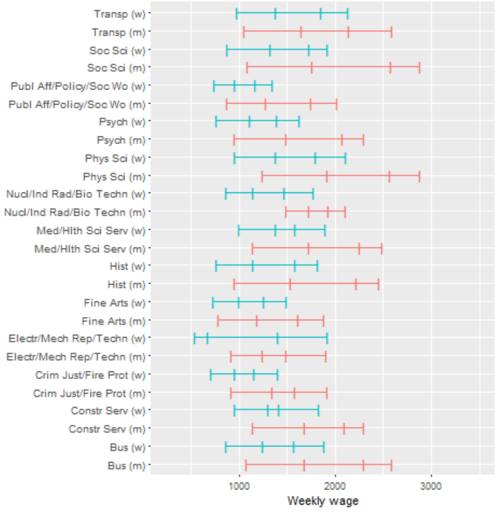


Figure 5.9: **Income according to college major.**

Plots on the left side refer to the high school degree subgroup, plots on the right side to the bachelor's degree subgroup. The lines in a plot indicate the 0.25-quantile, the median, the mean and the 0.75-quantile of the income distribution in a category.

5.7.4 Additional Results

5.7.4.1 Simultaneous Confidence Bands for Partial Effects

Figures 5.10 and 5.11 present selected effects of socio-economic variables on the magnitude of the gender wage gap with joint 0.95 confidence bands (black bounds) for the two subgroups. Effects indicate significant changes in the wage gap compared to the baseline category indicated by the vertical gray line. Baseline categories are: never married; no biological, adopted or stepchildren at home aged 4 or younger; no biological, adopted or stepchildren at home aged 18 or younger; White; not a veteran; wholesale trade (industry); management, business, science and arts (occupation); 35 to 40 hours work each week; and education administration and teaching (college major, bachelor's degree subgroup only).

Note that the results of the joint significance test do not necessarily match with those of the joint confidence bands. As explained in the main text, the test results are obtained using the Romano-Wolf stepdown procedure, which can lead to power improvements as compared to the simultaneous confidence bands. The construction of the confidence bands does not involve a stepdown procedure.

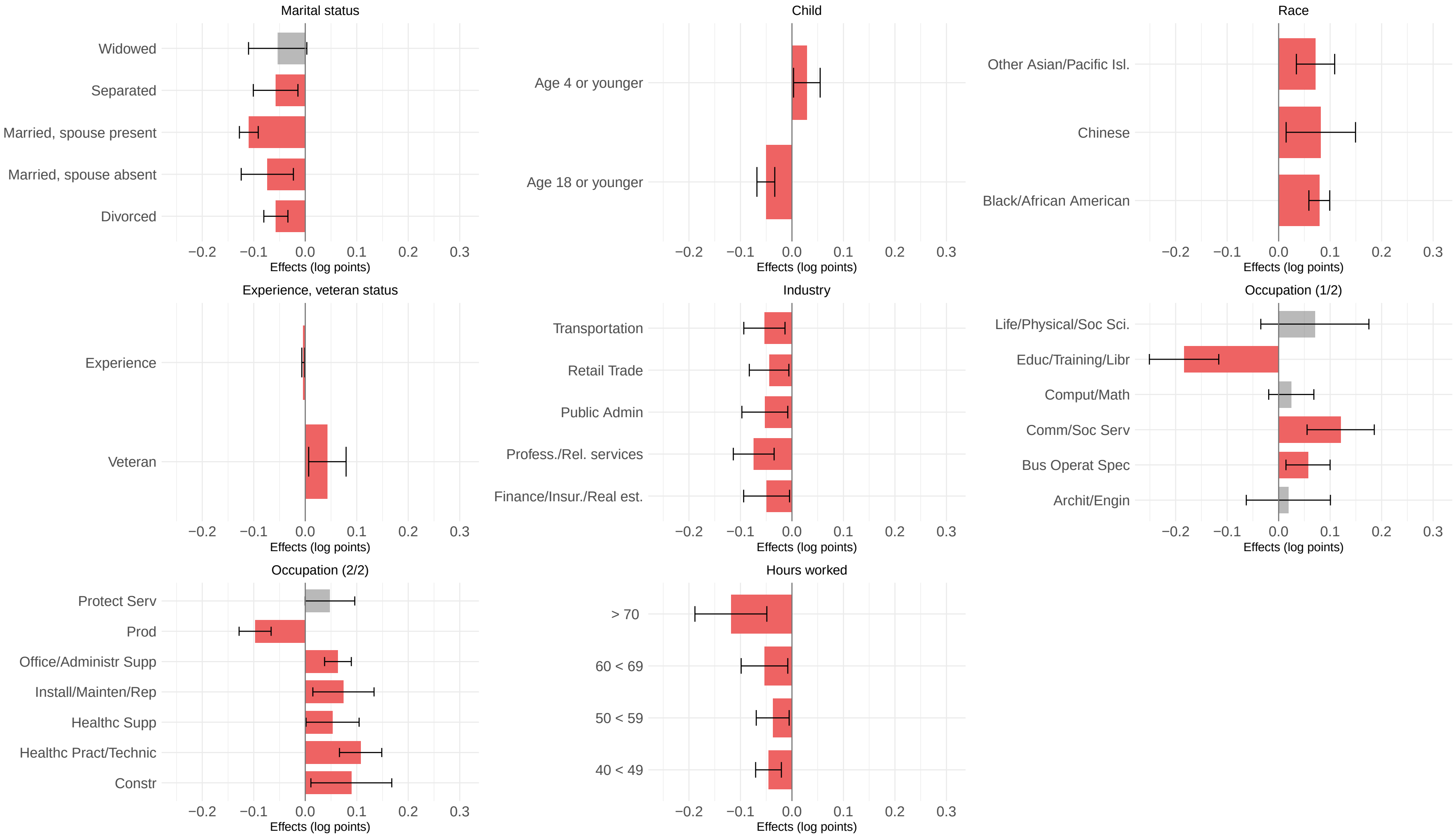


Figure 5.10: **Selected results, high school degree subgroup. Marginal effects with joint 0.95-confidence bands.**

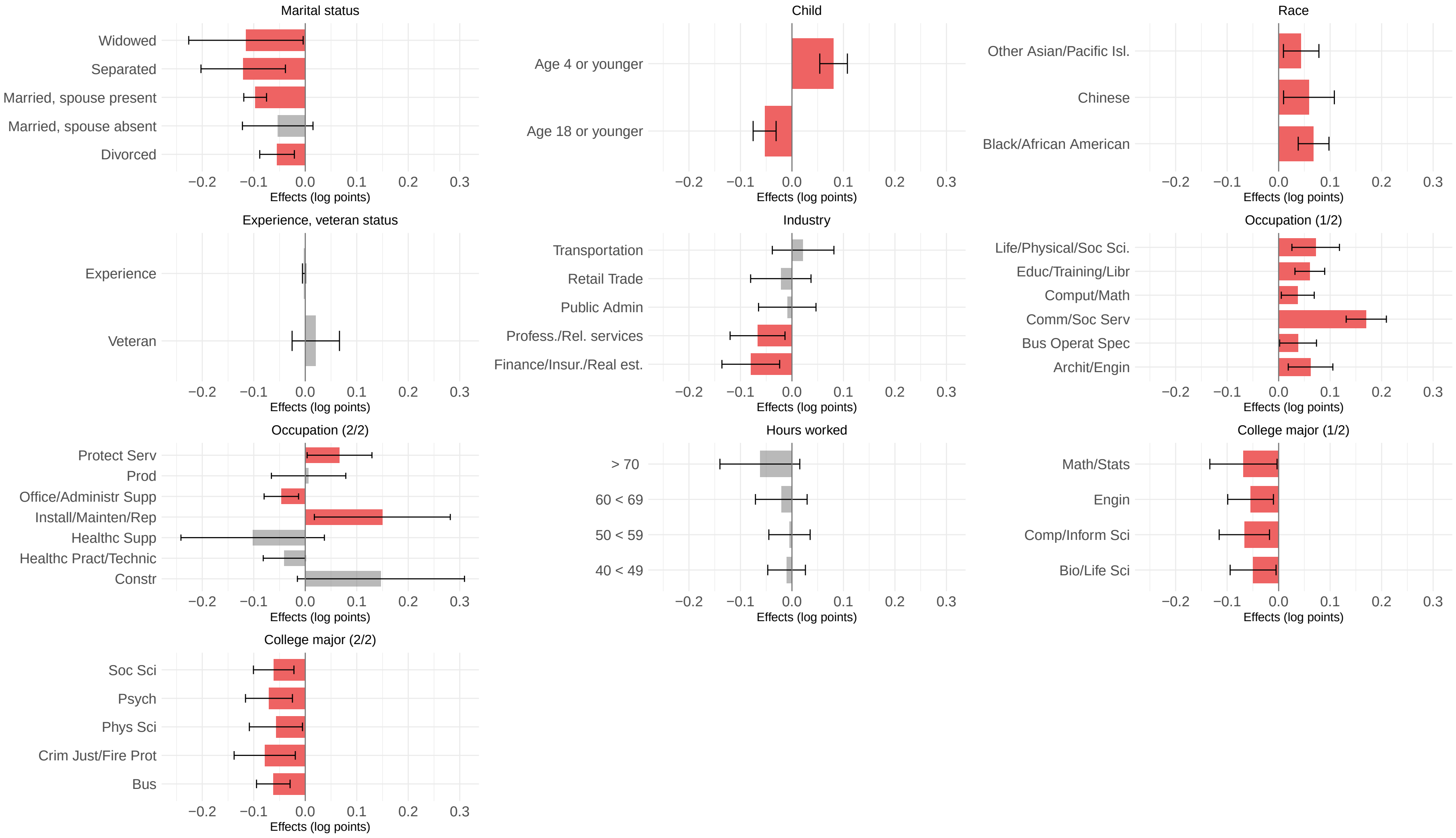


Figure 5.11: **Selected results, bachelor's degree subgroup. Marginal effects with joint 0.95-confidence bands.**

5.7.4.2 Full Result Tables

Table 5.5 to 5.10 present all estimates irrespective of their significance. p-values are obtained from a joint significance tests of all β_j with $j = 1, \ldots, p1$, coefficients in $\beta(x_i)$ from Equation (5.2) using the multiplier bootstrap procedure suggested in Belloni et al. (2014a) with 1000 repetitions.

Variable	Estimate	p-value
constant	-0.0463	0.9070
Marital status		
Married, spouse present	-0.1096	0.0000
Married, spouse absent	-0.0737	0.0010
Separated	-0.0575	0.0030
Divorced	-0.0571	0.0000
Widowed	-0.0536	0.0700
English language ability		
Does not speak English	0.0550	0.1600
Yes, speaks very well	0.0111	0.9200
Yes, speaks well	0.0172	0.8850
Yes, but not well	0.0303	0.3400
Race, ethnicity		
Black/African American/Negro	0.0789	0.0000
Chinese	0.0819	0.0100
Other Asian or Pacific Islander	0.0716	0.0000
Hispanic	0.0115	0.9200
Veteran status		
Veteran	0.0429	0.0140
Industry		
AGRI	-0.0419	0.8540
MINING	-0.0656	0.8540
CONSTR	-0.0511	0.1330
MANUF	-0.0283	0.4020
TRANS	-0.0535	0.0030
RETAIL	-0.0444	0.0150
FINANCE	-0.0493	0.0180
BUISREPSERV	-0.0433	0.0640
PERSON	-0.0384	0.3860
ENTER	-0.0281	0.9200
PROFE	-0.0742	0.0000
ADMIN	-0.0527	0.0140
MILIT	0.1145	0.2650

Table 5.5: **Complete double lasso results (1/3), high school degree data.**

Variable	Estimate	p-value
Occupation		
Bus Operat Spec	0.0571	0.0030
Financ Spec	-0.0127	0.9760
Comput/Math	0.0246	0.7830
Archit/Engin	0.0189	0.9760
Technic	0.0419	0.9200
Life/Physical/Soc Sci.	0.0703	0.5360
Comm/Soc Serv	0.1205	0.0000
Legal	0.1042	0.3400
Educ/Training/Libr	-0.1836	0.0000
Arts/Design/Entert/Sports/Media	-0.0304	0.9200
Healthc Pract/Technic	0.1075	0.0000
Healthc Supp	0.0530	0.0420
Protect Serv	0.0479	0.0510
Food Prepar/Serving	0.0290	0.2360
Build/Grounds Clean/Mainten	-0.0108	0.9710
Pers Care/Serv	0.0200	0.9200
Sales	-0.0187	0.7210
Office/Administr Supp	0.0635	0.0000
Farm/Fish/Forestry	0.0286	0.9760
Constr	0.0895	0.0150
Extract	-0.1448	0.9600
Install/Mainten/Rep	0.0742	0.0070
Prod	-0.0974	0.0000
Transp	-0.0085	0.9760
Milit Specific	0.0391	0.9760
U.S. Census region		
Middle Atlantic Division	-0.0110	0.9580
East North Central Div.	-0.0086	0.9760
West North Central Div.	-0.0065	0.9760
South Atlantic Division	0.0016	0.9820
East South Central Div.	-0.0254	0.5560
West South Central Div.	-0.0311	0.1070
Mountain Division	-0.0010	0.9820
Pacific Division	0.0200	0.6990

Table 5.6: **Complete double lasso results (2/3), high school degree data.**

Variable	Estimate	p-value
Metropolitan statistical area		
msa	0.0139	0.4120
Child		
Age 18 or younger	-0.0507	0.0000
Age 4 or younger	0.0289	0.0180
Usual hours worked per week		
40 to 49	-0.0456	0.0000
50 to 59	-0.0374	0.0150
60 to 69	-0.0534	0.0150
> 70	-0.1186	0.0000
Years of education		
yos	-0.0026	0.9200
Experience		
exp	-0.0040	0.0010
exp2	0.0000	0.3180

Table 5.7: **Complete double lasso results (3/3), high school degree data.**

Tables 5.5 to 5.14 present complete results from post-lasso estimation using double selection (double lasso) obtained for the high school degree subsample. p-values are obtained from a joint test of all β_j coefficients in $\beta(x_i)$ from Equation (5.2) using the multiplier bootstrap procedure suggested in Belloni et al. (2014a) with 1000 repetitions in combination with the stepdown procedure of Romano and Wolf (2005a).

Variable	Estimate	p-value
constant	0.0428	1.0000
Marital status		
Married, spouse present	-0.0973	0.0000
Married, spouse absent	-0.0535	0.2630
Separated	-0.1205	0.0000
Divorced	-0.0548	0.0000
Widowed	-0.1152	0.0110
English language ability		
Does not speak English	0.0221	1.0000
Yes, speaks very well	-0.0022	1.0000
Yes, speaks well	0.0392	0.3720
Yes, but not well	0.0030	1.0000
Race, ethnicity		
Black/African American/Negro	0.0679	0.0000
Chinese	0.0589	0.0020
Other Asian or Pacific Islander	0.0437	0.0010
Hispanic	0.0070	1.0000
Veteran status		
Veteran	0.0204	0.9930
Industry		
AGRI	-0.0655	0.8910
MINING	-0.0672	0.9680
CONSTR	-0.0310	0.9990
MANUF	-0.0040	1.0000
TRANS	0.0217	1.0000
RETAIL	-0.0216	1.0000
FINANCE	-0.0799	0.0000
BUISREPSERV	-0.0557	0.0450
PERSON	-0.0564	0.7250
ENTER	-0.0630	0.6300
PROFE	-0.0668	0.0010
ADMIN	-0.0091	1.0000
MILIT	0.1167	0.2040

Table 5.8: **Complete double lasso results (1/4), bachelor's degree data.**

Variable	Estimate	p-value
Occupation		
Bus Operat Spec	0.0377	0.0110
Financ Spec	-0.0348	0.0690
Comput/Math	0.0372	0.0020
Archit/Engin	0.0620	0.0000
Technic	0.1126	0.5270
Life/Physical/Soc Sci.	0.0719	0.0000
Comm/Soc Serv	0.1702	0.0000
Legal	0.0495	0.0810
Educ/Training/Libr	0.0606	0.0000
Arts/Design/Entert/Sports/Media	0.0469	0.0330
Healthc Pract/Technic	-0.0407	0.0260
Healthc Supp	-0.1022	0.3950
Protect Serv	0.0666	0.0110
Food Prepar/Serving	-0.0011	1.0000
Build/Grounds Clean/Mainten	-0.0248	1.0000
Pers Care/Serv	-0.0287	1.0000
Sales	-0.0162	0.9980
Office/Administr Supp	-0.0465	0.0000
Farm/Fish/Forestry	-0.0498	1.0000
Constr	0.1469	0.0810
Install/Mainten/Rep	0.1496	0.0020
Prod	0.0065	1.0000
Transp	0.0228	1.0000
Milit Specific	-0.0799	0.9930
U.S. census region		
Middle Atlantic Division	-0.0140	0.9980
East North Central Div.	-0.0108	1.0000
West North Central Div.	-0.0240	0.8980
South Atlantic Division	-0.0117	1.0000
East South Central Div.	-0.0374	0.2120
West South Central Div.	-0.0346	0.0810
Mountain Division	-0.0078	1.0000
Pacific Division	-0.0117	1.0000

Table 5.9: **Complete double lasso results (2/4), bachelor's degree data.**

Variable	Estimate	p-value
Metropolitan statistcal area		
msa	0.0214	0.5010
Child		
Age 18 or younger	-0.0531	0.0000
Age 4 or younger	0.0809	0.0000
Usual hours worked per week		
40 to 49	-0.0104	1.0000
50 to 59	-0.0048	1.0000
60 to 69	-0.0207	0.9980
> 70	-0.0623	0.2150
Years of education		
yos	0.0056	0.1560
Experience		
exp	-0.0024	0.2770
exp2	-0.0000	0.9270
College major		
Agri	-0.0388	0.9640
Envir/Nat Res	-0.0332	0.9980
Archit	-0.0316	0.9990
Area/Ethnic/Civiliz Stud	-0.0172	1.0000
Comm	-0.0133	1.0000
Comm Tech	-0.0173	1.0000
Comp/Inform Sci	-0.0666	0.0000
Cosmet Serv/Culin Arts	0.1138	0.9790
Engin	-0.0545	0.0010
Engin Techn	0.0357	1.0000
Ling/Foreign Lang	-0.0267	1.0000
Fam/Consum Sci	-0.0322	1.0000
Law	-0.0953	0.9650
English/Lit/Compos	-0.0140	1.0000
Lib Arts/Hum	-0.0330	0.9930
Lib Sci	-0.0594	1.0000

Table 5.10: **Complete double lasso results (3/4), bachelor's degree data.**

Variable	Estimate	p-value
College major (continued)		
Bio/Life Sci	-0.0496	0.0040
Math/Stats	-0.0683	0.0110
Milit Techn	-0.0554	1.0000
Inter-/Multi-Disc Stud (gen)	-0.0851	0.1120
Phys Fit/Parks/Recr/Leis	0.0140	1.0000
Philos/Rel Stud	0.0054	1.0000
Theol/Rel Voc	0.0224	1.0000
Phys Sci	-0.0570	0.0040
Nucl/Ind Rad/Bio Techn	0.0834	1.0000
Psych	-0.0705	0.0000
Crim Just/Fire Prot	-0.0788	0.0000
Publ Aff/Policy/Soc Wo	-0.0720	0.0670
Soc Sci	-0.0613	0.0000
Constr Serv	-0.0982	0.9830
Electr/Mech Rep/Techn	-0.1450	0.9930
Transp	0.1077	0.9510
Fine Arts	-0.0378	0.2050
Med/Hlth Sci Serv	-0.0149	1.0000
Bus	-0.0621	0.0000
Hist	-0.0561	0.0440

Table 5.11: **Complete double lasso results (4/4), bachelor's degree data.**

Tables 5.8 to 5.11 present complete results from post-lasso estimation using double selection (double lasso) obtained for the bachelor's degree subsample. p-values are obtained from a joint test of all β_j coefficients in $\beta(x_i)$ from Equation (3) using the multiplier bootstrap procedure suggested in Belloni et al. (2014a) with 1000 repetitions.

5.7.4.3 Variation of the penalty λ

As a robustness check, we repeat our analysis with the constant c used in the determination of λ set to level $c = 0.5$ implying a penalization that is less severe. Further details on the penalty choice and its implementation can be found in Chernozhukov et al. (2016a).

Variable	Estimate	p-value
constant	-0.0028	0.9980
Marital status		
Married, spouse present	-0.1065	0.0000
Married, spouse absent	-0.0730	0.0000
Separated	-0.0560	0.0010
Divorced	-0.0482	0.0000
Widowed	-0.0600	0.0150
English language ability		
Does not speak English	0.0522	0.2120
Yes, speaks very well	0.0111	0.9620
Yes, speaks well	0.0181	0.8560
Yes, but not well	0.0219	0.7850
Race, ethnicity		
Black/African American/Negro	0.0756	0.0000
Chinese	0.0837	0.0020
Other Asian or Pacific Islander	0.0645	0.0000
Hispanic	0.0154	0.6880
Veteran status		
Veteran	0.0371	0.0230
Industry		
AGRI	-0.0479	0.7320
MINING	-0.1101	0.1150
CONSTR	-0.0551	0.0710
MANUF	-0.0218	0.7620
TRANS	-0.0460	0.0110
RETAIL	-0.0362	0.0660
FINANCE	-0.0500	0.0130
BUISREPSERV	-0.0481	0.0150
PERSON	-0.0356	0.5200
ENTER	-0.0256	0.9620
PROFE	-0.0652	0.0000
ADMIN	-0.0583	0.0010
MILIT	0.0491	0.9670

Table 5.12: **Complete double lasso results, lasso with $c = 0.5$ (1/3), high school degree data.**

Variable	Estimate	p-value
Occupation		
Bus Operat Spec	0.0561	0.0010
Financ Spec	-0.0095	0.9950
Comput/Math	0.0300	0.4600
Archit/Engin	0.0174	0.9920
Technic	0.0410	0.9520
Life/Physical/Soc Sci.	0.0711	0.5260
Comm/Soc Serv	0.1179	0.0000
Legal	0.1098	0.2020
Educ/Training/Libr	-0.1813	0.0000
Arts/Design/Entert/Sports/Media	-0.0206	0.9790
Healthc Pract/Technic	0.1019	0.0000
Healthc Supp	0.0535	0.0230
Protect Serv	0.0620	0.0010
Food Prepar/Serving	0.0151	0.9520
Build/Grounds Clean/Mainten	-0.0140	0.9620
Pers Care/Serv	0.0236	0.8940
Sales	-0.0320	0.0270
Office/Administr Supp	0.0622	0.0000
Farm/Fish/Forestry	0.0249	0.9920
Constr	0.0991	0.0010
Extract	-0.0000	1.0000
Install/Mainten/Rep	0.0709	0.0020
Prod	-0.0948	0.0000
Transp	-0.0099	0.9790
Milit Specific	0.0708	0.9670
U.S. Census region		
Middle Atlantic Division	-0.0062	0.9920
East North Central Div.	-0.0042	0.9950
West North Central Div.	-0.0014	0.9980
South Atlantic Division	0.0082	0.9880
East South Central Div.	-0.0210	0.7850
West South Central Div.	-0.0236	0.4890
Mountain Division	0.0044	0.9950
Pacific Division	0.0203	0.6510

Table 5.13: **Complete double lasso results, lasso with $c = 0.5$ (2/3), high school degree data.**

Variable	Estimate	p-value
Metropolitan statistical area		
msa	0.0147	0.3090
Child		
Age 18 or younger	-0.0433	0.0000
Age 4 or younger	0.0226	0.1130
Usual hours worked per week		
40 to 49	-0.0434	0.0000
50 to 59	-0.0331	0.0170
60 to 69	-0.0530	0.0070
> 70	-0.1232	0.0000
Years of education		
yos	-0.0045	0.4260
Experience		
exp	-0.0036	0.0020
exp2	0.0001	0.0070

Table 5.14: **Complete double lasso results, lasso with $c = 0.5$ (3/3), high school degree data.**

Tables 5.12 to 5.14 present complete results from post-lasso estimation using double selection (double lasso) obtained for the high school degree subsample. p-values are obtained from a joint test of all β_j coefficients in $\beta(x_i)$ from Equation (5.2) using the multiplier bootstrap procedure suggested in Belloni et al. (2014a) with 1000 repetitions in combination with the stepdown procedure of Romano and Wolf (2005a).

Variable	Estimate	p-value
constant	0.0507	1.0000
Marital status		
Married, spouse present	-0.1011	0.0000
Married, spouse absent	-0.0590	0.1750
Separated	-0.1208	0.0000
Divorced	-0.0491	0.0000
Widowed	-0.1273	0.0050
English language ability		
Does not speak English	0.0336	1.0000
Yes, speaks very well	-0.0054	1.0000
Yes, speaks well	0.0241	0.9940
Yes, but not well	0.0192	1.0000
Race, ethnicity		
Black/African American/Negro	0.0546	0.0000
Chinese	0.0569	0.0040
Other Asian or Pacific Islander	0.0349	0.0310
Hispanic	0.0055	1.0000
Veteran status		
Veteran	0.0191	1.0000
Industry		
AGRI	-0.1086	0.1130
MINING	-0.0497	1.0000
CONSTR	-0.0352	0.9980
MANUF	0.0009	1.0000
TRANS	0.0184	1.0000
RETAIL	-0.0092	1.0000
FINANCE	-0.0798	0.0000
BUISREPSERV	-0.0464	0.2980
PERSON	-0.0547	0.8200
ENTER	-0.0742	0.3730
PROFE	-0.0621	0.0050
ADMIN	-0.0081	1.0000
MILIT	0.0629	0.9910

Table 5.15: **Complete double lasso results, lasso with $c = 0.5$ (1/4), bachelor's degree data.**

Variable	Estimate	p-value
Occupation		
Bus Operat Spec	0.0287	0.2610
Financ Spec	-0.0416	0.0070
Comput/Math	0.0260	0.2500
Archit/Engin	0.0347	0.2760
Technic	0.0798	0.9820
Life/Physical/Soc Sci.	0.0539	0.0040
Comm/Soc Serv	0.1505	0.0000
Legal	0.0214	1.0000
Educ/Training/Libr	0.0333	0.0050
Arts/Design/Entert/Sports/Media	0.0321	0.6490
Healthc Pract/Technic	-0.0590	0.0000
Healthc Supp	-0.1086	0.2780
Protect Serv	0.0636	0.0340
Food Prepar/Serving	-0.0022	1.0000
Build/Grounds Clean/Mainten	-0.0257	1.0000
Pers Care/Serv	-0.0456	0.9910
Sales	-0.0049	1.0000
Office/Administr Supp	-0.0489	0.0000
Farm/Fish/Forestry	-0.0395	1.0000
Constr	0.1254	0.3730
Install/Mainten/Rep	0.1320	0.0470
Prod	0.0055	1.0000
Transp	0.0230	1.0000
Milit Specific	0.0012	1.0000
U.S. census region		
Middle Atlantic Division	-0.0071	1.0000
East North Central Div.	-0.0100	1.0000
West North Central Div.	-0.0109	1.0000
South Atlantic Division	-0.0050	1.0000
East South Central Div.	-0.0343	0.4120
West South Central Div.	-0.0294	0.3350
Mountain Division	-0.0055	1.0000
Pacific Division	-0.0049	1.0000

Table 5.16: **Complete double lasso results, lasso with $c = 0.5$ (2/4), bachelor's degree data.**

Variable	Estimate	p-value
Metropolitan statistcal area		
msa	0.0133	0.9960
Child		
Age 18 or younger	-0.0556	0.0000
Age 4 or younger	0.0834	0.0000
Usual hours worked per week		
40 to 49	-0.0067	1.0000
50 to 59	-0.0011	1.0000
60 to 69	-0.0079	1.0000
> 70	-0.0511	0.6750
Years of education		
yos	0.0020	1.0000
Experience		
exp	-0.0042	0.0000
exp2	0.0000	1.0000
College major		
Agri	-0.0468	0.8640
Envir/Nat Res	-0.0117	1.0000
Archit	-0.0254	1.0000
Area/Ethnic/Civiliz Stud	-0.0208	1.0000
Comm	-0.0028	1.0000
Comm Tech	-0.0770	0.9900
Comp/Inform Sci	-0.0608	0.0010
Cosmet Serv/Culin Arts	0.0228	1.0000
Engin	-0.0555	0.0010
Engin Techn	-0.0261	1.0000
Ling/Foreign Lang	-0.0198	1.0000
Fam/Consum Sci	-0.0217	1.0000
Law	-0.1686	0.2140
English/Lit/Compos	-0.0028	1.0000
Lib Arts/Hum	-0.0289	1.0000
Lib Sci	-0.1427	1.0000

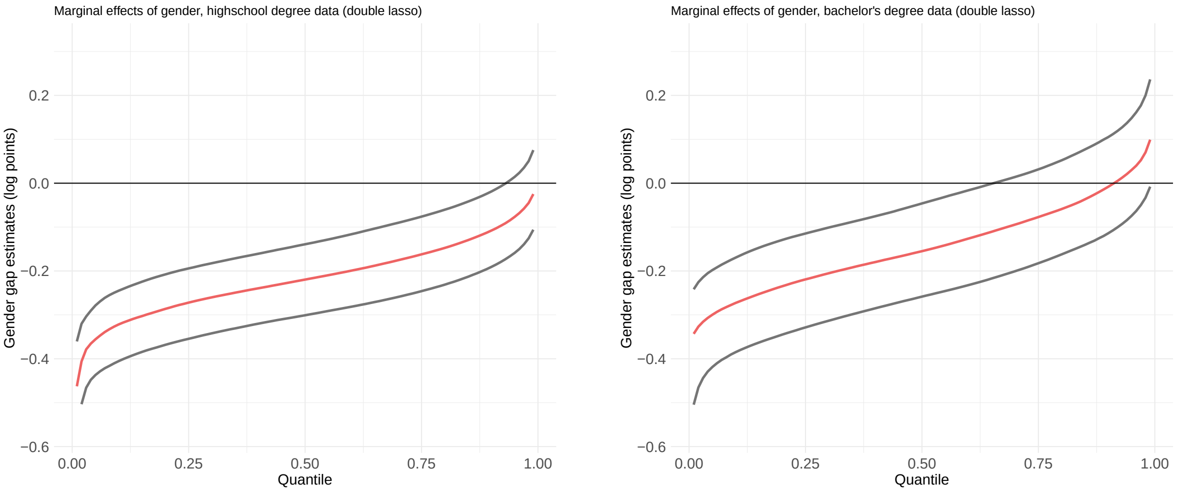
Table 5.17: **Complete double lasso results, lasso with $c = 0.5$ (3/4), bachelor's degree data.**

Variable	Estimate	p-value
College major (continued)		
Bio/Life Sci	-0.0458	0.0240
Math/Stats	-0.0588	0.1250
Milit Techn	0.1227	1.0000
Inter-/Multi-Disc Stud (gen)	-0.0867	0.1300
Phys Fit/Parks/Recr/Leis	0.0084	1.0000
Philos/Rel Stud	0.0045	1.0000
Theol/Rel Voc	0.0216	1.0000
Phys Sci	-0.0532	0.0230
Nucl/Ind Rad/Bio Techn	0.1160	1.0000
Psych	-0.0550	0.0020
Crim Just/Fire Prot	-0.0731	0.0020
Publ Aff/Policy/Soc Wo	-0.0598	0.2810
Soc Sci	-0.0565	0.0000
Constr Serv	-0.1081	0.9730
Electr/Mech Rep/Techn	-0.1754	1.0000
Transp	0.0512	1.0000
Fine Arts	-0.0303	0.6810
Med/Hlth Sci Serv	-0.0174	1.0000
Bus	-0.0633	0.0000
Hist	-0.0393	0.6520

Table 5.18: **Complete double lasso results, lasso with $c = 0.5$ (4/4), bachelor's degree data.**

Tables 5.15 to 5.18 present complete results from post-lasso estimation using double selection (double lasso) obtained for the bachelor's degree subsample. p-values are obtained from a joint test of all β_j coefficients in $\beta(x_i)$ from Equation (3) using the multiplier bootstrap procedure suggested in Belloni et al. (2014a) with 1000 repetitions.

Panel A: Quantiles of effects with corresponding confidence bounds, double lasso, $c = 0.5$.



Panel B: Quantiles of effects with corresponding confidence bounds, OLS.

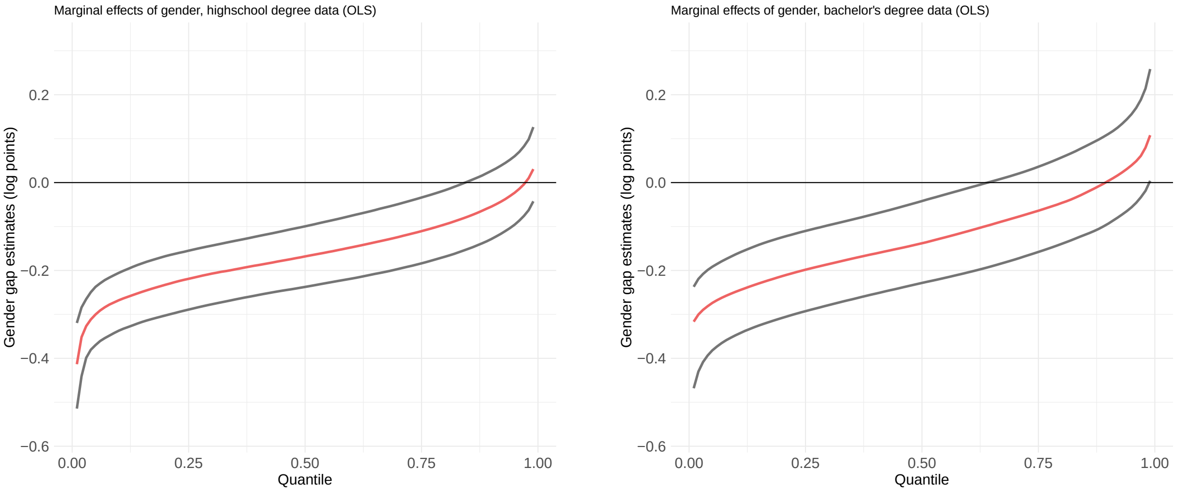


Figure 5.12: **Quantiles of effects with corresponding confidence bounds, $c = 0.5$.**

As a robustness check, the quantile plots in Figure 5.12 of the main text were reproduced with a variation of the penalty λ. The constant c that is used in the determination of λ was set to level $c = 0.5$ instead of $c = 1.1$ in the main analysis. The choice of a smaller constant c corresponds to decreasing the penalty parameter λ. Confidence bands are obtained from the multiplier bootstrap procedure with 500 repetitions.

Chapter 6

Insights from Optimal Pandemic Shielding in a Multi-Group SEIR Framework

We will never get tired of saying that the best way out of this pandemic is to take a comprehensive approach. [...] Not testing alone. Not physical distancing alone. Not contact tracing alone. Not masks alone. Do it all.

Tedros Adhamom Ghebreyesus, WHO Director-General, July 1, 2020

6.1 Introduction

The COVID-19 pandemic constitutes one of the largest threats in recent decades to the health and economic welfare of populations globally. A key challenge for policy makers everywhere is to prevent SARS-CoV-2 infections while avoiding economic losses of a magnitude that would result, in the long run, in an unacceptable level of negative effects on population health and well-being. Policy makers in most countries have reacted to the pandemic by imposing strict lockdown policies. In some countries and regions, strict lockdowns have remained in effect for many months or have been reimposed after initially being relaxed. Although such policies have slowed the spread of the virus by reducing social interactions, the more severe lockdowns have been accompanied by a large decline in economic activity. While protecting health and saving lives must, of course, take the highest priority, an optimal policy has to weigh both health and economic losses – that is, keeping mortality as low as possible, on the one hand, and mitigating an economic downturn on the other. In doing so, the goal is to identify a so-called efficient frontier – in other words, possible combinations of measures that that achieve a certain, ideally very low level of population mortality with minimal economic loss or vice versa. Once an efficient frontier for a set of different lockdown strategies has been constructed, a comparison of these strategies allows policy makers to achieve efficiency gains. The point on the efficient frontier that is considered desirable is a decision that must be made by policy makers and, ideally, society as a whole.

In a recent contribution, Acemoglu et al. (2020) extend the classical SIR model, which is well-known from the epidemiological literature, by explicitly incorporating the trade-off that policy makers must consider in times of the pandemic. The authors derive the efficient frontier for different policies and show that efficiency gains can be achieved by targeting lockdown policies at different age groups, each of which is, in turn, characterized by different productivity and mortality risks.[1] In a setting calibrated to the U.S. population and economy, they show that protecting the most vulnerable group (i.e., those aged 65 and older) with stricter shielding rules (i.e., targeted shielding) is associated with fewer losses than a blanket shielding policy (also referred to as a uniform shielding, i.e., a policy that applies equally to all groups). Acemoglu et al. (2020) briefly mention and discuss a potential extension of the multi-group SIR model to the SEIR case. Here, we continue their analysis and analyze a variety of policy measures within the SEIR model. We explicitly state the key equations of this model and calibrate it to social interaction patterns as estimated in Klepac et al. (2020).

In this paper, we consider a model that is calibrated to Germany – that is, we adjust it to the country's demographic and economic characteristics, as well as its system of health care provision. Germany and the U.S. differ in many regards, such as the demographic structure of the population, age-specific employment and income patterns, and the capacities of the health system. We present the results of the model and discuss various policy measures, such as group distancing, test strategies, contact tracing, and combinations of these. We also discuss in detail how a targeted policy, protecting vulnerable groups like old people, might be implemented in practice and discuss some policy examples.

Mortality from COVID-19 is particularly high among older people, Ferguson et al. (2020), whose productivity is relatively low. Hence, a targeted shielding policy that limits face-to-face contacts with persons aged 65 or older might lead to lower mortality in this population group and less damage to the economy. Additionally, a set of potentially voluntary policies that reduce transmission rates and social contacts could, in principle, be considered as an alternative to age-targeted shielding. Indeed, in our analysis, we find that testing, contact tracing, group distancing and improved conditions for working from home help to reduce the economic costs of the pandemic and the intensity and duration of age-targeted shielding. Moreover, if these measures are combined in a comprehensive approach as described in the initial quote by Tedros Adhamom Ghebreyesus, population mortality and economic outcomes improve substantially. Throughout our analysis, the efficiency gains associated with age-targeting remain relatively stable and sizable, and we recommend exploiting these gains by improving conditions for individuals at high risk, for example by providing services such as special shopping or consultation hours for older people, as well as testing capacities for those who have contact with high-risk groups to decrease the probability of infections.

The rest of this paper is structured as follows: In Section 6.2 we briefly introduce the multi-group SEIR model. In Section 6.3 we describe our specification of the parameters for the SEIR model for Germany. Section 6.4 presents the results and describes the optimal policies comprising measures such as group distancing, testing, contact tracing and improved medical treatment. Finally, a conclusion summarizes the results and makes a range of policy recommendations.

Because there is still so much that we do not know about SARS-CoV-2, including the transmission rate, mortality rates and aspects related to immunity, all of the results reported throughout the paper must be interpreted with caution. As in the study by Acemoglu et al. (2020), we do not focus on presenting absolute quantitative results, such as GDP forecasts, but rather qualitative insights into potential policy measures that are considered in variation-of-parameters analyses.

Literature review

The classical SIR and SEIR models are used widely in epidemiology and described in many standard textbooks. Driven by the COVID-19 crisis, various extensions of the standard epidemiological models have been developed and modified to consider economic factors. For example, Brotherhood et al. (2020)

---

[1] Acemoglu et al. (2020) employ the term "lockdown" to denote policies that limit social interactions, such as leisure activities or face-to-face interactions at work. In the following, we will refer to these policies as "shielding" measures to underscore the underlying concept of protecting people with higher mortality risks due to higher age or comorbidities.

include individual choices about the amount of time spent on activities outside the house, such as work or consumption, to the standard SIR epidemiological model. These activities are associated with externalities, i.e., higher risk of transmission to and from others. The model also incorporates heterogeneity in terms of age and different policy measures, such as testing or quarantines. Berger et al. (2020) provide an extended SEIR model focusing on testing and quarantine measures and thereby explicitly address the imperfect information that arises due to the fact that cases can be symptomatic or asymptomatic. A recent study by Grimm et al. (2020) extends a classical SEIR model by introducing a high and low risk group that differ, for example, in hospitalization and mortality rates. Their study focuses on the evolution of infected, recovered and deceased, i.e., the epidemiological aspects of the SEIR model in a parametrization calibrated to Germany. While a blanket shielding policy (i.e., for the entire population) is, of course, the optimal way to protect everyone from infection, the associated economic losses might become substantial. The multi-group SEIR model incorporates economic costs that arise due to sick leave, productivity losses when individuals work from home and discounted lifetime income losses from deaths due to COVID-19. Moreover, important indirect health consequences are associated with strict shielding measures, such as missed appointments for other conditions, less exercise, mental health issues, increased alcohol consumption, social isolation and increased levels of domestic abuse. While these indirect, non-pecuniary costs are not incorporated in our study, it might be useful to model them in future work.

We build on the work of Acemoglu et al. (2020), who study targeted shielding policies in a multi-group SIR model, and thereby address the trade-off between mortality and economic losses. They consider two possible targeting strategies: finding separate, optimal shielding policies for the young, middle-aged and senior groups (the so-called “fully targeted” policy) or imposing two separate shielding policies, one for the senior group and the other for the young and the middle-aged (so-called “semi-targeted” shielding). In their baseline results, semi-targeted policies are associated with substantial efficiency gains that cannot be improved substantially by fully targeted policies.

While Acemoglu et al. (2020) analyze the optimal policy for the U.S., we extend their framework and calibrate it to Germany. Our baseline model is a SEIR model that incorporates contact patterns as estimated by Klepac et al. (2020), who evaluate data from the BBC pandemic project in 2017 and 2018. Moreover, we consider a broader set of policy measures, such as testing and contact tracing, as well as various forms of group distancing.

6.2 Multi-Group SEIR Model

In this section, we briefly describe a SEIR model based on Acemoglu et al. (2020), who focus in their analysis on the SIR model and state that their conclusions also hold for the SEIR version. For an in-depth discussion with additional information on the theoretical set up of the original SIR model, we refer to Acemoglu et al. (2020). One of the major features of the framework is that it allows the population to be partitioned into subgroups that are heterogeneous in terms of their productivity and mortality rates. In particular, we consider the following three subgroups: young (20-49 years), middle-aged (50-64 years) and senior citizens (65+ years). Accordingly, there are age-group specific compartments for susceptible (S_j), infectious (I_j), recovered (R_j) and deceased (D_j) persons, with $j = y, m, s$ referring to the young, middle-aged and senior groups. The epidemiological SEIR model extends the SIR model by the compartment of exposed individuals $-$ that is, those who have been infected by the virus but whose infection is not yet sufficiently severe that they have symptoms or are infectious. Hence, the model considered in the following incorporates a compartment E_j for each age group in addition to compartments S_j, I_j, and R_j

at each point in time $t \in [0, \infty)$.

$$S_j(t) + E_j(t) + I_j(t) + R_j(t) + D_j(t) = N_j.$$

N_j is the number of initial members in each group, $j = y, m, s$. The compartment structure of a two-group SEIR model is illustrated in Figure 6.1 with the red arrows indicating the paths of transmissions through contacts of infectious and susceptible.

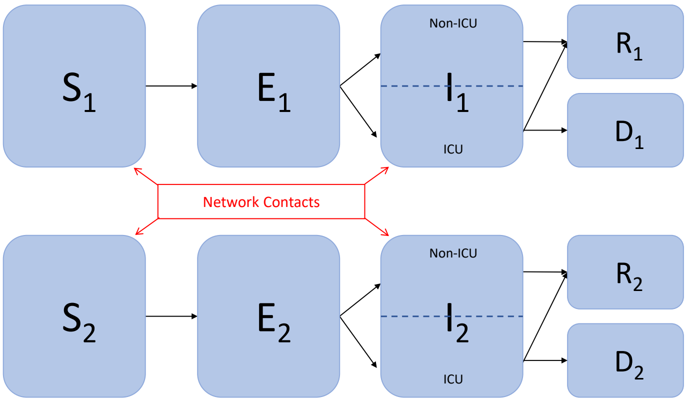


Figure 6.1: **Compartments in a two-group SEIR model.**

The red arrows illustrate the potential channels of infections through physical contacts.

Without any policy intervention that enforces shielding of the population or isolation of those who are infected, the (gross) number of new infections in the segment of exposed (E_j) and infectious (I_j) is governed by the following equations

$$\text{New exposed in group } j = M_j(S, E, I, R; \alpha) \cdot \beta \cdot S_j \cdot \sum_k \rho_{jk} I_k \tag{6.1}$$

$$\text{New infected in group } j = \gamma_j^I \cdot E_j, \tag{6.2}$$

where $\{\rho_{jk}\}$ are parameters for the contact rate between group j and k and $M_j(\cdot)$ refers to a matching technology, with $M_j(\cdot) = 1$ if $\alpha = 2$ which is our baseline case. The parameter β denotes the transmission rate from contacts between individuals in I_j and S_j and γ_j^E is the exit rate from the latent state to the infectious state.

6.2.1 Model Assumptions

In this section we describe and discuss the model assumptions.

Infection, ICU, Fatality and Recovery

In the SEIR model described above, a transmission of SARS-CoV-2 arises through contact of susceptible individuals with infectious individuals. After an average latent period $\frac{1}{\gamma_j^E}$, they become infectious themselves. Individuals in compartment I_j may require ICU care. We assume for simplicity that a need for ICU is apparent immediately after entering state I_j. ICU patients either recover with Poisson rate δ_j^r or die at Poisson rate δ_j^d. Non-ICU patients will always recover at Poisson rate γ_j^I. The death rate can vary with total ICU needs relative to capacity. We assume that

$$\gamma_j = \delta_j^d(t) + \delta_j^r(t).$$

This means that the proportions of ICU and non-ICU patients among the infected do not change over time in group j. $H_j(t)$ denotes the number of individuals needing ICU care at time t in group j, so that $H_j(t) = \iota_j I_j(t)$. $H(t) = \sum_j H_j(t)$ is the total need for ICU. The probability of death is a non-decreasing function of the number of patients, such that the probability of death will rise if the capacity is exceeded:

$$\delta_j^d(t) = \psi_j(H(t)),$$

for a given non-decreasing function ψ_j.

Testing, Contact Tracing and Isolation

Detection and isolation of infected individuals is not perfect. In the SIR model, Acemoglu et al. (2020) denote the probability that an individual in compartment I_j is not detected and put in isolation by η_j. In their analysis, comparative statics are performed to illustrate the consequences of variation of η_j, for example due to intensified testing. Incorporating the group of exposed (E_j) in a SEIR model allows tests to be performed for those who have had contact with an infected person. This setting could be considered a simplified form of contact tracing, for instance enabled by a smartphone application that records physical contacts. Hence, quarantining those who have been in contact with infected individuals might enable policy makers to exclude these infected but not yet infectious individuals from social interactions. Accordingly, we denote the probability that a person in compartment E_j or I_j is not detected and isolated by η_j^E and η_j^I, respectively, and thereby avoid including additional state variables. In this manner, we can model the fact that only those infected who have not been detected and isolated in stage E_j or I_j contribute to the spread of the disease via their contacts.

Shielding and Physical Distancing

Shielding policies describe all measures that reduce the rate of transmission of infections in social and business life and physical distancing. The productivity of members of j is w_j without shielding and $\xi_j w_j$ with shielding, with $\xi_j \in [0, 1]$. $L_j(t) = 1$ refers to a full shielding policy and $L_j(t) = 0$ to a situation without any restrictions to social interactions. $L_j(t) \in (0, 1)$ would be partial shielding, for example by shielding a (potentially randomly and independently drawn) fraction of the population. It is assumed that shielding cannot be perfectly enforced and that, with shielding, the effective reduction in social interaction is only $1 - \theta_j L_j(t)$ with $\theta_j < 1$.

Contact Rates

We implement a version of the SEIR model that incorporates social interaction patterns to capture the major findings in Klepac et al. (2020) – that is, high rates of interaction within and by the group of young and decreasing intensity of interactions with age. The study evaluates large-scale data on the frequency and intensity of social interactions that were collected in the BBC Pandemic project in the UK in 2017 and 2018 and make it possible to derive age-specific contact rates. To model the group interaction within and between groups, let denote ρ^0_{jk} the elements of the contact matrix

$$\rho^0 = \begin{pmatrix} 1.0 & 0.5 & 0.4 \\ 0.5 & 0.6 & 0.4 \\ 0.4 & 0.4 & 0.5 \end{pmatrix},$$

with the first row and column referring to the young group, the middle row and column referring to the middle-aged group and the third row and column referring to the senior citizen group.[2] The contact estimates of Klepac et al. (2020) refer to a pre-pandemic setting and, hence, constitute the benchmark scenario for comparison to social distancing policies. To incorporate voluntary reductions of physical contacts, we base our baseline results in Section 6.4 on a rescaled contact matrix that presumes a 25% reduction in physical contacts.

$$\rho = 0.75 \cdot \rho^0 = \begin{pmatrix} 0.750 & 0.375 & 0.300 \\ 0.375 & 0.450 & 0.300 \\ 0.300 & 0.300 & 0.375 \end{pmatrix}, \tag{6.3}$$

Incorporating more realistic contact patterns in the SEIR model with multiple groups is important for evaluating policy measures that are targeted at different age groups. For example, lower rates of contact between the vulnerable group (i.e., senior citizens) and younger people might allow for less intense shielding patterns.

Physical Distancing, Face Masks and Additional Hygiene Measures

Various mandatory or voluntary policies can be employed to reduce the transmission rate of SARS-CoV-2. These measures range from a general reduction in face-to-face or physical contacts (for example, by imposing strict physical distancing measures that apply equally to all age groups) or specific interventions that aim to protect especially those who are most vulnerable. The latter include, for example, a reduction in face-to-face contacts with senior citizens – for instance by placing restrictions on visits to nursing homes or prescribing mandatory (reusable or disposable) face masks during for contacts with senior citizens. For example, Chu et al. (2020) undertook a systematic review and meta-analysis of studies that examined the effectiveness of face masks and physical distancing for COVID-19 and related diseases (e.g., MERS and SARS). Accordingly lower transmission rates are associated with greater physical distance and the use of N95 face masks and comparable respirators rather than disposable surgical masks. There are a huge number of potential policy measures that aim to reduce the transmission of SARS-CoV-2, all of which can be employed in combination. We list a few examples of such measures in Section 6.4.3. Something that all of these measures have in common is that they effectively change or rescale the elements in the contact matrix ρ. In our analysis, we focus mainly on two variants of group distancing, namely (i) so-called uniform group distancing, which effectively reduces the contact rates in ρ for all groups (corresponding to a multiplication of the matrix (corresponding to a multiplication of the matrix ρ with a scalar ν), and (ii) group distancing policies with a focus on the vulnerable that refer only to interactions with the group of seniors and the elements ρ_{sj} with $j = y, m, s$, and ρ_{js}, respectively. Moreover, it is possible to simulate settings in which the level of interactions within the senior group might be left unchanged, thus reducing the impact on daily interactions with others at the same age.

---

[2] An example: The entry of a contact matrix ρ_{23} represents the contact rate that applies to interactions of members of the middle-aged and the senior age group, i.e., ρ_{ms}. Due to symmetry of the matrix, it holds that $\rho_{23} = \rho_{32}$.

Improved Conditions for Working from Home

Working from home can be an effective way to reduce the costs of the pandemic and of shielding policies. To host a scenario with improved conditions for working from home, we (i) implement a parameter constellation with respect to the contact rates within and between the young and middle-aged group and between these groups and the senior group and (ii) decrease the productivity loss associated with working from home, ξ_j.We believe that this captures some aspects of working from home in that those who are most likely to be employed can reduce their social interactions with lower economic losses. Changes in terms of (i) are imposed by scaling the entries of the contact matrix $\rho_{yy}, \rho_{mm}, \rho_{ym}$ by a factor $1-\pi_1$ and a scaling the contact rates ρ_{ys} and ρ_{ms} by $1-\pi_2$ with $\pi_1 < \pi_2$.

Vaccine and Cure

Acemoglu et al. (2020) assume that a vaccine and an effective drug for all infected individuals becomes available at some date T and that full immunity is achieved and maintained after an infection.[3] In our analysis, we will evaluate changes in T resulting from a faster development of a vaccine - for example after one year or six months.

Currently, there are various treatments for COVID-19 that have been approved or are being evaluated in clinical trials. We assess the implications of a medical treatment with respect to the optimal shielding policy. Put simply, a new treatment could have any of the following three effects: (i) reduce the length of hospitalization, (ii) reduce the probability of dying from COVID-19, (iii) reduce the probability that an infection with SARS-CoV-2 becomes severe. We will focus on the availability of a treatment that leads to a reduction in mortality from COVID-19 for the group of senior citizens because most deaths and severe cases have been observed in this age group (e.g., as reported for Germany in RKI (2020b)).

6.2.2 Dynamics in the MG-SEIR Model

If vaccine and cure are unavailable, the number of individuals in the exposed compartment for group j evolves according to the differential equations for all $t \in (0,T)$

$$\dot{E}_j = M_j(S,E,I,R,L;\alpha)\beta(1-\theta_j L_j)S_j \sum_k \rho_{jk}\eta_k^E\eta_k^I(1-\theta_k L_k)I_k - \gamma_j^E E_j,$$

for nonnegative β and contact coefficients ρ_{jk} and where

$$M_j(S,E,I,R,L;\alpha) \equiv \left(\sum_k \rho_{jk}\left[(S_k+\eta_k^E E_k+\eta_k^E\eta_k^I I_k+(1-\kappa_k)R_k)(1-\theta_j L_k)+\kappa_k R_k\right]\right)^{\alpha-2}.$$

In the quadratic case $M_j(S,E,I,R,L)=1$. The parameter κ_j refers to the share of recovered individuals that can return to work and social life while being exempted from shielding policies due to immunity.[4] Setting $\eta_j^E = 1$ for all j refers to a setting where it is not possible to test and isolate exposed individuals. However, a value $\eta_j^E < 1$ means that the effective number of individuals who contribute to further spread of the disease can be reduced by contact tracing and isolating those who have been exposed.

The rest of the laws of motion for $t \in (0,T)$ are

$$\dot{S}_j \quad = \quad -\dot{E}_j - \gamma_j^E E_j, \tag{6.4}$$

---

[3] In line with Acemoglu et al. (2020) we focus on the case with deterministic arrival of a vaccine.

[4] We acknowledge that there is not yet a consensus on whether individuals become immune to SARS-CoV-2 after an infection and whether such immunity, if achieved, is maintained for a substantial period. The empirical evidence on both points is mixed. We follow the baseline setting in Acemoglu et al. (2020) with $\kappa_j = 1$ for all j and repeat the robustness checks with setting $\kappa_j = 0$ for all j. The main conclusion remains valid and results are omitted for the sake of brevity.

$$\begin{aligned}\dot{I}_j &= \gamma_j^E E_k - \gamma_j^I I_j, &(6.5)\\ \dot{D}_j &= \delta_j^d H_j, &(6.6)\\ \dot{R}_j &= \delta_j^r H_j + \gamma_j^I (I_j - H_j), &(6.7)\end{aligned}$$

where $H_j = \iota_j I_j$ denotes the number of ICU patients in group j. After discovery of a vaccine and cure at T, every individual alive is in the recovered category.

6.2.3 Efficient Frontier

The government can control the degree of shielding $L_j(t)$ for each group j at any point in time $t \in [0, T)$. In particular we will compare uniform policies (i.e., blanket policies with $L_j(t) = L(t)$) and group-specific (i.e., targeted) policies. The goal of the social planner is to minimize the overall costs of the pandemic, which consist of two parts:

1. Lives Lost $= \sum_j D_j(T)$.
2. Economic Losses $= \int_0^T \sum_j \Psi_j(t)dt$.

The economic losses for group j are given by

$$\begin{aligned}\Psi_j(t) &= (1-\xi_j)w_j S_j(t) L_j(t) + &(6.8)\\ &+ (1-\xi_j)w_j E_j(t)(1-\eta_k^E(1-L_j(t))) + \\ &+ (1-\xi_j)w_j I_j(t)(1-\eta_k^E\eta_k^I(1-L_j(t))) + \\ &+ (1-\xi_j)w_j(1-\kappa_j)R_j(t)L_j(t) + \\ &+ w_j\Delta_j\iota_j\delta_j^d(t)I_j(t),\end{aligned}$$

where the second term refers to the income loss of exposed individuals under shielding. The third term in the economic cost function is now adjusted to the case with the testing and isolation of exposed individuals, as well. Δ_j captures the present discounted value of a group j member's remaining employment time until retirement, which is lost due to death. The objective function is a weighted sum of both losses with weight factor χ and the task is to choose a shielding policy which minimizes

$$\int_0^T \sum_j \Psi_j(t)dt + \chi \sum_j D_j(T).$$

Varying the values for χ makes it possible to identify the efficient frontier - in other words, to find the policy that minimizes the objective function for a given χ. Hence, the policy recommendations that can be obtained from an analysis of the efficient frontiers do not depend heavily on a specific choice of χ but rather reflect the difficult trade-off that policy makers face in the pandemic (Acemoglu et al., 2020).

6.3 Specification and Calibration

Before we discuss optimal shielding policies in the multi-group SEIR model, we will first comment on how we set and calibrated the parameters for Germany. We will present adaptations of country-specific parameters that would also apply to a calibration of the initial multi-group SIR model in Acemoglu et al. (2020). These parameters refer to demographic and economic conditions, as well as to characteristics of health care provision in Germany. Second, we will discuss the adaptations of the SIR model parameters to a SEIR version based on information from the Robert Koch Institute (RKI) as of July 2020. Finally, we will comment on the calibration of the basic reproduction number R_0.

Parameter	US	GER
	1. Socio-demographic and economic	
Population shares $\{N_y, N_m, N_o\}$	{0.53, 0.26, 0.21}	{0.46, 0.28, 0.26}
Per capita income $\{\omega_y, \omega_m, \omega_o\}$	{1.00, 1.00, 0.26}	{1.00, 1.00, 0.085}
Remaining years in empl. $\{\Delta_y, \Delta_m, \Delta_o\}$	{32.50, 10.00, 2.50}	{32.43, 10.44, 2.50}
	2. Health-care related	
Mortality penalty λ	1.00	$\{0.20, 0.40, \underline{0.60}, 0.80\}$
ICU constraint $\bar{H}(t)$	{0.016, 0.020}	{0.020, 0.030, 0.040}

Table 6.1: **Parameters for the United States and Germany.**

The underlined mortality parameter indicates the choice in the baseline setting. The remaining values for λ and $\bar{H}(t)$ are used in robustness checks.

6.3.1 Country-Specific Parameters

Calibration of Socio-Demographic and Economic Parameters

Germany has a demographic com- position that is substantially different from that of the U.S. In particular, the share of the group aged 65 and older is larger and that of the young group is smaller than in the U.S. For example, the median age in the U.S. is around 38 (United States Census Bureau, 2019) years whereas it is around 45 years in Germany (Bundeszentrale für politische Bildung und Statistisches Bundesamt, 2018). Using data from German micro census from 2018 as provided by the German Federal Statistical Office (Statistisches Bundesamt, 2019; Statistisches Bundesamt, 2020), we calculated the remaining lifetime earnings as displayed in Table 6.1 assuming retirement at age 67.

An interesting difference that we observed in the comparison of Germany and the U.S. is the distinct employment patterns in the group aged 65 and above. Whereas approximately 20% of individuals in this group are still employed in the U.S., the corresponding share for Germany amounts only to around 7%, leading to the re-weighted per-capita earnings in Table 6.1. In both countries, the median earnings are relatively similar for those who are employed in the middle-aged group and the senior groups.

The demographic distribution of the population in Germany implies that the share of persons who have a higher risk of dying from COVID-19 is relatively large. Thus, uniform shielding policies that aim to keep mortality in the entire population at a low level are expected to be more costly in terms of economic damage. At the same time, the group of senior citizens accounts for a relatively low share of GDP, implying that targeted policies are more favorable. Shielding targeted only towards the elderly therefore makes it possible to reduce overall mortality while allowing the younger and economically more productive groups to continue working.

Calibration of Health and Medical Variables

Calibrating the model in terms of parameters that are related to health care provision is challenging - for example due to limited comparability of hospital capacities and their dynamic expansion in reaction to the pandemic (Organisation for Economic Cooperation and Development (OECD), 2020). We performed various variations to parameters of the original SIR model of Acemoglu et al. (2020) and its SEIR version

	US	GER
Acute Care beds/10,000 pop. (OECD, 2020)	24.00	60.00
ICU beds/10,000 pop. (OECD, 2020)	2.58	3.39
ICU beds/10,000 pop. (AHA, 2020, DIVI, 2020)	3.13	3.89

Table 6.2: **Hospital beds and ICU beds.**

Source: OECD (2020), AHA (2020), DIVI (2020). AHA (2020) refers to adult ICU beds and population only.

which are provided, in part, in the Appendix and chose one of these parameter configurations as a baseline setting in our analysis as described in the following.

Health care provision in Germany is considerably different from that in the U.S. In a recent report, the Organisation for Economic Cooperation and Development (OECD, 2020) compares health care provision across different countries. We list the numbers of hospital and ICU beds for the U.S. and Germany in Table 6.2. Due to the dynamic expansion of hospital capacities during the COVID-19 pandemic in both countries, we add more recent, constantly updated data from the American Hospital Association (AHA, 2020) and the German Interdisciplinary Association for Intensive and Emergency Medicine (DIVI, 2020). Compared to the OECD data, the number of ICU beds reported by AHA and DIVI has increased by around 21% in the U.S. and around 15% in Germany. The number of ICU beds is frequently reported to be one of the crucial measures of whether countries are able to keep mortality from COVID-19 low. According to the report by OECD (2020), Germany is the country with the highest ICU capacities among all OECD members. Germany not only has more ICU beds per capita than the U.S.; other measures, such as the number of hospital beds or coverage with public health insurance (OECD, 2020), suggest that the health care system in Germany has comparably greater capacities (per capita) than that of the U.S.

To take account of these differences, we adjust the parameter λ, which enters the relationship of the daily mortality rate, δ_j^d, and hospital capacities at time t, $H(t)$, to a default value $\lambda = 0.6$, which is smaller than $\lambda = 1$ as chosen in the analysis of Acemoglu et al. (2020).

$$\delta_j^d = \underline{\delta}_j^d \cdot [1 + \lambda \cdot H(t)], \tag{6.9}$$

where $\bar{\delta}_j^d$ is the baseline mortality rate for group j with $\delta_y = 0.001$, $\delta_m = 0.01$ and $\delta_s = 0.06$.[5] We refer to Figure 6.18 in the Appendix for illustrations of the variation of health-provision-related parameters. An alternative to specifying the parameter λ would be to impose a hard ICU constraint by enforcing $H(t) < \bar{H}(t)$ as was done in the original study of Acemoglu et al. (2020). This would reflect more generous capacities than in the U.S.

Allowing for a less sensitive relationship between mortality and ICU needs (i.e., lowering the value of λ in (6.9)) reflects the policy maker being able to achieve lower mortality rates at a given (possibly high) number of infections. Similarly, a higher bound on available ICU beds (i.e., increasing $\bar{H}(t)$) implies that the policy maker faces a trade-off between mortality and economic damage under relaxed capacity constraints. We performed several variations with respect to the health-related parameters and refer to some examples illustrated in Figure 6.18 in the Appendix. These changes can all be summarized generally

---

[5] These mortality rates are based on Ferguson et al. (2020). We repeat the robustness check performed in Acemoglu et al. (2020) with $\delta_s = 0.12$ and confirm that the main conclusions remain unchanged. We omit the resulting plot for the sake of brevity because we provide two robustness checks with regard to a *lower* mortality rate in Section 6.4.2.

as restrictions to the set of possible options that are available to policy makers in situations in which infection rates are high.

6.3.2 Parameters for the SEIR Model

The multi-Group SEIR model is able to adapt to some characteristics specific to SARS-CoV-2. For example, one characteristic of the virus is that infections are frequently caused by some exposure to infectious persons through personal contact. To model the period spent in states E (i.e., carrying the virus but not exhibiting symptoms and neither being infectious) and I (i.e., potentially exhibiting symptoms and being infectious), we base the rates γ_j^E and γ_j^I on the conclusions of the Robert Koch Institute as provided in the RKI COVID-19 report (RKI, 2020a). In our analysis, we assume that the latent period is 6 days ($\gamma_j^E = \frac{1}{6}$) and the infectious period is 9 days ($\gamma_j^I = \frac{1}{9}$).

6.3.3 Calibrating R_0

The parameter β has been calibrated to match a basic reproduction number $R_0 = 2.4$ under the parameter constellation as described above. Setting $R_0 = 2.4$ corresponds to the lower bound on R_0 as reported by the RKI (2020a) as of July 2020. The calibration is performed in a setting without any policy intervention based on the contact matrix ρ_0, i.e., no shielding, i.e., $L_j = 0$, no testing and isolation, i.e., $\eta_j^I = 1$, and no contact tracing, i.e., $\eta_j^E = 1$, is imposed for any j in an almost entirely susceptible population.

6.4 Results and Optimal Policies

In this section, we present our results and discuss the optimal policies that can be derived from our model. We will first refer to the efficient frontier according to the German parametrization and then illustrate the effectiveness of various policies. Lastly, we will comment on the implementation of these policies in practice.

6.4.1 Efficient Frontier

Adapting the model to the socio-demographic, economic and health-care-related parameters for Germany leads to the baseline policy frontier shown in Figure 6.2. In line with the results reported by Acemoglu et al. (2020), the economic cost of shielding can be reduced substantially at a given mortality level by employing targeted shielding policies.[6] Due to the non-uniform contact rates, fully targeted policies provide improvements over semi-targeted policies, similar to the findings of Acemoglu et al. (2020). However, in many cases, these improvements are moderate to small. Assuming that the costs of implementing fully targeted policies are non-negligible and are likely to outweigh their gains, we will focus mostly on the comparison of semi-targeted and uniform policies in the following.

6.4.2 Optimal Policy

In the following, we shed light on the effectiveness of targeting shielding towards different age groups and the combination of this with additional measures, such as testing activities. Similar to Acemoglu et al. (2020), we analyze the impact of various policy measures by varying parameters and making comparisons to a baseline setting. In this benchmark, the parameters are chosen in line with the German socio-demographic and economic calibration discussed in Section 6.3. When presenting the results, we will

---

[6] Analogously to Acemoglu et al. (2020), we will refer to all policies that involve a different degree of shielding L_j for at least one group j generally as *targeted* policies. We follow the distinction of *fully targeted* policies with shielding intensities L_j that are determined separately for all three groups and *semi-targeted* policies that distinguish only one level for the senior group and one level that applies to the young and middle-aged group.

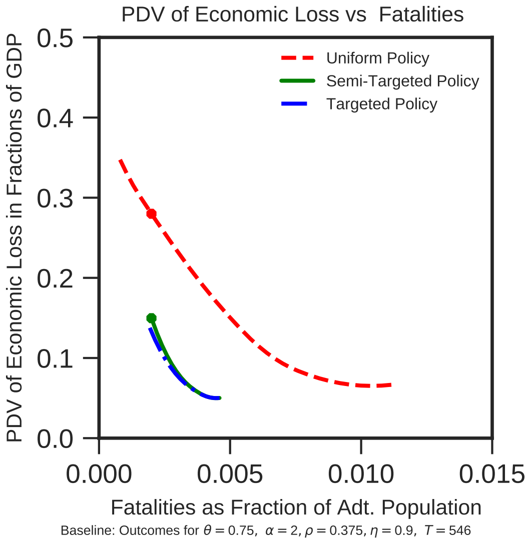


Figure 6.2: **Efficient frontier in SEIR model, baseline setting.**

The baseline setting is based on the adjusted contact matrix ρ in Equation (6.3) and demographic and health care parameters adjusted to Germany in Table 6.1.

frequently refer to a safety-focused scenario entailing policies that do not allow population mortality to exceed 0.2%. In figures with policy frontiers, the results that correspond to this setting are indicated by a dot on the respective policy frontier line.

The policy measures considered in the following refer to improving testing for those in compartment I_j (referred to as testing) and those in compartment E (referred to as contact tracing), two variants of group distancing, improved conditions for working from home, and a combination of these. In addition, we analyze the implications of a medical treatment that makes it possible to lower the mortality rate for those in ICU treatment, as well as of a vaccine arriving early. We list additional results and robustness checks in the Appendix.

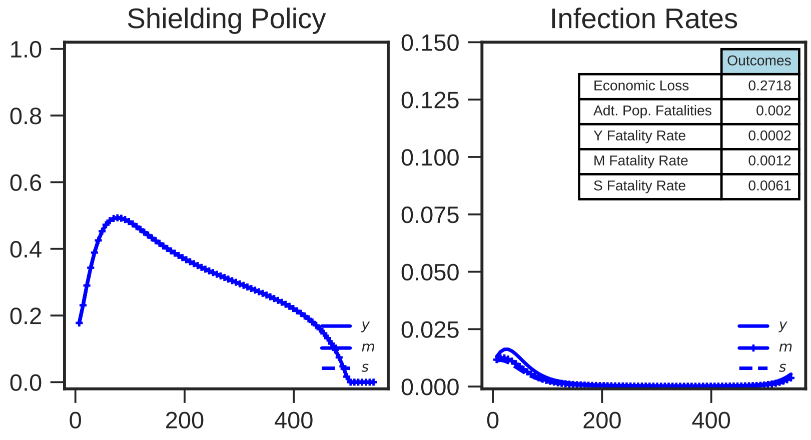
Figure 6.3 illustrates the optimal policy in the baseline setting with uniform and semi-targeted shielding. The results with regard to the economic loss at the fixed mortality level of 0.2% illustrate the gains that can be achieved by targeted shielding. In the baseline case with semi-targeted policies, a high shielding intensity is imposed on the elderly until a vaccine arrives, whereas the intensity for the other groups is lowered gradually after an initial peak. Figure 6.4 illustrates the evolution of the share of uninfected in each age group and the reproduction rate over time. Semi-targeted shielding policies as illustrated in Panel (ii) are associated with different infection rates across the age groups. Hence, the share of infected in the vulnerable group is relatively low whereas infections are more prevalent in the group of young. However, if uniform policies are considered (Panel (ii)) the variation in the share of uninfected across the age groups is much smaller.

The Effect of Physical Distancing, Face Masks and Additional Hygiene Measures

In general, the policy maker could reduce the intensity and duration of the shielding policy if the transmission rates in personal contacts could be decreased, for example by a voluntary limitation of contacts or reducing the transmission probability by wearing face masks, as described for instance in Chu et al. (2020). Among other recommendations, we provide a list of possible mandatory or voluntary group distancing measures in Section 6.4.3.

We consider two variants of group distancing by manipulating the entries of the contact matrix ρ. The

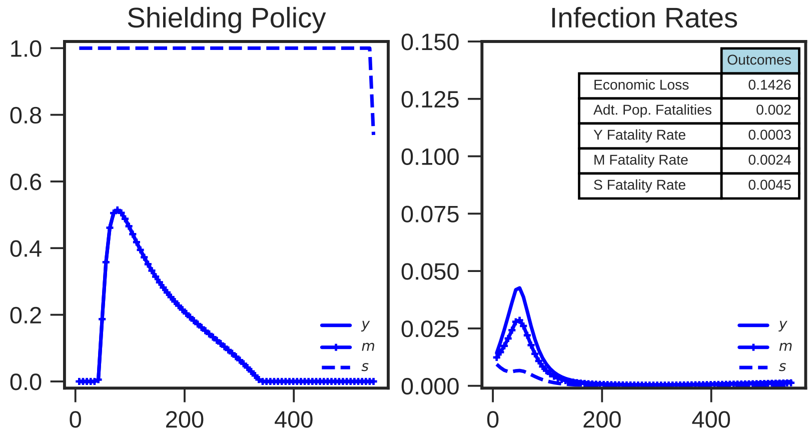
(i)



	Outcomes
Economic Loss	0.2718
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0002
M Fatality Rate	0.0012
S Fatality Rate	0.0061

Base: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

(ii)



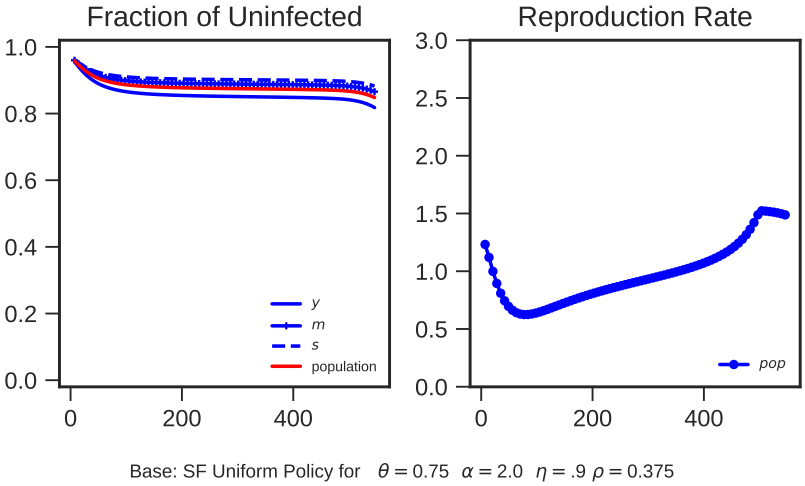
	Outcomes
Economic Loss	0.1426
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0003
M Fatality Rate	0.0024
S Fatality Rate	0.0045

Base: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

Figure 6.3: **Optimal uniform shielding policy, baseline setting.**

Panel (i): Optimal uniform shielding policy with safety focus, baseline setting. Panel (ii): Optimal semi-targeted shielding policy with safety focus, baseline setting.

(i)



(ii)

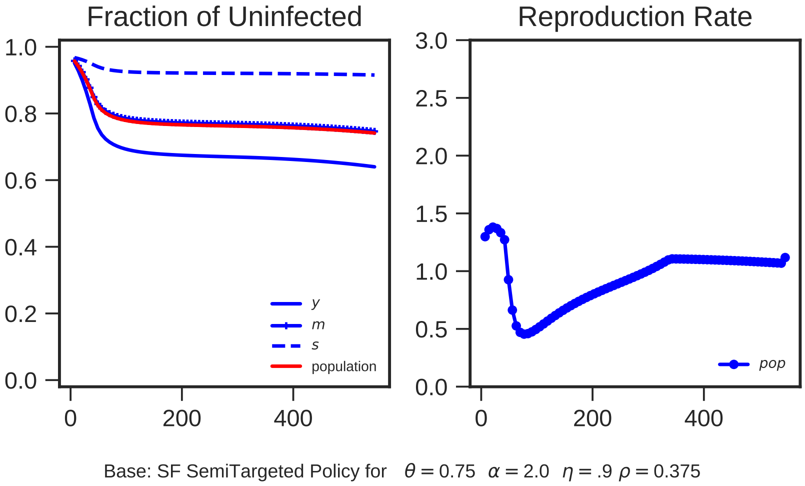


Figure 6.4: **Share of uninfected and reproduction rate, baseline setting.**

Share of uninfected (left) and reproduction rate $R(t)$ (right) in the baseline setting with safety focus. Panel (i): Optimal uniform shielding policy. Panel (ii): Optimal semi-targeted shielding policy.

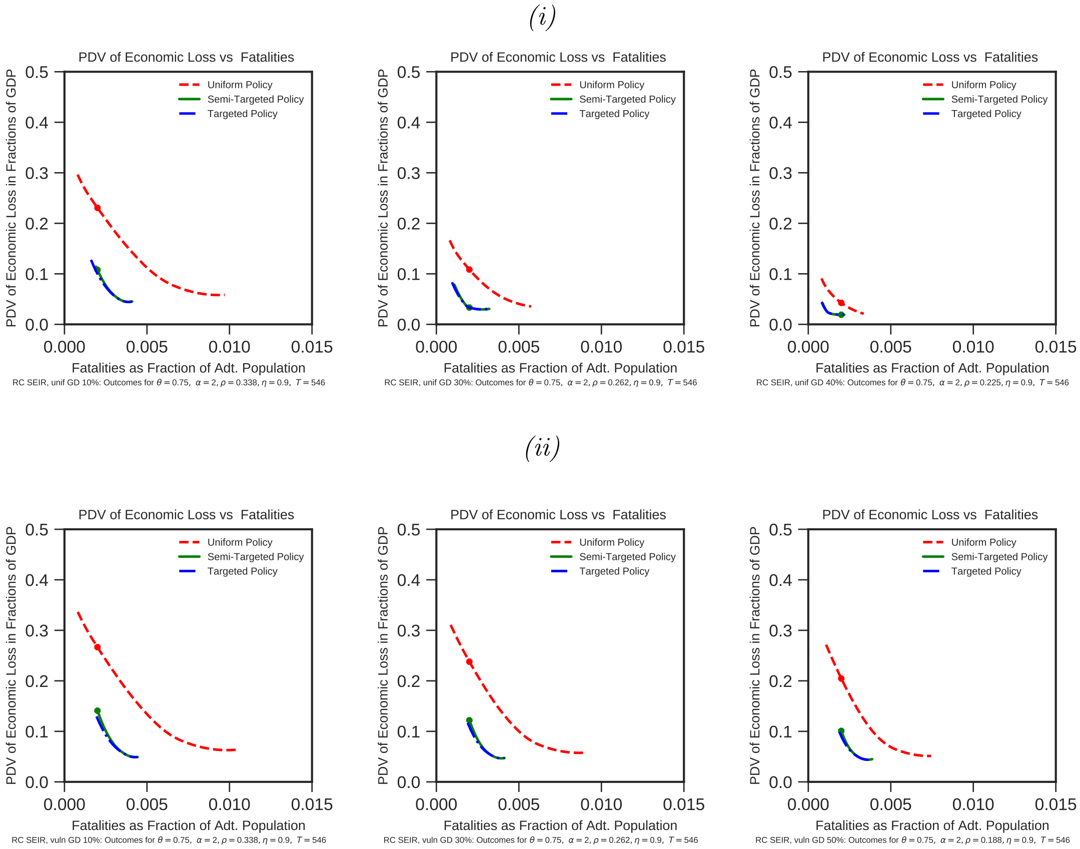
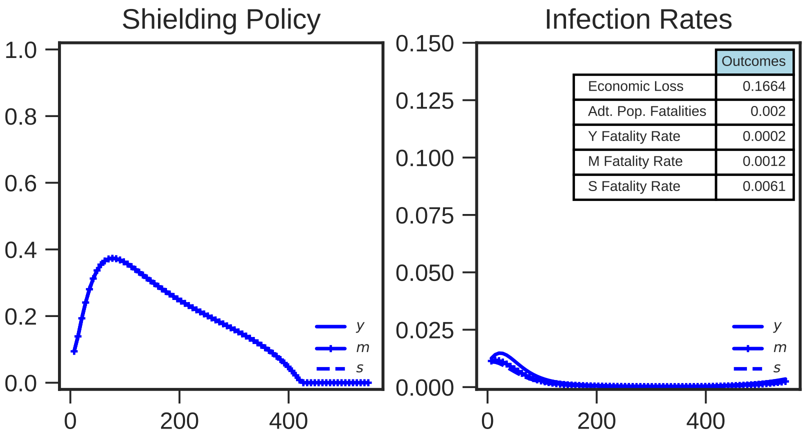


Figure 6.5: **Policy frontiers with group distancing or reduced transmission rates.**

Policy frontiers with group distancing or reduced transmission rates between and by age groups. Panel (*i*): Uniform reduction in all contact rates in the contact matrix ρ of 10% (left), 30% (center) and 40% (right). Panel (*ii*): Group distancing focusing on the most vulnerable group (i.e., the group of those aged 65+) with a reduction in the between-group contact rates ρ_{ys}, ρ_{ms} of 10% (left), 30% (center) and 50% (right).

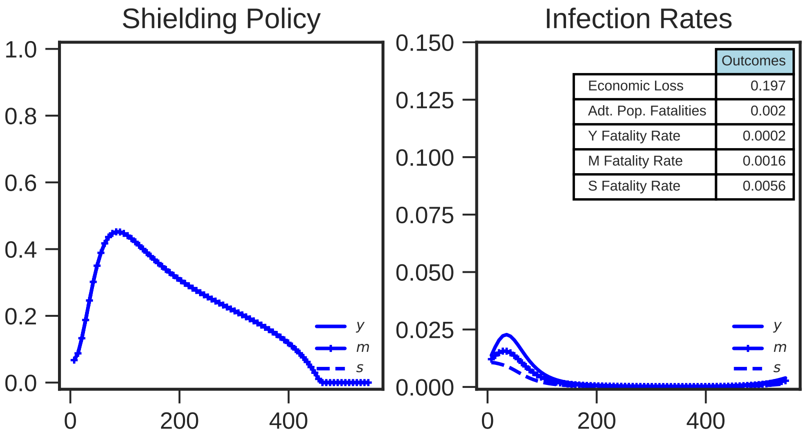
(i)



	Outcomes
Economic Loss	0.1664
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0002
M Fatality Rate	0.0012
S Fatality Rate	0.0061

unif GD 20%: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.3$

(ii)



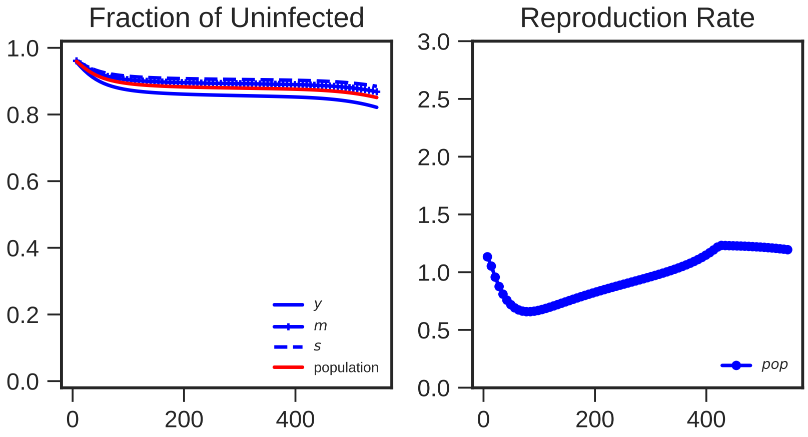
	Outcomes
Economic Loss	0.197
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0002
M Fatality Rate	0.0016
S Fatality Rate	0.0056

vuln GD 50%: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

Figure 6.6: **Optimal policy with safety focus and group distancing, uniform shielding.**

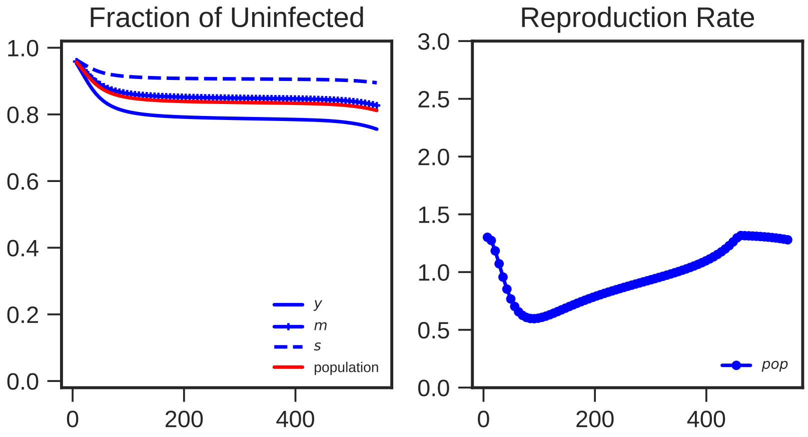
Panel (i): Uniform group distancing with 20% reduction in social interactions between and by all groups. Panel (ii): Group distancing with focus on interactions with vulnerable groups and reduction in contact rates ρ_{ys}, ρ_{ms} by 50%.

(i)



unif GD 20%: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.3$

(ii)

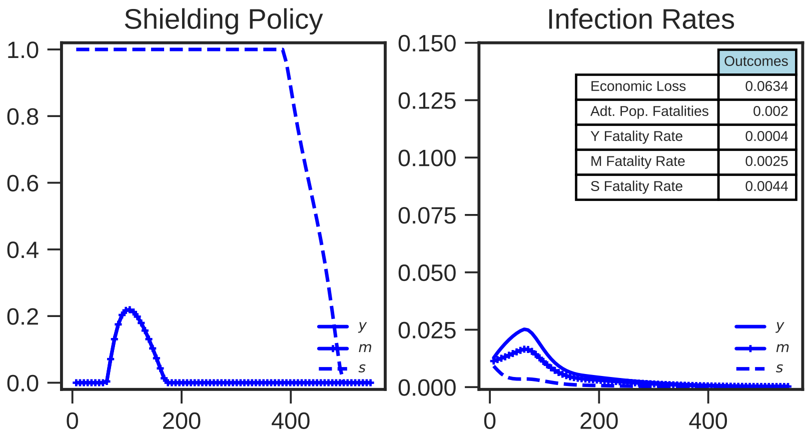


vuln GD 50%: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

Figure 6.7: **Share of uninfected and reproduction rate with group distancing, uniform shielding.**

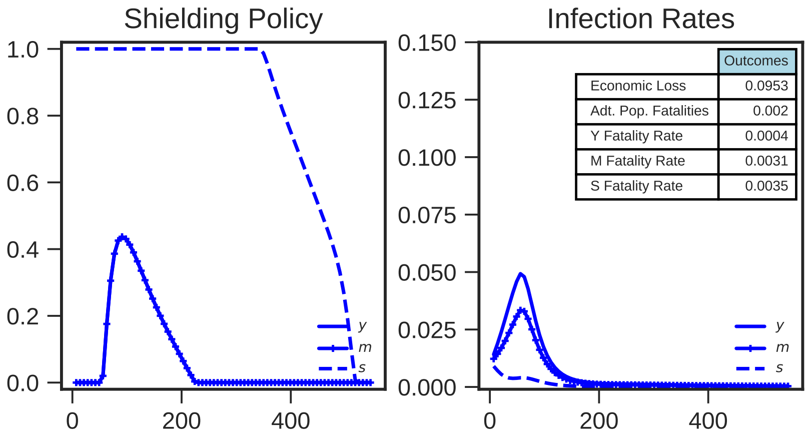
Share of uninfected (left) and reproduction rate $R(t)$ (right) in the setting with safety focus and group distancing. Uniform shielding policies. Panel (i): Uniform group distancing with 20% reduction in social interactions between and by all groups. Panel (ii): Group distancing with focus on interactions with vulnerable groups and reduction in contact rates ρ_{ys}, ρ_{ms} by 50%.

(i)



unif GD 20%: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.3$

(ii)

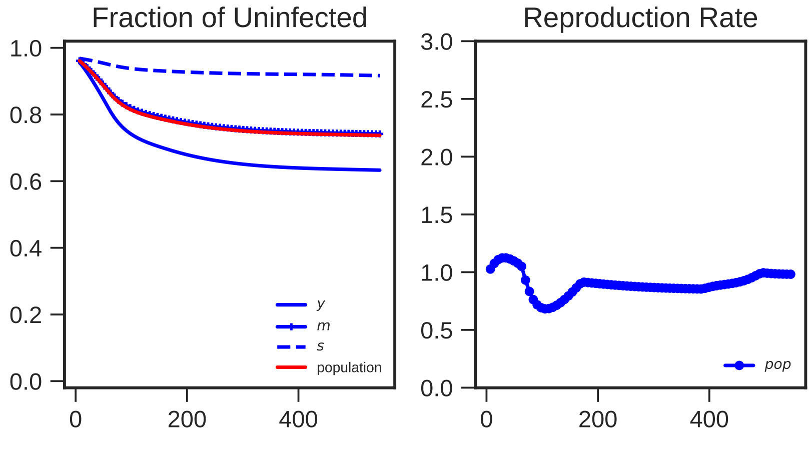


vuln GD 50%: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

Figure 6.8: **Optimal policy with safety focus and group distancing, semi-targeted shielding.**

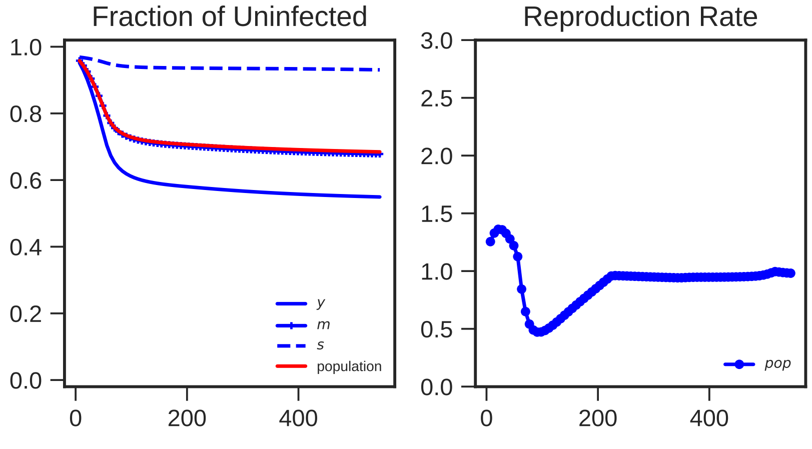
Panel (i): Uniform group distancing with 20% reduction in social interactions between and by all groups. Panel (ii): Group distancing with focus on interactions with vulnerable groups and reduction in contact rates ρ_{ys}, ρ_{ms} by 50%.

(i)



unif GD 20%: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.3$

(ii)



vuln GD 50%: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

Figure 6.9: **Share of uninfected and reproduction rate group distancing, semi-targeted shielding.**

Share of uninfected (left) and reproduction rate $R(t)$ (right) in the setting with safety focus and group distancing. Semi-targeted shielding policies. Panel (i): Uniform group distancing with 20% reduction in social interactions between and within all groups. Panel (ii): Group distancing with focus on interactions with vulnerable groups and reduction in contact rates ρ_{ys}, ρ_{ms} by 50%.

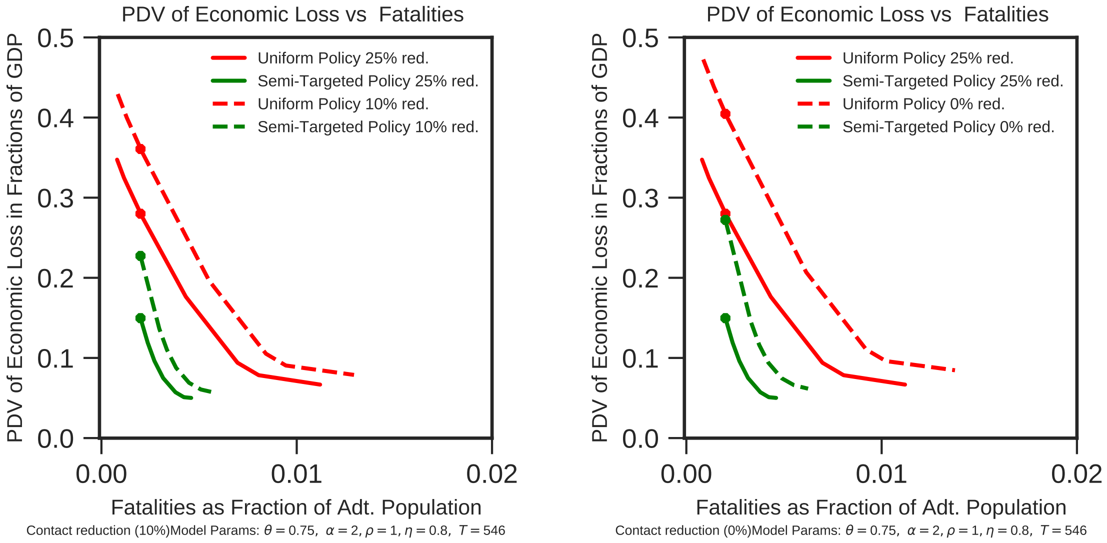


Figure 6.10: **Policy frontiers with alternative contact rate adjustment.**

Reduction of contact rates in ρ^0 by 10% (left), i.e., $\rho = 0.9 \cdot \rho^0$, and without any reduction of contact rates in ρ^0, i.e., $\rho = \rho^0$ (right). Solid lines refer to baseline scenario with $\rho = 0.75 \cdot \rho^0$.

first scenario of mandatory or voluntary group distancing we consider refers to a uniform reduction in contact rates in ρ. This setting could be considered similar to a general call to the population to reduce all personal interactions, irrespective of the age or vulnerability of the persons considered. The second scenario we consider refers to a change in the contact rates with respect to the senior group – in other words ρ_{ys} and ρ_{ms} and leaving all other entries in ρ unchanged. This scenario corresponds to "breaking the infection chain" with regard to the vulnerable group. In this scenario, the within-group and between-group contacts for the young and middle-age groups are left unchanged. Moreover, in the setting considered, we also leave the contacts within the group of senior citizens unchanged and thus attempt to model a scenario with an impact on the daily contacts of the elderly that is as low as possible. Figure 6.5 illustrates the policy frontier corresponding to changes in the contact matrix ρ according to uniform group distancing policy (panel (i)) and group distancing focusing on the vulnerable (panel (ii)). As expected, scaling all entries in ρ simultaneously is substantially more effective in reducing transmissions than is targeted group distancing. However, the social and psychological costs of a uniform group distancing policy are probably high and panel (ii) in Figure 6.5 shows that a reduced, but targeted approach might also help mitigate the health and economic costs of the pandemic.

The results in Figures 6.6 and 6.8 illustrate that group distancing can reduce the intensity and duration of uniform shielding policies while mitigating the economic damage. Substantial efficiency gains can be achieved by targeting shielding towards the separate groups. Comparing (i) a uniform physical group distancing policy (corresponding to a 20% reduction in contact rates across all groups, shown at the top of Figures 6.6 and 6.8) and (ii) targeted group distancing towards the vulnerable (corresponding to a 50% reduction in contact rates between young and middle group and the senior citizens, shown at the bottom of Figures 6.6 and 6.8) illustrates that an intense reduction in contacts and/or transmission rates between the vulnerable group and the other age groups can be an effective tool for mitigating the health and economic consequences of the COVID-19 pandemic if combined with targeted shielding. Figures 6.7 and 6.9 present the evolution of the share of susceptibles and the reproduction rate over time and shed light on the epidemiological consequences of uniform group distancing (Panel (i) in Figure 6.7 and Figure 6.9) and a group distancing policy with a focus on the vulnerable (Panel (ii) in Figure 6.7 and Figure

6.9). Hence, a more targeted form of group distancing is associated with a greater difference in terms of age-group specific infection rates. Similar to the baseline scenario, the variation in terms of age-specific infection rates is higher if semi-targeted policies are considered.

As a robustness check, we employ two settings with reduced physical distancing and illustrate the corresponding frontiers in Figure 6.10. Doing so, we intend to illustrate the consequences if individuals are less compliant to physical distancing guidelines, for example, because they underestimate the risk of transmissions.

The Effect of Testing and Contact Tracing

Improved testing and isolation with respect to infectious individuals refers to a reduction in the probability η_j^I from the baseline value of $\eta_j^I = 0.9$ − that is, the probability that an infectious individual is not detected and isolated to avoid subsequent infections.[7] A second testing strategy could refer to those who have had contact with the infectious individuals − that is, decreasing the probability that someone who was exposed to an infectious person is not detected and isolated, η^E, with default value $\eta_j^E = 1$.

Figure 6.11 illustrates the beneficial implications of improved testing with regard to persons in state I_j (panel (*i*)), improved testing and tracing for persons in E_j (panel (*ii*)) and a combination of these measures (panel (*iii*)). A reduction of η^I allows the policy frontier to be shifted closer to the origin, and therefore for efficiency gains to be realized compared to the baseline setting with $\eta_j^I = 0.9$. The corresponding frontier is shown in the first plot (on the left) of panel (*i*). Similar conclusions can be drawn for the contact tracing policy as illustrated in panel (*ii*) of Figure 6.11. However, simultaneously improving tests both for the infectious and tests for the exposed leads to a substantial improvement in the menu of potential alternatives for policy makers. For example, the safety-focused scenario indicated by the dot on the frontiers involves substantially lower economic costs if the probabilities for undetected infections are reduced to $\eta_j^I = \eta_j^E = 0.8$.

The Effect of Improved Conditions for Working From Home

In addition to voluntary or mandatory group distancing and test and trace policies, governments could provide incentives to promote working from home. To implement improved working from home conditions, we consider a setting with fewer physical interactions and increased productivity at home. Figure 6.12 illustrates two scenarios with (i) $\pi_1 = 20\%$ and $\pi_2 = 5\%$ and (ii) $\pi_1 = 30\%$ and $\pi_2 = 10\%$. In both settings, the efficiency loss is reduced by 10 percentage points − that is, in the baseline setting, the productivity loss under shielding was set to 70% and is now changed to 60%. Panel (*i*) of Figure 6.12 shows the policy frontiers that correspond to setting (i) (left) and setting (ii) (right). Panel (*ii*) illustrates the optimal semi-targeted shielding policy with a safety focus. We can see that improved conditions for working from home make it possible to reduce substantially the economic costs associated with the pandemic and with shielding. Moreover, better conditions for working from home make it possible to reduce the duration and intensity of shielding measures as compared to the baseline setting.

A Comprehensive Approach

The positive effect of improved testing and group distancing can be amplified if these measures are combined with other measures to form a comprehensive approach. As illustrated in panel (*i*) of Figure 6.13, combining improved testing and contact tracing with group distancing focusing on interactions of the other groups with the group of senior citizens allows the policy frontier to be shifted closer to the origin. According to the efficient frontiers in Figure 6.13 (panel(*i*)) and the optimal policies in panel (*ii*),

---

[7] In our analysis, we will only focus on changes in η_j^I and η_j^E that apply equally to all groups, e.g., consider cases with $\eta_y^I = \eta_m^I = \eta_s^I = 0.9$.

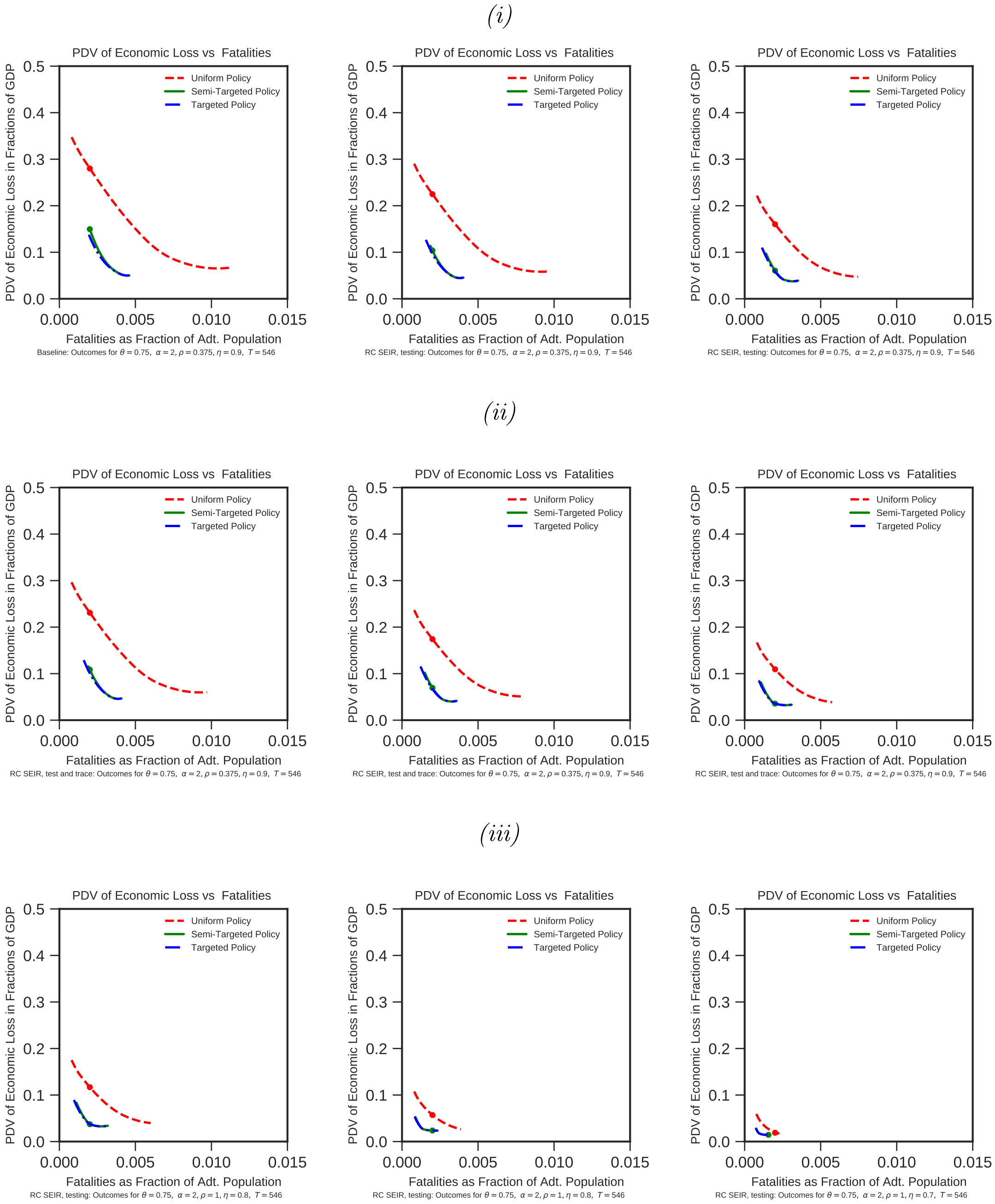
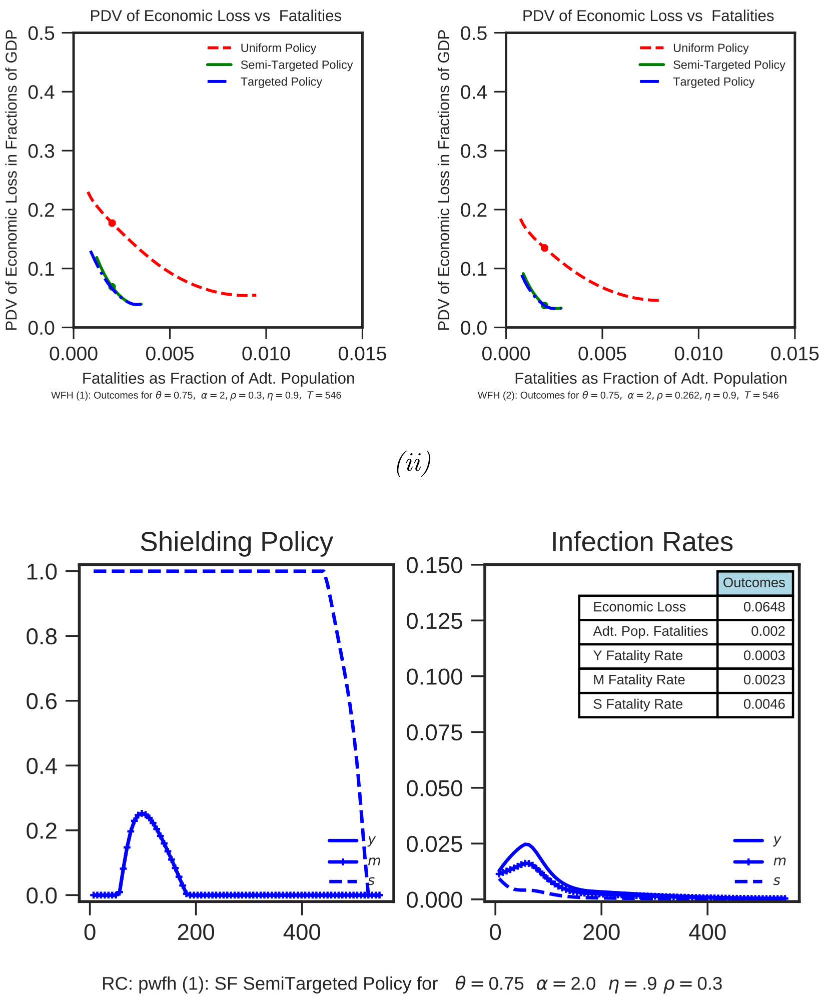


Figure 6.11: **Policy frontiers with improved testing and isolation.**

Panel (*i*): Tests for infectious persons, parameters in order from left to right $(\eta_j^I, \eta_j^E) = (0.9, 1), (0.8, 1), (0.7, 1)$. Panel (*ii*): Improved test and trace policy for exposed individuals, parameters $(\eta_j^I, \eta_j^E) = (0.9, 0.9), (0.9, 0.8), (0.9, 0.7)$. Panel (*iii*): Combination of testing infectious and test and trace policy with parameters $(\eta_j^I, \eta_j^E) = (0.8, 0.8), (0.7, 0.8), (0.7, 0.7)$.

(i)



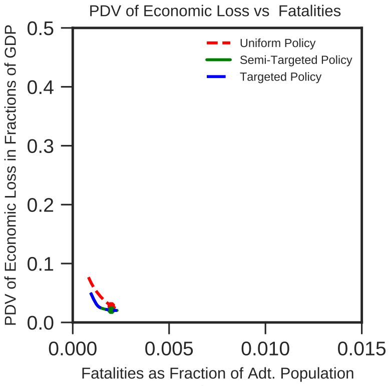
	Outcomes
Economic Loss	0.0648
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0003
M Fatality Rate	0.0023
S Fatality Rate	0.0046

RC: pwfh (1): SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.3$

Figure 6.12: **Policy frontiers with improved conditions for working from home.**

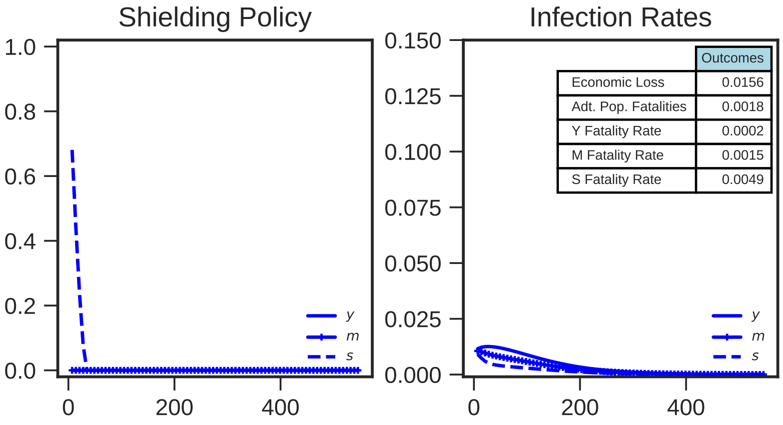
Panel (*i*): Frontiers with two variants of improved conditions for working from home, with $\pi_1 = 20\%, \pi_2 = 5\%, \xi = 0.4$ (left) and $\pi_1 = 30\%, \pi_2 = 10\%, \xi = 0.4$ (right). Panel (*ii*): Optimal semi-targeted policy with safety focus with scaling $\pi_1 = 20\%, \pi_2 = 5\%, \xi = 0.4$.

(i)



RC SEIR, Combined Measures: Outcomes for $\theta = 0.75$, $\alpha = 2$, $\rho = 0.375$, $\eta = 0.7$, $T = 546$

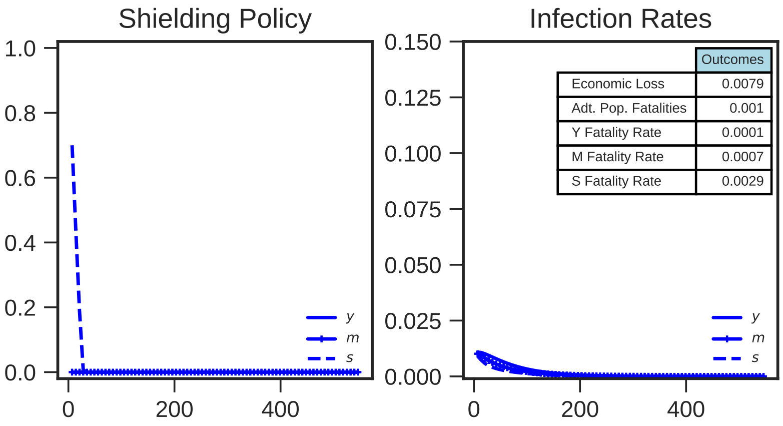
(ii)



	Outcomes
Economic Loss	0.0156
Adt. Pop. Fatalities	0.0018
Y Fatality Rate	0.0002
M Fatality Rate	0.0015
S Fatality Rate	0.0049

RC SEIR, combined measures: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

(iii)



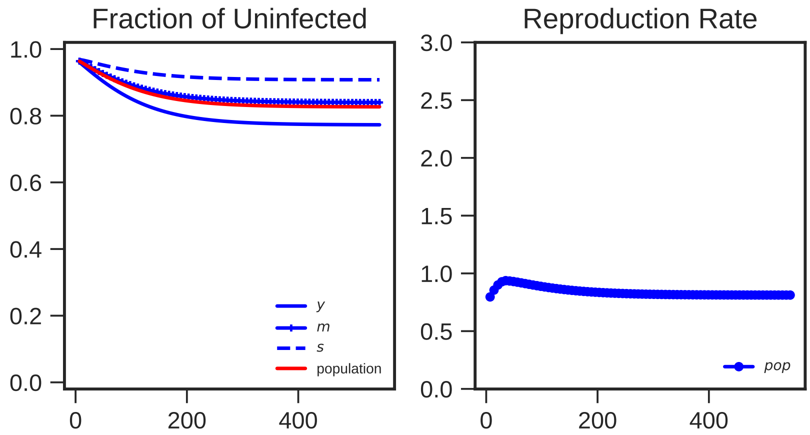
	Outcomes
Economic Loss	0.0079
Adt. Pop. Fatalities	0.001
Y Fatality Rate	0.0001
M Fatality Rate	0.0007
S Fatality Rate	0.0029

RC SEIR, combined measures, wfh: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.3$

Figure 6.13: **Policy frontiers and optimal policy with combination of policy measures.**

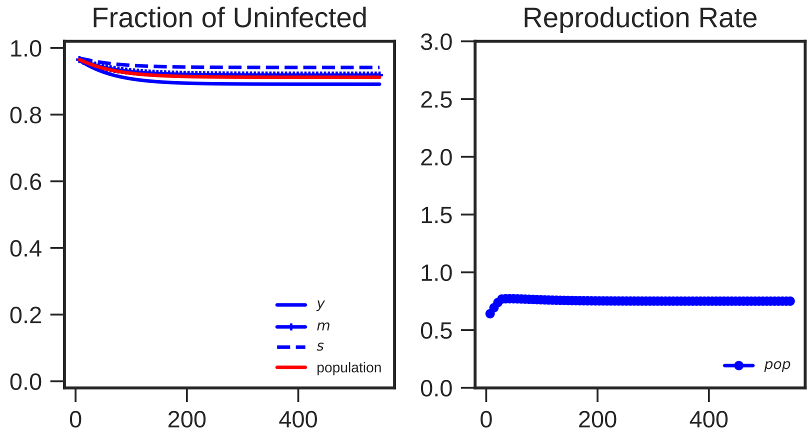
Efficient frontier and optimal semi-targeted shielding policy with safety focus with two variants of the comprehensive approach. Panel (*i*) and (*ii*): Improved testing and isolation for infected ($\eta_I = 0.7$) and exposed ($\eta_E = 0.8$), reduced contact rates for interactions with the senior group ($\rho_{ys} = \rho_{ms} = 0.2$). Panel (*iii*): Improved testing and isolation for infected ($\eta_I = 0.7$) and exposed ($\eta_E = 0.8$), reduced contact rates for interactions with the senior group ($\rho_{ys} = \rho_{ms} = 0.2$), and improved conditions for working from home ($\pi_1 = 20\%, \pi_2 = 5\%, \xi = 0.4$.)

(i)



RC SEIR, combined measures: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

(ii)



RC SEIR, combined measures, wfh: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.3$

Figure 6.14: **Share of uninfected and reproduction rate with combination of policy measures**

Share of uninfected (left) and reproduction rate $R(t)$ (right) in the setting with comprehensive approach and semi-targeted shielding policies. Panel (i): Improved testing and isolation for infected ($\eta_I = 0.7$) and exposed ($\eta_E = 0.8$), reduced contact rates for interactions with the senior group ($\rho_{ys} = \rho_{ms} = 0.2$). Panel (ii): Improved testing and isolation for infected ($\eta_I = 0.7$) and exposed ($\eta_E = 0.8$), reduced contact rates for interactions with the senior group ($\rho_{ys} = \rho_{ms} = 0.2$), and improved conditions for working from home ($\pi_1 = 20\%, \pi_2 = 5\%, \xi = 0.4$).

policy makers can almost refrain entirely from imposing shielding rules. Optimally, only a short shielding is imposed on the senior group at the early stage of the pandemic. The optimal policies associated with uniform shielding as presented in the Appendix, Figure 6.19 do not involve any shielding at comparable economic damage and a slightly higher mortality in the population.

Panel (*ii*) of Figure 6.13 illustrates that a comprehensive approach that also includes improved conditions for working from home, with $\pi_1 = 20\%$ and $\pi_2 = 5\%$, allows for even lower economic damages at a very short shielding phase. Moreover, the result on the infection rates and the reproduction rate in panel (*ii*) of Figure 6.14 show that the combined approach with improved conditions for working from home help to reduce the share of infected in all age groups. Furthermore, it can be observed that the comprehensive approach allows the reproduction rate being kept below the critical threshold of 1.

The Effect of Improved Medication and Treatment

Finally, we assess the effect associated with an improved treatment for COVID-19, which corresponds with a 30% and 50% lower baseline mortality rate for the group of senior citizens, $\underline{\delta}_s^d$. For example, a recent study by Horby and Landray (2020) shows that treatment with dexamethasone can reduce the mortality of severe hospitalized cases by up to one third. We acknowledge that the different effects of an approved drug for treating COVID-19 patients, as described in (i) to (iii) in Section 6.2.1 (Vaccine and Cure), might lead to different results in terms of optimal policies. Acemoglu et al. (2020) provide robustness checks by increasing the mortality rate for the senior group and also varying this group's per-capita income. By doing so, they conclude that the efficiency gains of targeted policies arise due to the high vulnerability rather than the low productivity of that group. Hence, effective medical treatments might soften the distinction between the vulnerable group and the groups with lower mortality risk.

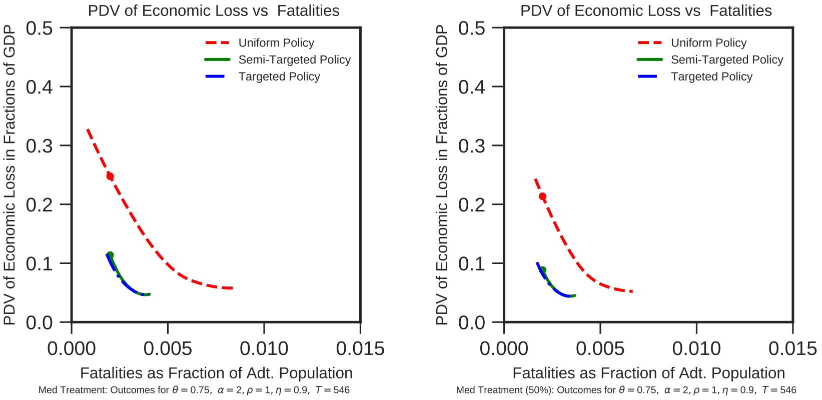


Figure 6.15: **Policy frontier with improved medical treatment.**

Efficient frontier in SEIR model with improved medical treatment corresponding to 30% (left) and 50% (right) lower mortality for the senior group.

Comparing the efficient frontier with improved treatment in Figure 6.15 with that in the baseline setting in Figure 6.2 illustrates that the economic costs at a given mortality level can be reduced substantially. At the same time, the distance between the frontiers of targeted and uniform shielding policies becomes smaller, which is in line with the observation in Acemoglu et al. (2020) that the efficiency gains of targeting accrue due to high vulnerability.[8] However, even with a substantially improved medical treatment that leads to a 50% lower mortality among the senior group, targeted shielding is still associated with considerable efficiency gains compared to uniform approaches.

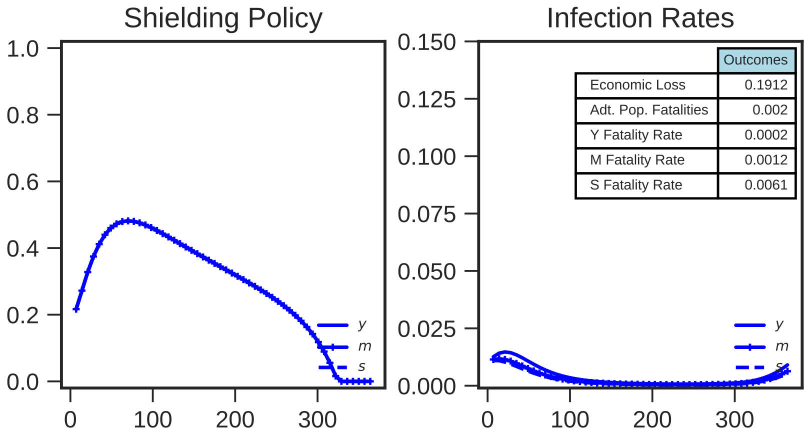
---

[8] We perform various robustness checks (results omitted) with respect to the income parameters ω_j and mortality rates $\underline{\delta}_s^d$ and confirm the conclusions in Acemoglu et al. (2020).

The Effect of a Vaccine Arriving Early

Since the early phase of the pandemic, governments around the world have encouraged research activities to develop a vaccine for SARS-CoV-2. In the initial study by Acemoglu et al. (2020) and the settings considered so far, we make a deterministic assumption that a vaccine arrives in 1.5 years. Figures 6.16 and 6.17 illustrate the optimal uniform and semi-targeted shielding policies if an effective vaccine is available after one year and after six months, respectively. The results highlight the economic importance of an effective vaccine being available early because this would substantially reduce the loss in GDP, which in the baseline scenario decreases by approximately 26% under uniform shielding policies and 13% under semi-targeted policies if a vaccine becomes available after 1.5 years. If, in contrast, a vaccine becomes available in one year, the loss under uniform shielding reduces to 18% and 9% under semi-targeted policies. In the scenario that a vaccine becomes available after six months, these numbers are 8% and 5%, respectively. As a consequence of a shorter period T, the shielding policies are maintained over a shorter time span, wherease their intensity does not change substantially.

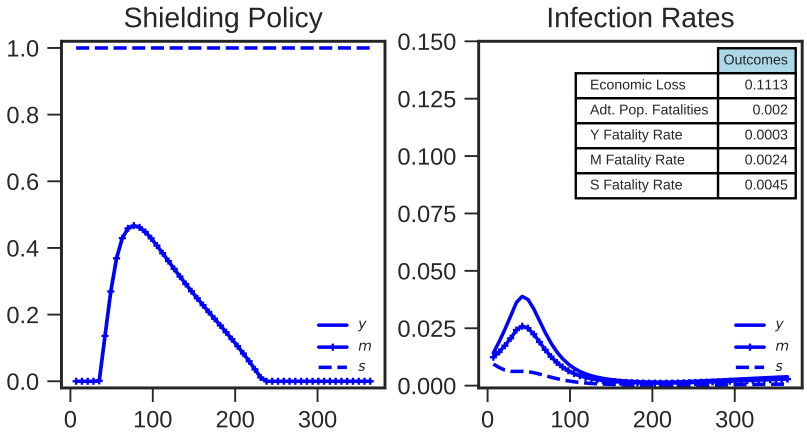
(i)



	Outcomes
Economic Loss	0.1912
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0002
M Fatality Rate	0.0012
S Fatality Rate	0.0061

Vaccine in 1 yr: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

(ii)



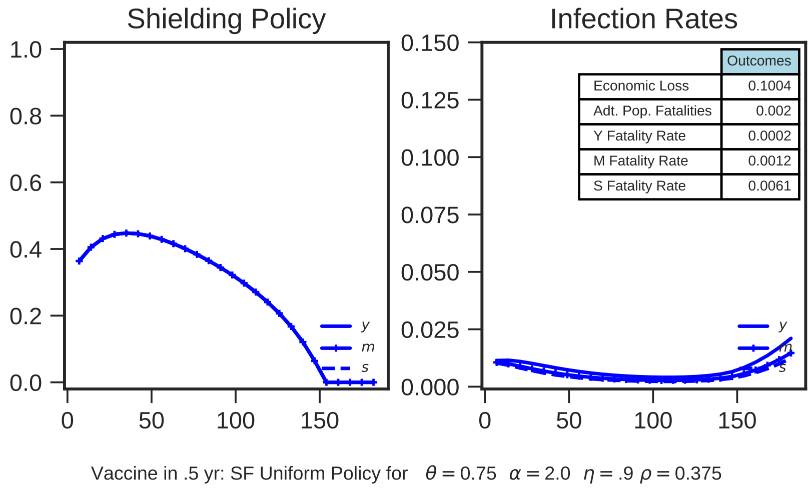
	Outcomes
Economic Loss	0.1113
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0003
M Fatality Rate	0.0024
S Fatality Rate	0.0045

Vaccine in 1 yr: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

Figure 6.16: **Optimal shielding policy with a vaccine arrival after one year.**

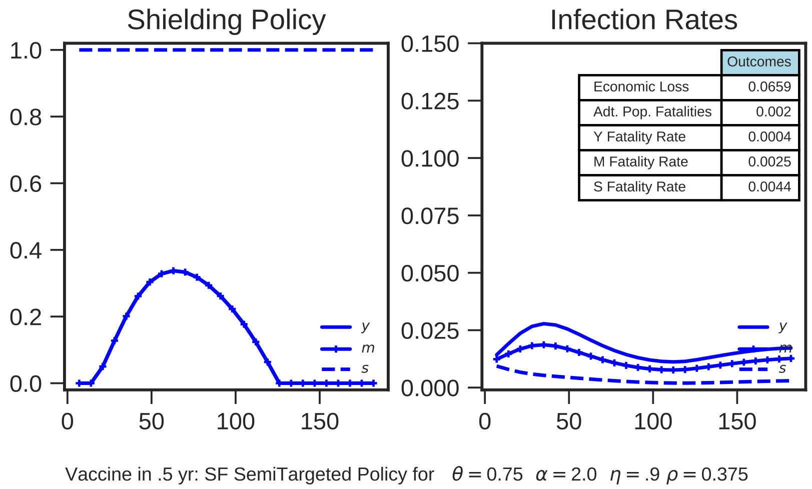
Panel (i): Optimal uniform policy. Panel (ii): Optimal semi-targeted policy.

(i)



	Outcomes
Economic Loss	0.1004
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0002
M Fatality Rate	0.0012
S Fatality Rate	0.0061

(ii)



	Outcomes
Economic Loss	0.0659
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0004
M Fatality Rate	0.0025
S Fatality Rate	0.0044

Figure 6.17: **Optimal shielding policy with a vaccine arriving after six months.**

Panel (*i*): Optimal uniform policy. Panel (*ii*): Optimal semi-targeted policy.

6.4.3 Implementation of Optimal Policies

The results of the model indicate that it is favorable to protect vulnerable groups, in particular persons at higher ages, while relieving other groups of the shielding measures. Of course such a policy must be implemented with a sense of proportion and be supported by accompanying measures to reduce the burden as much as possible. We propose that implementing such a policy should take into account the following measures and points:

1. Prioritize the use of masks (N-95/FFP2; surgical masks) and personal protection equipment among seniors and other individuals with comorbidities. These groups and contact persons (like nurses) should be equipped with masks that are of high quality.
2. Limit and reduce potential transmissions by decreasing the number of contacts with persons who are at higher risk, in particular with those who are likely to be exposed to other infectious individuals. The risk of transmission could be reduced by generally requiring people to cover their nose and mouth with simple disposable or reusable face masks when interacting with the elderly or other vulnerable groups, or by reducing the risk of individuals being infectious at the time of the contact, for example through intensive testing and requiring quarantine or reduced contacts during a defined period before visits to nursing homes.
3. Set up special shopping and medical consultation hours for the elderly and vulnerable to allow them to do shopping for daily essentials and attend important medical appointments. This has been practiced in the U.S., UK and other countries. Also encouraging the use and potentially the expansion of various home-delivery services could be valuable in this regard.
4. Ensure older people who are still participating in the labor market and other high-risk individuals are able to work from home easily, for example by providing them with the appropriate equipment, infrastructure and training.
5. Provide additional benefits and compensation, such as job guarantees and prioritized paid leave, for employees who behave in a socially responsible way, caring for the elderly or other vulnerable individuals. For example, health insurers could offer monetary benefits to individuals who commit themselves to reducing social interactions with others in order to care for individuals at high risk
6. Provide mental health and social support via teleconferencing and other safe means of interaction, particular through online consultation hours and tele-medicine, as well as video-conferencing systems in nursing homes so that residents can stay in contact with their families.
7. Implement a stay-at-home policy for older people on a voluntary basis. High compliance with this policy might be achieved through an incentive scheme. Given the right incentives a senior citizen should follow a stay-at-home policy in his or her own interest.
8. Provide frequent, easy-to-understand and non-contradictory information and communication and assistance to members of vulnerable groups who live in their own home; also, create incentives for members of the young and middle age groups to protect the vulnerable members of society.
9. Frequently update shielding policies according to new scientific evidence on the transmission of SARS-CoV-2 in order to increase the efficiency of such measures, reduce economic costs and achieve higher compliance with group distancing recommendations.

6.5 Conclusion

In this paper, we adopt the extended SIR model of Acemoglu et al. (2020) to Germany. Germany differs from the U.S. both in its socio-demographics and its system of health care coverage and provision. The model allows for a comparison of the impact of different policies both on survival rates and economic losses, thus providing policy makers with information to derive optimal policies. We evaluate several scenarios in a quantitative manner and find that semi-targeted shielding makes it possible to achieve efficiency gains, which might be used to fund measures that improve the conditions of vulnerable groups, such as senior citizens and people with comorbidities. Most importantly, we find that the intensity and duration of shielding policies can be reduced by employing additional measures, such as group distancing, testing and contact tracing. Indeed, a comprehensive approach that combines these measures and implements them simultaneously can keep both economic losses and population mortality at a low level – even with uniform shielding measures. Lastly, we highlight the importance of finding effective medical treatments and of timely vaccine development.

There are several extensions of our analysis that could be considered in future research. First, the estimates on contact rates that are used in the baseline setting are based on a study from the UK and might be re-adjusted to country-specific contact patterns. Additional work could be performed to provide comparable data for other countries, including Germany and the U.S. Second, the SEIR model incorporates contact tracing by including a parameter on the probability that a person who was exposed to an infectious individual is tested and isolated. This approach allowed us to maintain a relatively concise model structure. A more complex structure might involve separate compartments for exposed and infectious individuals who are either in quarantine or not in quarantine, allowing the social interactions between these two groups to be modeled. The model in Grimm et al. (2020) is an example of such an evolved compartment structure. Moreover, the infectiousness of individuals could be modeled in a more granular way. Several studies, such as Grimm et al. (2020), distinguish between symptomatic, asymptomatic and severe cases and allow for transmissions of SARS-CoV-2 by asymptomatic cases. Alternatively, the SEIR model in Berger et al. (2020) allows for infectiousness of the exposed individuals. Lastly, as soon as more information is available on whether people develop long-term immunity to SARS-CoV-2 after infection, this might be used to adapt the SEIR model to a SEIRS structure.

6.6 Appendix

Additional Results

Reducing the parameter λ rotates the policy frontier to the left, bringing the menu of policy choices involving high mortality rates closer to the bliss point as illustrated in Figure 6.18, panel (i). Because targeted policies can reduce the mortality in the adult population effectively, reducing λ has an effect in particular on the choice of uniform policies.

Relaxing the hard ICU constraint allows lower mortality rates to be achieved at a given level of economic damage as can be concluded from Figure 6.18, panel (ii). Increasing the bound in the capacity constraint from $\bar{H}(t) = 0.02$ to $\bar{H}(t) = 0.04$ brings the policy frontier closer to the case without binding ICU constraints. This change can be observed for uniform shielding policies, whereas the impact on targeted policies is smaller and only observable for the case with a relatively tight constraint $\bar{H}(t) = 0.02$.

(i)

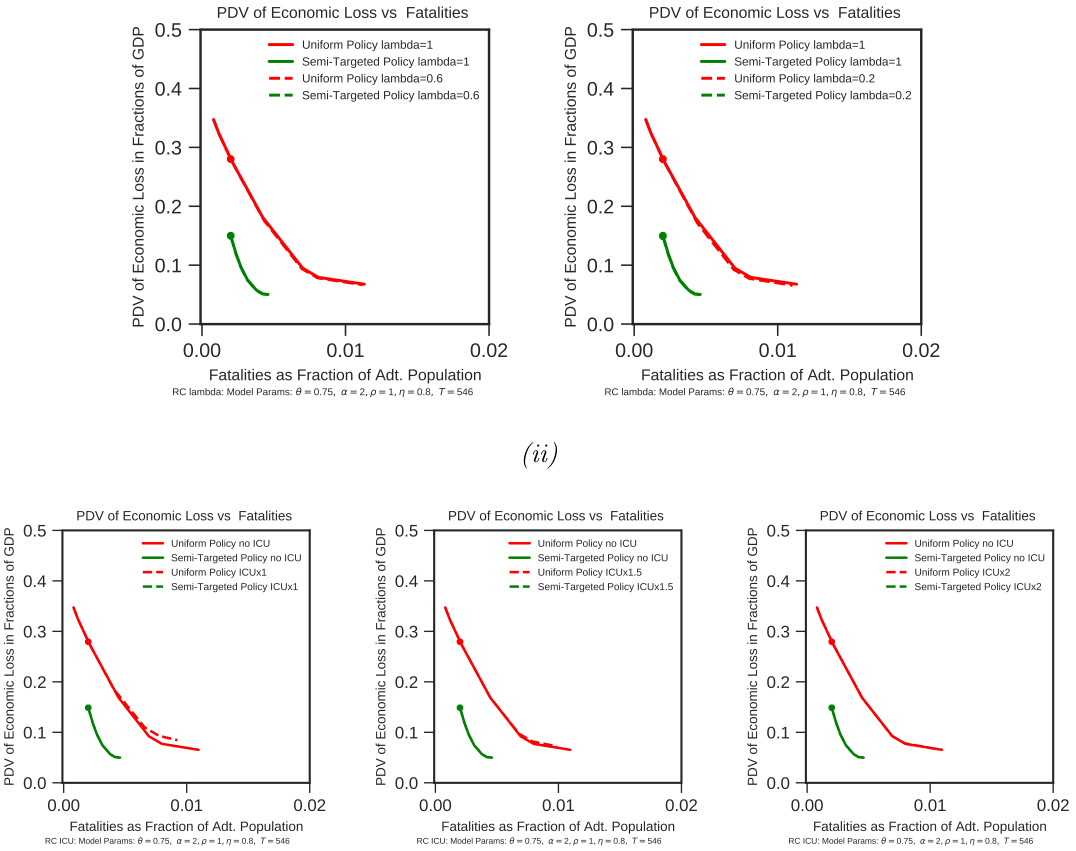
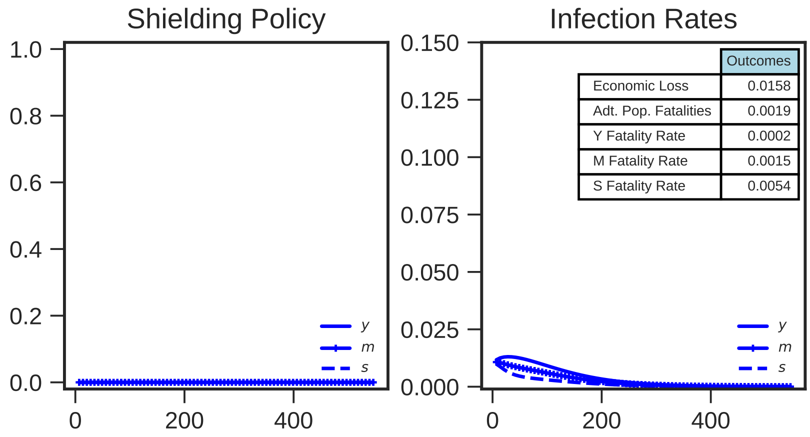


Figure 6.18: **Policy frontiers, variation of the parameter λ and the ICU capacity constraint.**

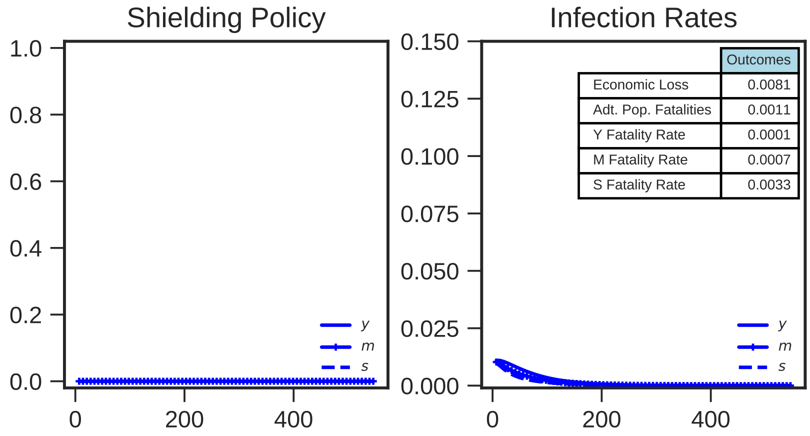
Panel (i): Variation in $\lambda = \lambda'$ in Equation (6.9), with $\lambda' = 0.6$ (left), $\lambda' = 0.2$ (right). The solid line refers to the case with $\lambda = 1$. The dashed lines refer to the policy frontier with the $\lambda = \lambda'$. Panel (ii): Variation of the ICU constraint. The solid line indicates the optimal policy frontier without a binding ICU constraint. Dashed lines refer to binding ICU constraints with $\bar{H}(t) = 0.02$ (left), $\bar{H}(t) = 0.03$ (center), and $\bar{H}(t) = 0.04$ (right).

(i)



RC SEIR, combined measures: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

(ii)

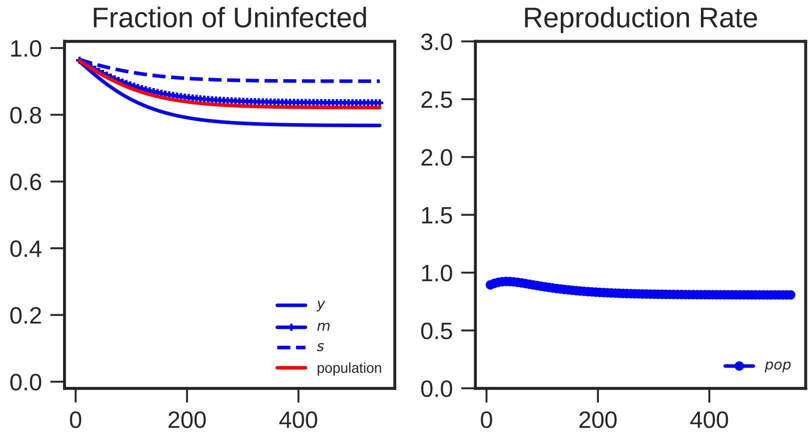


RC SEIR, combined measures, wfh: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.3$

Figure 6.19: **Optimal policy with combination of policy measures, uniform shielding.**

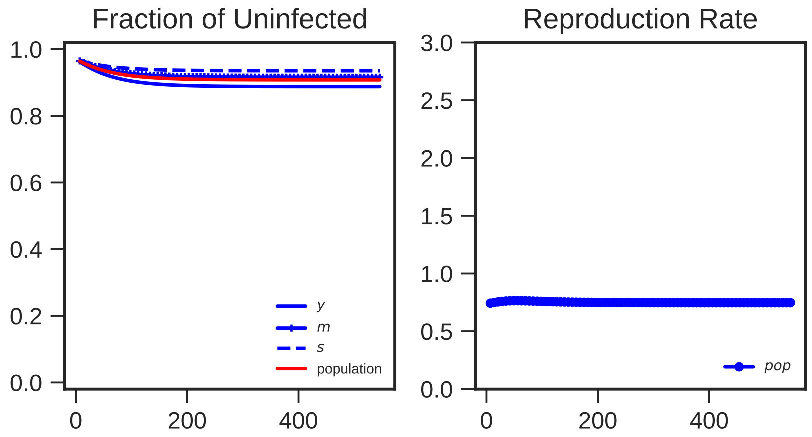
Optimal uniform policy with safety focus and comprehensive approach. Panel (i): Improved testing and isolation for infected ($\eta_I = 0.7$) and exposed ($\eta_E = 0.8$), reduced contact rates for interactions with the senior group ($\rho_{ys} = \rho_{ms} = 0.2$). Panel (ii): Improved testing and isolation for infected ($\eta_I = 0.7$) and exposed ($\eta_E = 0.8$), reduced contact rates for interactions with the senior group ($\rho_{ys} = \rho_{ms} = 0.2$), and improved conditions for working from home ($\pi_1 = 20\%, \pi_2 = 5\%, \xi = 0.4$.)

(i)



RC SEIR, combined measures: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

(ii)



RC SEIR, combined measures, wfh: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.3$

Figure 6.20: **Share of uninfected and reproduction rate with combination of policy measures, uniform shielding.**

Share of uninfected (left) and reproduction rate $R(t)$ (right) in the setting with combination of policy measures (comprehensive approach). Uniform shielding policies. Panel (*i*): Improved testing and isolation for infected ($\eta_I = 0.7$) and exposed ($\eta_E = 0.8$), reduced contact rates for interactions with the senior group ($\rho_{ys} = \rho_{ms} = 0.2$). Panel (*ii*): Improved testing and isolation for infected ($\eta_I = 0.7$) and exposed ($\eta_E = 0.8$), reduced contact rates for interactions with the senior group ($\rho_{ys} = \rho_{ms} = 0.2$), and improved conditions for working from home ($\pi_1 = 20\%, \pi_2 = 5\%, \xi = 0.4$).

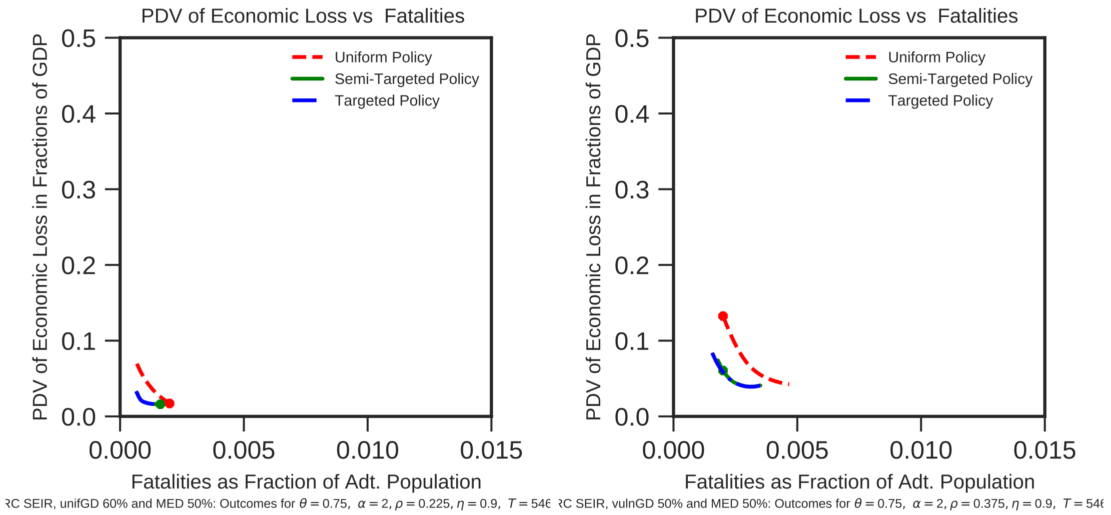
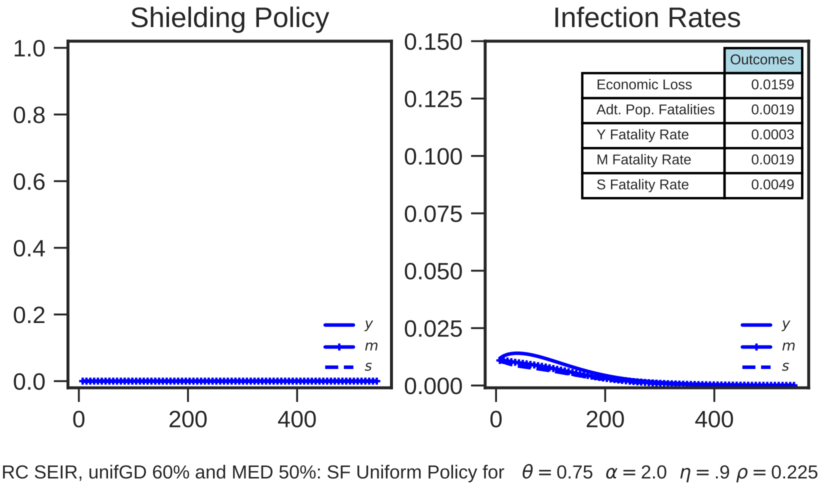


Figure 6.21: **Policy frontiers, combination of improved medical treatment and group distancing.**

Combination of group distancing and decreasd mortality of the elderly by 50% due to improved medical treatment. Efficient frontier with uniform distancing policy (left) and distancing targeted towards the vulnerable (right).

(i)



(ii)

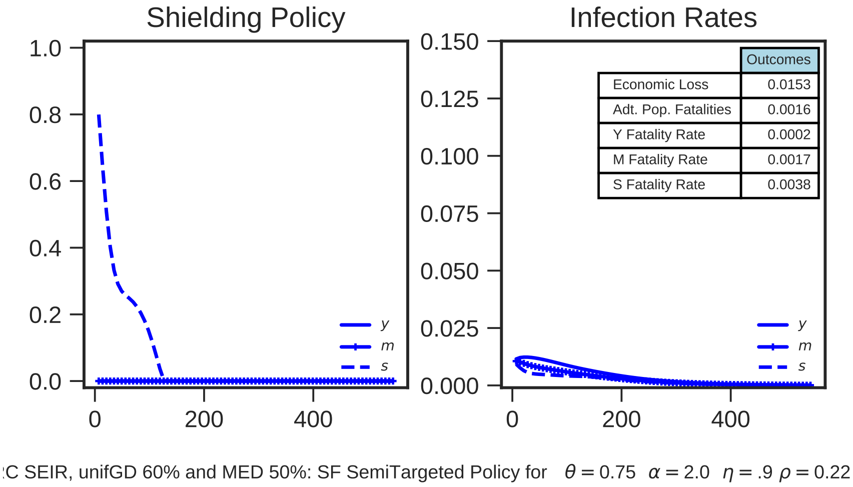
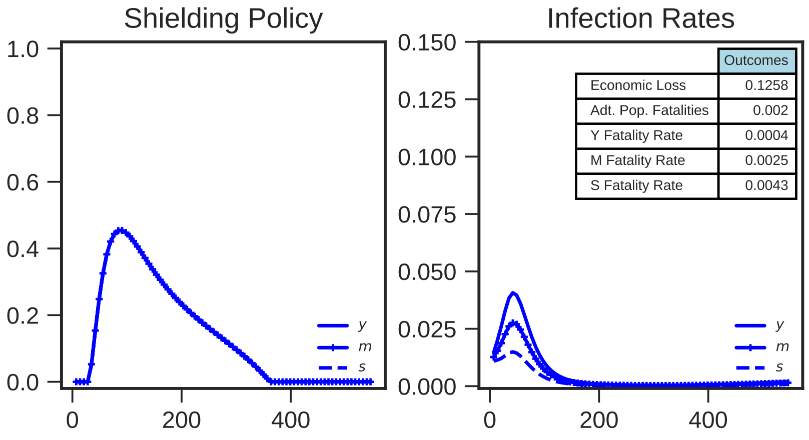


Figure 6.22: **Optimal policy, combination of improved medical treamtent and uniform group distancing policy.**

Combination of a uniform group distancing policy (reduction of all elements in ρ by 40%) and improved medical treamtent (50% lower mortality rate for the senior group). Panel (*i*): Optimal semi-targeted shielding policy. Panel (*ii*): Optimal semi-targeted shielding policy.

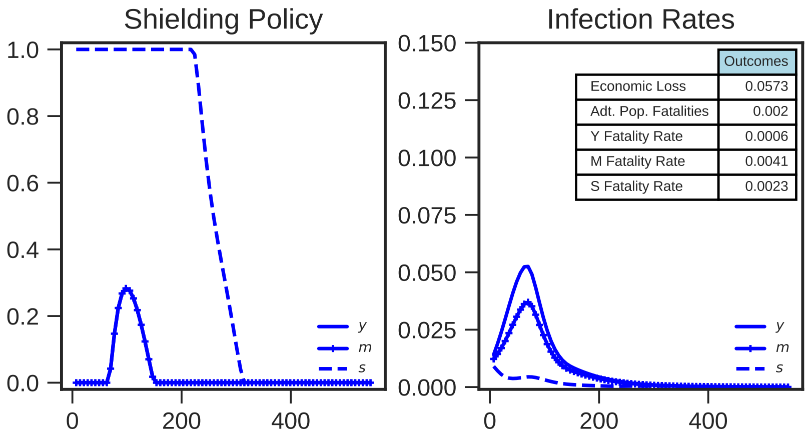
(i)



	Outcomes
Economic Loss	0.1258
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0004
M Fatality Rate	0.0025
S Fatality Rate	0.0043

RC SEIR, vulnGD 50% and MED 50%: SF Uniform Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.375$

(ii)



	Outcomes
Economic Loss	0.0573
Adt. Pop. Fatalities	0.002
Y Fatality Rate	0.0006
M Fatality Rate	0.0041
S Fatality Rate	0.0023

C SEIR, vulnGD 50% and MED 50%: SF SemiTargeted Policy for $\theta = 0.75$ $\alpha = 2.0$ $\eta = .9$ $\rho = 0.37$

Figure 6.23: **Optimal policy, combination of improved medical treamtent and group distancing towards the vulnerable.**

Combination of group distancing towards the vulnerable (reduction of contact rates ρ_{ys}, ρ_{ms} by 50%) and improved medical treatment (50% lower mortality rate for the senior group). Panel (i): Optimal uniform policy. Panel (ii): Optimal semi-targeted policy

Chapter 7

Conclusive Summary

The first four papers presented in this doctoral thesis provide implementations and empirical applications of the double machine learning framework. Chapter 2 provides an introduction to double machine learning as implemented in the R package and its python twin. The survey and the examples provided in Chapter 3 illustrate how valid simultaneous inference can be performed in high-dimensional settings. The study presented in Chapter 4 provides an estimation framework for valid inference in high-dimensional additive models. An analysis of a heterogeneous gender wage gap for full-time and year-round employees in the U.S. is presented in Chapter 5. Chapter 6 presents a framework for optimal shielding policies in a pandemic.

Each of the presented studies can be extended in future work. For example, by providing a flexible and easily extendable implementation of the double machine learning framework in Chapter 2, we hope to encourage empirical researchers and practitioners to use machine learning based methods for estimation of causal effects. Whereas the theoretical framework has been established in a sequence of papers, for example Belloni et al. (2014c) and Chernozhukov et al. (2018a), many practical questions still remain unanswered. Future studies might assess the role of important ingredients of the double machine learning estimators in empirical applications such as the sample splitting schedules, possible reweighting and refinement of the estimators and optimal tuning rules for the machine learning methods. Moreover, the set of causal models being implemented can be extended in the future. Regarding the inference procedure considered in Chapter 4, the considered class of models my be further generalized, for example, by relaxing the assumption of additivity. The gender wage gap analysis in Chapter 5 may be extended in various regards. For example, it may be interesting to consider a broader sample definition, covering the group of part-time employees, which plays an important role in gender inequality in earnings. Furthermore, the heterogeneity analysis might be translated to other important topics in economics such as labor market participation or returns to schooling. Finally, the framework considered in Chapter 6 may be extended in terms of the crucial model ingredients. For example, the contact patterns may be endogenized and, hence, made dependent on the current and past numbers of infections. Similarly, the detection technologies may depend on the number of infections. A more general extension would be to use the modeling approach that balances economic and public health costs to the optimal distribution of a vaccine.

Bibliography

Acemoglu, D., V. Chernozhukov, I. Werning, and M. D. Whinston (2020). *Optimal targeted lockdowns in a multi-group SIR model.* Working Paper 27102. National Bureau of Economic Research. URL: `http://www.nber.org/papers/w27102`.

Albrecht, J., M. A. Bronson, P. S. Thoursie, and S. Vroman (2018). "The career dynamics of high-skilled women and men: Evidence from Sweden". In: *European Economic Review* 105, pp. 83–102.

American Hospital Association (AHA) (2020). *Fast facts on U.S. hospitals, 2020.* `https://www.aha.org/statistics/fast-facts-us-hospitals` (accessed July 3, 2020).

Angelov, N., P. Johansson, and E. Lindahl (2016). "Parenthood and the gender gap in pay". In: *Journal of Labor Economics* 34.3, pp. 545–579.

Athey, S., J. Tibshirani, and S. Wager (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178. URL: `https://cran.r-project.org/package=grf`.

Bach, P., V. Chernozhukov, M. S. Kurz, and M. Spindler (2021). *DoubleML – An Object-Oriented Implementation of Double Machine Learning in Python.* arXiv:2104.03220. arXiv: `2104.03220 [stat.ML]`.

Bach, P., V. Chernozhukov, and M. Spindler (2018a). "Closing the U.S. gender wage gap requires understanding its heterogeneity". arXiv:1812.04345. URL: `https://arxiv.org/abs/1812.04345`.

Bach, P., V. Chernozhukov, and M. Spindler (2018b). "Valid simultaneous inference in high-dimensional settings (with the hdm package for R)". In: *arXiv preprint arXiv:1809.04951.*

Barber, R. F. and E. J. Candès (2015). "Controlling the false discovery rate via knockoffs". In: *Annals of Statistics* 43.5, pp. 2055–2085. URL: `https://doi.org/10.1214/15-AOS1337`.

Barber, R. F. and E. J. Candès (2019). "A knockoff filter for high-dimensional selective inference". In: *Annals of Statistics* 47.5, pp. 2504–2537. URL: `https://doi.org/10.1214/18-AOS1755`.

Baxter, E. (2015). *How the gender wage gap differs by occupation.* Center for American Progress.; Available at `https://www.americanprogress.org/issues/women/news/2015/04/14/110959/how-the-gender-wage-gap-differs-by-occupation/` (April 14, 2015).

Becker, M., M. Binder, B. Bischl, M. Lang, F. Pfisterer, N. G. Reich, J. Richter, P. Schratz, and R. Sonabend (Mar. 2021). *mlr3 book.* URL: `https://mlr3book.mlr-org.com`.

Becker, M., M. Lang, J. Richter, B. Bischl, and D. Schalk (2020). *mlr3tuning: Tuning for 'mlr3'.* R package version 0.7.0. URL: `https://CRAN.R-project.org/package=mlr3tuning`.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). "Sparse models and methods for optimal instruments with an application to eminent domain". In: *Econometrica* 80.6, pp. 2369–2429. URL: `http://dx.doi.org/10.3982/ECTA9626`.

Belloni, A. and V. Chernozhukov (2011). "High dimensional sparse econometric models: An introduction". In: *Inverse Problems and High-Dimensional Estimation.* Springer, pp. 121–156.

Belloni, A. and V. Chernozhukov (2013). "Least squares after model selection in high-dimensional sparse models". In: *Bernoulli* 19.2, pp. 521–547. URL: `http://dx.doi.org/10.3150/11-BEJ410`.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2018). “Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework”. In: *Annals of Statistics* 46.6B, pp. 3643–3675.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). “Program evaluation and causal inference with high-dimensional data”. In: *Econometrica* 85.1, pp. 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2011). “Inference for high-dimensional sparse econometric models”. In: *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010.* III:245–295.

Belloni, A., V. Chernozhukov, and K. Kato (2014a). “Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems”. In: *Biometrika* 102.1, 77–94. URL: `http://dx.doi.org/10.1093/biomet/asu056`.

Belloni, A., V. Chernozhukov, and L. Wang (2014b). “Pivotal estimation via square-root lasso in non-parametric regression”. In: *The Annals of Statistics* 42.2, pp. 757–788.

Belloni, A., V. Chernozukov, and C. Hansen (2014c). “Inference on treatment effects after selection among high-dimensional controls”. In: *The Review of Economic Studies* 81.2 (287), pp. 608–650. URL: `http://www.jstor.org/stable/43551575`.

Benjamini, Y. and Y. Hochberg (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300.

Berger, D. W., K. F. Herkenhoff, and S. Mongey (2020). *An SEIR infectious disease model with testing and conditional quarantine*. Working Paper. National Bureau of Economic Research.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Johns Hopkins University Press Baltimore.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). “Simultaneous analysis of Lasso and Dantzig selector”. In: *Annals of Statistics* 37.4, pp. 1705–1732.

Bilias, Y. (2000). “Sequential testing of duration data: The case of the Pennsylvania ‘reemployment bonus’ experiment”. In: *Journal of Applied Econometrics* 15.6, pp. 575–594.

Blau, F. D. and L. M. Kahn (2017). “The gender wage gap: Extent, trends, and explanations”. In: *Journal of Economic Literature* 55.3, pp. 789–865. URL: `http://www.aeaweb.org/articles?id=10.1257/jel.20160995`.

Blinder, A. S. (1973). “Wage discrimination: Reduced form and structural Estimates”. In: *Journal of Human Resources*, pp. 436–455.

Breiman, L. (2001). “Random Forests”. In: *Machine Learning* 45 (1), pp. 5–32. URL: `http://dx.doi.org/10.1023/A:1010933404324`.

Brotherhood, L., P. Kircher, C. Santos, and M. Tertilt (2020). *An economic model of the COVID-19 epidemic: The importance of testing and age-specific policies*. CRC TR 224 Discussion Paper Series crctr224_2020_175. University of Bonn and University of Mannheim, Germany. URL: `https://ideas.repec.org/p/bon/boncrc/crctr224_2020_175.html`.

Bühlmann, P. and S. van De Geer (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer-Verlag, New York.

Bühlmann, P. and S. van de Geer (2015). “High-dimensional inference in misspecified linear models”. In: *Electronic Journal of Statistics* 9.1, pp. 1449–1473.

Bundeszentrale für politische Bildung und Statistisches Bundesamt (2018). *Ein Sozialbericht für die Bundesrepublik Deutschland, Datenreport 2018 Bd. 2018.*

Bütikofer, A., S. Jensen, and K. G. Salvanes (2018). “The role of parenthood on the gender gap among top earners”. In: *European Economic Review* 109. Gender Differences in the Labor Market, pp. 103 –123. URL: `http://www.sciencedirect.com/science/article/pii/S0014292118300874`.

Candès, E., Y. Fan, L. Janson, and J. Lv (2018). “Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection”. In: *Journal of the Royal Statistical Society: Series B* 80.3, pp. 551–577.

Chang, W. (2020). *R6: Encapsulated classes with reference semantics*. R package version 2.5.0. URL: `https://CRAN.R-project.org/package=R6`.

Chen, R. Y., A. Gittens, and J. A. Tropp (2012). “The masked sample covariance estimator: an analysis using matrix concentration inequalities”. In: *Information and Inference: A Journal of the IMA* 1.1, pp. 2–20.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018a). “Double/debiased machine learning for treatment and structural parameters”. In: *The Econometrics Journal* 21.1, pp. C1–C68. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12097`.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). “Gaussian approximation of suprema of empirical processes”. In: *The Annals of Statistics* 42.4, pp. 1564–1597.

Chernozhukov, V., D. Chetverikov, K. Kato, et al. (2013a). “Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors”. In: *The Annals of Statistics* 41.6, pp. 2786–2819.

Chernozhukov, V., I. Fernández-Val, and Y. Luo (2018b). “The sorted effects method: Discovering heterogeneous effects beyond their averages”. In: *Econometrica* 86.6, pp. 1911–1938.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013b). “Inference on counterfactual distributions”. In: *Econometrica* 81.6, pp. 2205–2268.

Chernozhukov, V., C. Hansen, and M. Spindler (2016a). “hdm: High-dimensional metrics”. In: *R Journal* 8.2, pp. 185–199. URL: `https://journal.r-project.org/archive/2016/RJ-2016-040/index.html`.

Chernozhukov, V., C. Hansen, and M. Spindler (2016b). “High-dimensional metrics in R”. In: *arXiv preprint arXiv:1603.01700*.

Chernozhukov, V., C. Hansen, and M. Spindler (2015a). “Post-selection and post-regularization inference in linear models with many controls and instruments”. In: *American Economic Review* 105.5, pp. 486–90.

Chernozhukov, V., C. Hansen, and M. Spindler (2015b). “Valid post-selection and post-regularization inference: An elementary, general approach”. In: *Annual Review of Economics* 7.1, pp. 649–688. eprint: `https://doi.org/10.1146/annurev-economics-012315-015826`. URL: `https://doi.org/10.1146/annurev-economics-012315-015826`.

Chernozhukov, V., C. Hansen, and M. Spindler (2018+). “Heterogeneity in the (high-dimensional) Mincer equation (working title)”. In: *Work in Progress*.

Chu, D. K., E. A. Akl, S. Duda, K. Solo, S. Yaacoub, H. J. Schünemann, A. El-harakeh, A. Bognanni, T. Lotfi, M. Loeb, et al. (2020). “Physical Distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: A systematic review and meta-analysis”. In: *The Lancet*.

Claeskens, G. and I. Keilegom (2003). “Bootstrap confidence bands for regression curves and their derivatives”. In: *Annals of Statistics* 31.

Correll, S. J., S. Benard, and I. Paik (2007). “Getting a job: Is there a motherhood penalty?” In: *American Journal of Sociology* 112.5, pp. 1297–1338.

Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin (DIVI) (2020). *Daily report DIVI Intensivregister (July 3, 2020)*. Report.

Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen (2015). “High-dimensional inference: Confidence intervals, p-values and R-Software hdi”. In: *Statistical Science* 30.4, pp. 533–558.

Doksum, K. and A. Samarov (1995). "Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression". In: *The Annals of Statistics* 23.5, pp. 1443–1473.

Dowle, M. and A. Srinivasan (2020). *data.table: Extension of 'data.frame'*. R package version 1.13.2. URL: `https://CRAN.R-project.org/package=data.table`.

Fan, J. and W. Zhang (2000). "Simultaneous confidence bands and hypothesis testing in varying-coefficient models". In: *Scandinavian Journal of Statistics* 27.4, pp. 715–731. URL: `http://www.jstor.org/stable/4616637`.

Farbmacher, H., R. Guber, and S. Klaassen (2020). "Instrument validity tests with causal forests". In: *Journal of Business and Economic Statistics*.

Ferguson, N., D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg, et al. (2020). "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand". In.

Fortin, N., T. Lemieux, and S. Firpo (2011). "Decomposition methods in economics". In: *Handbook of Labor Economics* 4, pp. 1–102.

Friedman, J. H. and W. Stuetzle (1981). "Projection pursuit regression". In: *Journal of the American Statistical Association* 76.376, pp. 817–823.

Gimenez, J. R. and J. Zou (2019). "Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization". In: *Proceedings of Machine Learning Research*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 2184–2192. URL: `http://proceedings.mlr.press/v89/gimenez19b.html`.

Goldin, C. (2014). "A grand gender convergence: Its last chapter". In: *The American Economic Review* 104.4, pp. 1091–1119.

Gregory, K., E. Mammen, and M. Wahl (2016). "Statistical inference in sparse high-dimensional additive models". In: URL: `https://arxiv.org/abs/1603.07632`.

Grimm, V., F. Mengel, and M. Schmidt (2020). "Extensions of the SEIR model for the analysis of tailored social distancing and tracing approaches to cope with COVID-19". In: *medRxiv*. URL: `https://www.medrxiv.org/content/early/2020/04/29/2020.04.24.20078113`.

Harrison Jr, D. and D. L. Rubinfeld (1978). "Hedonic housing prices and the demand for clean air". In: *Journal of Environmental Economics and Management* 5.1, pp. 81–102.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Vol. 43. Chapman and Hall, Ltd., London.

Holm, S. (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian Journal of Statistics*, pp. 65–70.

Horby, P. and M. Landray (2020). *Low-cost dexamethasone reduces death by up to one third in hospitalised patients with severe respiratory complications of COVID-19*. RECOVERY Trial Press Release. URL: `\url{https://www.ox.ac.uk/news/2020-06-16-low-cost-dexamethasone-reduces-death-one-thirdhospitalised-patients-severe}`.

Härdle, W. (1989). "Asymptotic maximal deviation of M-smoothers". In: *Journal of Multivariate Analysis* 29.2, pp. 163–179. URL: `https://ideas.repec.org/a/eee/jmvana/v29y1989i2p163-179.html`.

Huang, J., J. Horowitz, and F. Wei (Aug. 2010). "Variable selection in nonparametric additive models". In: *Annals of Statistics* 38, pp. 2282–2313.

Imbens, G. W. and J. D. Angrist (1994). "Identification and estimation of local average treatment effects". In: *Econometrica* 62.2, pp. 467–475. URL: `http://www.jstor.org/stable/2951620`.

Janson, L. and W. Su (2016). "Familywise error rate control via knockoffs". In: *Electronic Journal of Statistics* 10.1, pp. 960–975. URL: `https://doi.org/10.1214/16-EJS1129`.

Javanmard, A. and A. Montanari (2014). “Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory”. In: *IEEE Transactions on Information Theory* 60.10, pp. 6522–6554.

Javanmard, A. and A. Montanari (2018). “Debiasing the lasso: Optimal sample size for gaussian designs”. In: *The Annals of Statistics* 46.6A, pp. 2593–2622.

Kato, K. (2012). “Two-step estimation of high dimensional additive models”. In: URL: `https://arxiv.org/abs/1207.5313`.

Klaassen, S., J. Kück, M. Spindler, and V. Chernozhukov (2018). “Uniform inference in high-dimensional gaussian graphical models”. In: *arXiv preprint arXiv:1808.10532*.

Klepac, P., A. J. Kucharski, A. J. Conlan, S. Kissler, M. Tang, H. Fry, and J. R. Gog (2020). “Contacts in context: Large-scale setting-specific social mixing matrices from the BBC pandemic project”. In: *medRxiv*.

Kleven, H., C. Landais, and J. E. Søgaard (2019). “Children and gender inequality: Evidence from Denmark”. In: *American Economic Journal: Applied Economics* 11.4, pp. 181–209.

Knaus, M. C. (2018). “A double machine learning approach to estimate the effects of musical practice on student’s skills”. In: *arXiv preprint arXiv:1805.10300*. URL: `https://github.com/MCKnaus/dmlmt`.

Knaus, M. C. (2020). “Double machine learning based program evaluation under unconfoundedness”. In: *arXiv preprint arXiv:2003.03191*. URL: `https://github.com/MCKnaus/causalDML`.

Koltchinskii, V. and M. Yuan (Dec. 2010). “Sparsity in multiple kernel learning”. In: *Annals of Statistics* 38.6, pp. 3660–3695. URL: `https://doi.org/10.1214/10-AOS825`.

Kong, E. and Y. Xia (2012). “A single-index quantile regression model and its estimation”. In: *Econometric Theory* 28.4, pp. 730–768.

Kozbur, D. (Mar. 2015). “Inference in additively separable models with a high dimensional conditioning set”. In: *SSRN Electronic Journal*.

Kurz, M. S. (2021). “Distributed Double Machine Learning with a Serverless Architecture”. In: ICPE ’21. Virtual Event, France: Association for Computing Machinery, pp. 27–33.

Lang, M., Q. Au, S. Coors, and P. Schratz (2020a). *mlr3learners: Recommended learners for ’mlr3’*. R package version 0.4.3. URL: `https://CRAN.R-project.org/package=mlr3learners`.

Lang, M., M. Binder, J. Richter, P. Schratz, F. Pfisterer, S. Coors, Q. Au, G. Casalicchio, L. Kotthoff, and B. Bischl (2019). “mlr3: A modern object-oriented machine learning framework in R”. In: *Journal of Open Source Software*. URL: `https://joss.theoj.org/papers/10.21105/joss.01903`.

Lang, M., B. Bischl, J. Richter, X. Sun, and M. Binder (2020b). *paradox: Define and work with parameter spaces for complex algorithms*. R package version 0.7.1. URL: `https://CRAN.R-project.org/package=paradox`.

Leeb, H. and B. M. Pötscher (2008). “Recent developments in model selection and related areas”. In: *Econometric Theory* 24.2, pp. 319–322. URL: `http://dx.doi.org/10.1017/S0266466608080134`.

Leisch, F. and E. Dimitriadou (2010). *mlbench: Machine learning benchmark problems*. R package version 2.1-1.

Lin, Y. and H. H. Zhang (Oct. 2006). “Component selection and smoothing in multivariate nonparametric regression”. In: *Annals of Statistics* 34.5, pp. 2272–2297. URL: `https://doi.org/10.1214/009053606000000722`.

List, J. A., A. M. Shaikh, and Y. Xu (2019). “Multiple hypothesis testing in experimental economics”. In: *Experimental Economics* 22.4, pp. 773–793.

Lou, Y., J. Bien, R. Caruana, and J. Gehrke (2016). “Sparse partially linear additive models”. In: *Journal of Computational and Graphical Statistics* 25.4, pp. 1126–1140.

Lu, J., M. Kolar, and H. Liu (2020). “Kernel meets sieve: Post-regularization confidence bands for sparse additive model”. In: *Journal of the American Statistical Association* 0.ja, pp. 1–16. URL: `https://doi.org/10.1080/01621459.2019.1689984`.

Manning, A. and J. Swaffield (2008). “The gender gap in early-career wage growth”. In: *The Economic Journal* 118.530, pp. 983–1024.

Meier, L., S. Van de Geer, and P. Bühlmann (2009). “High-dimensional additive modeling”. In: *Annals of Statistics* 37.6B, pp. 3779–3821.

Microsoft Research (2019). *EconML: A Python package for ML-based heterogeneous treatment effects estimation.* https://github.com/microsoft/EconML. Version 0.x.

Mueller, G. and E. Plug (2006). “Estimating the effect of personality on male and female earnings”. In: *ILR Review* 60.1, pp. 3–22.

Nelson, L. (2016). *Equal pay day: The most unequal jobs in America.* Vox Media. `https://www.vox.com/2016/4/12/11413246/equal-pay-women-jobs` (April 12, 2016).

Neumark, D., R. J. Bank, and K. D. Van Nort (1996). “Sex discrimination in restaurant hiring: An audit study”. In: *The Quarterly Journal of Economics* 111.3, pp. 915–941.

Newey, W. K. (1994). “The asymptotic variance of semiparametric estimators”. In: *Econometrica: Journal of the Econometric Society*, pp. 1349–1382.

Newman, D. J., S. Hettich, C. L. Blake, and C. J. Merz (1998). *UCI Repository of machine learning databases.* URL: `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Neyman, J. (1959). “Optimal asymptotic tests of composite statistical hypotheses”. In: *Probability and statistics*, pp. 57–213.

Oaxaca, R. (1973). “Male-female wage differentials in urban labor markets”. In: *International Economic Review*, pp. 693–709.

Organisation for Economic Cooperation and Development (2020). *Beyond containment: Health systems responses to COVID-19 in the OECD.* Report.

Overberg, P. and J. Adamy (2016). *What’s your pay gap?* The Wall Street Journal. 17 May 2016; `http://graphics.wsj.com/gender-pay-gap/`.

Patterson, E. and M. Sesia (2018). *knockoff: The Knockoff filter for controlled variable selection.* R package version 0.3.2. URL: `https://CRAN.R-project.org/package=knockoff`.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay (2011). “Scikit-learn: Machine learning in python”. In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. URL: `http://jmlr.org/papers/v12/pedregosa11a.html`.

Permutt, T. and J. R. Hebel (1989). “Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight”. In: *Biometrics*, pp. 619–622.

Petersen, A., D. Witten, and N. Simon (2016). “Fused lasso additive model”. In: *Journal of Computational and Graphical Statistics* 25.4. PMID: 28239246, pp. 1005–1025.

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2009). “Sparse additive models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.5, pp. 1009–1030. URL: `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2009.00718.x`.

Robert-Koch-Institut (RKI) (2020a). *SARS-CoV-2 Steckbrief zur Coronavirus-Krankheit-2019 (COVID-19) Stand: 10.07.2020.* `https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Steckbrief.html`.

Robert-Koch-Institut (RKI) (2020b). *Täglicher Lagebericht des RKI zur Coronavirus-Krankheit-2019 (COVID-19), Stand: 10.08.2020.* `https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/2020-08-10-de.pdf?__blob=publicationFile`.

Robins, J. M. and A. Rotnitzky (1995). "Semiparametric efficiency in multivariate regression models with missing data". In: *Journal of the American Statistical Association* 90.429, pp. 122–129.

Robinson, P. M. (1988). "Root-N-consistent semiparametric regression". In: *Econometrica: Journal of the Econometric Society*, pp. 931–954.

Romano, J. P. and M. Wolf (2005a). "Exact and approximate stepdown methods for multiple hypothesis testing". In: *Journal of the American Statistical Association* 100.469, pp. 94–108.

Romano, J. P. and M. Wolf (2005b). "Stepwise multiple testing as formalized data snooping". In: *Econometrica* 73.4, pp. 1237–1282. URL: `http://dx.doi.org/10.1111/j.1468-0262.2005.00615.x`.

Romano, J. P. and M. Wolf (2016). "Efficient computation of adjusted p-values for resampling-based stepdown multiple testing". In: *Statistics & Probability Letters* 113, pp. 38 –40. URL: `http://www.sciencedirect.com/science/article/pii/S0167715216000389`.

Ruggles, S., S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas, and M. Sobek (2020). "IPUMS USA: Version 10.0 [dataset]". In: *Minneapolis, MN: IPUMS.* URL: `https://doi.org/10.18128/D010.V10.0`.

Sardy, S. and P. Tseng (2004). "AMlet, RAMlet, and GAMlet: Automatic nonlinear fitting of additive models, robust and generalized, with wavelets". In: *Journal of Computational and Graphical Statistics* 13.2, pp. 283–309. URL: `http://www.jstor.org/stable/1391177`.

Schick, A. (1986). "On asymptotically efficient estimation in semiparametric models". In: *The Annals of Statistics*, pp. 1139–1151.

Sigle-Rushton, W. and J. Waldfogel (2007). "Motherhood and women's earnings in Anglo-American, Continental European, and Nordic countries". In: *Feminist Economics* 13.2, pp. 55–91. URL: `\url{https://doi.org/10.1080/13545700601184849}`.

Sonabend, R. and P. Schratz (2020). *mlr3extralearners: Extra learners For mlr3.* R package version 0.3.0.9000.

Statistisches Bundesamt (2019). *Bevölkerung und Erwerbstätigkeit, Erwerbsbeteiligung der Bevölkerung, Ergebnisse des Mikrozensus zum Arbeitsmarkt.*

Statistisches Bundesamt (2020). *Bevölkerung und Erwerbstätigkeit, Bevölkerungsfortschreibung auf Grundlage des Zensus 2011.*

Stone, C. J. (June 1985). "Additive regression and other nonparametric models". In: *Annals of Statistics* 13.2, pp. 689–705. URL: `https://doi.org/10.1214/aos/1176349548`.

Sun, J. and C. R. Loader (Sept. 1994). "Simultaneous confidence bands for linear regression and smoothing". In: *Annals of Statistics* 22.3, pp. 1328–1345. URL: `https://doi.org/10.1214/aos/1176325631`.

Szitas, J. (2019). *postDoubleR: Post double selection with double machine learning.* URL: `https://CRAN.R-project.org/package=postDoubleR`.

The American Association of University Women (2018). "The simple truth about the gender pay gap". In: *Spring 2018 Edition.* Ed. by K. Bibler.

Tibshirani, J., S. Athey, and S. Wager (2020). *grf: Generalized random forests.* R package version 1.2.0. URL: `https://CRAN.R-project.org/package=grf`.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". In: *J. Roy. Statist. Soc. Ser. B* 58, pp. 267–288.

United States Census Bureau (2019). *Population estimates show aging across race groups differs.* `https://www.census.gov/newsroom/press-releases/2019/estimates-characteristics.html` (accessed July 10, 2020).

U.S. Bureau of Labor Statistics (2017). *Highlights of women's earnings in 2016.* Report 1069. URL: `\url{https://www.bls.gov/opub/reports/womens-earnings/2016/pdf/home.pdf}`.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (June 2014). "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *Annals of Statistics* 42.3, pp. 1166–1202. URL: `https://doi.org/10.1214/14-AOS1221`.

Van der Laan, M. J. and S. Rose (2011). *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic statistics.* Vol. 3. Cambridge university press.

Van der Vaart, A. W. and J. A. Wellner (1996). "Weak convergence". In: *Weak convergence and empirical processes.* Springer, pp. 16–28.

Vasel, K. (2017). *5 things to know about the gender pay gap.* CNN. 4 April 2017; `https://money.cnn.com/2017/04/04/pf/equal-pay-day-gender-pay-gap/index.html`.

Waldfogel, J. (1998). "Understanding the 'family gap' in pay for women with children". In: *The Journal of Economic Perspectives* 12.1, pp. 137–156. URL: `http://www.jstor.org/stable/2646943`.

Wickham, H. (2019). *Advanced R.* CRC press.

Wood, S. N. (2017). *Generalized additive models: an introduction with R.* CRC press.

Wright, M. N. and A. Ziegler (2017). "ranger: A fast implementation of random forests for high dimensional data in C++ and R". In: *Journal of Statistical Software* 77.1, pp. 1–17.

Wright, P. G. (1928). *Tariff on animal and vegetable oils.* Macmillan Company, New York.

Zhang, C.-H. and S. S. Zhang (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 217–242. URL: `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12026`.

Zhang, W. and H. Peng (2010). "Simultaneous confidence band and hypothesis test in generalised varying-coefficient models". In: *Journal of Multivariate Analysis* 101.7, pp. 1656 –1680. URL: `http://www.sciencedirect.com/science/article/pii/S0047259X10000539`.

Appendices

A.1 Statement of Personal Contribution Pursuant to §6(4) PromO

		Personal contribution		
Resear Project	**Co-Autors**	**Conception/Design**	**Execution**	**Reporting**
DoubleML - An Object-Oriented Implementation of Double Machine Learning in R (Chapter 2)	Victor Chernozhukov Malte S. Kurz Martin Spindler	• Definition of research question • Contribution to conception of package structure and class organization	• Literature and software research and review • Implementation in R and (partly) in python • Code examples and simulation studies • Visualization of results	• Contribution to user guide and package documentation • Prepared manuscript
Valid Simultaneous Inference in High-Dimensional Settings with the hdm Package for R (Chapter 3)	Victor Chernozhukov Martin Spindler	• Definition of research question	• Literature research and review • Implementation of algorithm, simulation study and real-data example • Summary, illustration and interpretation of results	• Prepared package documentation • Prepared reproducible data examples • Prepared manuscript
Uniform Inference in High-Dimensional Additive Models (Chapter 4)	Sven Klaassen Jannis Kueck Martin Spindler	• Definition of research question	• Literature research • Implementation of algorithm and simulation study • Empirical example (boston housing data) • Visualization and interpretation of results	• Prepared manuscript

		Personal contribution		
Resear Project	**Co-Autors**	**Conception/Design**	**Execution**	**Reporting**
Heterogeneity in the U.S. Gender Wage Gap (Chapter 5)	Victor Chernozhukov Martin Spindler	• Definition of research question	• Literature research • Data preparation • Implementation and empirical analysis • Visualization of results	• Prepared manuscript • Presentation at workshops: IAB Causal ML Workshop (Erlangen-Nuremberg, 2020), SMILES Summer School (Moscow, 2020), HCHE Center Day (Hamburg, 2019), German Economic Association (Leipzig, 2019), Brownbag Seminar TU Dresden (Dresden, 2018), German Statistical Week, Young Researcher Seminar (Linz, 2018), Machine Learning in Economics and Econometrics (Munich, 2018).
Insights from Optimal Pandemic Shielding in a Multi-Group SEIR Framework (Chapter 6)	Victor Chernozhukov Martin Spindler	• Definition of research question	• Literature research • Implementation: Extension of code of Acemoglu et al. (2020, A multi-risk SIR model with optimally targeted lockdown, NBER Working Paper) • Adaption of model setting to Germany • Model adaption to SEIR-Framework and extensions • Visualization and interpretation of results	• Prepared manuscript • Presentation at HCHE Center Day (Hamburg, 2020)

A.2 Short Summary of Papers Pursuant to §6(6) PromO

Short summary in English language

DoubleML - An Object-Oriented Implementation of Double Machine Learning in R (Chapter 2)

The R package `DoubleML` implements the double/debiased machine learning framework of Chernozhukov et al. (2018). It provides functionalities to estimate parameters in causal models based on machine learning methods. The double machine learning framework consist of three key ingredients: Neyman orthogonality, high-quality machine learning estimation and sample splitting. Estimation of nuisance components can be performed by various state-of-the-art machine learning methods that are available in the `mlr3` ecosystem. `DoubleML` makes it possible to perform inference in a variety of causal models, including partially linear and interactive regression models and their extensions to instrumental variable estimation. The object-oriented implementation of `DoubleML` enables a high flexibility for the model specification and makes it easily extendable. This paper serves as an introduction to the double machine learning framework and the R package `DoubleML`. In reproducible code examples with simulated and real data sets, we demonstrate how `DoubleML` users can perform valid inference based on machine learning methods.

Valid Simultaneous Inference in High-Dimensional Settings with the hdm Package for R (Chapter 3)

Due to the increasing availability of high-dimensional empirical applications, researchers and practictioners across all disciplines frequently encounter situations where they have to test many hypotheses at the same time. Addressing multiple testing issues and methodological shortcomings of classical linear regression becomes essential to obtain reliable results. This paper provides a selective review of methods to perform simultaneous inference in high-dimensional settings. It does so by summarizing inferential approaches based on regularized estimation in combination with classical methods to correct for multiple testing. Moreover, we conduct a simulation study to compare the methods in a high-dimensional setting. Finally, we illustrate how the R package `hdm` can be used to perform valid simultaneous inference in a replicable real-data example in the context of the gender wage gap.

Uniform Inference in High-Dimensional Additive Models (Chapter 4)

We develop a method for uniformly valid confidence bands of a nonparametric component f_1 in the additive model $Y = f_1(X_1) + \ldots + f_p(X_p) + \varepsilon$ in a high-dimensional setting. We employ sieve estimation and embed it in a high-dimensional Z-estimation framework that allows us to construct uniformly valid confidence bands for the first component f_1. Our study extends the existing results for inference in high-dimensional additive models and clarifies the required assumptions. In a setting where the number of regressors p may increase with the sample size, a sparsity assumption is critical for our analysis. Moreover, we run simulation studies that show that our proposed method delivers reliable results concerning the estimation and coverage properties even in small samples. Finally, we illustrate our procedure in an empirical application demonstrating the implementation and the use of the proposed method in practice.

Heterogeneity in the U.S. Gender Wage Gap (Chapter 5)

As a measure of gender inequality, the gender wage gap has come to play an important role both in academic research and the public debate. In 2016, the majority of full-time employed women in the U.S. earned significantly less than comparable men. The extent to which women were affected by gender inequality in earnings, however, depended greatly on socio-economic characteristics, such as marital status or educational attainment. In this paper, we analyze data from the 2016 American Community Survey using a high-dimensional wage regression and applying double lasso to quantify heterogeneity in the gender wage gap. We find that the wage gap varied substantially across women and that the magnitude of the gap varied primarily by marital status, having children at home, race, occupation, industry, and educational attainment. We recommend that policy makers use these insights to design policies that will reduce discrimination and unequal pay more effectively.

Insights from Optimal Pandemic Shielding in a Multi-Group SEIR Framework (Chapter 6)

The COVID-19 pandemic constitutes one of the largest threats in recent decades to the health and economic welfare of populations globally. In this paper, we analyze different types of policy measures designed to fight the spread of the virus and minimize economic losses. Our analysis builds on a multi-group SEIR model, which extends the multi-group SIR model introduced by Acemoglu et al. (2020). We adjust the underlying social interaction patterns and consider an extended set of policy measures. The model is calibrated for Germany. Despite the trade-off between COVID-19 prevention and economic activity that is inherent to shielding policies, our results show that efficiency gains can be achieved by targeting such policies towards different age groups. Alternative policies such as physical distancing can be employed to reduce the degree of targeting and the intensity and duration of shielding. Our results show that a comprehensive approach that combines multiple policy measures simultaneously can effectively mitigate population mortality and economic harm.

Kurzzusammenfassung in deutscher Sprache

DoubleML - Eine objekt-orientierte Implementierung des Double Machine Learning Ansatzes in R (Kapitel 2)

Das R-Paket `DoubleML` ist eine Implementierung des Double/Debiased Machine Learning Ansatzes von Chernozhukov et al. (2018). Es bietet Funktionalitäten zur Schätzung struktureller Parameter in kausalen Modellen basierend auf Methoden des maschinellen Lernens. Das Double Machine Learning Framework besitzt drei wesentliche Bestandteile: Neyman-Orthogonalität, hochqualitative Schätzung mittels Maschinellen Lernens und Sample Splitting. Die Schätzung der Nuisance-Komponenten kann mit zahlreichen state-of-the-art Methoden des Maschinellen Lernens durchgeführt werden, die im `mlr3` Ökosystem verfügbar sind. `DoubleML` ermöglicht die Durchführung von Inferenz in einer Vielzahl kausaler Modelle, einschließlich des partiell-linearen und des interaktiven Regressionsmodells, sowie deren Erweiterungen zur Schätzung mit Instrumentenvariablen. Die objekt-orientierte Implementierung von `DoubleML` ermöglicht eine hohe Flexibilität hinsichtlich der Modellspezifikation und erleichtert zusätzliche Erweiterungen. Diese Studie bietet eine Einführung in das Double Machine Learning Framework und das R Paket `DoubleML`. In reproduzierbaren Code-Beispielen mit realen und simulierten Daten wird veranschaulicht, wie Benutzer von `DoubleML` valide Inferenz auf Basis von Methoden des Maschinellen Lernens durchführen können.

Valide simultane Inferenz in hochdimensionalen Modellen mit dem R-Paket hdm (Kapitel 3)

Aufgrund der zunehmenden Verfügbarkeit hochdimensionaler empirischer Anwendungen, sehen sich Forscher und Anwender sämtlicher Disziplinen immer häufiger mit Situationen konfrontiert, in denen gleichzeitig eine Vielzahl an Hypothesen getestet werden muss. Die Berücksichtigung simultaner Inferenz, sowie Grenzen klassischer linearer Regressionsverfahren gewinnt zunehmend an Bedeutung, um belastbare Ergebnisse zu erhalten. Die vorliegende Studie bietet eine selektive Literaturübersicht zu Methoden simultaner Inferenz in hochdimensionalen Modellen. Dabei werden Inferenzansätze, welche auf Regularisierungsmethoden basieren, sowie klassische Korrekturverfahren für mutliples Testen vorgestellt. Darüber hinaus wird im Rahmen einer Simulationsstudie die Performanz der vorgestellten Methoden miteinander verglichen. Abschließend wird in der Arbeit vorgestellt, inwiefern das R-Paket `hdm` dazu verwendet werden kann, um Methoden für valide simultane Inferenz zu verwenden. Die Verwendung dieser Methoden wird in einem realen Datenbeispiel veranschaulicht.

Gleichmäßige Inferenz in hochdimensionalen additiven Modellen (Kapitel 4)

Wir entwickeln eine Methode für gleichmäßig valide Konfidenzbänder für eine nichtparametrische Komponente f_1 in einem hochdimensionalen additiven Modell der Form $Y = f_1(X_1) + \ldots + f_p(X_p) + \varepsilon$. Dafür verwenden wir Sieve-Schätzung und binden diese in eine hochdimensionale Z-Schätungsumgebung ein, um gleichmäßig valide Konfidenzbänder für die Komponente f_1 zu konstruieren. Unsere Studie erweitert bestehende Ergebnisse für Inferenz in hochdimensionalen additiven Modellen und erläutert die zugrundeliegenden Annahmen. In einer Situation, in der die Anzahl an Regressoren p mit dem Stichprobenumfang steigen kann, kommt einer Sparsity-Annahme eine zentrale Bedeutung zu. In einer Simulationsstudie wird gezeigt, dass das vorgeschlagene Schätzverfahren belastbare Ergebnisse in endlichen, sowie in kleinen Stichproben liefert. Abschließend veranschaulichen wir unser Verfahren anhand einer empirischen Anwendung, die die Implementierung und Anwendung der vorgeschlagenen Methode in der Praxis demonstriert.

Heterogenität im Gender Pay Gap in den USA (Kapitel 5)

Als ein Maß für geschlechtsspezifische Ungleichheit spielt der Gender Pay Gap eine wichtige Rolle - sowohl in der akademischen Forschung als auch in der öffentlichen Debatte. Im Jahr 2016 verdiente die Mehrheit der vollzeitbeschäftigten Frauen in den USA signifikant weniger als vergleichbare Männer. Das Ausmaß, in dem Frauen von geschlechtsspezifischen Lohnungleichheit betroffen sind, hängt dabei zu einem großen Teil von sozioökonomischen Charakteristika ab, zum Beispiel von Familienstand oder Bildungshintergrund. In dieser Studie analysieren wir Daten des 2016 American Community Survey, um Heterogentität im Gender Pay Gap zu quantifizieren. Dazu wird eine hochdimensionale Lohnregression mithilfe des Double-Lasso-Schätzers geschätzt. Unsere Ergebnisse deuten darauf hin, dass der Gender Pay Gap deutlich von Frau zu Frau variiert und vor allem davon bestimmt wird, in welchen Familienstand eine Frau lebt, ob sie gemeinsam mit Kindern im Haushalt wohnt, welcher Ethnie sie angehört, in welchem Beruf oder in welcher Industrie sie arbeitet, sowie von ihrem Bildungshintergrund. We empfehlen Entscheidungsträgern in der Politik, diese Erkenntnisse zu berücksichtigen, um Politikmaßnahmen abzuleiten, die Diskriminierung und Lohnungleichheit effektiv reduzieren.

Erkenntnisse aus optimalen Shielding-Maßnahmen während einer Pandemie in einem Multi-Gruppen SEIR Modell (Kapitel 6)

Die COVID-19 Pandemie stellt eine der größten Herausforderungen für die Gesundheit und den Wohlstand der weltweiten Bevölkerung in den vergangenen Jahren dar. In dieser Studie analysieren wir verschiedene Politikmaßnahmen, die die Verbreitung des Virus eindämmen und die ökonomischen Folgeschäden minimieren. Unsere Analyse basiert auf einem Multi-Gruppen SEIR Modell, welches eine Erweiterung zu einem kürzlich entwickelten SIR-Modell in Acemoglu et al. (2020) darstellt. Wir nehmen Anpassungen hinsichtlich der zugrundeliegenden sozialen Interaktionsmuster vor und betrachten eine erweiterte Auswahl an betrachteten Politikmaßnahmen. Die Kalibrierung des Modells bezieht sich auf Deutschland. Trotz des Zielkonflikts, der Shielding-Maßnanhmen zu eigen ist und welcher zwischen der Prävention neuer COVID-19 Fälle und gleichzeitig entstehender ökonomischer Schäden besteht, können Effiziengewinne erreicht werden. Diese sind möglich, wenn sich Shielding-Maßnahmen differenziert an unterschiedliche Altersgruppen richten. Alternative Politikmaßnahmen wie zum Beispiel Physical Distancing können dazu beitragen, dass das Ausmaß der Unterscheidung nach Altersgruppen reduziert wird. Unsere Ergebnisse zeigen zudem, dass ein kombinierter Ansatz, der verschiedene Maßnahmen miteinander verbindet, dazu beitragen kann, die Mortalität in der Gesamtbevölkerung bei gleichzeitig niedrigeren ökonomischen Einbußen zu begrenzen.

A.3 List of Publications Pursuant to §6 (6) PromO

Journal article	**Publication status**
Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2020). DoubleML - An Object-Oriented Implementation of Double Machine Learning in R	Working Paper
Bach, P., Chernozhukov, V., and Spindler, M. (2018). Valid Simultaneous Inference in High-Dimensional Settings with the hdm Package for R	Working Paper
Bach, P., Klaassen, S., Kueck, J. and Spindler, M. (2020). Uniform Inference in High-Dimensional Additive Models	Working Paper
Bach, P., Chernozhukov, V., and Spindler, M. (2018). Heterogeneity in the U.S. Gender Wage Gap	Working Paper
Bach, P., Chernozhukov, V., and Spindler, M. (2020). Insights from Optimal Pandemic Shielding in a Multi-Group SEIR Framework	Working Paper

Affidavit

Hiermit erkläre ich, Philipp Simeon Bach, an Eides statt, dass ich die Dissertation mit dem Titel

Applications in High-Dimensional Econometrics

selbständig – und bei einer Zusammenarbeit mit anderen Wissenschaftlerinnen und Wissenschaftlern gemäß den beigefügten Darstellungen nach §6 Abs. 4 der Promotionsordnung der Fakultät der Betriebswirtschaft vom 9. Juli 2014 – verfasst habe und keine anderen als die von mir angegebenen Hilfsmittel benutzt habe. Die den herangezogenen Werken wörtlich oder sinngemäß entnommenen Stellen sind als solche gekennzeichnet.

Ich versichere, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe und die Arbeit nicht schon in einem früheren Promotionsverfahren im In- oder Ausland angenommen oder als ungenügend beurteilt worden ist.