Ludwig-Maximilians-Universität München

Thesis for the Master of Science in Physics

# Bayesian Component Separation for Tomography

Bayesianische Komponentenseparation in Tomographie


Robin Dehde

September 11, 2017

Supervised by PD Dr. Torsten Ensslin

**Abstract**

While medical imaging devices have constantly improved over the last thirty years, one part of them has remained virtually untouched: The reconstruction algorithm. Even though the underlying physics of the measurement process is very well understood and large amounts of high fidelity data is available, not all of the available information is being used in imaging devices today (e.g. characteristic absorption spectra of certain materials within the body, like bone). This thesis aims to advance the information harvesting in medical tomography by reconstructing several image components, like tissue and bone, that are correlated on certain spatial scales and anti-correlated on others. To this end the information field theory framework is used. It is shown that leveraging the cross correlation between components can increase the performance of the imaging algorithm.

# Contents

# Chapter 1

# Introduction

In recent years algorithms like deep learning have received an extraordinary amount of media and research attention. Progress in computational power has given even conceptually rather simple algorithms the power to outperform humans in many regards. Yet there are fields where algorithms still suffer from a lack of attention: "Despite major advances in x-ray sources, detector arrays, gantry mechanical design and especially computer performance, one component of computed tomography (CT) scanners has remained virtually constant for the past 25 years—the reconstruction algorithm." [1].

How can one improve the reconstruction algorithm? Generally speaking, algorithm A might be able to outperform another algorithm B (given the same data) when more information about the setup of the experiment is given to A. The reconstruction algorithm that is still in use today in medical imaging (namely the backprojection algorithm) uses merely the most basic knowledge about the imaging setup. For example, whether there is a leg or a walnut in the scanner is not known by the algorithm nor does it matter to its reconstruction. While this makes for great generalizability, all the knowledge about the human body is neglegted. Yet a high volume of information is available and could possibly improve the reconstruction. So why not use it?

Another important question is what the gains could be from improving the used algorithms and whether they justify the costs in terms of development, testing and deployment efforts necessary. Ideally, an improved algorithm could lead to reduced patient radiation exposure necessary in the scanning process of several imaging techniques such as positron emission tomography (PET) or computed tomography (CT), thereby decreasing the risk of diseases that are linked to radiation exposure.

This line of reasoning is the starting point of this thesis. Using the methods of information field theory (IFT) it is aimed to improve medical tomography algorithms.

In the following chapter, the theoretical fundament for this thesis will be laid down. Bayesian inference based on the results of a measurement process will be discussed and IFT introduced. Chapter 3 contains the application of the theory on a medical imaging tomography setup. A component reconstruction algorithm is derived for cross-correlated components. Consecutively chapter 4 presents the implementation of this algorithm and results obtained with it on simulated measurement data. Chapter 5 concludes and provides an outlook on future work in this area.

# Chapter 2

# Statistical inference

## 2.1 The measurement process

The measurement process is the bread and butter of experimental physics. Generally speaking, the physicist is interested in accumulating evidence in favor or against a theory, or in determining a certain parameter within a theory. The latter will be the main focus in the following.

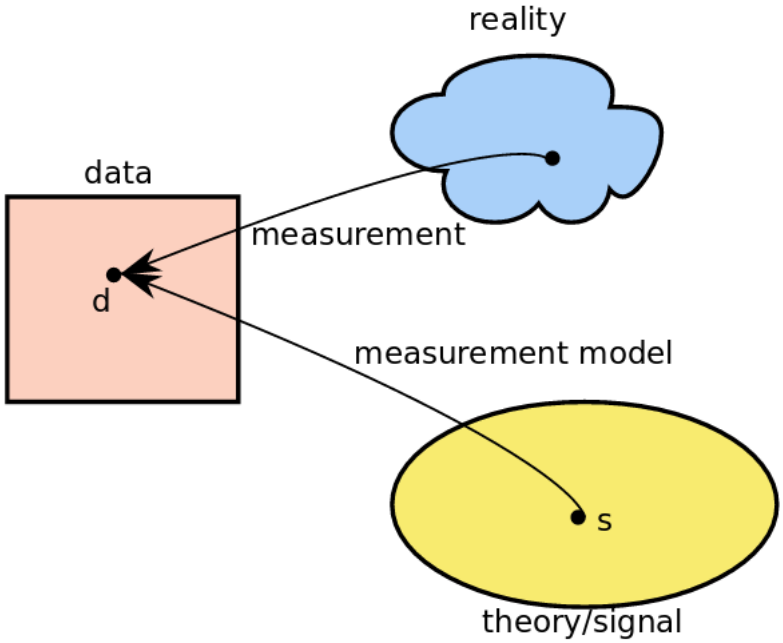



**Figure 2.1:** The measurement process. Reality is measured in an experiment and creates the data d. It is expected that a measurement model is capable of reproducing the same data d. How to fit model parameters towards data is a task of statistical inference. Picture taken from [2]

In order to be able to determine parameters, the physicist designs an experiment to make an observation of the reality that we experience (see figure 2.1). This observation is called measurement and everything that was measured (e.g. a number of photons that arrieved in a detector) is carefully documented. This documentation results in the data d. So how can we now make sense of this data, or more specifically, how can we link the data to the free parameters? This link is provided by the measurement model (also called data model). It

represents the theory for which we want to determine free parameters. Making sense of the data then means: Finding the signal s that has lead to the measurement data d.

This problem is complicated by the fact that every data that was measured was corrupted by some sort of noise in the measurement process. Noise is simply any fluctuation in the data that cannot be predicted by the model even when knowing all the free parameters s. In fact, noise is not correlated to the signal. The model can merely say that noise is to be expected, but it is not possible to predict the exact value of the noise. This introduces uncertainty into the measurement process.

In mathematics, uncertainty is dealt with using the tools of probability theory. Therefore, now the task is: Finding the probability of the signals having a certain value, given the data d. Probability theory tells us, that the quantity we are looking for is the posterior probability of certain signal values s, given the data d: $\mathcal{P}(s|d)$. Hence, we want to infer this posterior probability. This will be done using Bayes' theorem.

## 2.2 Bayesian inference

As we just found out, we want to infer the posterior probability $\mathcal{P}(s|d)$. How can we calculate it? Bayes' theorem answers this question in a logically consistent way. When this theorem is used for the inference, the process is referred to as Bayesian inference.

Bayes' theorem is described in the language of probability theory. Hence, in order to discuss it, the basic mathematical terms should be easily recollected by the reader. First of all, what is a probability? The probability of a statement 'A' is denoted $\mathcal{P}('A')$ and assigns the statement 'A' a number between 0 and 1, referring to the probability that the statement 'A' is true. In the following, $\mathcal{P}(s)$ and $\mathcal{P}(d)$ will appear. The former is to be interpreted as 'the probability that the parameters have certain values s', which is actually a shorthand for $\mathcal{P}('$The parameter of the model have exactly the value s$')$ (keep in mind that for continuous a continuous parameter space $\mathcal{P}(s)$ is a density and the probability for an exact value s is mostly 0). For $\mathcal{P}(d)$ this translates to $\mathcal{P}('$The measured data is exactly d$')$. Thereby, $\mathcal{P}(s)$ and $\mathcal{P}(d)$ assign every possible parameter value (or possible data) a certain probability.

Furthermore, probabilities that we assign to certain parameters change when our knowledge changes. For example, when before the physicist's experiment we assigned the probability $\mathcal{P}(s)$ to the values s of the parameters, we will assign the conditional probability $\mathcal{P}(s|d)$ to the values s of the parameters, after we have taken into account a certain dataset d. Thereby, the dataset d has changed our beliefs about the probability for every possible parameter value set s.

Lastly, one can state the so called joint probability of two statements at once as $\mathcal{P}(s, d)$. The relationship between this joint probability and the former expressions is captured in the product rule of probabilities:

$$\mathcal{P}(s, d) = \mathcal{P}(s|d)\mathcal{P}(d) = \mathcal{P}(d|s)\mathcal{P}(s) \tag{2.1}$$

Using this rule, Bayes' theorem can now be constructed:

$$\mathcal{P}(s|d) = \frac{\mathcal{P}(d|s)\mathcal{P}(s)}{\mathcal{P}(d)} \tag{2.2}$$

The posterior $\mathcal{P}(s|d)$ assigns a configuration of parameters s a probability, given the data d. It is calculated by considering subjective beliefs about the signal, represented through the so-called

prior distribution $\mathcal{P}(s)$, as well as the model likelihood $\mathcal{P}(d|s)$ and a probability distribution $\mathcal{P}(d)$ called evidence, which gives the probability to measure exactly the data d. In order to infer the signal, these three quantities need to be considered.

The model likelihood assigns a probability to the measured data, given the structure of the chosen model and its parameters s. In a measurement process, the data model can be expressed as:

$$d = R(s) + n \tag{2.3}$$

where $R(s)$ is referred to as the response of the signal and $n$ is a random number that represents the ever present noise in the data. The response can be understood as the functional dependency of the quantity measured through the data of the signal. In case of a linear response, i.e. $R(s) = R_{xy}s_y$, the inference scheme naturally leads to the Wiener filter solution [2].

The prior distribution $\mathcal{P}(s)$ should encode the knowledge about the signal prior to the measurement (therefore the name 'prior distribution'). Generally speaking, priors can enforce better convergence of the minimization scheme and enhance results by inhibiting overfitting. Of course, a badly chosen prior might as well corrupt the inference.

Lastly, the evidence needs to be calculated. This is achieved by marginalizing over the signal.

$$\mathcal{P}(d) = \int ds\mathcal{P}(s,d) = \int ds\mathcal{P}(d|s)\mathcal{P}(s) \tag{2.4}$$

Yet this leads to an integral for continuous spaces which are far from easy to calculate, especially in high dimensional parameter spaces. But the evidence is merely a normalization constant and can be circumvented. Without it, one can still argue where the mean and mode of the distributions are. Section 2.4 will introduce three methods to achieve so.

## 2.3 Information Field Theory

Information field theory (IFT) lays the fundament to update beliefs on fields in the presence of data optimally regarding the information that data and beliefs contain, and makes it possible and straightforward to create reasonable and informed field reconstruction schemes for a wide scope of applications.[3] Its main ingredients are logically consistent probabilistic reasoning (mostly represented in the strict and consequent use of Bayes' theorem), information as its ultimate tool for evaluation (Shannon entropy, KL-divergence and more, introduced in section 2.4.3) as well as the use of the grand repository of methods to be found in physicists' field theories (Feynman diagrams, operator formalism, Gibbs free energy and many more). In the following, the basic language of information field theory will be shortly introduced.

### Probability distributions over continuous fields

When dealing with probability distribution functions over continuous fields, it is interesting to note how they are defined. As an example take the common multivariate Gaussian distribution:

$$\mathcal{G}(s-m,D) = \frac{1}{|2\pi D|^{1/2}}\, e^{-\frac{1}{2}(s-m)^\dagger D^{-1}(s-m)} \tag{2.5}$$

where $s$ and $m$ are $N$-dimensional vectors and $D$ is a $N \times N$ matrix, with $N$ being the number of dimensions. Going from these discrete quantities to the continuous case then simply involves

taking the limit of $N$ to infinity. A way to imagine this is to let the entry $s_i$ of the vector $s$ describe one pixel of an image, where the pixels are chosen to be the same size. Taking the limit of $N$ to infinity can then be imagined as increasing the resolution of the image to an infinitely precise resolution. In accordance with this mental picture, taking the limit of N to infinity does not change the form of the distribution and will turn vectors into fields as well as matrices into operators (as is common practice in field theories) [2]. The multiplication of vectors then becomes the scalar product of fields:

$$s^\dagger j = \int dx\ \bar{s}(x) j(x) \tag{2.6}$$

while the matrix multiplication $Ds$ becomes the operator $D$ acting on the field $s$:

$$Ds = \int dy\ D(x,y) s(y) \tag{2.7}$$

### Information Hamiltonian and partition function

Two further expressions that will appear throughout this thesis are the information Hamiltonian and the partition function. Both these expressions are analogous to statistical physics where a connection between energies and probabilities is established, too. The posterior distribution can be written as:

$$\mathcal{P}(s|d) = \frac{\mathcal{P}(s,d)}{\mathcal{P}(d)} = \frac{1}{Z} e^{-\mathcal{H}(s,d)} \tag{2.8}$$

where $Z$ and $\mathcal{H}(s,d)$ denote the partition function and information Hamiltonian respectively. The partition function is just the evidence. The connection between all these terms is best described in the following expressions.

$$\mathcal{P}(d) = \int \mathcal{D}s\ \mathcal{P}(d,s) = \int \mathcal{D}s\ e^{-\mathcal{H}(s,d)} = Z(d) \tag{2.9}$$

Since the evidence is merely the normalization constant, it carries no information about the location of mean or mode. Information about mean and mode is coded in the information Hamiltonian.

### Operator calculus for IFT

A very convenient tool for calculating Gaussian expectation values is the operator calculus developed by Leike and Ensslin [7]. Otherwise tedious analytical calculations become much simpler with the help of it.

A general task of the operator calculus is to calculate the expectation value

$$\langle f(s) \rangle_{\mathcal{G}(s-m,D)} \tag{2.10}$$

For the simple case $f(s) = s$ the expectation value can be easily solved with a substitution ($s' = s - m$).

$$\langle s_x \rangle_{\mathcal{G}(s-m,D)} = \langle s'_x + m_x \rangle_{\mathcal{G}(s',D)} = m_x \tag{2.11}$$

where it was used that the first moment of a zero mean Gaussian is 0. It can be shown that calculating Gaussian expectation values in this way is equivalent to splitting up each random number vector $s$ into an annihilation operator ($a$) and creation operator ($b$) [7]. They are defined as:

$$a_x = \int dy D_{x,y} \frac{\delta}{\delta m_y} \tag{2.12}$$

$$b_x = m_x \tag{2.13}$$

where $m_x$ is the mean of the Gaussian distribution. For example, the above expectation value can be calculated in the following way:

$$\begin{aligned}
\langle s_x \rangle_{\mathcal{G}(s-m,D)} &= \left\langle \int dy D_{x,y} \frac{\delta}{\delta m_y} + m_x \right\rangle_{\mathcal{G}(s-m,D)} = \\
&= \int \mathcal{D}s \left( \int dy D_{x,y} \frac{\delta}{\delta m_y} + m_x \right) \mathcal{G}(s-m,D) \\
&= \left( \int dy D_{x,y} \frac{\delta}{\delta m_y} + m_x \right) \langle 1 \rangle_{\mathcal{G}(s-m,D)} \\
&\equiv (a_x + b_x) \langle 1 \rangle_{\mathcal{G}(s-m,D)} \\
&= m_x
\end{aligned} \tag{2.14}$$

In the last row, it was used that the annihilation operator will vanish when acting on a term independent of m while b is merely the mean of the Gaussian $m$. For this example the classical way of calculating the expectation value was simple though and introducing the formalism did not simplify the calculation. But this changes for more complex functions like the exponential of $s$:

$$\langle \exp(s_x) \rangle = \exp(a_x + b_x) \langle 1 \rangle \tag{2.15}$$

For further simplification, the Baker-Campbell-Hausdorff formula is necessary [8]:

$$\begin{aligned}
e^{b_x + a_{x'}} &= \exp\left( -\frac{1}{2} [b_x, a_{x'}] \right) e^{b_x} e^{a_{x'}} \\
&= \exp\left( +\frac{1}{2} D_{x,x'} \right) e^{b_x} e^{a_{x'}}
\end{aligned} \tag{2.16}$$

The last line was achieved by using the commutator relation between the annihilation and creation operator $[b_x, a_{x'}] = -D_{x,x'}$. Now for calculating 2.15 one only needs to rewrite the exponential of the annihilation operator in its Taylor series form. The only term that does not contain an annihilation operator in the series is the starting 1. All the other terms cancel with the expectation value $\langle 1 \rangle$.

$$\begin{aligned}
\exp(a_x + b_x) \langle 1 \rangle &= \exp\left( +\frac{1}{2} D_{x,x'} \right) e^{b_x} e^{a_{x'}} \langle 1 \rangle = \\
&= \exp\left( +\frac{1}{2} D_{x,x'} \right) e^{b_x} \langle 1 \rangle = \\
&= e^{\frac{1}{2} D_{x,x'}} e^{m_x}
\end{aligned} \tag{2.17}$$

The classical analytical solution takes much more effort. This will be a handy tool for all Gaussian expectation values in the calculations in chapter 3.

## 2.4 Posterior distribution

As mentioned, the principal goal of Bayesian inference is to determine the exact posterior distribution. The distribution could then be used to directly assign posterior probabilities to parameter configurations s, or to calculate mean, maximum and any other quantities of interest. Yet calculating the distribution directly is highly challenging, as it is necessary to evaluate the partition function. Rarely is it possible to perform this analytically. Two common approaches to solving this problem are numerical methods like sampling and approximations of the posterior of some sort. Three approaches are discussed in the following: Sampling, maximum a posteriori and approximating the posterior by a Gaussian.

### 2.4.1 Sampling from the posterior

If it is deemed necessary to get a nearly exact posterior distribution, it might make sense to sample the posterior. This approach is computationally extremely demanding and therefore only makes sense in certain cases. The idea of sampling is to draw samples from the posterior distribution and then use these samples to calculate the expectation value of any function under the sample posterior.

A widely used method for this is the Metropolis-Hastings algorithm, a Markov chain Monte Carlo (MCMC) algorithm. This algorithm can provide any probability distribution $\mathcal{P}(x)$ when provided with a function $f(x)$ that is proportional to $\mathcal{P}(x)$. This thereby makes it possible to skip calculating the partition function. Therefore drawing from the joint distribution of signal and data allows to replace the posterior with a set of N samples $s_i$:

$$\widetilde{P}(s|d) = \frac{1}{N}\sum_{i=0}^{N}\delta(s - s_i) \tag{2.18}$$

This posterior can now be used to calculate the expectation value of any function of s according to the Glivenko-Cantelli theorem.

$$\langle f(s)\rangle = \int \mathcal{D}s\ f(s)P(s|d) \approx \frac{1}{N}\sum_{i=0}^{N} f(s_i) \tag{2.19}$$

Computational difficulties arise for several reasons. Generally speaking, a large number of samples is necessary to have a good estimate for the posterior and this number increases exponentially for higher dimensions. Moreover, the Metropolis-Hastings algorithm needs to converge towards the equilibrium distribution, as samples do not necessarily approach the wanted posterior if some parts of the phase space are not reached by the MCMC. Furthermore, Markov chains draw correlated samples which reduce the effective number of independent samples. Therefore, this method needs to be carefully considered for its computational hunger. While it might make sense to use it for a one time only reconstruction, it most certainly is too interminable for repeated use in medical imaging (at least with current price-performance ratio).

### 2.4.2 Maxmimum a posteriori

The maximum a posteriori (MAP) is on of the simplest ways to approximate the posterior and hence a good starting point. It is also used in this thesis. This approach reduces the posterior distribution to a single point: It's maximum and therefore the most likely posterior value. Therefore, the approach approximates the posterior by a $\delta$ distribution at the location of the maximum. In this approach, the task becomes to find the MAP, which boils down to maximising $P(s, d)$ with respect to $s$.

The maximum can be found by differentiating the information Hamiltonian $H(s,d)$ with respect to the signals and using a gradient descent scheme in order to find its minimum (which corresponds to the maximum of $P(s,d)$). The big advantage of this approach is that it is easily derived and relatively simple to implement, especially for complicated Hamiltonians. In some cases, the mean and MAP will also coincide. Yet in the case of a multimodal distribution, the mode of the distribution is not necessarily representative of the distribution. Moreover, uncertainty information is lost, though it can be approximated by the Laplace approximation which uses the curvature of the Hamiltonian in its minimum to approximate the posterior by a Gaussian distribution that is centered around the MAP and has the curvature of the Hamiltonian as its covariance.

### 2.4.3 Approximating via Kullback-Leibler divergence

When it seems plausible to approximate the posterior by some other distribution than a $\delta$-function, the question becomes how to fit the approximation towards the actual posterior in the best way possible. In information theory, the 'best way possible' is translated into 'with the least loss of information'. The amount of relative information loss between distributions can be quantified by the Kullback-Leibler divergence [4]:

$$\mathrm{KL}\left(P(s|d)||\widetilde{P}(s|d)\right) \equiv \int \mathcal{D}s P(s|d) \ln\left(\frac{P(s|d)}{\widetilde{P}(s|d)}\right) \tag{2.20}$$

where $P(s|d)$ and $\widetilde{P}(s|d)$ denote the actual posterior and approximate posterior respectively. It is interesting to note that the KL-divergence is not symmetric, as can easily be seen in the definition above (exchange of the distributions does not result in the same expression). Minimizing the approximative posterior to the real posterior corresponds to the expression above. Unfortunately, this divergence poses the problem of having to calculate an expectation value under the more complex, but correct posterior. An approximative way around this is to exchange the arguments of the KL, so that the integral goes over the simpler approximative posterior $\widetilde{P}(s|d)$ This divergence now has to be minimized with respect to all parameters of the approximate posterior. The choice of an approximating function, therefore, has a major effect on the quality of the resulting posterior.

In this thesis it was tried to approximate the posterior with a Gaussian of the form:

$$\widetilde{P}(s|d) = \mathcal{G}(s-m, D) \tag{2.21}$$

where m and D denote the mean and covariance matrix respectively. This leads to the following KL-divergence:

$$\begin{aligned}
\mathrm{KL}\left(P(s|d)||\widetilde{P}(s|d)\right) &= \int Ds \widetilde{P}(s|d) \ln\left(\frac{\widetilde{P}(s|d)}{P(s|d)}\right) && (2.22)\\
&\widehat{=} \left\langle H(s,d)\right\rangle_{\widetilde{P}(s|d)} - \left\langle \widetilde{H}(s|d)\right\rangle_{\widetilde{P}(s|d)} && (2.23)\\
&\equiv \left\langle H(s,d)\right\rangle_{\widetilde{P}(s|d)} - \mathcal{S}(m,D|d) && (2.24)
\end{aligned}$$

Here it was used that $\left\langle \ln(P(s|d))\right\rangle_{\widetilde{P}(s|d)} = \left\langle -H(s,d)\right\rangle_{\widetilde{P}(s|d)}$ up to a constant term which can be neglected, as it drops later when the KL is differentiated. Furthermore, the expectation value of the negative logarithm of the Gaussian under the Gaussian itself was identified as the Shannon entropy of the Gaussian. Moreover, the last expression was identified as the Gibb's energy [5]:

$$G(m, D) = U(m|d) - T\mathcal{S}(m, D|d) \tag{2.25}$$

where $U$ is referred to as the inner energy, which is just the expectation value of the Hamiltonian from above. The temperature $T$ is equal to 1 in the statement before. In order to express all terms in the Gibb's energy, the Shannon entropy needs to be determined. For a Gaussian approximate posterior it is:

$$\begin{aligned}
\mathcal{S}(m, D|d) &= \left\langle \widetilde{H}(s|d) \right\rangle_{\widetilde{P}(s|d)} && (2.26)\\
&= \left\langle \frac{1}{2}(s-m)^\dagger S^{-1}(s-m) \right\rangle_{\widetilde{P}(s|d)} + \frac{1}{2}\left[\ln|2\pi S|\right] && (2.27)\\
&= \frac{1}{2}\,\mathrm{Tr}\left[\left\langle (s-m)(s-m)^\dagger \right\rangle S^{-1}\right] + \frac{1}{2}\,\mathrm{Tr}[\ln(2\pi S)] && (2.28)\\
&= \frac{1}{2}\,\mathrm{Tr}\left[\mathbb{1} + \ln(2\pi S)\right] && (2.29)
\end{aligned}$$

Here it was used that traces allow for cyclical permutations, which let the remaining expectation value cancel itself with $S^{-1}$ to a unit matrix. Moreover the identity $\ln|A| = \mathrm{Tr}[\ln(A)]$ for any positive definite, hermitian matrix $A$ was employed.

In order to find an expression for the KL-divergence, the last task is to find the expectation value $\langle H(s,d)\rangle_{\widetilde{P}(s|d)}$, which is specific to the information Hamiltonian and will be done in section 3.2. When this is done, the Gaussian needs to be fitted to the real posterior. This is achieved by minimizing the Gibbs' energy with respect to both $m$ and $D$. Again this minimization cannot be done analytically in most cases, but instead needs a numerical minimization scheme.

## 2.5 Minimization scheme

In this thesis, the MAP is estimated in order to approximate the posterior. Finding the corresponding minimum energy or maximum probability is most of the time impossible to do analytically though. Instead, the minimum is found with numerical iterative schemes. In this thesis, Newton's method will be applied. Instead of blindly following the gradient of the information Hamiltonian to lead to its minimum, steps are also guided by the second derivative. Taking the second derivative into account leads to more reasonable steps (compared to when just using the gradient of the Hamiltonian) and thereby improves convergence behavior of the minimization scheme. Put in pseudo code this boils down to:

1. Initialize necessary parameters
2. While convergence criterium is false:
   (a) Build implicit covariance operator: $D$ (see 3.30)
   (b) Calculate force on m: $f_m$ (see 3.29)
   (c) Calculate Newton step: $\Delta m = D f_m$
   (d) Take Newton step: $m = m + \Delta m$

This minimization scheme reliably and quickly converges for the setup used in this thesis. The convergence criterium was a certain number of steps to go. Alternatively one could consult a relative energy difference from one step to another.

This chapter ends the introduction of the necessary theory. Next, the theory will be applied on the medical tomography setup. Modifications of the basic setup will be discussed and finally, the tools that have just been described will be applied.

# Chapter 3

# Medical tomography setup

## 3.1 Basic Data model

This thesis treats a tomography problem which is encountered in several imaging devices like the single photon electron computed tomography (SPECT) and positron electron tomography (PET) [6]. One can imagine a computed tomography-like scenario, as shown in figure 3.1.

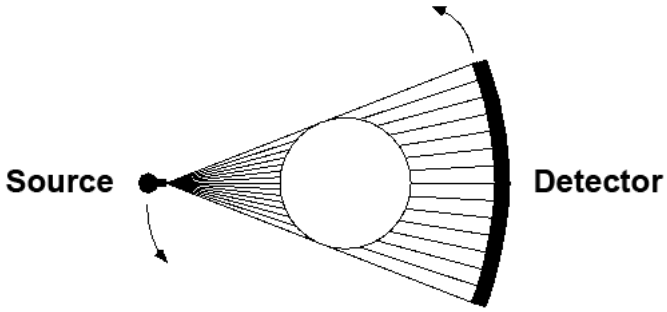



**Figure 3.1:** The source emits a number of photons. They travel through a body (shown as a white circle) where some of these photons are absorbed (absorptions are described by the absorption factor) or scattered (resulting in a background). The detector then counts how many photons reached all the way through. Picture from [6]

The line that a photon traveled along from source to detector is called line of sight (LOS). For each LOS, it is assumed that a number of photons were sent-in (of which the statistical mean is known) and that it is measured how many photons reached the detector. Source and detector are rotated around the body in order to observe the body under different geometries and to obtain complementary information. The data then consists of the number of photons measured by the detector for each LOS as well as the path these photons took.

Given that it is known that on average a certain number of photons $b_i$ was emitted, it is expected that on average $\lambda_i$ photons are measured in the detector. The actual data $d_i$ that is measured is expected to be Poisson distributed around its expectation value $\lambda_i$.

The physics happening inside the device can be described with Beer's law of absorption

plus a background $r_i$. [6]

$$\langle d_i \rangle_{(d|s)} = \lambda_i(E) = b_i(E)\ \exp\left(-\int \mathrm{d}x^3 \int \mathrm{d}l\ \delta(\vec{x} - \vec{x}_i + l\hat{n}_i)\ s(\vec{x}, E)\right) + r_{i,E} \tag{3.1}$$

The integrals define the exact line of sight taken for this datapoint. It has the length $l$ and points in the direction $\hat{n}_i$. $s(\vec{x}, E)$ denotes the signal, which here is the position and energy dependent photon absorption coefficient. It specifies the strength of absorption at a certain position $\vec{x}$ for photons of energy E. The background $r_{i,E}$ occurs due to photons scattering inside the device. The background is kept in calculations for future work but is neither simulated in the mock data nor is it applied to the reconstruction algorithm.

Since the process follows Poissonian statistics, the likelihood for this setup is:

$$\mathrm{P}\left(d|\lambda(\varphi)\right)\ =\ \prod_i \frac{\lambda_i^{d_i}\ e^{-\lambda_i}}{d_i!}\ , \tag{3.2}$$

where $i$ denotes one particular scanning setup, hence one path that one ray went along until it was detected. The likelihood now covers the basic physics (Beer's law) and statistics (Poisson distribution) of this medical tomography setup.

### 3.1.1 Model improvements and further assumptions

As this thesis aims to advance current medical algorithms, more information needs to exploited by our reconstruction algorithm. A prior needs to be specified, and several assumptions will be made in the following.

In an absorption process, a certain component, like bone, should show a characteristic absorption spectrum, no matter what bone it is. That is, bones could be assumed to typically consist of a characteristic mixture of atoms. This makes it reasonable to expect the material bone to show a characteristic energy dependent absorption coefficient $\mu(E)$. This knowledge can be leveraged to turn this reconstruction scheme into a component separation algorithm. Yet a simple 0/1, bone or no bone approach would be too simplistic to capture medical situations. Healed injuries can increase bone density, just as medical conditions can decrease or increase bone density at certain points in the body. Therefore, it is necessary to include variable bone densities. Nonetheless, it is assumed that while the density of bones may vary, the absorption coefficient of the component 'bones' is characteristic and does not (or not too much). This should not only be true for bones, but also for other components of the body, like tissue. This can be incorporated into our modeling in the following way:

$$s(\vec{x}, E) = \sum_c \mu_c(E) s_c(\vec{x})\ , \tag{3.3}$$

where $\mu_c(E)$ denotes the absorption coefficient of component $c$ at Energy $E$ and $s_c(\vec{x})$ will be reffered to as component-signal. The absorption coefficient is assumed to be known for all components. From now on $s$ will be referred to as absorption factor and $\mu$ as absorption coefficient in order to distinguish the two.

Moreover, now that the absorption factor is split up into components, it is straightforward to assume that the component-signal $s_c(\vec{x})$ is strictly positive. Furthermore, it is expected that sudden jumps in the coefficient will appear in places where for example bone ends and tissue starts. Both properties can be modeled by parametrizing the signal as the exponential of an underlying field. This ensures positivity for all signal values. Furthermore, it makes larger

steps in the order of magnitude of the field easier to achieve for the algorithm. Hence, the underlying field $\varphi(\vec{x})$ is introduced.

$$s_c(\vec{x}) = \exp(\varphi_c(\vec{x})) \tag{3.4}$$

With these modifications the expectation value $\lambda_i$ becomes:

$$\begin{aligned} \lambda_i(E) &= b_i(E) \exp\left( - \int \mathrm{d}x^3 \int \mathrm{d}l \; \delta(\vec{x} - \vec{x}_i + l\hat{n}_i) \sum_c \mu_c(E) e^{\varphi_c(\vec{x})} \right) + r_{i,E} \\ &\equiv b_i \exp\left[ -R_i^{\dagger_x} \left( \mu^{\dagger_c} e^{\varphi} \right) \right] + r_i \equiv \lambda_i \end{aligned} \tag{3.5}$$

In the second part of this equation, shorthand notations were introduced. Firstly, the energy dependence of $\lambda$ is not explicitly stated from now on, as it would only unnecessarily extend the notation and is not required for the following calculations. Secondly, throughout this thesis the following definitions hold:

$$a^\dagger b = \int \mathrm{d}\vec{x} \sum_c a_x^c b_x^c, \qquad a^{\dagger_c} b = \sum_c a^c b^c, \qquad a^{\dagger_x} b = \int \mathrm{d}\vec{x} \; a_x b_x$$

Furthermore, $R_{i,x} \equiv \int \mathrm{d}l \delta(\vec{x} - \vec{x}_i + l\hat{n}_i)$ determines the exact line of sight to integrate over. To avoid later confusions it is important to note what exactly $R_{i,x}$ does. It is an operator that takes in a signal space field and returns a data space vector. If, as in $\lambda_i$, it appears as $R_i$ then it is to be treated as the $i$-th vector within the $R_{i,x}$ matrix. Such a vector $R_i$ has as many $x$-entries as there are pixels in the discretized signal space (needs to be discretized in order to get it into computer memory). Each entry contains a weight that tells us if the $i$-th LOS went across this pixel, and if yes, then how much space it covered within the pixel (minimally just covering the edge of a pixel, maximally taking the full pixel diagonal). Therefore $R_i^{\dagger_x} \left( \mu^{\dagger_c} e^{\varphi} \right)$ resembles the inner product of fields (first summation over components, then over signal space) and returns a scalar. This is expected, since the exponential of the latter needs to return a scalar in eq. 3.5.

Lastly the prior needs to be discussed. $\varphi$ is assumed to be homogenous in space and to follow a Gaussian distribution with zero mean and a known covariance $\Phi$.

$$\mathcal{P}(\varphi) = \mathcal{G}(\varphi, \Phi) \tag{3.6}$$

$\varphi$ is a vector that has the field of a certain component as entries. $\Phi$ describes the covariance for different lengths of one component with itself, but also the cross correlation between the components. This raises the question: How are these components cross-correlated? And can this be used for reconstruction? Starting out again with the bone and tissue example it is reasonable to expect certain cross-correlations depending on the length scale under consideration. Bone and tissue are often direct neighbors, yet they do not appear in the exact same place. In terms of distance correlation, this should translate to a positive large scale correlation and a negative mid to short scale correlation. It is assumed that the entire covariance matrix $\Phi$ is known. Since it is generally possible to determine typical power spectra for certain components and cross-correlation out of the huge amounts of data that have been created in medical imaging, these assumptions seem plausible. $\Phi$ represents information about the general structure of the components of the human body, and it is intelligible to leverage this knowledge in a reconstruction scheme.

With these further assumptions the final information Hamiltonian is:

$$\mathcal{H}(\varphi, d) = \mathcal{H}(d|\varphi) + \mathcal{H}(\varphi) = \tag{3.7}$$

$$= \sum_i \left[\ln(d_i!) + \lambda_i(\varphi) - d_i \ln \lambda_i(\varphi)\right] + \frac{1}{2}\varphi^\dagger \Phi^{-1}\varphi + \frac{1}{2}\ln|2\pi\Phi| \tag{3.8}$$

As a sidenote, it is straightforward to add another measurement with photons at Energy $E'$ that result in the data $d'$ to this Hamiltonian. This simply enlarges the sum over all i by the amount of the new data $d'$ with corresponding values $\lambda'_i$. This will be used in the implementation later on. For now, with the information Hamiltonian available, it is possible to derive approximations for the posterior.

## 3.2 Gaussian approximation of posterior

The first approach pursued during this thesis was a Gaussian approximation. In order to calculate the Gibbs free energy of the information Hamiltonian (at $T = 1$):

$$G(m, D) = U(m|d) - \mathcal{S}(m, D|d) \tag{3.9}$$

$$U(m) = \langle \mathcal{H}(d, \varphi)\rangle_{(s|d)} \tag{3.10}$$

the expectation value of the Hamiltonian needs to be computed. This will be done in the following in consecutive steps and approximations. Starting with the (approximate) posterior expectation value of $\lambda_i$:

$$\langle \lambda_i \rangle_{\mathcal{G}(\varphi - m, D)} = b_i^{\dagger_E} \left\langle \exp\left[-R_i^{\dagger_x}\left(\mu^{\dagger_c} e^{\varphi}\right)\right]\right\rangle_{\mathcal{G}(\varphi-m,D)} + r_i \tag{3.11}$$

Note that this expectation value cannot be solved analytically in a straightforward manner. But it can be approximated by splitting $\varphi$ up into its mean value plus the variation around its mean: $\varphi = m + \delta\varphi$.

$$\left\langle \exp\left[-R_i^{\dagger_x}\left(\mu^{\dagger_c} e^{\varphi}\right)\right]\right\rangle_{\mathcal{G}(\varphi-m,D)} = \left\langle \exp\left[-\int \mathrm{d}x R_{i,x}\left(\mu^{\dagger_c}(e^{\varphi})_x\right)\right]\right\rangle_{\mathcal{G}(\varphi-m,D)} \tag{3.12}$$

$$= \left\langle \exp\left[-\int \mathrm{d}x R_{i,x}\left(\mu^{\dagger_c}(e^{m}e^{\delta\varphi})_x\right)\right]\right\rangle_{\mathcal{G}(\delta\varphi,D)} \tag{3.13}$$

$$= \left\langle \exp\left[-R_i^{\dagger_x}\left(\mu^{\dagger_c}\widehat{e^m}e^{\delta\varphi}\right)\right]\right\rangle_{\mathcal{G}(\delta\varphi,D)} \tag{3.14}$$

When splitting the exponential function up in this way a new notation needs to be introduced. A hat on a field effectively turns that field into a matrix with the field components on the diagonal and zero everywhere else, $\widehat{e^m}_{x,x'} = (e^m)_x \delta_{x,x'}$. Later on, there will also appear two hats over a field. Every hat thereby codes for a delta function, therefore $\widehat{\widehat{e^m}}_{x,x'} = (e^m)_x \delta_{x,x'}\delta_{x,x''}$. Now the second exponential function can be Taylor expanded and taken to first order in $\delta\varphi$:

$$\left\langle \exp\left[-R_i^{\dagger_x}\left(\mu^{\dagger_c}\widehat{e^m}e^{\delta\varphi}\right)\right]\right\rangle_{\mathcal{G}(\delta\varphi,D)} \approx \exp\left[-R_i^{\dagger_x}\left(\mu^{\dagger_c}e^m\right)\right]\left\langle \exp\left[-R_i^{\dagger_x}\left(\mu^{\dagger_c}\widehat{e^m}\delta\varphi\right)\right]\right\rangle_{\mathcal{G}(\delta\varphi,D)}$$

Only the right term is now dependent on the $\varphi$ within the Gaussian. But this term can be expressed as:

$$\left\langle \exp\left[-R_i^{\dagger_x}\left(\mu^{\dagger_c}\widehat{e^m}\delta\varphi\right)\right]\right\rangle_{\mathcal{G}(\delta\varphi,D)} = \exp\left(\frac{1}{2}\left(R_i^{\dagger_x}\mu\widehat{e^m}\right)^\dagger D\left(R_i^{\dagger_x}\mu\widehat{e^m}\right)\right) \tag{3.15}$$

using the operator formalism developed by Leike and Enßlin (see section 2.3). It is applied in the following for 2 components.

$$\left\langle \exp\left[-R_i^{\dagger x}\left(\mu^{\dagger c}\widehat{e^m}\delta\varphi\right)\right]\right\rangle_{\mathcal{G}(\delta\varphi,D)} =$$

$$\begin{aligned}
&= \exp\left[-R_{i,x}^{\dagger x}\left(\mu^{\dagger c}\widehat{e^m}\,(a+b)\right)\right]\langle 1\rangle_{\mathcal{G}} \\
&= \exp\left(-\int \mathrm{d}\vec{x} R_{i,x}\mu_{c_1} e^{m_x^{c_1}} b_x^{c_1} - \int \mathrm{d}\vec{x}' R_{i,x'}\mu_{c_1} e^{m_{x'}^{c_1}} a_{x'}^{c_1}\right. \\
&\qquad \left. - \int \mathrm{d}\vec{x}'' R_{i,x''}\mu_{c_2} e^{m_{x''}^{c_2}} b_{x''}^{c_2} - \int \mathrm{d}\vec{x}''' R_{i,x'''}\mu_{c_2} e^{m_{x'''}^{c_2}} a_{x'''}^{c_2}\right)\langle 1\rangle_{\mathcal{G}} \\
&= \exp\left(-\int \mathrm{d}\vec{x} R_{i,x}\mu_{c_1} e^{m_x^{c_1}} b_x^{c_1} - \int \mathrm{d}\vec{x}' R_{i,x'}\mu_{c_1} e^{m_{x'}^{c_1}} a_{x'}^{c_1} - \int \mathrm{d}\vec{x}'' R_{i,x''}\mu_{c_2} e^{m_{x''}^{c_2}} b_{x''}^{c_2}\right) \\
&\times \exp\left(-\frac{1}{2}\left[\int \mathrm{d}\vec{x} R_{i,x}\mu_{c_1} e^{m_x^{c_1}} b_x^{c_1} + \int \mathrm{d}\vec{x}' R_{i,x'}\mu_{c_1} e^{m_{x'}^{c_1}} a_{x'}^{c_1}\right.\right. \\
&\qquad \left.\left. + \int \mathrm{d}\vec{x}'' R_{i,x''}\mu_{c_2} e^{m_{x''}^{c_2}} b_{x''}^{c_2}\ ,\ \int \mathrm{d}\vec{x}''' R_{i,x'''}\mu_{c_2} e^{m_{x'''}^{c_2}} a_{x'''}^{c_2}\right]\right) \\
&\times \exp\left(-\int \mathrm{d}\vec{x}''' R_{i,x'''}\mu_{c_2} e^{m_{x'''}^{c_2}} a_{x'''}^{c_2}\right)\langle 1\rangle_{\mathcal{G}}
\end{aligned}$$

Here the Baker-Campbell-Hausdorff formula was used to create the last equation:

$$\begin{aligned}
e^{b_x^c + a_{x'}^{c'}} &= \exp\left(-\frac{1}{2}\left[b_x^c, a_{x'}^{c'}\right]\right) e^{b_x^c} e^{a_{x'}^{c'}} \\
&= \exp\left(+\frac{1}{2}(D_{x,x'})_{c,c'}\right) e^{b_x^c} e^{a_{x'}^{c'}}
\end{aligned} \tag{3.16}$$

Using the commutator relation $\left[b_x^c, a_{x'}^{c'}\right] = -(D_{x,x'})_{c,c'}$, the commutator term above can be further simplified.

$$\left[\int_L \mathrm{d}x R_{i,x} e^{m_x^{c_1}}\mu_{c_1} b_x^{c_1} + \int_L \mathrm{d}x' R_{i,x'} e^{m_{x'}^{c_1}}\mu_{c_1} a_{x'}^{c_1} + \int_L \mathrm{d}x'' R_{i,x''} e^{m_{x''}^{c_2}}\mu_{c_2} b_{x''}^{c_2}\ ,\ \int_L \mathrm{d}x''' R_{i,x'''} e^{m_{x'''}^{c_2}}\mu_{c_2} a_{x'''}^{c_2}\right]$$

$$\begin{aligned}
&= -\sum_c \int_L \mathrm{d}x \int_L \mathrm{d}x''' \left(R_{i,x} e^{m_x^c}\mu_c\right)(D_{x,x'''})_{c,c_2}\left(R_{i,x'''} e^{m_{x'''}^{c_2}}\mu_{c_2}\right) \\
&= -\sum_c \left(R_i^{\dagger x}\mu_c\widehat{e^{m^c}}\right)^{\dagger_x}(D)_{c,c_2}\left(R_i^{\dagger x}\mu_{c_2}\widehat{e^{m^{c_2}}}\right)
\end{aligned}$$

Now swapping the last annihilation operator along the creation operators yields:

$$\begin{aligned}
&= \exp\left(-\int_L \mathrm{d}x R_{i,x}\mu_{c_1} b_x^{c_1} - \int_L \mathrm{d}x'' R_{i,x''}\mu_{c_2} b_{x''}^{c_2}\right) \\
&\times \exp\left(\frac{1}{2}\left(R_i^{\dagger x}\mu\widehat{e^m}\right)^{\dagger} D\left(R_i^{\dagger x}\mu\widehat{e^m}\right)\right) \\
&\times \exp\left(-\int_L \mathrm{d}x' R_{i,x'}\mu_{c_1} a_{x'''}^{c_1}\right)\exp\left(-\int_L \mathrm{d}x''' R_{i,x'''}\mu_{c_2} a_{x'''}^{c_2}\right)\langle 1\rangle_{\mathcal{G}}
\end{aligned}$$

Taylor expanding the exponential functions containing the annihilation operators and letting these operators annihilate as well as using $b_x^c = 0$ only leaves the middle row in the last

expression standing. All other exponential functions become 1, since $(a_x^c)^n \langle 1 \rangle_{\mathcal{G}} = 0$ for any integer n . This calculation can be done for arbitrarily many components.

According to the above calculations, the posterior expectation value of $\lambda$ approximately is:

$$\langle \lambda_i \rangle_{\mathcal{G}(\varphi - m, D)} \approx b_i^{\dagger E} \exp\left( -R_i^{\dagger x} \left( \mu^{\dagger c} e^m \right) + \frac{1}{2} \left( R_i^{\dagger x} \mu^{\dagger c} \widehat{e^m} \right)^\dagger D \left( R_i^{\dagger x} \mu^{\dagger c} \widehat{e^m} \right) \right) + r_i \;\equiv\; \lambda_i'(m)$$

This is the value of $\lambda$ at the posterior mean plus an uncertainty correction. Another term within the likelihood that needs to be computed is the logarithm of lambda. Again approximations are necessary. In order to be able to reasonably apply the Taylor expansion of the logarithm, the expression is first expanded.

$$\begin{aligned} \ln(\lambda_i(\varphi)) &= \ln\left( \lambda_i(\varphi) \frac{\lambda_i(m)}{\lambda_i(m)} - 1 + 1 \right) \\ &= \ln(\lambda_i(m)) + \ln\left( \left[ \frac{\lambda_i(\varphi)}{\lambda_i(m)} - 1 \right] + 1 \right) \end{aligned} \tag{3.17}$$

Now it is assumed that the term in square brackets above is a small number, thereby empowering us to Taylor expand the second logarithm and only use its first term. This yields the expectation value up to first order in $\lambda_i(\varphi)$:

$$\langle \ln(\lambda_i(\varphi)) \rangle_{\mathcal{G}(\varphi - m, D)} \;\approx\; \ln(\lambda_i(m)) + \frac{\lambda_i'(m)}{\lambda_i(m)} - 1 \tag{3.18}$$

For calculating the expectation value it was also used that $\langle \lambda_i(\varphi) \rangle_{\mathcal{G}(\varphi - m, D)} \approx \lambda_i'(m)$. Next, the expectation value of the prior needs to be computed. As it is just a squared $\varphi$ under a Gaussian, the calculation is straightforward.

$$\begin{aligned} \left\langle \varphi^\dagger \Phi^{-1} \varphi \right\rangle_{\mathcal{G}(\varphi - m, D)} &= \sum_{c,c'} \int \mathrm{d}x \int \mathrm{d}y \left[ m_y^{c'} m_x^c \left( \Phi_{x,y}^{-1} \right)_{c,c'} + (D_{y,x})_{c',c} \left( \Phi_{x,y}^{-1} \right)_{c,c'} \right] \\ &= \mathrm{Tr}\left[ m m^\dagger \Phi^{-1} + D \, \Phi^{-1} \right] \\ &= m^\dagger \Phi^{-1} m + \mathrm{Tr}\left[ D \, \Phi^{-1} \right] \end{aligned}$$

where the cyclical property of the trace was leveraged. The last missing term is the Shannon entropy, which was already calculated in section 2.4.3 to be:

$$\mathcal{S}(m, D|d) \;=\; \frac{1}{2} \mathrm{Tr}[\mathbb{1} + \ln(2\pi D)] \tag{3.19}$$

Combining all the parts then yields the Gibbs energy.

$$\begin{aligned} G(m, D) \;\approx\; & \sum_i \left( \ln(d_i!) + d_i - d_i \ln(\lambda_i(m)) + \lambda_i'(m) \left( 1 - \frac{d_i}{\lambda_i(m)} \right) \right) \\ + \;& m^\dagger \Phi^{-1} m + \mathrm{Tr}\left[ D \, \Phi^{-1} \right] - \frac{1}{2} \mathrm{Tr}[\mathbb{1} + \ln(2\pi D)] + \frac{1}{2} \ln |2\pi \Phi| \end{aligned} \tag{3.20}$$

This then resembles the KL divergence between the real posterior and an approximate Gaussian with mean m and covariance D. In order to determine the optimal approximation, the Gibbs energy must be minimized with respect to both $m$ and $D$.

### Minimization: Gradient and curvature

As described in 2.5 Newton's method was chosen for determining the minimum energy with respect to m. Hence, gradient and curvature need to be determined. As before, both quantities will be developed step by step. This derivation starts with the gradient:

$$f_m = \frac{\delta G(m, D)}{\delta m} \tag{3.21}$$

$$\tag{3.22}$$

First the expectation value $\lambda_i'(m)$ is differentiated. In component-wise notation, this results in

$$\frac{\partial}{\partial m_z^{c^*}} \lambda_i'(m) = (\lambda_i'(m) - r_i) \left( \int \mathrm{d}y \sum_{c'} \left( R_{i,z} \mu_{c^*} e^{m_z^{c^*}} \right) (D_{z,y})_{c^*,c'} \left( e^{m_y^{c'}} \mu_{c'} R_{i,y} \right) - \left( R_{i,z} \mu_{c^*} e^{m_z^{c^*}} \right) \right)$$

while for $\lambda_i(m)$ the result is:

$$\frac{\partial}{\partial m_z^{c^*}} \lambda_i(m) \;=\; -(\lambda_i(m) - r_i) R_{i,z} \mu_{c^*} e^{m_z^{c^*}}$$

Continuing with the logarithm of $\lambda_i(m)$ yields:

$$\frac{\partial}{\partial m_z^{c^*}} \ln(\lambda_i(m)) \;=\; -\frac{\lambda_i(m) - r_i}{\lambda_i(m)} R_{i,z} \mu_{c^*} e^{m_z^{c^*}}$$

These derivatives can now be used to compute the last missing part for the likelihood contribution to the gradient.

$$\frac{\partial}{\partial m_z^{c^*}} \left( \frac{\lambda_i'(m)}{\lambda_i(m)} \right) =$$

$$\begin{aligned}
&= \frac{\lambda_i(m)(\lambda_i'(m) - r_i) \left( \int \mathrm{d}y \sum_{c'} \left( R_{i,z} \mu_{c^*} e^{m_z^{c^*}} \right) (D_{z,y})_{c^*,c'} \left( e^{m_y^{c'}} \mu_{c'} R_{i,y} \right) - \left( R_{i,z} \mu_{c^*} e^{m_z^{c^*}} \right) \right)}{(\lambda_i(m))^2} \\
&+ \frac{\lambda_i'(m)(\lambda_i(m) - r_i) R_{i,z} \mu_{c^*} e^{m_z^{c^*}}}{(\lambda_i(m))^2} \\
&= R_{i,z} \mu_{c^*} e^{m_z^{c^*}} \frac{(\lambda_i'(m) - r_i) \left( \int \mathrm{d}y \sum_c (D_{z,y})_{c^*,c} e^{m_y^c} \mu_c R_{i,y} - 1 \right) + \lambda_i'(m) \left( 1 - \frac{r_i}{\lambda_i(m)} \right)}{\lambda_i(m)}
\end{aligned}$$

The contribution of the prior to the gradient is:

$$\frac{\partial}{\partial m_z^{c^*}} \left\langle \frac{1}{2} \varphi^\dagger \Phi^{-1} \varphi \right\rangle_{\mathcal{G}(\varphi - m, D)} = \int \mathrm{d}y \sum_c \left( \Phi_{z,y}^{-1} \right)_{c^*,c} m_y^c$$

Combining all these parts then finally yields the gradient with respect to $m$. In the hope of providing more clarity, it is provided below in component-wise notation as well as in vector notation.

$$
\begin{aligned}
(f_m)_z^{c^*} &= \sum_i \left[ d_i \frac{\lambda_i(m) - r_i}{\lambda_i(m)} - (\lambda_i'(m) - r_i) \right] \left( \int \mathrm{d}z' R_{i,z'} \mu_{c^*} e^{m_{z'}^{c^*}} \delta_{z',z} \right)_z^{c^*} + \\
&+ \sum_i d_i \left[ \frac{\lambda_i'(m) - r_i}{\lambda_i} - \frac{\lambda_i'(m)\left(\lambda_i(m) - r_i\right)}{\lambda_i(m)^2} \right] \left( \int \mathrm{d}z' R_{i,z'} \mu_{c^*} e^{m_{z'}^{c^*}} \delta_{z',z} \right)_z^{c^*} \\
&+ \sum_i \left[ (\lambda_i'(m) - r_i) \left( 1 - \frac{d_i}{\lambda_i(m)} \right) \right] \left( \int \mathrm{d}z' R_{i,z'} \mu_{c*} e^{m_{z'}} \delta_{z',z} \right) \int \mathrm{d}y \sum_c (D_{z,y})_{c,c*} \left( R_{i,y} \mu_c e^{m_y} \right) \\
&+ \int \mathrm{d}y \sum_c \left( \Phi_{z,y}^{-1} \right)_{c^*,c} m_y^c
\end{aligned} \tag{3.23}
$$

$$
\begin{aligned}
f_m &= \sum_i \left[ d_i \frac{\lambda_i(m) - r_i}{\lambda_i(m)} - (\lambda_i'(m) - r_i) \right] R_i^{\dagger x} \left( \mu^{\dagger c} \widehat{e^m} \right) + \\
&+ \sum_i d_i \left[ \frac{\lambda_i'(m) - r_i}{\lambda_i} - \frac{\lambda_i'(m)\left(\lambda_i(m) - r_i\right)}{\lambda_i(m)^2} \right] R_i^{\dagger x} \left( \mu^{\dagger c} \widehat{e^m} \right) \\
&+ \sum_i \left[ (\lambda_i'(m) - r_i) \left( 1 - \frac{d_i}{\lambda_i(m)} \right) \right] \left( R_i^{\dagger x} \mu \widehat{\widehat{e^m}} \right)^\dagger D \left( R_i^{\dagger x} \mu \widehat{e^m} \right) \\
&+ \Phi^{-1} m
\end{aligned} \tag{3.24}
$$

Next up is the curvature or the covariance of the Gaussian. Given a current value for the mean, the covariance matrix can be determined by minimizing the Gibbs energy for that mean with respect to the covariance operator. Hence we differentiate the gradient of G with respect to the covariance operator D and then set it to 0.

$$
\begin{aligned}
((f_D)_{x,y})_{c^*,c^{**}} &= \sum_i (\lambda_i'(m) - r_i) \left( 1 - \frac{d_i}{\lambda_i(m)} \right) \frac{1}{2} R_{i,x} \mu_{c^*} e^{m_x^{c^*}} e^{m_y^{c^{**}}} \mu_{c^{**}} R_{i,y} \\
&+ \frac{1}{2} \left( (\Phi_{x,y}^{-1})_{c^*,c^{**}} - (D_{x,y}^{-1})_{c^*,c^{**}} \right)
\end{aligned} \tag{3.25}
$$

$$
f_D = \sum_i \frac{1}{2} \left( \widehat{R_i}\ \mu e^m \right) \frac{(\lambda_i'(m) - r_i)}{\lambda_i(m)} \left( \lambda_i(m) - d_i \right) \left( \widehat{R_i}\ \mu e^m \right)^\dagger + \frac{1}{2} \left( \Phi^{-1} - D^{-1} \right) \tag{3.26}
$$

Setting the gradient to 0 then results in:

$$
(D_{x,y}^{-1})_{c^*,c^{**}} = (\Phi_{x,y}^{-1})_{c^*,c^{**}} + \sum_i (\lambda_i'(m) - r_i) \left( 1 - \frac{d_i}{\lambda_i(m)} \right) \frac{1}{2} R_{i,x} \mu_{c^*} e^{m_x^{c^*}} e^{m_y^{c^{**}}} \mu_{c^{**}} R_{i,y} \tag{3.27}
$$

$$
D^{-1} = \Phi^{-1} + \sum_i \left( R_i^\dagger \mu \widehat{e^m} \right) \frac{(\lambda_i'(m) - r_i)}{\lambda_i(m)} \left( \lambda_i(m) - d_i \right) \left( R_i^{\dagger x} \mu \widehat{e^m} \right)^\dagger
$$

Therefore, for determining $D^{-1}$ it was not necessary to use Newton's method, as the optimum can be analytically found. This ends the derivation for a reconstruction algorithm that employs a Gaussian approximation. This solution was not used in the end, due to complications it causes for the implementation. $\lambda_i'(m)$, which appears frequently in gradient and curvature, is highly demanding to compute. In it appears the term:

$$
\frac{1}{2} \left( R_i^{\dagger x} \mu \widehat{e^m} \right)^\dagger D \left( R_i^{\dagger x} \mu \widehat{e^m} \right) \tag{3.28}
$$

For the covariance operator, there is only an implicit representation of $D^{-1}$ available. A direct inversion of $D^{-1}$ is not computationally feasible. Normally such a situation is approached with the conjugate gradient method ( see section 4.1). With quite some effort, it is possible to compute the matrix multiplication result of $D\exp(m)$ given merely $D^{-1}$ and $\exp(m)$. Yet we are not capable of determining $D$ in the process. So this time-consuming method has to be done every time D appears in the calculations. Unfortunately, this conjugate gradient method would have to be used for every data point in order to calculate $\lambda_i$. Using some other methods or approximations this might be solvable though. However, as the scope of this theis is to investigate the impact of including knowledge on inter-component correlations, and not on optimal numerical solution strategies, it was decided to settle for the maximum a posteriori approximation of the posterior.

## 3.3 Maximum a posteriori

Finding the maximum a posteriori (MAP) is quite simple compared to the Gaussian approximation, as the difficult calculation of the expectation values falls away. All that is needed here to feed Newton's method is the gradient of the information Hamiltonian and its curvature, the latter being equivalent to the Laplace approximation of the covariance. The derivative of the Hamiltonian with respect to $\varphi$ is:

$$\frac{\partial H(\lambda(\varphi),d)}{\partial \varphi_z^{c^*}} = -\sum_i \left(\lambda_i - d_i + r_i\left(\frac{d_i}{\lambda_i}-1\right)\right) R_{i,z}\mu_{c^*} e^{\varphi_z^{c^*}} + \int \mathrm{d}y \sum_c \left(\Phi^{-1}_{z,y}\right)_{c^*,c} \varphi_y^c \tag{3.29}$$

The Laplace approximation of the covariance is:

$$\begin{aligned} D^{-1} &= \frac{\partial H(\lambda,d)}{\partial\varphi\partial\varphi^\dagger} = \\ &= \sum_i\sum_j \left(\widehat{e^{\varphi}}\mu R_i\right)\left(\lambda_i(\varphi) - r_i\left(1+\frac{d_i}{\lambda_i(\varphi)}\right) + r_i^2\frac{d_i}{\lambda_i^2(\varphi)}\right)\delta_{i,j}\left(R_j\mu\ \widehat{e^{\varphi}}\right) + \\ &+ \sum_i \left(d_i - \lambda_i(\varphi) - r_i\left(1+\frac{d_i}{\lambda_i(\varphi)}\right)\right)\left(R_i\mu\ \widehat{\widehat{e^{\varphi}}}\right) \\ &+ \Phi^{-1} \end{aligned} \tag{3.30}$$

Thankfully, the MAP results cause no major computational problems. The first derivative of the information Hamiltonian will seek a balance between the likelihood and prior contributions. The likelihood contribution takes into account the difference between $\lambda_i$ and $d_i$, thereby comparing how many photons would be measured according to the current $\varphi$ with the number of measured photons $d_i$. Yet the likelihood contribution does not contain any information about the relative information from one component to the other. This is only achieved through the cross-correlation in the prior, which will prove to be useful.

When the covariance is applied within Newton's method, it weighs the force on the parameters $\varphi$ with the inverse of the exponential of their current value, their absorption coefficient $\mu$ and the number of photons $\lambda_i$ that is expected, given the current value of $\varphi$. Of course the prior also contains valuable information about the current curvature.

With help of the derivations in this chapter so far, a tomography setup dataset could theoretically be reconstructed now. Yet it is always a good idea to first deliver a proof of concept before daring to work with real data and its extra challenges.

## 3.4 Drawing correlated random fields

To demonstrate the capabilities of this algorithm, mock data for each component field $\varphi_c(\vec{x})$ needs to be created. These fields need to follow a predefined correlation structure $\Phi$ of the prior, as this information is assumed to be known and used in the reconstruction scheme. Hence, the task of this section will be to explore how to draw correlated random fields.

Thinking again of the bone and tissue example for two components, a distance based correlation between components is sought after. A distance based covariance matrix naturally arises by demanding statistical homogeneity of the fields. In a field that is statistically homogenous in space, the correlation between the field value of one point and another only depends on the distance and direction of the vector that separates these points. Therefore covariance matrix entries $\Phi_{x,x'}$ are a function dependent on this connecting vector, by definition of statistical homogeneity.

$$\Phi_{x,x'} = f(x - x') \tag{3.31}$$

Fields with this property have a diagonal covariance matrix in Fourier space according to the Wiener-Khintchin theorem for statistically homogenous processes [9]:

$$\Phi_{k,k'} = (2\pi)^u \delta(k - k') P_\Phi(k) \tag{3.32}$$

where $u$ is the number of dimensions. $P_\Phi(k)$ is referred to as the power spectrum of the field. As will be seen in the following, the covariance between components will also be diagonal in Fourierspace [10].

The goal here is to draw two real-valued random fields. Using the statistical homogeneity of the fields, the following properties of the power spectra of these fields can be shown to exist [11]:

$$\begin{aligned}
\langle \mathrm{Re}[\varphi_c(k)]\, \mathrm{Re}[\varphi_{c'}(k')]\rangle &= \langle \mathrm{Re}[\varphi_{c'}(k)]\, \mathrm{Re}[\varphi_c(k')]\rangle \\
&= (2\pi)^u \delta(k' - k)\ \mathrm{Re}\left[(P_\Phi)_{cc'}(k)\right] && (3.33)\\
\langle \mathrm{Re}[\varphi_c(k)]\, \mathrm{Im}[\varphi_{c'}(k')]\rangle &= -\langle \mathrm{Im}[\varphi_c(k)]\, \mathrm{Re}[\varphi_{c'}(k')]\rangle \\
&= (2\pi)^u \delta(k' - k)\ \mathrm{Im}\left[(P_\Phi)_{cc'}(k)\right] && (3.34)
\end{aligned}$$

These equations give the relationships between the real and imaginary components of $\Phi_{c,c'}(\mathrm{k})$ between components $c$ and $c'$. Again only the same modes of the power spectra interact. Together with the statement from above, $\Phi_{c,c'}(k)$ can now be constructed from power spectra. Only for this section the imaginary and real part will be split up into their own matrix entries, hence $\vec{\varphi}(k) = (\mathrm{Re}\,\varphi_c(k), \mathrm{Im}\,\varphi_c(k), \mathrm{Re}\,\varphi_{c'}(k), \mathrm{Im}\,\varphi_{c'}(k))$. Then:

$$\Phi(k) = \begin{bmatrix}
(P_\Phi(k))_{cc} & 0 & \mathrm{Re}\left[(P_\Phi)_{cc'}(k)\right] & -\mathrm{Im}\left[(P_\Phi)_{cc'}(k)\right] \\
0 & (P_\Phi(k))_{cc} & \mathrm{Im}\left[(P_\Phi)_{cc'}(k)\right] & \mathrm{Re}\left[(P_\Phi)_{cc'}(k)\right] \\
\mathrm{Re}\left[(P_\Phi)_{cc'}(k)\right] & \mathrm{Im}\left[(P_\Phi)_{cc'}(k)\right] & (P_\Phi(k))_{c'c'} & 0 \\
-\mathrm{Im}\left[(P_\Phi)_{cc'}(k)\right] & \mathrm{Re}\left[(P_\Phi)_{cc'}(k)\right] & 0 & (P_\Phi(k))_{c'c'}
\end{bmatrix} \tag{3.35}$$

This relationship now allows drawing correlated random numbers. First, $\Phi_{c,c'}(\mathrm{k})$ needs to be decomposed into a lower and upper triangular matrix (LU decomposition) in the following way.

$$\Phi = LL^T \tag{3.36}$$

where $L$ is a lower triangular matrix. Then $L$ times a zero mean unit variance Gaussian random vector $\vec{\rho}$ will yield the correlated random fields for the mode k.

$$\vec{\varphi}(k) = (\mathrm{Re}\,\varphi_c(k), \mathrm{Im}\,\varphi_c(k), \mathrm{Re}\,\varphi_{c'}(k), \mathrm{Im}\,\varphi_{c'}(k)) = \mathrm{L}(k)\vec{\rho} \tag{3.37}$$

For two components, this can be done analytically. Generally, this could also be done by a Cholesky decomposition, performed automatically by the computer for every $k$-value. Solving this analytically for two correlated fields results in the matrix:

$$L = \begin{bmatrix} \sqrt{S_{11}} & 0 & 0 & 0 \\ 0 & \sqrt{S_{11}} & 0 & 0 \\ \frac{\mathrm{Re}[S_{12}]}{\sqrt{S_{11}}} & \frac{\mathrm{Im}[S_{12}]}{\sqrt{S_{11}}} & \sqrt{S_{22} - \frac{S_{12}^2}{S_{11}}} & 0 \\ -\frac{\mathrm{Im}[S_{12}]}{\sqrt{S_{11}}} & \frac{\mathrm{Re}[S_{12}]}{\sqrt{S_{11}}} & 0 & \sqrt{S_{22} - \frac{S_{12}^2}{S_{11}}} \end{bmatrix} \tag{3.38}$$

which must be fulfilled for every value of k. To simplify the overview over the matrix, $(P_\Phi(k))_{cc} \equiv S_{11}$, $(P_\Phi(k))_{c'c'} \equiv S_{22}$ and $(P_\Phi(k))_{cc'} \equiv S_{1,2}$. Now the last task is to identify $S_{1,2}$. According to the definition of covariance and correlation this is just:

$$(P_\Phi(k))_{cc'} = \mathrm{corr}(c, c')(k)\sqrt{(P_\Phi(k))_{cc}(P_\Phi(k))_{c'c'}} \tag{3.39}$$

Therefore, we only need to define a desired correlation function in $k$-space and plug it into the method above and this will result in two fields in k-space that follow this correlation. This then was the last missing piece for the theoretical part of this algorithm. Now we are capable to draw correlated random fields and know how to numerically reconstruct them from data. In the following the implementation of this algorithm will be discussed.

# Chapter 4

# Implementation and Results

Within this chapter the implementation of this reconstruction algorithm will be discussed. All necessary parameters to create the mock data are stated and resulting signal fields and data presented. Lastly, reconstruction results are presented and discussed for different correlation structures, for reconstructions with and without using the cross-correlation as prior knowledge and different noise levels.

## 4.1 Numerical details

This algorithm was implemented in Python, using the package numerical information field theory for python (NIFTy3) [12]. It allows to easily implement algorithms that are formulated in the language of information field theory. Among its many features is the use of the conjugate gradient method for determining the result of an inverse operator acting on a field, given only an implicit representation of the operator. This was used for the covariance operator $D$.

The conjugate gradient method needs the operator it is applied on to be positive definite. Given equation 3.30 this is not always the case. The term allowing for negative matrix entries with $(d_i - \lambda_i)$ needs to be regulated in some way. One possibility is to set this difference to $\max(0, (d_i - \lambda_i))$. During use within this thesis this led to much higher computation time for the conjugate gradient method, yet this effect was not countered by quicker convergence. Therefore, this approach was dismissed. The second possibility is to exclude this term altogether. As the covariance here is mostly a tool for quicker and more reliable convergence, but does not influence the final result in the MAP approximation, discarding problematic terms is reasonable.

Regarding the implementation of the algorithm, the functionality of NIFTy needed to be complemented in several ways. Generally, NIFTy was at the time not capable of dealing with component fields. Therefore, 2 new operator classes (one respectively for prior covariance $\Phi$ and Laplace approximation of the covariance $D$) needed to be implemented, which are capable of taking a component field as input and yielding the same as output and handling the inputs accordingly. This allowed leveraging the built-in functionality of NIFTy such as the conjugate gradient method.

To create mock data, the method of chapter 3.4 was implemented as a Python class. It takes as input power spectra of up to two fields together with a correlation function in $k$-space and ouputs up to two fields that are cross correlated as expected.

With these fields at hand now, the LOS data needed to be created. To achieve so, the NIFTy LOSResponse class was used. In order to use it for integration, start and end points as

in a medical tomography setup were necessary. This was realized in a Python class that takes as inputs the number of LOS per angle and the number of shots that should be made evenly between 0 and 180 degrees. It then ouputs numpy arrays with starts and ends for each LOS as necessary to feed the LOSResponse.

## 4.2 Mock data

Before creating mock data several parameters need to be set. Both the ground truth signal and the reconstructed map for each signal and component were chosen to have a resolution of $256^2$ pixels (2-dimensional reconstruction setup). In order to create the data from the ground truth, a rotated line of sight setup was constructed with 250 lines per angle and 100 shots at equidistant angles within a range from 0 to 180 degrees. This results in 25.000 data points, each point containing a photon count.

The number of photons sent-in was adjusted until the reconstruction reached a reasonable resolution. For 3 out of the 5 setups presented here the sent-in photon number was set to 5 million on average (referred to as low noise setup). The last two setups were chosen to have a lower count of 4 million photons on average (referred to as higher noise setup).

The data was assumed to be measured at two distinct photon energies for two components. For 2 energy channels and 2 components 4 $\mu$ values need to be chosen (see table 4.1).

| | $\mu_1$ | $\mu_2$ |
|---|---|---|
| $E1$ | 1.25 | 0.25 |
| $E2$ | 0.5 | 0.4 |

**Table 4.1:** The values for the absorption coefficients $\mu$ for the two components at energies E1 and E2. This creates an energy channel where one component clearly dominates the other (inspired by the domination of bones over tissue), while the other energy channel barely makes it possible to distinguish the two.

The variance of each component is set as a power spectrum for each component. In order to create signal fields that show at least a bit similarity to bones and tissue (that is, rather broader structures for bones), fastly decaying power spectra were chosen ( $k^{-5}$). Combining all this with an appropriate correlation structure allows data to be created according to the data model (see equation 3.1). The following sections show the details of signal and data for two correlation structures.

### 4.2.1 Straight leap correlation

The most straightforward way to encode the previously described bone-tissue correlation is a straight jump from positive to negative inter-component correlation. Since the power spectra of the two signal fields are falling very quickly, the correlation leap needs to take place within the first few $k$-modes, otherwise the positive correlation dominates and resulting signals deviate strongly from the imagined structure. While watching out for meaningful resulting signal fields, the inter-component correlation as shown in figure 4.1 was chosen. This resulted in the signal fields shown in figure 4.2.

Now, taking into account the $\mu$ values for each energy channel, figure 4.3 shows representations of what underlying signal fields the algorithm actually can 'see' in the data for each energy

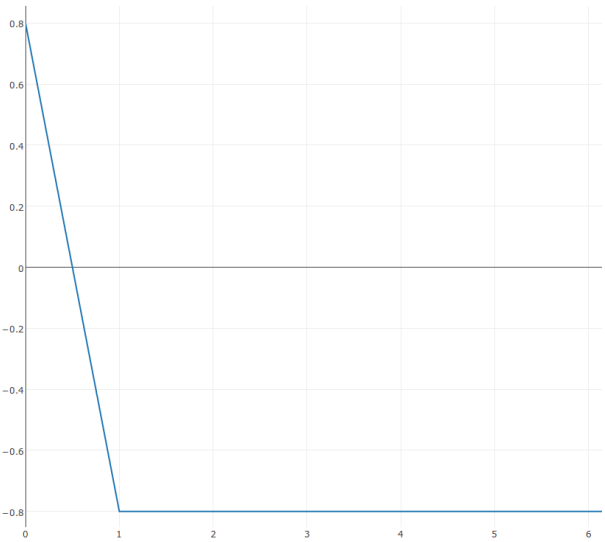
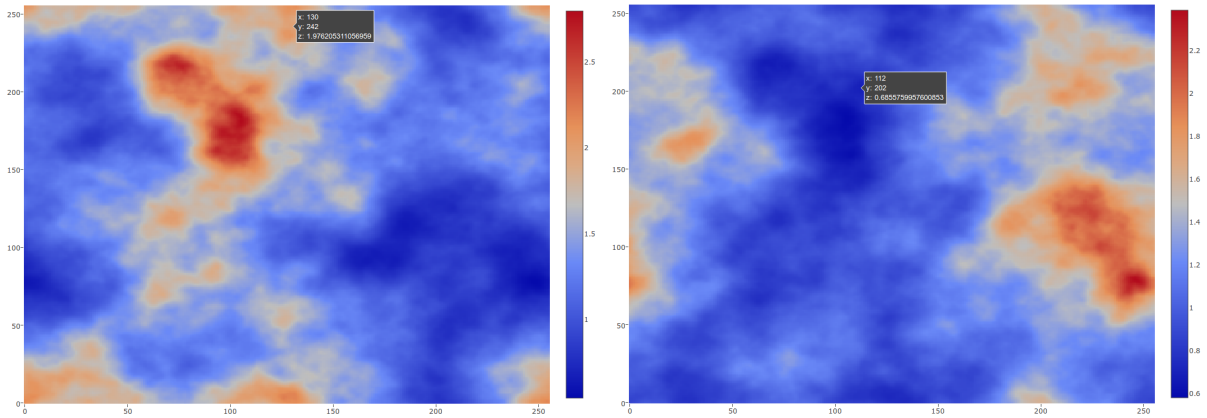


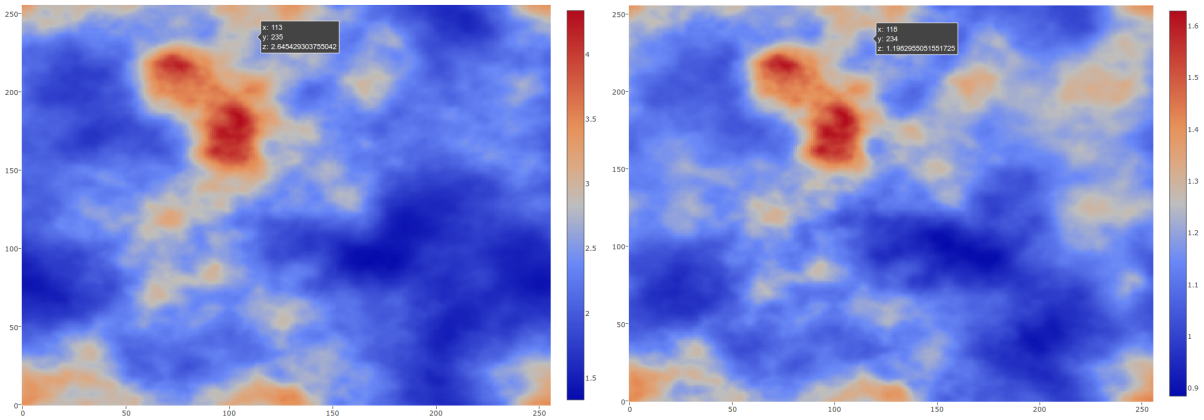

**Figure 4.1:** Inter-component correlation structure. Only the 1st mode has a positive correlation of 0.8, all the other modes are negatively correlated with -0.8. The $y$-axis shows the correlation while the $x$-axis shows the Fourier space modes $k$. They actually reach up to 181 ($= \frac{1}{\sqrt{2}}256$). but are truncated to the area of interest



**(a)** $\exp(\varphi_1)$.

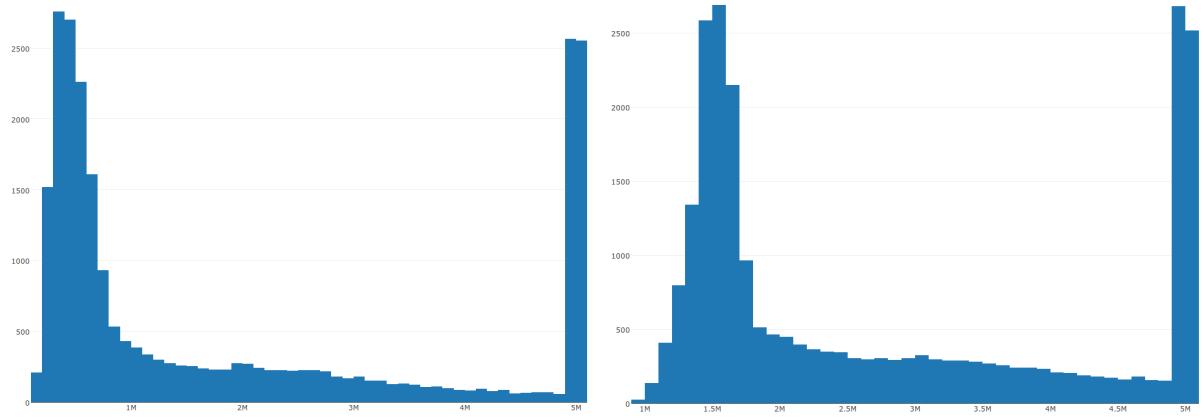
**(b)** $\exp(\varphi_2)$.

**Figure 4.2:** The space dependent fields for both components, created with the correlation structure 4.1. Red codes for high and blue for low values.



**(a)** $\exp(\varphi_1)\mu_1(E1) + \exp(\varphi_2)\mu_2(E1)$. **(b)** $\exp(\varphi_1)\mu_1(E2) + \exp(\varphi_2)\mu_2(E2)$.

**Figure 4.3:** The maps of the position dependent absorption factors for both components combined. Each plot stands for one energy channel. Component 1 clearly dominates both pictures.



**(a)** $\exp(\varphi_1)\mu_1(E1) + \exp(\varphi_2)\mu_2(E1)$. **(b)** $\exp(\varphi_1)\mu_1(E2) + \exp(\varphi_2)\mu_2(E2)$.

**Figure 4.4:** Histograms of the data created with the LOS setup of the underlying combined fields in figure 4.3. Both plots show 2 peaks. The first peak (to the left in both pictures) is due to most photons being absorbed in most of the lines of sight. The right peak is due to 'empty' lines of sight.

channel. Figure 4.3 shows that component 1 clearly dominates both energy channels. While in energy channel E1 component 2 is basically negligible, it appears much stronger in energy channel E2. The pictures represent the position dependent total absorption factor for each energy channel. Apllying the line of sight setup and response to these fields then yields the expectation values for each LOS. Consecutively for each value a Poissonian random number is drawn with the respective expectation value. This results in the data histogram shown in figure 4.4.

The histograms visualize how many LOS were created within a certain photon count interval. For energy channel E1, most LOS photon counts lie around 0.5 million, therefore roughly 90% of photons were absorbed for these LOS, whereas for E2 most counts accumulate around 1.5 million counts. For both channels the graph shows a second maximum around the 5 million sent in photons. This is because the rotating scanning setup stretches farther than the area that actually contains the signal. Thereby, part of the scanning setup does not measure the underlying field. But this should also be expected for real scanning devices.

### 4.2.2 Sigmoidal correlation with high and low noise

For this correlation structure the reconstruction is not only divided into known and unknown correlation structures (which will be referred to as informed and uninformed setup respectively), but also setups with high and low sent-in photon numbers. The intention is to better showcase the power of the prior knowledge and to examine whether prior knowledge could actually compensate for a lower radiation exposure. This results in the two datasets shown in figure 4.8 (5 million photons sent-in) and figure 4.9 (4 million photons sent-in).

In order to demonstrate the effect of a different correlation structure, that is also closer to the imagined bone tissue correlation, a sigmoidal correlation (see figure 4.5 was used to create mock data, which slowly transitions from a correlation of $+0.37$ down to $-0.80$ over the range of roughly 20 $k$-modes.

As the same random seed is used, the new correlation only changes the second component, as can be seen in figure 4.6. These signals, combined with the respective $\mu$ values than lead to the effective total absorption visualized in figure 4.7. The scanning setup together with Poissonian noise then creates the data, which is visualized in 4.8 in form of a histogram. As before, the combined signals show a strong domination by the first component and the data histograms show two main peaks, respectively.

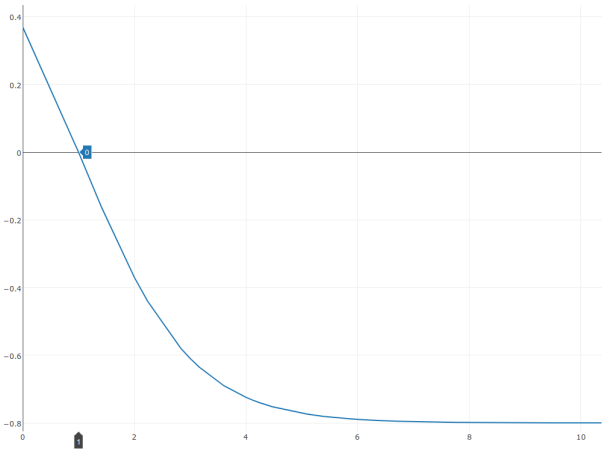
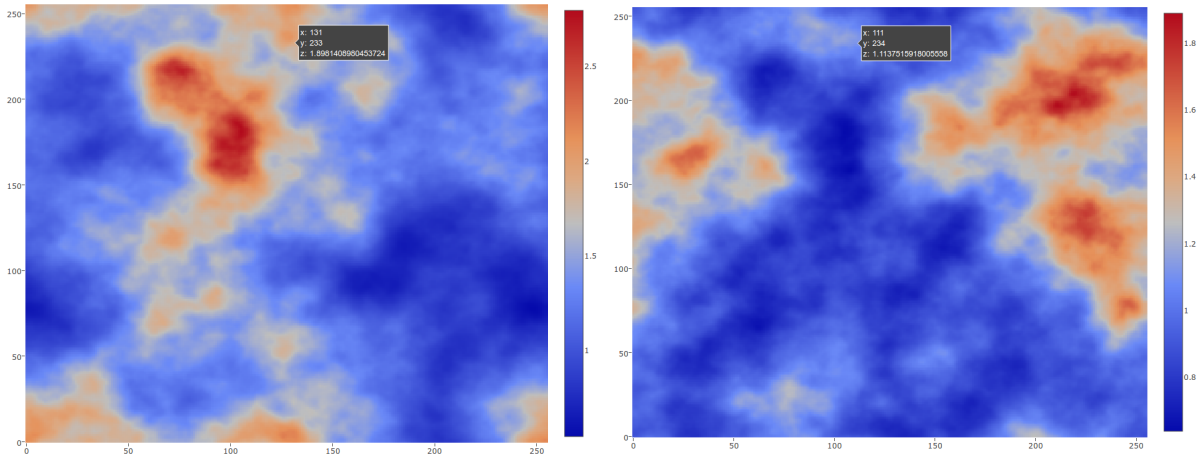
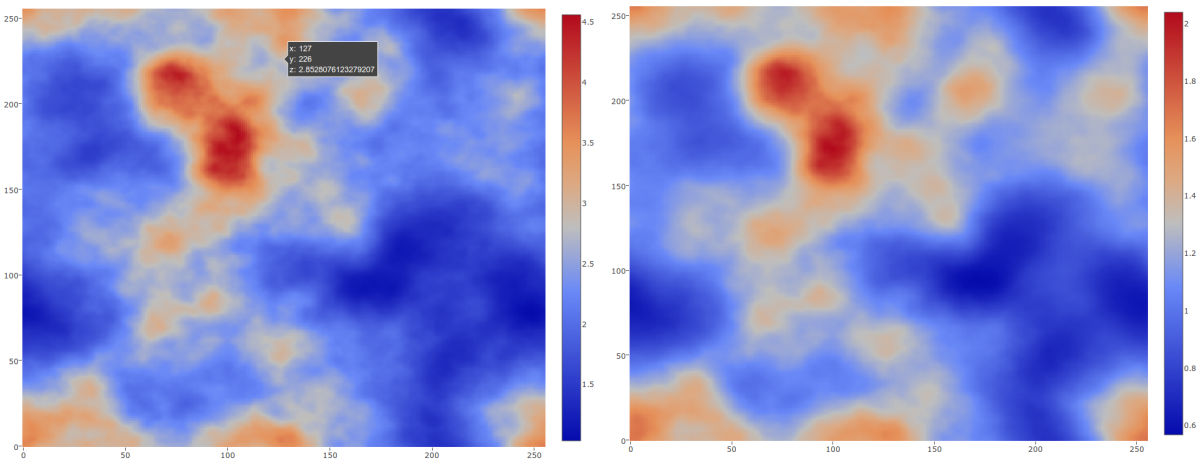


**Figure 4.5:** Correlation structure in k-space for the sigmoidal correlation setup. The first $k$-mode has a positive correlation of 0.369. Over the next roughly 20 $k$-modes the correlation transitions to a value of $-0.8$



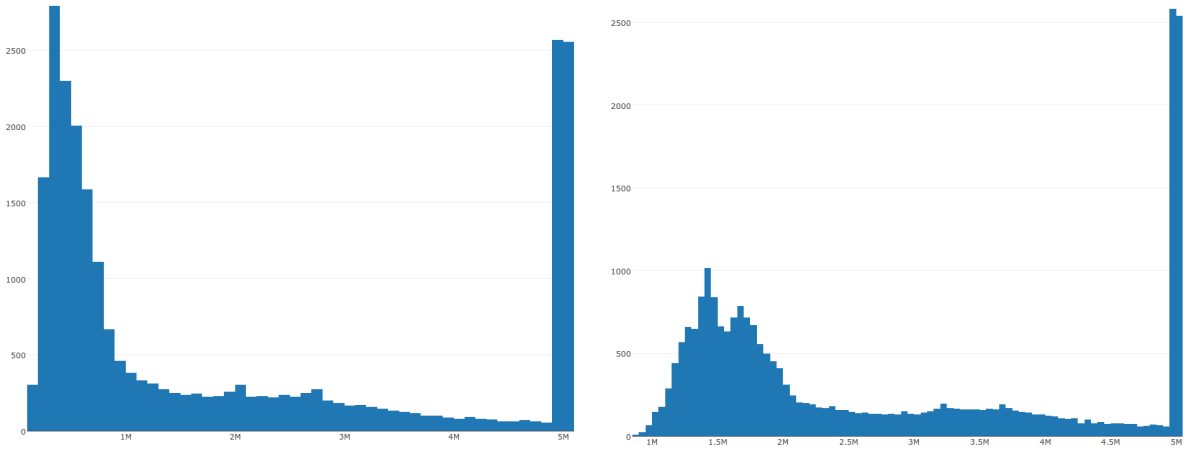
**(a)** $\exp(\varphi_1)$. **(b)** $\exp(\varphi_2)$.

**Figure 4.6:** Position dependent heat-map of the signal fields for the sigmoidal correlation setup.



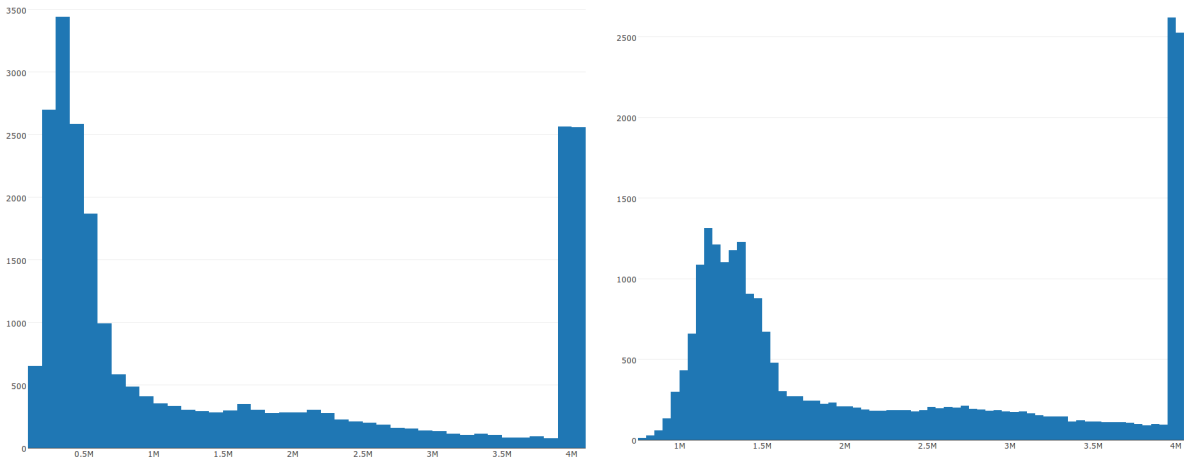
**(a)** $\exp(\varphi_1)\mu_1(E1) + \exp(\varphi_2)\mu_2(E1)$. **(b)** $\exp(\varphi_1)\mu_1(E2) + \exp(\varphi_2)\mu_2(E2)$.

**Figure 4.7:** The maps of the position dependent absorption factors for both components combined. Each plot stands for one energy channel. Component 1 again dominates both pictures.




**(a)** $\exp(\varphi_1)\mu_1(E1) + \exp(\varphi_2)\mu_2(E1)$. **(b)** $\exp(\varphi_1)\mu_1(E2) + \exp(\varphi_2)\mu_2(E2)$.

**Figure 4.8:** 5 million photon illumination case. The graphs show histograms of the data created with the LOS setup of the underlying combined fields in figure 4.7. The graph for the E2 channel shows a subdivision of the left peak into two peaks.




**(a)** $\exp(\varphi_1)\mu_1(E1) + \exp(\varphi_2)\mu_2(E1)$. **(b)** $\exp(\varphi_1)\mu_1(E2) + \exp(\varphi_2)\mu_2(E2)$.

**Figure 4.9:** 4 million photon illumination case. The graphs show histograms of the data created with the LOS setup of the underlying combined fields in figure 4.7. The subdivision of the E2 channel peak into underpeaks is less visible for this higher noise setup. The counts of photons reach farther to a 0 count, as is to be expected.

## 4.3 Results

In the following, it will be argued that:

- The introduced algorithm is generally capable of reliably reconstructing the position dependent absorption factor and splitting it up into 2 components, given 2 measurements with different characteristic absorption coefficients $\mu$ for each component and energy. This is achieved with or without the given knowledge of the cross-correlation.
- The cross-correlation is valuable for identifying components, in the presence of 2 energy channels. It is capable of increasing the resolution especially for the harder to determine component 2 which has lower (hence harder to detect) absorption coefficients $\mu$.
- The information within the cross-correlation is capable to achieve better reconstructions even when the incoming radiation is cut by 20% from 5 down to 4 million photons per line of sight. Again especially the second component benefits.
- The information gain through the cross-correlation is not fixed to the exact setup of the correlation structure but holds for different (though similar) correlation structures.

This will be demonstrated with reconstructions of the two correlation functions, which each was solved with (referred to as informed case) and without (referred to as uninformed case) prior knowledge about the cross-correlation between the components. For the sigmoidal informed case, the noise is increased by decreasing the photon illumination by 20%.

### 4.3.1 Straight leap correlation

**No cross-correlation given in prior (uninformed)**

Figures 4.10 and 4.12 show the reconstructions of the algorithm next to the ground truth for each component, respectively. The rough structure of the components is well represented. No major area is left out for the first component. The highest and lowest values in the ground truth reach 2.75 and 0.63, while the reconstruction covers an interval from 2.53 down to 0.74. These differences are likely due to the rougher resolution of the image, thereby smoothening maxima and minima. Small scale details are visibly lost.

For the second component, this resolution is even rougher. The ground truth interval goes from 2.4 down to 0.58 while the reconstruction reaches 1.96 and 0.82 respectively. The finer structure is barely visible. A larger dot at the lower right edge is only weakly represented.

A last interesting question to ask for this setup is whether this reconstruction would create the same data as was given to the algorithm. This is answered in figure 4.14 for each energy channel respectively. There, every expectation value that was calculated in the mock data creation $\lambda(\varphi)$ is compared relatively to the expectation values $\lambda(m)$:

$$\Delta(\varphi, m) = \frac{\lambda(\varphi) - \lambda(m)}{\lambda(\varphi)} \tag{4.1}$$

$\Delta$ thereby denotes the relative difference between the ground truth expectation values and the reconstruction expectation values. This is shown in 4.14a and 4.14b for the respective energy channels.

In figure 4.14a 15588 LOS out of 25000 (62.3%) lie within an interval of relative deviation from $-1\%$ to $+1\%$. Accounting for the 'empty' LOS (the lines that do not touch the

signal area) boils down to subtracting roughly 5000 lines, yielding 10588 out of 20000 or 52.9%. Therefore, more than half datapoints would show less than 1% deviation with no more than a maximal 7.8% deviation for channel E1.

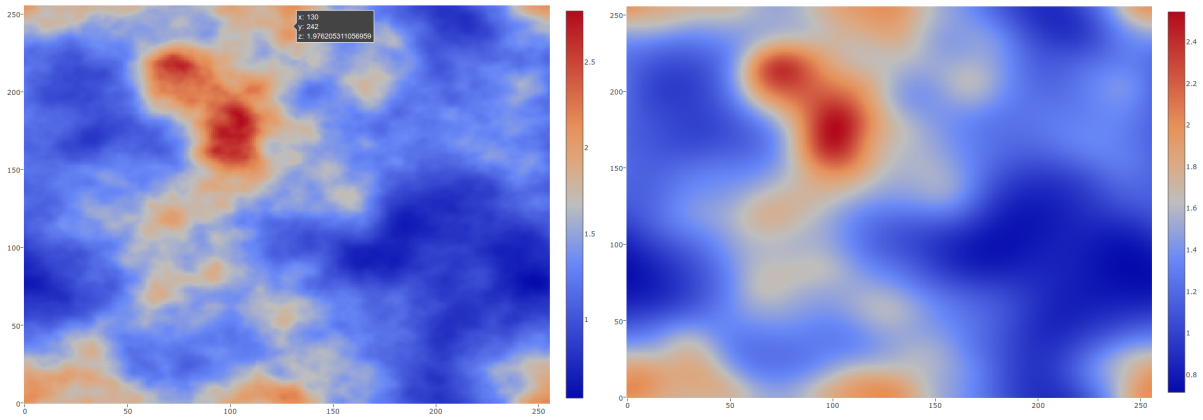
For channel E2, the interval from $-0.01$ to $+0.01$ contains 22013 out of 25000 (88.0% ) or accounted for 'empty' LOS 17013 out of 20000 (85.0%). The farthest outlier shows a relative deviation of 3.7%. Therefore, the algorithm was arguably successful in reconstructing a signal that can describe the original data that the algorithm was fed. This is interesting because it shows that the second energy channel is easier to satisfy with imperfect solutions of the reconstructed map. This is due to the lower absorption coefficients $\mu$: Since fewer photons are absorbed at each point it is harder to determine what a good reconstruction would look like.

However imperfect, these reconstruction maps demonstrate the capability of the algorithm to leverage two energy channel measurements to reconstruct and separate two components even in the absence of prior knowledge of the cross-correlation.

#### Cross correlation given in prior (informed)

For the following reconstruction, the cross-correlation structure was given to the algorithm within the prior. The components compared to the ground truth are to be found in 4.11 and 4.13. The interval of the first component wider than its correlation-uninformed counterpart, stretching from 2.59 down to 0.68 (ground truth values: 2.75 and 0.63; the last reconstruction achieved 2.53 down to 0.74). This could be a hint of a better reconstruction of details that is visually confirmed in figure 4.11 and further discussed in the direct comparison of the two in the following section.

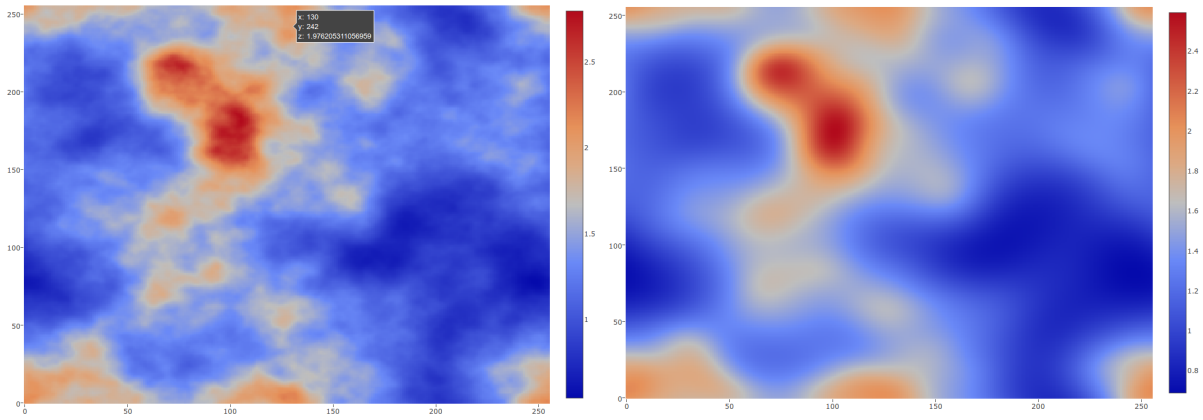
The second component reaches a maximum of 1.95 and minimum of 0.70 (ground truth 2.4 and 0.58; last reconstruction at 1.96 and 0.82). The lower extrema are quite probably due to the multitude of negative correlations between the components. But this should not distract from the fact that the overall structure of the second component seems to be captured more precisely. Even the dot at the lower right edge that was nearly missed before is now clearly visible.



**(a)** Ground truth for component 1.

**(b)** Reconstructed signal for component 1.

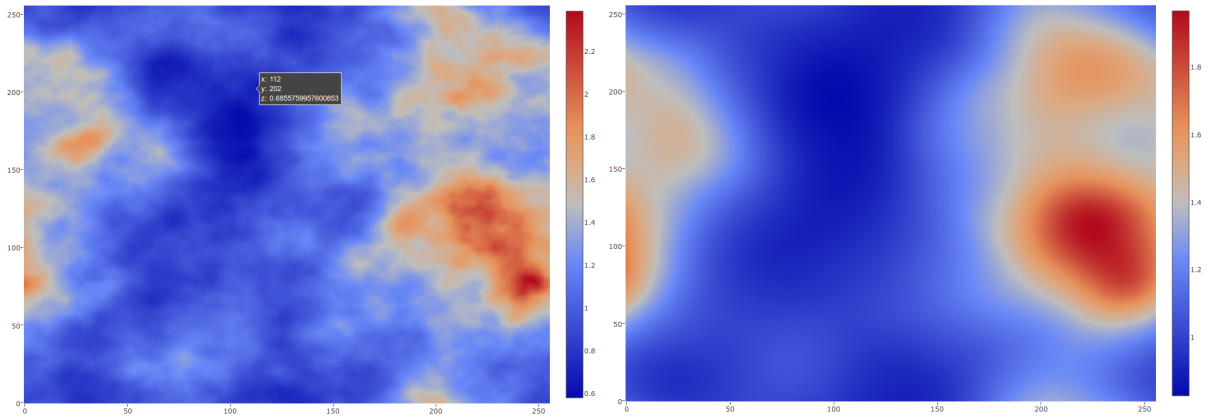
**Figure 4.10:** Comparison between ground truth and reconstruction for component 1. Large and mid-scale structures are well reconstructed.



**(a)** Ground truth for component 1.

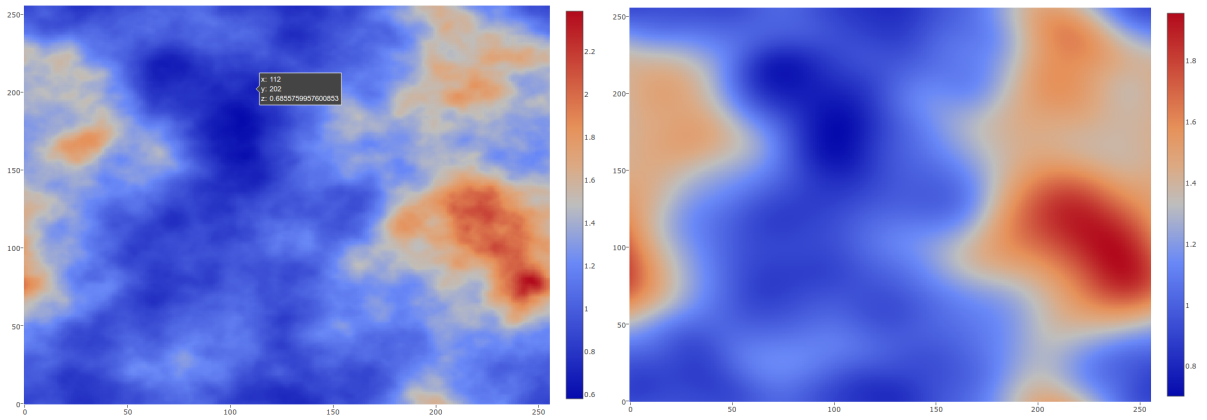
**(b)** Reconstructed signal for component 1.

**Figure 4.11:** Comparison between ground truth and reconstruction with known cross-correlation for component 1. Large scale structures are well reconstructed. The cross-correlation gives a slight boost to the reconstruction performance, visible only at close comparison (colour transition happens faster).



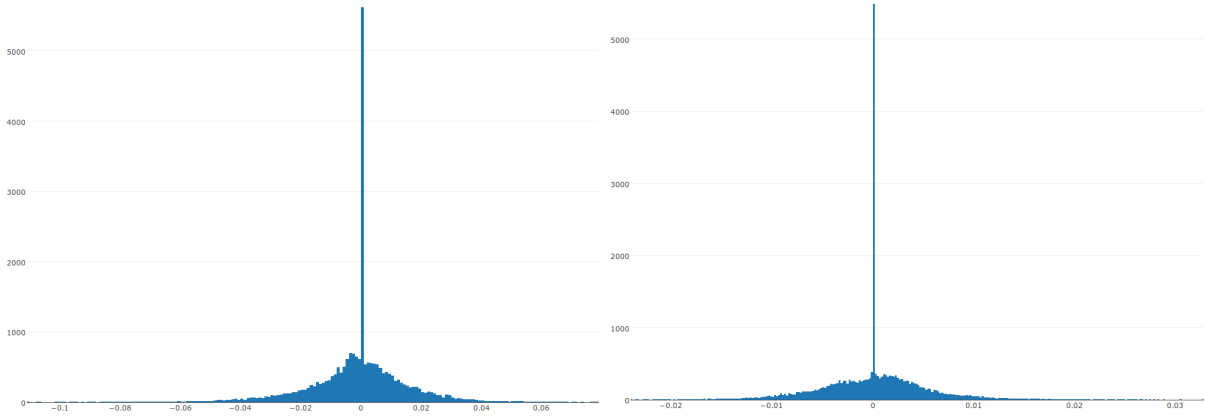
**(a)** Ground truth for component 2. **(b)** Reconstructed signal for component 2.

**Figure 4.12:** Comparison between ground truth and reconstruction for component 2. Large scale structures are successfully reconstructed.



**(a)** Ground truth for component 2. **(b)** Reconstructed signal for component 2.

**Figure 4.13:** Comparison between ground truth and reconstruction for component 2. The second component is very clearly quite different from the reconstruction before. Many more details are visible compared to the uninformedsetup.




**(a)** Comparison to noiseless data for channel E1 **(b)** Comparison to noiseless data for channel E2

**Figure 4.14:** Relative deviation between the expectation values $\lambda_i(\varphi)$ created from ground truth signal to expectation values created from the reconstruction $\lambda_i(m)$ according to equation 4.1

### Comparison between setups

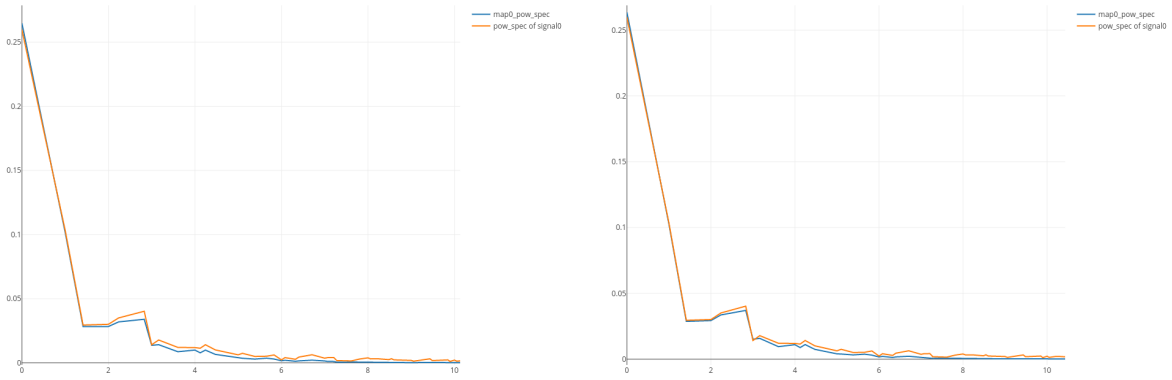
After the first visual comparison between reconstructions, it is helpful to compare the $k$-space power spectra of the fields. Here, only the first few $k$-modes are of high relevance, since the power spectra that created the fields are fastly converging to 0 with $k^{-5}$ and the reconstructed maps do the same even faster due to the prior and the difficulty of retrieving small scale modes in a noisy setup. But one should be aware that for creating the power spectrum, the $k$-modes are averaged for each $k$ value. This poses no problem for the original signal field since it is diagonal in Fourier space. This is not necessarily true though for the reconstruction. Still, this makes it possible to compare the reconstructions from only a few number differences (instead of judging reconstruction quality from $256^2$ pixels).

In figure 4.15, each plot contains a reconstruction power spectrum (blue) and the original power spectrum (orange) of the first component. The left reconstruction was done without prior knowledge about the cross-correlation. It is clearly visible that both reconstructions fit well to the original power spectrum. Especially the first few $k$-modes are nearly exactly correct. For the higher $k$-modes, the right reconstruction is slightly closer to the original, thereby affirming the former discussion.

The second component is compared in the same way as above in figure 4.16. Here again, both reconstructions are quite the same until the k value 2. After that value, the left reconstruction tends to 0 too fast. This can be seen in the reconstructions as the missing detail in the maps. For the reconstruction with prior knowledge about the cross-correlation the power spectrum better covers higher $k$-modes. Overall it is clear though that for both reconstructions the first component was much better captured, due to the higher absorption coefficients $\mu$. But this is probably also the reason for the effectiveness of the prior: The relatively good reconstruction of the first component even for higher $k$-values allows the cross-correlation to infer more details about the second component. This might be interesting to investigate and confirm in future work: Whether a badly recovered first component limits the usefulness of the cross-correlation.

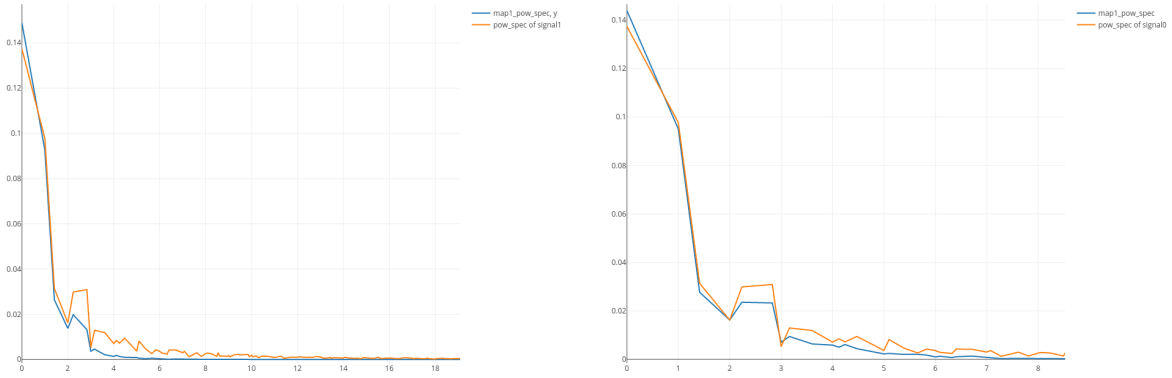
In conclusion, the reconstruction including the cross-correlation information performs better for this setup. Especially the harder-to-capture second component benefits from the advanced

prior knowledge. This means that using the informed setup does increase reconstruction performance.




**(a)** Cross-correlation unknown **(b)** Cross-correlation known

**Figure 4.15:** Power spectra for component 1. The orange graph codes for the ground truth power spectrum while blue is the spectrum obtained from the reconstruction map delivered by the algorithm. The left graph shows the reconstruction without prior knowledge of the cross-correlation. The right graph shows a slightly better performance in reconstruction.




**(a)** Cross-correlation unknown **(b)** Cross-correlation known

**Figure 4.16:** Power spectra for component 2. The orange graph codes for the ground truth power spectrum while blue is the spectrum obtained from the reconstruction map delivered by the algorithm. The left graph shows the reconstruction without prior knowledge of the cross-correlation. This figure shows the influence of the cross-correlation most clearly: Reconstruction results are quite visibly improved in the middle to small scale $k$-modes.

### 4.3.2 Sigmoidal correlation with high and low noise

In the following, the reconstructions for the 3 sigmoidal cross-correlation cases (uninformed low noise, uninformed higher noise and informed higher noise) are presented. Each reconstruction map will again be compared to the ground truth, followed by a comparison between the uninformed low noise and informed higher noise case.

#### Uninformed, low noise setup

As before, the first component is well captured in the reconstruction. Only smaller details fall away, as to be seen in figure 4.17. There is no real difference visible to the straight leap case except for the higher values in the first component (ground truth and reconstruction) due to the longer beginning of positive correlation between the components.

For the second component (figure 4.20) again only the larger scale structure is really captured well. The values for the ground truth are lower here compared to the straight correlation setup. This should make the reconstruction of this component even harder, though this is hard to judge from the position space images.

The overall performance of the algorithm for this different setup seems to be the same as for the uninformed straight leap setup, no matter the different underlying correlation structure of the signals.

#### Uninformed, higher noise setup

This setup was chosen to demonstrate how much information is lost by lowering the photon count by 20% from 5 million down to 4 million. Figures 4.18 and 4.21 present the results. The plots show that the first component did hardly loose structure. Component two, on the other hand, has visibly lost details. The general structure of the 'red islands' tends to be rounder, which is in contrast to the sharpness of the ground truth for these higher values of the second component.

This shows that lowering the sent-in photon count does not decrease reconstruction quality by much, though the performance loss is notable. It will now be interesting to see how this compares the to informed case.

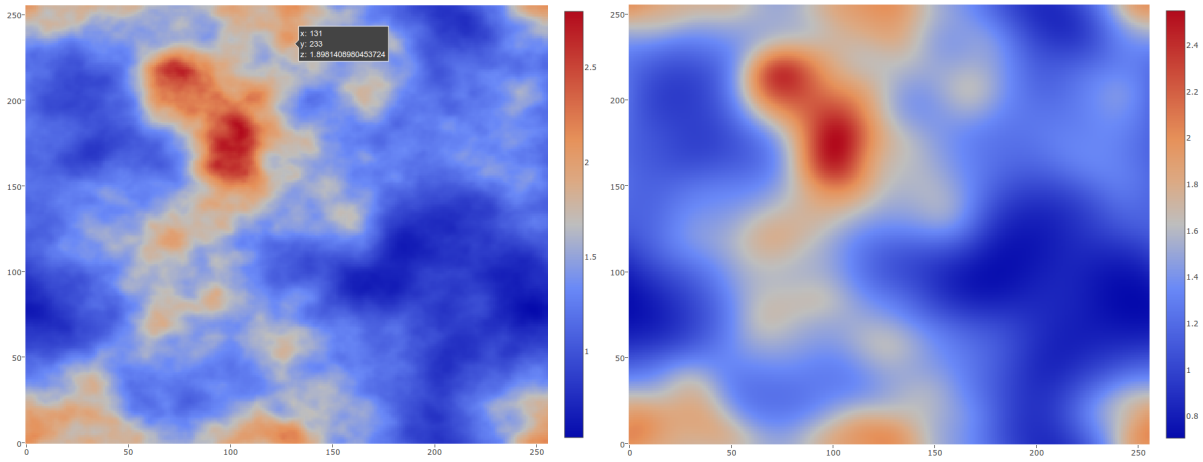
#### Informed, higher noise setup

Figures 4.19 and 4.22 show the reconstruction for the informed noisier case. By now it should come as no surprise that the reconstruction, especially for the second component, is superior to the uninformed case for equal noise. The first component is slightly better reconstructed at a close comparison between the two (the contours become sharper defined).

Interestingly, the first component reconstruction seems to still be slightly better than it was for the uninformed case with a million more photons sent-in. This becomes evident only when comparing the two close to each other. The contours of the informed reconstruction are a bit sharper for the first component.

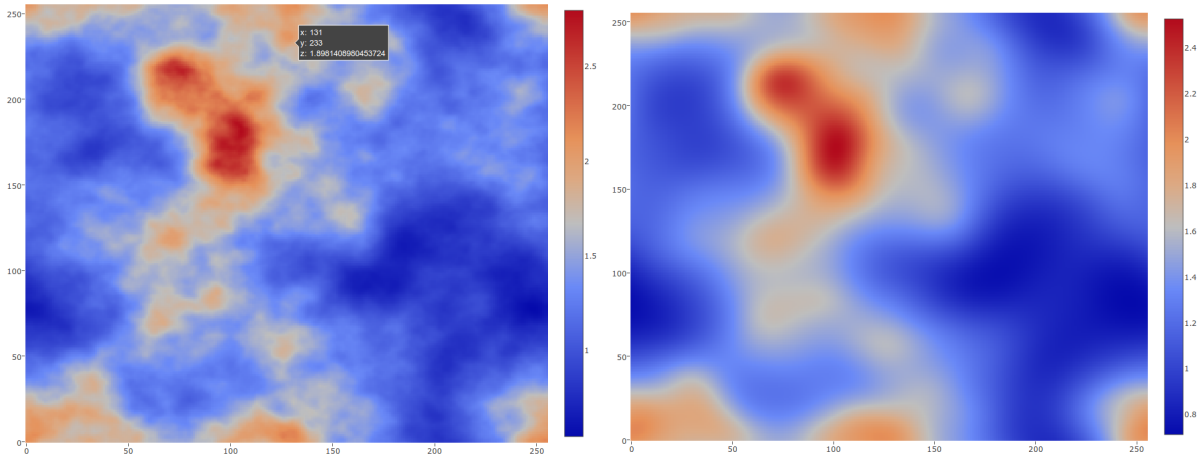
For the second component (figure 4.22) this comparison is much easier: The help of the additional prior information clearly dominates over the additional information through the extra photons given to the setup before. Again much more detail is displayed, not only in the shape of the red islands but also the blue, hard to capture areas of low absorption. This is

especially of value since the second component was the main weakness of the uninformed higher noise setup.



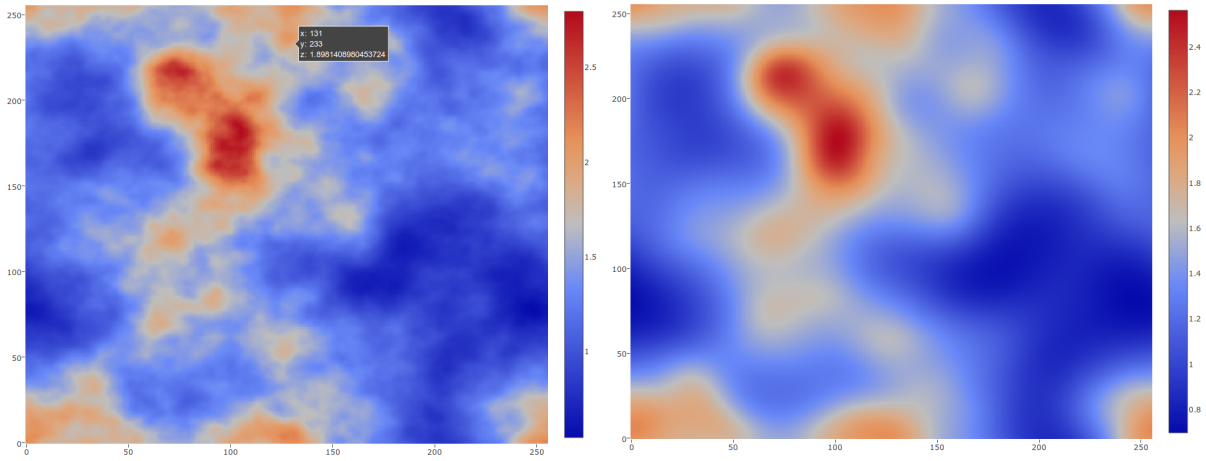

**(a)** Ground truth **(b)** Uninformed low noise reconstruction

**Figure 4.17:** Comparison between ground truth and reconstruction for component 1 for the sigmoidal cross-correlation between signals. As this field is identical to the first component in the straight leap correlation, differences to 4.11 are neither observed nor expected, as for both cases the correlation structure is not given to the reconstruction algorithm.



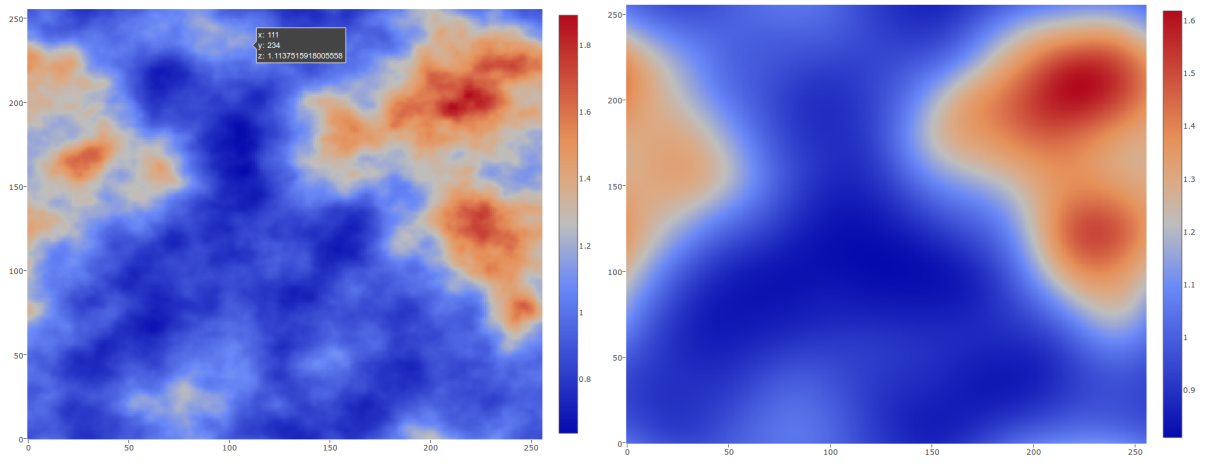

**(a)** Ground truth **(b)** Uninformed, noisy case reconstruction

**Figure 4.18:** Comparison between ground truth and reconstruction for component 1 for the sigmoidal cross-correlation between signals. As this field is identical to the first component in the straight leap correlation, differences to 4.11 are neither observed nor expected, as for both cases the correlation structure is not given to the reconstruction algorithm.



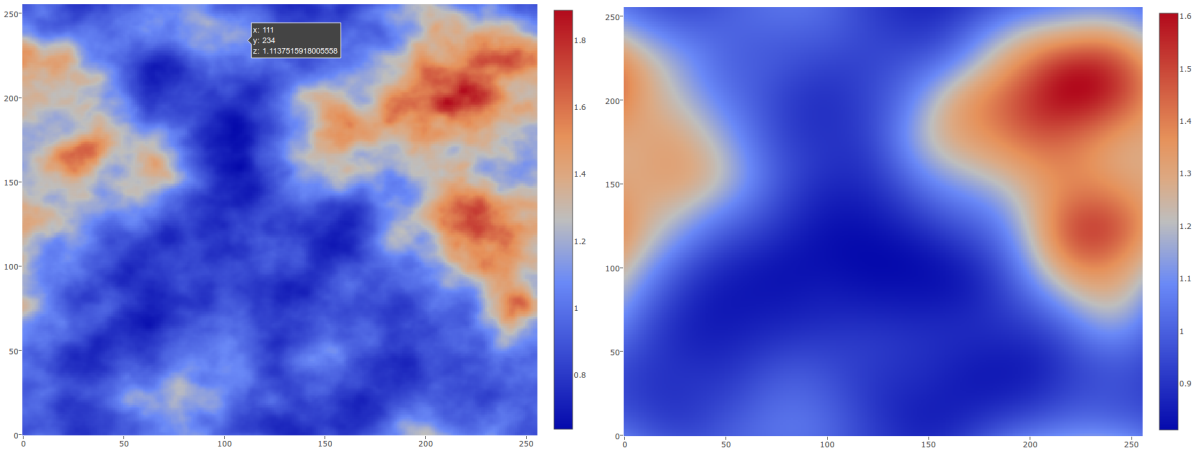

**(a)** Ground truth **(b)** Informed, noisy case reconstruction

**Figure 4.19:** Comparison between ground truth and reconstruction for component 1 for the case of known prior sigmoidal cross-correlation between signals. This reconstruction is highly similar to the reconstruction with known straight leap correlation. A close comparison between the two even reveals that the reconstruction in figure 4.11 is just slightly better in small areas.



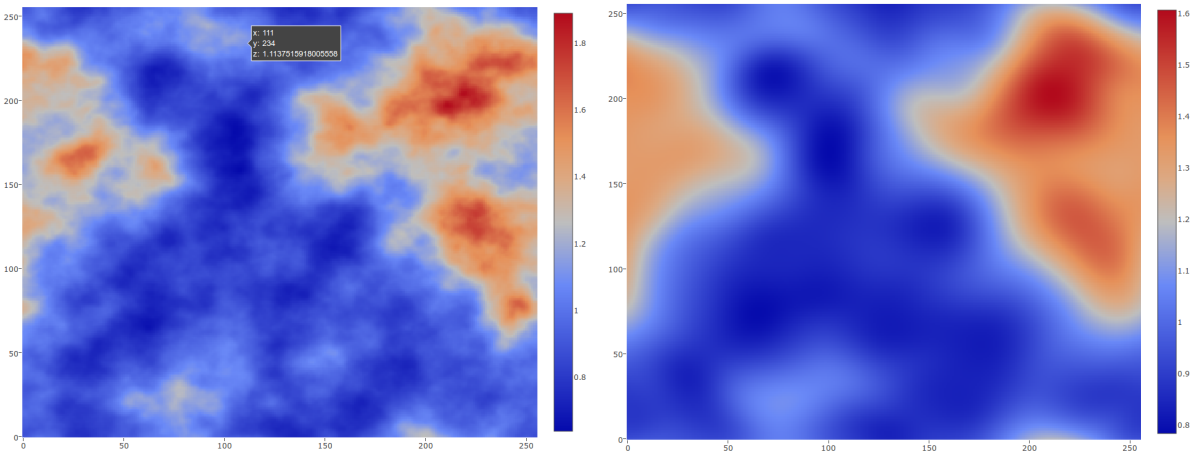

**(a)** Ground truth **(b)** Uninformed low noise reconstruction

**Figure 4.20:** Comparison between ground truth and reconstruction for component 2 (unknown prior sigmoidal correlation case). Similar to the straight leap correlation case without prior knowledge of the cross-correlation, the second component is only detected on large scales. Again large scales are well detected though.



**(a)** Ground truth **(b)** Uninformed, noisy case reconstruction

**Figure 4.21:** Comparison between ground truth and reconstruction for component 2 (known prior sigmoidal correlation case). Similar to the straight leap correlation case without prior knowledge of the cross-correlation, the second component is only detected on large scales. Again large scales are well detected though.



**(a)** Ground truth **(b)** Informed, noisy case reconstruction

**Figure 4.22:** Comparison between ground truth and reconstruction for component 2 (known prior sigmoidal correlation case). The difference to the reconstruction before is again striking. Much more detail is captured in this reconstruction.

**Comparison between setups**

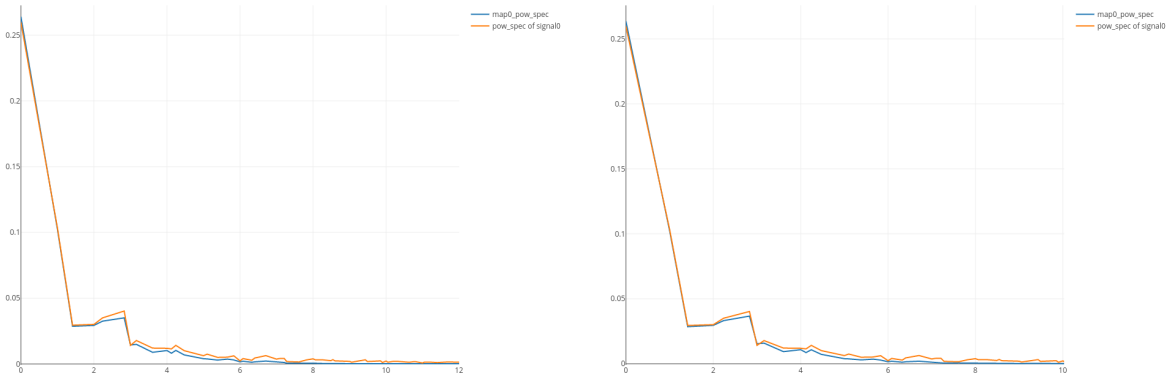
Since the results so far have shown, that the last case (higher noise informed case) is superior even to the reconstruction at a million more photons per line of sight (lower noise uninformed), this discussion compares only these two, that is the high noise informed and lower noise unformed cases. The power spectra comparisons shown in figures 4.23 and 4.24 confirm this observations. For both components, the informed reconstruction better represents the underlying ground truth signal.

For the first component, there is just a minor difference visible, similar to the straight leap correlation setup before (power spectrum after a $k$-value of 2 follows the ground truth more closely).

For the second component, it is rather evident that the informed reconstruction outperformed the lower noise uninformed case. Before the $k$-value 2, both graphs are nearly identical. Both overestimate the 0th mode and underestimate the 1st mode. Yet, after a k value of 2, the right graph in figure 4.24 is much closer to the ground truth. The given cross-correlation again enables the reconstruction to better achieve higher $k$-modes and thereby smaller scale structures.

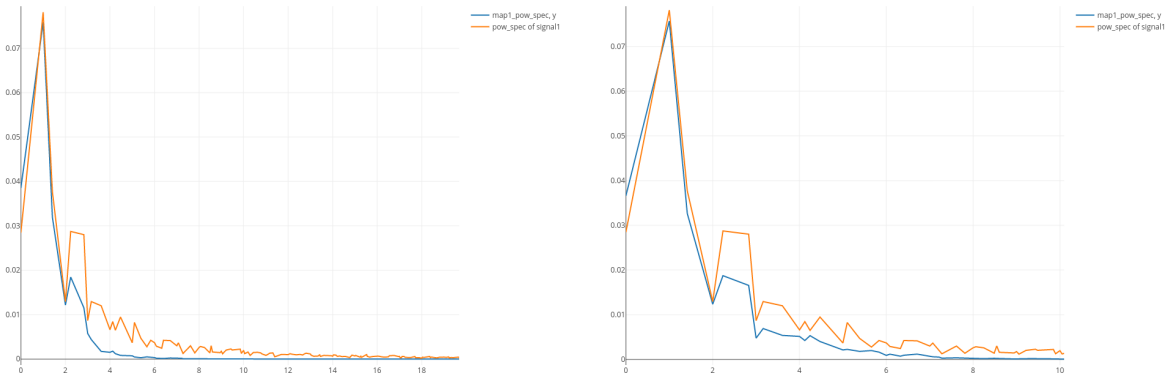
This means that the prior knowledge of the cross-correlation really enables the experimenter to deliberately choose a lower radiation exposure while being able to expect the same quality reconstruction. But two important issues need to be considered. Firstly, the quality of the images did not worsen by a big factor from one noise scenario to the next. At a lower incoming photon rate, the relative difference of incoming photons that the cross-correlation can compensate is very likely smaller than demonstrated here. At least when lowering the sent-in photon count becomes visible in the reconstruction of the first component, the effects of the cross-correlation should decrease (judging from the data presented here, though this needs further investigation in order to confirm this).

Secondly, the case demonstrated here is not yet exactly what was demanded in the introduction (achieving a better reconstruction resolution for the same photon illumination with an improved algorithm). For the algorithm presented here, the reconstruction task was first complicated by introducing components into the reconstruction in order to then leverage knowledge about the components. Usually, in medical imaging, the result of a scan is just a picture of the total absorption factor. The identification of bone and other tissue has to be done by the radiologist. The task outlined in the introduction would be to create an equal quality picture of the total absorption factor with less radiation exposure. This remains to be shown in future work. But this is definitely a promising start. The possibility that an algorithm might automatically deliver labels of the respective components together with the reconstruction is an interesting improvement to the reconstruction process.




**(a)** Uninformed, low noise case **(b)** Informed, noisy case

**Figure 4.23:** Power spectra for component 1 for the two preceding reconstruction maps. The orange graph codes for the ground truth power spectrum while blue is the spectrum obtained from the reconstruction maps delivered by the algorithm. The left graph shows the reconstruction without prior knowledge of the cross-correlation and at 5 million photon illumination. The right graph again shows a slightly better performance in reconstruction.




**(a)** Uninformed, low noise case **(b)** Informed, noisy case

**Figure 4.24:** Power spectra for component 2 for the two preceding reconstruction maps. The orange graph codes for the ground truth power spectrum while blue is the spectrum obtained from the reconstruction maps delivered by the algorithm. The left graph shows the reconstruction without prior knowledge of the cross-correlation and at 5 million photon illumination (uninformed low noise case). Again the prior influence is striking in the middle to small scale $k$-modes, where the blue graph is more similar to the orange graph.

# Chapter 5

# Conclusion and Outlook

In this thesis, a Bayesian component separation algorithm was derived for a medical tomography inspired setup using the language and framework of information field theory. This algorithm was implemented in Python utilizing especially the NIFTy package. The components were inferred using two energy channels with different (and assumed to be known) absorption factors for each component. Additionally, it was examined how the prior knowledge of the inter-component cross-correlation can improve the reconstruction quality.

It was shown that the algorithm developed here is capable of reconstructing two components from two distinct energy channel measurements for different correlation structures. Moreover, evidence is provided that prior knowledge about the inter-component cross-correlation improves the reconstruction quality. The last setup comparison then showed that lowering the sent-in photon count could be compensated for by the information contained in the cross-correlation.

Much work remains to be done to turn this work into a reliable reconstruction algorithm for real medical tomography. Since the main goal is to achieve a better or equal reconstruction resolution at the same or less radiation exposure, further investigations of the behavior of the reconstructions under changed noise levels will be of interest. That could show whether the contribution of the cross-correlation is higher, lower or the same, depending on the radiation exposure (the overall amount of photons sent-in).

Furthermore, the background was completely ignored in this thesis. Future work could use techniques to estimate the background and take it into consideration in the signal reconstruction.

To further establish the algorithm, it should be tested on real medical tomography data. As a transition to this, a certain CT setup could be used to determine real absorption coefficients for bone and tissue and also adjust the number of photons that are emitted for each energy channel. Then high quality data from high radiaton observations of dead human corpses could be studied to find the correlation structure of and between these two components, resulting in a realistic cross-correlation. This could result in a mock data setup that is much closer to reality. Then the transition to real data with its noise and background would be the next logical step.

Furthermore, it would be interesting to see how the reconstruction handles more than two components. Each new component introduces a lot of new free parameters, yet it can also bring with it information with the inter-component cross-correlation. It would be interesting to know what an optimal number of components would be. This is probably highly dependent on the absorption coefficients of the additional components though, therefore a more realistic setup (absorption coefficients $\mu$ determined from real data) should be pursued first.

Lastly, comparing the performance of this algorithm to a classical filtered backprojection could be a good benchmark for evaluating the performance of the presented algorithm.

# Bibliography


[1] Xiaochuan Pan et al., *Why do commercial CT scanners still employ traditional, filtered back-projection for image reconstruction?*, PubMed Central (PMC), january 2009.

[2] Torsten Ensslin, *Information Theory and Information Field Theory Lecture Notes*, http://wwwmpa.mpa-garching.mpg.de/ ensslin/lectures/lectures.html, last visited 30/08/17.

[3] Ensslin et al., *Information field theory for cosmological perturbation reconstruction and non-linear signal analysis*, arXiv0806.3474, september 2009.

[4] Leike and Ensslin, *Optimal Belief Approximation*, arXiv1610.09018, august 2017.

[5] Ensslin and Knollmueller, *Correlated signal inference by free energy exploration*, arXiv1612.08406, february 2017.

[6] Jeffrey A. Fessler, *Statistical Image Reconstruction Methods for Transmission Tomography*, SPIE Handbook of Medical Imaging, Vol. 1, 2000, november 2002.

[7] Leike and Ensslin, *Operator Calculus for Information Field Theory*, arXiv1605.00660, october 2016.

[8] J. E. Campbell, *Proceedings of the London Mathematical Society* 1, 14, (1897)

[9] Khintchin, A., *Korrelationstheorie der stationären stochastischen Prozesse.*, Mathematische Annalen, 109(1):604–615, 1934

[10] Robin et al., *Cross-Correlated Random Field Generation With the Direct Fourier Transform Method*, Water Resources Research, Vol.29, No. 7, 1993

[11] Rosenblatt, *Random Processes*, Oxford University Press, 1962

[12] Steininger et al., *NIFTy 3 – Numerical Information Field Theory*, arXiv1708.01073, august 2017.

# List of Figures

# Acknowledgements

I would like to thank all the people that contributed towards this thesis. Foremost my thanks go to Torsten for many discussions about data model and calculations. Furthermore, Reimar and Jakob were a great help for for resolving problems in calculations and implementations. Moreover, Theo and Jakob provided helpful advice for how to best implement the new classes into the nifty framework. Lastly, I want to thank Torsten for the opportunity to broaden and deepen my understanding of information theory and machine learning, not only through this thesis, but also by letting me contribute as a tutor for the lecture. The thesis brought many coding challenges with it and was very valuable for learning plenty of lessons.

# Declaration

Hiermit erkläre ich, dass ich die Arbeit selbstständig verfasst habe, und dass ich keine anderen Quellen und Hilfsmittel, als die angegebenen, benutzt habe.

---

München, Robin Dehde