Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT FÜR MATHEMATIK, INFORMATIK UND NATURWISSENSCHAFTEN

INSTITUT FÜR EXPERIMENTALPHYSIK

BAYESIAN METHODS FOR XFI

Dissertation zur Erlangung des Doktorgrades
an der Universität Hamburg

vorgelegt von

Jonas Joachim Kreidelmeyer

Hamburg, 2026

Gutachter:innen und Prüfungskommission

Gutacher der Dissertation	Prof. Dr. Florian Grüner
	Prof. Dr. Rene Werner
Prüfungskommission	Prof. Dr. Florian Grüner
	Prof. Dr. Rene Werner
	Prof. Dr. Wolfgang Parak
	Prof. Dr. Markus Drescher
	Prof. Dr. Ludwig Mathey
Vorsitzender Prüfungskommission	Prof. Dr. Wolfgang Parak
Vorsitzender Promotionsausschuss Physik	Prof. Dr. Johannes Haller
Leiter des Fachbereichs Physik	Prof Dr. Markus Drescher
Dekan der MIN Fakultät	Prof Dr. Norbert Ritter

Datum der Disputation

Hamburg, 2026-03-18

Danksagung

Zunächst möchte ich mich bei Prof. Grüner bedanken, für das Vertrauen, mich mehr oder weniger machen zu lassen was ich für interessant und wichtig hielt. Bei Theresa Staufer, dass sie mir den ganzen weg über für alle Fragen beiseite stand und kritisches Feedback geliefert hat. Und bei der Prüfungskommission, dass sie sich die Zeit genommen haben sich mit meiner Arbeit auseinanderzusetzen.

Weiterhin möchte ich mich bei Florian Ziegler, Jannis Haak und Catharina Mewes bedanken, dass ihr über die letzten vier Jahre das Büro zu einem Ort gemacht habt, an dem man gerne ist, sowie für die zahlreichen in promptu Brainstormingrunden. Dank geht außerdem an alle die die Arbeit korrekturgelesen haben, besondres Thorben Schmirander und Theresa Staufer, durch eure Rückmeldungen habt ihr geholfen aus einem rohen Dokument so etwas wie eine lesbare Arbeit zu machen. Und letztlich an meine Freunde und Familie die mich über den ganzen Weg bestärkt und unterstützt haben.

Abstract

X-ray fluorescence imaging (XFI) has the potential to become a highly advantageous medical imaging technique. It offers functional imaging over arbitrary time intervals into significant tissue depth and with high spatial resolution. There are two key steps in processing XFI data, the evaluation of the energy spectra at a given point and the reconstruction of the image. Both of these tasks can be interpreted as inference tasks in the statistical sense. One powerful statistical framework for inference is given by the field of Bayesian statistics which revolves around updating beliefs using new information. This work attempts to extend the established XFI processing pipeline using various Bayesian techniques.

Firstly, this work discusses how x-ray fluorescence (XRF) spectra can be modeled as Poisson variables over a linear combination of base curves. This model then allows simple predictions for measurements at different tracer concentrations and dose levels, which is particularly useful for planning measurements using simulation data. This model is also validated against simulation data.

Next, linear curve fitting of spectral models is reexamined, with least squares, weighted least squares and the first order correction to the Poisson likelihood being studied. Additionally this work shows that projecting the uncertainty of the energy-bin-wise posterior into the parameter space provides a robust and useful uncertainty estimate. The error estimation for signal isolation via spectral subtraction is also examined. Here, the Bayesian posterior leads to a correction of the uncertainty estimate that effectively removes singular behavior in the low count regime. It is shown that both of these revised uncertainty estimates produce strictly more powerful z-tests for signal detection than the established methods. Finally a new method for incorporating data from multiple detectors into spectral processing is proposed.

To handle quantitative 3d image reconstruction several methods are evaluated. It is shown that the reconstruction of depth information using multiple detectors, but only one projection, is generally not feasible. However, hierarchical generative Bayesian models are capable of reconstructing 3d images from very few (e.g. eleven) projections at reasonable resolution with additional detectors being able to reduce the required dose at equal image quality. The uncertainty estimate this model provides are also shown to be usable for quantitative work. The resolution of the reconstruction is shown to be at the expected limits induced by the data dimensionality itself.

Zusammenfassung

Röntgen-Fluoreszenz-Bildgebung (XFI) hat das Potenzial, sich zu einer leistungsfähigen medizinischen Bildgebungsmethode zu entwickeln. Sie ermöglicht funktionelle Bildgebung über beliebige Zeiträume und mit hoher räumlicher Auflösung. Die Verarbeitung von XFI-Daten umfasst zwei zentrale Schritte: die Auswertung der Energiespektren an einem gegebenen Punkt und die Bildrekonstruktion. Beide Aufgaben lassen sich als statistische Inferenzprobleme auffassen. Ein besonders leistungsfähiger Rahmen für derartige Probleme wird durch die Bayes'sche Statistik bereitgestellt, deren Kern das Aktualisieren von Wissensständen anhand neuer Informationen bildet. Diese Arbeit erweitert die bestehende XFI-Datenverarbeitung um verschiedene Bayes'sche Methoden.

Zunächst wird gezeigt, wie sich Röntgenfluoreszenz-(XRF)-Spektren als Poisson-verteilte Zufallsvariablen über Linearkombinationen modellieren lassen. Dieses Modell erlaubt einfache Vorhersagen für Messungen bei unterschiedlichen Markerkonzentrationen und Strahlendosen und eignet sich insbesondere zur Versuchsplanung anhand von Simulationsdaten. Die Modellannahmen werden anhand von Monte-Carlo Simulationen validiert.

Anschließend wird die lineare Anpassung von Spektralmodellen untersucht. Dabei werden verschiedene Methoden betrachtet. Es wird gezeigt, dass die Projektion der Varianz der energiespezifischen Posteriorverteilung in den Parameterraum eine robuste und aussagekräftige Unsicherheitsabschätzung liefert. Die Fehlerabschätzung bei der Signalisolierung durch spektrale Subtraktion wird ebenfalls analysiert. Hier führt der Bayes'sche Ansatz zu einer Korrektur, die singuläres Verhalten im Bereich geringer Photonzahlen beseitigt. Beide korrigierten Unsicherheitsabschätzungen ermöglichen strengere und sensitivere Tests zur Signaldetektion als bisherige Verfahren. Darüber hinaus wird eine neue Methode zur gemeinsamen spektralen Auswertung von Messdaten mehrerer Detektoren vorgestellt.

Für die quantitative 3D-Bildrekonstruktion werden verschiedene Verfahren untersucht. Es zeigt sich, dass eine Tiefenrekonstruktion mit mehreren Detektoren, jedoch nur einer Projektion, im Allgemeinen nicht möglich ist. Hierarchische generative Bayes-Modelle erlauben allerdings die Rekonstruktion dreidimensionaler Bilder aus sehr wenigen (z.B. elf) Projektionen bei angemessener Auflösung, wobei zusätzliche Detektoren die benötigte Strahlendosis bei gleicher Bildqualität verringern können. Die von diesen Modellen gelieferten Unsicherheitsabschätzungen sind zudem für quantitative Auswertungen geeignet. Die erreichbare Rekonstruktionsauflösung entspricht dabei den durch die Dimensionalität des Datenraums gesetzten theoretischen Grenzen.

CONTENTS

List of Symbols

$\vec{a}$	vector valued object
$\mathbf{M}$	matrix valued object
$\mathbb{1}$	identity matrix
$\vec{\mathbb{1}}$	vector containing only ones
$a := b$	a is defined as b
$f(x) \propto g(x)$	$f(x)$ is proportional to $g(x)$
$\vec{a} \circ \vec{b}$	point-wise multiplication of $\vec{a}$ and $\vec{b}$
$\vec{a} \oslash \vec{b}$	point-wise division of $\vec{a}$ and $\vec{b}$
$\vec{a}^{\circ\vec{b}}$	point-wise exponentiation of $\vec{a}$ with $\vec{b}$
$\otimes$	tensor product
$\mathrm{diag}\,(\,\vec{x}\,)$	diagonal matrix with entries of $\vec{x}$
$\mathrm{diag}\,(\,\mathbf{M}\,)$	vector with the diagonal entries of $\mathbf{M}$
$P(x)$	probability / likelihood of x
$y\|x$	y conditional on x (usually in $P(y\|x)$)
$\mathcal{N}\,[\,\vec{\mu}, \Sigma\,]$	normal distribution with mean $\vec{\mu}$ and covariance Σ
$\mathrm{Poiss}\,[\,\lambda\,]$	Poisson distribution with rate λ
$\Gamma\,[\,\alpha, \beta\,]$	gamma distribution with shape α and rate β
$\mathrm{NegBin}\,[\,r, p\,]$	negative binomial distribution for r failures and probability p
$\delta(x)$	dirac delta or "degenerate" distribution
$\mathcal{U}\,[\,I\,]$	uniform distribution over I
$\mathrm{StudT}\,[\,k\,]$	Student's t distribution with k degrees of freedom
$\chi^2\,[\,k\,]$	χ^2 distribution with k degrees of freedom, if k is not specified the reduced χ^2 distribution is meant
$x \sim \mathcal{D}$	x is distributed according to $\mathcal{D}$
$\hat{x}$	an estimator for x
$\mathbb{E}\,[\,x\,]$	expectation value of x
$\mathrm{Var}\,[\,x\,]$	variance of x
$\mathrm{Cov}\,[\,\vec{x}\,]$	covariance of $\vec{x}$
$D_{\mathrm{KL}}\,(\,P\|\|Q\,)$	Kullback-Leibler divergence between distributions P and Q
$\bar{\mathbb{R}}$	The extended reals $\mathbb{R} \cup \{-\infty, \infty\}$
$\mathbb{R}_{\geq 0}$	The non-negative reals
$\mathbb{R}_{>0}$	The positive reals

CHAPTER 1

INTRODUCTION

Imaging plays a central role in modern medical practice and research [36]. Commonly used diagnostic modalities like planar x-ray imaging are most well known and wide spread, but there is also a vide variety of modalities targeted at preclinical research [35]. The imaging modality this thesis focusses on belongs to the latter category.

Preclinical research is highly diverse, it includes the early stages of drug development and studying how diseases progress. A key component of preclinical research is that of the animal study, which provides a stepping stone between cell research and application to humans. [35]

Medical images, including preclinical imaging, can provide insights on several scales. These are often grouped into two main categories: structural/anatomical and functional. The former provide primarily high quality information about the structure of the body itself, while the latter provides information about the location where a certain process happens or some molecule is present. Though most modalities are capable of providing at least some information from both categories, they are generally primarily aimed at one or the other. [37]

Some of the most common functional preclinical imaging modalities are positron emission tomography (PET), single photon emission tomography (SPECT), optical bioluminescence and optical fluorescence imaging. PET and SPECT provide quantitative information, but are limited to time scales of minutes or hours after injection. The optical methods on the other hand are often considered to be non-quantitative, but have no inherent limits on the time scale of visualized processes. [38]

X-ray fluorescence imaging (XFI) is currently being developed as a new imaging modality that closes this capability gap, offering time independent, quantitative images for small animals or in vitro samples. [12] [15] [16]

The reconstruction of a quantitative image can be interpreted as an inverse problem, where the input to a known system has to be approximated by its output [39]. This thesis improves upon the data processing methods currently used in XFI by applying standard as well as adapted statistical methods.

Bayesian methods are well suited to this task as they allow for prior information to be incorporated, potentially turning underdetermined problems into overdetermined ones [40]. The Bayesian approach also includes uncertainty estimation with is critical for fully quantitative work [41]. General Bayesian inverse problems are often numerically intractable, so approximate methods are needed. Generative models offer one possible solution that has been proven effective in many contexts [41].

Chapter 2 first discusses how XFI fits into the current medical imaging environment. Then the physics of XFI is described. Finally, the current state of the art in XFI techniques is described.

This work specifically focusses on applying different Bayesian techniques to XFI. Chapter 3 first introduces the Bayesian perspective, then describes these techniques along with the necessary theory behind them.

In chapter 4 the methods for processing XFI spectra are studied. First an abstracted spectral model is presented which lays the foundation for further analysis. Next this model is used to predict future measurements from simulation data and to augment datasets with synthetic data. The same simple model is then used to describe and compare various methods for isolating the signal from the background in XFI spectra. The aforementioned synthetic data augmentation is used to quantitatively compare the performance of the signal isolation techniques. Additionally, two options for incorporating data from multiple detectors at once are presented. Finally the current method for performing a statistical test for signal detection is improved upon.

Chapter 5 discusses three techniques for reconstruction of 3D images. First, the standard technique of tomography is described and the impact of XFI-specific corrections evaluated. Next, it is shown that the problem of 3D reconstruction is underdetermined if only data from a single projection is available. Lastly, a generative model tailored to this task is constructed and its performance evaluated. The novelty of section 5.3 in this work is the application of hierarchical, interpretable generative models to the XFI context, as well as the models themselves which incorporate domain knowledge and the implementation of a framework for this task optimized for gradient descent.

CHAPTER 2

X-RAY FLUORESCENCE IMAGING (XFI)

2.1 Imaging and Measurement

The goal of XFI, as with any imaging or measurement technique, is to obtain information about some object. This is accomplished by measuring a signal that relates to properties of the object in a known way. Real measurement processes are always subject to noise. One thus distinguishes between three different levels, the underlying properties, the observables which are determined by these properties through the relevant physics and finally the data which takes the form of a realization of a random variable parameterized by the observables. [75]

2.1.1. Imaging Functions

The physics of a given imaging process describe how the object X of interest relates to the observable Y. This process is deterministic and can thus be fully described by a function $f : X \mapsto Y$. The stochastic parts are treated separately, relating the actually observed data D to the observables Y. Imaging functions are generally not injective, so multiple objects may map onto the same observable. The inverse problem of finding the object given an observable is thus generally underdetermined. [39]

An illustrative example is a projection like $Y = f(X(x, y, z)) \propto \int X(x, y, z)dz$ which is part of the imaging functions of both x-ray imaging (see subsection 2.2.3) and XFI (see subsection 2.4.2). One can see that the projection introduces an unconstrained degree of freedom.

There are several ways of approaching this issue. One is to alter the observation technique such that the imaging function becomes invertible, this is the approach taken in idealized CT imaging (subsection 2.2.3). However, in reality this will only ever be approximately possible as the data space will always be finite dimensional while the space of general functions over space is infinite dimensional [39]. An alternative is to discretize the object space, to reduce the problem to finite dimensions. This is done implicitly in most modern image reconstruction techniques [26].

2.1.2. Noise Models

The stochastic components of a measurement process relate the observables Y to the finally observed data D. This can be expressed mathematically as D being a realization of a random variable parameterized by Y written as $P(D|Y)$. [39]

In many medical imaging applications a key stochastic component is that of photon counting. In SPECT, PET, CT ([38]) and XFI ([12]) the data consists of photon counts. Photon counts are discrete events happening over continuous time and are thus Poisson distributed, meaning the observable must take the form of an expectation value for these photon counts [39].

2.2 Common Medical Imaging Modalities

2.2.1. Anatomical Imaging

The first broad class of medical imaging is ***anatomical imaging***. This class focusses on resolving the anatomical structures in the body, typically bones and organs. It can be used to find pathological abnormalities like bone fractures or solid tumors, but the anatomical information can also be useful to guide surgeries or provide a reference for other types of imaging. This is the most commonly used type of medical imaging. [18]. For this work the most relevant anatomical imaging techniques are planar x-ray imaging and CT-imaging.

2.2.2. Planar X-Ray Imaging

Planar X-Ray Imaging is a technique in which a part of the body is illuminated by an x-ray source and the transmitted radiation is projected onto a sensitive screen. Today, this is often done with semiconductor detectors, before that, fluorescent screens and photographic film were commonly used. This method is easy to implement, requiring minimal materials and software; hence it is widespread and can, for instance, be found in many dental practices. [18]

The physics which impart the transmitted x-rays with anatomical information is simple absorption in the subject. The Lambert-Beer law (compare subsection 2.3.3) can thus be used to model this process [42]. For each length element along the beam direction dz, the beam intensity S is proportionally attenuated by a factor $S(z+dz) = S(z)\exp(-\mu dz)$ where μ is called the linear attenuation coefficient. This makes it connivent to express the imaging function in logarithmic terms. After integrating out the z component one finds:

$$\log S(x,y) = \log S_0 \int -\mu(x,y,z)dz \tag{2.1}$$

The linear attenuation coefficient μ depends on the elements in a material and the overall density, so it carries information about tissue types and densities. Planar x-ray can be understood as a projection of those attenuation coefficients.

Because it requires significant differences in μ, planar x-ray imaging is primarily used to resolve bone and teeth, but can also resolve some organs and tumors. We can understand planar x-ray as an image of the integrated attenuation coefficients. These attenuation coefficients encode information about the subject and usually are a good proxy for density, especially in objects of homogeneous elemental makeup.

2.2.3. Computed Tomography (CT)

Computed Tomography (CT) builds on planar x-ray by acquiring many such projections with the subject (or equivalently source and detector) rotated to various angles. From these projections a three dimensional image can be computed (hence the name). CT is generally more sensitive to differences in μ, because fully reconstructed CT images are not integrated over depth. This higher fidelity also translates to being more sensitive to different types of tissues. [42]

CT imaging works by taking x-ray transmission images from many different directions. In the idealized case the signal is characterized as a function $S(\theta, s)$ where θ is the angle the object or measurement setup is rotated by and s is an offset along the image plane. If we look at the Lambert-Beer law for a beam path going through the object, we see a structure like

$$S = S_0 \exp\left(-\int \mu(x)dx\right)$$

where μ is a map of the linear attenuation coefficients, I_0 is the beam intensity and I is the intensity that leaves the object. This structure is fairly complex, so a transformation is applied to make the signal function S and the object O follow a simpler relation. Specifically,

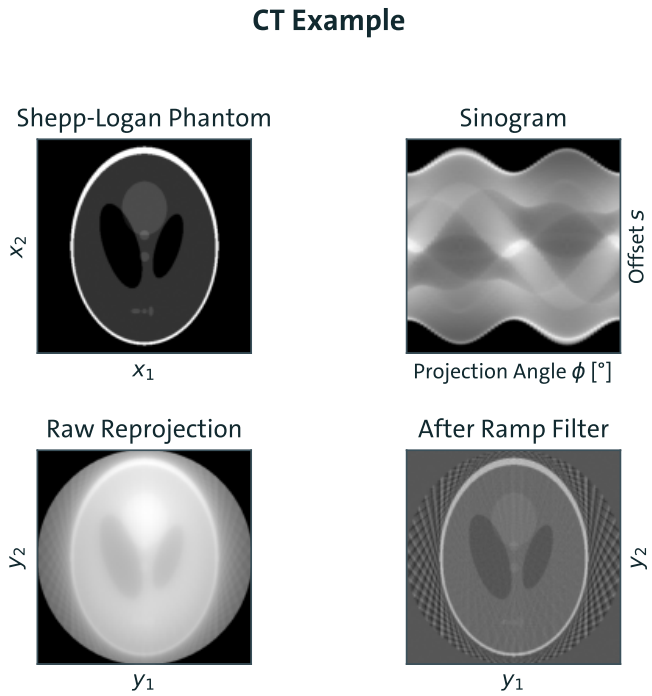

Figure 2.1. Example of a CT Image: Given an object in x-space (top left) one can find the corresponding sinogram (top right) in ϕ, s coordinates via the signal equation. Backprojection (bottom left) yields a blurry image in y-space. Applying a ramp filter (bottom right) can greatly reduce that blur.

$S(\theta, s) = -\ln(I(\theta, s)/I_0)$ and $O(\vec{x}) = \mu(\vec{x})$. For CT imaging the recorded signal S is called a sinogram, because a point-like object produces a sinusoidal signal. An example of a sinogram can be seen in subsection 2.2.3 top right. So the idealized CT data is a continuous ensemble of planar x-ray images. [42]

The idealized signal equation then becomes

$$S(\theta, s) = \int O(\vec{x}(t))dt\,,\ \vec{x}(t) = s\begin{pmatrix}\cos(\theta)\\ \sin(\theta)\end{pmatrix} + t\begin{pmatrix}-\sin(\theta)\\ \cos(\theta)\end{pmatrix} =: s\vec{\theta} + t\vec{\theta}^{\perp} \tag{2.2}$$

This transform is known as the Radon transform [24].

While this relation is an integral transform, it does preserve the dimensionality of the object, so one might expect that it could be invertible and indeed it is. The naive attempt at inverting this would be to simply project the signal back along the line it came from with equal intensity. Using $B(\vec{y})$ as the result of this process gives us

$$\begin{aligned} B(\vec{y}) &= \int_0^{2\pi} S(\phi, \vec{\phi}\cdot\vec{y})d\phi = \int_0^{2\pi}\int O\left((\vec{\phi}\cdot\vec{y})\vec{\phi} + t\vec{\phi}^{\perp}\right)dtd\phi \\ &= \int_0^{2\pi}\iint O(\vec{x})\delta(\vec{\phi}\cdot\vec{x} - \vec{\phi}\cdot\vec{y})\,d^2x\,d\phi \\ &= \iint O(\vec{x})\int_0^{2\pi}\delta(\vec{\phi}\cdot(\vec{x}-\vec{y}))\,d\phi\,d^2x =: \iint O(\vec{x})K(\vec{x}-\vec{y})\,d^2x \end{aligned}$$

Hence, the naive back-projection actually produces an image showing the object O convolved with some kernel K. Finding K is a straightforward integral over a delta distribution in polar

coordinates

$$K(\vec{u}) = \int_0^{2\pi} \delta(\vec{\phi}\cdot\vec{u})d\phi = \int_0^{2\pi} \delta\left(\vec{\phi}\cdot|u|\vec{\theta}\right)d\phi = \frac{1}{|u|}\int_0^{2\pi} \delta\left(\vec{\phi}\cdot\vec{\theta}\right)d\phi = \frac{2}{|u|}$$

where θ was used for the argument of $\vec{u}$.

To get an exact inverse Radon transform one thus would have to perform a deconvolution with K. The standard way to do deconvolutions of translation invariant kernels is by using the Fourier convolution theorem ([59]):

$$\begin{aligned}\mathcal{F}[O * K] &= \mathcal{F}[O]\cdot\mathcal{F}[K]\\ \Rightarrow \mathcal{F}[B]/\mathcal{F}[K] &= \mathcal{F}[O]\end{aligned}$$

The Fourier transform of $2/|u|$ is a bit tricky as it requires regularization (e.g. with $e^{-|\epsilon|r}$ followed by $\lim_{\epsilon\to 0}$), but computing it yields:

$$\mathcal{F}[K](\vec{k}) = \frac{4\pi}{|\vec{k}|} \Rightarrow 1/\mathcal{F}[K](\vec{k}) = \frac{|\vec{k}|}{4\pi}$$

So the deconvolution step of the inverse Radon corresponds to a ramp filter in Fourier space. [24]

In practice this is not done as two separate steps. Instead one does the back-projection in Fourier space as well and filters before transforming to position space again. This is possible because both the Fourier and Radon integrals are well behaved, allowing for Fubini's theorem to be applied, enabling a swap of the order of operations [24]. This process is called filtered back-projection and is the standard method for reconstructing CT images [18]. The impact of deconvolution of the ramp filter can be seen by comparing the lower left and lower right of subsection 2.2.3. One issue with this method is that the ramp filter may make the Fourier transform singular and if the signal is noisy will introduce over sharpening. So in real world use the ramp filter is typically cut off or regularized otherwise. [25]

2.2.3.1. Tomosynthesis

Above we saw that the integral over the projection angles in the reconstruction ends up defining the convolution kernel K which relates the object and the image after back-projection. The first interesting observation is that if instead of integrating over the full circle we only integrated over one half of it, the result would only differ by a constant factor of 2. So it is not actually necessary to record a full 2π sinogram, half of it suffices. If even less of the sinogram is recorded, a term depending on the direction of $\vec{u}$ would be introduced which can lead to blur or artifacts, but the image might still be usable for specific applications. This is especially true if care is taken in selecting a kernel for the deconvolution. This imaging technique is called tomosynthesis and has a variety of applications in medical practice, like low dose mammography. [7]

2.2.3.2. Issues

The first problem which arises when trying to use these concepts in reality is that of discretization of the sinogram; one does not record a continuous signal or processes the signals as functions, instead a set of discrete images is recorded which are processed as finite dimensional vectors [74]. As was discussed with tomosynthesis this will change the character of K in complex ways, if the number of angles is low for example (even if they cover 2π), one tends to see radially decaying star-shaped artifacts around high density objects. This problem can be tackled by taking a sufficient number of images of sufficient resolution and, again, by carefully choosing the deconvolution kernel.

Another issue is that for very strong attenuation I might be zero, which breaks the Radon relation between the signal and the object because the signal S is not defined. Such strong attenuation might for example occur when there are pieces of metal in the imaged area. Even

when I is only very close to 0, so-called photon-starvation might lead to starburst artifacts that decay slowly with distance from the intrusion, degrading large portions of the slice.

Another assumption that can be broken in reality is that μ only depends on the position in the object. For monochromatic x-rays or energy-resolved detectors this is true, but most CT imaging uses a polychromatic source and energy integrated detectors. The issue arises because the linear attenuation coefficient depends on photon energy, so as the beam passes through the material its spectrum changes and thus the effective linear attenuation coefficient. This phenomenon is called beam hardening and introduces a deviation from the Radon transform similar to an attenuation term, manifesting as an artificial suppression of the attenuation coefficients towards the center of the object and around high density intrusions [60].

2.2.3.3. Alternate Reconstruction Techniques

There are a variety of other techniques available to reconstruct 3D images from sinograms, which can be divided into three categories: iterative matching, learned back transform and Bayesian methods. [26]

The iterative matching methods try to find an object $\hat{O}$ such that the image it produces has a signal $\hat{S}$ which is as close as possible to the recorded signal S. More generally one can say that $\hat{O}$ is minimizing a loss function which includes the difference between $\hat{S}$ and S, but may also include other terms which can serve to obey physical constraints or regularize $\hat{O}$. These methods can avoid many of the artifacts discussed above, but the ill-posedness of the inverse still imposes uncertainty which is not reflected in the output. Additionally these methods can be computationally expensive in comparison with filtered back-projection. [26]

Learned inverses are machine learning algorithms that are trained to produce reconstructed images from the measured data. These approaches often include hybrid functionality where parts of the process like Fourier transforms are performed classically and as part of the architecture. An alternate framing would be that machine learning algorithms are used to replace or augment filtered back-projection at specific points in the process. While training is computationally expensive, the inference is fairly cheap and this approach can still avoid many artifacts. Hence making variations of this approach very attractive for clinical use and indeed there already are several such systems used in clinical practice. [8]

Bayesian methods are diverse in and of themselves. The simplest are maximum posterior methods which produce a point prediction that maximizes the posterior, composed of the likelihood to produce the observed signal and prior assumptions. This method can be reframed as a special type of loss function for the iterative approaches with the likelihood serving as the measure of difference between $\hat{S}$ and S. These generally belong to the class of model-based iterative reconstruction because they require a projection from image space to projection space. Simpler methods that only model photon statistics and do not contain priors in image space also exist, if they do not require a forward projection they are classed as (hybrid) statistical iterative reconstruction methods. [26]

2.2.3.4. Spectral CT

One can extend normal CT imaging by using an energy-resolved detector. In that case, the observables, signals etc. all gain a dependence on the photon energy E. In particular, rather than just gaining access to $\mu(\vec{r})$ this gives access to $\mu(\vec{r}, E)$ which helps to identify the difference between materials. Using prior knowledge on what materials/tissue types are expected to be present in the subject, aids in the identification of those. This holds true, even if energy differentiation is limited; this method can be used with as little as two or four different energy bins. Using the materials, a full map of $\mu(\vec{r}, E)$ can then be constructed for arbitrary energies. Spectral CT images are thus especially interesting for situations where an image is used to calculate photon interactions in the subject, like in treatment planning. [43]

2.2.4. Functional Imaging

In contrast to anatomical imaging, functional imaging attempts to highlight features based on their function. In most cases this is done by using a tracer that is visible in the image and accumulates in regions where a given process occurs. Most functional imaging is done via emission imaging ([18]), which works by using tracers that emit radiation which is detected outside the subject. These tracers are typically embedded into a molecule which interacts with the biological processes inside the subject such that their distribution carries information about where those processes occur. [18]

2.2.5. Single Photon Emission Tomography (SPECT)

In SPECT, the tracers are radioisotopes that emit high energy photons as part of their decay. An example would be ^{99m}Tc which is a metastable isomer of technetium 99 with a half-life of 6 h and which decays to the stable isomer by emission of an ≈ 140 keV photon. The half-life plays a very important role in tracer selection. Any decay contributes to the total radiation dose the subject is exposed to. And these occur from the moment of injection and fall off in intensity over the next few half-lives until after several half-lives essentially all tracer decayed. This means that, in order to be dose-efficient one needs to observe the subject for several half-lives and quickly start imaging after injection. SPECT is typically dose limited, so dose-efficiency is a primary concern. This means one needs to use reasonably short half-life tracers, as long measurement times pose many issues. These range from the simple practical or logistical issues to the very fundamental ones, like metabolic processes eventually leading to tracers being excreted. On the other hand short decay times make the tracer production logistically challenging as the time between production of the radioisotopes and injection must be minimized. Additionally it is impossible to observe processes which occur over several half-lives as essentially no signal will be emitted after that time. This makes SPECT challenging to use for long-term accumulation or kinetics studies. The image is typically acquired by using gamma cameras with collimators which are rotated around the subject. The reconstruction can be done through back-projection or various iterative methods. [18]

2.2.6. Positron Emission Tomography (PET)

PET is very similar to SPECT in many regards, including the limitations induced by tracer decay. However, instead of a single photon being emitted from a decay, PET uses a β^+ decay which creates a positron which annihilates, creating two photons at 511 keV that propagate in opposing directions. The subject is typically surrounded by a ring of scintillator-based detectors, giving great angular coverage. Collimators are not needed because the back-projection can use the line connecting two simultaneous detections. PET can also use light atoms like ^{18}F which can be directly incorporated into relevant molecules. For example an OH group in glucose can be replaced by ^{18}F, this ^{18}Fluoro-Desoxy-Glucose (FDG) behaves similar to glucose in the body and can thus be used to study metabolic activity and glucose uptake. [18]

2.2.7. Bioluminescence

Bioluminescence is also a type of emission imaging, but instead of using radioisotopes the emission is generated by biological processes. These processes are typically enzyme mediated oxidations, one example of such an enzyme is luciferase. Thus, the technique typically relies on labeled, luminescent, cells. These cells can for instance be modified pathogens or cancer cell-lines which enables direct measurements of the disease burden. (Near-)visible light has a high scattering and absorption cross-sections in tissue. Therefore, bioluminescence imaging is limited to the scale of centimeters in tissue-depth, this restricts the in vivo applicability

to small mammals or tissues close to the skin. Additionally, coherent scattering creates an upper bound for the maximum possible resolution depending on the depth in the tissue. [19]

2.2.8. Optical Fluorescence

Fluorescence imaging is in principle similar to bioluminescence, however, instead of luminescent proteins, fluorescent ones are imaged. The fluorescence is typically excited with violet to ultraviolet light, meaning that tissue penetration is minimal. Many proteins can fluoresce under some conditions, often in different colors. The dependence on external conditions can be used to develop highly specific markers. [20]

2.2.9. The Role of X-Ray Fluorescence Imaging

X-ray fluorescence imaging (XFI) is the focus of this work. The details of this method will be described in section 2.4. XFI is a functional imaging modality that can use many elements as tracers. It will be shown that, unlike in SPECT and PET, these tracers are stable, meaning there are no restrictions on imaging timelines induced by tracer choice. And unlike optical fluorescence, at appropriate tracer choice, the attenuation and coherent scattering in the tissue is minimal, meaning it can be used in depth. With these advantages XFI enables a new domain of possible measurements in preclinical use.

2.3 Light-Matter Interactions

To understand XFI, one needs to understand how light interacts with matter. In particular one needs to understand both how fluorescence is generated (interaction with atoms), the processes dominating attenuation of fluorescence (photoelectric effect) and the spectral background (Compton scattering) [21]. These processes take place on scales ranging from the subatomic (Compton scattering with free electrons) to the mesoscopic (photoelectric effect being impacted by the surroundings of an atomic). No single model of light is capable of describing all of these interactions in a cohesive language.

As an electromagnetic object, light can couple to things that are electrically charged, have an electric or magnetic moment or are polarizable [78]. This means one can go through the hierarchy of models from the subatomic to a description of the macroscopic and identify all levels at which one of those is present. Not all of the interactions are relevant for this work, but interactions from all levels are. The list which is presented here should thus not be taken as an exhaustive description of light-matter interactions, but as a summary of the most relevant ones for x-ray fluorescence measurements.

2.3.1. Compton Scattering

The model of choice to study Compton scattering is quantum electrodynamics, or QED for short. In QED light mediates the interactions between charged particles [28]. The simplest interaction light can take part in is elastic scattering between a charged electron and light, i.e. the photon and electron exchanging some energy and momentum. This process is called Compton scattering and for a free electron it is simple enough that the differential scattering cross-section $\frac{d\sigma}{d\Omega}$ can be calculated from first principles using QED [28]. The result is the famous Klein-Nishina formula

$$\frac{d\sigma}{d\Omega} = \frac{r_e^2}{2} l^2 \left[l + \frac{1}{l} - \sin^2(\theta) \right] \tag{2.3}$$

$$l = \frac{\lambda}{\lambda'} = \frac{E_{\gamma'}}{E_{\gamma'}} = \frac{1}{1 + \frac{E_\gamma}{m_e c^2}(1 - \cos(\theta))} \tag{2.4}$$

where θ is the angle between the incident and exiting photon and $r_e \approx 3\,\text{fm}$ is the classical electron radius, λ is the wavelength of the light, E_γ the corresponding energy, c the speed of light in a vacuum and $'$ indicates the state after the scattering event [28].

For high photon energies and large scatter angles this differential cross-section can be approximated as

$$\frac{d\sigma}{d\Omega} \approx \frac{r_e^2}{2} \frac{1}{1 + \frac{E_\gamma}{m_e c^2}(1 - \cos(\theta))}$$

which suggests that the scattering probability for meaningful scattering angles decreases like $\frac{m_e c^2}{E_\gamma}$. So for very high energies Compton scattering is disfavored. In the low energy limit the Klein-Nishina formula reduces to the Thompson scattering cross-section

$$\frac{d\sigma}{d\Omega} = \frac{r_e^2}{2}(1 + \cos^2(\theta)) \tag{2.5}$$

which shows that the likelihood converges to a finite value in the low energy limit. Interestingly the angular dependence in equation (2.3) shifts as relativistic effects become relevant, where the Thompson cross-section (equation (2.5)) is symmetric between forward and backward scattering (as only $\cos^2$ appears), the high energy limit strongly favors the forward direction. While this model studies free electrons, it also provides a good approximation for bound electrons, like those in an atom, provided that the scattering energy $E_\gamma - E_{\gamma'} \gg V$ where $V$is the binding energy, i.e., the initial electron is approximately at rest and the resulting electron is free. The corrections induced this way are usually treated separately, like a Doppler broadening from the electron's initial momentum distribution. Compton scattering is one of the most important background components in XFI ([15]), especially for lines that are at or above half the incident energy. Because Compton scattering leads to ionization it can also lead to signal production (as will be discussed in the following sections), though ionization of the inner shells is relatively rare compared to competing processes. [23]

2.3.2. Interactions with Atoms

While Compton scattering can be studied using free electrons, other processes can only be described at the atomic scale.

2.3.2.1. The Hydrogen Atom

The simplest possible atom is the idealized hydrogen atom with a single electron and a nucleus modeled as an immobile point charge. An accurate description of the electron in the hydrogen atom was a major achievement of early quantum mechanics, but interestingly also presents the most complex atom that can be studied exactly. Larger atoms are thus usually treated as hydrogen atoms with small corrections.

The full model takes the form of a Schrödinger equation in spherical coordinates with the reduced mass μ and electron charge e [51]

$$E\psi = \left[-\frac{\hbar^2}{2\mu}\Delta - \frac{e^2}{4\pi\epsilon_0 r}\right]\psi \tag{2.6}$$

Where $\hbar$ is the reduce Planck constant, ψ is the wave function, E is the energy of the state, Δ is the Laplace operator, ϵ_0 is the electrostatic constant and r is the radial coordinate. The solution of this equation are (up to normalization)

$$\psi = f_{n,l}(\rho) Y_l^m(\theta, \phi) \tag{2.7}$$

$$f_{n,l}(\rho) = e^{-\rho/2} \rho^l L_{n-l-1}^{2l+1}(\rho)$$

where Y_l^m are the spherical harmonics, L_{n-l-1}^{2l+1} the generalized Laguerre polynomials, $\rho = \frac{2r}{n a_0^*}$ and a_0^* the Bohr radius. From this solution we can see that the states the electron can take in a hydrogen atom are indexed with three quantum numbers n, l, m. n describes the energy of the electron, l indexes the total angular momentum and m the z component of the angular momentum. They follow the relation $1 \leq n > l \geq |m|$, so with $n = 1$ only $m = 0 = l$ is admissible while $n = 1$ also allows for $l = 1, m = 0$ and $l = 1, m = \pm 1$ for four states in total. The individual states are called atomic orbitals.

Orbitals with $l > 0$ have an intrinsic magnetic moment which provides an additional way for light to interact with. Additionally the orbitals are slightly altered in the presence of an external electromagnetic field (corresponding to a correction to the radial potential), leading to polarizability. [51]

2.3.2.2. Wigner-Weisskopf and Atomic Lines

The most important way the coupling between orbitals and the light field manifests is in atomic lines. Wigner and Weisskopf ([22]) formulated a model to describe how excited atomic states decay under emission of light using Dirac's then recently published theory of light. The model studies an atom in a hohlraum (which can later be taken to the continuum limit), thus the light is quantized to countable modes indexed by i with energies E_i. The full system is then described by studying the joint state $|Q, (N_i)\rangle$ where Q describes the state of the atom and (N_i) is a vector which indicates how many light quanta are in mode i for all modes. The corresponding Hamiltonian is given by:

$$\begin{aligned} H_{WW} |Q, (N_i)\rangle &= \sum_i N_i E_i + E_Q + \sum_{i,Q'} (A_{i,Q'} + A_{i,Q'}^\dagger) \\ A_{i,Q'} &= g^{(i)}_{Q,Q'} |Q'\rangle \langle Q| \hat{a}_i \\ g^{(i)}_{Q,Q'} &\propto \langle Q| \hat{d} \cdot \hat{e}_i |Q'\rangle \end{aligned}$$

So the total Hamiltonian consists of the energy of all the light modes $\sum_i N_i E_i$, the energy of the state the atom is in E_Q and an interacting term with $A_{i,Q'}$ built from the ladder operators. This interaction describes emitting or absorbing a light quantum to change the state of the atom, $A_{i,Q'}$ represents the atom emitting a photon with index i while transitioning from Q to Q'. The amplitude of this interaction g depends on the dipole of the atom and the polarization of the light. [22]

This simple model is sufficient to indicate which transitions can and which cannot emit light and which transitions are dominant. We find that out of the transitions starting with the atom in the ground state, except for a vacancy in the $n = 1$ orbital two options dominate, those where a $n = 2, l = 1$ electron fills the vacancy and those where an $n = 3, l = 1$ electron does so. These are called the K_α and K_β line respectively and are generally the brightest lines in sufficiently excited atoms [76]. For elements from approximately $Z = 30$ (zinc) to $Z = 80$ (mercury) these lines are in the range of $10 - 70\,\text{keV}$ [23]. Decays like $n = 3, l = 2 \to n = 1$ have very low probability of occurring due to the $\langle Q| \hat{d} \cdot \hat{e}_i |Q'\rangle$ term [23]. The next brightest lines are called the L lines and fill vacancies with $n = 2$. Again the largest contribution comes from one n up, i.e. electrons from $n = 3$ followed by $n = 4$ and the dipole transitions, namely those where l changes by 1 [23]. This leaves five particularly relevant L lines which usually merge into two L_α and L_β in energy spectra. The L lines cross the 10 keV threshold around $Z = 70$ (ytterbium) and never reach higher than around 20 keV [23]. Lines starting with a vacancy in $n = 3$ never go beyond 3 keV [23] and are thus not relevant for this work. These lines all have very short life-times and thus can be described as a fluorescence phenomenon. These photons are what the signal in XRF is composed of.

2.3.2.3. Rayleigh Scattering

Where spontaneous emission came from the coupling of the light field to the inherent dipole moment, the light field can also couple to the induced dipole, or the polarizability of the atom. This obviously requires an external field to begin with and thus describes a scattering process rather than emission. The elastic scattering off of a polarizable object is called Rayleigh scattering. For soft x-rays and UV the whole atom can manifest a real dipole, for higher energy x-rays the corresponding deformations are more complex and instantaneous, which also leads to lower scattering amplitudes at higher energies. In XRF applications the signal spectrum often features a distinct Rayleigh peak at the incident energy resulting from forward scattering in air. [29]

2.3.2.4. The Atomic Shell Model

For some processes the angular momenta do not play a role, thus n becomes the only relevant quantum number. As the radial component f in equation (2.7) is indexed by n, but the angular components L are not, this creates a picture of "shells", each corresponding to a different energy and allowing for up to $2n^2$ electrons in that shell. The lowest energy state with N electrons thus is one where all shells are filled one by one, starting with $n = 1$ up to the outermost occupied shell $n = k$ where $(k-1)k(2k+1) < 3N \leq k(k+1)(2k+1)$.

This approximate model of the atom greatly simplifies the atomic Hamiltonian, diagonalizing it. The cost of this description is the loss of the spatial distribution and associated values like dipole-moment and polarizability. But to describe the energies of atomic lines, this model is sufficient. For historical reasons these shells are indexed $K, L, M, N, ...$ and correspond to the rows of the periodic table. These shell indices are also where the line names in so-called x-ray notation like K_α come from. The Roman letter indexes the shell of the vacancy, the Greek letter counts how many shells are transited through. So K_β corresponds to a vacancy in the K shell being filled with an electron from the M shell. [30]

2.3.2.5. Photoelectric Ionisation

In the free electron interaction (Compton scattering), it is impossible to fully absorb the photon due to energy and momentum conservation. When one tries to conserve energy, the amount of momentum after the collision is always too large.

However, when an electron from within an atom is excited to the continuum, i.e. the atom is ionized, an additional degree of freedom is introduced. If the ionized atom and electron move in opposing directions both quantities can be conserved at the same time. In crystals this would create a phonon, in solutions and gases atomic movement, both of which eventually thermalize, becoming heat. This process is called the photoelectric effect. Due to the complex nature of the momentum absorption one typically needs a model of not just the atom itself, but also the larger scale structure it is embedded in to estimate cross-sections. For typical XRF applications photoelectric ionization dominates over ionization by Compton scattering in the inner shells. [30]

2.3.2.6. X-Ray Fluorescence

X-ray fluorescence is a compound process, where first an inner electron inside an atom is ejected via Compton scattering or the photoelectric effect. The resulting vacancy is then filled by an electron from a higher shell leading to emission. The previous sections discussed the individual processes involved. The number of ionizations of a given shell is just the sum of the ionizations through the photoelectric effect and Compton scattering, as such it is linear in both the number of atoms involved and the number of incident photons. [30]

The number of emissions of a particular line given a specific shell ionization is also linear, and if a line can be produced through different ionizations, the total number is again just their sum. Thus one can group the entire process into one fluorescence cross-section per atomic emission line. This cross-section generally depends not just on the incident energy, but also the type of material the atom is embedded in, however, they are available in reference tables like x-ray lib [23].

2.3.3. Lambert-Beer Law

The phenomena discussed above are all almost perfectly linear in the domain relevant to XFI. This means that one can speak of interaction cross-sections or attenuation coefficients in the traditional sense; when a photon travels a short distance dx through a material, there is a probability of $\mu\, dx$ that an interaction occurs where μ is called the linear attenuation coefficient. This is more rigorously expressed in the differential form

$$dI = -\mu dx \Rightarrow I(x) = I_0 e^{-\mu x}$$

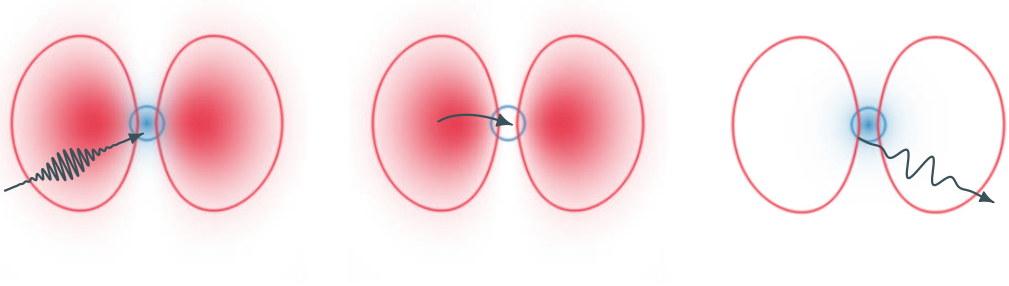

Figure 2.2. Fluorescence Process: an incident x-ray photon ionizes an electron from an inner shell (left). The atom is now in an unstable intermediate state (center). An outer electron fills the inner vacancy, emitting a photon corresponding to the difference in potential energy of those shells (right).

where I is the number of photons that have not interacted yet and $I_0 = I(0)$ is the initial photon number. This is called the Lambert-Beer law. [31]

With processes like photoelectric absorption the photon fully disappears after the interaction, so one can apply the Lambert-Beer law directly to the flux of the incident beam and get a reasonable picture of the impact of photoelectric interaction with it.

2.4 X-Ray Fluorescence Measurements and Imaging

This section and the section on x-ray fluorescence imaging lean on [11] where not stated otherwise.

X-ray fluorescence measurements work by illuminating a subject with an external beam. This creates excited states through ionization which relax to produce secondary characteristic radiation at x-ray energies. That characteristic radiation is observed and can be used to identify the elements in the illuminated subject area.

X-ray fluorescence imaging generally describes spatially resolved XRF measurements. For example when a subject is scanned by a small beam, or when an image is projected through a pinhole (see [13]). So it should not be thought of as an entirely different category, but a subset of XRF measurements. This work will occasionally use terms like "XRF tomography" which is an XFI technique.

2.4.1. XRF Components

In order to conduct an XRF measurement one needs three primary components: incident x-rays that induce the fluorescence, a tracer substance that fluoresces and a detector to record the fluorescence.

The ***incident beam*** is used to excite the tracer atoms in the measured subject. It needs to have an energy sufficient to ionize the shell into which the desired relaxation occurs. Photons from the primary beam can find their way to the detector through scattering. The spectrum of the beam should therefore ideally be chosen such that the scattered spectrum at the detector has minimal overlap with the energy region where the fluorescence line of interest is located. Furthermore, if time is limited, the total flux needs to be high enough to accumulate a significant signal in the allotted time. Lastly, the incident beam is often used to gain spatial information, as fluorescence photons will be primarily produced in the illuminated area. In that case the beam size limits the spatial resolution of the acquired data. [12]

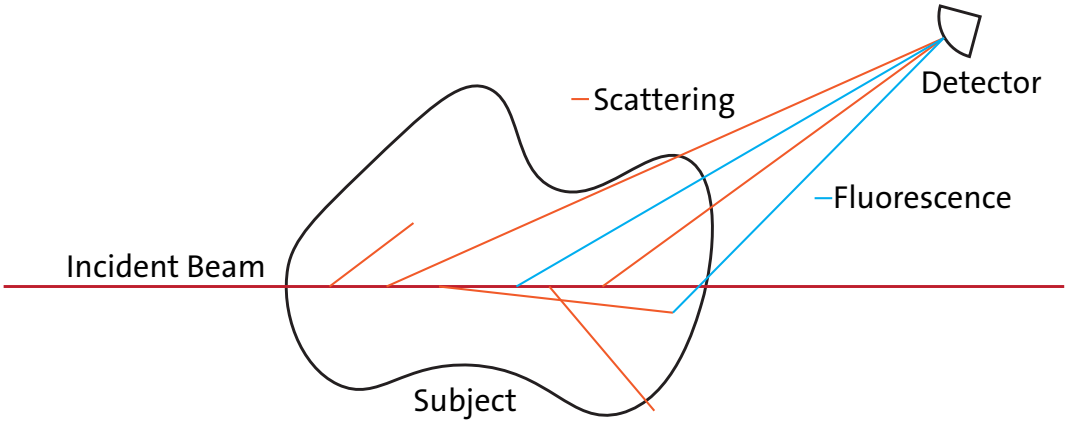

Figure 2.3. Schematic of a Typical XFI Setup: an incident x-ray beam (dark red) is used to illuminate a subject (center, black). Some incident photons scatter (orange) and some create fluorescence (blue). The detector (top right) records a combination of both components.

The ***tracer elements*** are the atoms that produce the signal fluorescence. They need to be chosen based on two key characteristics: compatibility with the subject (for example in biological measurements) and the energy of its lines. The line energy should ideally be located in an energy region with minimal background, but it needs to be large enough that a significant fraction of the characteristic radiation can escape the subject and low enough that the incident beam can ionize the lower shell. Additionally the line should have a good fluorescence cross-section at the given incident energy. [12]

The ***detectors*** record the fluorescence signal, but will usually also be hit by a lot of scattered photons from the incident beam. In order to gain useful information from the recorded data, two things become necessary: the detection needs to happen at reasonable efficiency and the number of background photons that cannot be distinguished from signal photons should be as small as possible. The former is primarily driven by the absorption cross-section for photons of the signal energy and the thickness of the detector. The latter is given by a combination of energy resolution (fluorescence lines are very sharp compared to the broad scattering background) and limiting the background induced by detector imperfections. To reduce dose or measurement time they should be as large as possible. Some detectors also record spatial information which can be useful for imaging measurements. [12]

This complex interdependence means that finding a combination that is well suited to a specific application is not trivial and different applications can lead to drastically different implementations. For example [13] describes primarily surface scanning applications, so the line energies can be as low as a few keV without problematic attenuation. In applications like [16], where lines like the 19 keV Ruthenium K_α were used, one can still shape incident beams using x-ray optics at good efficiency [52], as well as very high efficiency and resolution (near theoretical limit) silicon detectors [53]. In contrast [12] discusses larger subjects and where the tracer can be at depths beyond 10 cm, requiring much higher fluorescence energies (like Gold K_α at 69 keV) and thus beam energies where optics typically have very poor efficiency. [12]

2.4.2. The XFI Imaging Function

The imaging function for XFI can be found by analyzing the event chain leading to the detection of a fluorescence photon. This work focusses on scanning beam techniques, so the corresponding imaging function will be described following [21]. First, a number N_0 of primary photons is emitted. Next the photon needs to make it to the site of fluorescence without being absorbed in the material along the way which happens at a probability T_i. The probability to produce a fluorescence photon at a given site is simply given by the

linear attenuation coefficient and the line element $\mu_f dz = \left(\frac{\mu}{\rho}\right)_f \rho dz$ where ρ is the local density of tracer material and $\left(\frac{\mu}{\rho}\right)_f$ is the specific linear attenuation coefficient which can be found in reference tables [23]. The produced fluorescence photon then must leave the subject without being absorbed which has the probability T_o. It then must hit the detector, the probability of which depends on the solid angle coverage $\frac{A}{4\pi r^2}$ where r is the distance from the fluorescence site and A is the effective detector area. Finally the hit must be converted into a detection, the probability of which is just the detector efficiency ϵ. The total number of detected fluorescence photons N_f will be the superposition over all possible fluorescence sites, so integration over the coordinate along the beam z is needed.

This finally leads to the form found in [21]:

$$N_f = N_0 \epsilon \int T_i(z) \left(\frac{\mu}{\rho}\right)_f \rho(z) T_o(z) \frac{A}{4\pi r(z)^2} dz \tag{2.8}$$

The expected number of fluorescence photons N_f is an observable, while $\rho(z)$ is generally the object of interest. The data, meaning the registered fluorescence counts in a detector, is then Poisson distributed around that expectation value.

2.4.3. Typical Parameters or Preclinical XFI

This work primarily focusses on XFI aimed at preclinical measurements of mice. This means that line energies need to be at least around 20 keV for the signal to escape the subject [12]. This limits the usable elements to those with $Z \geq 40$ and their respective K-lines. Possible x-ray sources include filtered x-ray tubes and undulator sources at synchrotron facilities [32]. Using the tungsten K_β peak from an x-ray tube for example produces a 60 keV beam which can excite the K-shell of elements with $Z < 70$ [52]. However, with those sources the background due to scattered primary photons can become dominant over potential signals around 30−40 keV, so in practice, using tracers heavier than iodine ($Z = 53$) is not particularly beneficial[12].

2.4.4. Anatomy of XRF Spectra

Figure 2.5 shows a simulated spectrum with the parameters discussed in subsection 2.4.3 with palladium and tellurium as tracers, a glass cuvette as the subject and a beam approximating the characteristics of the P21 beamline at PETRA III [34], which uses undulators to generate near monochromatic radiation from high energy electrons tapped from the PETRA storage ring. The simulation allows for events to be labeled according to their origin which is helpful for visualizing at which energies various processes become relevant. One can see that everything above 35 keV is dominated by the Compton scattered primaries. The region below 10 keV is primarily filled with events of the same type, but where the detector did not capture the full photon energy, leading to an incorrect energy measurement (compare figure 2.5). This leaves $15-30$ keV as the ideal region to detect tracer signals. In this region we find a low background level. Additionally some fluorescence peaks that do not relate to the subject can be found there, for example silver and tin from solder.

2.4.5. State of the Art for XFI Methods

2.4.5.1. Processing Spectra

To obtain the signal strength from an XRF spectrum one needs to estimate how many background counts are in the signal region. One way of classifying these methods is by whether they can work on a single spectrum or need a reference measurement, and by whether they use local or non-local correlations in the spectrum.

The first method in use is the ***subtraction method***, as seen in [14], where it is used to image gold nanoparticles in mice. The same subject is measured with and without the tracer.

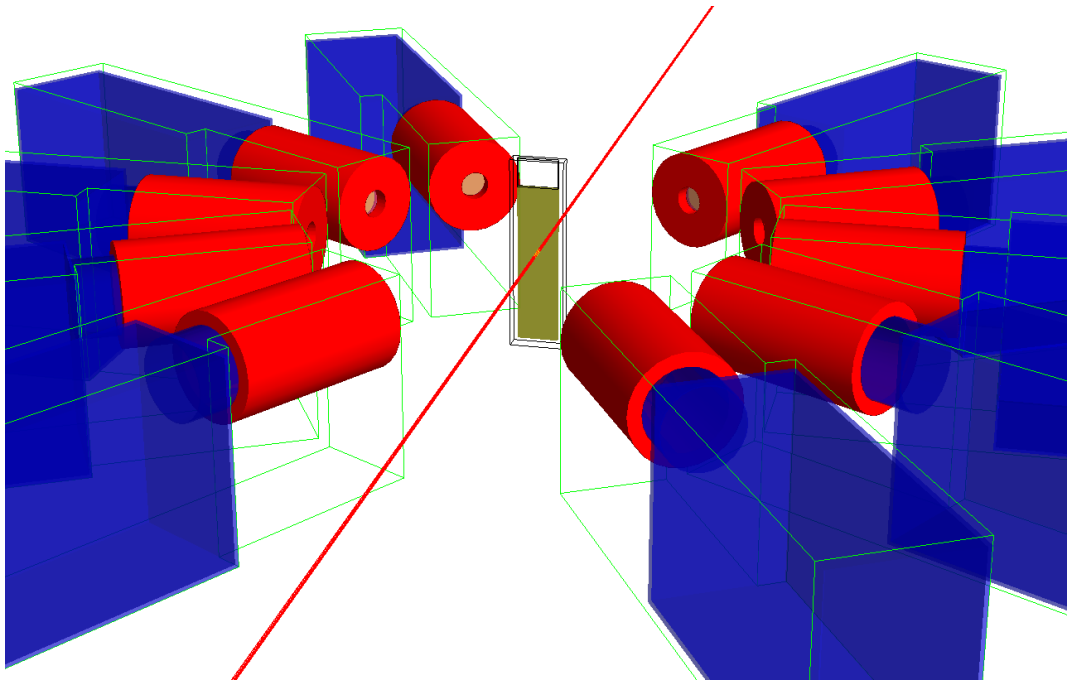

Figure 2.4. Measurement Apparatus in Simulation: A typical measurement setup was implemented in GEANT4, consisting of the incident beam (red line from bottom left to top right), ten detectors (blue and red boxes arranged in a circle) and the subject, here a glass cuvette containing an aqueous solution of tracer elements (transparent and yellow in the center). One of the spectra recorded by a detector is shown in figure 2.5

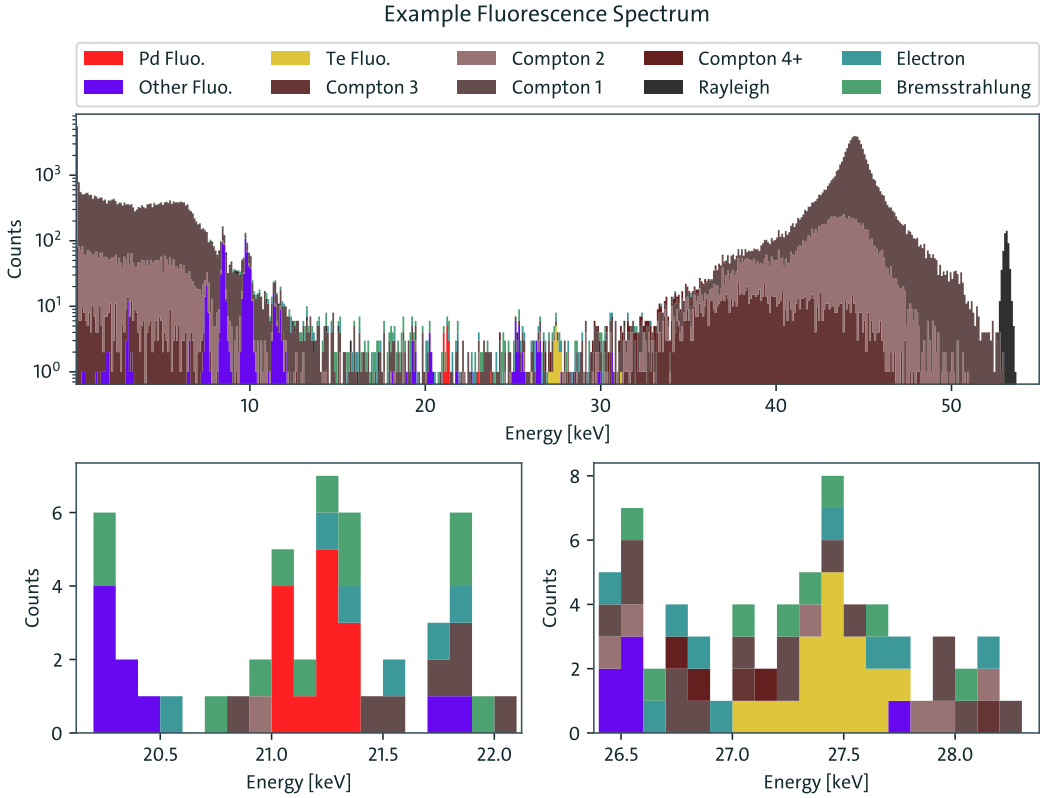

Figure 2.5. Example XRF spectrum: at high and low energies Compton scattered primary photons dominate, while the background level at medium energies is quite low allowing for fluorescence lines to become clearly visible. The measurement setup used to generate that data is shown in figure 2.4.

The spectra are subsequently subtracted. This method thus uses a reference measurement and local correlations. An advantageous property of this process is, that it can operate independent of energy resolution and involves no arbitrary choice of the experimenter in the evaluation. However, it requires scanning each subject twice, which can be challenging for dose or time constrained measurements.

An alternative is to first acquire a ***calibration curve***, as discussed in [16]. This method operates by correlating the relevant background region to a part of the spectrum that does not overlap with signal. Afterwards a subtraction of the estimated background is performed. The advantage over the subtraction method is that the individual subject only needs to be measured once, however it requires many reference measurements of similar subjects to build trust in the calibration curve and may struggle with more complex background components that have a non-linear relation to the reference component. A common choice for the reference region are photons that Compton scatter exactly once [16], as this results in a predictable and distinct peak in the spectrum. This method uses non-local correlations in the spectrum and requires a reference measurement.

2.4.5.2. Curve fitting

Curve fitting the spectrum in the signal region uses local correlation and does not require a reference measurement, which is a significant advantage over the other methods. This method is described in [17] where it is used for measuring iodine in murine thyroids. However, it requires a choice to be made regarding the basis functions used in the fit and how they are parameterized.

The last quadrant would be filled by using a priori knowledge about correlations like the ones calibrated through measurement in [16], but this has not yet been demonstrated in practice.

2.4.5.3. 2D Image Reconstruction

There are two primary ways to acquire 2D images in XFI. In the ***projection method*** the entire subject is illuminated at the same time. The fluorescence is then focussed by an aperture (for example a pinhole) onto a spatially resolved detector (compare figure 2.6 (a)). In the most basic form with a pinhole aperture the process is quite wasteful as a sharp image requires a small pinhole and thus most of the signal is rejected. The key advantages are that it can be carried out with a high divergence / large focal spot source and that the imaging time is solely determined by the required dose. Coded apertures like multi-pinhole ones can be used to reduce the rejection rate and/or improve resolution. [13]

The other important option is the ***scanning method*** whereby the subject is scanned with a pencil beam. This allows for detectors without spatial resolution and in principle arbitrarily large solid angle coverage to be used. The resolution of the process is determined by the beam size and scanning step width. Not being restricted to small solid angles means that scanning has the potential for much higher dose efficiency compared to projection. But as each scan point corresponds to a physical arrangement of the imaging setup, the measurement time also becomes limited by the speed of the mechanical components. For low energy, XFI beams can also be focussed down to µm size, granting this method very high spatial resolution. [13]

2.4.5.4. 3D Image Reconstruction

The most common method is ***tomography*** [12] [21] where, similar to CT, several 2D images are captured from different angles and combined into a final 3D image. The details of how tomography is reconstructed can be found in section 5.1.

Jung et al. show that the ***projection method*** can also be used to capture 3D images directly [15]. This method uses a fan beam, i.e. one that covers the $x-z$ plane. Fluorescence is thus only excited in a single y slice. A pinhole collimator and pixelated area detector are then placed above or below the object (centered on the y axis) with the area of the detector being co-planar to the $x-z$ plane. This projects an image of the excited slice onto the

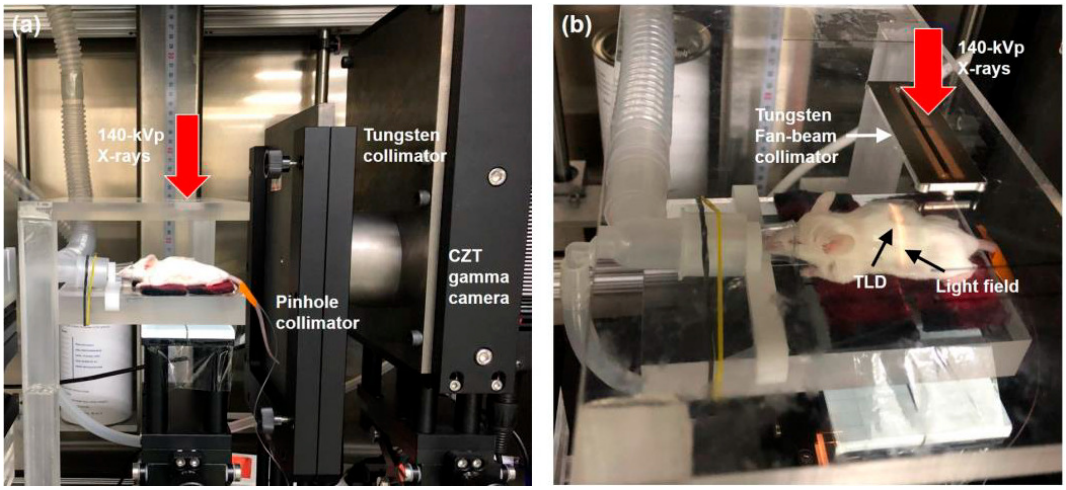

Figure 2.6. Imaging Setup for Projection Method: the imaging setup used for projection 3D images taken from [14]. (a) shows a side view with the fanbeam source above and the pinhole-collimated pixelated detector to the right of the subject. (b) shows a closeup of the subject with the fanbeam collimator.

detector. The pixels of the detector correspond to x, z positions and the y offset is determined by where the fanbeam strikes the object giving full 3D information. In this method there is an inherent tradeoff between dose efficiency and spatial resolution, as a smaller pinhole increases resolution, but decreases the effective sensitivity and vice versa. An example of such an imaging stage can be seen in figure 2.6.

2.4.5.5. Detection Tests

When one wants to ascribe a level of confidence for a detection of fluorescence signal one typically performs a one sided null-hypothesis test under the assumption of a point prediction for the background level which is then standardized and expressed as a z-test (compare subsection 3.1.5). Assuming the background is a Poisson distributed variable around b and there being no fluorescence signal, the number of total photons T is distributed as $T \sim \text{Poiss}[b] \approx \mathcal{N}[b, b]$. The observed realization of T is t and we define the "observed fluorescence" as $f := t - b$. This allows us to write the z-score for the observation compactly as

$$z = \frac{t-b}{\sqrt{b}} = \frac{f}{\sqrt{b}} \tag{2.9}$$

This test was used for example in [17]. The point-prediction for b is typically obtained from the observation itself using the methods discussed above.

It is worth noting explicitly that this test ignores uncertainty in the estimation of b. Extensions and alternatives to this method are discussed in section 4.6.

2.4.5.6. Use of Simulations

In the development of XFI as well as in measurement planning simulation is an important tool. [21], [54], [52] make extensive use of Monte Carlo physics simulations to conduct feasibility studies. These works use GEANT4 [27] which is a framework for simulating the interactions of particles with energies above keV with matter. These simulations evolve each particle separately [27], so the modeled behavior is inherently linear.

2.4.6. Comparison to Other Modalities

XFI is a functional imaging modality [12], meaning that it makes sense to compare it to methods like protein-based fluorescence/luminescence and methods based on radioactive tracers like SPECT and PET [18]. In comparison to the visible or UV light of bioluminescence and fluorescence, x-rays are much less impacted by the surrounding tissues, especially when

it comes to coherent scattering, allowing for high resolution reconstruction in depth. The tracers are not radioactive and may even be fully inert, meaning that no temporal limitations like in PET or SPECT arise. The main disadvantage of XFI compared to those methods is that the required equipment is currently not widely available and that measurements under dose and time constraints can be challenging depending on the tracer concentration [32]. Because the stimulation occurs with an external beam that has low interaction cross-sections in the tissue, the maximum possible resolution is determined by the beam size, not the inherent characteristics of the process or scattering [13], as in PET or the optical processes respectively [18].

2.5 Applications of XFI

Being such a flexible and powerful imaging technique there are many interesting applications of XFI, even if one limits the scope to preclinical research. This section attempts to briefly summarize a few case studies with interesting characteristics.

[17] shows XFI being used to localize the endogenous iodine in the thyroids of mice. The relevant region of the mice were scanned using a beam from an undulator source. An energy resolved detector was placed at an angle to the beam axis. The signal was estimated using spectral fitting. Finally, images of the integrated concentration in $\mathrm{ng\,mm^{-2}}$ were produced from the fitted number of fluorescence photons. This work is interesting as this would not be possible using methods that rely on special markers, at least not without extreme effort like replacing all iodine the mouse is exposed to for an extended period. Thus, showcasing one of the key advantages of XFI - that tracers do not need to be specially prepared, all atoms of appropriate elements can be utilized. Using a collimated beam, images of 200 µm resolution were showing details of the thyroids, thus demonstrating how resolution can be arbitrarily improved by reducing beam size using collimators.

[32] is especially interesting because it shows how XFI can be performed on a living subject, called in vivo, to capture relevant immunological processes that cannot be obtained through invasive methods. The measurement is fairly similar to [17], but instead of endogenous iodine, this experiment mapped the distribution of iodine-labeled immune cells. Another interesting aspect is that another detector on the beam axis, behind the subject was used to simultaneously acquire a transmission image. The simultaneous acquisition makes it possible to easily identify how the distribution of labeled cells corresponds to the mouse's anatomy.

In [16] another in vivo XFI experiment was performed. Here a ruthenium liposome complex was used as the tracer which is fluorescent both in the visible and x-ray domain. This allowed the authors to capture both optical fluorescence and XRF images. They used a liquid metal jet x-ray source which allowed for high resolution scans. Correlations between the background in the signal region and the Compton scattered incident photons were used to eliminate the background from the signal region. Quantitative CT scans of the organs were performed and compared to reference techniques like mass spectroscopy. This validated both the quantitative nature and the spatial detection capability of XFI.

In [33] a similar XFI method was performed to aid in the development of a process to avoid unwanted accumulation of nanoparticles. This also allowed the authors to locate a tumor in the mouse and perform in vivo XFCT in the tumor region. An important property of XFI, which this work makes uses, is that it is insensitive to the molecular scale. They thus could modify their procedure and nano-particles without potentially effecting the efficacy of XFI, aiding with comparability of results from different protocols.

[14] is another application of XFI in vivo, also concerning the distribution of nanoparticles. They used the projection method with a gamma camera to acquire the XFI image and the subtraction method to isolate the signal. This process makes it possible to acquire images very quickly as the scan only happens over a single axis. For registration

with the anatomy they used the Compton scattered primary photons which are inherently captured in almost all XFI applications and correlate with density in the subject similar to the attenuation mapped in CT images. They again showed good agreement with mass spectroscopy reference measurements.

2.6 Open Problems of XFI

2.6.1. Error Estimation

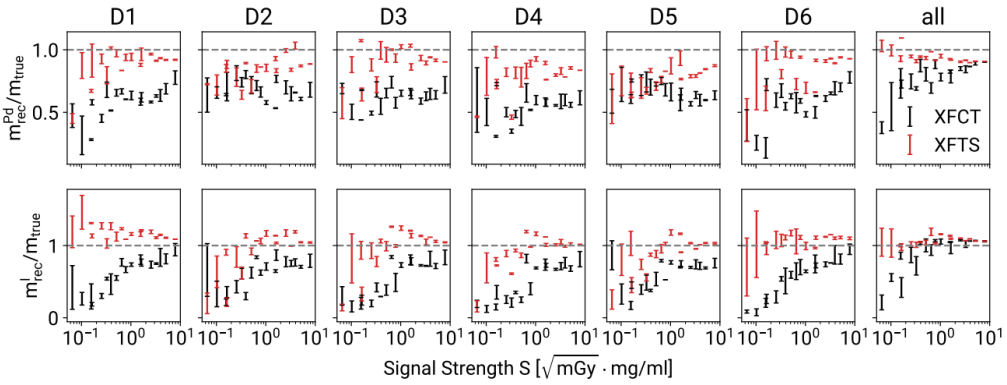

Figure 2.7. Underestimated Uncertainties: The top row of this plot from [21] shows estimated concentrations relative to ground truth values using two reconstruction techniques. Especially for XFCT (black) the the concentration is systematically underestimated with the uncertainty estimate not capturing that error.

Quantitative methods require uncertainty estimation to be interpretable, which generally requires more assumptions care than the point prediction itself [39]. The XFI literature currently does not show a consensus of methodology and many papers do not even offer uncertainty quantification. In other publications like [21] it is apparent that the errors are systematically underestimated, for example in the top row of Figure 2.7 from [21], one can see dozens of data points deviating from the true value by several uncertainty intervals. Even different reconstruction techniques applied to the same data disagree by several standard deviations in that plot. This behavior is especially apparent at low signal strengths. Later parts of this work will show that this is likely because of the usage of an error model with overly strong assumptions (see chapter 4).

2.6.2. Multi Detector Usage

Both the scanning and projection method for image acquisition can in principle use multiple detectors at the same time. However, in the reconstruction the data is usually processed individually for each one and only higher level results are combined, if at all [54]. An example can be seen in [17] where the entire analysis is performed with a single detector. Using multiple detectors in the reconstruction provides higher data efficiency and could thus reduce the required dose. Possible approaches to this problem are discussed in section 4.5.

2.6.3. Quantitative 3D Reconstruction

There are several techniques for 3D reconstruction like the tomosynthesis or tomography presented in [21]. However, as discussed above, these either do not provide an error estimation at all, or systematically underestimate uncertainty. Furthermore a method of 3D reconstruction that is more efficient with data could be of significant benefit, as in vivo 3D XFI is currently difficult due to dose and time constraints. Two novel approaches to this problem can be found in chapter 5.

CHAPTER 3

BAYESIAN METHODS

3.1 Bayesian Statistics

The goal of this thesis is to evaluate, improve and develop techniques for extracting relevant information from XFI data. A powerful statistical framework for these inference tasks is Bayesian statistics. The key idea is to use Bayes theorem (subsection 3.1.2) to update ones beliefs systematically to incorporate the new data. This framework lends itself to so-called logical probability, where probabilities are understood as generalizations of binary true-false which express subjective degrees of confidence, rather than long run frequencies [2].

3.1.1. Conditional Probabilities

Conditional probabilities in the logical frame describe the degree of confidence one has in a proposition Y given that one knows that X is true. This idea can be formalized using the language of mathematical statistics, where X and Y are realized as functions over some base-space Ω. The natural language statement above can be represented mathematically as

$$\begin{aligned} \Omega_{|X} =& \{o | X(o), o \in \Omega\} \\ \Sigma_{|X} =& \{S \cap \Omega_{|X} | S \in \Sigma\} \\ P_{|X} :& \Sigma_{|X} \to \bar{\mathbb{R}}_{\geq 0} \\ & S \mapsto P(S)/P(X) \end{aligned}$$

where $\Omega_{|X}$ is the space of base elements which correspond to X being true, $\Sigma_{|X}$ being the corresponding σ-algebra (an intermediate structure necessary for the definition of a probability measure) and $P_{|X}$ being the probability measure for Y over that space. Simply put, one restricts the space of all possibilities to those that match the provided information. The probability assigned to each of the possibilities must then be normalized correspondingly by a factor of $P(X)$. The probability of Y is then also normalized in the same way. [1]

An equivalent way to write this is simply [1]:

$$P(Y|X) := \frac{P(Y \wedge X)}{P(X)} \tag{3.1}$$

It is straight forward to check that both agree

$$P_{|X}(Y) = P(Y|X)$$

3.1.2. Bayes Theorem

One interesting thing about equation (3.1) is that it is symmetric in X and Y, meaning we can start with the joint and express it as two different conditionals

$$P(X|Y)P(Y) = P(X \wedge Y) = P(Y|X)P(X)$$

$$\Rightarrow P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{3.2}$$

This equation is called Bayes theorem [1]. It follows almost trivially from the modern definition of conditional probabilities, but it is extremely powerful because it allows for the inversion of conditionals. This allows us to use evidence to update our beliefs systematically and without risking confusing $P(X|Y)$ for $P(Y|X)$ which is the well known prosecutors fallacy [44].

Relating this idea to scientific data reads as: $P(Y|X)$ being "the probability of the hypothesis given the data" (also called the posterior), $P(X|Y)$ being "the probability of the data given the hypothesis" (or likelihood) and $P(Y)$ being "the prior probability for Y" i.e. the degree of confidence in Y before the data X became available. While $P(Y|X)$ is generally what the scientists wants to know, $P(X|Y)$ is generally what a model (like the three stage models this work uses to describe medical imaging) will provide.

3.1.3. Moments and Cumulants

An important set of properties of a distributed variable X, that finds a lot of use in statistics, are the moments $\mu_n[X]$ around c. While moments are not enough to fully constrain a distribution, they are powerful enough to select for a distribution with the correct behavior in the relevant region [1]:

$$\mu_n = \mathbb{E}\left[\,(X-c)^n\,\right]$$

where $\mathbb{E}\left[\,\cdot\,\right]$ denotes an expectation value. If no reference point is specified it is usually assumed to be $c = 0$. Moments around $c = \mathbb{E}\left[\,X\,\right]$ are called central moments.

The first moment is just the expectation value $\mathbb{E}\left[\,X\,\right]$ itself, also called the mean and written as $\mu[X]$ or just μ. The second central moment is called the variance Σ or $\mathrm{Var}\left[\,X\,\right]$. The standard deviation is defined as the square root of the variance $\sigma = \sqrt{\mathrm{Var}\left[\,X\,\right]}$. Moments are also often standardized, the n-th standardized moment $\tilde{\mu}_n[X]$ is given by

$$\tilde{\mu}_n[X] := \frac{\mathbb{E}\left[\,(X-\mu)^n\,\right]}{\sigma^n} = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^n\right]$$

Therefore, the standardized moments can also be interpreted as the moments of a standardized random variable $X^* = \frac{X-\mu}{\sigma}$. [1]

The moments of a distribution carry a lot of information about the shape of the density. The first moment indicates where the relevant part is roughly located, the second indicates the width, the third the asymmetry and the fourth the thickness of the tails.

As is often the case with sequences, it is useful to define a generating function for the moments, i.e. a function such that its derivatives generate the moments. One way to do this is with the function $\varphi_X(t)$ with $i^n\left[\frac{d^n}{dt^n}\varphi_X\right](0) = \mu_n[X]$ which is called the characteristic function of X. One such function is

$$\varphi_X(t) := \mathbb{E}\left[\,\exp(-itX)\,\right]$$

This choice expands to the Fourier transform of the density of X if X admits one. Note that $\mathbb{E}\left[\,\exp(itX)\,\right]$ is more common in literature [1] leading to mismatching signs in odd-order terms, while this work uses the given sign convention because it makes the relation to Fourier transforms more explicit.

Moments can also be generalized to vectorial random variables

$$\mu_n = \mathbb{E}\left[\,(\vec{X}-\vec{c})^{\otimes n}\,\right]$$
$$\varphi_X(\vec{t}) = \mathbb{E}\left[\,\exp\!\left(-i\vec{t}\cdot\vec{X}\right)\,\right]$$

where $\cdot^{\otimes}n$ denotes repeated tensor products. The second central moment again is particularly important. It is called the covariance $\Sigma = \mathrm{Cov}\left[\,\vec{X}\,\right]$ in its matrix form to distinguish it from the scalar variance (using scalar instead of tensor product) which is the trace of

the covariance. Standardization also generalizes, but X^* is not uniquely determined by X, instead depending on a choice of invertible S such that $\Sigma = SS^T$ giving $X^* = S^{-1}(X - \mu)$. This process removes correlations between different elements of X and is often called whitening. [1]

An alternative to the moments are the cumulants κ_n. They are generated by the second characteristic function $\psi_X := \log \varphi_X$ and

$$\kappa_n = i^n \left[\frac{d^n}{dt^n} \psi_X \right] (0)$$

One thing that is convenient about the cumulants is that they do not contain contributions from lower cumulants, so while with the moments we had to define the variance as the second *central* moment, we find that $\kappa_2[X] = \mathrm{Var}\,[\,X\,]$, so no centering is required. [45]

If we have two independent random variables X, Y and $Z = X + Y$, we can see that

$$p_Z(z) = \int p_{Y|X=x}(z - x|x) p_X(x) dx = \int p_Y(z - x) p_X(x) dx = [p_Y * p_X](z)$$

where the $*$ indicates convolution. Because the characteristic function is a form of Fourier transform it thus follows that $\varphi_Z(t) = \varphi_X(t)\varphi_Y(t)$ and finally $\psi_Z(t) = \psi_X(t) + \psi_Y(t)$ meaning that cumulants are additive:

$$\kappa_n[X + Y] = \kappa_n[X] + \kappa_n[Y]$$

Similarly one finds that $\psi_{cX}(t) = \psi_X(ct)$ and $\kappa_n[cX] = c^n \kappa_n[X]$. [45] This is what enables Gaussian error progression, but will also be used in this work to quantify deviations from normal distributions.

The first few cumulants will play an important role for approximate inference in this work, where an exact Bayesian posterior is too difficult to find, but finding its first few cumulants is possible.

3.1.4. Common Types of Random Variables

This work uses several well known distributions, which are introduced here.

3.1.4.1. Normal Distribution

The normal distribution has a preferred place in statistics because for any independent realizations x_i of a distribution with finite mean μ and variance σ^2, their average $\frac{1}{n}\sum_{i=1}^{n} x_i$ converges to the normal $\mathcal{N}\,[\,\mu, \sigma/\sqrt{n}\,]$. [1]

The density of the normal distribution can be found to be [1]

$$\mathcal{N}\,[\,\mu, \sigma^2\,]\,(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

With the multivariate (vectorial) generalization:

$$\mathcal{N}\,[\,\vec{\mu}, \Sigma\,]\,(\vec{x}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$$

where Σ is the covariance matrix, which must be positive-definite and square.

The normal distribution is important in this work because of a combination of two facts. Firstly, many distributions like the Γ distribution become approximately normal in certain limits. Secondly, the linear functions of normal distributions are themselves normally distributed with a simple closed form relation [1]. These two properties together mean that in certain limiting cases the normal distribution can be used to approximate the result of linear functions of other distributions as well which will be used for processing spectra.

3.1.4.2. Poisson Distribution

The Poisson distribution Poiss [λ] with rate λ is a discrete random variable. The defining Poisson process is the number of discrete events happening over a continuous interval with uniform likelihood and fixed expectation value λ. This is often discussed as a limit of a binomial process. The argument goes that we divide the interval up into N equal chunks. If the chunks are small enough we expect at most one event to happen in the chunk. The probability of such an event being found in a given interval is $p = \lambda/N$. Thus the total number of events would follow a binomial distribution B$[N, p]$. As $N \to \infty$ this discretized version of the Poisson process should converge to the continuous case and thus we can find the Poisson likelihood as

$$\begin{aligned} P_\lambda(k) &= \lim_{N\to\infty} \binom{N}{k} (\lambda/N)^k (1-\lambda/N)^{N-k} \\ &= \lim_{N\to\infty} \left[\prod_{i=0}^{k} \left(1 - \frac{i}{N}\right)\right] \left(1 - \frac{\lambda}{N}\right)^N \frac{\lambda^k}{k!} \\ &= \frac{\lambda^k e^{-\lambda}}{k!} \end{aligned}$$

An interesting property of the Poisson distribution is that the variance equals the mean equal λ. [1]

An example of a Poisson process is photon detections as they happen as discrete events over continuous time. As such, the number of counts in a photon energy histogram (e.g. in an XRF spectrum) is Poisson distributed. Because of this Poisson noise plays a central role in quantitative medical imaging in general [18] and this thesis in particular.

3.1.4.3. Γ Distribution

In Bayesian statistics it is often useful to use very general parametric families of distributions, meaning those where a wide range of possible distributions can be generated by simply adjusting the parameters. The Γ distribution is one such family. It includes the χ^2 and exponential distribution as special cases and normal distribution as a limit. It is parameterized by the shape α and rate β and has the probability density

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

with mean $\mu = \alpha/\beta$ and variance $\sigma^2 = \alpha/\beta^2$. [1]

More relevant to this work, the Γ distribution has a special relation to the Poisson distribution in Bayesian statistics. If one uses a Γ distributed prior for the rate parameter $\lambda \sim \Gamma[\alpha, \beta]$, one finds that after updating that belief for an observation $k \sim$ Poiss [λ] that the posterior is $\lambda \sim \Gamma[\alpha + k, \beta + 1]$. The complexity of Bayes theorem reduces to a simple update rule for the parameters. This relation is described by saying Γ is the conjugate prior to the Poisson distribution (compare equation (4.1), [1]).

3.1.4.4. Negative Binomial Distribution

Similar to a binomial random experiment, one can imagine performing random experiments with only two outcomes, success and failure with success rate p (so called Bernoulli trials) until one finds r successes, the number of total failures k then follows a negative binomial distribution NegBin [r, p]. It has the probability mass function of

$$P(k) = \binom{k+r-1}{k} (1-p)^k p^r$$

and its mean is $\frac{r(1-p)}{p}$ and variance is $\frac{r(1-p)}{p^2}$. [1]

In this work however, it appears because of its relation to the Poisson distribution. If we sample rate parameters from a Γ distribution (e.g. the posterior for it after an experiment) and then sample a Poisson distribution around that the final result follows a negative binomial distribution:

$$\begin{aligned} \lambda &\sim \Gamma\left[\,\alpha, \beta\,\right] \\ k &\sim \text{Poiss}\left[\,\lambda\,\right] \\ \Rightarrow k &\sim \text{NegBin}\left[\,\alpha, \frac{\beta}{1+\beta}\,\right] \end{aligned}$$

This is relevant for example when a realization from a Poisson distribution is given and attempting to form a belief about what the next realization will be. Generally one would say that the negative binomial is the posterior predictive for Poisson with a Γ prior [1].

Later chapters will use this fact to generate en ensemble of plausible future data from an observation (subsection 4.1.2, section 4.6).

3.1.4.5. χ^2 Distribution

The χ^2 distribution with ν degrees of freedom is the distribution of the sum of the squares of ν standard normal random variables. This can be found to be a special case of the Γ distribution, namely $\Gamma\left[\,\nu/2, 1/2\,\right]$. Even more generally if $q \sim \chi^2\left[\,\nu\,\right]$ then $c \cdot q \sim \Gamma\left[\,\nu/2, 1/2c\,\right]$. An important special case of the latter is the reduced χ^2 distribution $\chi^2_{\text{red.}}$ where $c = 1\nu$, which has a fixed mean of 1 and variance of 2ν. [1]

This distribution most commonly comes up when comparing a data sequence to a model predicting means and standard deviations for each data point where the noise is assumed to be independent and normal. One can then standardize the individual errors and sum their squares to get a χ^2 distributed variable. The degrees of freedom are then the number of data samples minus the number of model parameters. Dividing by the number of degrees of freedom makes comparison between models of different parameter numbers easier (see subsection 3.1.5.3). [1]

3.1.5. Statistical Tests

Bayes theorem provided us with a framework to assign confidence in a hypothesis given an observation. However, the full posterior for a complex proposition is not always tractable to compute. In those cases it is often useful to calculate the confidence in a proposition that is either necessary or sufficient for the original hypothesis and easier to compute. These surrogate propositions are often called tests. The most common type of tests use necessary conditions, corresponding to an upper bound of the posterior. This means that most tests are built to falsify a hypothesis rather than validating them. This is because such an upper bound can often be found without evaluating large model or parameter spaces. [1]

3.1.5.1. z-Test

The simplest test leans heavily on the law of large numbers and central limit theorem. Suppose we have n independent and identically distributed samples $x_i \sim X$ and our hypothesis is that $x_i \sim Y$ where Y has mean μ_y and variance σ^2. For large n or if Y is normal, the hypothesis implies that the sample mean $\bar{x} := \frac{1}{n}\sum_i x_i$ would be distributed according to $\bar{x} \sim \mathcal{N}\left[\,\mu_y, \sigma^2/n\,\right]$. We can standardize this distribution to $z := \sqrt{n}\frac{\bar{x}-\mu_y}{\sigma}$ where the expectation would be that z is standard normal. This value is called the test statistic of the z-test. The structure of computing a value from the samples which under the assumption of the model follows a known distribution is how most tests work. [1]

Because z is standard normal, its value provides a universal reference for how unusual a result is, for example values like $z = 10$ should be exceedingly rare and likely suggest that the assumption was wrong. When we want to convert a z-score to an actual probability we must make a decision on whether we ask "how likely is it to observe a value this strange, regardless

of its sign?" or restrict it to one side - asking "how likely is it to observe a value this large?" or "... this small?" instead. The former case is more common in general and corresponds to the famous $67, 95, 99.7\%$ rule where those probabilities correspond to $z = 1, 2, 3$ respectively, meaning that observing $|z| > 3$ has a likelihood of only about $< 0.5\%$.

z-tests are very common and other tests often convert their test statistics to a z-value for easy comparison. Even so, the z-test has a few limitations. Firstly, on the assumption that we have a large enough sample size that the central limit theorem makes normality of the sample mean a good approximation. Secondly, we require knowing both the mean and standard deviation of the model distribution a priori. Lastly, this test only provides an upper bound on the likelihood $P(D|A)$, it neither considers the prior and evidence, nor does it provide an exact value. A good example where this fails would be testing the model that a coin is perfectly fair, meaning heads and tails occur equally often and results are uncorrelated and then finding a data where heads and tails perfectly alternate in a long sequence. The data is exceedingly unlikely under the hypothesis, but the z-test cannot tell. But where it is suitable it provides an easy to implement and strong test, making it especially useful for testing models that can be formulated fully a priori, like comparing data to a physics model and where the distribution Y is (approximately) normal to begin with to reduce the required sample sizes. [1]

3.1.5.2. Student t-Test

The Student t-test, originally published by William Gosset under the pseudonym Student, attempts to remove the z-test's requirement for a priori knowledge of σ. Here the sample standard deviation s with $s^2 := \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$ instead of population variance (true variance of the distribution of x) is used, giving the test statistic $t := \sqrt{n}\frac{\bar{x}-\mu}{s} \sim \mathrm{StudT}\,[\,n-1\,]$. For small n this distribution has more mass further from the mean than the normal distribution (fatter tails or positive excess kurtosis) which is expected because less total information is available. For large n it converges to a normal distribution as would be expected under the central limit theorem. The excess kurtosis (fourth cumulant) is $\frac{6}{n-5}$, so when significantly more than eleven samples are available $t \approx z$. [1]

This test is one of the most common tests used in statistics. Only needing a point prediction from the comparison model makes it a lot more suited for cases where the noise in a system is not known a priori and the condition for the test statistic to follow the test distribution is actually even weaker than independent and identically distributed normality or central limit theorem compatible sample size, with certain symmetry conditions being sufficient as well. This makes the t-test well suited for use with small sample sizes where a z-test would be overly confident. [1]

3.1.5.3. χ^2-Test

A lot of models produce many (approximately) normally distributed variables, like a fit for a curve. For each of them a z-score can be computed. As discussed before, usually the square of z is much more interesting than the signed value, so an obvious candidate would be $\chi^2 := \sum_i z_i^2 \sim \chi^2\,[\,n\,]$. Often the standardized $\tilde{\chi}^2 := \frac{\chi^2 - n}{\sqrt{2n}}$ or reduced value $\chi^2_{\mathrm{red}} = \frac{\chi^2}{n}$ is used to more easily compare results with different numbers of degrees of freedom. The excess kurtosis is $12/k$, so similar to the Student-t for significantly more than twelve samples it converges to $\tilde{\chi}^2 \approx z$. [1]

When the model is fit to the data, the number of degrees of freedom of the data is reduced by the number of fit parameters, and we find

$$\tilde{\chi}^2 = \frac{\chi^2 - n + p}{\sqrt{2(n-p)}}$$

$$\chi^2_{\mathrm{red}} = \frac{\chi^2}{n-p}$$

Here p is the number of free parameters of the model. This is very important because in models with $p \approx n$ we expect the data to be fit much better, for example, a polynomial model could always exactly match p points. [1]

This test is well suited to compare fits or predictions of a curve to the data if only one sample per curve point is available. This makes it widely applicable in physics. However, similar to the z-test it requires a priori characterization of the noise. For this reason χ^2 tests will be used to evaluate certain models (subsection 4.5.1.2).

3.2 Key Concepts of Information Theory

As discussed before, full Bayesian inference is often not computationally tractable ([1]). There are many possible solutions to this problem, but this thesis will lean on information theory for two reasons. Firstly, information theory provides a wide set of powerful tools that have become key enablers in many domains ([46]). Secondly, information theory is closely tied to modern thermodynamics ([10]), so many powerful intuitions translate well for physicists.

3.2.1. Shannon Information

Bayes theorem (equation (3.2)) tells us how to update probabilistic beliefs given new data, but this process is multiplicative. Using negative logarithmic probabilities $I(A) := -\log P(A)$ yields an additive form that is often easier to work with:

$$\begin{aligned} I(A|D) &= I(D|A) + I(A) - I(D) \\ &= I(D|A) + I(A) + \text{const.} \end{aligned}$$

where the constant $I(D)$ is fully determined by the normalization criterion. So the Bayesian updating process reduces to adding something, a negative logarithmic probability, to our prior. Negative logarithmic probabilities are also called (Shannon) information. This name is used because it captures the notion that the data presents us with new information (in the colloquial sense of the word) that we can add to the previously known information. The negative sign ensures that very unlikely events carry large amounts of information and vice versa. [46]

One important property of information is that the outcome of minimum information is the outcome of maximum probability, meaning minimizing information is equivalent to maximizing probability. This is very useful because information is often easier to estimate (especially when constants can be neglected) and because the logarithm often increases numerical stability. [46]

3.2.2. Entropy

Shannon information tells us about how much information one needs to describe a specific sample from a given distribution. When trying to make a more general statement about entire probability distributions it would thus make sense to look at the expected information

$$H = \mathbb{E}[I]$$

Higher entropy means that the information carried by an individual outcome tends to be very large, while low entropy means that we expect the outcome to carry very little extra information. Loosely, the higher entropy distribution is less informative by itself, while the lower entropy distribution also informs us a lot about the possible outcomes. This means that large uncertainty should correspond to high entropy. The prior that entropy should be maximized is called the maximum entropy principle. [46]

We can also measure entropy of the distribution Q relative to the distribution P over a shared space, meaning how much information you need on average to describe an outcome

sampled according to P in terms of Q:

$$H(P,Q) = \mathop{\mathbb{E}}_{x\sim P}\left[\, I_Q(x) \,\right]$$

This is called the cross-entropy [1]. When $P = Q$ this simplifies to the entropy of P. Increasing $H(P,Q)$ indicates increasing differences between P and Q. In the most extreme case, there is an outcome that P allows, but which Q excludes. This causes the cross-entropy to diverge. However, the cross-entropy does not provide a good notion of "distance" in the space of probability distribution because $H(P,Q) \geq H(P)$, so importantly $H(P,P) = H(P) \neq 0$. This also means that only differences in cross-entropy are meaningful, as there is not an inherent reference point. [46]

So while the cross-entropy provides a good measure of differences if we want to compare different Q, but not a very useful one when we want to understand different P. For example $H(P,Q)$ makes a good loss function for gradient-based techniques optimizing Q (as constants do not matter), but not for P, as the offset would not be constant anymore. In general one cannot just change positions of the two, because one needs to be able to calculate or estimate the expectation value over P, meaning that one needs access to P directly.

3.2.3. Kullback-Leibler Divergence

One way to mitigate this issue is to just subtract off $H(P)$ from $H(P,Q)$ yielding

$$\begin{aligned} D_{\mathrm{KL}}\left(\, P||Q \,\right) = H(P,Q) - H(P) &= \mathop{\mathbb{E}}_{P}\left[\, I_Q - I_P \,\right] \\ &= \int \log(p(x)) - \log(q(x))\, dP(x) \end{aligned}$$

which is called the Kullback-Leibler (KL) divergence. It has the metric-like properties that $D_{\mathrm{KL}}\left(\, P||Q \,\right) = 0 \equiv P = Q$ and $D_{\mathrm{KL}}\left(\, P||Q \,\right) > 0$ meaning it can be used as an optimization target for either of the two distributions, as well as giving values of it an absolute reference point. This makes the KL-divergence an excellent measure of difference, but it also requires access to the densities of both distributions and being able to find an expectation value with respect to one of them. [46]

3.2.4. Evidence Lower Bound

A special case, which is central to this work, is when Q is a posterior distribution for x given data d, using Bayes theorem for information we can write:

$$\begin{aligned} D_{\mathrm{KL}}(P||Q(x|d)) &= \mathop{\mathbb{E}}_{x\sim P}\left[\, I_Q(x|d) - I_P(x) \,\right] \\ &= \mathop{\mathbb{E}}_{x\sim P}\left[\, I(d|x) + I_Q(x) - I_P(x) + I_Q(d) \,\right] \\ &= \mathop{\mathbb{E}}_{x\sim P}\left[\, I_Q(d|x) + I_Q(x) - I_P(x) \,\right] + I_Q(d) \\ &= \mathop{\mathbb{E}}_{x\sim P}\left[\, I_Q(d|x) + I_Q(x) - I_P(x) \,\right] + \text{const.} =: \mathcal{L}_Q[P] \end{aligned}$$

Here, $I_Q(x)$ is the prior information and $I_Q(d|x)$ is the likelihood to observe the data and $I_Q(d)$ is the evidence. Since $D_{\mathrm{KL}}\left(\, P||Q \,\right) > 0$ we can conclude that $\mathcal{L} < I_Q(d)$, so $\mathcal{L}$ is a lower bound of the evidence, it is therefore called the "evidence lower bound" or ELBO for short. The ELBO is generally a lot more accessible than the KL-divergence. For one the prior is a choice and thus can be chosen to be accessible. The likelihood is an expression of the underlying model of the data and the inference target and thus also generally accessible. In comparison, the full KL-divergence to the true posterior would also include the difficult evidence term. This makes it a very popular choice as an optimization target for finding approximate posterior distributions, or more generally for approximating distributions where one is only accessible implicitly. [47]

A key part of this work is the development of a dedicated generative inference model for Bayesian inference in the XFI context. This model relies on the EBLO as an optimization target.

3.2.5. Information Criteria

As discussed before two of the key tasks in statistical inference are hypothesis testing and model selection. The former can often be reduced to the latter by selecting between a model that assumes the hypothesis and one that assumes its negation. So it is sufficient to find a way to compare the goodness of two models.

A common version of this problem is parametric approximation. Here one has access to some data $d_i, i \leq n$ which are exactly described by some unknown distribution $f(x)$. One attempts to describe this data given a parameterized distribution $p(x|\vec{\theta})$ where $\vec{\theta}$ are the parameters. It is important to note that one does not need to assume that f is in this parametric space, but one can try to approximate f with this parametric family nonetheless. Now one attempts to find the distribution over the parameters $p(\vec{\theta})$ such that $p(x) = \int p(x|\vec{\theta})p(\vec{\theta})\, d\theta$ approximates the target distribution $f(x)$. [1]

Finding the full distribution $p(\vec{\theta})$ is often not feasible. What is usually possible however, is finding a specific set of parameters $\vec{\theta}^*$ which approximate the distribution well (for example by minimizing the KL-divergence with respect to θ). If one wants to compare two different models, possibly with different numbers of parameters, one would need a way to approximate the KL-divergence of the full models to the true distribution by only using the single "optimal" set of parameters and the finite samples, without exploring the entire parameter space. But for comparison purposes it is enough to find the KL-divergence up to a shared constant. [4]

One way of doing this is to expand around $\vec{\theta}^*$ which ultimately leads to

$$D_{\mathrm{KL}}\left(\, f||p\,\right) \approx 2\left(k - \log p\left(\vec{d}|\vec{\theta}^*\right)\right) + \text{const.}$$
$$\mathrm{AIC} := 2\left(k - \log p\left(\vec{d}|\vec{\theta}^*\right)\right)$$

where $k = \dim \vec{\theta}$ [4]. This is called the Akaike Information Criterion (AIC) [4]. We can see that the AIC not only favors models that fit the data well, but also models with fewer parameters. This is similar to Occam's razor, but Akaike can be understood as providing a derivation of that heuristic from first principles. The reason for that can be understood as the volume growing faster with radius in higher dimensional spaces, meaning that the expected distance from the region of high likelihood is larger in those spaces [46].

A common criticism of the AIC is that it is not asymptotically consistent in the estimate model degree ([4]), meaning that as we get more and more data the model it favors might end up having more and more parameters. From the conceptualization of the AIC we can see that this criticism is flawed in that it assumes the AIC attempts to find a "true" model rather than just an estimation from a given space (in which case there is no reason to expect asymptotic consistency). Nonetheless this is a useful property for a model selection criterion to have. Bozdogan [4] modifies the AIC to give it that property, providing the consistent AIC (CAIC)

$$\mathrm{CAIC} := k\left(\log(n) + 1\right) - 2\log p\left(\vec{d}|\vec{\theta}^*\right)$$

where n is the dimension of the $\vec{d}$. We can see that the CAIC favors lower model degrees for $n \geq 3$, and especially in the high n limit. The CAIC is in a way "less pure" than the AIC as the choice of $\log(n)$ is at least somewhat arbitrary, though motivated.

Another common information criterion that looks similar on the surface is the Bayesian information criterion (BIC) [3]. The key distinction in assumptions is that we consider the true distribution to be part of the parameterized space. The BIC attempts to estimate $\log p(\vec{\theta}^*|\vec{d})$ up to a constant which corresponds to estimating the Bayes factors

$p(\vec{d}|M_1)/p(\vec{d}|M_2)$ for any two candidate models M_1 and M_2 that are both parameterized. The BIC can be found to be

$$\text{BIC} := k\log(n) - 2\log p\left(\vec{d}|\vec{\theta}^*\right)$$

For data of dimension $n = 8$ or higher the BIC favors lower dimensional models than the AIC [3]. [4] There are many other possible choices each of which uses a slightly different weighting between k and $p\left(\vec{d}|\vec{\theta}^*\right)$ and come from slightly different ways to conceptualize model selection. [3] provides a larger overview with even more candidates. This work favors the AIC and the BIC as standard choices depending on context and preferred behavior in the large n limit. For example, when fitting a complex background function it might be desirable that higher and higher model degrees be used as the data becomes sufficiently strong to properly constrain such a model.

To make comparison to more classical methods easier and provide motivated choices for cutoff values of the information criteria in testing we can consider the case where two single variable Gaussian models with equal variance are compared. The classical method would compare the $z = \frac{\mu - d}{\sigma}$ values. This is a useful comparison case as tests against other distributions are also often converted to z-scores. We can see that $\log p = -\frac{z^2}{2} + \text{const.}$ where the constants would cancel in the model comparison. This suggests that thresholds like 4 or 10 would make sense to be roughly comparable to 95% or 99.9% confidence in the class of $2\log p$ criteria. [1]

3.3 Generative Models

One of the key contributions of this work is a generative Bayesian model for XFI image reconstruction tasks. At its heart a generative model is any statistical model that is able to generate samples from the distribution it is modeling.

Modeling a distribution is sometimes possible without having access to the densities directly, especially in high dimensional spaces. For example, for any given matrix $\mathbf{M}$ we can construct a model $X = \mathbf{M}\lambda$ where λ is a standard normal variable. But calculating the density of X requires inverting $\mathbf{M}\mathbf{M}^T$ which may not be tractable for large numbers of dimensions. Alternatively the modeling function may be non-linear which could lead to difficult normalization integrals.

Generative models have recently received a lot of attention in the Deep Learning community and the general public alike through the success of model classes like diffusion models with Stable Diffusion or auto-regressive transformers with the GPT family [56]. Generative models in deep learning are however much more diverse and older, including other model classes like variational autoencoders (VAEs), generative adversarial networks (GANs) and normalizing flows. [5] provides a broad overview and categorization of model classes and lists some of their successes, like improvements in MRI reconstruction.

3.3.1. Normalizing Flows

Normalizing flows serve as the best comparison point for the model used in this work. Normalizing flows are parameterized functions that map between the target space, which is to be modeled, and a standardized space with a fixed density (for example n-dimensional standard normal). The parameters are later optimized in the training process. The important thing is that the flows are by construction bijective with a known inverse and have easy to calculate Jacobi determinants [48]. These are the determinants of the Jacobian of the function between latent and target space and whose logarithms are important in the estimation of the KL-divergence (subsection 3.2.3) or the ELBO (subsection 3.2.4). This allows for the explicit calculation of the density in the target space through the change of variable relation. In Yazdani's classification ([5]) this makes them explicit density models.

The classic way to do this is

$$\begin{aligned} f_\theta(x_1, \dots, x_{2n}) &= (x_1, \dots, x_n, s_1 x_{n+1} + t_1, \dots, s_n x_{2n} + t_n) \\ \vec{s} &= \exp(g_\theta(x_1, \dots, x_n)) \\ \vec{t} &= h_\theta(x_1, \dots, x_n) \end{aligned}$$

where g_θ and h_θ are arbitrary learned functions [48]. This is invertible, because the parameters, which are used to generate the scale and offset, are not transformed, meaning they can be calculated from the output. The exponential ensures that s is positive to guarantee bijectivity and also means the log Jacobi determinant is just a sum over the output of g. To achieve sufficient generality many of these transforms are applied with different partitions of the indices [48].

Normalizing flows have been used for many tasks including image modeling and variational inference [48]. The latter is essentially the same task this work attempts. The issue with flows in this context is that, in their most general form, they are black box models. The model parameters and intermediate outputs do not carry physical meaning, so it is difficult to formulate priors over them [49]. However, the key idea of a forward model that has bijectivity guaranteed and logarithmic Jacobi determinants available by construction is maintained.

3.3.2. ELBO Loss

Many generative models, like normalizing flows, can be understood as a mapping f_θ between a known base space Λ and the data space $\mathcal{D}$. Normalizing flows and VAEs are typically trained by taking samples from $\mathcal{D}$ and then choosing the model weights such that $f_\theta(\mathcal{D}) \approx \Lambda$. As discussed in subsection 3.2.3 the KL-divergence provides a useful measure of fitness that can be estimated from just the samples. For flows especially the ELBO (subsection 3.2.4) is often chosen because it is easier to calculate when one has direct access to the logarithmic Jacobi determinants [50].

3.3.3. Variational Inference

One application of generative models, which is also what this work attempts, is variational inference (VI). The goal of VI is to estimate the posterior distribution $p(x|d)$ of some parameters x from space X given data d using a parameterized family of probability distributions $p_\theta(x)$. This can be reframed as finding a $f : \Lambda \to X$ from a base space with known density (like standard normal) with equal dimension to X, such that $p(f_\theta(\lambda)) \approx p(x|d)$ [50]. This is in essence a similar problem to that which VAEs and normalizing flows attempt to solve, but instead of having access to samples from the target space we have access to the log-likelihood up to a constant

$$\log p(x|d) = \log p(x) + \log p(d|x) + \text{const.}$$

where $\log p(x)$ is a prior we freely choose and $\log p(d|x)$ is the likelihood of observing the data given the sample from x. [6] provides a general overview of VI methods of this class.

The method which this work uses is to take samples λ_i from Λ, apply the parameterized transform f_θ to get samples $x_i = f_\theta(\lambda_i)$ and then calculate the ELBO-loss as

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i) + \log p(d|x_i) - \log p(\lambda) + \log ||J_{f_\theta}|| \tag{3.3}$$

$$= \hat{H}(X) - \hat{H}(D, \hat{D}) \tag{3.4}$$

where $\hat{H}$ is the (cross) entropy estimate through sampling and $\hat{D}$ is the approximate data, obtained by pushing X through the observation model. The entropy framing also shows us what the model is trying to balance, the goodness of the data fit with not letting X collapse to its mode.

The structure of the model consists of a base space Λ, a mapping to the space of inference targets X, a mapping to the space that parametrizes the data distribution Y and a log likelihood function for the data

$$\begin{aligned} f_\theta : \Lambda &\to X \\ g : X &\to Y \\ \log p(d|\cdot) : Y &\to D \end{aligned}$$

In order for the model to be well behaved f needs to be bijective and differentiable. To be able to train the model, g needs to be implicitly differentiable (using the training backend), $f_\theta(\lambda)$ needs to be differentiable in θ and the logarithmic Jacobi determinant with respect to λ needs to efficient to compute.

This work chooses not to use normalizing flows with neural networks as the parameterized function, instead using a hierarchical approach with transforms chosen a priori for their properties to bring in domain knowledge. This allows for the use of priors for abstract properties of X like the length scale on which it varies. To accomplish that specific example, one adds a parameter φ to the inference that characterizes the over which X typically varies. Then, as part of f, one applies a transform T_φ which induces correlations of that length scale. The same process can be used to form priors over other properties, like by how many orders of magnitude the minima and maxima of x differ, constraints to a given value range etc. This makes the hierarchical approach very flexible and powerful, and partially addresses some of the black box concerns of flows.

An important distinction between the type of model used in this work and variational auto-encoders or normalizing flows for tasks like image generation is the space in which the ELBO loss is computed and the direction of data flow in training. Said models in literature typically project data samples into the latent space and estimate the KL-divergence between the transformed data distribution and the target latent space distribution [48]. The model in this work starts with a predefined latent space, transforms it into the observable space and estimates the KL-divergence between that distribution over the observables and the Bayesian posterior for the observables given the data.

CHAPTER 4

PROCESSING XFI SPECTRA

One of the key tasks in XFI measurements is processing the spectra. Detectors only provide numbers of counts of in a given energy range, but the experimenter is generally interested in differentiating between signal and background [12]. This chapter studies the statistics of the established methods used for this task, which leads to new correction terms that are not currently used in the literature. It further proposes how Bayesian methods can be used to perform this task and evaluate the performance of the existing and new methods. Additionally, a method for predicting future measurements from simulation data is developed, which is also used to generate synthetic data for the evaluation task mentioned before. The new statistical evaluation methods are also shown to be capable of efficiently using data from multiple detectors simultaneously.

4.1 Spectral Scaling

4.1.1. Linearities

A common class of problems involves making predictions about future measurements using simulated data. The observables which are predicted might be the number of fluorescence photons or the background level. This section shows how that task can be performed in a Bayesian way by exploiting the linearities of the relevant physics.

As discussed in section 2.3 all processes we generally consider are single photon or linear processes. That is, we can study the physics of each initial photon independently and without considering the total number of photons of the flux or flux density. For the simulation this is axiomatically true as GEANT4 simulates only one particle at a time ([27]). This suggests that we can model the expected number of photons of a given type (counts in a bin, of a class …) λ_X as a linear function of the total number of incoming photons N_0. As a corollary the same is true for the measurement time given a constant flux etc.

$$\lambda_X = c_X N_0$$

A similar argument can be made regarding the number of fluorescence photons as a function of the tracer concentration. We make the assumption that the stopping power of the material is primarily coming from components that are independent of the tracer so $\kappa_{\text{Total}} \cdot \rho_{\text{Total}} = \text{const.}$. This assumption is valid for low concentrations of tracer compared to the total density of the material. In the medical XFI context we typically deal with tracer concentrations $\mathcal{O}(\mu\text{g}\,\text{cm}^{-3})$ ([12]) and material densities $\mathcal{O}(\text{g}\,\text{cm}^{-3})$ (density of water or tissue) so the approximation generally holds. A validation of this linearity approximation can be seen figure 4.1, with the $\chi^2_{\text{red.}}$ of the linear approximation for all data collected here being $\chi^2_{\text{red.}} \approx 1.009$. This means that the linear fit is about as good as one would expect given the inherent noise in the data, which can be interpreted as a validation of the closed form posterior.

Using the Lambert-Beer law (subsection 2.3.3) we see that the number of photons that can cause fluorescence at a given depth scales linearly with the number of incoming photons N_0 and under the assumption of low tracer concentrations is independent of the tracer concentration. Using the definition of the fluorescence mass attenuation coefficient κ_{Fluo} we can now write the differential equation for the production of fluorescence photons

$$dN_{\text{Fluo}} = N_0 \cdot \kappa_{\text{Fluo}} \cdot \rho_{\text{Fluo}} \cdot c(l) dl$$

where $c(l)$ is some function arising from the exponential Lambert-Beer decay, but is independent of ρ_{Fluo}. Integrating immediately shows that

$$N_{\text{Fluo}} \propto N_0 \cdot \kappa_{\text{Fluo}} \cdot \rho_{\text{Fluo}}$$

So given two fluorescence spectra arising from different (low) concentrations we also expect them to be scaled by the ratio of their concentrations.

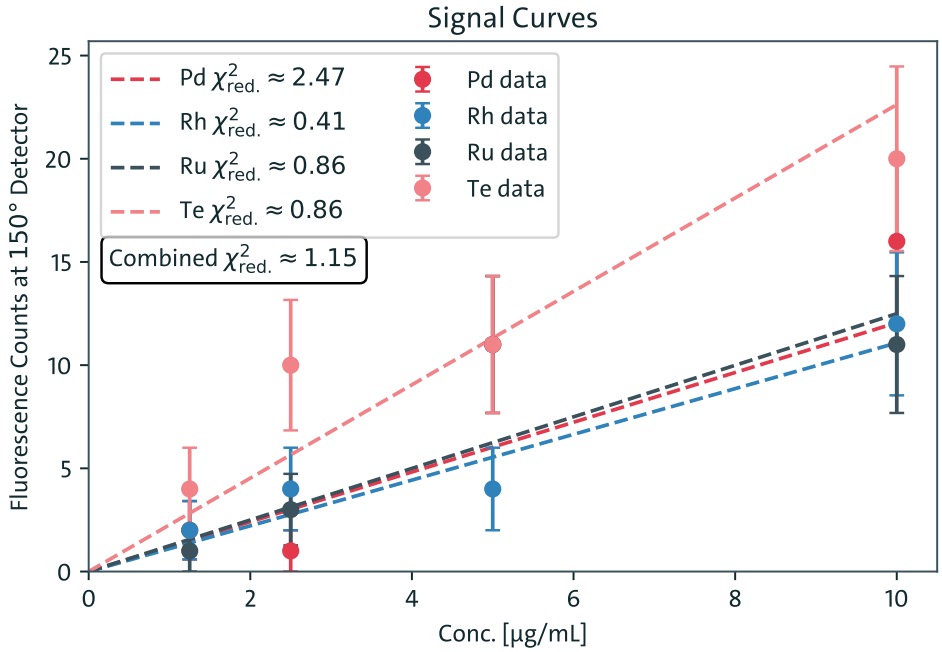

Figure 4.1. Fluorescence - Concentration Linearity: At low concentrations the fluorescence counts scale linearly with tracer concentration. Combining the χ^2 over all ten detectors yields $\chi^2 \approx 1.009$, validating the linearity.

4.1.2. Simulations and Predictions

Consider a simulation using $1.0\,\text{mg}\,\text{cm}^{-3}$ tracer concentration and 10^{10} initial photons with the goal to predict the spectrum at $0.1\,\text{mg}\,\text{cm}^{-3}$ tracer and $5 \cdot 10^8$ photons. Simply scaling the fluorescence spectrum down by a factor of $10 \cdot 20$ and the background spectrum by 20 is not sufficient however, as this point prediction would not capture the inherent uncertainties. One can perform a fully Bayesian calculation to alleviate that concern.

We use $\hat{\cdot}$ to mark the variables for the case we try to predict for and normal symbols to indicate the available data. The model we use is:

$$\begin{aligned} N_{\text{Fluo}} &= \alpha_{\text{Fluo}} \cdot N_0 \cdot \rho_{\text{Fluo}} \\ \hat{N}_{\text{Fluo}} &= \alpha_{\text{Fluo}} \cdot \hat{N}_0 \cdot \rho_{\hat{\text{Fluo}}} \\ N_{\text{Total}} &= \alpha_{\text{Total}} \cdot N_0 \\ \hat{N}_{\text{Total}} &= \alpha_{\text{Total}} \cdot \hat{N}_0 \end{aligned}$$

where α are two constants that are the same between the two measurements. They capture almost the entire physics and geometry of the measurement setup used in the modeled measurement. We then use the data from the simulation to try and constrain their values.

The Bayesian calculation for constraining α_{Total} using N_{Total} is:

$$\begin{aligned} P(\alpha_{\text{Total}}|N_{\text{Total}}) &= \frac{P(N_{\text{Total}}|\alpha_{\text{Total}})P(\alpha_{\text{Total}})}{P(N_{\text{Total}})} = \frac{P(N_{\text{Total}}|\alpha_{\text{Total}})}{\int P(N_{\text{Total}}|\alpha_{\text{Total}})d\alpha_{\text{Total}}} \\ &= \frac{\text{Poiss}\left[\, N_0 \cdot \alpha_{\text{Total}} \,\right](N_{\text{Total}})}{\int \text{Poiss}\left[\, N_0 \cdot \alpha_{\text{Total}} \,\right](N_{\text{Total}})d\alpha_{\text{Total}}} = \frac{(N_0 \cdot \alpha_{\text{Total}})^{N_{\text{Total}}} \cdot e^{-N_0 \cdot \alpha_{\text{Total}}}}{\int (N_0\alpha_{\text{Total}})^{N_{\text{Total}}} e^{-N_0 \cdot \alpha_{\text{Total}}} d\alpha_{\text{Total}}} \\ \Rightarrow P(\alpha_{\text{Total}}|N_{\text{Total}}) &= \Gamma\left[\, N_{\text{Total}} + 1, N_0 \,\right](\alpha_{\text{Total}}) \end{aligned} \tag{4.1}$$

Where, in order, Bayes theorem, the uninformative prior $P(\alpha_{\text{Total}}) = \text{const.}$ and the definition of the Gamma probability density function (PDF) was used. The integral is possible to compute analytically, but one can also look at the scaling of the expression with α_{Total} and see that it follows the same behavior as the Gamma PDF, thus, from the normalization requirement it then follows that both must be identical. The final result then gives a closed form for the posterior distribution of α_{Total}, which captures both the expected behavior and uncertainty from Poisson noise in the simulation.

The more general statement would be that

$$d \sim \text{Poiss}\left[\, c\lambda \,\right] \Rightarrow P(\lambda|d) = \Gamma\left[\, d + 1, c \,\right](\lambda) \tag{4.2}$$

given an uninformative prior for λ.

Applying the same process to α_{Fluo} yields

$$P(\alpha_{\text{Fluo}}|N_{\text{Fluo}}) = \Gamma\left[\, N_{\text{Fluo}} + 1, N_0 \cdot \rho_{\text{Fluo}} \,\right](\alpha_{\text{Fluo}})$$

meaning that the posteriors for both model parameters are now known.

This result is helpful to the experimenter, because not only can the entirety of the relevant physics be condensed into a single value (or one value per bin, spectrum etc.), but we can also analytically give the posterior for this value in a closed form. And we can find this value with very good accuracy before doing any measurements because we can capture the physics dominating the measurement process in Monte Carlo simulations ([27]) with arbitrarily large photon numbers (only constrained by available compute capacity) and higher concentrations (only constrained by demanding linearity, compare figure 4.6).

The next step is using the $\hat{}$-model to make a prediction for a different measurement. To do this we write down the probability for a number $\hat{N}_{\text{Total}}$ given that α_{Total} takes a specific value (which is just the Poisson probability mass function), and then marginalize over the α_{Total}s.

$$\begin{aligned} P(\hat{N}_{\text{Total}}|N_{\text{Total}}) &= \int P(\hat{N}_{\text{Total}}|\alpha_{\text{Total}})P(\alpha_{\text{Total}}|N_{\text{Total}})d\alpha_{\text{Total}} \\ &= \int \frac{e^{-\hat{N}_0 \cdot \alpha_{\text{Total}}}(\hat{N}_0 \cdot \alpha_{\text{Total}})^{\hat{N}_{\text{Total}}}}{\hat{N}_{\text{Total}}!} \frac{e^{-N_0 \cdot \alpha_{\text{Total}}}(N_0 \cdot \alpha_{\text{Total}})^{N_{\text{Total}}}}{N_{\text{Total}}!} N_0 d\alpha_{\text{Total}} \\ &= \frac{\hat{N}_0^{\hat{N}_{\text{Total}}} {N_0}^{N_{\text{Total}}+1}}{\hat{N}_{\text{Total}}! N_{\text{Total}}!} \frac{(\hat{N}_{\text{Total}} + N_{\text{Total}})!}{(N_0 + \hat{N}_0)^{(\hat{N}_{\text{Total}}+N_{\text{Total}}+1)}} \\ &= \text{NegBin}\left[\, N_t + 1, \frac{N_0}{N_0 + \hat{N}_0} \,\right](\hat{N}_{\text{Total}}) \\ &= \text{NegBin}\left[\, N_t + 1, \frac{1}{1+q} \,\right](\hat{N}_{\text{Total}}) \;,\; q := \frac{\hat{N}_0}{N_0} \end{aligned}$$

And analogously

$$P(\hat{N}_{\text{Fluo}}|N_{\text{Fluo}}) = \text{NegBin}\left[N_{\text{Fluo}} + 1, \frac{1}{1+q} \right](\hat{N}_{\text{Fluo}}) \ , \ q := \frac{\hat{\rho}_{\text{Fluo}}\hat{N}_0}{\rho_{\text{Fluo}}N_0} \tag{4.3}$$

where $\text{NegBin}[r, p]$ is the negative binomial distribution. This is the posterior predictive for the count-observables.

The behavior of these distributions is as one might expect. Firstly, if N_0 is very large, the variance of the posterior for α becomes small, i.e. with lots of data we can characterize the physics very well. Secondly, for the scaling when q is small (that we are basing the prediction on much more data than we want to predict for), the variance approaches a simple Poisson variance and for very large q (i.e. large extrapolation) the variance explodes - large extrapolation comes with great uncertainty.

4.1.3. Application: Detection Limits

Samples from the posterior predictive can be interpreted as plausible results of future measurements, because they represent taking a setup with α values consistent with the simulation and performing a measurement with Poisson noise on that. So they can be used to test any relevant follow-on processing.

Suppose we were planning a measurement which will use $\hat{N}_0$ incident photons and we wanted whether this number was sufficient such that a given post processing tool will find a significant result. We can simulate the measurement setup with N_0 incident photons. Ideally we would choose $N_0 \gg \hat{N}_0$, but we might be limited by computational resources. The simulation finds a spectrum with N_i counts in the i-th energy bin. We can then write the posterior for the spectrum as $\hat{N}_i \sim \text{NegBin}\left[N_i, \frac{N_0}{N_0+\hat{N}_0} \right]$. Next we draw an ensemble of samples from that posterior predictive. The spectral processing can then be applied to the samples individually. This will be modeled as a general function $f(\hat{N}_i)$. By looking at the values this function takes over a large number of posterior predictive samples, we can study what we expect the post processing to find.

For many plausible significance functions f the asymptotic behavior will be approximately proportional to the concentration and square root of the incident photons: $f \propto \sqrt{N_0}c$ in the low signal regime because here the error is approximately independent of c and scales with $\sqrt{N}$ while the signal scales linearly with both. If we use a significance function and a threshold we can thus approximate the value locally by that scaling and then run a search using the Newton method [77] for the concentration or incident photon count which satisfy the distribution of f we are looking for.

4.1.3.1. Example

In this example uses the scaling method described above to find the detection limit for the tracer concentration in a particular setup. The setup studied here uses iodine as the tracer in an aqueous solution in a 350 µL cuvette, 10^{10} incident photons at 53 keV, ten detectors placed in 30° steps between ±30° and ±150° at 6 cm distance (compare figure 2.4). As the spectral processing function of interest, the measurement significance defined in [21] was used. For the purpose of this section it is sufficient to treat the measurement significance as a black box function f of the spectrum.

To tackle this problem a typical measurement setup using those parameters was simulated for $0.1\,\text{mg}\,\text{mL}^{-1}$ iodine solution and 10^{11} incident photons. The spectrum is subdivided into the signal component (iodine fluorescence) and the background component (everything else). These spectra are shown in figure 4.2.

The goal is to find the detection limit, defined as the lowest tracer concentration such that 99% confidence detection significance is 95% likely to occur, i.e. the detection limit is the concentration were 95% of samples of T produce an $f(T) \geq 99\%$.

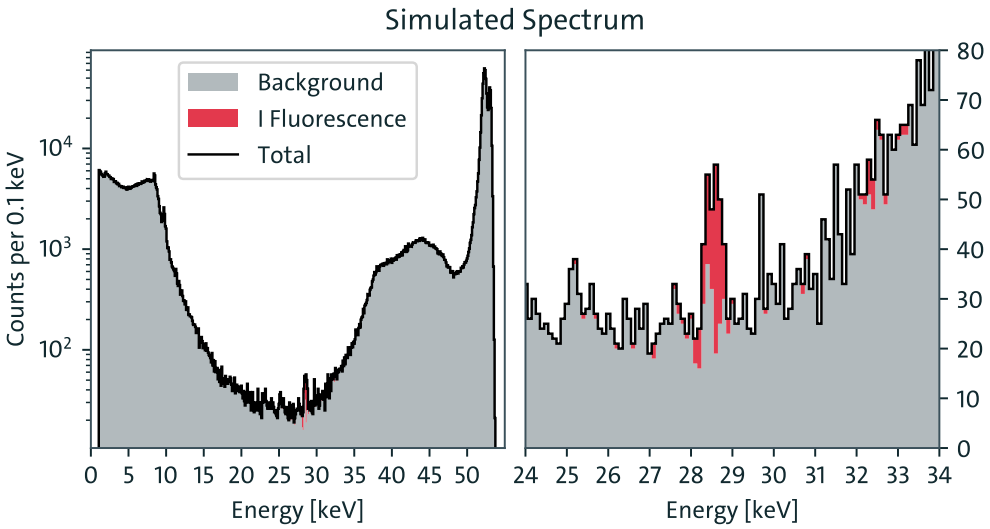

Figure 4.2. Example of a Simulation Output: Event histories can be tracked in detail during a simulation, allowing for the division of events and thus spectrum entries into a signal component S (the iodine fluorescence) and a background component B. In a real experiment only the sum of these two subspectra (Total) would be available.

The total spectrum is composed of the signal spectrum S and the background spectrum B with $T = S + B$ where the background scales linearly with the incident photons and the signal with incident photons and concentration.

We can now use the result above to produce plausible signal and background spectra for the measurement with the same parameters, but at various concentrations. It is highly computationally efficient that this can be done by just sampling from a known distribution rather than running a full Monte Carlo simulation for each number of incident photons and concentration, because the former is constant time in the number of incident photons while the latter scales linearly with it. After samples of S and B from the posterior predictives are generated, samples of T can be generated by adding them together. These can now be passed to the post processing giving samples $f(T)$. For each concentration we can now look at the distribution of $f(T)$, where the variance in this distribution represents being more or less lucky with the noise. We can choose a threshold for $f(T)$ to call a detection, due to the variance inherent in the process, at each concentration some fraction of samples will be above the threshold and some below. So we need to choose a different threshold for how likely we require a detection to be.

Rather than a detection probability function we use the significance produced by fitting the data with a linear model where $f(T) = \hat{N}_f / \Delta\hat{N}_f$, where $\hat{N}_f$ is the number of fluorescence photons estimated by the fit and $\Delta\hat{N}_f$ is the estimated error of that value. Firstly, this significance function is evaluated for all generated samples of T, then their distribution is approximated by Gaussian function by calculating the sample mean and standard deviation. We hope to find that $\mu_f(c, N_0) - 2\sigma_f(c, N_0) = 3$ but for the generated samples at $c = 10\,\mu\mathrm{g\,mL^{-1}}$ and $N_0 = 10^{10}$ we find $\mu_f(10\,\mu\mathrm{g\,mL^{-1}}, 10^{10}) = 2.0$ and $\sigma_f(10\,\mu\mathrm{g\,mL^{-1}}, 10^{10}) = 0.8$. Now we use the approximate scaling of these values to estimate the critical concentration to be $c = 90\,\mu\mathrm{g\,mL^{-1}}$ which yields 9.2 ± 0.7, iterating the process again yields $35\,\mu\mathrm{g\,mL^{-1}}$ with 5.1 ± 0.7 and finally $29\,\mu\mathrm{g\,mL^{-1}}$ with $Z = 4.5 \pm 0.8$. The final significance distribution can be seen in figure 4.3, where the process successfully yielded the desired threshold.

Practically this means that we could use a simulation with $N_0 = 10^{11}$ and $c = 10\,\mu\mathrm{g\,mL^{-1}}$ to produce the prediction that we are 95% confident to find at least 3σ significance if we measured $29\,\mu\mathrm{g\,mL^{-1}}$ in the same setup.

As a comparison 49 simulations of this scenario with $29\,\mu\mathrm{g\,mL^{-1}}$ concentration and 10^{10} photons were performed. A priori we expect that the distribution of the fit significance

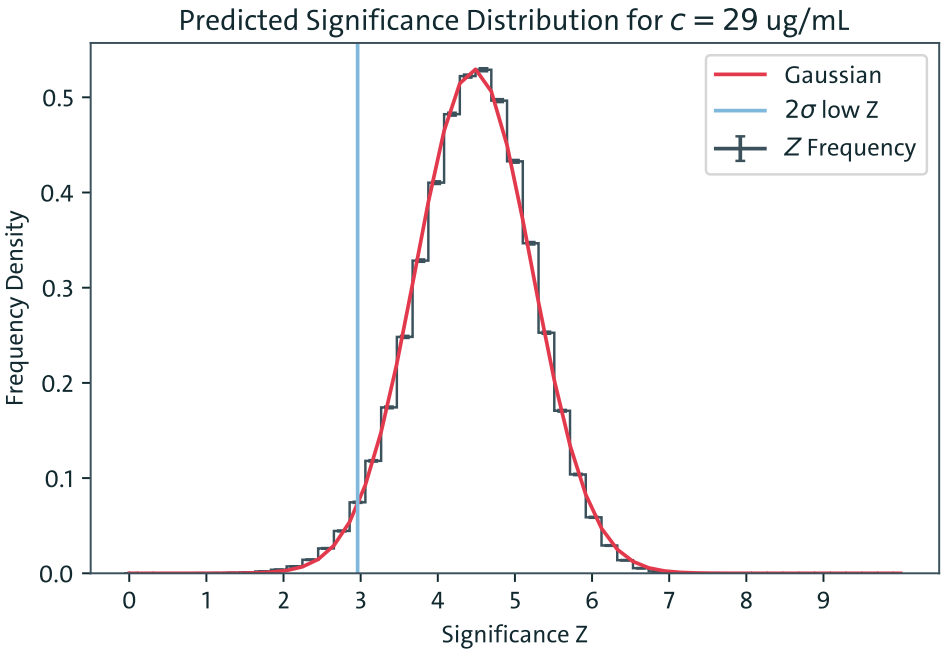

Figure 4.3. Significance Distribution: The expected significance distribution for future measurements was numerically estimated. The distribution is approximately Gaussian. The concentration of $c = 29\,\mu g\,mL^{-1}$ was found to produce $Z \geq 3$ in 95% of cases.

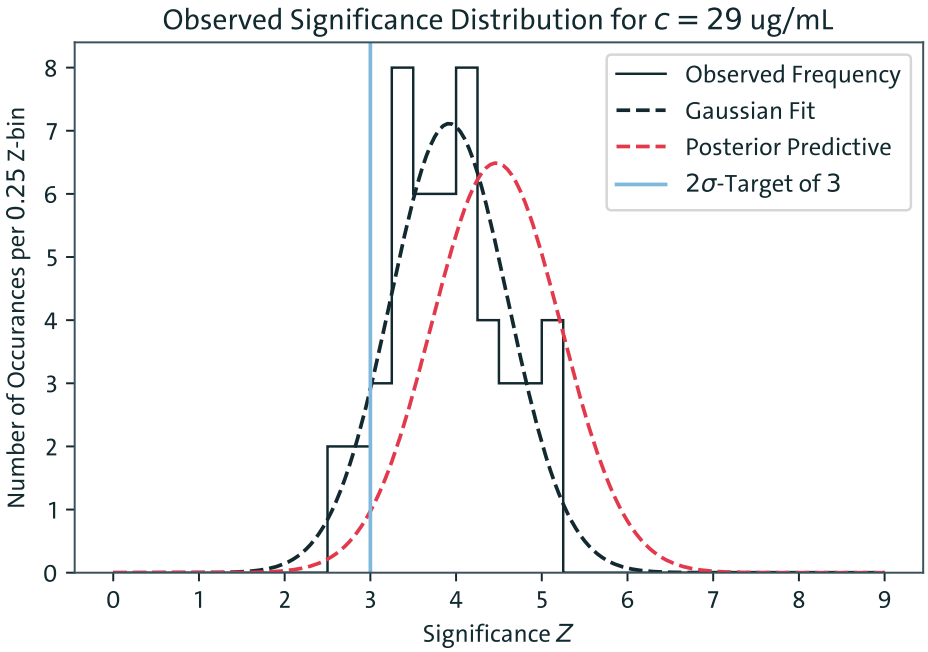

Figure 4.4. Observed Significance Distribution: The experiment was carried out 49 times in Monte Carlo simulations. The observed distribution of Z scores are compared to the posterior predictive for them. The posterior predictive has a slightly larger standard deviation (as it captures additional uncertainty) and a slight shift (because the sample used to generate it had a slightly higher Z than the population mean). The attempt to find a concentration for which most experiments produce $Z \geq 3$ was successful.

is a bit narrower than that produced from the single simulation at 10^{11} photons. This is because the distribution of actual simulation only has the Poisson noise of that run, while the method using the posterior predictive also has to contend with the uncertainty from the individual run it was generated from. This added uncertainty expresses that we expect the actual distribution to be slightly shifted compared to the posterior predictive by an amount that depends on how close the sample we got for the 10^{11} photon run was to the actual expectation value. As discussed before we expect 95% of simulations to produce $Z \geq 3$ (and thus $f \geq 99\%$). Whether a simulation produces a significance above the threshold is a binary outcome, and the outcomes of each simulation have identical probability, but are independent of each other. So the number of outcomes above the threshold should be predicted as a binomial distribution. Evaluating gives us the prediction that 47 ± 2 out of the 49 simulations should find $Z \geq 3$.

The histogram over the 49 simulations (seen in figure 4.2) can be interpreted as a frequentist (rather than Bayesian) alternative to the method outlined in this section. Out of the 49 simulations 45 produced $Z \geq 3$ and thus $f \geq 99\%$, which aligns very well with the prediction from the posterior predictive. Interpreting it as a prediction of the probability of a measurement to produce a significant result, which would yield $92\% \pm 14\%$, compared to the $95\% \pm 3\%$ from the Bayesian approach. The frequentist approach used five times the compute resources (because five times as many photons were simulated) and produces a noisier result (greater uncertainty in the predicted rate of significance). And as expected the distribution is a bit narrower and slightly shifted. Additionally, the two Gaussian fits overlap with a majority of their mass, indicating that the two methods produce results that are compatible.

4.2 Spectral Curve Fitting

As discussed in chapter 2, a key task in XFI is to separate the signal and background components of the spectrum, and a standard approach of doing this is to perform a curve fit of the spectrum in the signal region. The following section studies this approach, focussing especially on Bayesian correction terms and uncertainty quantification.

4.2.1. Equal Weight Gaussian Least Squares

The simplest method to fit a linear model to data is least squares fitting [57]. To describe it we can understand the model that is being fit as a linear transformation $\mathbf{M}$ mapping a parameter vector $\vec{p}$ onto a corresponding prediction for a vector of observables $\vec{o}$ while the data $\vec{d}$ is understood as a realization of a random variable parameterized by the observables. Because the model is a linear transformation, $\mathbf{M}$ can be represented as a matrix, or more generally, a tensor of rank ≥ 2 and at least one observable and one parameter index. In the simplest possible case $\vec{d} \sim \mathcal{N}\left[\vec{o}, \sigma^2 \cdot \mathbb{1}\right]$ and $P(\vec{o}) \propto$ const.. This allows us to write

$$-\log p(\vec{o}|\vec{d}) = \text{const.} + \frac{1}{2\sigma^2}(\vec{d}-\vec{o})^2 \tag{4.4}$$

which suggests that the maximum likelihood estimator for $\vec{p}$ is the one that minimizes $(\vec{d}-\vec{o})^2$. The minimum can be found through differentiation

$$\begin{aligned}\vec{0} &\stackrel{!}{=} \nabla_{\vec{p}}(\vec{d}-\vec{o})^T(\vec{d}-\vec{o}) = \nabla_{\vec{p}}(\vec{d}-\mathbf{M}\vec{p})^T(\vec{d}-\mathbf{M}\vec{p}) = 2\mathbf{M}^T(\mathbf{M}\vec{p}-\vec{d}) \\ \Leftrightarrow &\mathbf{M}^T\mathbf{M}\vec{p} = \mathbf{M}^T\vec{d} \Leftrightarrow \vec{p} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\vec{d} =: \mathbf{M}^+\vec{d}\end{aligned}$$

where the last step introduced the Moore-Penrose pseudo inverse $\mathbf{M}^+ := (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$ [57]. This calculation is entirely analogous for the relevant tensors as the least squares problem is a natural full contraction.

Any matrix can be represented as the decomposition $\mathbf{M} = \mathbf{USV}^T$ where $\mathbf{U}, \mathbf{V}$ are square, orthogonal matrices, i.e. $\mathbf{U}^T\mathbf{U} = \mathbb{1}$ and $\mathbf{V}^T\mathbf{V} = \mathbb{1}$ and $\mathbf{S} = \delta_{i,j}s_i$ is a diagonal, non-negative matrix, but in general not square [58]. It is clear that the diagonal elements of $\mathbf{S}$ are also the non-negative square roots of the eigenvalues of $\mathbf{M}^T\mathbf{M}$ because both $\mathbf{U}$ and $\mathbf{V}$ are unitary. Calculating $\mathbf{M}^+$ with this representation yields

$$\mathbf{M}^+ = \mathbf{V}(\mathbf{SS}^T)^{-1}\mathbf{S}^T\mathbf{U}^T = \mathbf{VS}^+\mathbf{U}^T$$

where for diagonal matrices $\mathbf{S}^+_{i,j} := \delta_{i,j}(1/s_i)$ was used. This can be generalized to matrices with eigenvalues of 0 by setting $\mathbf{S}^+_{i,j}$ to 0 as well [58]. A 0 eigenvalue indicates that the least squares solution is not unique, spanning a space of dimension equal to the number of 0 eigenvalues, i.e. the dimension of the kernel of $\mathbf{M}$.

This pseudo inverse turns the problem of finding the maximum likelihood estimator $\hat{\vec{p}} = \mathbf{M}^+\vec{d}$ of the parameter vector $\vec{p}$ into a simple linear function of the data $\vec{d}$. In the derivation of the pseudo inverse we assumed that $\vec{d}$ was normal with a constant and diagonal covariance. However, if we strictly understand $\hat{\vec{p}}$ as a function of the data, we can calculate how sensitive it is to noise in the data irrespective of whether the noise on the data satisfies that assumption. In particular

$$\mathrm{Cov}\left[\hat{\vec{p}}\right] = \mathbf{M}^+\mathrm{Cov}\left[\vec{d}\right](\mathbf{M}^+)^T$$

And if $\mathrm{Cov}\left[\vec{d}\right]$ is diagonal as well (as is the case for Poisson noise for example) we further find

$$\mathrm{diag}\left(\mathrm{Cov}\left[\hat{\vec{p}}\right]\right) = (\mathbf{M}^+)^{\circ 2}\,\mathrm{diag}\left(\mathrm{Cov}\left[\vec{d}\right]\right)$$

where $(\mathbf{M}^+)^{\circ 2}$ indicates the pointwise square of the pseudo inverse of $\mathbf{M}$. This means that the variance of the individual components of the estimate $\hat{\vec{p}}$ are also accessible as a simple linear transformation of $\vec{d}$. This can also serve as an uncertainty estimate for the stochastic error.

Notably, $\mathbf{M}^+$ and $(\mathbf{M}^+)^{\circ 2}$ only depend on the choice of model, meaning that if we want to perform many fits with the same model the expensive calculation of $\mathbf{M}^+$ only needs to be done once leaving only the matrix-vector multiplication for each individual fit.

In summary this method provides a simple and computationally efficient way to fit a linear model to data including an uncertainty estimate.

4.2.2. Weighted Least Squares

One generalization of the problem that leads to equation (4.4) is to study an arbitrary covariance matrix for $\vec{d}$, which analogously yields [57]

$$\hat{\vec{p}} = (\mathbf{M}^T\Sigma^{-1}\mathbf{M})^{-1}\mathbf{M}^T\Sigma^{-1}\vec{d} \tag{4.5}$$

In the case that $\Sigma = \mathrm{diag}(\vec{\sigma}^{\circ 2})$ is diagonal we can introduce $\widetilde{\mathbf{M}} = \mathrm{diag}(\vec{\sigma}^{\circ -1})\mathbf{M}$ and $\tilde{\vec{d}} = \mathrm{diag}(\vec{\sigma}^{\circ -1})\vec{d}$ which simplifies this to

$$\hat{\vec{p}} = \widetilde{\mathbf{M}}^+\tilde{\vec{d}} \tag{4.6}$$

However, $\mathrm{diag}(\vec{\sigma}^{\circ -1})$ may vary from fit to fit. This means that a large part of the computational benefit for large data sets is lost as the inversion process needs to be repeated for each fit individually. Computing the covariance of the prediction yields

$$\mathrm{Cov}\left[\hat{\vec{p}}\right] = \widetilde{\mathbf{M}}^+\left(\widetilde{\mathbf{M}}^+\right)^T \tag{4.7}$$

The process of weighting the model matrix and data is sometimes called whitening because it transforms the data to white Gaussian noise [57]. However, we are still working with the

assumption of the data components being independently and normally distributed. Alternatively to the normality constraint this can be interpreted as a second order estimate using the first two moments of the data.

4.2.3. Poisson Correction

If we try to use the same method to find an exact MLE for Poisson data we are presented with

$$\mathbf{M}^T \left(\vec{d} \oslash \mathbf{M}\hat{\vec{p}} - \vec{\mathbb{1}} \right) = \vec{\mathbb{0}}$$

where element-wise division $\oslash$ is used and $\vec{\mathbb{1}}$ is the one vector that shares a size with $\vec{d}$. This equation does not have a closed form solution, it would have be to solved via the Newton method or similar processes, especially considering that there is also the implicit domain restriction that $\mathbf{M}\hat{\vec{p}} \geq 0$.

There are several methods to proceed from here, trying to get an estimate nonetheless. The easiest would be to use the results from above directly (with $\vec{\sigma} = \sqrt{\vec{d}}$ for the weighted case) which is equivalent to approximating the Poisson distribution of the data by a Gaussian. This is quite reasonable especially if the typical values in $\vec{d}$ are large.

Another approach is to start with $\mathbf{M}^+\vec{d}$ and then perform one step of the Newton method to get a better estimate. Using $\mathbf{R} := \mathbb{1} - \mathbf{M}\mathbf{M}^+$ (i.e. the residue of the projection on the space represented by the model) for compactness, this becomes

$$\hat{\vec{p}} = \mathbf{M}^+ \left(\mathbb{1} + \operatorname{diag}\left(\left(\mathbf{M}\mathbf{M}^+\vec{d} \right) \oslash \vec{d} \right) \mathbf{R} \right) \vec{d} = \left(\mathbf{M}^+ + \mathbf{C} \right) \vec{d}$$

making it quite explicit that the size of this first correction depends on $\mathbf{R}$ - how well $\mathbf{M}^+$ inverts $\mathbf{M}$ in general and $(\mathbf{M}\mathbf{M}^+\vec{d}) \oslash \vec{d}$ - how well the model represents $\vec{d}$ in relative terms. One interesting fact is that if $(\mathbf{M}\mathbf{M}^+\vec{d}) \oslash \vec{d} \approx \vec{\mathbb{1}}$ this becomes

$$\mathbf{C} \approx \mathbf{M}^+\mathbf{R} = \mathbf{M}^+ - \mathbf{M}^+\mathbf{M}\mathbf{M}^+ = \mathbb{0}$$

This provides a good test for whether a model is suited to fit a given Poisson dataset with the previously outlined methods. Another property of $\mathbf{C}$ is that finding it for any given data point requires no pseudo inversion or decomposition, only linear or point-wise operations with pre-computable objects. The downside is that as $\hat{\vec{p}}$ now depends on $\vec{d}$ in non-linear ways, estimating the uncertainty of $\hat{\vec{p}}$ in closed form becomes much harder. Another issue arises when the number of counts in the data includes low numbers, and especially 0. One can see that the correction may become singular and $\mathbf{M}\mathbf{M}^+\vec{d}$ may even become negative. These issues arise with all three linear fit models outlined so far in these cases.

4.2.4. Posterior Projection

One way to alleviate this issue is to reframe the problem in a Bayesian sense rather than trying to find the $\hat{p}$ that maximized the likelihood of seeing the data. By far the simplest method is to first find posteriors for $\vec{\lambda}$ with $\vec{d} \sim \text{Poiss}\left[\vec{\lambda} \right]$ which are treated as independent. This follows the same process as outlined in section 4.1 and gives the closed form $\vec{\lambda} \sim \Gamma\left[\vec{d} + 1, 1 \right]$ if all data points are equally weighted. Now, instead of using $\vec{p}$ to directly fit the data, we instead try to find $\vec{p} \sim \mathcal{N}$ such that $\mathbf{M}\vec{p} \approx \vec{\lambda}$ in a distributional sense (the predicted distribution from the fit model approximating the point-wise posterior).

Scaling factors, like detector dead-time (the proportion of time that a detector is not ready to receive events) or incident photon flux variation, can also be incorporated using this approach (without using concepts like "fractional counts" as one often does with the previous methods). The resulting posterior is $\vec{\lambda} \sim \Gamma\left[\vec{d} + 1, \vec{\beta} \right]$ where $\vec{\beta}$ are the scaling factors. Naively one might think that just by using different scaling factors, the uncertainty of the posterior can be arbitrarily suppressed, as $\text{Var}\left[\vec{\lambda} \right] = \operatorname{diag}\left((\vec{d} + 1) \oslash \vec{\beta}^{\circ 2} \right)$, however,

this rescaling does not change the shape of the distribution, i.e. the relative uncertainty, only the scale of $\vec{\lambda}$ changes, which is similar to stating $\vec{\lambda}$ in units of s^{-1}.

If we want $\mathbf{M}\vec{p} \approx \vec{\lambda}$ in a distributional sense we might start by instead fitting a Gaussian approximation of $\vec{\lambda}$ i.e. $\mathcal{N}\left[\vec{d}+1, \mathrm{diag}\left((\vec{d}+1) \oslash \vec{\beta}^{\circ 2}\right)\right]$. This is solved exactly like discussed at the beginning of this section and produces

$$\vec{p} \sim \mathcal{N}\left[\mathbf{M}^{+}(\vec{d}+1), (\mathbf{M}^{+})^{T}\mathrm{diag}\left((\vec{d}+1) \oslash \vec{\beta}^{\circ 2}\right)\mathbf{M}^{+}\right] \tag{4.8}$$

$$\vec{p} \sim \mathcal{N}\left[\widetilde{\mathbf{M}^{+}}\vec{\beta}, \left(\widetilde{\mathbf{M}^{+}}\right)^{T}\widetilde{\mathbf{M}^{+}}\right] \text{ , } \widetilde{\mathbf{M}^{+}} = \left(\mathrm{diag}\left(\vec{\beta} \oslash \sqrt{(\vec{d}+1)}\right)\mathbf{M}\right)^{+} \tag{4.9}$$

respectively, depending on whether the approximation $\vec{d} \approx$ const. was used. The Gaussian approximation of $\vec{\lambda}$ is reasonable if $\vec{d} \gtrsim 5$ and very good if $\vec{d} \gg 5$ as skewness and excess kurtosis of Γ are $2/\sqrt{\vec{d}+1}$ and $6/(\vec{d}+1)$. On the other hand, the approximation that $\vec{d}$ itself is Gaussian only starts to become good when the discreteness of the Poisson distribution becomes small relative to its mean, i.e. $\vec{d} \gg \sqrt{\vec{d}}$.

The approximation in equation (4.8) shares all the useful properties of the result in subsection 4.2.1 while also alleviating the singularity issues for 0 counts and only requiring that the number of counts in each treated bin is similar (which typically implies the absolute numerical values for the Gaussian shape of Γ already). For a typical spectrum this is the case in any region where the background dominates and a small section of the data can be fitted, for example in the region of heavy metal L-lines.

The method summarized by equation (4.9) shares the same relative tradeoffs to equation (4.8) as described in subsection 4.2.2, but it is more suited to regions where the background is very low as is often seen for K-lines of elements in the fifth period.

4.2.5. Log Linear Models

In the literature it is not uncommon to see models like $\vec{d} \sim \mathrm{Poiss}\left[\exp(\mathbf{M}\vec{p})\right]$. In general this has several useful properties. For one, the domain restriction issue is solved, as exp is positive definite. Additionally the gradient of the likelihood becomes slightly easier to work with, involving an exponential rather than a logarithm which can stabilize numerical methods. Though the MLE still does not have a closed form.

For fitting XFI spectra this is problematic in a different way. A key property of spectra is that they are additive - if both background and signal are present in an energy bin, their counts will add up. So if the goal of fitting the data is discriminating between signal and background, ultimately we will always have to come back to the model $\vec{d} \sim \mathrm{Poiss}\left[\lambda_{\mathrm{Signal}} + \lambda_{\mathrm{Background}}\right]$ even if a different model is used as in intermediate. If an exponential (usually log normal) is used first and $\lambda_{\mathrm{Signal}}$ is treated as a prediction of that model the problems discussed earlier will typically arise again in this last step.

4.2.6. Summary

Fitting linear models to Poisson data is not a trivial task as the log-likelihood has a more complex dependence on the data error than the simple square error of normally distributed data. To get a closed form result one has to use approximations or additional assumptions. Alternatively an iterative approach could be used analogous to subsection 5.2.2. Fitting becomes especially problematic if the model matrix has small eigenvalues, the data includes points with very few counts or the model represents the data poorly. Some of the problems of fitting the data directly can be alleviated by instead fitting the point-wise rate parameter posterior, which in effect is the same as adding 1 to each data point before fitting. This is especially useful for error estimation where projection of the posterior uncertainties into parameter space provides very useful error estimates.

4.3 Subtraction Method

The most important intermediate goal of processing XFI spectra is to separate the background from the fluorescence signal. The most intuitive way to do this might be to simply subtract the background from the signal. If one had access to the actual number of counts in the background this would indeed be a simple task. In practice, this is not possible to obtain. The next best thing is an independent realization of the same background that clouds the measurement containing the signal. This method is used for example in [15]. An alternative would be to use a realization of something different (i.e. another region of the spectrum) with a known relationship to the background, which is used by [16].

The simplest model would be $B, B' \sim \text{Poiss}[\lambda_b]$ being two independent realizations of the background, $S \sim \text{Poiss}[\lambda_s]$ the signal and $T = B + S \sim \text{Poiss}[\lambda_b + \lambda_s] = \text{Poiss}[\lambda_t]$ being the total counts with signal and background. The available data would be T and B' and the model parameters λ_b and λ_s. We can use this model to study the naive approach of estimating $\hat{\lambda_s} = T - B'$:

$$\begin{aligned}
&\hat{\lambda_s} = T - B = S + B - B' \\
\Rightarrow &\mathbb{E}\left[\hat{\lambda_s}\right] = \mathbb{E}[S] + \mathbb{E}[B] - \mathbb{E}[B'] = \lambda_s \\
\Rightarrow &\text{Var}\left[\hat{\lambda_s}\right] = \text{Var}[S] + \text{Var}[B] + \text{Var}[B'] = \lambda_t + \lambda_b \\
\Rightarrow &\widehat{\text{Var}}\left[\hat{\lambda_s}\right] = \hat{\lambda_t} + \hat{\lambda_b} = T + B'
\end{aligned}$$

Thus, this approach produces an unbiased estimator for λ_s and λ_b (which is just B) that in turn also provide an estimate for the variance of $\hat{\lambda_s}$. The estimate of the variance is $\approx T + B$ rather than $T - B$ which would be the estimate of the variance of S itself, rather than a characterization of the quality of the estimator used.

Similar to subsection 4.2.4 the simplest extension to Bayesian terms is to first find the posteriors for λ_t and λ_b and then use them in place of the naive estimators in the same process:

$$\begin{aligned}
&\hat{\lambda_s} := \hat{\lambda_t} - \hat{\lambda_b} = (T + 1) - (B' + 1) = T - B' \\
\Rightarrow &\mathbb{E}\left[\hat{\lambda_s}\right] = \mathbb{E}[T - B'] = \lambda_s \\
\Rightarrow &\widehat{\text{Var}}\left[\hat{\lambda_s}\right] = T + B' + 2
\end{aligned}$$

So again we find an unbiased estimator and an estimation for the uncertainty. If we can use a Gaussian approximation for the Γ posteriors (i.e. we have $T, B \gtrsim 6$), we can estimate the posterior for λ_s as the Gaussian

$$P(\lambda_s | T, B) \approx \mathcal{N}[T - B, T + B + 2](\lambda_s) \tag{4.10}$$

The $+2$ comes from the $\Gamma[1, 0]$ prior being used for both λ_t and λ_b. In the mean the offsets of $+1$ cancel due to subtraction, but in the variance they both contribute.

The most obvious advantage compared to the naive option is that the variance estimate does not become singular when no events are measured at all. This is desirable, even if we have seen neither background nor any other photons in a given measurement and we cannot conclude that it is completely impossible that $\lambda_s > 0$ which the naive estimator would claim.

4.4 Tracer Quantification

The previous sections discussed several methods for estimating the number of observed signal (fluorescence) photons S or their rate parameter λ_s with $F \sim \text{Poiss}[\lambda_s]$. When tracer mass in the beam path is low enough that self-absorption is minimal compared to competing

effects and the flux is low enough that no saturation occurs, one can see that $\lambda_s \sim m$ where m is the tracer mass. More specifically

$$\lambda_s = I_0 T_{\text{in}} \left(\frac{\mu}{\rho}\right)_F m T_{\text{out}} \frac{A_{\text{eff}}}{4\pi r^2}$$

where I_0 is the incident flux in photons per area, $\left(\frac{\mu}{\rho}\right)_F$ is the linear mass attenuation coefficient for the fluorescence line of interest, T is the transmission into and out of the object, A_{eff} is the effective detector area and r is the distance to the detector. [21]

Because λ_s and m or c have a linear relationship it is mathematically straight forward to convert an estimate of one to the other if the constants are known. It is generally difficult to determine every term precisely, but they can usually be estimated from the measurement parameters. Using error estimates for the individual parameters of that measurement model allows for error propagation to the final result.

4.5 Using Multiple Detectors

A central property of Bayesian analysis is that it provides a clear and consistent framework for incorporating multiple pieces of evidence or data. The simplest way to see this is by looking at the logarithmic form of Bayes theorem with an inference target x and two pieces of data d_1 and d_2

$$\log p(x|d_1, d_2) = \log p(d_1, d_2|x) + \log p(x) - \log p(d_1, d_2)$$

In most physically relevant cases $d_1|x$ and $d_2|x$ are independent, meaning they only correlate in so far as they depend on the same parameter. Another way of phrasing this is that while the signal depends on the same thing the noise component is independent. This implies

$$\begin{aligned} \log p(d_1, d_2|x) &= \log p(d_1|x) + \log p(d_2|x) \\ \Rightarrow \log p(x|d_1, d_2) &= \log p(d_1|x) + \log p(d_2|x) + \log p(x) - \log p(d_1, d_2) \end{aligned}$$

Again, $\log p(d_1, d_2)$ is constant in x and we also have a normalization constraint so it suffices to just write

$$\log p(x|d_1, d_2) = \log p(d_1|x) + \log p(d_2|x) + \log p(x) + \text{const.} \tag{4.11}$$

Whether the update is performed all at once or one after another does not matter.

Another way to interpret this is via Bayesian updating, instead of considering two pieces of simultaneous evidence we first update our belief (the probability distribution of x) according to one piece, then the other piece of data i.e. using the posterior of the first measurement as the prior for the second:

$$\begin{aligned} \log p(x|d_1, d_2) &= \log p(d_2|x, d_1) + \log p(x|d_1) + \log p(d_2|d_1) \\ &= \log p(d_2|x, d_1) + \log p(d_1|x) + \log p(x) + \log p(d_1) + \log p(d_2|d_1) \\ &= \log p(d_2|x, d_1) + \log p(d_1|x) + \log p(x) + \log p(d_1, d_2) \end{aligned}$$

With the assumption of conditional independence of d_1 and d_2 equation (4.11) can be recovered.

4.5.1. Linear Modeling

In XFI measurements we can typically use multiple detectors (as shown in [21]). Rather than doing the data analysis on the data collected from each detector individually, as [21] does, the Bayesian frame allows us to combine the data effectively. One simple framework to accomplish this, is to model the signal component as being linearly related (for all detectors it is proportional to the tracer mass) and the background being independent. At first we

will look at the problem without the background to formulate an approach. Writing this model in distributional form yields

$$d_i \sim \text{Poiss}\left[\, \epsilon_i x \,\right]$$

where i indexes the detectors, d is the data, x the inference target that the data shares and ϵ_i a constant which varies from detector to detector.

4.5.1.1. A Priori ϵ

In some cases ϵ_i can be approximated a priori because it only contains geometrical information. This is equivalent to saying that all contributions to the signal which cannot be estimated like that are shared between all detectors and thus can be offloaded into x. An example for this would be attempting to estimate the beam flux using a priori guesses for the product of the detector solid angle coverage, detector efficiency and fluorescence cross section using a known reference target.

Using the result from equation (4.2) and equation (4.11) we find that

$$P\left(x|\vec{d}\right) = \Gamma\left[\, 1 + \sum_i d_i, \sum_i \epsilon_i \,\right](x)$$

where $\vec{d}$ was used to represent all d_i. Of course it is still useful to see whether the individual measurements are compatible with each other as this closed form posterior follows from a fairly strong modeling assumption, especially with the a priori ϵ_i.

4.5.1.2. Reference Measurement

If we have access to two measurements, one reference measurement with a known relation between the Poisson rate and the detector constant and one measurement used for inference we could model it as

$$\vec{d_r} \sim \text{Poiss}\left[\, \vec{\epsilon r} \,\right] \;,\; \vec{d_m} \sim \text{Poiss}\left[\, \vec{\epsilon} x \,\right]$$

In general we might only be interested in x, so we have to marginalize the joint posterior $p(x, \vec{\epsilon}|\vec{d_r}, \vec{d_m})$ over $\vec{\epsilon}$

$$\begin{aligned}
p(x|\vec{d_r}, \vec{d_m}) &= \int p(x, \vec{\epsilon}|\vec{d_r}, \vec{d_m}) d\vec{\epsilon} \\
&= \int p(\vec{d_r}, \vec{d_m}|x, \vec{\epsilon}) p(x, \vec{\epsilon}) / p(\vec{d_r}, \vec{d_m}) d\vec{\epsilon} \\
&\propto \int p(\vec{d_m}|x, \vec{\epsilon}) p(x) p(\vec{d_r}|\vec{\epsilon}) d\vec{\epsilon} \\
&\propto \int p(x|\vec{d_m}, \vec{\epsilon}) p(\vec{\epsilon}|\vec{d_r}) d\vec{\epsilon} = \mathop{\mathbb{E}}_{\vec{\epsilon}|\vec{d_r}} \left[\, p(x|\vec{d_m}, \vec{\epsilon}) \,\right]
\end{aligned}$$

The final form is not a normalized probability and thus not the full posterior. However, this form tells us how we can generate samples from the full posterior. First, we sample $\vec{\epsilon}$ from the posterior $p(\vec{\epsilon}|\vec{d_r})$ and then sample x from the conditional posterior $p(x|\vec{d_m}, \vec{\epsilon})$. Repeating this many times will yield good estimations for the moments of $x|\vec{d_m}, \vec{d_r}$ via the sample moments. This also includes the mean and variance in specific which can be used as an estimator and corresponding uncertainty estimate for x.

Reference measurements can often be made arbitrarily informative, by increasing measurement duration, as they do not tend to be constrained by dose or time. If the reference measurement provides very strong evidence we have $p(\vec{\epsilon}|\vec{d_r}) \approx \delta\left(\vec{\epsilon} - \hat{\vec{\epsilon}}\right)$ and thus

$$p(x|\vec{d_r}, \vec{d_m}) \propto \int p(x|\vec{d_m}, \vec{\epsilon}) \delta\left(\vec{\epsilon} - \hat{\vec{\epsilon}}\right) d\vec{\epsilon} = p(x|\vec{d_m}, \hat{\vec{\epsilon}})$$

but because $p(x|\vec{d_m}, \hat{\vec{\epsilon}})$ is a properly normalized distribution we actually have full equality rather than just proportionality here. This result is very unsurprising as there is no meaningful difference between knowing a parameter a posteriori with absolute certainty from a different measurement and knowing it a priori. Therefore, this limit can be treated just as the case in the previous section with the same closed form, but with $\vec{\epsilon} = (\vec{d_r} + 1) \oslash \vec{r}$ (the mean of the gamma posterior).

The use of multiple detectors reduces the total uncertainty approximately by the root of the number of detectors. This can be seen by calculating the standard deviation of the final Gamma posterior directly:

$$
\begin{aligned}
x &\sim \Gamma\left[1 + \sum_i d_i, \sum_i \epsilon_i\right] \\
\Rightarrow \mathrm{Var}[x] &= \frac{1 + \sum_i d_i}{\sum_i \epsilon_i} = \frac{\bar{d}}{\bar{\epsilon}}\left(n_d + \frac{1}{\bar{d} n_d}\right) \\
\Rightarrow \sigma_x &\propto \sqrt{n_d + \frac{1}{\bar{d} n_d}}
\end{aligned}
$$

where $\bar{\cdot}$ indicates the mean over the detectors and n_d the number of detectors and where typically $\bar{d} n_d \gg 1$.

4.5.1.3. Example

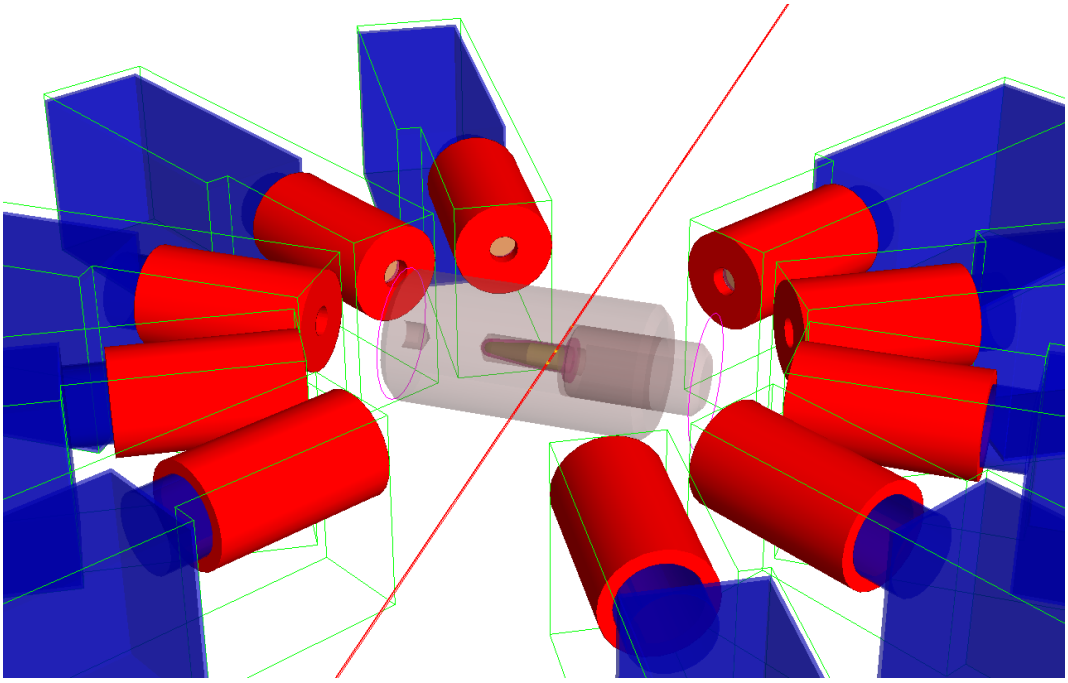

Figure 4.5. Measurement Apparatus in Simulation: A typical measurement setup was implemented in GEANT4, consisting of the incident beam (red line from bottom left to top right), ten detectors (blue and red boxes arranged in a circle) and the subject, here a polymethylmetacrylate phantom containing an Eppendorf tube with tracer solution (transparent and yellow in the center).

To obtain the data for this example, a PMMA cylinder with 1.5 cm radius containing a 0.5 mL Eppendorf tube filled with a tracer solution with palladium, iodine and gold was simulated (see figure 4.5). The concentration of the tracer concentration was varied between $0.025 - 1.0\,\mathrm{g\,L^{-1}}$ with six different values. Ten detectors arranged in a circle around the cylinder at 6 cm distance were used. It was assumed that a method for finding the observed number of fluorescence photons given a measurement is available. The data was thus fully represented by the observed number of fluorescence photons for each combination of the

three elements, ten detectors and six concentrations. The data for $0.5\,\mathrm{g\,L^{-1}}$ was taken to represent the reference measurement.

As a first step the linearity assumption is tested using a Student-T test for the Pearson correlation coefficient ([1]). For each pair of distinct detectors one such test is performed. In all cases a strong correlation is found. The minimum p value (likelihood of a correlation being that strong under the assumption of independence) was $p \approx 5 \cdot 10^{-6}$. We can thus proceed with the previously outlined method which uses linear correlation as an assumption.

As discussed in section 4.1, the posterior for the $\epsilon_{d,e}$ for each detector and element was chosen to be

$$p(\epsilon_{d,e}|\vec{d}) = \Gamma\left[\, d_{d,e,0.5\,\mathrm{g\,L^{-1}}} + 1, 0.5\,\mathrm{g\,L^{-1}} \,\right]$$

Afterwards the previously discussed method for generating samples from the posterior for the concentration was applied using the data for the other five concentrations.

$$\epsilon_{d,e} \sim \Gamma\left[\, d_{d,e,0.5\,\mathrm{g\,L^{-1}}} + 1, 0.5\,\mathrm{g\,L^{-1}} \,\right]$$

$$x_{d,e,c} \sim \Gamma\left[\sum_d d_{d,e,c} + 1, \sum_d \epsilon_{d,e} \right]$$

Then mean and standard deviation for the posterior samples for the concentrations $x_{d,e,c}$ were calculated using the standard formulas for the cumulants of the Γ distribution ([1]).

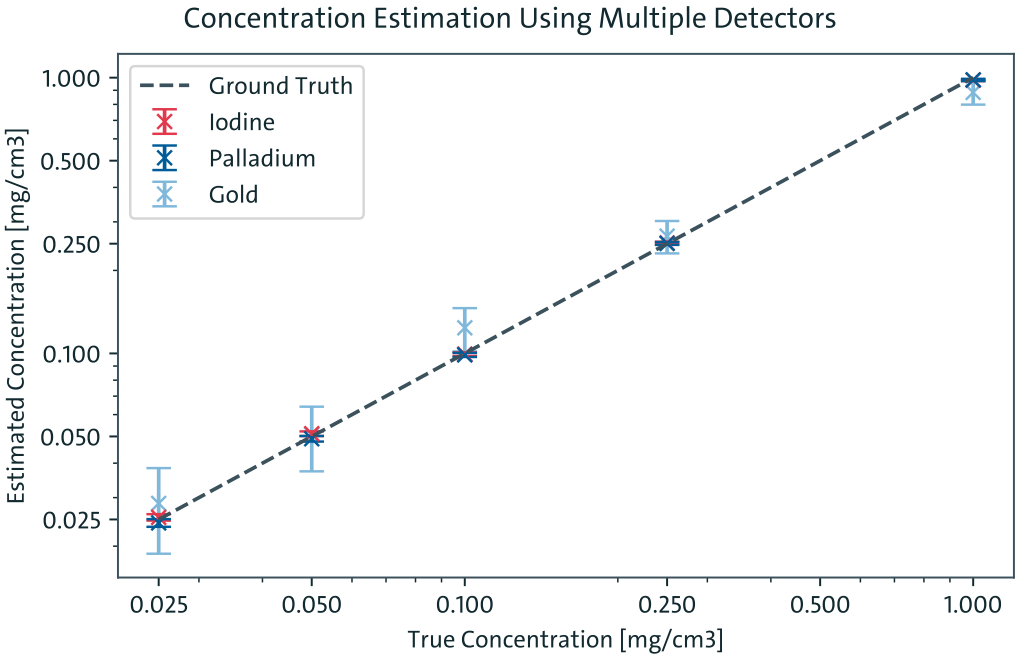

Figure 4.6. Concentration Estimate Using Multiple Detectors: The estimated concentration using data from all ten detectors was plotted against the ground truth value set in the simulation. The error estimates are well calibrated ($\chi^2_{\mathrm{red.}} \approx 1.13$) and the estimate is very precise for palladium and iodine.

The results of this concentration estimation can be seen in figure 4.6 as well as in relative terms in table 4.1. The error estimate is well calibrated with $\chi^2_{\mathrm{red.}} \approx 1.13$. For the two lighter elements the uncertainty was also quite small, partially due to the reduction of a factors ≈ 3 from using all the detectors rather than just a single one. In the case of Gold the errors are quite large. This is because the vast majority of L-line fluorescence is absorbed within the object, resulting in the relative uncertainty from Poisson error in the reference measurement alone being between 15% and 70% depending on the detector, depending on the detector angle which impacts the length of the path from the fluorescence site through the object. However, the significant reduction in uncertainty from using all the available data remains.

Conc.	I		Pd		Au	
	obs.	est.	obs.	est.	obs.	est.
0.025	1.8%	2.7%	-2.8%	3.1%	14.3%	39.2%
0.050	2.3%	2.0%	-1.9%	2.3%	1.5%	26.7%
0.100	-0.6%	1.4%	-1.2%	1.7%	23.9%	22.1%
0.250	0.6%	1.0%	0.2%	1.2%	6.5%	14.3%
1.000	-2.1%	0.7%	-1.8%	0.9%	-11.6%	8.6%

Table 4.1. Relative Errors of Reconstructed Concentration: the relative observed deviation between ground truth and reconstruction (obs.) and the estimated relative uncertainty (est.) is shown for each of the three elements (iodine, palladium and gold) and each concentration (Conc. in $\mathrm{g\,L^{-1}}$). The error estimates are well calibrated and the process is very precise for palladium and Iodine. The large errors for Gold stem from strong attenuation of the L-line fluorescence in the object.

4.6 Detection Tests

After having performed a measurement and having fit the data a common question is whether one is certain that any tracer was present at all. This can be framed in two ways: as a one sided null-hypothesis test with the null-hypothesis being that there is no tracer present, or as a model selection test where we determine whether a model with or without tracer is better suited to express the data.

Several statistical tests were introduced in subsection 3.1.5, most of which can be applied to XRF spectra using the models discussed in section 4.2. The test statistics then reduce to being functions of the fit parameters, the reconstruction error or the data likelihood. Each test statistic produces a family of tests parameterized by a threshold for the test statistic.

4.6.1. Evaluation Methodology

Evaluating different tests requires a dataset with known labels regarding fluorescence content. This dataset should be chosen such that the failure rates of the tests can be evaluated in a meaningful way, but also resemble typical real-world data.

To ground the spectra in reality a GEANT4 simulation was performed to obtain a baseline spectrum. The subject consisted of aqueous Iodine solution in a $350\,\mu\mathrm{L}$ cuvette, a $53\,\mathrm{keV}$ incident beam and silicon drift detectors modeled after [53]. The setup can be seen in figure 2.4, note that iodine instead of various metals was used, so the signal region overlaps with third order compton scattering, a spectrum is shown in figure 4.8. Next, to enlarge the dataset to sufficient size, linear spectral scaling and sampling from the resulting posterior predictive was used to generate a dataset with $5 \cdot 10^6$ spectra corresponding to random numbers of incident photons around 10^8 and random concentrations around $100\,\mu\mathrm{g\,mL^{-1}}$.

All tests used the same models for signal and background with the former being a Gaussian at known location and the latter being a polynomial. Weighted least squares fitting was used to find the model parameters and their respective uncertainties.

4.6.2. State of the Art

The current state of the art, as discussed in chapter 2, is to perform a one-sided z-test of $\hat{f} \sim \mathcal{N}\left[0, \hat{b}\right]$, i.e.

$$z := \hat{f}/\sqrt{\hat{b}}$$

where $\hat{f}$ and $\hat{b}$ are the signal and background estimates in the signal region acquired through fitting. This can be interpreted as a subtraction of the background in the regions without a signal component from the total signal. This method implicitly adds the extra assumption, that the background, which is fit, and the background, which is overlaying the signal, are

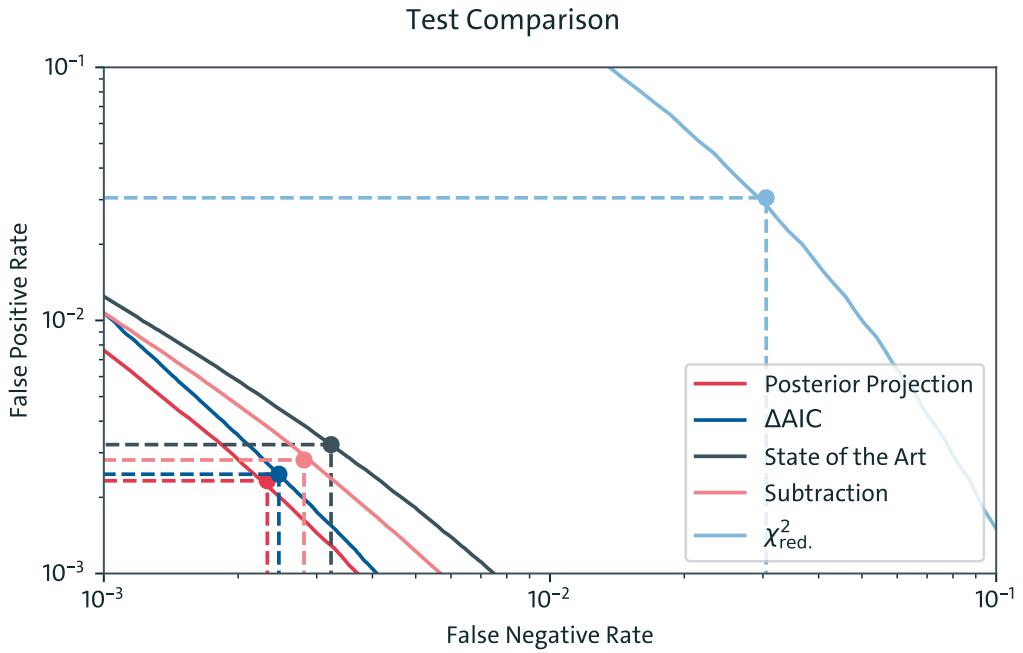

Figure 4.7. Comparison of different tests: Error rate curves for the current state of the art, posterior projection, the low count limit of subtraction, the AIC and the $\chi^2_{\text{red.}}$ of the null-hypothesis model. The shown curves were acquired using $5 \cdot 10^6$ synthetic data samples of a cuvette with iodine solution or water with random numbers of incident photons around 10^8 and random concentrations around $100\,\mu g\,mL^{-1}$. The dashed lines indicate the equal error rate point. A test can be said to dominate another if its curve is further to the lower left. Posterior projection appears to be the best choice for this task.

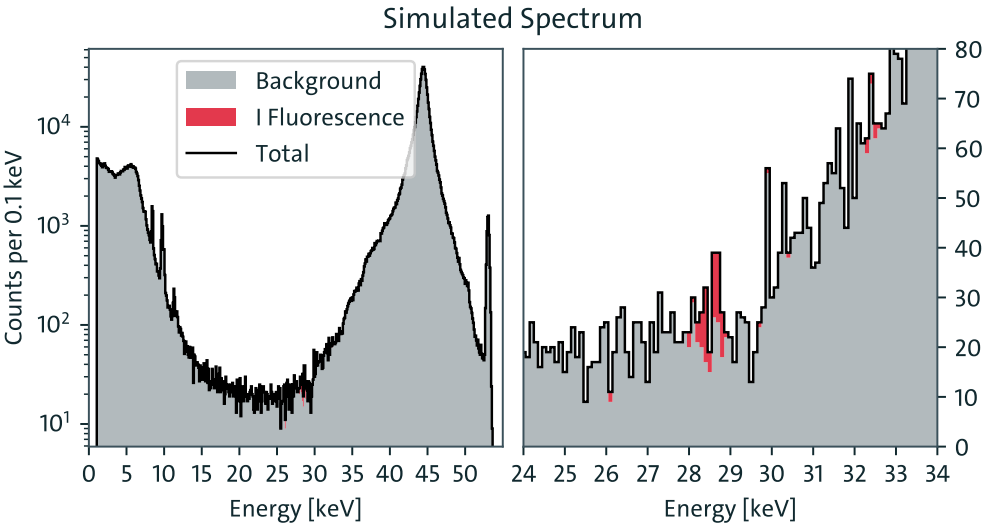

Figure 4.8. Evaluation Spectrum: Example of a spectrum used for test evaluation. The iodine fluorescence signal (red) is small compared to the background (grey), even in the signal region. The background intensity additionally follows a non-linear profile.

identical. The equal error rate (EER), i.e. the error rate where false positive and false negative results are equally likely, was at 0.32% in this task.

This test is in common usage in the literature and was already described in subsection 2.4.5.5. It was however outperformed by the two other z-tests using Bayesian estimates for σ.

4.6.3. Subtraction Method

The state of the art method can be improved upon by using the results in section 4.3 to remove the assumption that the background, which is fit, and the background overlapping the signal, are equal. Instead they can be treated as independent but identically distributed. This changes the test statistic to

$$z := \frac{\hat{f}}{\sqrt{2\hat{b} + 2}}$$

That this is a more accurate estimate of the true uncertainty can be seen by the fact that this test strictly dominates the state of the art, meaning that, at the same sensitivity as the state of the art, this test is more specific and vice versa. The lower EER of 0.28% also reflects this fact.

4.6.4. χ^2 Null-Hypothesis Test

The most straight-forward null-hypothesis (NH) test would be to fit a linear model without a signal component to the data and see whether the fit is sufficiently accurate, for example by performing a χ^2_{red}-test. This would be a null-hypothesis test in the truest sense, as this test only dismisses one very specific null-hypothesis. Importantly, we do not consider whether the hypothesis we are actually interested in (the data containing a specific fluorescence line) represents the data better like we do in other tests.

Figure 4.7 shows that this ends up making the χ^2 NH test unsuitable for this specific application as better tests exist, because all other tests strictly dominate this one. The comparatively very high EER of 3.0% also suggests the same.

4.6.5. Posterior Projection

As discussed in section 4.2, a linear model can project the uncertainty induced by Poisson noise on the data to the parameter space. Thus this process can provide both μ_f, the point prediction for the most likely level of fluorescence, and σ_f, an uncertainty estimate for that parameter. Using these two a z-test can be performed with

$$z := \frac{\mu_f}{\sigma_f}$$

For this dataset the posterior projection z-test strictly dominated all other tests and had an EER of 0.25%. An important requirement for this test is that the model must be linear to be exact and the model to be differentiable to obtain an approximate σ_f. Other methods discussed here have no such requirements.

4.6.6. Akaike Information Criterion

One can treat the task of testing whether a signal contains fluorescence as a model selection task. If models with a fluorescence component are better suited to the data, then it is reasonable to assume that such a component is also present in the data. This means that information criteria like the AIC and BIC discussed in subsection 3.2.5 can be used to perform the test by comparing the ΔAIC between the background models and the full model against a threshold. As this comparison only uses a single background and single signal model the difference between AIC and BIC disappears in the error rate curves. With an EER of 0.25%, the ΔAIC performed slightly worse than posterior projection, however, this test is far more general and could in principle be performed on a large ensemble of models, including non-linear ones. Figure 4.7 also shows that the AIC is particularly suitable to high-specificity applications, as the curve false positive rate as a function of false negative rate is steeper than for the other tests.

4.7 Summary

Various statistical techniques for the processing and evaluation of XFI spectra were summarized. For many applications a mix of classical and Bayesian techniques was shown to provide a good compromise between computational tractability and accuracy. In some cases like the subtraction method, the Bayesian approach induces a correction to the uncertainty estimate that avoids divergence in the low count limit. No single solution fit for all purposes exists, rather one needs to select a spectral model and signal extraction method based on the specific application and data availability.

CHAPTER 5

3D RECONSTRUCTION

5.1 XRF CT

The most intuitive method for creating 3D images from XFI measurements would be to reconstruct them exactly the same way as a transmission CT image. This section discusses XRF CT, studying a possible correction to the standard back-projection method stemming from the differences between XRF and CT measurement geometries.

5.1.1. XRF CT

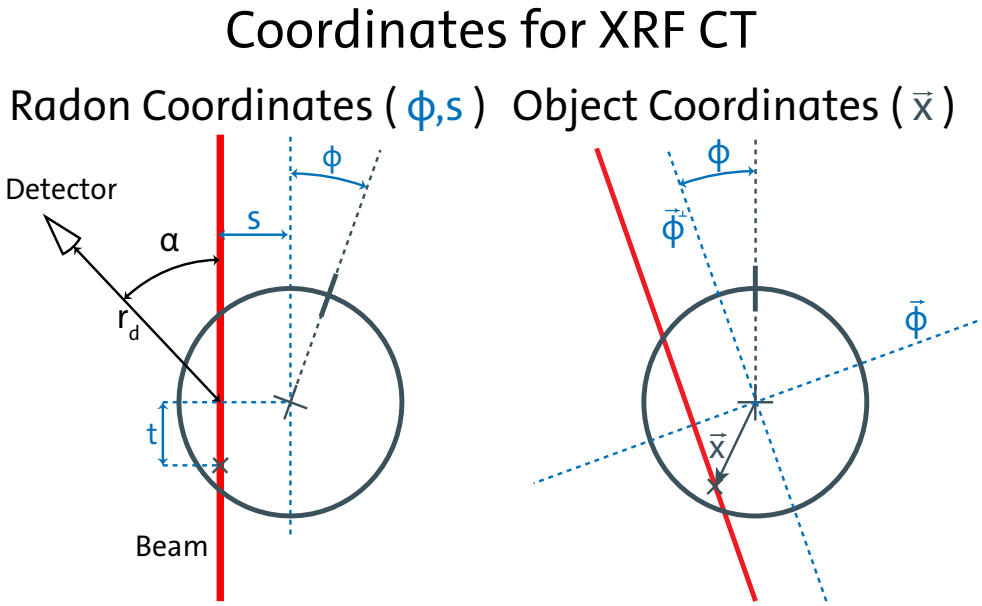

Figure 5.1. Schematic of XRF CT: The Radon transform converts from the Object coordinates (right) to the Radon coordinates (left) by integrating along the beam coordinate t. In standard CT images the detector angle $\alpha = 0$, while it is generally non-zero in XRF tomography.

XRF CT is typically performed analogously to traditional CT imaging (see subsection 2.2.3) [21]. However, unlike in CT, the detector is not in line with the beam, being offset by an angle α. Using the same geometrical approach as in CT, but with the signal

function of XFI ([21]), one can find the signal function for XRF CT:

$$\begin{aligned} &S(E_{\text{out}}, \theta, s) \\ &= \int_0^\infty dE_{\text{in}} \int_0^\infty dt\ I_0(E_{\text{in}}) T_{\text{in}} \mu_F(E_{\text{in}}, E_{\text{out}}, \vec{x}(t)) T_{\text{out}} \frac{A_D}{4\pi r_D(t)^2} \\ &=: I_0 \int_0^\infty dt \int_0^\infty dE_{\text{in}}\ \mu_F(E_{\text{in}}, E_{\text{out}}, x(t)) C(t, E_{\text{out}}, E_{\text{in}}) \\ &=: \int_0^\infty dt\ O(\vec{x}(t), E_{\text{out}}) \end{aligned}$$

$$T_{\text{in}} := e^{-\int_0^t dt' \mu_T(E_{\text{in}}, \vec{x}(t'))}$$
$$T_{\text{out}} := e^{-\int_0^\infty dt' \mu_T(E_{\text{out}}, \vec{y}(t,t'))}$$

where symbols were used to be consistent with the Radon description in subsection 2.2.3, μ_F is the linear fluorescence cross-section, A_D is the detector area, $r_D(t)$ is the distance from $\vec{x}(t)$ to the detector, $\vec{y}(t)$ parametrizes the path from $\vec{x}(t)$ to the detector, E_{in} is the incident energy and E_{out} is the outgoing energy. The expression reads as the incident beam being attenuated on the way to $\vec{x}(t)$, then a fraction of that beam induces fluorescence at that position, this fluorescence is attenuated on the way out and only a portion based on the solid angle hits the detector. Collecting terms based on their dependencies and integrating over E_{in} gives the more compact form below where C is constant in μ_F and collects all the non-fluorescence components. If we were to impose the Radon structure we would then have the object function be a convolution between C and μ_F in the energy domain. [21]

The most straightforward method for reconstruction is to treat all E_{out} as independent and first find $O(\vec{x}, E_{\text{out}})$ (the energy-resolved linear fluorescence coefficients) using the standard inverse Radon transform discussed in subsection 2.2.3 to find μ_F and then using known cross-section data to calculate back to ρ_F, the density of the tracer in each location. C can be either estimated a priori, using approximate geometry and the Lambert-Beer law, or computed through line integration from a spectral CT image acquired at the same time. The former is more common in practice (see [21]). Usually it is more convenient to integrate over E_{out} in an energy interval around the fluorescence peaks instead of performing the reconstruction in an energy-resolved manner [21]. The advantage of using this approach is twofold: it offers a way to combine data from multiple energy bins and it also allows for the usage of spectral fitting techniques (see section 4.2) to obtain the fluorescence signal from the spectrum.

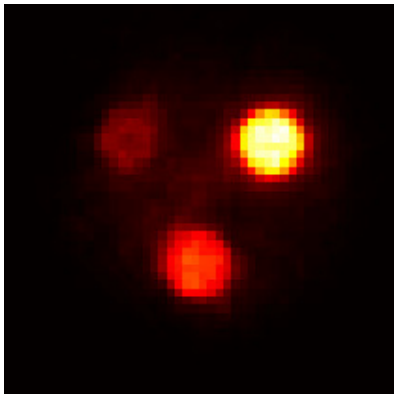

Figure 5.2. An XRF Tomography Image: this image shows a reconstruction from measured XRF data of a phantom with three inserts. The data was acquired by scanning the subject multiple times and rotating in between, the reconstruction used back-projection. Cropped from [21].

Performing simultaneous acquisition of a spectral CT image requires a polychromatic narrow pencil beam source with a peak of sufficiently high energy such that the subject is penetrated, an energy-resolved detector measuring transmission and an energy-resolved detector at an angle where the fluorescence produces a good signal to noise ratio (see chapter 2). Neither of these requirements is trivial, especially while maintaining in vivo conform measurement times, meaning the source needs to have a high flux.

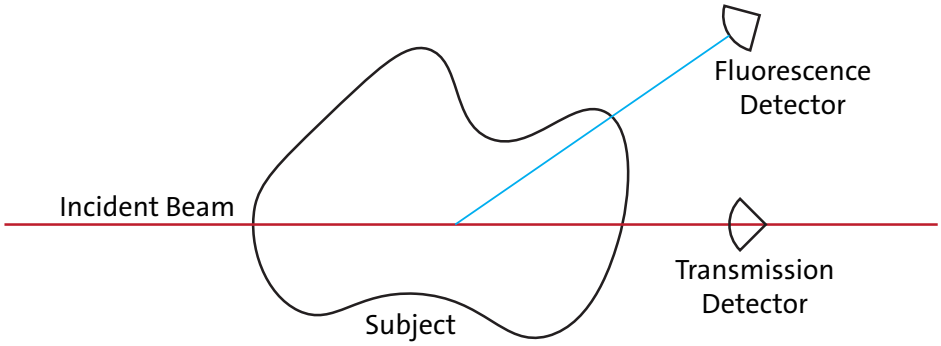

Figure 5.3. Setup for Co-acquisition of XFI and CT: an incident x-ray beam (dark red) is used to scan the subject (center). The transmission is captured by a detector on the beam axis (right), another detector captures the fluorescence (top right). From the former signal a spectral CT image can be reconstructed.

A difficult question with this approach is how one obtains reasonable uncertainty estimates of $\rho(\vec{x})$. One could treat it as a function of the random variable S and propagate the uncertainty, but that would not capture the structure imposed on ρ from the reconstruction process, only the effect of Poisson noise, for example the impact of starburst artifacts. It also involves two dependent reconstruction steps, with the first one being spectral CT, the second one for the XRF CT which uses the former. Therefore, the uncertainty would be propagated through multiple Fourier transforms and spectral filters which are strictly speaking ill-posed and can introduce vanishing or diverging determinants, which would make the uncertainty estimate unusable or unreliable.

The other major issue is that of dose and measurement time. Performing a full or half rotation per scanned slice increases the number of scan positions by a corresponding factor, for example that of 31 angular positions. If the dose is kept constant compared to a 2D XRF image one ends up in the very low count regime of the Poisson distribution in the individual spectra which is not ideal for the linear spectral models. Additionally, the time to move the measurement setup into the correct positions can quickly add up, leading to much longer imaging times than total dose and flux would suggest. Both aspects are problematic for in vivo measurements which are both time and dose constrained.

5.1.1.1. The Projection Kernel

The derivation in subsection 2.2.3 showed how one can find the kernel capturing the impact of backprojection from a sinogram. In XRF measurements this is more complicated because the detector is not in line with the incident beam, but at an angle. This induces a dependence of the solid angle that the detector covers from any given position in the imaging process. Specifically the distance to and alignment with the detector area change with the coordinate along the beam. This coordinate only appears inside the inner integral and thus the impact cannot be removed through pre- or postprocessing. We can find the

kernel through the same process as before.

$$B(\vec{y}) = \int_0^{2\pi} S(\phi, \vec{\phi} \cdot \vec{y}) d\phi \tag{5.1}$$

$$= \int_0^{2\pi} \int O((\vec{\phi \cdot} \vec{y})\vec{\phi} + t\vec{\phi}^{\perp}) P(t) dt d\phi$$

$$= \int O(\vec{x}) \int_0^{2\pi} P(\vec{\phi}^{\perp} \cdot \vec{x}) \delta\left(\vec{\phi} \cdot (\vec{x} - \vec{y})\right) d\phi d^2\vec{x}$$

$$= \int O(\vec{x}) \frac{1}{|\vec{x} - \vec{y}|} \left(P(\frac{\vec{y} \cdot \vec{x}}{|\vec{x} - \vec{y}|}) + P(-\frac{\vec{y} \cdot \vec{x}}{|\vec{x} - \vec{y}|}) \right) d^2\vec{x} \tag{5.2}$$

$$B(\vec{y}) =: \int O(\vec{x}) K_{\text{combined}}(\vec{x}, \vec{y}) d^2\vec{x}$$

with

$$K_{\text{combined}}(\vec{x}, \vec{y}) = \frac{1}{|\vec{x} - \vec{y}|} \left(P(\vec{\theta} \cdot \vec{x}) + P(-\vec{\theta} \cdot \vec{x}) \right) \tag{5.3}$$

where the projection P can be calculated from the t coordinate and the dector distance R_d and angle α relative to the origin using basic geometry

$$\tau := \frac{t}{R_d}$$

$$P_{\text{area}}(t) = \sqrt{1 - \frac{\sin^2(\alpha)}{\frac{1}{\tau^2} + \frac{2\cos(\alpha)}{\tau} + 1}}$$

$$P_{\text{distance}}(t) = \frac{1}{1 + 2\tau\cos(\alpha) + \tau^2} \tag{5.4}$$

$$P(t) = P_{\text{area}}(t) \cdot P_{\text{distance}}(t) = \frac{\sqrt{1 - \frac{\sin^2(\alpha)}{\frac{1}{\tau^2} + \frac{2\cos(\alpha)}{\tau} + 1}}}{1 + 2\tau\cos(\alpha) + \tau^2} \tag{5.5}$$

Importantly, K_{combined} (equation (5.3)) is not translation invariant as it depends on $\vec{x}$ and $\vec{y}$ independently, not just on $\vec{x} - \vec{y}$. Fourier deconvolution like with the pure ramp filter is thus not possible. This is problematic for applications where the impact of the projection kernel is a leading contribution to reconstruction error, because inversion of general convolution kernels is not always possible [61].

For common values like with $(t_{\max} - t_{\min}) \approx R_d$ i.e. the detector distance from the origin being around the maximum diameter of the object, the impact of the projection kernel is however minimal. The simulated raw images obtained through convolution with the full kernel and the pure CT kernel differ only on the order of a few percent. One can see in figure 5.4 that there is essentially no difference between the reconstructions from the combined kernel and only the standard CT kernel using the ramp filter.

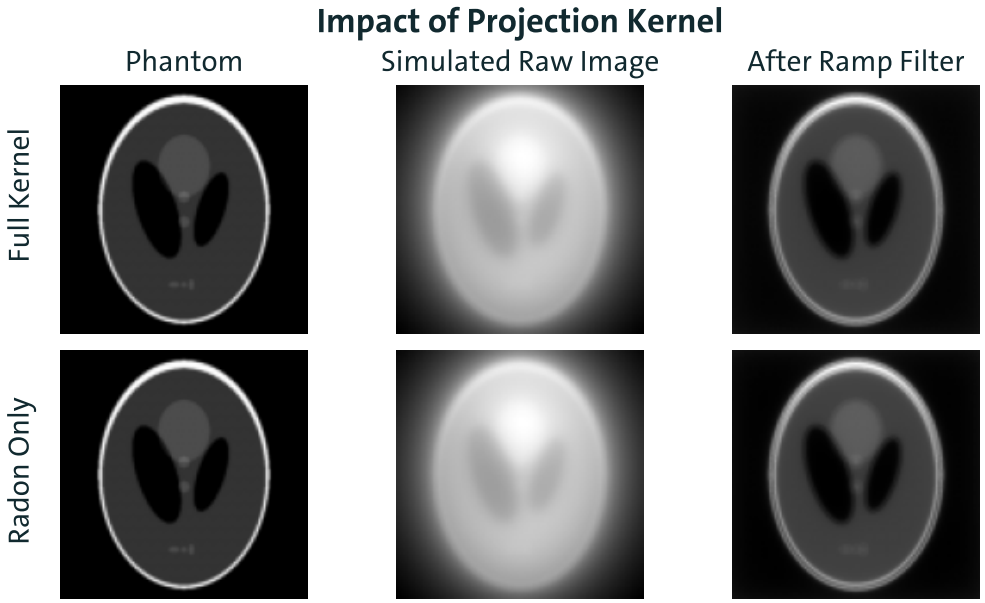

Figure 5.4. Impact of the Projection Kernel: The imaging kernels of standard Radon (bottom row) and the one including the projection as well (top row) were applied to an image of a phantom (central column). A ramp filter was then applied to that simulated raw image (right column). The top right image represents ignoring the projection kernel in reconstruction. The induced error is minimal as no difference in the reconstruction after the ramp filter can be seen. The object was chosen to be $3 \times 3\,\text{cm}$ in dimensions and the distance to the detector to be $6\,\text{cm}$. The top and bottom row appear near identical, indicating that the impact of the XRF specific kernel components is minimal.

5.2 Reconstruction from a Single Projection

In trying to address the issues mentioned above, an attempt was made to reconstruct XFI images from a single projection. This might seem impossible due to the dimensionality of the problem - reconstructing a 3D image from 2D data. The key idea was to use detectors in multiple different locations to give the 2D data a third, discrete, "detector" dimension. For example, with ten detectors and 100×25 scan points the data would have a shape of $(100, 25, 10)$. So one might think that this data would be sufficient to reconstruct up to ten positions in the depth direction while keeping the direction in the other two directions. It will be shown that this does not work because the projection is singular.

5.2.1. Projection Matrix

The imaging equation for a single projection can be found by substituting a fixed ϕ into equation (5.2), yielding $S_i = \int dz\; C_i(z)\mu_F(z)$, where $z = t\cos(\phi)$ is the coordinate along the beam and i indexes the detectors. Discretizing this in z gives the problem a linear form

$$\vec{S} = \mathbf{C}\vec{\mu}_F$$

where each component of $\vec{S}$ is the signal in one detector and the component of $\vec{\mu}_F$ is the linear fluorescence coefficient at one z-position. $\mathbf{C}$ acts as a projection matrix between the two spaces and encodes the physics of the problem. If $\mathbf{C}$ were to be known, one could find a (weighted) least squares fit for $\vec{\mu}_F$ with the exact same methods as discussed in section 4.2.

$\mathbf{C}$ can in principle be found by using equation (2.8) to compute how much signal each detector is expected to receive from each z-position. As discussed in subsection 5.1.1, the attenuation correction would still require either strong assumptions about the subject, or

access to a spectral CT. But unlike with XFI tomography, spectral CT cannot be acquired in the same imaging process anymore as no sinogram is recorded. However, usually the solid angle term (equation (5.5)) dominates because attenuation is weak ([17]), so the problem can still be approached by first studying only the solid angle component. In this simplification $\mathbf{C}_{ij} \approx c d_{ij}^{-2}$ where c is a constant and d_{ij} is the distance from the location i to detector j. Using the law of cosines this can be computed as $d_{ij}^2 = R_i^2 + z_j^2 + 2R_i z_j \cos(\alpha_i)$. Combining this yields

$$\begin{aligned}\mathbf{C}_{ij} &= \frac{A_d}{4\pi\left(R_i^2 + z_j^2 + 2R_i z_j \cos(\alpha_i)\right)} T_{\text{in}}(z_i) T_{\text{out}}(z_i) \Delta z \\ &=: c \frac{1}{R_i^2 + z_j^2 + 2R_i z_j \cos(\alpha_i)} \end{aligned} \tag{5.6}$$

Here, the attenuation terms T_{in} and T_{out} are assumed to be slowly varying with z, α and are thus grouped into the shared constant.

The model was further simplified to have M detectors equally spaced on a circle around the origin with radius R and N z-positions equally spaced between $-R/2$ and $R/2$. A factor of R^{-2} can now also be absorbed into the shared constant leaving a unitless version of the matrix with $\zeta := z/R$ ranging from -0.5 to 0.5

$$\begin{aligned}\tilde{\mathbf{C}}_{ij} &:= \frac{R^2}{c}\mathbf{C}_{ij} = \frac{1}{1 + \zeta_j^2 + 2\zeta_j \cos(\alpha_i)} \\ &= \frac{1}{1 + (\frac{j-1}{M-1} - 0.5)^2 + 2(\frac{j-1}{M-1} - 0.5)\cos\left(\frac{2\pi i}{M+1}\right)}\end{aligned}$$

This simplified version only depends on the number of detectors M and number of z-positions N.

An important criterion is the noise sensitivity of the resulting reconstruction. Let $\vec{\lambda}$ be the expected number of fluorescence photons in each detector, $\vec{F}$ be the recorded number of photons and $\hat{\vec{\mu}}$ the estimate for the linear fluorescence coefficients. The model can be written as:

$$\vec{\lambda} = c\tilde{\mathbf{C}}\vec{\mu}$$

$$\vec{F} = \vec{\lambda} + \vec{\epsilon} \tag{5.7}$$

$$\hat{\vec{\mu}} = \tilde{\mathbf{C}}^\dagger \vec{F} = \tilde{\mathbf{C}}^\dagger \tilde{\mathbf{C}} \vec{\mu} + \tilde{\mathbf{C}}^\dagger \vec{\epsilon} \tag{5.8}$$

The pseudo-inverse $\tilde{\mathbf{C}}^\dagger$ projects the noise $\vec{\epsilon}$ onto the model parameters $\hat{\vec{\mu}}$. The noise in the detectors is independent, so the noise vector $\vec{\epsilon}$ can be assumed to be white. White noise remains white under all unitary transformations [1]. As discussed before, $\tilde{\mathbf{C}}^\dagger = V\mathbf{S}^{-1}U^T$ where $\mathbf{S}$ is a non-square diagonal matrix containing the singular values of $\tilde{\mathbf{C}}$ and U, V are unitary. The noise amplification is then primarily driven by the largest eigenvalue of $\mathbf{S}^{-1}$, which is the inverse of the smallest eigenvalue of $\mathbf{S}$. In the most extreme case $\tilde{\mathbf{C}}$ is singular, meaning it has at least one singular value of 0 in which case the problem becomes ill-posed. If $\tilde{\mathbf{C}}$ had a smallest singular value of 0.01, one noise component would be amplified by a factor of 100.

The singular value decomposition was computed for values up to $N \leq M \leq 2$ using pytorch [62]. The minimum singular value is shown in figure 5.5. For the number of z-positions $N \geq 7$ the smalles singular value was below 10^{-3} irrespective of the number of detectors used. This suggests that the single projection method is highly sensitive to noise.

To illustrate the impact of this sensitivity a projection matrix $\mathbf{C}$ of the previously explained structure was constructed with $R = 3$ and $M = N = 11$. The true object $\vec{o}$ was chosen to follow a Gaussian profile in z. The expectation value of the signal was the projection of the object according to the projection matrix $\mathbf{C}\vec{o}$. Afterwards white Gaussian noise

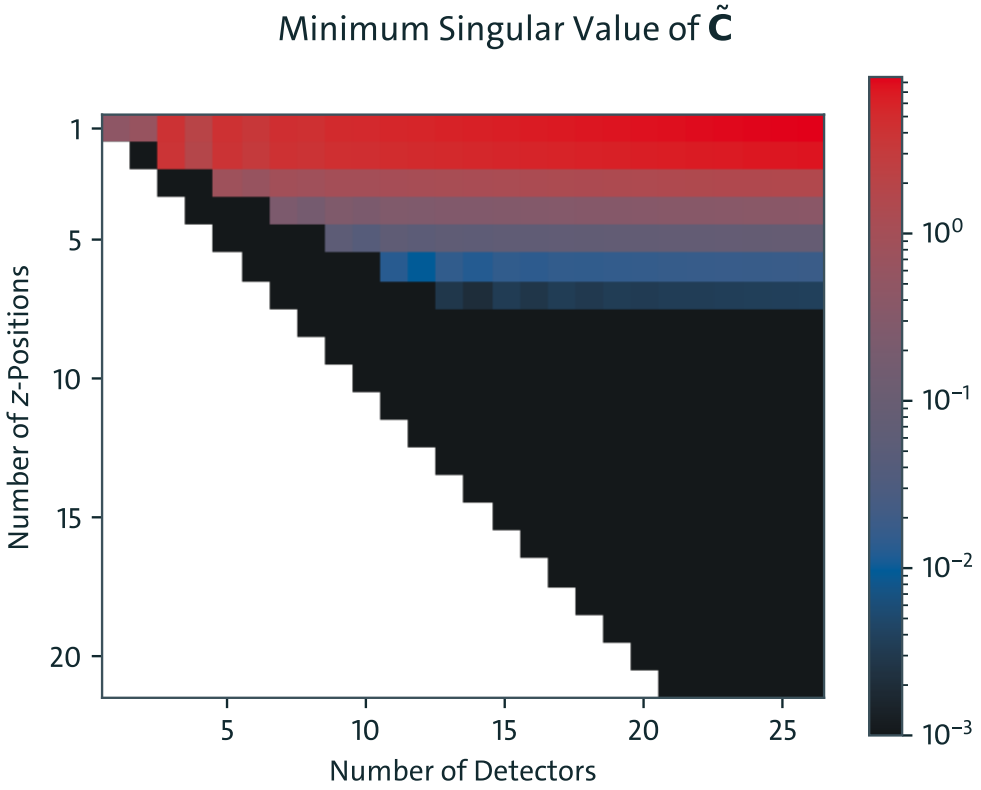

Figure 5.5. Minimum Singular Values of the Simplified Projection Matrix $\tilde{\mathbf{C}}$: The smallest singular value of the simplified projection matrix quickly converges to 0 as the number of z-positions increases. This indicates that the reconstruction becomes highly sensitive to noise.

$\vec{\epsilon} \sim \mathcal{N}\left[\vec{0}, 10^{-6}\mathbb{1}\right]$ was added pointwise to the signal to emulate noisy data with the relative noise strength being $\mathcal{O}\left(10^{-4}\right)$. The data was then projected back to z space using the inverse of $\mathbf{C}$. Here $N = M$ and the smallest singular value was not exactly 0, so $\mathbf{C}^{+} = \mathbf{C}^{-1}$ was a full matrix inverse. This smallest singular value was $\mathcal{O}\left(10^{-11}\right)$. In addition to the reprojection of the noisy data $\mathbf{C}^{+}\left(\mathbf{C}\vec{o}\vec{\epsilon}\right)$, $\mathbf{C}^{+}\mathbf{C}\vec{o}$ was computed to show that the reconstruction error primarily resulted from noise amplification, not from imprecise inversion. The result can be seen in figure 5.6. The reconstructed depth profile has a high frequency component which is 10^{6} times stronger than the scale of the target object.

5.2.1.1. Regularization

One way to avoid large diagonal elements in $\mathbf{C}^{+}\left(\mathbf{C}^{+}\right)^{T}$ is to regularize $\mathbf{C}$ before taking the pseudo inverse [63], which is performing an operation on $\mathbf{C}$ that ensures that no singular values are below a given threshold. The simplest form of regularization is to use the approximate projection matrix $\mathbf{M}$ with

$$\mathbf{M} = \mathbf{C} + \iota\mathbb{1}$$

where ι is a small scalar, typically smaller than the root of the typical diagonal entries of $\mathbf{C}\mathbf{C}^{T}$. This ensures that $\mathbf{M}^{+}$ is still a reasonable approximate inverse of $\mathbf{C}$.

With the extremely low noise component of $\approx 0.01\%$ in the data, this simple regularization is sufficent to obtain a reconstruction that meaningfully resembles the object. This can be seen in figure 5.7. Smaller ι increase the noise sensitivity and thus uncertainty of the reconstruction while larger values increase the systematic error (which here manifests as a broadening of the peak). Instead of picking a particular noise realization as before, the Gaussian random variable itself is projected back into z space and the means and standard deviation are displayed. The 1σ envelope indicates the region where a typical reconstruction

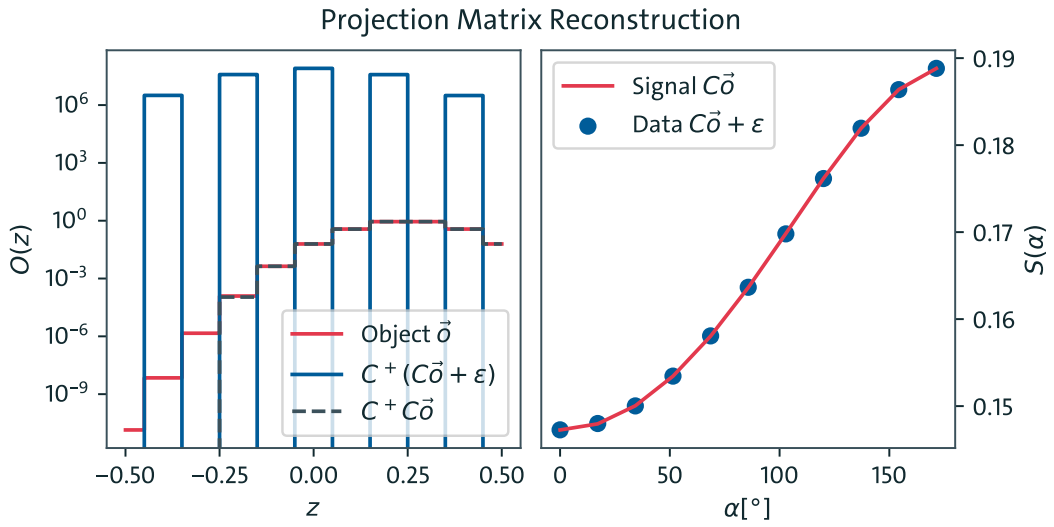

Figure 5.6. Single Projection Reconstruction: $\mathbf{C}^\dagger\mathbf{C}\vec{o}$ and $\vec{o}$ are identical, so when no noise is present in the system, the object can be reconstructed exactly. $\mathbf{C}^\dagger(\mathbf{C}\vec{o}+\vec{\epsilon})$ shows oscillations which are 10^6 times larger than the object scale, indicating complete failure of the reconstruction. The noise component is small relative to the signal strength, as can be seen by the data points $\mathbf{C}\vec{o}+\vec{\epsilon}$ being close to the idealized curve of $\mathbf{C}\vec{o}$. C is the matrix projecting from z space to α space and $\mathbf{C}^+$ its inverse. The plot in z space (left) uses a logarithmic y-axis due to the large difference in scale of $\vec{o}$ and $\mathbf{C}^\dagger(\mathbf{C}\vec{o}+\vec{\epsilon})$.

might be found. This makes it easier to see the tradeoff between accuracy and precision, however, it also hides the nature of the frequency spectrum deviation. Whether $\mathbf{C}$ is square or overdetermined again has only a negligible effect. In the 1% noise case $\iota = 10^{-3}$ shows accepatble systematic error while the statistical uncertainty is significantly greater than the object scale, which suggests that the reconstruction is not usable. With $\iota = 10^{-2}$ the statistical error is of a similar order as the object, but the systematic error induces a blurr with a width comparable to the total z size of the object, which also limits the usefulness in a potential application.

5.2.2. Iterative Maximum Posterior

Instead of directly projecting the data from detector space back onto the z-space one can attempt to find a depth distribution that minimizes a loss functional. If that functional is the negative log likelihood to observe the data given the depth distribution of the tracer, one produces an iterative approximation of the maximum likelihood estimator (MLE). A major advantage of an iterative approach is that it allows for non-linear functions in the model, for example a positivity constraint can be introduced by using $\log(\vec{m}u)$ or $\log(\exp(\vec{\mu})+1)$ as the target and then transforming back to μ. This can already eliminate the kind of behavior seen in matrix inversion.

There are many options to find such extrema, one of the more efficient ones is gradient descent, though this approach in general might only find the closest local minimum rather than a global one [64]. One way to try and mitigate this issue is to start with an ensemble of initializations and perform the optimization for each of them. If all produce the same result it is very likely a global or very large local extremum [65]. One can further improve the accuracy of the ensemble by weighing the results by their loss, for example with $\exp(-\mathcal{L})$ to obtain an ensemble that captures that some solutions are better than others and should thus be considered more strongly. This is an approximation of posterior importance weighting [66]. As this would not effect the case where all converge to the same result it is sensible to include their weighting in any case.

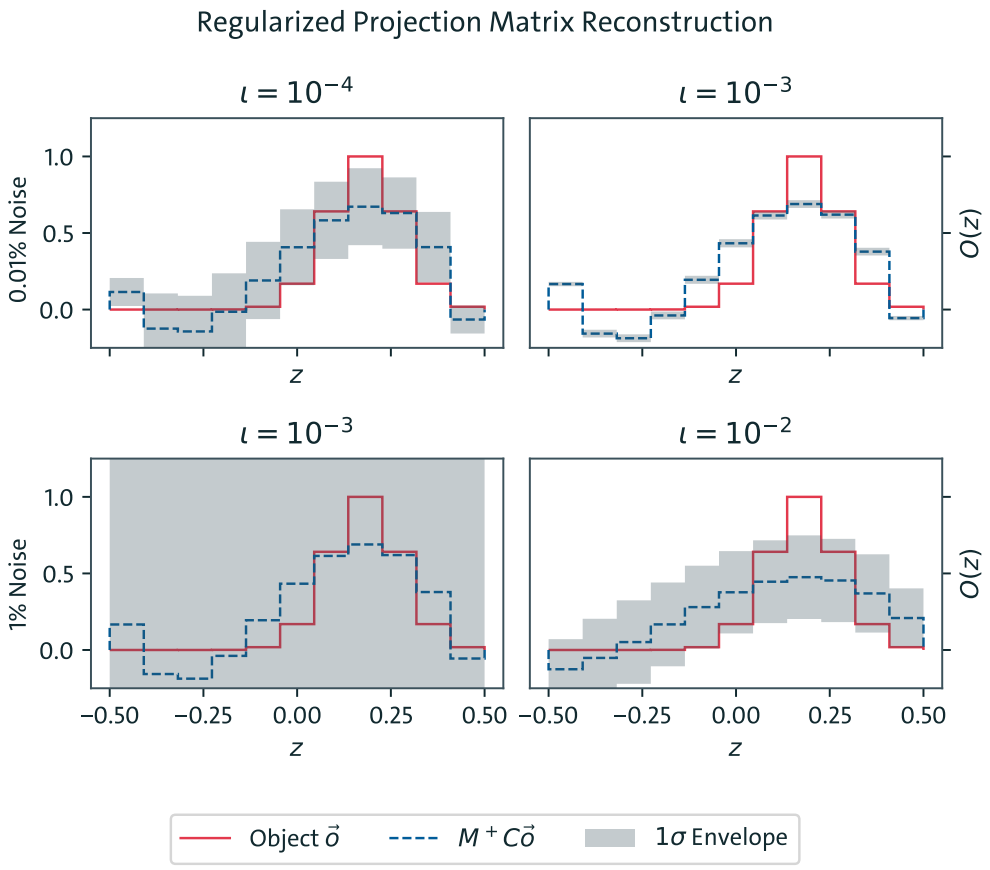

Figure 5.7. Example for Regularized Single Projection Reconstruction: The mean and standard deviation of the reconstruction are plotted together with the true object. The reconstruction is treated as the back-projection of the data as a random variable that is centered around the true signal with Gaussian noise. At 1% noise-level regularization does not offer a reasonable compromise.

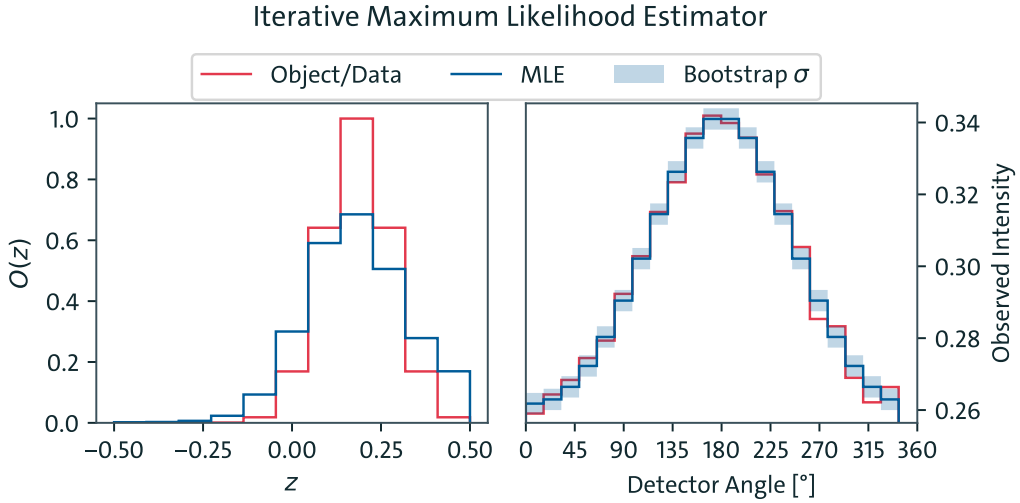

Figure 5.8. Example for Iterative Maximum Likelihood Estimation: The left hand side shows the ground truth depth distribution (red) and the MLE approximation for it (blue). The right hand side shows the data (red) and the 1σ interval of the posterior predictive using the initialization ensemble technique. The reconstructed distribution matches the observed data closely, and produces a depth distribution that is comparable to the ground truth. However, different initializations are not sufficient to produce an uncertainty estimate capable of capturing the systematic error (the ground truth is not in the uncertainty interval).

Example results from this process can be seen in figure 5.8. The projection to the detector space is kept the same as in the previous section with 21 detector angles, 11 depth positions and 1% relative noise level. The MLE does a reasonable job of approximating the true distribution and the different initializations converge to the same solution. The data is very well described by the projection of the MLE to the detector space and the noise.

An interesting observation about this approach is that it can attempt to find a z distribution of arbitrary resolution i.e. arbitrary numbers of z positions, not just less z positions than detectors. This might make the problem strongly under-determined which indicates that we need another assumption to select the solution by - a Bayesian prior. Adding a log prior to the log likelihood guarantees that the loss functional has gradients matching the log posterior, which allows an iterative approximation of the posterior maximum.

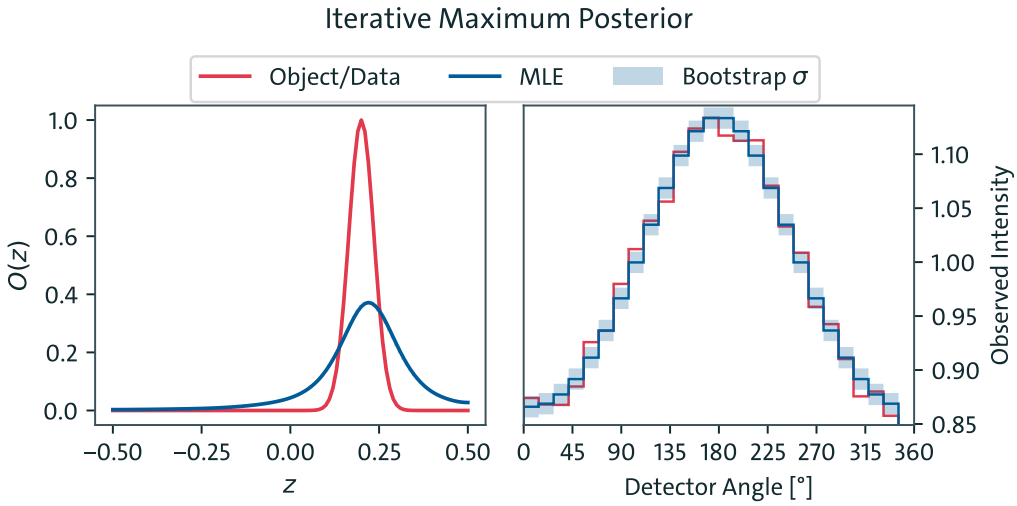

Figure 5.9. Iterative Maximum Posterior Estimation: The left hand side shows the ground truth distribution (red) along with the reconsruction using the iterative maximum posterior technique (blue). The right hand side shows the data (red) along with the posterior predictive (blue). The data is within the uncertainty range of the posterior predictive, indicating that the data is approximated well. The depth distribution shows similar general behavior with a peak located around $z = 0.2$, but is less sharp than the ground truth. The addition of a prior allowed for reconstruction of a continuous depth distribution using only a discrete set of data points. Here a very weakly informative Gaussian prior in Fourier space around $(1 + k^{5/2})^{-1}$ was used.

An example for this can be seen in figure 5.9. The object that was optimized was the cosine transform ([67]) of the distribution with $N = 25$ frequency components.

$$\mu_i = \sum_{j=0}^{N} \cos\left(\frac{\pi}{N}\frac{2i+1}{2}j\right)$$

This transform is closely related to the real Fourier transform and can similarly be evaluated for continuous i. Being in frequency-space allows the prior to depend on the frequency spectrum, in this case the prior was chosen to be a Gaussian with $\sigma = 100$ around the Matérn spectrum $(1 + k^{5/2})^{-1}$ [68] where k is the spatial frequency. The Matérn spectrum supresses high frequency modes, making the prior favor blurrier solutions in the absense of strong data. This supression is only polynomial, unlike a Gaussian, making this prior less informative regarding those high frequency modes. This bias towards spread out solutions ensures that rapidly oscillating solutions like the ones produces by the matrix inverse are supressed, favoring more plausible distributions.

A particularly challenging case is that of multi-modal depth distributions (figure 5.10). Even in a best case scenario with a large number of detectors and low noise level, the two peaks of the object function could not be resolved, being reconstructed as a single

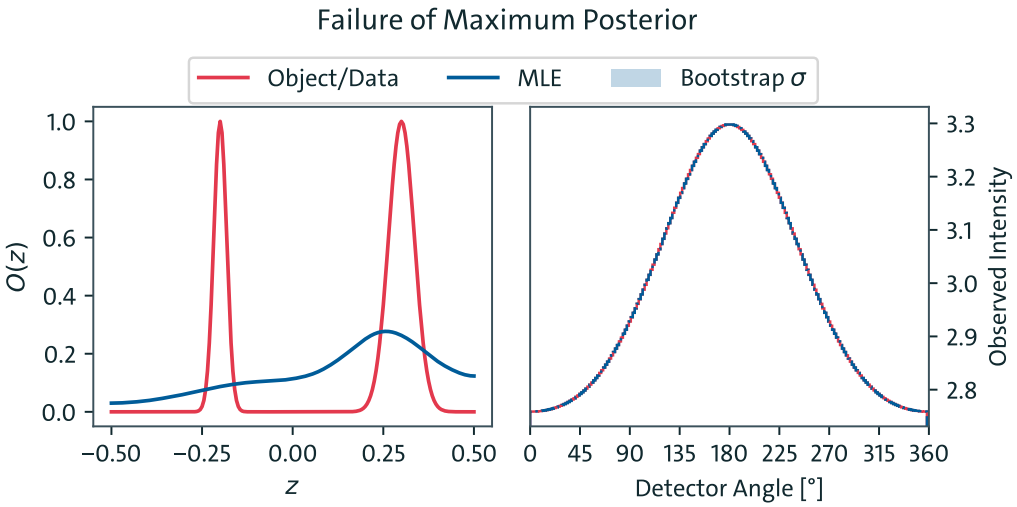

Figure 5.10. Problem Case for Iterative Maximum Posterior Estimation: Using the same technique as in figure 5.9 a multi-modal depth distribution was reconstructed. The data is again approximated well (right), though in contrast to the previous examples unrealistically low noise and large detector numbers were used to produce a best case scenario. The sharp multi-modal depth distribution was reconstructed as a highly spread out one (left).

highly spread out function. This suggests that the iterative maximum posterior approach is not suitable for image reconstruction, while the missing uncertainty estimation makes it unsuitable for quantitative work.

5.3 Generative Modeling for Single Projection Reconstruction

To gain an uncertainty estimate one can use variational Bayesian inference, where the posterior is approximated using a family of parameterized distributions. These distributions can then be fit using gradient descent with a loss function that includes both the quality of the data fit, but also the uncertainty estimate. One standard candidate for a loss function like this is the Evidence Lower Bound (ELBO, see subsection 3.2.4). The exact ELBO is generally hard to compute, but it is easy to estimate from samples. So this approach generates samples from an approximate posterior distribution to iteratively fit that distribution to the true posterior, making it a generative model. Not only do these generative models contain an uncertainty estimate, but they can also be used in ensembles, making the overall model more expressive and being able to capture multi-modal distributions (here meaning multiple different solutions that are not in a shared neighborhood, not solutions with multiple peaks).

This model is a straightforward generalization of the model used in the iterative approach for single projection reconstruction. The generative part of the model can be expressed as

$$\vec{\lambda} \sim \mathcal{N}\left[\vec{\mu}, \operatorname{diag}\left(\vec{\sigma}\right)\right]$$
$$O(z) = \operatorname{softplus}\left(\cos\left(z\vec{k}\right) \cdot \vec{\lambda}\right) := \log\left(\exp\left(\cos\left(z\vec{k}\right) \cdot \vec{\lambda}\right) + 1\right)$$

where the cosine is applied point-wise and $\vec{k}$ are the spatial frequencies corresponding to each element of the latent space vector $\vec{\lambda}$. This latent space is purely abstract and has no clear physical meaning. Because the uncertainty is already captured by the model, a bias towards low frequencies is not strictly required. Nonetheless, priors are always useful and physical processes often favor low frequency configurations. Here the prior is chosen to be much weaker than before to more clearly show the effect of the generalization.

Figure 5.11 shows the results under the same conditions as figure 5.9, with an ensemble of 100 models which differ only by initializations of the parameters $\vec{\mu}$ and $\vec{\sigma}$. The high frequency oscillation of the 1σ uncertainty interval of the reconstructed distribution can be attributed

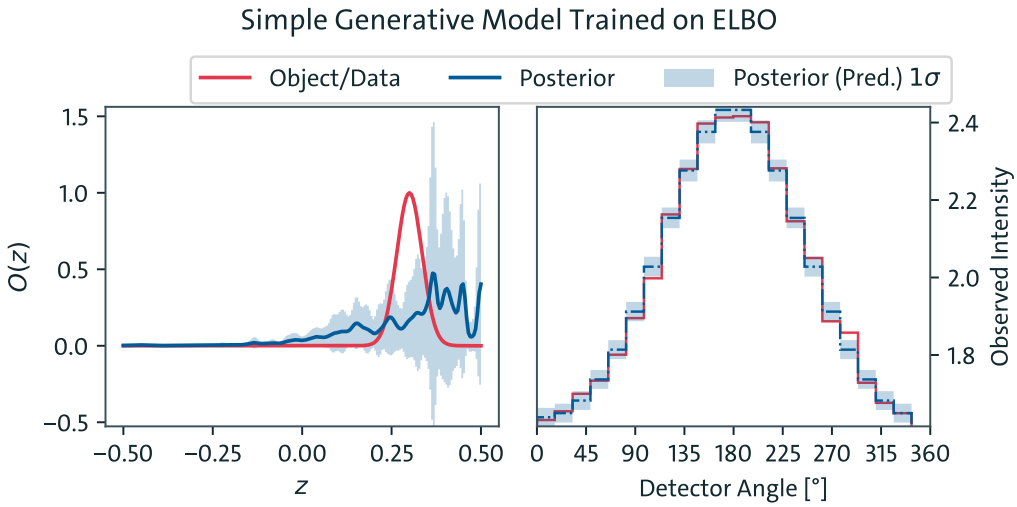

Figure 5.11. Generative Model Reconstruction from Single Projection: A generative Model using independent and normally distributed parameters in frequency space is tested under the same parameters as that of figure 5.9. The generative model ensemble captures the uncertainty in the actual shape of the distribution.

to the small ensemble size, which insufficiently probes the parameter space, larger ensembles would provide smoother results but are computationally more expensive. Nonetheless, one can see (Figure 5.11, right) that the generative model captured the uncertainty of the shape of the tracer distribution which experiments with the maximum posterior also indicated. This is further evidence that the data from only one projection angle but with multiple detectors carries insufficient information to reconstruct a depth distribution irrespective of the reconstruction technique.

5.4 Generative Modeling for Tomosynthesis

When one has access to multiple projection angles the data has the form of a discrete sinogram. This process is inherently two dimensional, so the reconstruction cannot be performed along the z dimension alone, but on a two dimensional $x-z$ slice through the object. The reconstructions in this section will thus be two dimensional images rather than depth curves. The next subsections will first outline the model that is used, along with properties of it that would be expected a priori. Afterwards several tests will be performed using that model. These evaluations will attempt to show that the reconstruction is successful and the uncertainty estimation well calibrated, further, that the the model is robust against the data imperfections which are not included in the forward model and is able to quantify the resolution of the spatial reconstruction.

5.4.1. Imaging Function

As discussed in subsection 5.1.1, the imaging equation for XRF CT is

$$S(\mu_F) = \int_0^\infty dE_{\text{in}} \int_0^\infty dt\; I_0 e^{-\int_0^t dt' \mu_T} \mu_F e^{-\int_0^\infty dt' \mu_T} \frac{A_D}{4\pi {r_D}^2} \tag{5.9}$$

$$\to S_i = \left(\int_0^\infty dE_{\text{in}}\; I_0 e^{-\int_0^{t_j} dt' \mu_T} e^{-\int_0^\infty dt' \mu_T} \frac{A_D}{4\pi {r_D}^2} \Delta t\right)_{ij} (\mu_F)_j =: \mathbf{P}\mu_F \tag{5.10}$$

where S denotes the signal strength (expectation value of recorded fluorescence events) and μ_F the linear fluorescence coefficients, which is itself linear with the tracer density. The former is the observable, the latter the inference target. Equation (5.10) shows a discretized

form where each j index corresponds to a position in space and each i index to a projection angle.

Ignoring second order effects like the μ_F contribution to μ_T, this expression is not only differentiable in μ_F, but linear. The coefficients depend on the μ_T map (which results from the shape, density and elemental makeup of the object and might be available from simultaneous spectral CT) and the positions of the detectors. The relation between the tracer concentration at each position and the observable expectation value for the number of recorded photons is then linear, fulfilling the differentiability requirement for generative modeling (see section 3.3).

When using this imaging function in the fitting of a generative model is is necessary to compute the sinogram of a given tracer distribution at each step. This means that optimizing the speed of sinogram calculation is crucial, as it usually takes thousands of steps to fit such a model. Additionally one needs to optimize for memory size of the objects involved at each step of the computation. The full integrand of equation (5.9) would be a rank 5 tensor before even considering the potential energy dimension. With 10 detectors, 40 angles, 100 offsets and a 50×50 map for the tracer distribution, this intermediate tensor would have 10^8 entries. We additionally require around $10 - 100$ samples for stability, so with 16 bytes per number the direct representation of the integrand would reach several to tens of gigabytes in size. But for training one needs to propagate gradients further increasing memory needs.

To help with both issues one can use the linearity of the discretized form of equation (5.10)

$$S(\mu_F) = \mathbf{P}\mu_F$$

where $\mathbf{P}$ is a (3, 2) tensor mapping $\mathbb{R}^{50\times50} \to \mathbb{R}^{10\times40\times100}$ in the example above. The first important realization is that, in the linear low concentration regime, $\mathbf{P}$ does not depend on μ_F and thus can be precomputed, which solves the speed issue. The next is that $\mathbf{P}$ is sparse, meaning most entries are 0. this allows storage in the form of a list of non-zero indices along with the corresponding weights rather than having to store the full tensor which solves the memory size issue. Note that this optimization does not work if μ_T is also an inference target because $\mathbf{P}$ has an explicit dependence on μ_T.

To calculate the projection matrix $\mathbf{P}$ one first needs to find the grid cells, which are irradiated by the beam, for any given projection, as well as the path length through each cell, denoted by Δt in equation (5.10). Next, one needs to find how much of the fluorescence emitted by a grid cell hits the detector (the solid angle term $\frac{A_D}{4\pi {r_D}^2}$ in equation (5.10)). Lastly, one needs to find how much the beam is attenuated on the way to the cell and the fluorescence from the cell to the detector (attenuation terms $e^{-\int_0^{t_j} dt' \mu_T} e^{-\int_0^{\infty} dt' \mu_T}$ in equation (5.10)).

The first problem is purely geometric, it does not depend on the distribution whatsoever, only on the offsets and angles used in the measurement. It can be solved by implementing the Siddon algorithm [71] which yields a sparse $(2, 2)$ tensor representation of $\mathbf{P}$ with the indices representing ((offset, angle), (x index, z index)). The second problem is also geometric. The combination of depth along the beam (t value), angle, offset and detector is sufficient to find the distance to the detector using the law of cosines. This extends the tensor $\mathbf{P}$ to ((offset, angle, detector), (x index, z index)) which is the final form.

In the following discussion the attenuation component is ignored in the reconstruction models, which is reasonable for higher energy lines like iodine K-lines, objects of mouse size and detectors close to the beam axis [17]. This is partly to keep comparability to the previous examples and partly to isolate possible error sources. The expected effect of ignoring this would be radial suppression/elevation of reconstructed tracer levels depending on whether the detector is in the forward or backward direction.

5.4.1.1. Spatial Model

The other component such a generative model needs is a mapping between the unconstrained latent space and the inference target. As discussed in section 3.3 this mapping must be bijective and have known logarithmic Jacobi determinants.

While it would be possible to work with individual pixels, similar to the attempt to reconstruct depth from a single offset and angle, it is generally beneficial to include spatial correlations in the model. In general it makes sense to formulate a prior that favors low gradients, meaning that high gradients in the reconstruction require meaningful evidence in the data. A class of models which has this property are Gaussian processes [69]. Such processes can be understood as functions $t \mapsto X(t)$ where $X(t)$ is a random variable such that $\vec{X}_n = (X(t_i))_{i \leq n}$ is a multivariate Gaussian. For generative models it makes sense to consider primarily the case where $\vec{X}_n = \mathbf{T}(t_1, \dots, t_n)\vec{\lambda}$ where $\vec{\lambda} \sim \mathcal{N}\left[\vec{\mu}, \mathrm{diag}\left(\sigma\right)\right]$, meaning $\vec{X}$ is generated by linearly transforming Gaussian white noise. $\mathbf{T}_{ij}$ indicates how much $X(t_i)$ and $X(t_j)$ covary. As discussed before our application needs $\log\det\mathbf{T}$ to be computed in an efficient manner.

This work focuses on the first order auto-regressive process AR(1) [70] with even spacing (i.e. $t_i - t_j = \Delta t(j-i)$). Here one has the form

$$\mathbf{T}_{ij} = \begin{cases} \sqrt{1-\varphi^2}\,\varphi^{|i-j|}, & j > 0 \\ 0, & j \leq 0 \end{cases}$$
$$\varphi = e^{-\Delta t/\tau}$$

where τ gives the coherence length. As one uses even spacings, φ itself can be treated as the "smoothness" parameter and a prior can be formulated for it (favoring longer correlation lengths/stronger correlations). For this prior a beta distribution $\mathrm{Beta}[N, 1]$ ([1]) with strong bias towards low values of φ is used. φ is then treated as a free parameter of the model and fit to the data. AR(1) is a one dimensional process, in order to obtain a 2D spatial correlation (for a slice through the object) the process is applied independently along the x and z axes. Using this approach only axis-aligned correlations are induced. One would expect this to result in stripes aligned with the axes in the reconstruction. The other issue with it is that the variance only converges to a stable value after a number of steps proportional to $(1-\varphi)^{-1}$, meaning for smooth objects there might be a noticeable difference between the top left and bottom right area of the reconstruction. AR(1) is still used, because its parameter φ has a direct relation to the coherence length, can be applied in real space and has $\log\det\mathbf{T} = 0$ (meaning it is also invertible by construction).

Some alternatives to AR(1) which are fit for this application do exist, like the Matérn covariance which is diagonal in Fourier space and weak suppression of high frequencies. This allows the logarithmic determinant to be calculated efficiently and in a numerically stable way.

So summarize, the model has the following parameters:

- φ - a scalar in $(0, 1)$ giving the decay of the covariance from pixel to pixel
- $\vec{\mu}$ - a 2D tensor indicating how much the tracer concentration typically increases from one pixel to the neighbor
- $\vec{\sigma}$ - a positive 2D tensor indicating how much that increase varies. This converges to 0 as the data becomes strongly informative.

This spatial model provides us with candidate tracer distributions and the approximate model entropy as well as the mean log prior. To fit the model to the data one also requires access to the log likelihood to observe the data given a tracer distribution.

5.4.1.2. Application to Synthetic Data

To test whether the method works at all, a target tracer distribution is chosen and synthetic data is generated using the same sinogram calculation as is used during reconstruction.

This also provides a reliable ground truth against which the uncertainty estimation can be evaluated quantitatively. The measurement function used 31 projection angles between 0 and 360°, 1 mm offset steps, 5 detectors per side, evenly spaced in $[30°, 150°]$ and $[-30°, -150°]$ respectively and $\approx 2 \cdot 10^9$ incident photons. The object is constructed to emulate a phantom consisting of a cylinder filled with weak tracer solution and three inserts of varying concentration between 0 and $3\,\mathrm{mg\,cm^{-3}}$. This corresponds to slightly stronger data than that in [21] and also excludes the attenuation from the data generation.

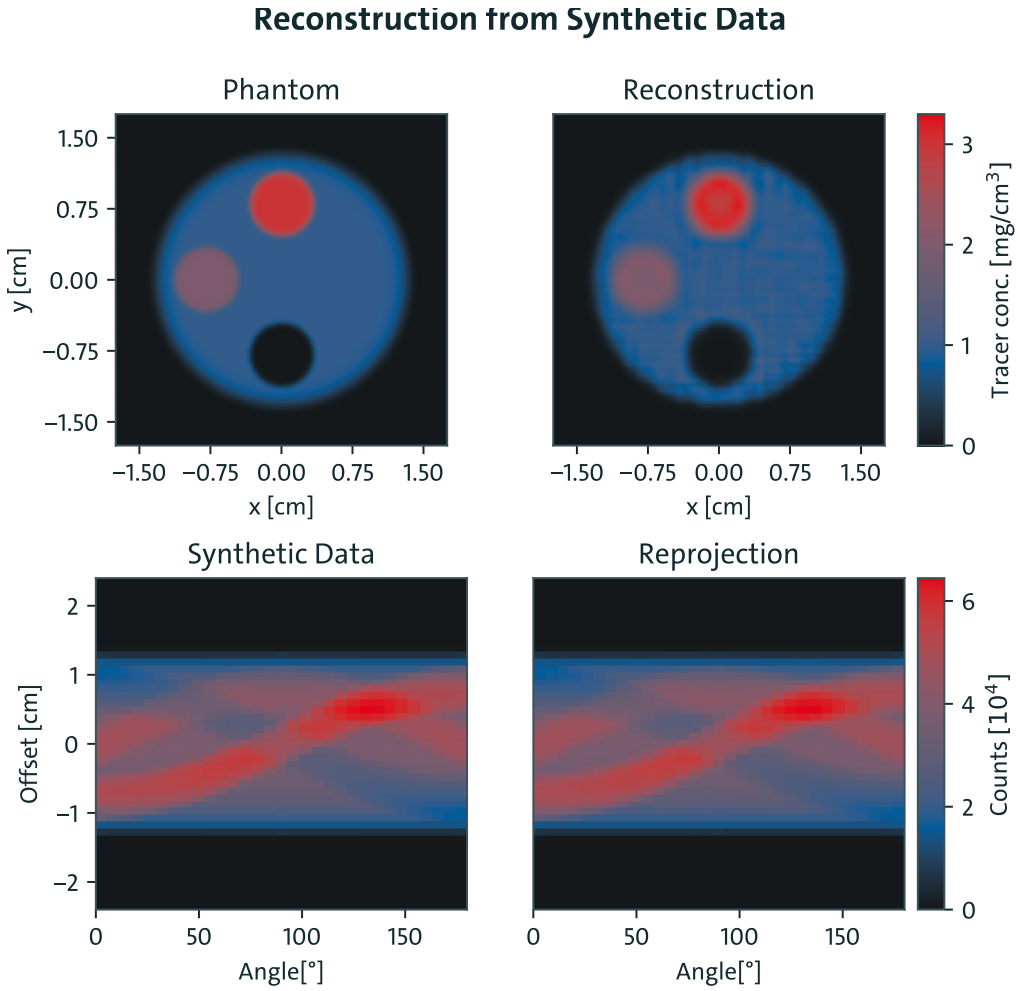

Figure 5.12. Reconstruction from Synthetic Tomography Data: The model reconstruction of the tracer distribution (top right) closely matches the ground truth (top left) and the data (bottom left) is almost perfectly described by the model (bottom right). The stripe pattern appears in the reconstruction as an artifact of the AR(1) process for spatial coherence. The sinograms (bottom row) are summed over all detectors and the reconstructions show the posterior mean.

The results can be seen in figure 5.12. The reconstruction closely matches the ground truth, except for the line artifacts from the AR(1) coherence function. The reprojection back to data space (i.e. the posterior predictive mean) is indistinguishable from the data. However, this is only the case and also only desirable if the data has very low noise. Otherwise, the reprojection should appear like a denoised version of the data. Visually inspecting the reconstruction and comparing to the ground truth serves as a qualitative assessment of the success of the model for reconstruction. With that one can conclude that the reconstruction is successful.

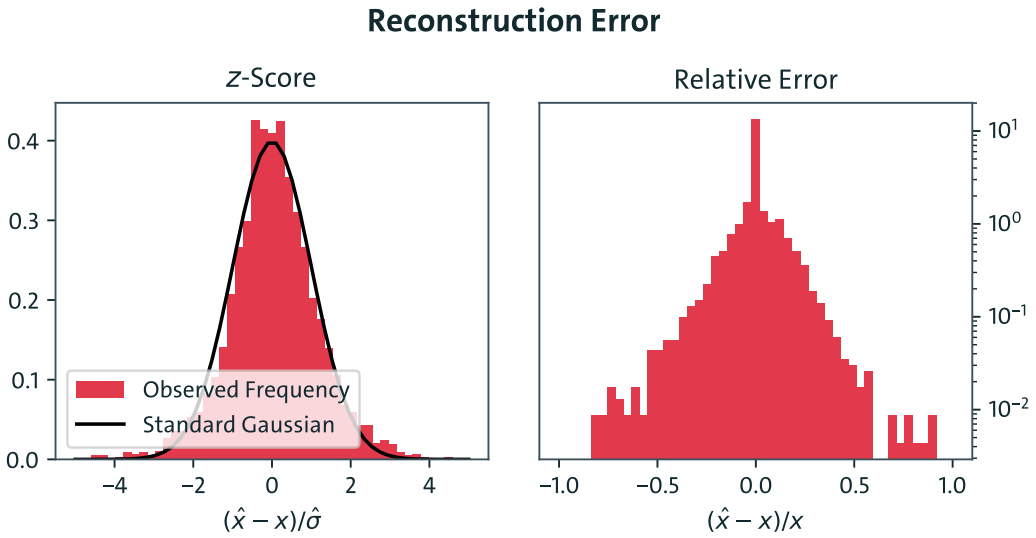

Figure 5.13. Reconstruction Error from Synthetic Data: Left - histogram of the z-score of the ground truth value from the posterior mean and standard deviation. The z-score is approximately Gaussian. Right - logarithmic histogram of the reconstruction error relative to the ground truth value. The distribution appears composite with a component of perfect reconstruction, a Gaussian component (paraboloid in the log histogram) and a component dominating in the tails.

To evaluate the reconstruction quantitatively one can study the distribution of the reconstruction errors over the pixels inside the object (here defined as pixels where the reconstruction yields $0.01\,\mathrm{mg\,cm^{-3}}$ or higher). The z-score on the left of figure 5.13 shows a Gaussian distribution as desired, indicating that the uncertainty estimate is well calibrated. However, the relative error distribution (right) shows an interesting composite behavior. The first component appears to be a δ-peak at 0, which corresponds to areas of near perfect reconstruction. The second component appears to be vaguely Gaussian, seen as a paraboloid shape between -25% and 25% relative error. However, there is also a component of very significant relative error, particularly in overestimations. These predominantly come from areas on the edges where the smoothness bias induces larger relative deviations, but as these are very infrequent, they are not considered to be problematic.

5.4.1.3. Application to Monte Carlo Data

Showing robustness against effects that are not modeled requires a more complete simulation of the imaging process. To do this the same phantom is now implemented in a GEANT4 simulation with the same projections being acquired. The tracer element is chosen to be iodine. The beam parameters were chosen to emulate the P21 beamline of PETRA III at DESY [72], meaning it is a 1 mm box beam at monochromatic 53 keV. The phantom is constructed as a water cylinder with tracer added to various areas. This means that attenuation through the water would now be simulated in the data acquisition, but not modeled in the forward process for the reconstruction model. Additionally, the beam shape is also not represented in the forward model. These two effects should produce a slight bias towards tracer at the edges of the object and a slight size increase of $\mathcal{O}(2\,\mathrm{mm})$ diameter, because the beam size manifests as a box-blur on the offset component of the sinogram. Only $3 \cdot 10^8$ photons per scan position, corresponding to $\approx 100\,\mathrm{mGy}$, were used this time, which is a more realistic real world value, but also means that the Poisson noise is approximately 3 times stronger in relative terms. It is worth noting that the synthetic phantom before and this phantom do not necessarily match exactly, therefore no evaluations that require a perfect ground truth reference are performed.

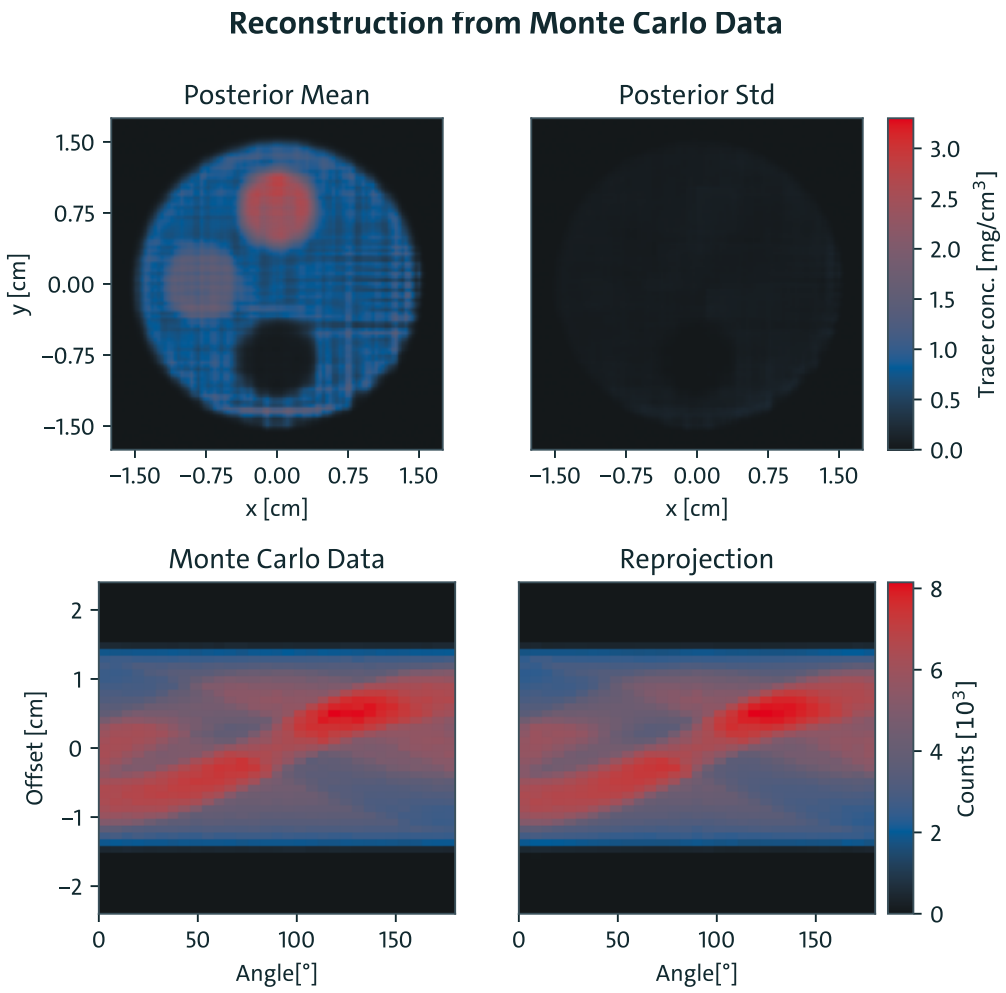

Figure 5.14. Reconstruction from Monte Carlo Tomography Data: The data (bottom left) consists of a sinogram, which is obtained through Monte Carlo simulation using GEANT 4, $3 \cdot 10^8$ instead of $\approx 2 \cdot 10^9$ photons per point and attenuation from water in the data generation. The reconstruction of the tracer distribution (top left) is noisier than in figure 5.12. The error estimate (top right) is relatively small compared to the reconstructed concentrations. The expected signal intensities calculated from the reconstructed distribution (bottom right) matches the data almost perfectly.

The results can be seen in figure 5.14. The expected effects of top left corner bias, stripe artifacts and a slight blur, all occur. The stripe artifacts in particular appear amplified compared to the previous experiment shown in figure 5.12, which is probably due to the likelihood function having weaker gradients, as those scale with the photon number. The reprojection still matches the data very closely and the reconstruction is very usable nevertheless.

5.4.1.4. Resolution Phantoms

To test the reconstruction resolution, a phantom is simulated that consisted of alternating lines of high concentration ($0.1\,\mathrm{mg\,cm^{-3}}$), low concentration ($0.025\,\mathrm{mg\,cm^{-3}}$) iodine solution and pure water. The lines has decreasing widths from 4 mm to 0.1 mm. To avoid mistaking line artifacts for reconstructed features the object is slightly rotated with respect to the cardinal axes. The smallest reconstructed line or gap is somewhere between 0.8 mm to 1.0 mm, depending on what one considers to be a distinct line. This is comparable to the resolution limit induced by beam size and scan step size. The reconstructed image is shown in figure 5.15

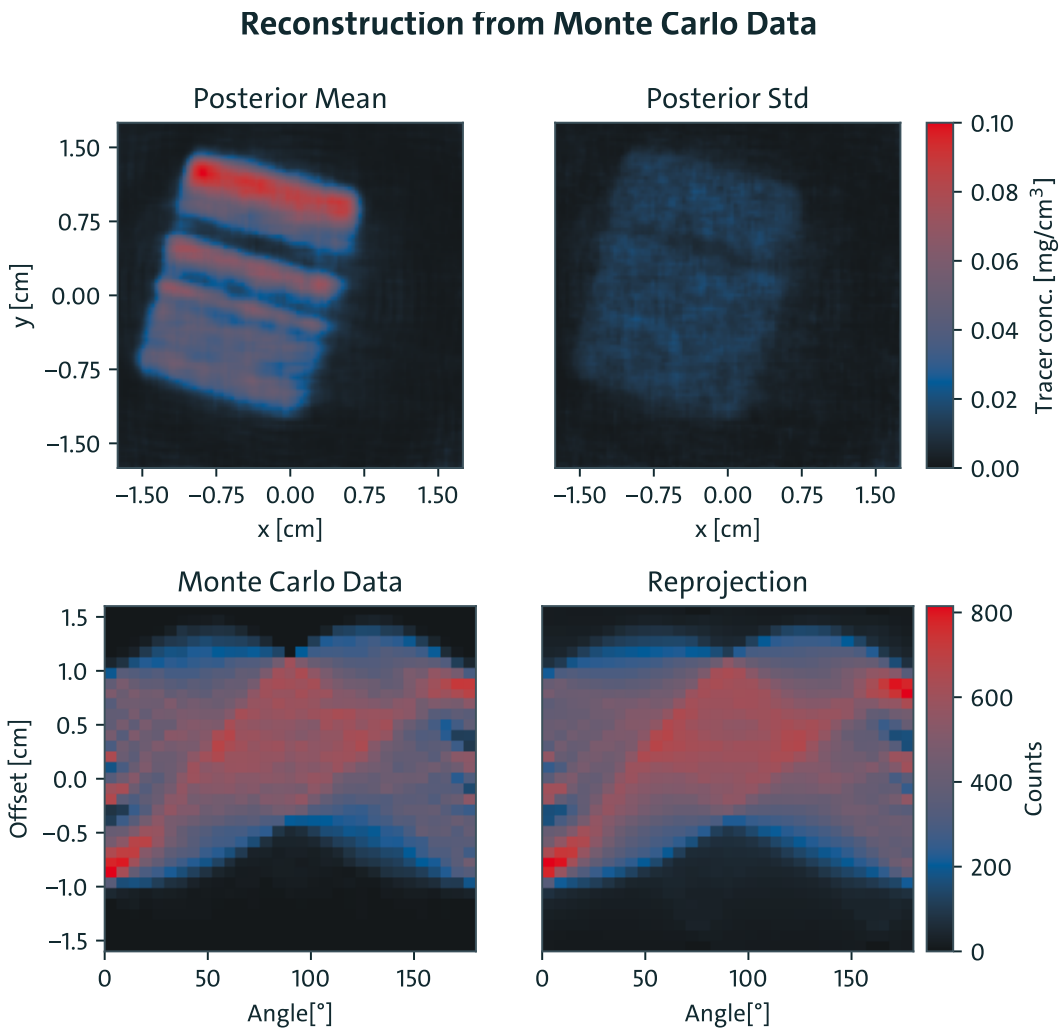

Figure 5.15. Monte Carlo Resolution Test: A simple line phantom is reconstructed using Monte Carlo simulated data. This method is resilient against the factors like attenuation that were ignored in the model, but are present in the real world and physics simulations. The resolution suffered slightly, however, one of the effects represented in the simulation and not the model is a beam size of 1 mm which is expected to contribute a meaningful box blur.

Another resolution phantom is tested with synthetic data to obtain a reference without perturbations of the beam size. Here the smallest feature is 0.6 mm to 0.8 mm, slightly smaller than the step size of the scan. This aligns well with the largest component in Fourier space with a significant contribution around 1.5 mm^{-1}. The results can be seen in figure 5.16.

5.4.1.5. Impact of Detector Number on Resolution

Using the synthetic resolution test as a baseline, the same process is repeated with the same beam and tracer configuration with one, five, ten and 100 detectors. The impact on resolution is minimal with one detector already yielding approximately 1.0 mm resolution, while five, ten and 100 all yielded approximately 0.7 mm resolution. However, the additional data made the reconstructions much less noisy. This can be seen in figure 5.17.

5.4.1.6. Undersampling

In contrast to standard back-projection, this method does not require every pixel to be hit in every projection. This means that it is possible to undersample the sinogram, recording fewer angles and offsets than one might expect to be necessary. In back-projection the rule of thumb is that the sinogram should have about the same number of pixels as the final

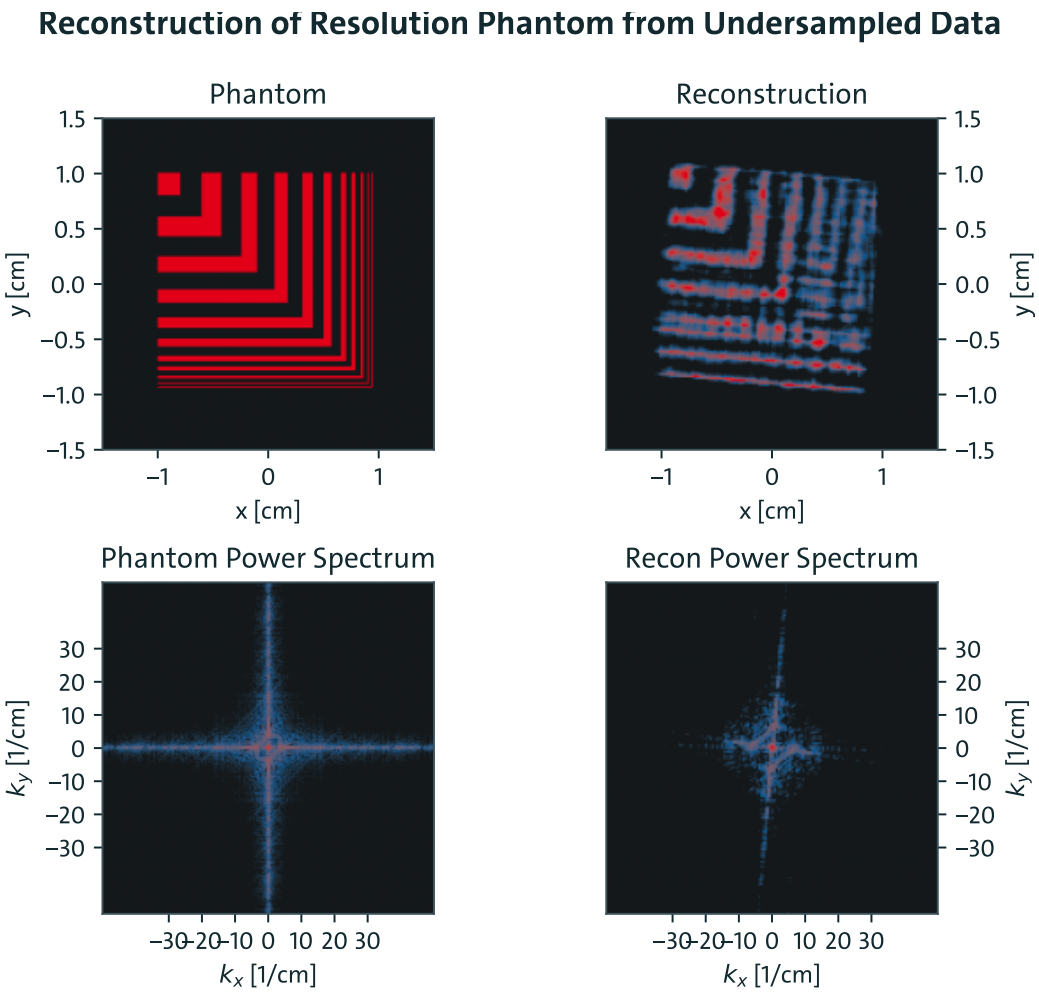

Figure 5.16. Synthetic Resolution Test: The reconstruction of another line phantom (top) along with its Fourier domain representation (bottom) is used to get another estimate for the reconstruction resolution. The tilt is induced artificially to avoid grid aligned artifacts being confused for reconstructed features. The power spectrum is represented by $\log(1 + |P(k)|)$ to emphasize which regions contribute meaningfully to the image. The finest lines of the actual phantom are only one pixel wide, thus the full Fourier space contributes.

reconstructed image, which can be seen by dimensional analysis. For a 3 cm object to be reconstructed at 1 mm resolution, this would suggest $\sqrt{2} \cdot 3\,\text{cm}/1\,\text{mm} \approx 50$ offsets (in order to capture the diagonal direction of the image with sufficient resolution) and ≈ 30 angles. When reconstructing from fewer angles or offsets, the result is expected to be roughly equivalent to a blur corresponding to a cutoff in frequency space. With 25 angles and 15 angles one would expect a cut to the lowest frequency quadrant of frequency space as that would preserve the dimensionality. A rough analog would be a Gaussian blur with $\sigma = 2$ pixels.

Performing the reconstruction using these parameters, one finds that this heuristic also applies to this reconstruction algorithm. The results can be seen in figure 5.18. Thus, when keeping the number of offsets constant and when tolerating a resolution of ≈ 3.5 mm and for a 3 cm object, one can drastically reduce the acquisition time, especially in the regime where motor movement contributes significantly to it.

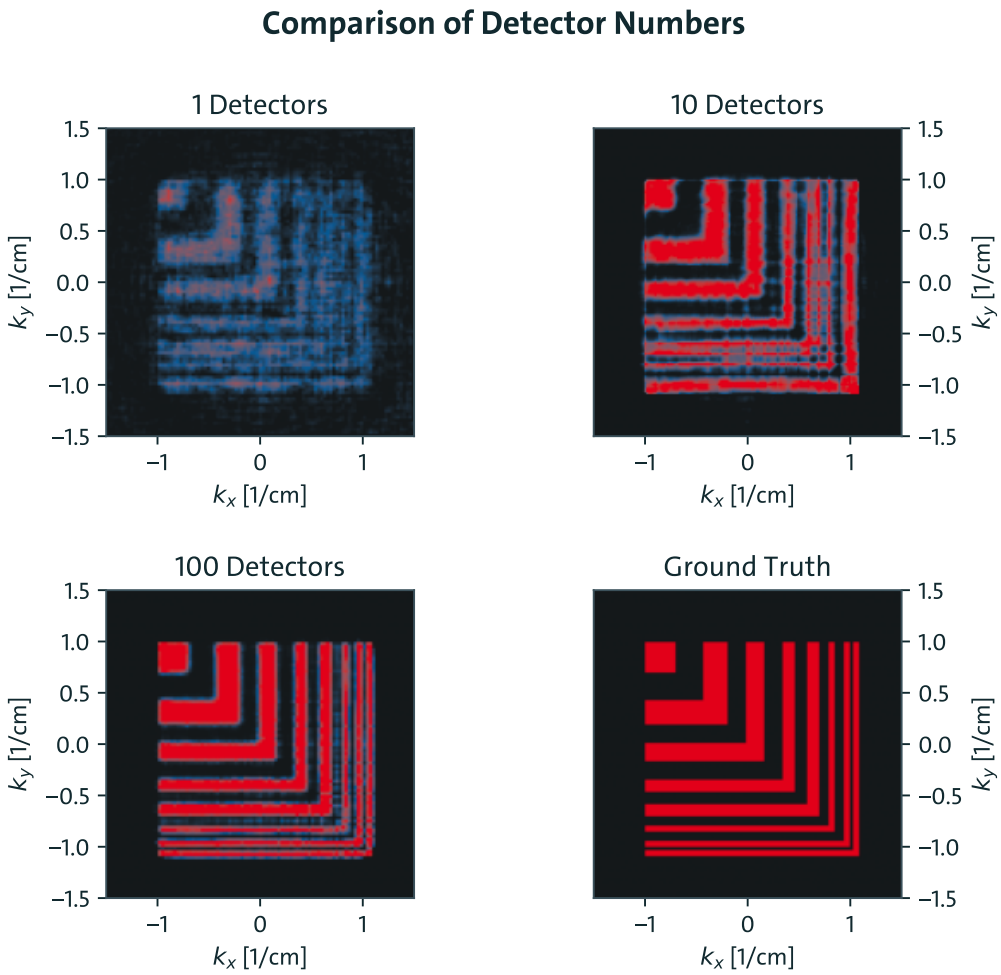

Figure 5.17. Impact of Detector Number: The same reconstruction as in figure 5.16 is performed using one (top left), ten (top right) and one hundred (bottom left) detectors. The ground truth (bottom right) is shown for comparison. The noisiness of the images decreases as the number of detectors increases.

5.5 Summary

Several techniques for 3D reconstruction of XRF images were evaluated. While projection from a single projection proved infeasible and tomography/tomosynthesis struggles to provide reliable error quantification, generative models were more successful. Generative models fit to tomosynthesis/tomography style data with multiple detectors and can perform reconstruction of the 3D tracer distribution matching the ground truth qualitatively and having well calibrated errors quantitatively. The Monte Carlo-based experiments showed that the reconstruction can be done with 10^{10} photons per square millimeter shadow area which corresponds to doses seen in in vivo conform 2D scans [17]. Currently, one of the critical limitations to in vivo conform tomography scans are the hardware-imposed time overhead for moving to all required positions. This could be solved by either reducing the number of needed projections and potentially angular range needed (not studied in this work), or by generalizing to continuous scan operations. Both of which can be implemented in the current modeling approach, but further work would be needed to show successful reconstructions. Future work should also experiment with covariance kernels other than AR(1) and attenuation modeling to avoid the artifacts discussed above.

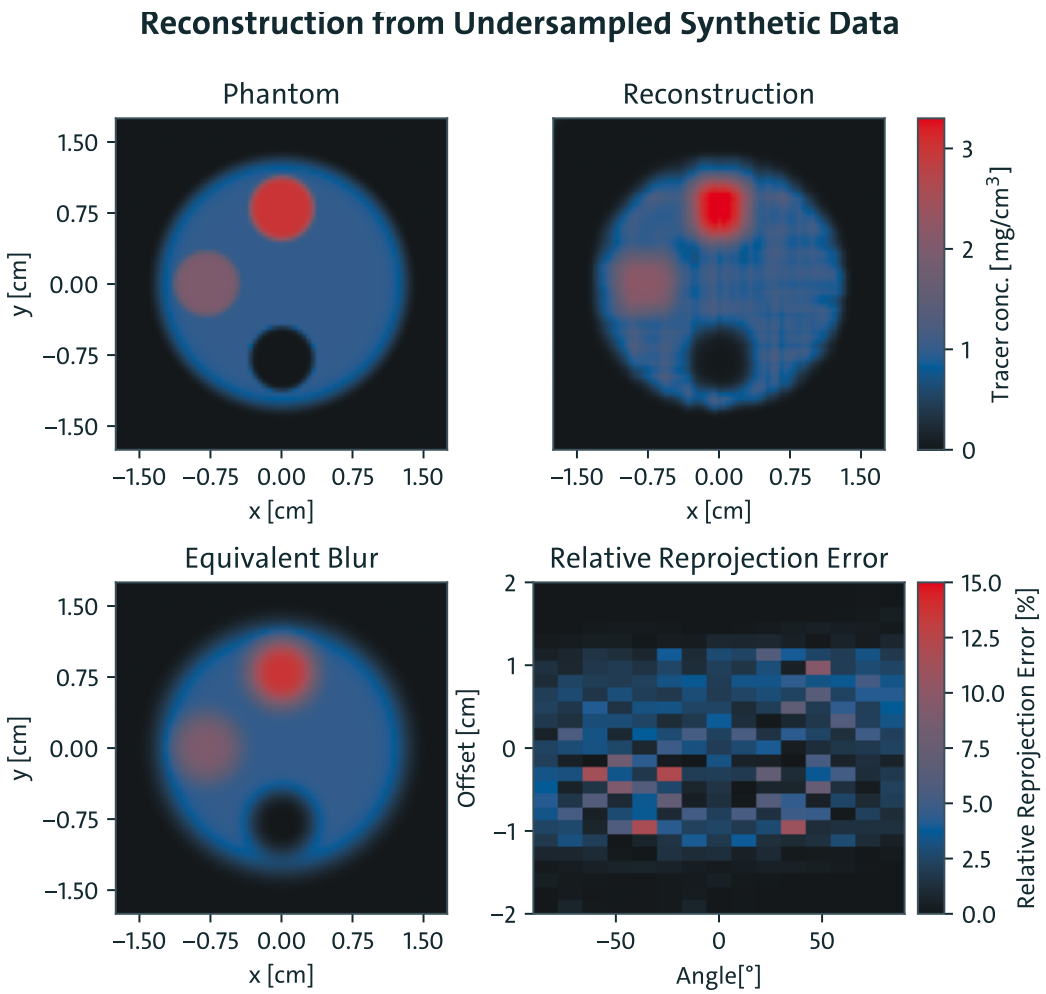

Figure 5.18. Impact of Undersampling: Using fewer projection angles and offsets reduces the resolution of the reconstruction in a way roughly comparable to a blur with equivalent loss of information. However, the less informative data also makes the biases of the spatial filter (vertical and horizontal lines) more apparent. Only 15 angles were used in the reconstruction.

CHAPTER 6

CONCLUSION

Bayesian statistics of Poisson variables was used to develop techniques to help plan future experiments on the basis of Monte Carlo simulations. The same tools were used to study two common methods for extracting the fluorescence signal from spectra, direct subtraction and (weighted) least squares fitting. Concepts from information theory, particulary the AIC and BIC were used to describe an alternate method for testing for a tracer signal based on model comparison rather than simple null hypothesis tests.

A fully Bayesian generative modeling framework was developed which can describe both the spectral and imaging level as well as potentially unify them. This framework was used to reconstruct images acquired through Monte Carlo simulations. Using various virtual phantoms this reconstruction technique was benchmarked for in vivo conforming radiation doses.

Bayesian or partially Bayesian techniques are well-suited to various tasks in XFI, namely signal extraction from spectra, 3D image reconstruction and combining the data from multiple detectors. While full Bayesian analysis is always the most accurate and reliable, it is often sufficient and more convenient to use this approach only in parts of the analysis. The improvements over the state of the art techniques are particularly stark when the quality of the data is poor, but improvements could be shown for most cases in most parts of the analysis pipeline.

Further work is needed to explore the breadth of the techniques outlined in this work, in order to fully utilize their full potential for XFI image reconstruction. This is especially true for the generative modeling frame where significant improvements are likely possible through more sophisticated models, such as using inference in frequency space to gain access to better suited spatial covariance functions. This modeling frame should also be applied to spectral data and possibly even to the full 4D (3D scan + 1D spectrum) data.

BIBLIOGRAPHY

[1] Mark J Schervish. *Theory of statistics*. Springer series in statistics. Springer, New York, NY, 1995 edition, December 2012.

[2] Richard T Cox. Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13, 1946.

[3] Jiawei Zhang, Yuhong Yang, and Jie Ding. Information criteria for model selection. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(5):e1607, 2023.

[4] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[5] Shamim Yazdani, Akansha Singh, Nripsuta Saxena, Zichong Wang, Avash Palikhe, Deng Pan, Umapada Pal, Jie Yang, and Wenbin Zhang. Generative ai in depth: A survey of recent advances, model variants, and real-world applications. *Journal of Big Data*, 12 (1):1–43, 2025.

[6] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

[7] Ioannis Sechopoulos. A review of breast tomosynthesis. part ii. image reconstruction, processing and analysis, and advanced applications. *Medical physics*, 40(1):014302, 2013.

[8] Timothy P Szczykutowicz, Giuseppe V Toia, Amar Dhanantwari, and Brian Nett. A review of deep learning ct reconstruction: concepts, limitations, and promise in clinical practice. *Current Radiology Reports*, 10(9):101–115, 2022.

[9] Lennart R Koetzier, Domenico Mastrodicasa, Timothy P Szczykutowicz, Niels R van der Werf, Adam S Wang, Veit Sandfort, Aart J van der Molen, Dominik Fleischmann, and Martin J Willemink. Deep learning image reconstruction for ct: technical principles and clinical prospects. *Radiology*, 306(3):e221257, 2023.

[10] Christopher Aubin. *Statistical Thermodynamics: An Information Theory Approach*. John Wiley & Sons, 2024.

[11] M Jake Pushie, Ingrid J Pickering, Malgorzata Korbas, Mark J Hackett, and Graham N George. Elemental and chemically specific x-ray fluorescence imaging of biological systems. *Chemical reviews*, 114(17):8499–8541, 2014.

[12] Theresa Staufer and Florian Grüner. Review of development and recent advances in biomedical x-ray fluorescence imaging. *International Journal of Molecular Sciences*, 24 (13):10990, 2023.

[13] Valentin Viktorovich Lider. X-ray fluorescence imaging. *Physics-Uspekhi*, 61(10):980, 2018.

[14] Seongmoon Jung, Taeyun Kim, Wooseung Lee, Hyejin Kim, Hyun Suk Kim, Hyung-Jun Im, and Sung-Joon Ye. Dynamic in vivo x-ray fluorescence imaging of gold in living mice exposed to gold nanoparticles. *IEEE transactions on medical imaging*, 39(2):526–533, 2019.

[15] Seongmoon Jung, Jimin Lee, Hyungjoo Cho, Taeyun Kim, and Sung-Joon Ye. Compton background elimination for in vivo x-ray fluorescence imaging of gold nanoparticles using convolutional neural network. *IEEE Transactions on Nuclear Science*, 67(11):2311–2320, 2020.

[16] Giovanni Marco Saladino, Po-Han Chao, Bertha Brodin, Shyh-Dar Li, and Hans Martin Hertz. Liposome biodistribution mapping with in vivo x-ray fluorescence imaging. *Nanoscale*, 16(37):17404–17411, 2024.

[17] Christian Körnig, Theresa Staufer, Oliver Schmutzler, Tanja Bedke, Andres Machicote, Beibei Liu, Yang Liu, Elisabetta Gargioni, Neus Feliu, Wolfgang J Parak, et al. In-situ x-ray fluorescence imaging of the endogenous iodine distribution in murine thyroids. *Scientific reports*, 12(1):2903, 2022.

[18] Andreas Maier, Stefan Steidl, Vincent Christlein, and Joachim Hornegger. *Medical imaging systems: An introductory guide*. Springer, 2018.

[19] Ruxana T Sadikot and Timothy S Blackwell. Bioluminescence imaging. *Proceedings of the American Thoracic Society*, 2(6):537–540, 2005.

[20] STEFAN Andersson-Engels, KATARINA Svanberg, and SUNE Svanberg. Fluorescence imaging in medical diagnostics. *Chap*, 10:265–305, 2009.

[21] Christian Gabriel Körnig. *Pilot studies for quantitative 2D and 3D X-ray fluorescence imaging*. PhD thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2022.

[22] Victor Weisskopf and Eugene Wigner. Berechnung der natürlichen linienbreite auf grund der diracschen lichttheorie. *Zeitschrift für Physik*, 63(1):54–73, 1930.

[23] Antonio Brunetti, M Sanchez Del Rio, Bruno Golosio, Alexandre Simionovici, and Andrea Somogyi. A library for x-ray–matter interaction cross sections for x-ray fluorescence applications. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 59(10-11):1725–1731, 2004.

[24] Sigurdur Helgason and S Helgason. *The radon transform*, volume 2. Springer, 1980.

[25] Marleen Vonder, Monique D Dorrius, and Rozemarijn Vliegenthart. Latest ct technologies in lung cancer screening: protocols and radiation dose reduction. *Translational lung cancer research*, 10(2):1154, 2021.

[26] Wolfram Stiller. Basics of iterative reconstruction methods in computed tomography: a vendor-independent overview. *European journal of radiology*, 109:147–154, 2018.

[27] Sea Agostinelli, John Allison, K al Amako, John Apostolakis, Henrique Araujo, Pedro Arce, Makoto Asai, D Axen, Swagato Banerjee, GJNI Barrand, et al. Geant4—a simulation toolkit. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003.

[28] Walter Greiner and Joachim Reinhardt. Quantum-electrodynamical processes. In *Quantum Electrodynamics*, pages 77–241. Springer, 2009.

[29] Andrew T Young. Rayleigh scattering. *Phys. Today*, 35(1):42–48, 1982.

[30] Mitchel Weissbluth. *Atoms and molecules*. Elsevier, 2012.

[31] J Michael Hollas. *Modern spectroscopy*. John Wiley & Sons, 2004.

[32] Theresa Staufer, Christian Körnig, Beibei Liu, Yang Liu, Clarissa Lanzloth, Oliver Schmutzler, Tanja Bedke, Andres Machicote, Wolfgang J Parak, Neus Feliu, et al. Enabling x-ray fluorescence imaging for in vivo immune cell tracking. *Scientific reports*, 13(1):11505, 2023.

[33] Giovanni M Saladino, Bertha Brodin, Ronak Kakadiya, Muhammet S Toprak, and Hans M Hertz. Iterative nanoparticle bioengineering enabled by x-ray fluorescence imaging. *Science Advances*, 10(12):eadl2267, 2024.

[34] Z Hegedüs, T Müller, J Hektor, Emanuel Larsson, T Bäcker, S Haas, ALC Conceiçao, S Gutschmidt, and U Lienert. Imaging modalities at the swedish materials science beamline at petra iii. In *IOP Conference Series: Materials Science and Engineering*, volume 580, page 012032. IOP Publishing, 2019.

[35] Lídia Cunha, Ildiko Horvath, Sara Ferreira, Joana Lemos, Pedro Costa, Domingos Vieira, Dániel S Veres, Krisztián Szigeti, Teresa Summavielle, Domokos Máthé, et al. Preclinical imaging: an essential ally in modern biosciences. *Molecular diagnosis & therapy*, 18(2):153–173, 2014.

[36] T Jack. The importance of medical imaging in modern healthcare. *J Adv Med Res*, 6 (2):1000030, 2023.

[37] Siddharth Venkatraman and Capt Tej Ishaan Mehta. Overview of preclinical research. In *Translational Interventional Radiology*, pages 15–19. Elsevier, 2023.

[38] Ali Douraghy and Arion F Chatziioannou. Preclinical imaging. *Basic Sciences of Nuclear Medicine*, pages 379–413, 2010.

[39] Qingping Zhou, Tengchao Yu, Xiaoqun Zhang, and Jinglai Li. Bayesian inference and uncertainty quantification for medical image reconstruction with poisson data. *SIAM Journal on Imaging Sciences*, 13(1):29–52, 2020.

[40] Kenneth Lange. Overview of bayesian methods in image reconstruction. In *Digital image synthesis and inverse optics*, volume 1351, pages 270–287. SPIE, 1990.

[41] Pakshal Bohra, Thanh-an Pham, Jonathan Dong, and Michael Unser. Bayesian inversion for nonlinear imaging models using deep generative priors. *IEEE Transactions on Computational Imaging*, 8:1237–1249, 2023.

[42] Mikael Jensen and Jens E Wilhjelm. X-ray imaging: Fundamentals and planar imaging, 2006.

[43] Aaron So and Savvas Nicolaou. Spectral computed tomography: fundamental principles and recent developments. *Korean journal of radiology*, 22(1):86, 2020.

[44] William C Thompson and Edward L Schumann. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. In *Expert evidence and scientific proof in criminal trials*, pages 371–391. Routledge, 2017.

[45] Edmund A Cornish and Ronald A Fisher. Moments and cumulants in the specification of distributions. *Revue de l'Institut international de Statistique*, pages 307–320, 1938.

[46] Ali Mohammad-Djafari. Entropy, information theory, information geometry and bayesian inference in data, signal and image processing and inverse problems. *Entropy*, 17(6):3989–4027, 2015.

[47] Lars Ruthotto and Eldad Haber. An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2):e202100008, 2021.

[48] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[49] Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semi-supervised learning with normalizing flows. In *International conference on machine learning*, pages 4615–4630. PMLR, 2020.

[50] Abhinav Agrawal, Daniel R Sheldon, and Justin Domke. Advances in black-box vi: Normalizing flows, importance weighting, and optimization. *Advances in Neural Information Processing Systems*, 33:17358–17369, 2020.

[51] Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[52] K Kumar, M Fachet, and C Hoeschen. Pp23. 04 x-ray fluorescence imaging using bragg diffractionbased focusing with graphite optics: A simulation study. *Physica Medica*, 125:103826, 2024.

[53] RH Redus, AC Huber, P Bennett, E Kostamo, and PT Törmä. Improving sensitivity of silicon detectors for x-ray spectroscopy. In *2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 1–4. IEEE, 2022.

[54] Florian Blumendorf. *Background reduction for XFI with human-sized phantoms*. PhD thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2019.

[55] Karel Proesmans, Jannik Ehrich, and John Bechhoefer. Finite-time landauer principle. *arXiv preprint arXiv:2006.03242*, 2020.

[56] Staphord Bengesi, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers. *IEEe Access*, 12:

69812–69837, 2024.
[57] Åke Björck. Least squares methods. *Handbook of numerical analysis*, 1:465–652, 1990.
[58] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.
[59] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020.
[60] Gert Van Gompel, Katrien Van Slambrouck, Michel Defrise, K Joost Batenburg, Johan De Mey, Jan Sijbers, and Johan Nuyts. Iterative correction of beam hardening artifacts in ct. *Medical physics*, 38(S1):S36–S49, 2011.
[61] MA Jaswon. A review of the theory. *Basic Principles and Applications*, pages 13–40, 1984.
[62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
[63] Per Christian Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998.
[64] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.
[65] Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical science*, 11(3):189–228, 1996.
[66] Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.
[67] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 2006.
[68] Marc G Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2(Dec):299–312, 2001.
[69] David JC MacKay et al. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168:133–166, 1998.
[70] Ngai Hang Chan and Lanh Tat Tran. On the first-order autoregressive process with infinite variance. *Econometric Theory*, 5(3):354–362, 1989.
[71] Robert L Siddon. Fast calculation of the exact radiological path for a three-dimensional ct array. *Medical physics*, 12(2):252–255, 1985.
[72] Martin v. Zimmermann, Oleh Ivashko, Fernando Igoa Saldaña, Jiatu Liu, Philipp Glaevecke, Olof Gutowski, Rüdiger Nowak, Katharina Köhler, Björn Winkler, Andreas Schöps, Horst Schulte-Schrepping, and Ann-Christin Dippel. P21.1 at petra iii – a high-energy x-ray diffraction beamline for physics and chemistry. *Journal of Synchrotron Radiation*, 32(3):802–814, April 2025. ISSN 1600-5775. doi: 10.1107/s1600577525002826. URL http://dx.doi.org/10.1107/S1600577525002826.
[73] Aleksandr Mihajlovič Denisov. *Elements of the theory of inverse problems*, volume 14. VSP, 1999.
[74] Csaba Olasz. Addressing discretization artifacts in tomography by accessing and balancing pixel coverage of projections. In *International Workshop on Combinatorial Image Analysis*, pages 182–205. Springer, 2024.

[75] Andreas Kirsch et al. *An introduction to the mathematical theory of inverse problems*, volume 120. Springer, 2011.

[76] R Jenkins, R Manne, R Robin, and C Senemaud. Iupac—nomenclature system for x-ray spectroscopy. *X-Ray Spectrometry*, 20(3):149–155, 1991.

[77] Jorge J Moré and Danny C Sorensen. Newton's method. Technical report, Argonne National Lab.(ANL), Argonne, IL (United States), 1982.

[78] David Halliday, Robert Resnick, and Kenneth S Krane. *Physics, Volume 2*. John Wiley & Sons, 2010.

EIDESSTATTLICHE VERSICHERUNG

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Ich versichere weiterhin, dass dieses gebundene Exemplar der Dissertation und das in elektronischer Form eingereichte Dissertationsexemplar (über den Docata-Upload) und das bei der Fakultät (zuständiges Studienbüro bzw. Promotionsbüro Physik) zur Archivierung eingereichte gedruckte gebundene Exemplar der Dissertationsschrift identisch sind.

I, the undersigned, hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated. If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

I further declare that this bound copy of the dissertation and the dissertation submitted in electronic form (via the Docata upload) and the printed bound copy of the dissertation submitted to the faculty (responsible Academic Office or the Doctoral Office Physics) for archiving are identical.

Hamburg, den 2026-05-27 _____

Jonas Kreidelmeyer

APPENDIX A

IMPLEMENTATION DETAILS

A.1 Framework Design

A generative model works by sampling some base distribution and transforming the output. This framework implements these using two main concepts, parameters and transforms. Transforms are functions of parameters, which can be parameterized themselves. For example an affine function $\vec{f}(\vec{x}) = \mathbf{M}\vec{x} + \vec{b}$ might have the all the components of $\mathbb{M}$ and $\vec{b}$ as parameters. Transforms are also used to enforce constraints, for example the softplus function $f(x) = \log(1 + \exp(x))$ can be used to guarantee positivity. Parameters come in two flavors, base parameters and distributed parameters. The former are the objects that gradient descent is performed on, they are tensors. The latter represent samples from distributions, these can either be from base distributions like a standard normal, or outputs of sub-models. This last option is what allows for arbitrarily complex hierarchical models.

Because the goal is to fit the model to data using the ELBO loss using gradient descent, a differentiable and computationally efficient way to estimate the model entropy must be provided. The relevant parts of the model are the entropy of base distributions and the log determinant of the Jacobians of the transforms. This latter part could in theory be estimated using implicit differentiations, but that would require calculating n^2 gradients where n is the dimension of the transformed object, the determinant of an $n \times n$ matrix in a differentiable manner and finally passing gradients through both during training, which is extremely expensive. Instead transforms are carefully chosen to admit closed form expressions of the log Jacobi determinant given the transform parameters. This restriction is what ultimately makes the modeling approach of this work computationally feasible.

Priors implemented as log-likelihood functions which are optionally attached to parameters. Whenever a parameter is sampled, the prior log-likelihood is calculated and passed up through the model together with the Jacobi determinants and summed up at each stage. This allows for priors to be attached to arbitrary combinations of parameters at different levels of the model, which represents a flexible way to add parameter level domain knowledge. Other domain knowledge is implemented in the choice of transforms by enforcing constraints with them or favoring specific types of outputs.

This framework was built on top of pytorch ([62]) to handle vectorized computations and gradient descent. Parameters and transforms are implemented as torch Modules with the corresponding forward and call functions and exposing their learnable parameters to the optimizer backend.

A.2 3D Reconstruction Model

From domain knowledge we know that the tracer concentrations are positive definite and typically spatially smooth almost everywhere. This suggests that a sensible model would for example be generating Gaussian white noise, then smoothing followed by a softplus

transform to ensure positivity. The base model parameter are then the μ and σ of the white noise and the correlation parameter of the $AR(1)$ smoothing ϕ.

The concentration point estimate is encoded in the μ and the uncertainties in σ. A prior is formulated on ϕ to favor higher smoothness, in this case $\beta[N, 1]$ with $N \gg 1$ was used, the specific value of N was found to have minimal impact on the final reconstruction. ϕ was initialized with 0.9 corresponding to a full width half maximum correlation of 13 pixels or 0.6 mm. To help with convergence ϕ was only added to the training parameters on after 20% of the training steps. This is only part of the prior necessary to favor smoothness though, if μ was completely free its values could show high frequency, large magnitude fluctuations which would negate the smoothing. As the smoothing is bijective, this is clearly always the case. To handle this a prior of $\mathcal{N}[0, \zeta]$ was added to μ where ζ is a hyperparameter capturing the typical numerical scale of the problem, the specific value was again found to have minimal impact on model performance.

APPENDIX B

USE OF LARGE LANGUAGE MODELS

In composing this work, large language models (LLMs), in particular OpenAI's GPT models accessible through UHH-ChatGPT were used for proof reading and for performing descriptive search. This appendix discusses how LLMs were used was done, how associated risks were mitigated and subjective evaluations of those methods.

B.1 Proofreading

During the writing process LLMs were used to proof read chunks of the text. These chunks where typically of section size. Before the text was submitted in a prompt the LLM was conditioned using the following prompt:

> I will provide LaTeX text for proofreading. Please go through it paragraph by paragraph and provide feedback using the following tiered system with emojis:
>
> Error – for actual mathematical mistakes, incorrect derivations, or misstatements.
>
> Suggestion – for improvements, clarifications, more precise phrasing, or style refinements.
>
> Typo – for spelling, small notation, or formatting errors.
>
> Literature - something that needs a citation, ideally recommend something to cite (review articles > books > primary research)
>
> Missing context - Concepts which should be introduced elsewhere in the work
>
> At the end of each subsection or logical block, include a short summary in a clearly marked ' Feedback' with professor-style commentary, highlighting rigor, clarity, and readability. Focus on providing actionable feedback without rewriting the text unless explicitly requested. Adjust your feedback style to match the subject area (mathematics, ML, physics, etc.) as context dictates. Be critical, but not unfair or aggressive. Pay special attention that all arguments are sound and all mathematics is correct.

The suggestions generated by the LLM were then evaluated and when appropriate the text was improved manually based on them. This caught a lot of typos, non-ideal latex and improved some formulations. Mistakes in the content itself were not found. Usually most suggestions by the LLM were not deemed to be genuine improvements and thus not acted upon, but due to the bullet point nature of the output this did not cause significant time loss. All in all this was found to be very helpful for improving language and catching typos that still produced valid english words or typos in the equations where syntax highlighting cannot catch them. As all text was still generated by myself it is unlikely that this caused issues for the content of this work. No text in this thesis was generated by an LLM.

B.2 Descriptive Search

Occasionally I had the problem that I wanted to look for something that I did not know the standard lingo for. In those situations it was helpful to describe to the LLM what I was looking and asking it to provide me with a pointer to where I could go search for more information. These questions were usually worded like “how do you call”, “where can I find out more about” etc. The response usually includes keywords that can then be used to perform classical web and literature searches much more easily.

This process was found to be extraordinarily helpful, it manages to avoid hallucinations because the results I use still come from a conventional search, but still allows me to be wordy, imprecise and mathematical in describing what I am looking for. No information used in this thesis was taken from LLM generation.