Diplomarbeit

Bayesian Sampling Methods for Optical Coherence Tomography

ausgeführt am
Institut für Nachrichtentechnik und Hochfrequenztechnik (E 389)
der Technischen Universität Wien

unter der Leitung von

Univ.-Ass. Dipl.-Ing. Clemens Novak
Ao. Univ.-Prof. Dipl.-Ing. Dr. Franz Hlawatsch

von
Georg Kail
Schottenring 19/10
A-1010 Wien

Wien, Dezember 2007 _____

Acknowledgements

The advice and help I have received from my supervisors in the course of writing this thesis deserves great appreciation.

I am grateful to Dipl.-Ing. Clemens Novak for his continuous support and guidance, for countless improvements of the thesis, and for always being available for discussions.

I am indebted to Prof. Franz Hlawatsch for encouraging my work and showing great interest in it, and for his active contributions.

My thanks go to Dipl.-Ing. Bernd Hofer for introducing me to the topic and supplying up-to-date information.

Finally, I would like to thank my parents, who have enabled me to achieve everything I have, and her who has been filling my days with optimism.

Abstract

Bayesian sampling methods constitute a powerful tool for signal detection and estimation. They exploit the virtues of Bayesian estimation, concise modeling of the problem and efficient use of all prior information available, while being computationally feasible and apt for efficient implementation. Therefore they offer practical solutions to a wide field of problems in statistical signal processing [1].

This thesis describes the application of a Bayesian sampling method, the Gibbs sampler [2], on signals from the domain of Optical Coherence Tomography (OCT) [3]. OCT is a state-of-the-art imaging technology, which is used for high-resolution imaging of the outer layer of biological tissue. The task of extracting the relevant information from OCT images is described as a joint detection and estimation problem by formulating a statistical system model which represents the physical processes involved in the generation of OCT images, as well as the known statistical properties of the tissue.

With the help of this model, a Gibbs sampler is developed with the purpose of locating and quantifying steps in the depth profile of the tissue's refractive index, based on the distorted observed signal, i.e. the OCT image. The system model and the respective algorithms for synthesis as well as detection and estimation of OCT signals are gradually refined in order to achieve more realistic models for OCT signals.

In order to compare the estimator's performance with an existing method, the Single Most Likely Replacement (SMLR) algorithm, which has been used successfully in the domain of OCT [4], is adapted to the newly defined system model.

All algorithms derived and discussed in the course of this thesis are applied to synthetic as well as real OCT signals to compare their outcome.

Zusammenfassung

Bayes'sche Stichprobenverfahren stellen ein wirkungsvolles Werkzeug für die Signaldetektion und -schätzung dar. Sie nützen die Stärken der Bayes'schen Schätzung, eine umfassende Modellierung des Problems sowie die Nutzung der gesamten verfügbaren a priori-Information, sind aber dabei rechnerisch durchführbar und geeignet für eine effiziente Implementierung. Somit bieten sie praktische Lösungen für ein weites Problemfeld in der statistischen Signalverarbeitung [1].

Diese Arbeit beschreibt die Anwendung eines Bayes'schen Stichprobenverfahrens, des Gibbs-Samplers [2], auf Signale aus dem Gebiet der Optischen Kohärenztomographie (OCT) [3]. OCT ist ein hochmodernes Bildgebungsverfahren, das zur hochauflösenden Abbildung der äußeren Schicht biologischen Gewebes verwendet wird. Die Aufgabenstellung, die gesuchte Information aus OCT-Bildern auszulesen, wird als kombiniertes Detektions- und Schätzproblem beschrieben. Dazu wird ein statistisches Modell des Systems formuliert, welches die physikalischen Vorgänge beim Entstehen eines OCT-Bildes sowie die statistischen Eigenschaften des Gewebes beschreibt.

Mithilfe dieses Modells wird ein Gibbs-Sampler entwickelt, dessen Aufgabe es ist, aufgrund des verzerrten beobachteten Signals, d.h. des OCT-Bildes, Stufen im Tiefenprofil des Brechungsindex des Gewebes zu finden und zu messen. Das Systemmodell und die entsprechenden Algorithmen zur Synthese als auch zur Detektion und Schätzung von OCT-Signalen werden schrittweise verfeinert und angepasst, um realistischere Modelle für OCT-Signale zu erreichen.

Um die Leistungsfähigkeit des Schätzers mit einer vorhandenen Methode zu vergleichen, wird der Single Most Likely Replacement (SMLR) -Algorithmus, der schon erfolgreich im Gebiet der OCT angewendet wurde [4], an das neu definierte Systemmodell angepasst.

Alle im Zuge dieser Arbeit hergeleiteten und besprochenen Algorithmen werden auf synthetische sowie auf reale OCT-Signale angewendet, um ihre Ergebnisse zu vergleichen.

Contents

Chapter 1

Introduction

Optical Coherence Tomography (OCT) [3] has found its unchallenged place in the field of medical diagnosis due to a number of specific strengths, most prominently among them its high depth resolution at a respectable penetration depth, and its quality as a non-invasive imaging method. However, the images produced by OCT and used for diagnosis are largely influenced by the process of their generation and various physiological effects, and far from displaying purely the desired information. Extracting this information is the task of signal processing which is performed on the raw OCT scans.

The problem can be - and has been [4] - described as that of detecting and estimating a sparse signal from an observed signal which evolved from the former by convolution. The method discussed here, the Gibbs sampler [2], seems promising because it claims to be of tolerable complexity, while at the same time not being a priori suboptimal. Gibbs samplers enjoy great popularity in signal processing, in fact they have been proposed as a solution for numerous similar problems [5–7]. However, particularly our constraint to sparse signals requires a special treatment.

The intention of this thesis is to give an introduction to the field of Bayesian sampling first, then to the domain of OCT and its specific requirements for the estimator. Subsequently, a solution for joint detection and estimation is to be derived using a Gibbs sampler. The Single Most Likely Replacement algorithm, used on OCT signals in [4], is to be refined and used as a reference to assess the performance of the new method.

Chapter 2

Bayesian Sampling Methods

2.1 Bayesian Estimation

2.1.1 Basic Concept

The basic framework of estimation, a set of parameters $\boldsymbol{\theta}$ which cannot be observed directly, and a set of observed quantities $\mathbf{y}$, which depend on the parameters, is set into a statistical context in the Bayesian estimation theory. The explicit quantification of uncertainty in terms of probability distributions, namely for describing the parameter, is the essential characteristic of Bayesian estimation [8]. Both the parameters $\boldsymbol{\theta}$ and the observation $\mathbf{y}$ are modeled as random, and some statistical knowledge about them is available, which is ultimately expressed in the posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ [8].

Since this statistical knowledge is usually not available in the shape of a precise distribution, it is useful to define a model of the problem domain, which summarizes all the a priori knowledge and supplements it with additional assumptions to allow the derivation of a posterior distribution. The latter describes the interdependence of parameters and observation, or, more precisely, the dependence of the parameters' probability distribution on the observation. Estimates are obtained by inference from the data, i.e. by inserting $\mathbf{y}$ into $f(\boldsymbol{\theta}|\mathbf{y})$ and calculating statistics of the distribution.

The interrelation of parameters and observation is illustrated in Fig. 2.1.

2.1.2 Estimators

Two standard estimators that will be used in our context are the MMSE estimator and the MAP estimator.

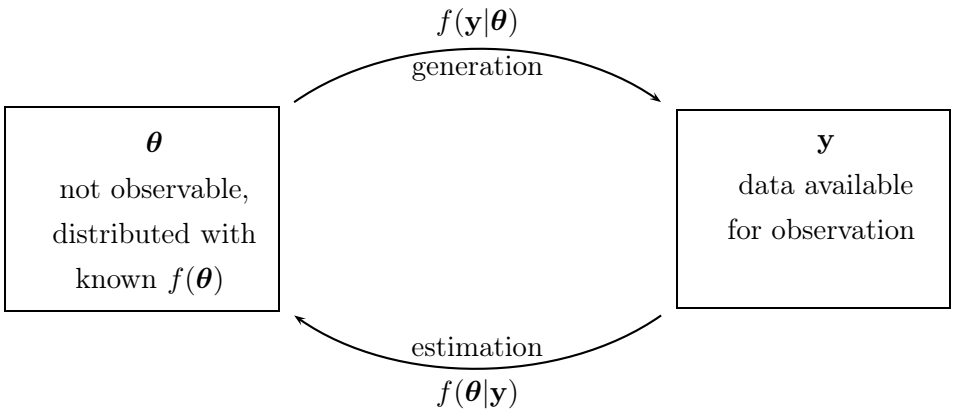


Figure 2.1: Interdependence of parameters and observation in a Bayesian setting.

The Minimum Mean Square Error (MMSE) estimator minimizes the expectation of the squared error, which leads to the mean of the posterior distribution:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\mathrm{MMSE}}(\mathbf{y}) &= \arg\min_{\hat{\boldsymbol{\theta}}} \mathrm{E}\{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2 \mid \mathbf{y}\} \\
&= \arg\min_{\hat{\boldsymbol{\theta}}} \mathrm{E}\{\|(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}}) - (\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}})\|^2 \mid \mathbf{y}\} \\
&= \arg\min_{\hat{\boldsymbol{\theta}}} \mathrm{E}\{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}}\|^2 + \|\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}}\|^2 - 2(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}})^{\mathrm{H}}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}}) \mid \mathbf{y}\} \\
&= \arg\min_{\hat{\boldsymbol{\theta}}} \Big(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}}\|^2 + \underbrace{\mathrm{E}\{\|\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}}\|^2 \mid \mathbf{y}\}}_{\text{independent of } \hat{\boldsymbol{\theta}}} - 2(\hat{\boldsymbol{\theta}}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}})^{\mathrm{H}} \underbrace{\mathrm{E}\{\boldsymbol{\theta}-\boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}} \mid \mathbf{y}\}}_{=\mathbf{0}}\Big) \\
&= \boldsymbol{\mu}_{\boldsymbol{\theta}|\mathbf{y}} = \mathrm{E}\{\boldsymbol{\theta}|\mathbf{y}\}.
\end{aligned}$$

The Maximum A Posteriori (MAP) estimator is even more intuitive; it finds the argument that maximizes the posterior probability (or probability density, for continuous distributions):

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}(\mathbf{y}) = \arg\max_{\hat{\boldsymbol{\theta}}} f(\hat{\boldsymbol{\theta}}|\mathbf{y}).$$

2.1.3 Deriving the Posterior Distribution

The statistical framework necessary for deriving the posterior distribution is expressed as a joint probability distribution $f(\boldsymbol{\theta}, \mathbf{y})$, which must be known or assumed for estimation. Alternatively, the two distributions $f(\boldsymbol{\theta})$ and $f(\mathbf{y}|\boldsymbol{\theta})$ may be given. These formulations are equivalent due to the relations

$$f(\boldsymbol{\theta}, \mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \qquad \Leftrightarrow \qquad \begin{cases} f(\boldsymbol{\theta}) & = \int f(\boldsymbol{\theta}, \mathbf{y}) \mathrm{d}\mathbf{y} \\ f(\mathbf{y}|\boldsymbol{\theta}) & = \frac{f(\boldsymbol{\theta}, \mathbf{y})}{f(\boldsymbol{\theta})}. \end{cases}$$

Most times, defining $f(\boldsymbol{\theta})$ and $f(\mathbf{y}|\boldsymbol{\theta})$ is the most intuitive way of formulating the statistical framework of the problem, considering that $\boldsymbol{\theta}$ is stochastically transformed into $\mathbf{y}$, which is then observed in lack of direct access to $\boldsymbol{\theta}$. This process is reflected by the distributions as follows:

- $f(\boldsymbol{\theta})$, the prior distribution of $\boldsymbol{\theta}$, contains the statistical knowledge about $\boldsymbol{\theta}$ independent of $\mathbf{y}$
- $f(\mathbf{y}|\boldsymbol{\theta})$, the distribution of $\mathbf{y}$ given $\boldsymbol{\theta}$, or the likelihood function, describes how $\mathbf{y}$ emerges from $\boldsymbol{\theta}$

The distribution that is ultimately used by all Bayesian estimators, $f(\boldsymbol{\theta}|\mathbf{y})$, can be derived from this information using Bayes' rule:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\boldsymbol{\theta},\mathbf{y})}{f(\mathbf{y})} = \frac{f(\boldsymbol{\theta},\mathbf{y})}{\int f(\boldsymbol{\theta'},\mathbf{y})\mathrm{d}\boldsymbol{\theta'}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta'})f(\boldsymbol{\theta'})\mathrm{d}\boldsymbol{\theta'}}. \tag{2.1}$$

2.1.4 Simplifications for Calculation

Statistically Independent Parameters. We can split the parameters $\boldsymbol{\theta}$ into the parameters $\boldsymbol{\theta}_1$, which we are currently interested in, and the remaining parameters $\boldsymbol{\theta}_2$, which are assumed to be known:

$$f(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2) = \frac{f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\mathbf{y})}{\int f(\boldsymbol{\theta}'_1,\boldsymbol{\theta}_2,\mathbf{y})\mathrm{d}\boldsymbol{\theta}'_1} = \frac{f(\mathbf{y}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)}{\int f(\mathbf{y}|\boldsymbol{\theta}'_1,\boldsymbol{\theta}_2)f(\boldsymbol{\theta}'_1,\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}'_1}. \tag{2.2}$$

If these two sets are statistically independent, the calculation can be simplified:

$$f(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2) = \frac{f(\mathbf{y}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_1)f(\boldsymbol{\theta}_2)}{\int f(\mathbf{y}|\boldsymbol{\theta}'_1,\boldsymbol{\theta}_2)f(\boldsymbol{\theta}'_1)f(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}'_1} = \frac{f(\mathbf{y}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_1)}{\int f(\mathbf{y}|\boldsymbol{\theta}'_1,\boldsymbol{\theta}_2)f(\boldsymbol{\theta}'_1)\mathrm{d}\boldsymbol{\theta}'_1}. \tag{2.3}$$

Conjugate Prior Distributions. Conjugate prior distributions of a given likelihood function are distributions which retain their shape when multiplied to the respective likelihood function. Therefore a parameter distributed with a conjugate prior distribution will have a posterior distribution of the same family [9]. Knowing this in advance considerably simplifies its calculation because we can neglect normalizing constants (including the entire denominator in (2.1) ff.) and need to find only the shape parameters of the respective distribution.

It should be noted that the likelihood function, while usually being written as a function of the observation, is generally a function of a different shape with respect to the parameter. For example, $f(y|\mu,\sigma^2) = \mathcal{N}(\mu,\sigma^2)$ is, coincidentally, a Gaussian function of the parameter μ, but not - as is the general case - of the parameter σ^2. Furthermore, the simplicity of this example should not distract from the fact that both μ and σ^2 may be arbitrary functions of the parameter, resulting in yet different families of conjugate priors. Despite the notation of the likelihood function as a

conditional distribution of the observation it should be noted that, in general, it does not have the properties of a probability distribution with respect to the parameter.

While the prior family is sufficiently determined by the shape of the likelihood function with respect to the parameter, it is often more intuitive (and therefore common use in lookup tables [10]) to specify the shape of the likelihood function with respect to the observation and the dependence of its shape variables on the parameter.

Example 2.1.4 To take a more detailed look at another example, let us assume a simple estimation problem with a parameter θ, a known fading factor c and additive zero-mean Gaussian noise. The observation can then be expressed as

$$y = c\ \theta + n.$$

Its likelihood function is fully determined by the above assumptions:

$$f(y|\theta) = \mathcal{N}(c\ \theta, \sigma_n^2), \tag{2.4}$$

where σ_n^2 denotes the noise variance. A normal distribution with respect to y, the likelihood function now has to be examined as a function of θ:

$$f(y|\theta) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y - c\ \theta)^2}{2\sigma_n^2}\right) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(\theta - y/c)^2}{2\sigma_n^2/c^2}\right).$$

As a consequence of our model assumptions, this is again a Gaussian function, with mean y/c and variance σ_n^2/c^2, albeit not normalized. The shape of this function in turn determines the set of conjugate prior distributions for θ as all distributions which retain their shape (with respect to θ) when multiplied with the likelihood function. In other words, the multiplication may only change the shape variables of the distribution and introduce a constant factor, which will be eliminated by normalization. Obviously in this simple case, the normal distribution is a conjugate prior, since its multiplication with a Gaussian function will do precisely that. Defining the prior as

$$f(\theta) = \mathcal{N}(\mu_\theta, \sigma_\theta^2)$$

yields a posterior distribution

$$\begin{aligned} f(\theta|y) \propto f(y|\theta)f(\theta) &= \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(\theta - y/c)^2}{2\sigma_n^2/c^2}\right) \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{(\theta-\mu_\theta)^2}{2\sigma_\theta^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\left(\frac{c^2}{\sigma_n^2}+\frac{1}{\sigma_\theta^2}\right)\theta^2 - 2\left(\frac{c\,y}{\sigma_n^2}+\frac{\mu_\theta}{\sigma_\theta^2}\right)\theta\right]\right) \\ &\propto \exp\left(-\frac{(\theta-\mu_{\text{post}})^2}{2\sigma_{\text{post}}^2}\right) \\ \Rightarrow f(\theta|y) &= \mathcal{N}(\mu_{\text{post}}, \sigma_{\text{post}}^2) \qquad (2.5) \\ &\text{with} \quad \sigma_{\text{post}}^2 = \left(\frac{c^2}{\sigma_n^2}+\frac{1}{\sigma_\theta^2}\right)^{-1} \quad \text{and} \quad \mu_{\text{post}} = \sigma_{\text{post}}^2\left(\frac{c\,y}{\sigma_n^2}+\frac{\mu_\theta}{\sigma_\theta^2}\right) \end{aligned}$$

In practice, the conjugate prior can usually not be recognized as easily just by looking at the likelihood function. Instead, we would look it up in a table of standard likelihood functions and their respective conjugate prior distributions. To stick to our example, we would find the solution by looking for "Normal distribution" (for the conditional distribution of the observation) and μ/c (for the parameter). □

The above definition of conjugate priors relies on statistically independent parameters as in (2.3). If this is not the case the shape of the conjugate priors depends not only on the likelihood function but also on the conditional distribution of the other parameters given the parameter of interest. This becomes obvious when comparing (2.2) and (2.3).

2.2 Monte Carlo Sampling Methods

Monte Carlo sampling methods aim at solving the problem of calculating statistics of complex distributions. They represent a numerical solution rather than an analytical one. They have been shown to be particularly useful in Bayesian estimation, since computational challenges with complex posterior distributions often limit the possibilities of their analytical evaluation [2].

The key idea of the Monte Carlo approach is to avoid calculating statistics of the distribution itself by drawing a large sample from it and calculating the statistics of this ensemble instead.

For example, the mean of the distribution can be approximated as:

$$\mathrm{E}\{\mathbf{x}\} \approx \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} \quad \text{with} \quad \mathbf{x}^{(i)} \sim f(\mathbf{x})$$

where N is the sample size.

The inherent error of this approximation can be made arbitrarily small if

- the sample size N is large enough,
- the sample $\{\mathbf{x}^{(i)}\}$ is sufficiently uncorrelated.

The term "sufficiently uncorrelated" deserves some further consideration: The condition would be optimally satisfied by an i.i.d. sample. However, given that the posterior distribution derived from the likelihood function and the prior is complex and possibly highdimensional, it will be challenging or even impossible to draw samples from it. Besides, the condition is less strict than demanding an i.i.d. sample. As long as the entire ensemble resembles the proportions of the desired distribution appropriately, correlation between samples is irrelevant [11]. Therefore, if the correlation does not restrict the samples from any part of the sample space, its effect can be compensated by a large enough sample size, leading to a tradeoff between the two above conditions. A comparably convenient way of generating an appropriate sample is by using a Markov chain with the stationary distribution $f(\mathbf{x})$, as will be explained in section 2.3.

Example 2.2 The MC approximation can easily be illustrated by an example. We will continue with the setup outlined in the example on page 6. Due to its simplicity any statistics can be calculated analytically, which makes the use of an MC estimator unnecessary, but at the same time this allows us to compare results and check its validity. The true MMSE estimate of θ is the mean of its posterior distribution (given in (2.5)):

$$\hat{\theta}_{\mathrm{MMSE}} = \mathrm{E}\{\theta|y\} = \mu_{\mathrm{post}} = \sigma^2_{\mathrm{post}} \left(\frac{c\ y}{\sigma_n^2} + \frac{\mu_\theta}{\sigma_\theta^2} \right) \quad \text{with} \quad \sigma^2_{\mathrm{post}} = \left(\frac{c^2}{\sigma_n^2} + \frac{1}{\sigma_\theta^2} \right)^{-1}$$

The MC approximation is the mean of a large number of samples from the posterior distribution $\mathcal{N}(\mu_{\mathrm{post}}, \sigma^2_{\mathrm{post}})$. In this simple case, samples can be obtained directly by i.i.d. draws from the distribution.
For comparison, we will also consider a non-Bayesian estimator, the Maximum Likelihood estimator, which ignores the prior information and instead maximizes the conditional probability (density) $f(y|\theta)$, i.e. the likelihood function, which was given in (2.4):

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_\theta f(y|\theta) = \arg\max_\theta \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y - c\ \theta)^2}{2\sigma_n^2} \right) = \arg\min_\theta (y - c\ \theta)^2 = \frac{y}{c}$$

Results of an exemplary implementation of these three estimators are shown in Fig. 2.2, comparing the resulting Mean Square Error averaged over several experiments. □

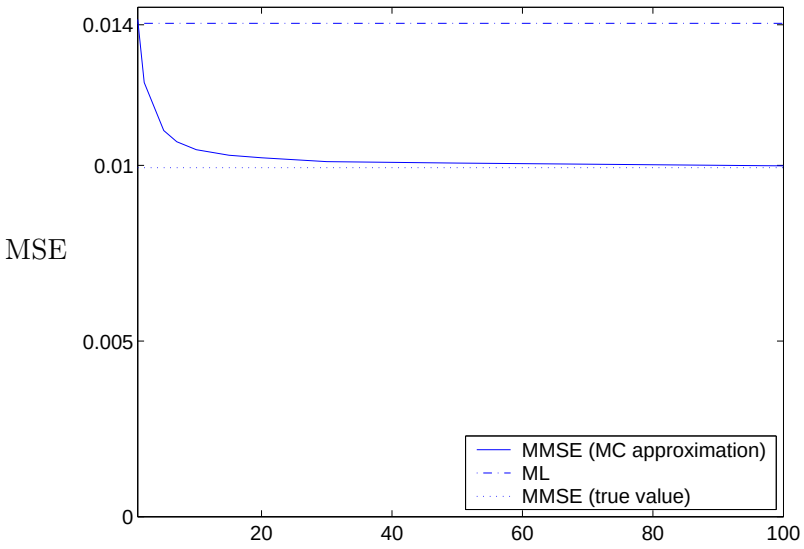


Figure 2.2: MSE obtained in experiments as described in Example 2.2, with $\mu_\theta = 0$, $\sigma_\theta^2 = 1$, $c = 1$, $\sigma_n^2 = 1$. With increasing sample size, the MC estimator approximates the true MMSE estimate, yielding a result that is clearly better than the non-Bayesian ML estimate.

Monte Carlo Integration. The original application of the Monte Carlo approach aimed at solving complex integrals by expressing them as an expectation, which could then be approximated by a sample average. If, for example, a function $s(t)$ can be expressed using a probability density function $f(t)$, such that

$$q(t)f(t) = \begin{cases} s(t) & \text{for } a \leq t \leq b \\ 0 & \text{otherwise,} \end{cases}$$

then its integral can be represented as follows:

$$\int_a^b s(t)\mathrm{d}t = \int_{-\infty}^{\infty} q(t)f(t)\mathrm{d}t = \mathrm{E}_{f(t)}\{q(t)\},$$

where $\mathrm{E}_{f(t)}\{\cdot\}$ denotes that t is distributed according to $f(t)$. Instead of this expectation, a large sample $\{t^{(i)}\}$ is drawn from $f(t)$, and its average is used as an approximation for the integral:

$$\int_a^b s(t)\mathrm{d}t = \mathrm{E}_{f(t)}\{q(t)\} \approx \frac{1}{N}\sum_{i=1}^{N} q\left(t^{(i)}\right).$$

Importance Sampling [2, 12]. In a similar manner we can approximate statistics of a distribution $\varphi(t)$ without drawing samples from $\varphi(t)$ itself, but rather from $f(t)$. As an example, we

shall derive the expectation of some function $q(t)$:

$$\int_{-\infty}^{\infty} q(t)\varphi(t)\mathrm{d}t = \int_{-\infty}^{\infty} q(t)\frac{\varphi(t)}{f(t)}f(t)\mathrm{d}t = \mathrm{E}_{f(t)}\left\{q(t)\frac{\varphi(t)}{f(t)}\right\} \approx \frac{1}{N}\sum_{i=1}^{N} q\left(t^{(i)}\right)\frac{\varphi\left(t^{(i)}\right)}{f\left(t^{(i)}\right)}.$$

The distribution of interest $\varphi(t)$ may be to complex for drawing samples from it. Importance sampling avoids this and requires only the evaluation of $\varphi(t)$ for the samples drawn from the more convenient distribution $f(t)$. The latter may be chosen arbitrarily, however it must not be zero for any t with $q(t)\varphi(t) > 0$.

2.3 Markov Chain Monte Carlo (MCMC) Methods

Markov Chain Monte Carlo methods [1] simulate direct draws from a distribution by implementing a Markov Chain, which repeatedly generates a new sample based on the previous sample value.

A Markov chain is a stochastic process with the specific property that given the present state, future states are conditionally independent of the past. A particular Markov chain is therefore defined by its transition probability distribution:

$$\mathbf{x}^{(i)} \quad \xrightarrow[f(\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)})]{} \quad \mathbf{x}^{(i+1)}.$$

While Markov Chain in the strict sense of the word refers to a process with discrete states and distributions, we will continue using the expression while describing the analogous concept with continuous distributions.

The transition distribution can be used to obtain the probability distribution of any future state, given the present one. These probability distributions at specific points in time are called the marginal distributions. The marginal distribution is linked to that of the previous state as follows:

$$f(\mathbf{x}^{(i+1)}) = \int f(\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)})f(\mathbf{x}^{(i)})\mathrm{d}\mathbf{x}^{(i)}. \tag{2.6}$$

If the marginal distribution remains unchanged after a transition, i.e. if $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(i+1)}$ have the same distribution $f(\mathbf{x})$, this is called a stationary distribution of the Markov chain.

Therefore, once a Markov chain has reached a stationary distribution, it will continue producing samples drawn from this distribution. This is precisely the behavior needed for Monte Carlo estimation. The key idea of MCMC methods is thus to find a transition distribution which has the desired distribution as its stationary distribution. The various MCMC algorithms suggest different concepts to find such a Markov Chain.

The conditions for a Markov chain to converge to a unique stationary distribution are as follows [13]:

- It is time-homogenous: the transition distribution $f(\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)})$ does not depend on the ("time") index i, which would obviously contradict the concept of a stationary distribution.
- It is irreducible: the transition between any two states occurs with non-zero probability after a finite number of steps:

  $$\forall \mathbf{a}, \mathbf{b} \Rightarrow \exists n \text{ such that } \mathrm{P}\{\mathbf{x}^{(i+n)} = \mathbf{a} | \mathbf{x}^{(i)} = \mathbf{b}\} > 0.$$

  In other words, the sample space is not divided into disjoint subspaces in the sense that a sample from one subspace cannot lead to any sample in a different subspace.
- It is aperiodic: The number of steps after which the chain may return to the same state with nonzero probability is not bound to be a multiple of some integer.
- It is positive-recurrent: the average number of steps between two occurrences of the same state is finite for some (and hence for all) states. This condition is equivalent with the existence of some stationary distribution.

In a MCMC context however, our goal is not to examine the properties of a given Markov chain (i.e. a given transition distribution), but rather to find an appropriate transition distribution $f(\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)})$ that has a given stationary distribution $f(\mathbf{x})$. To this end, having a sufficient condition for verifying the relation between two such distributions is very useful. This condition is found in the detailed balance equation:

$$f_{\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)}}(\mathbf{a}|\mathbf{b})\, f_{\mathbf{x}}(\mathbf{b}) = f_{\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)}}(\mathbf{b}|\mathbf{a})\, f_{\mathbf{x}}(\mathbf{a}) \quad \text{for all } \mathbf{a}, \mathbf{b}. \tag{2.7}$$

Applying this relation to (2.6) will lead to the result that $f(\mathbf{x}^{(i+1)})$ equals $f(\mathbf{x}^{(i)})$, which we used above as a definition of a stationary distribution.

If a Markov has a unique stationary distribution, the marginal distribution converges towards this stationary distribution, regardless of the initial state (or distribution). Therefore, once we have found an appropriate Markov Chain for our problem according to one of the MCMC algorithms, we can initialize it arbitrarily and rely on the fact that it will converge. However, the beginning of the chain will evidently not contain representative samples, since it will take a certain number of steps until the marginal distributions have converged to the stationary distributions. This is respected by defining a burn-in period, i.e. a number of samples at the beginning of the chain which are disregarded to avoid the influence of the initial state [11].

The samples produced by such a Markov chain are generally not statistically independent. However, irreducibility ensures that the whole state space is explored, while the appropriate multiplicity or density (for discrete or continuous distributions, respectively) of the samples is provided by the stationary distribution - both under the condition that the Markov chain is long enough for applying the law of large numbers.

A common concept to reduce correlation within the sample is thinning, i.e. regarding only every n-th sample in order to avoid the correlation of nearest samples [2].

Different MCMC methods vary in the way they create an appropriate transition distribution (aimed at a certain stationary distribution), with side conditions optimized for certain classes of problems.

2.4 The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm [2] is a MCMC method with a property that is very convenient for Bayesian estimation: the desired posterior distribution must be known only up to a constant scaling factor. This means we can avoid normalization and derive only the numerator of (2.1), ignoring the integral which is often hard to calculate. We will denote this non-normalized distribution by $\tilde{f}(\mathbf{x})$. In the context of Bayesian estimation, this is the non-normalized posterior distribution $\tilde{f}(\boldsymbol{\theta}|\mathbf{y})$, while our notation with $\mathbf{x}$ is intended to show the more general applicability.

The algorithm relies on drawing samples from a conditional distribution $g(\mathbf{w}|\mathbf{x})$, which can be chosen almost arbitrarily. It must however fulfill the condition that for all $\mathbf{w}$ with $\tilde{f}(\mathbf{w}) > 0$ there must be an $\mathbf{x}$ such that $\tilde{f}(\mathbf{x}) > 0$ and $g(\mathbf{w}|\mathbf{x}) > 0$.

Based on the current value $\mathbf{x}^{(i)}$, a sample $\mathbf{w}$ is generated from $g(\mathbf{w}|\mathbf{x}^{(i)})$. Then the so-called acceptance probability ρ is calculated as follows:

$$\rho = \min\left\{\frac{g(\mathbf{x}^{(i)}|\mathbf{w})\tilde{f}(\mathbf{w})}{g(\mathbf{w}|\mathbf{x}^{(i)})\tilde{f}(\mathbf{x}^{(i)})}, 1\right\}. \tag{2.8}$$

This acceptance probability is used to obtain the next sample $\mathbf{x}^{(i+1)}$:

$$\mathbf{x}^{(i+1)} = \begin{cases} \mathbf{w} & \text{with a probability } \rho \\ \mathbf{x}^{(i)} & \text{with a probability } 1-\rho. \end{cases}$$

It is now evident why the normalization of $\tilde{f}(\mathbf{x})$ can be omitted: it is only used in the fraction in (2.8), where any constant factors are canceled. The validity of the algorithm to produce a

Markov Chain with the stationary distribution $f(\mathbf{x})$ is not as obvious, but it can easily be shown by checking the detailed balance equation (2.7):

$$\underbrace{\min\left\{\frac{g(\mathbf{b}|\mathbf{a})\tilde{f}(\mathbf{a})}{g(\mathbf{a}|\mathbf{b})\tilde{f}(\mathbf{b})},1\right\}g(\mathbf{a}|\mathbf{b})}_{f_{\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)}}(\mathbf{a}|\mathbf{b})}\tilde{f}(\mathbf{b}) = \underbrace{\min\left\{\frac{g(\mathbf{a}|\mathbf{b})\tilde{f}(\mathbf{b})}{g(\mathbf{b}|\mathbf{a})\tilde{f}(\mathbf{a})},1\right\}g(\mathbf{b}|\mathbf{a})}_{f_{\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)}}(\mathbf{b}|\mathbf{a})}\tilde{f}(\mathbf{a}).$$

Since the two fractions are reciprocal to each other, one minimum must be 1, while the other must be the fraction, leading to the same result in both possible cases:

$$g(\mathbf{b}|\mathbf{a})\tilde{f}(\mathbf{a})g(\mathbf{a}|\mathbf{b})\tilde{f}(\mathbf{b}) = g(\mathbf{a}|\mathbf{b})\tilde{f}(\mathbf{b})g(\mathbf{b}|\mathbf{a})\tilde{f}(\mathbf{a}),$$

where all factors cancel each other.

One problem that is not solved by the Metropolis-Hastings algorithm is that of higher dimensionality: the distribution used for sampling, $g(\mathbf{w}|\mathbf{x})$, has the same dimension as the desired distribution.

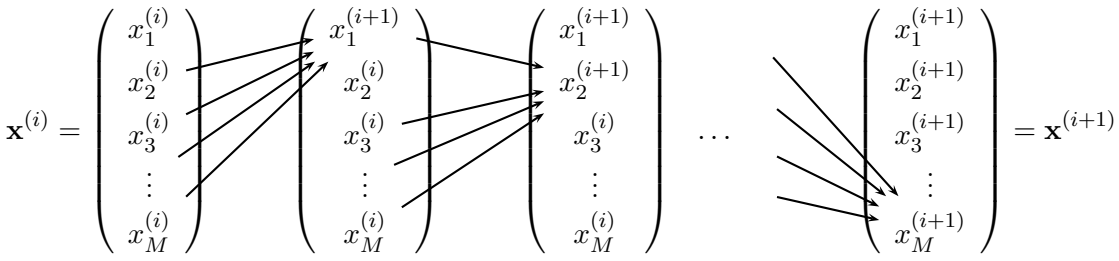
2.5 The Gibbs Sampler

The Gibbs sampler [2] is a MCMC method which can be used very efficiently if $\mathbf{x}^{(i)}$ - and consequently the desired stationary distribution $f(\mathbf{x})$ - is high-dimensional, which might make sampling from a transition distribution $f(\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)})$ computationally complex. The Gibbs sampler avoids deriving such an M-dimensional transition distribution. Instead, a step in the Markov chain is divided into M substeps using M univariate distributions $f(x_m|\mathbf{x}_{\sim m})$,

$$\begin{aligned}\text{where} \qquad \mathbf{x} &= [x_1, x_2, ..., x_M] \\ \mathbf{x}_{\sim m} &= [x_1, ..., x_{m-1}, x_{m+1}, ..., x_M].\end{aligned}$$

In each substep, a sample of x_m is drawn from this distribution $f(x_m|\mathbf{x}_{\sim m})$. As opposed to the Metropolis-Hastings algorithm the distributions used for sampling here are not arbitrary, they are derived from the desired stationary distribution $f(\mathbf{x})$. Also, we cannot avoid normalization in this algorithm. On the other hand, the joint stationary distribution is not needed at all for the Gibbs sampler, and in a Bayesian context it is often easier to directly derive the marginal posteriors (using (2.2)) rather than the joint posterior (using (2.1)). This is especially true in the case of statistically independent parameters, allowing the simplification (2.3). As mentioned for the Metropolis-Hastings algorithm, our notation using $\mathbf{x}$ is intended to show the more general applicability. In a Bayesian estimation problem, the corresponding distributions are $f(\boldsymbol{\theta}|\mathbf{y})$ and $f(\theta_m|\mathbf{y}, \boldsymbol{\theta}_{\sim m})$.

The values $\mathbf{x}_{\sim m}$ fed into the distribution for sampling x_m are the most recent values of each element. This means that at step $(i+1)$ and substep m (i.e. sampling $x_m^{(i+1)}$) the values used as conditional information are $x_1^{(i+1)}, \ldots, x_{m-1}^{(i+1)}, x_{m+1}^{(i)}, \ldots, x_M^{(i)}$. One step in the Markov chain can thus be depicted as follows:



The entire algorithm is depicted in Fig. 2.3, where a Markov chain of length N is used to implement an MMSE estimator. A burn-in period of B is observed, as explained earlier.

To show the validity of the resulting Markov chain, we will use a simplified notation, with $f_m(\cdot|\cdot)$ denoting $f(x_m|\mathbf{x}_{\sim m})$ and $[x_m, \mathbf{x}_{\sim m}]$ denoting the vector $\mathbf{x}$ composed of x_m and $\mathbf{x}_{\sim m}$. Each substep fulfills the detailed balance equation (2.7):

$$\begin{aligned} f_m(a|\mathbf{x}_{\sim m})\, f\left([b, \mathbf{x}_{\sim m}]\right) &= f_m(b|\mathbf{x}_{\sim m})\, f\left([a, \mathbf{x}_{\sim m}]\right) \\ \frac{f\left([a, \mathbf{x}_{\sim m}]\right)}{f(\mathbf{x}_{\sim m})}\, f\left([b, \mathbf{x}_{\sim m}]\right) &= \frac{f\left([b, \mathbf{x}_{\sim m}]\right)}{f(\mathbf{x}_{\sim m})}\, f\left([a, \mathbf{x}_{\sim m}]\right). \end{aligned} \tag{2.9}$$

For the sake of simplicity the remaining step in the proof of the validity will be shown for 3-dimensional samples, the extension to M dimensions is straight-forward. To show that the desired joint distribution is the stationary distribution of the Gibbs sampler's transition distribution we will insert the latter into (2.6):

$$\begin{aligned} f_{\mathbf{x}^{(i+1)}}(\mathbf{b}) &= \int\int\int f_{\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)}}(\mathbf{b}|\mathbf{a})\, f_{\mathbf{x}^{(i)}}(\mathbf{a})\, \mathrm{d}\mathbf{a} \\ f_{\mathbf{x}^{(i+1)}}(b_1, b_2, b_3) &= \int\int\int f_3(b_3|b_1, b_2)\, f_2(b_2|b_1, a_3)\, f_1(b_1|a_2, a_3)\, f_{\mathbf{x}^{(i)}}(a_1, a_2, a_3)\, \mathrm{d}a_1\, \mathrm{d}a_2\, \mathrm{d}a_3. \end{aligned}$$

Applying (2.9) to all three conditional distributions yields

$$\begin{aligned} f_{\mathbf{x}^{(i+1)}}(b_1,b_2,b_3) &= \int\int\int f_{\mathbf{x}^{(i)}}(b_1,b_2,b_3)\, f_3(a_3|b_1,b_2)\, f_2(a_2|b_1,a_3)\, f_1(a_1|a_2,a_3)\,\mathrm{d}a_1\,\mathrm{d}a_2\,\mathrm{d}a_3 \\ &= \int\int f_{\mathbf{x}^{(i)}}(b_1,b_2,b_3)\, f_3(a_3|b_1,b_2)\, f_2(a_2|b_1,a_3) \underbrace{\int f_1(a_1|a_2,a_3)\,\mathrm{d}a_1}_{=1}\,\mathrm{d}a_2\,\mathrm{d}a_3 \\ &= \int f_{\mathbf{x}^{(i)}}(b_1,b_2,b_3)\, f_3(a_3|b_1,b_2) \underbrace{\int f_2(a_2|b_1,a_3)\mathrm{d}a_2}_{=1}\,\mathrm{d}a_3 \\ &= f_{\mathbf{x}^{(i)}}(b_1,b_2,b_3) \underbrace{\int f_3(a_3|b_1,b_2)\mathrm{d}a_3}_{=1}, \end{aligned}$$

which readily shows that the desired joint distribution is the stationary distribution of the Gibbs sampler.

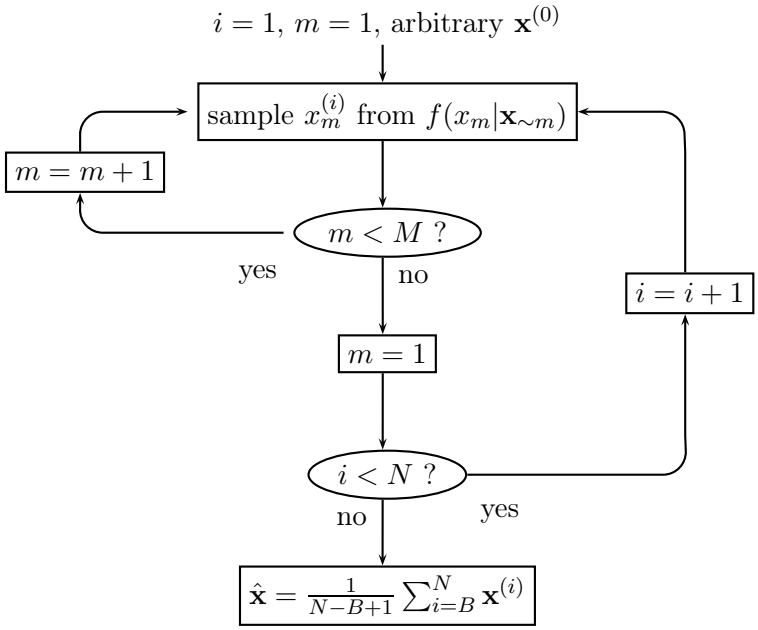


Figure 2.3: Flow chart of a MMSE estimator for $\mathbf{x}$ using a Gibbs sampler.

Example 2.5 To give a simple example for a Gibbs sampler implementation, let us assume an estimation problem with a 2-dimensional i.i.d. parameter vector $\boldsymbol{\theta}$ and a 2-dimensional observation vector $\mathbf{y}$ related as follows:

$$\mathbf{y} = \mathbf{C}\,\boldsymbol{\theta} + \mathbf{n},$$

where $\mathbf{C} = [\mathbf{c}_1\ \mathbf{c}_2]$ is known and the noise is white and Gaussian: $n_1, n_2 \sim \mathcal{N}(0, \sigma_n^2)$. Assigning the parameters a Gaussian prior $\theta_1, \theta_2 \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$ leads to the following joint posterior distribution:

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) = \frac{1}{2\pi\sigma_n^2} \exp\left(-\frac{\|\mathbf{y} - \mathbf{C}\ \boldsymbol{\theta}\|^2}{2\sigma_n^2}\right) \frac{1}{2\pi\sigma_\theta^2} \exp\left(-\frac{\|\boldsymbol{\theta} - \boldsymbol{\mu}_\theta\|^2}{2\sigma_\theta^2}\right)$$

$$f(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}}^2)$$

$$\text{with} \quad \boldsymbol{\Sigma}_{\text{post}}^2 = \left(\frac{\mathbf{C}^{\mathrm{T}}\mathbf{C}}{\sigma_n^2} + \frac{\mathbf{I}_2}{\sigma_\theta^2}\right)^{-1} \quad \text{and} \quad \boldsymbol{\mu}_{\text{post}} = \boldsymbol{\Sigma}_{\text{post}}^2 \left(\frac{\mathbf{C}^{\mathrm{T}}\mathbf{y}}{\sigma_n^2} + \frac{\boldsymbol{\mu}_\theta}{\sigma_\theta^2}\right)$$

In this simple case, any statistics of the joint posterior can be calculated analytically, which would make the use of a Gibbs sampler unnecessary, however allowing us to compare the results and check its validity. On top of that, the obvious extension of the problem to more dimensions would imply the inversion of a bigger matrix in the analytical case, which is avoided by the Gibbs sampler. The marginal posteriors can be derived as follows:

$$f(\theta_1|\mathbf{y}, \theta_2) \propto f(\mathbf{y}|\boldsymbol{\theta}) f(\theta_1) = \frac{1}{2\pi\sigma_n^2} \exp\left(-\frac{\|\mathbf{y} - \mathbf{c}_1\ \theta_1 - \mathbf{c}_2\ \theta_2\|^2}{2\sigma_n^2}\right) \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{(\theta_1 - \mu_\theta)^2}{2\sigma_\theta^2}\right)$$

$$f(\theta_1|\mathbf{y}, \theta_2) = \mathcal{N}(\mu_1, \sigma_1^2)$$

$$\text{with} \quad \sigma_1^2 = \left(\frac{\|\mathbf{c}_1\|^2}{\sigma_n^2} + \frac{1}{\sigma_\theta^2}\right)^{-1} \quad \text{and} \quad \mu_1 = \sigma_1^2 \left(\frac{\mathbf{c}_1^{\mathrm{T}}(\mathbf{y} - \mathbf{c}_2\ \theta_2)}{\sigma_n^2} + \frac{\mu_\theta}{\sigma_\theta^2}\right)$$

with the analogous result for θ_2:

$$f(\theta_2|\mathbf{y}, \theta_1) = \mathcal{N}(\mu_2, \sigma_2^2)$$

$$\text{with} \quad \sigma_2^2 = \left(\frac{\|\mathbf{c}_2\|^2}{\sigma_n^2} + \frac{1}{\sigma_\theta^2}\right)^{-1} \quad \text{and} \quad \mu_2 = \sigma_2^2 \left(\frac{\mathbf{c}_2^{\mathrm{T}}(\mathbf{y} - \mathbf{c}_1\ \theta_1)}{\sigma_n^2} + \frac{\mu_\theta}{\sigma_\theta^2}\right)$$

Following the Gibbs sampler algorithm, samples are drawn alternately from these two distributions, each time using the most recent sample as a condition for the next one. With increasing length of the chain, the empirical distribution of the samples converges towards the joint posterior distribution. Likewise, the estimates obtained from the sample converge. This is illustrated in Fig. 2.4 and compared to the non-Bayesian ML estimator, which is based on the likelihood function, neglecting the information given by the priors:

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} f(\mathbf{y}|\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{C}\ \boldsymbol{\theta}\|^2 = (\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{y}.$$

□

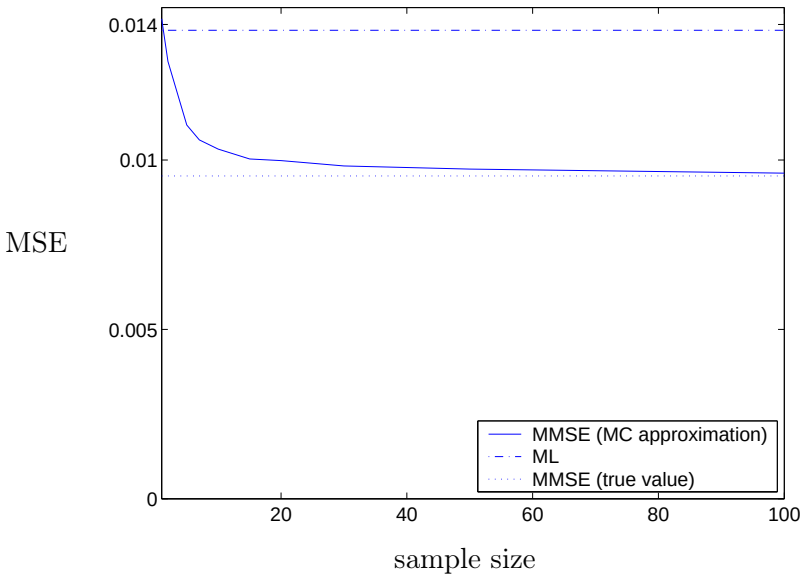


Figure 2.4: MSE for θ_1 obtained in experiments as described in Example 2.5, with $\mu_\theta = 0$, $\sigma_\theta^2 = 1$, $\mathbf{c}_1 = [1, 1]^{\mathrm{T}}$, $\mathbf{c}_2 = [0, 2]^{\mathrm{T}}$, $\sigma_n^2 = 1$. With increasing sample size, the MC estimator approximates the analytic MMSE estimate, yielding a result that is clearly better than the non-Bayesian ML estimate.

Chapter 3

Optical Coherence Tomography

3.1 Introduction

Optical Coherence Tomography is an interferometric imaging technique which is used to obtain cross-sectional views of the subsurface microstructure of biological tissue. Using broadband laser light, probing depths of up to 2 mm in the skin and over 2 cm in transparent tissues can be achieved [3]. Besides being non-invasive, another strength of OCT is its high resolution, reaching the sub-micrometer order [14].

Being a tomography technique, OCT generates slice images of 3-dimensional objects. To achieve this, several 1-dimensional depth scans have to be repeated at varying lateral positions. One depth scan typically contains about 16000 samples, while an image may be composed of roughly $400 - 600$ scans. Since one of the decisive strengths claimed for OCT is its aptitude for in vivo diagnosis, high scan speeds are necessary to obtain an entire image while the probe does not move.

A typical application of OCT is for obtaining images of the human retina. The method is very attractive for this task because the laser light propagates through the eyeball without significant distortion. This application shall therefore serve as an example throughout this survey. An OCT scan image of the retina can be seen in Fig. 3.1. An obvious property of the tissue which will of interest for signal processing is its layered structure.

A scheme of the setup for producing one depth scan is provided in Fig. 3.2. Each depth scan is accomplished by a Michelson-Moorley interferometer setup. Light emitted from a laser is split into a reference field, which is reflected by a mirror, and a sample field, which is reflected by the probe. In the photodiode, the superposition of the two reflected fields is detected.

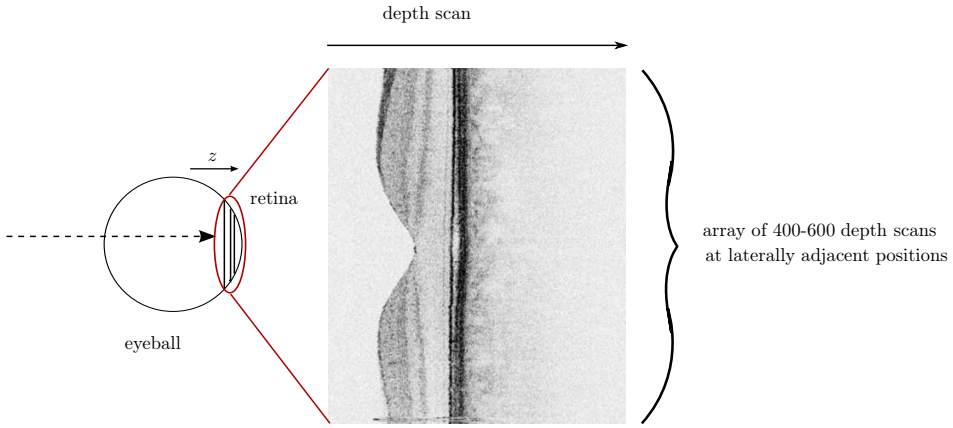


Figure 3.1: OCT image of the human retina.

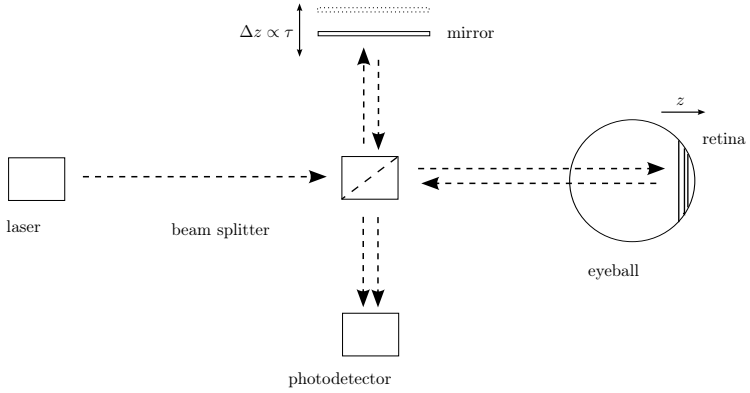


Figure 3.2: Scheme of the interferometric method producing one depth scan.

Denoting the reference field by E_r and the sample field by E_s, the light intensity at the photodetector can thus be expressed as

$$\begin{aligned} I(\tau) &= \left\langle \big(E_s(t) + E_r(t-\tau)\big)^2 \right\rangle_t \\ &= I_s + I_r + 2R_{sr}(\tau) \end{aligned} \tag{3.1}$$

$$\begin{aligned} \text{with} \qquad I_{r,s} &= \left\langle E_{r,s}^2 \right\rangle_t \qquad \text{(constant for all } \tau\text{)} \\ R_{sr}(\tau) &= \langle E_s(t)\, E_r(t-\tau) \rangle_t \,, \end{aligned} \tag{3.2}$$

where $\langle \cdot \rangle_t$ denotes the time average over an interval $[-T;T]$ much longer than the period duration of the field:

$$\langle x(t) \rangle_t = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t) \mathrm{d}t$$

The parameter τ in equations (3.1) ff. denotes a time lag introduced by the different paths if the argument t refers to a common time instant before the field is split.

The constant intensities are removed by a heterodyne measurement, and only the time-varying component is passed on for signal processing.

In a simplified model of the physical processes, which shall be discussed later in this section, the sample field can be described as the output of a linear time invariant system containing all the effects caused by interaction with the probe. The input to this system is equal to the reference field:

$$E_s(t) = \int_{-\infty}^{\infty} E_r(t-t')\, h_{\text{probe}}(t')\mathrm{d}t'.$$

Inserting this in (3.2) yields:

$$\begin{aligned}
\Rightarrow R_{sr}(\tau) &= \left\langle \int_{-\infty}^{\infty} E_r(t-t')\, h_{\text{probe}}(t')\mathrm{d}t'\, E_r(t-\tau) \right\rangle_t \\
&= \int_{-\infty}^{\infty} \langle E_r(t-t')\, E_r(t-\tau)\rangle_t\, h_{\text{probe}}(t')\mathrm{d}t' \\
&= \big(R_r * h_{\text{probe}}\big)(\tau) \\
&\quad \text{with } R_r(\tau) = \langle E_r(t)E_r(t-\tau)\rangle_t.
\end{aligned}$$

The time lag τ between E_r and E_s corresponds to a difference in the lengths of the two paths, which is controlled by the positioning the mirror (see Fig. 3.2). This spatial interpretation can be extended to $h_{\text{probe}}(t)$ such that interaction with the probe is expressed as a function of depth within the tissue, denoted as $h_{\text{probe}}(z)$.

Thus the observed signal

$$y(z) = R_{sr}(z) + n(z)$$

contains the autocorrelation function of the laser signal, convolved with $h_{\text{probe}}(z)$, which characterizes the probe, and some additional observation noise $n(z)$. An example is shown in Fig. 3.3.

Interaction between the probe and the electromagnetic field can be described as a combination of the following effects:

- reflection - At the borders of tissue layers with different refractive indices, the signal is reflected. This effect is visible in OCT images like Fig. 3.1 and will be discussed in more detail in the following section.
- absorption - Tissue with a nonzero conductivity absorbs energy from the field, reducing the signal magnitude. This effect is strongly frequency dependent.

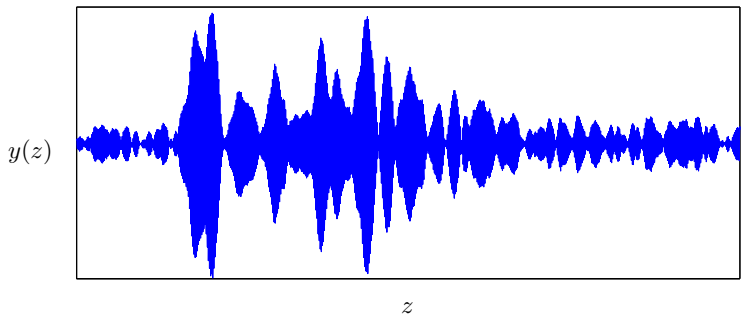


Figure 3.3: Observed signal of one depth scan.

- scattering - At small inhomogeneities in the probe the laser light gets scattered, again in a frequency dependent manner.

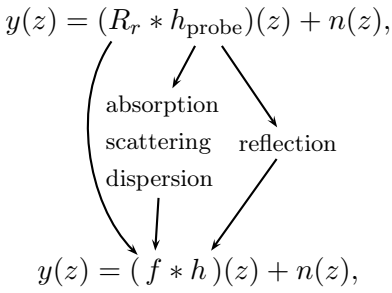
- dispersion - The signal components at different frequencies pass through the tissue at different speeds, depending on the dielectric properties of the material. The laser pulse is thereby broadened, which reduces the resolution of the image.

Each of these processes makes its particular contribution to the received sample field and can be exploited to obtain information about the tissue, e.g. the blood circulation or the type of tissue. However, in this survey, only the linear time invariant effect of reflection at layer borders, denoted as $h(z)$, is estimated. Absorption, scattering and dispersion are time variant processes. Throughout this survey, this time variance is assumed to be compensated to permit a time invariant signal model. Compensation itself is a research field of its own and beyond the scope of this thesis.

3.2 Signal Model

3.2.1 General Model and Adjustments

The aim of the signal model to be defined is to facilitate estimation of the desired information, which is contained in $h(z)$, as mentioned above. In order to obtain a model useful for estimation, the components of the received signal can be rearranged in such a way that the data is contained in one function, while the distorting processes are summarized in another one:



where $y(z)$ denotes the observed signal, $f(z)$ is the so-called fringe, $h(z)$ describes the reflections at layer boundaries and $n(z)$ denotes observation noise.

The newly defined fringe $f(z)$ contains the autocorrelation function of the laser signal as well as effects caused by the tissue (again, after compensation of time variant effects).

Sampling yields the discrete-time signal

$$y[k] = (f * h)[k] + n[k] \quad \text{with} \quad k = 1, ..., K,$$

which is illustrated in Fig. 3.4.

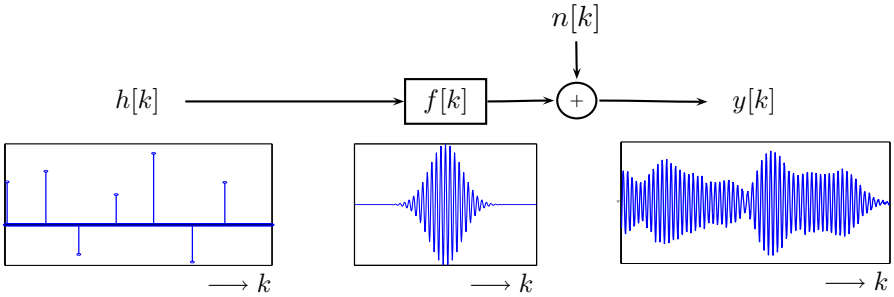


Figure 3.4: Signal model with fringe and retina function.

3.2.2 Modeling the Fringe

Using a parametric model for the fringe, its spectrum can be approximately represented by a Gaussian function around a nonzero center frequency [4]:

$$F(j\omega) \propto \exp\left(-\frac{(\omega - \omega_{\text{center}})^2}{\alpha^2_{\text{spectrum}}}\right).$$

Neglecting further distorting effects of sampling, the signal can therefore be expressed as

$$f[k] = \mathrm{e}^{-(\frac{k}{\alpha})^2} \cos(\omega_0 k),$$

with the two determining parameters α, describing the fringe width, and the normalized angular frequency ω_0, representing the center frequency. Alternative ways of modeling the fringe are addressed in chapter 7 but not pursued in this survey.

An example of a synthetic fringe is shown in Fig. 3.5.

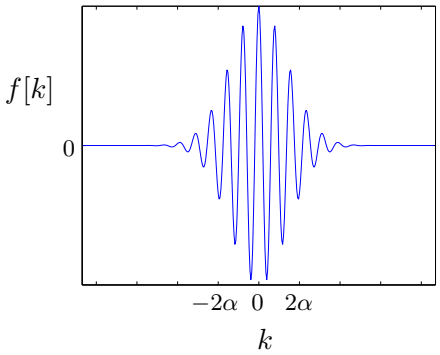


Figure 3.5: Synthetic example of the fringe.

3.2.3 Modeling the Retina

The property of the retina - or tissue in general - that shall be exploited for diagnoses is its composition of various layers. These layers have different refractive indices, which causes the electromagnetic field to be reflected at their boundaries. At the boundary of two layers with the respective refractive indices n_i, n_{i+1}, the reflection coefficient a_i describes the amplitude of the reflected wave relative to the incoming wave [15]:

$$a_i = \frac{n_i - n_{i+1}}{n_i + n_{i+1}}.$$

(A negative a_i, resulting from $n_i < n_{i+1}$, accounts for a phase change of 180° relative to the incoming wave.)

The entire probe can therefore be described by what will be called the retina function, $h(z)$:

$$h(z) = \sum_{i=1}^{N} a_i \delta(z - z_i),$$

where N is the number of layer boundaries in the tissue and z_i represents the respective depth at which the i-th boundary is located in the tissue.

In the discrete-time domain the retina function is defined as

$$h[k] = \sum_{i=1}^{N} a_i \, \delta[k - k_i]. \tag{3.3}$$

An example is shown in Fig. 3.6. The values of a_i are related to the size of the steps in the refractive index, while the values of k_i denote their positions.

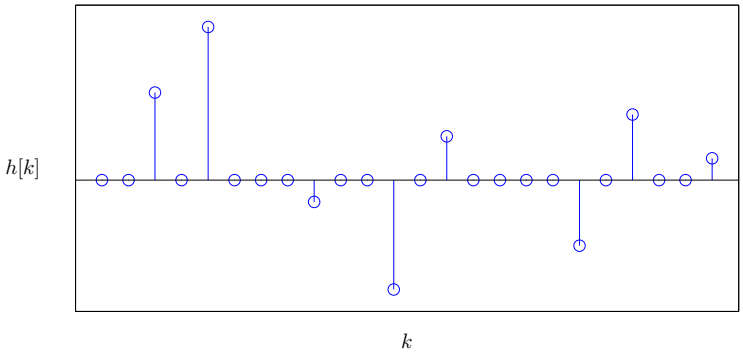


Figure 3.6: Synthetic example of the retina function.

With respect to future use of the model, it can be rewritten in an equivalent alternative formulation that will prove more efficient for estimation. The retina function is, to that end, split into two factors, one of which contains only information about the pulse positions, while the other one describes the respective amplitudes:

$$\begin{aligned} h[k] &= a[k]r[k] \\ \text{with} \quad r[k] &= \sum_i \delta[k-k_i] \in \{0,1\} \\ a[k] &= a_i \quad \text{for } k = k_i. \end{aligned} \tag{3.4}$$

Obviously, $r[k]$ is fully defined by the model, while the values of $a[k]$ at positions $k \notin \{k_i\}$ are not. They can be chosen arbitrarily, since they will not have any influence on estimation.

The virtues of this less intuitive way of describing $h[k]$ will be discussed in section 4.1.2, as Bayesian estimation methods are applied to the model. Comparable product models are common in Bayesian estimation methods and have been used to describe OCT signals [4].

The observed signal can now be expressed as

$$\begin{aligned} y[k] &= (f * h)[k] + n[k] \\ &= (f * (a \cdot r))[k] + n[k], \end{aligned} \tag{3.5}$$

or, in vector/matrix notation:

$$\begin{aligned} \mathbf{y} &= \mathbf{F}\,\mathbf{h} + \mathbf{n} \\ &= \mathbf{F}\,\mathrm{diag}(\mathbf{r})\,\mathbf{a} + \mathbf{n} \end{aligned} \tag{3.6}$$

with $\mathbf{F}$ ... $K \times K$ Toeplitz matrix containing the fringe: $[\mathbf{F}]_{ij} = f[i-j]$
$\mathbf{h}$ ... vector of length K containing the retina function
$\mathbf{n}$ ... noise vector of length K containing the observation noise

The aim of the estimation procedure to be proposed in the following problem will be to estimate the values of $\mathbf{r}$ and $\mathbf{a}$ from the observed signal $\mathbf{y}$.

Chapter 4

Application of Bayesian Sampling to OCT

4.1 Basic Model

The choice of prior distributions of the random variables contained in our model reflects our knowledge or assumptions about the variables. Since the processes involved in the real world application of OCT are not analytically understood but only modeled on some level of abstraction, the distributions are to some extent arbitrary. They are chosen based on observations of OCT signals, modeling of underlying physical processes and more general mathematical considerations.

4.1.1 The Likelihood Function

Before defining the prior distributions of our parameters it is useful to consider the likelihood function. From our system model,

$$\begin{aligned}\mathbf{y} &= \mathbf{F}\,\mathbf{h} + \mathbf{n} \\ &= \mathbf{F}\,\mathrm{diag}(\mathbf{r})\,\mathbf{a} + \mathbf{n},\end{aligned}$$

with $\mathbf{F}$ being deterministic and $\mathbf{r}$ and $\mathbf{a}$ being our parameters, it is obvious that the likelihood function is equivalent to the distribution of the noise $\mathbf{n}$, which is assumed to be zero-mean, Gaussian and white. The variance of the noise distribution is modeled to be random and added to our list of parameters to be estimated. This yields the likelihood function:

$$\begin{aligned}f_{\mathbf{y}|\mathbf{r},\mathbf{a},\sigma_n^2}(\mathbf{y}|\mathbf{r},\mathbf{a},\sigma_n^2) &= f_{\mathbf{n}|\sigma_n^2}(\mathbf{y} - \mathbf{F}\,\mathrm{diag}(\mathbf{r})\,\mathbf{a}\,|\,\sigma_n^2) \\ &= \mathcal{N}(\mathbf{F}\,\mathrm{diag}(\mathbf{r})\,\mathbf{a}\,,\,\sigma_n^2\mathbf{I}_K).\end{aligned}$$

4.1.2 Prior Distributions

The parameters for which we must now define a prior distribution are $\mathbf{r}$, $\mathbf{a}$ and σ_n^2. In a first approach we define both vectors $\mathbf{r}$ and $\mathbf{a}$ as i.i.d., which means than both r_k and a_k are not correlated with respect to the position k:

$$p(\mathbf{r}) = \prod_{k=1}^{K} p(r_k) \quad \text{and} \quad f(\mathbf{a}) = \prod_{k=1}^{K} f(a_k). \tag{4.1}$$

The splitting of the retina function h_k into a position function r_k and an amplitude function a_k as proposed in Chapter 2 is completely transparent to the Bayesian model and therefore perfectly arbitrary. It is made for the the sake of a more intelligible presentation of the problem and convenience of calculations. The product h_k however represents a physical quantity, and its distribution can therefore be judged as more or less appropriate.

Since h_k represents the reflectivities of boundaries between different tissue layers, it is of a sparse nature. This means that there is a finite (and rather high) probability p_0 that h_k will be 0 at position k. The magnitude of the reflectivities at nonzero positions however will be continuously distributed and therefore described by a density function rather than a mass function. Modeling h_k as a product of a discrete and a continuous variable allows us to separate detection of the pulse positions from estimation of the amplitudes, using efficient methods for the respective purpose. These methods are applied to the samples obtained from the Gibbs sampler. The sampling method itself, however, does not make any difference between the two types of variables except for the fact that the distribution is continuous in one case and discrete in the other.

The most straight-forward way of splitting h_k into two variables is to let one represent the positions of the layer boundaries, while the other represents their amplitudes of reflectivity. By designing the position function as a function of k, we implicitly give it a binary nature. Being a discrete variable, it is described by a probability mass function, which in the binary case consists of only two values.

The second factor of h_k is the amplitude function a_k, which represents the values of reflectivity of layer boundaries. Designing it as a function of k means that it contains more information than needed. The amplitudes are only defined at positions k where $r_k = 1$, all other entries of $\mathbf{a}$ are irrelevant. For a constructive choice of the prior distribution of a_k we may take into account that the magnitudes of reflectivity are usually well above the noise level. We should therefore use a distribution that favors amplitudes of some typical finite range over smaller ones. At the same time, the distribution should be symmetric around 0 to include negative values and keep the sign statistically independent of the magnitude.

These latter considerations point to a more practical albeit not as intuitive way of splitting h_k. By allowing r_k to assume not only the values 0 and 1 but also -1 with a symmetric discrete distribution, we can considerably simplify the distribution of a_k. It no longer needs to be symmetric around 0, and may have only one local maximum rather than two, which opens a wide range of standard distributions. This simplicity of the distributions is crucial because the distributions are not only used for evaluating the probability (density) of a given value but also for drawing samples, which can be achieved much more efficiently with standard distributions.

The distribution of r_k (depicted in Fig. 4.1) is therefore:

$$p(r_k) = \begin{cases} p_0 & \text{for } r_k = 0 \\ \frac{1-p_0}{2} & \text{for } r_k = 1 \\ \frac{1-p_0}{2} & \text{for } r_k = -1. \end{cases} \tag{4.2}$$

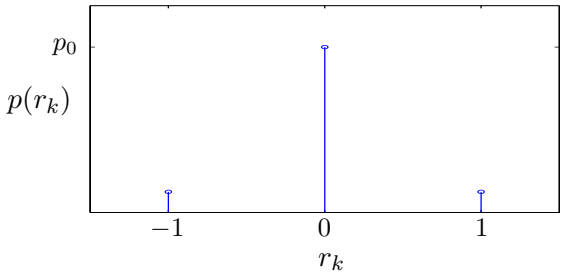


Figure 4.1: Prior distribution of r_k.

Our choice for the distribution of a_k is a normal distribution with nonzero mean:

$$a_k \sim \mathcal{N}(\mu_a, \sigma_a^2). \tag{4.3}$$

With its two parameters it can be adjusted to fit the problem, while not being too complex to handle. At the same time, the normal distribution has a decisive computational advantage in this context: it is a conjugate prior of the Gaussian likelihood function, which simplifies calculations, as explained in section 2.1.4.

The resulting distribution of $h_k = r_k a_k$ is given in Fig. 4.2 for the sake of completeness (it is not needed for any calculations).

The noise variance σ_n^2 is defined to be distributed according to an Inverse-Gamma distribution (as proposed in [5]), which is again a conjugate prior:

$$\sigma_n^2 \sim \mathcal{IG}(\xi, \eta) = \frac{\eta^\xi}{\Gamma(\xi)} (\sigma_n^2)^{-\xi-1} \exp\left(-\frac{\eta}{\sigma_n^2}\right) u(\sigma_n^2), \tag{4.4}$$

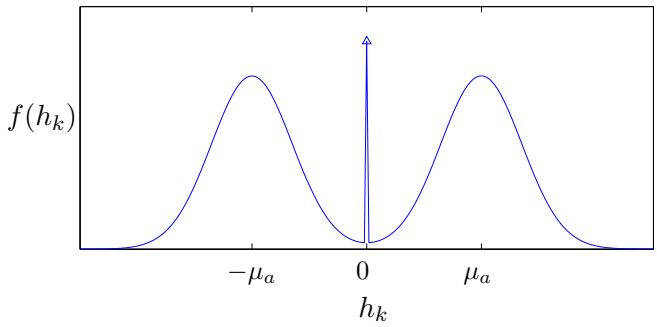


Figure 4.2: Prior distribution of $h_k = r_k a_k$, showing three characteristics: sparsity (dirac impulse at 0), suppression of small amplitudes and symmetry around 0.

where $u(x)$ denotes the unit step function. The Inverse-Gamma distribution has two shape parameters ξ and η, which have to be predefined. The mean of the distribution is $\frac{\eta}{\xi-1}$. The choice of this prior is motivated by computational convenience, while not contradicting any known physical properties of the estimation problem. Assumptions regarding the noise variance can be implemented by adjusting the shape parameters. An illustration of the distribution is given in Fig. 4.3.

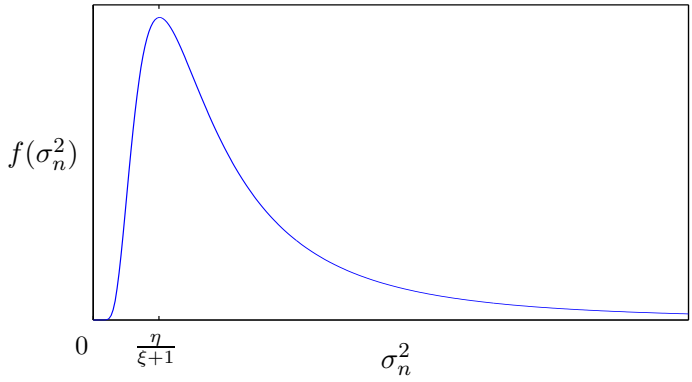


Figure 4.3: Prior distribution of σ_n^2.

In a further refinement of the model which is not discussed here, the hyperparameters, i.e. variables determining the exact shape of the parameters' priors (p_0 for r_k, μ_a and σ_a^2 for a_k and ξ and η for σ_n^2), could also be estimated in the iterative sampling process, therefore making the model a hierarchical model [16]. Defining a hierarchical model increases the robustness compared to fixed parameter priors. The Bayesian estimation framework makes this possible, since hyperparameters can be treated just like the parameters and estimated together with them. More specifically, the joint posterior distribution can be derived as a joint distribution of parameters and hyperparameters. While all our parameters so far were statistically independent of each other, this would evidently not be the case with hyperparameters. This, however, does not affect the estimation

procedure (apart from the fact that the simplification (2.3) is not applicable). An example for this procedure is presented in [6].

4.1.3 Posterior Distributions

In order to estimate the parameters we will implement a Gibbs sampler, which requires the posterior distributions of each parameter given all others and the observation. Using the above prior distributions, the posteriors can now be derived according to Bayes' rule. Since all our parameters are statistically independent (which is obvious from their prior distributions) we can use its simplified version (2.3). Furthermore, in our calculations it will prove sufficient to use a proportion rather than an equation, which allows us to drop the normalizing denominator:

$$f(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2) \propto f(\mathbf{y}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) f(\boldsymbol{\theta}_1). \tag{4.5}$$

A remark on notation for the following equations: Let $\mathbf{r}_{\sim i}$ denote the vector $\mathbf{r}$ reduced by the dimension i. $\mathbf{f}_k$ expresses the kth column of the matrix $\mathbf{F}$, while $\mathbf{F}_{\sim k}$ denotes the matrix $\mathbf{F}$ reduced by the kth column. For the sake of simpler notation, $\mathbf{h}$ is used for the equivalent $\mathrm{diag}(\mathbf{r})\,\mathbf{a}$.

The Position Function. The posterior distribution of r_k is a probability mass function with only three values. Deriving the discrete-value posterior analogously to (2.3) yields

$$p(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2) = \frac{f(\mathbf{y}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) p(\boldsymbol{\theta}_1)}{\sum_{\boldsymbol{\theta}'_1} f(\mathbf{y}|\boldsymbol{\theta}'_1,\boldsymbol{\theta}_2) p(\boldsymbol{\theta}'_1)}.$$

The sum in the denominator of this equation requires the evaluation of the term equal to the numerator for all possible realizations of $\boldsymbol{\theta}_1$, just to obtain a normalizing constant. Evidently, it is more efficient to derive a proportional distribution neglecting the normalizing constants and normalize it after all values are obtained. We will therefore use the discrete-value equivalent of (4.5):

$$p(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2) \propto f(\mathbf{y}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) p(\boldsymbol{\theta}_1), \tag{4.6}$$

inserting r_k for $\boldsymbol{\theta}_1$ and $(\mathbf{r}_{\sim k}, \mathbf{a}, \sigma_n^2)$ for $\boldsymbol{\theta}_2$. However, no assumptions about a_k can be made while r_k is being estimated, which makes a_k a nuisance parameter. The joint distribution must therefore be marginalized by integrating over all possible values of a_k. For the sake of shorter equations we write $p(r_k|\ldots)$ for $p(r_k|\mathbf{y}, \mathbf{r}_{\sim k}, \mathbf{a}_{\sim k}, \sigma_n^2)$. The posterior is derived as follows:

$$\begin{aligned} p(r_k|\ldots) &\propto \int f(\mathbf{y}|r_k, \mathbf{r}_{\sim k}, a_k, \mathbf{a}_{\sim k}, \sigma_n^2) f(a_k)\,\mathrm{d}a_k\, p(r_k) \\ &\propto \int \exp\left(-\frac{(\mathbf{y} - \mathbf{F}_{\sim k}\,\mathbf{h}_{\sim k} - \mathbf{f}_k\, r_k\, a_k)^{\mathrm{T}} (\mathbf{y} - \mathbf{F}_{\sim k}\,\mathbf{h}_{\sim k} - \mathbf{f}_k\, r_k\, a_k)}{2\sigma_n^2} \right) \exp\left(-\frac{(a_k - \mu_a)^2}{2\sigma_a^2} \right) \mathrm{d}a_k\, p(r_k), \end{aligned}$$

where all constants have been dropped. The two exponents can be rearranged to yield a complete square of the integrand and an error term independent of a_k:

$$p(r_k|\ldots) \propto \int \exp\left(-\frac{(a_k-\tilde{\mu}(r_k))^2}{2\tilde{\sigma}^2(r_k)}\right) \mathrm{d}a_k \\ \exp\left(\frac{\tilde{\mu}^2(r_k)}{2\tilde{\sigma}^2(r_k)} - \frac{(\mathbf{y}-\mathbf{F}_{\sim k}\,\mathbf{h}_{\sim k})^{\mathrm{T}}(\mathbf{y}-\mathbf{F}_{\sim k}\,\mathbf{h}_{\sim k})}{2\sigma^2} - \frac{\mu_a^2}{2\sigma_a^2}\right) p(r_k),$$

using $\tilde{\mu}$ and $\tilde{\sigma}^2$ as defined below. These are constants with respect to a_k, but not with respect to r_k. Integration yields

$$\begin{aligned} p(r_k|\ldots) &\propto \sqrt{2\pi}\tilde{\sigma}(r_k)\,\exp\left(\frac{\tilde{\mu}^2(r_k)}{2\tilde{\sigma}^2(r_k)} - \frac{\|\mathbf{y}-\mathbf{F}_{\sim k}\,\mathbf{h}_{\sim k}\|^2}{2\sigma^2} - \frac{\mu_a^2}{2\sigma_a^2}\right) p(r_k) \\ &\propto \tilde{\sigma}(r_k)\,\exp\left(\frac{\tilde{\mu}^2(r_k)}{2\tilde{\sigma}^2(r_k)}\right) p(r_k) \end{aligned} \tag{4.7}$$

with the two auxiliary functions of r_k

$$\tilde{\sigma}^2(r_k) = \left(\frac{r_k^2\|\mathbf{f}_k\|^2}{\sigma_n^2} + \frac{1}{\sigma_a^2}\right)^{-1} \quad \text{and} \quad \tilde{\mu}(r_k) = \tilde{\sigma}^2\left(\frac{r_k\mathbf{f}_k^{\mathrm{T}}(\mathbf{y}-\mathbf{F}_{\sim k}\,\mathbf{h}_{\sim k})}{\sigma_n^2} + \frac{\mu_a}{\sigma_a^2}\right).$$

Since r_k can only assume three different values we can now obtain the respective probabilities by evaluating this statement for $r_k = 0$, $r_k = 1$ and $r_k = -1$ and dividing them by their sum.

The Amplitude Function. With the prior distribution being a conjugate prior, we can ignore all constants and use (4.5). In compliance with the considerations in section 2.1.4, the posterior distribution of a_k is a normal distribution with only the mean and variance altered from the prior distribution:

$$\begin{aligned} f(a_k|\mathbf{y},\mathbf{r},\mathbf{a}_{\sim k},\sigma_n^2) &\propto f(\mathbf{y}|r_k,\mathbf{r}_{\sim k},a_k,\mathbf{a}_{\sim k},\sigma_n^2)\,f(a_k) \\ &\propto \exp\left(-\frac{(\mathbf{y}-\mathbf{F}_{\sim k}\,\mathbf{h}_{\sim k}-\mathbf{f}_k\,r_k\,a_k)^T(\mathbf{y}-\mathbf{F}_{\sim k}\,\mathbf{h}_{\sim k}-\mathbf{f}_k\,r_k\,a_k)}{2\sigma_n^2}\right)\exp\left(-\frac{(a_k-\mu_a)^2}{2\sigma_a^2}\right) \\ &= \exp\left(-\frac{(a_k-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right) \\ f(a_k|\mathbf{y},\mathbf{r},\mathbf{a}_{\sim k},\sigma_n^2) &= \mathcal{N}(\tilde{\mu},\tilde{\sigma}^2) \end{aligned} \tag{4.8}$$

with the same $\tilde{\sigma}^2$ and $\tilde{\mu}$ as used in (4.7).

The Noise Variance. The same strategy can be used to obtain the posterior distribution of σ_n^2, since it is another conjugate prior. The result is an Inverse-Gamma distribution with the

parameters ξ and η altered from the prior distribution:

$$
\begin{aligned}
f(\sigma_n^2|\mathbf{y},\mathbf{r},\mathbf{a}) &\propto f(\mathbf{y}|\mathbf{r},\mathbf{a},\sigma_n^2)\, f(\sigma_n^2) \\
&\propto \frac{1}{(2\pi\sigma_n^2)^{K/2}} \quad \exp\left(-\frac{\|\mathbf{y}-\mathbf{F}\,\mathbf{h}\|^2}{2\sigma_n^2}\right) \quad \frac{\eta^{\xi}}{\Gamma(\xi)}\,(\sigma_n^2)^{-\xi-1} \quad \exp\left(-\frac{\eta}{\sigma_n^2}\right) \quad u(\sigma_n^2) \\
&\propto (\sigma_n^2)^{-\xi-1-K/2} \quad \exp\left(-\frac{\eta+\frac{1}{2}\|\mathbf{y}-\mathbf{F}\,\mathbf{h}\|^2}{\sigma_n^2}\right) \quad u(\sigma_n^2) \\
f(\sigma_n^2|\mathbf{y},\mathbf{r},\mathbf{a}) &= \mathcal{IG}\left(\xi+\frac{K}{2}\,, \quad \eta+\frac{1}{2}\|\mathbf{y}-\mathbf{F}\,\mathbf{h}\|^2\right). \qquad (4.9)
\end{aligned}
$$

4.1.4 Gibbs Sampler Algorithm

Having derived the posterior distribution of each parameter, we can implement a Gibbs sampler to generate a large number of samples from the joint posterior distribution $f(\mathbf{r},\mathbf{a},\sigma_n^2|\mathbf{y})$. The steps of the algorithm are as follows:

- Initialization:
  - for $k = 1, 2, ..., K$
    sample $r_k^{(0)}$ and $a_k^{(0)}$ from the prior distributions (4.2) and (4.3)
  - sample $\sigma_n^{2\,(0)}$ from the prior distribution (4.4)
  - set $i = 1$
- Iteration:
  - for $k = 1, 2, ..., K$
    sample $r_k^{(i)}$ and $a_k^{(i)}$ from the posterior distributions (4.7) and (4.8) where the most recent samples are used for the conditional parameters[1]
  - sample $\sigma_n^{2\,(i)}$ from the posterior distribution (4.9) where the most recent samples are used for the conditional parameters
  - set $i = i + 1$

4.1.5 Detection and Estimation

While $\mathbf{a}$ and σ_n^2 have continuous probability densities and can therefore be estimated using one of the standard estimators mentioned in 2.1.2, the discrete distribution of $\mathbf{r}$ calls for detection methods.

[1] For example, both posteriors require the conditional parameter $\mathbf{r}_{\sim k}$. "Using the most recent samples" means that, at iteration i and position k, we insert $[r_1^{(i)} r_2^{(i)} \cdots r_{k-1}^{(i)} r_{k+1}^{(i-1)} \cdots r_K^{(i-1)}]$.

Adhering to the MMSE principle, we basically average over the sample to obtain our estimates. This is simple in the case of σ_n^2 but slightly more complicated for the other two parameters.

For an efficient detection and estimation of the retina function we will first resolve an ambiguity in our model: since the retina function is defined as the product $h_k = r_k a_k$, each value can be obtained in two different ways, because both r_k and a_k can be positive or negative. We will therefore manipulate the sample in such a way that $\tilde{a}_k$ contains the full information about the sign of h_k, while $\tilde{r}_k$ becomes a binary variable (which also facilitates detection):

$$\begin{aligned} \tilde{r}_k^{(i)} &= |r_k^{(i)}| \\ \tilde{a}_k^{(i)} &= a_k^{(i)} \operatorname{sgn}\left(r_k^{(i)}\right), \end{aligned}$$

where $\operatorname{sgn}(x)$ denotes the sign of x.

Averaging over the sample $\tilde{\mathbf{r}}^{(i)}$ yields a continuous function $\hat{\mathbf{r}}_{\text{cont}}$, which requires quantization to ultimately obtain the estimate $\hat{\mathbf{r}}$. Experiments have shown that in the context of this model it is sufficient to set a threshold at 0.5 and obtain each $\hat{r}_k$ by a separate quantization. In section 4.3 it will become necessary to derive a more sophisticated quantization method for the entire vector.

When estimating $\mathbf{a}$ it is important to account for the fact that $\tilde{a}_k^{(i)}$ contains arbitrary values at all positions k and iterations i where $\tilde{r}_k^{(i)} = 0$. Therefore, $\hat{a}_k$ is the average of all $\tilde{a}_k^{(i)}$ such that $\tilde{r}_k^{(i)} = 1$. If there is no such (i, k), the value of $\hat{a}_k$ is itself arbitrary, since $\hat{r}_k$ at this position is definitely 0.

4.2 Refined Model with Parameter Correlations

4.2.1 Model Changes

The assumption of all parameters being statistically independent may be more restrictive than desired. While it seems sensible that a_k or σ_n^2 should be independent from all other parameters, the pulse positions vector $\mathbf{r}$ should more appropriately be described by a joint distribution. More specifically, the pulse positions (i.e. nonzero entries of $\mathbf{r}$) should not be allowed to be too close to each other. This can be justified by two different arguments. On one hand, the layers of tissue represented by the retina function have a finite width. On the other hand, it is impossible to detect multiple pulses within an interval that is small compared to the fringe width, in the presence of noise.

We will therefore introduce a minimum interval length, or minimum distance $d_{\min}$ and adjust the prior distribution of $\mathbf{r}$ such that no two nonzero values can exist within an interval smaller than $d_{\min}$.

Fig. 4.4 and Fig. 4.5 show a typical proportion of $d_{\min}$ and the fringe width.

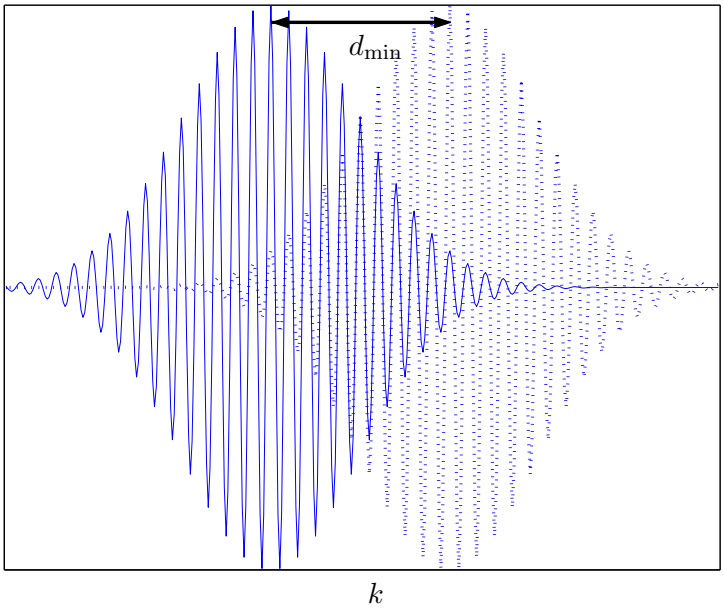


Figure 4.4: Two fringes at the minimum distance $d_{\min}$.

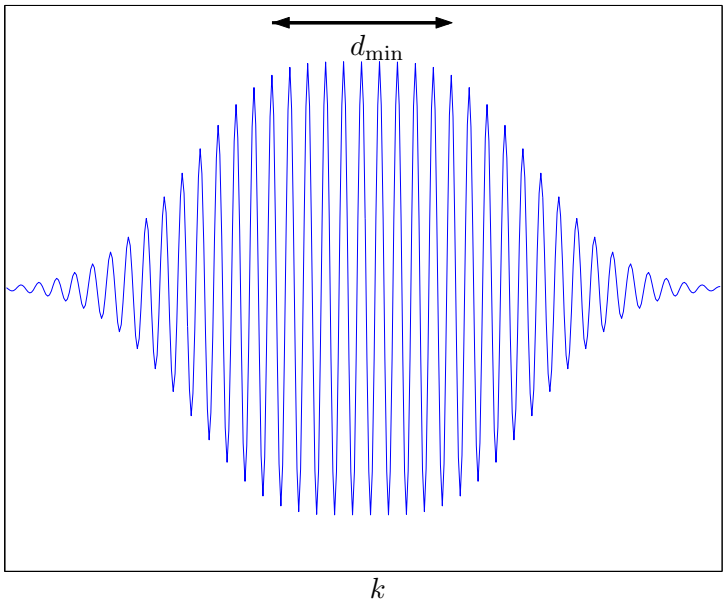


Figure 4.5: Example of the superposition of two fringes at the distance $d_{\min}$. The superposition of fringes at smaller distances may lead to detection errors.

4.2.2 Prior Distribution

Before defining the new prior distribution we will take another look at the prior distribution for statistically independent pulse positions by repeating (4.1) and (4.2):

$$p(\mathbf{r}) = \prod_{k=1}^{K} p(r_k) \qquad \text{and} \qquad p(r_k) = \begin{cases} p_0 & \text{for } r_k = 0 \\ \frac{1-p_0}{2} & \text{for } r_k = 1 \\ \frac{1-p_0}{2} & \text{for } r_k = -1. \end{cases}$$

This joint distribution of the vector $\mathbf{r}$ can be expressed directly by using L for the number of pulses (i.e. nonzero entries) contained in $\mathbf{r}$:

$$p(\mathbf{r}) = \left(\frac{1-p_0}{2}\right)^L p_0^{K-L}.$$

Equivalently we can define the interval lengths between two pulses: let $\mathbf{r}$ have its lth nonzero entry at position k and its $(l+1)$th nonzero entry at $k+d_l$. Then d_l is the length of the lth interval, as illustrated in Fig 4.6. This parameter d_l can be described by a probability mass function:

$$p(d_l) = p_{r_{k+1},\cdots,r_{k+d_l}|r_k}(0,\cdots,0,\pm 1|\pm 1) = p_0^{d_l-1}(1-p_0),$$

which is independent of k. Here, "± 1" means "nonzero", or equivalently "1 or -1". This distribution is a geometric distribution with the parameter p_0, as shown in Fig. 4.7.

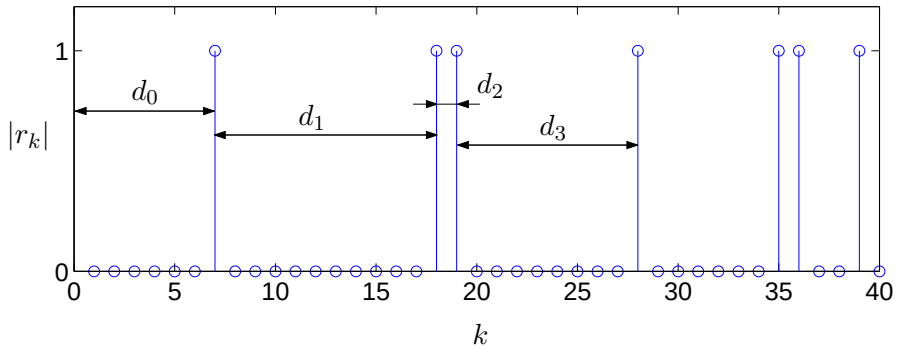


Figure 4.6: Exemplary position function and its interval lengths. The joint prior of $\mathbf{r}$ can be expressed as a function of the interval lengths, as done in (4.10).

If we define d_0 as the position of the first pulse in $\mathbf{r}$ and d_L such that $(K+1-d_L)$ is the position of the last pulse, then $p(\mathbf{r})$ can be expressed as:

$$p(\mathbf{r}) \propto p_0^{d_0-1} \prod_{l=1}^{L} p(d_l). \tag{4.10}$$

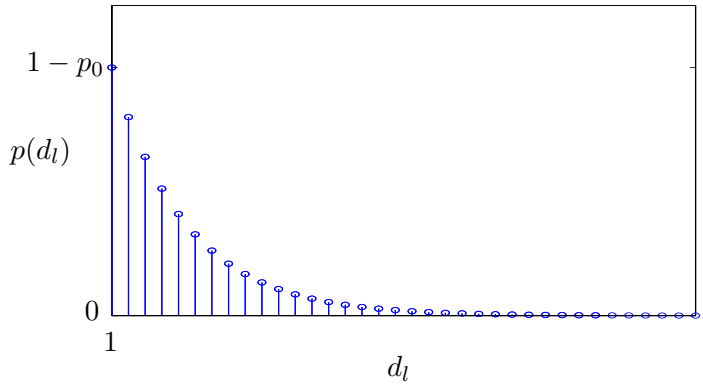


Figure 4.7: Prior distribution of d_l in the basic model (with i.i.d. positions).

(For equality, another scaling factor of $(\frac{1}{2})^L$ is needed because $p(d_l)$ does not distinguish between 1 and -1.)

Using these additional ways of describing $p(\mathbf{r})$ we will now adjust the distributions to introduce a minimum interval length $d_{\min}$. Our goal is to assign a probability of 0 to all occurrences of an interval smaller than $d_{\min}$.

The most obvious adjustment is that of $p(d_l)$ - it is shifted by $d_{\min}$:

$$p(d_l) = \begin{cases} 0 & \text{for } d_l < d_{\min} \\ p_0^{d_l - d_{\min} - 1}(1-p_0) & \text{else.} \end{cases} \tag{4.11}$$

The resulting distribution $p(d_l)$ is illustrated in Fig. 4.8.

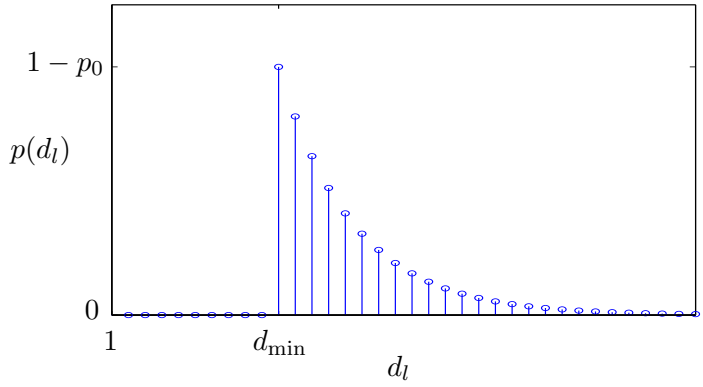


Figure 4.8: Prior distribution of d_l in the basic model (with minimum interval length $d_{\min}$).

Using the relation (4.10) we get the new joint prior distribution of $\mathbf{r}$:

$$p(\mathbf{r}) \propto \begin{cases} \left(\frac{1-p_0}{2}\right)^L p_0^{K-L} & \text{if } \forall l \in [1, L-1] : d_l \geq d_{\min} \\ 0 & \text{else.} \end{cases} \tag{4.12}$$

4.2.3 Posterior Distribution

In our basic model, the prior distributions of all parameters are independent from each other, whereas the posterior distributions are obviously not, because the Bayesian transformation introduces interrelations. Since the Gibbs sampler knows only the posterior distributions (using the priors for initialization is optional) it will also work with parameters that are a priori correlated. This means that in principle we can derive a new $p(r_k|\mathbf{y}, \mathbf{r}_{\sim k}, \mathbf{a}_{\sim k}, \sigma_n^2)$ and use it in the algorithm given in section 4.1.4. This new posterior can be derived analogously to (4.7) with the only difference that the factor $p(r_k)$ would have to be replaced by the joint prior $p(r_k, \mathbf{r}_{\sim k})$ defined in (4.12). This in turn means that the posterior probabilities would stay the same as in the i.i.d. case except for the case that there is a pulse in $\mathbf{r}_{\sim k}$ within the minimum distance from k. In that case the posterior probability for a pulse at k would be zero.

However, strong correlations of this type will strongly degrade the efficiency of the algorithm, since they counteract the basic idea of random jumps. For example, if at iteration i the sample $\mathbf{r}^{(i)}$ contains a pulse at position k, then it will be highly unlikely that in the subsequent samples this pulse gets shifted to a different position in the vicinity of k, because all the marginal posteriors of nearby positions assume the pulse at k as given and will therefore not produce new pulses. The only change possible is the removal of the pulse when sampling at the position k itself, which may by a very unlikely step, requiring a large number of iterations in order to happen. As mentioned in chapter 2, the tradeoff between parameter correlations and the required length of the Markov chain is a general property of MCMC methods. Since the correlations in the procedure proposed above are very strong, with a large number of deterministically predictable steps, we will not use this intuitive approach.

Alternatively, we could view the entire vector $\mathbf{r}$ as one single parameter, which would mean that all parameters are again statistically independent. By deriving a posterior distribution $p(\mathbf{r}|\mathbf{y}, \sigma_n^2)$ and drawing joint samples of the entire vector $\mathbf{r}$ we would thus avoid the problem mentioned above. However, we have to keep in mind that we cannot draw samples analytically, but only by evaluating the probabilities of all possible realizations. For the vector $\mathbf{r}$, this would imply evaluating 3^K probabilities (rather than 3 for each r_k), which would not be feasible for realistic lengths of $\mathbf{r}$.

A feasible solution to this problem can be found in the following approximation: When sampling r_k we split the vector $\mathbf{r}$ into two parts, $\mathbf{r}_{\mathrm{vic},k}$, which contains the vicinity of k and $\mathbf{r}_{\mathrm{rest},k}$, which contains the rest of $\mathbf{r}$. We then assume that the distribution of r_k depends only on $\mathbf{r}_{\mathrm{vic},k}$. This assumption is not entirely arbitrary: Due to the type of the correlation we introduced, the minimum interval length, the distribution of r_k is fully defined by the entries of $\mathbf{r}$ within a certain range

around r_k, rather than the whole vector:

$$p(r_k|\mathbf{r}_{\sim k}) = p(r_k|(\mathbf{r}_{\mathrm{vic},k}, \mathbf{r}_{\mathrm{rest},k})) = p(r_k|\mathbf{r}_{\mathrm{vic},k}), \tag{4.13}$$

which is true as long as $\mathbf{r}_{\mathrm{vic},k}$ contains all positions within the minimum interval length from k.

However, this is not sufficient for concluding that r_k is statistically independent of $\mathbf{r}_{\mathrm{rest},k}$:

$$\begin{aligned} p(\mathbf{r}) &= p(r_k, \mathbf{r}_{\mathrm{vic},k}, \mathbf{r}_{\mathrm{rest},k}) \\ &= p(r_k|\mathbf{r}_{\mathrm{vic},k}, \mathbf{r}_{\mathrm{rest},k}) p(\mathbf{r}_{\mathrm{vic},k}, \mathbf{r}_{\mathrm{rest},k}) \\ &= p(r_k|\mathbf{r}_{\mathrm{vic},k}) p(\mathbf{r}_{\mathrm{vic},k}, \mathbf{r}_{\mathrm{rest},k}) \\ &= p(r_k|\mathbf{r}_{\mathrm{vic},k}) p(\mathbf{r}_{\mathrm{vic},k}|\mathbf{r}_{\mathrm{rest},k}) p(\mathbf{r}_{\mathrm{rest},k}). \end{aligned}$$

In order to find a solution with manageable complexity we will neglect the dependency of $\mathbf{r}_{\mathrm{vic},k}$ on $\mathbf{r}_{\mathrm{rest},k}$:

$$p(\mathbf{r}) \approx p(r_k|\mathbf{r}_{\mathrm{vic},k}) p(\mathbf{r}_{\mathrm{vic},k}) p(\mathbf{r}_{\mathrm{rest},k}),$$

which makes r_k statistically independent of $\mathbf{r}_{\mathrm{rest},k}$.

This simplification allows us to sample r_k in the following way: We can first evaluate the probabilities of all possible instances of $\mathbf{r}_{\mathrm{vic},k}$, then marginalize this distribution to obtain the posterior of r_k. Within one iteration of the Gibbs algorithm we do this for each k proceeding from smaller to larger values. We will assume the positions smaller than k as given since they were evaluated before. We now define our vicinity interval as $[k, k + d_{\min} - 1]$. (To simplify the equations, $\mathbf{r}_{\mathrm{vic},k}$ will from here on also contain r_k.) The length of this vector is set to $d_{\min}$ because this is the minimal length for it to satisfy (4.13). At the same time, this length means that all realizations with more than one nonzero entry will have a probability of 0, limiting the number of possible realizations to $2d_{\min} + 1$.

This drastically reduces the workload from 3^K evaluations per Gibbs sampler iteration when sampling the entire vector $\mathbf{r}$ to $K(2d_{\min} + 1)$ when using our approximation. On the other hand, moving a pulse position within its vicinity becomes considerably more likely than it is when sampling r_k like in the uncorrelated case.

In the following equations we will also use $\mathbf{a}_{\mathrm{vic},k}$, $\mathbf{a}_{\mathrm{rest},k}$, $\mathbf{h}_{\mathrm{vic},k}$, $\mathbf{h}_{\mathrm{rest},k}$, $\mathbf{F}_{\mathrm{vic},k}$ and $\mathbf{F}_{\mathrm{rest},k}$, which are defined in analogy to $\mathbf{r}_{\mathrm{vic},k}$ and $\mathbf{r}_{\mathrm{rest},k}$. Furthermore, we use $\mathbf{1}_n$ to denote a vector of length n filled with ones.

As explained in section 4.1.3, we will neglect normalizing constants and derive a simpler function proportional to the desired distribution, using (4.6) and inserting $\mathbf{r}_{\mathrm{vic},k}$ for $\boldsymbol{\theta}_1$ and $(\mathbf{r}_{\mathrm{rest},k}, \mathbf{a}, \sigma_n^2)$

for $\boldsymbol{\theta}_2$. Again, $\mathbf{a}_{\mathrm{vic},k}$ is a nuisance parameter, since no assumptions can be made about it while $\mathbf{r}_{\mathrm{vic},k}$ is being estimated. The joint distribution is therefore marginalized by integrating over all possible values of $\mathbf{a}_{\mathrm{vic},k}$. For the sake of shorter equations we will write $p(\mathbf{r}_{\mathrm{vic},k}|\ldots)$ for $p(\mathbf{r}_{\mathrm{vic},k}|\mathbf{y}, \mathbf{r}_{\mathrm{rest},k}, \mathbf{a}_{\mathrm{rest},k}, \sigma_n^2)$. The derivation is mostly analogous to that in the basic model:

$$
\begin{aligned}
p(\mathbf{r}_{\mathrm{vic},k}|\ldots) &\propto \int f(\mathbf{y}|\mathbf{r}_{\mathrm{vic},k}, \mathbf{r}_{\mathrm{rest},k}, \mathbf{a}_{\mathrm{vic},k}, \mathbf{a}_{\mathrm{rest},k}, \sigma_n^2) f(\mathbf{a}_{\mathrm{vic},k})\, \mathrm{d}\mathbf{a}_{\mathrm{vic},k}\, p(\mathbf{r}_{\mathrm{vic},k}) \\
&\propto \int \exp\left(- \frac{\|\mathbf{y} - \mathbf{F}_{\mathrm{rest},k}\, \mathbf{h}_{\mathrm{rest},k} - \mathbf{F}_{\mathrm{vic},k}\, \mathrm{diag}(\mathbf{r}_{\mathrm{vic},k})\, \mathbf{a}_{\mathrm{vic},k}\|^2}{2\sigma_n^2} \right) \exp\left(- \frac{\|\mathbf{a}_{\mathrm{vic},k} - \mu_a \mathbf{1}_{d_{\min}}\|^2}{2\sigma_a^2} \right) \mathrm{d}\mathbf{a}_{\mathrm{vic},k}\, p(\mathbf{r}_{\mathrm{vic},k}),
\end{aligned}
$$

where all constants have been dropped. The two exponents can be rearranged to yield a complete square of the integrand and an error term independent of $\mathbf{a}_{\mathrm{vic},k}$:

$$
\begin{aligned}
p(\mathbf{r}_{\mathrm{vic},k}|\ldots) \propto \int \exp\left(- \frac{\|\mathbf{a}_{\mathrm{vic},k} - \tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})\|^2}{2\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})} \right) \mathrm{d}\mathbf{a}_{\mathrm{vic},k} \\
\exp\left(\frac{\|\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})\|^2}{2\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})} - \frac{(\mathbf{y} - \mathbf{F}_{\mathrm{rest},k}\, \mathbf{h}_{\mathrm{rest},k})^{\mathrm{T}} (\mathbf{y} - \mathbf{F}_{\mathrm{rest},k}\, \mathbf{h}_{\mathrm{rest},k})}{2\sigma_n^2} - \frac{\mu_a^2 d_{\min}}{2\sigma_a^2} \right) p(\mathbf{r}_{\mathrm{vic},k})
\end{aligned}
$$

using $\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})$ and $\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})$ as defined below. These are again constants with respect to the integrand, but not with respect to $\mathbf{r}_{\mathrm{vic},k}$. Integration yields

$$
\begin{aligned}
p(\mathbf{r}_{\mathrm{vic},k}|\ldots) &\propto \sqrt{2\pi}\tilde{\sigma}(\mathbf{r}_{\mathrm{vic},k}) \exp\left(\frac{\|\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})\|^2}{2\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})} - \frac{\|\mathbf{y} - \mathbf{F}_{\mathrm{rest},k}\, \mathbf{h}_{\mathrm{rest},k}\|^2}{2\sigma^2} - \frac{\mu_a^2 d_{\min}}{2\sigma_a^2} \right) p(\mathbf{r}_{\mathrm{vic},k}) \\
&\propto \tilde{\sigma}(\mathbf{r}_{\mathrm{vic},k}) \exp\left(\frac{\|\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})\|^2}{2\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})} \right) p(\mathbf{r}_{\mathrm{vic},k})
\end{aligned}
$$

with the two auxiliary functions of $\mathbf{r}_{\mathrm{vic},k}$

$$
\begin{aligned}
\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k}) &= \left(\frac{\mathrm{diag}(\mathbf{r}_{\mathrm{vic},k}) \mathbf{F}_{\mathrm{vic},k}^{\mathrm{T}} \mathbf{F}_{\mathrm{vic},k} \mathrm{diag}(\mathbf{r}_{\mathrm{vic},k})}{\sigma_n^2} + \frac{1}{\sigma_a^2} \right)^{-1} \\
\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k}) &= \tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k}) \left(\frac{\mathrm{diag}(\mathbf{r}_{\mathrm{vic},k}) \mathbf{F}_{\mathrm{vic},k}^{\mathrm{T}} (\mathbf{y} - \mathbf{F}_{\mathrm{rest},k}\, \mathbf{h}_{\mathrm{rest},k})}{\sigma_n^2} + \frac{\mu_a \mathbf{1}_{d_{\min}}}{\sigma_a^2} \right).
\end{aligned}
$$

To obtain the distribution of r_k, we will evaluate this statement for all $2d_{\min} + 1$ possible realizations of $\mathbf{r}_{\mathrm{vic},k}$, normalize it, and add up the resulting probabilities in an appropriate manner to derive the probabilities of the three possible realizations of r_k.

Having sampled r_k, we will proceed with the algorithm exactly as in the i.i.d. case, sampling a_k from (4.8), since the amplitudes are still assumed statistically independent of all other parameters.

Detection and estimation procedures still work in the same way as in the case of the basic model. Some further considerations of this are pursued in section 4.3.7.

4.3 Refined Model with Complex Amplitudes

4.3.1 Model Changes

The two models discussed so far focus on one property of the probe: the reflections of varying strength at the boundaries of tissue layers. In a more general approach, however, these layer boundaries may not only influence the amplitude but also the phase of the reflected wave, resulting in observations that cannot be described sufficiently well by our previous models.

Phase changes occur at boundaries of materials which are not perfectly absorption-free. Such materials are described by a complex refractive index n, which results in the reflection index itself being a complex number [15], since it is derived from the former as follows (cf. section 3.2.3):

$$a_i = \frac{n_i - n_{i+1}}{n_i + n_{i+1}}.$$

To include a representation of this effect, changes have to be made in the system model as well as the prior distributions.

4.3.2 System Model

In the basic system model described in section 3.2, the phase information of the reflected wave is determined by the definition of the fringe function, which was arbitrarily chosen in such a way that the maximum of the Gaussian envelope concurs with a local maximum of the cosine carrier signal. A more fundamental observation, however, than this specific relationship of envelope and phase, is the fact that such a relationship exists and that it is fixed and independent of the reflection process. In order to break this fixed relation, the system model must be refined in such a way that the phase information is contained in the retina function rather than the fringe function.

Using (3.3) and (3.5), the system can be described as follows:

$$\begin{aligned} y[k] &= \sum_{i=1}^{N} a_i \, f[k-k_i] + n[k] \\ &= \sum_{i=1}^{N} a_i \, \mathrm{e}^{-(\frac{k-k_i}{\alpha})^2} \cos(\omega_0(k-k_i)) + n[k]. \end{aligned}$$

In this expression, a phase change ϕ_i for each reflection can easily be inserted, which then

allows us to transform the system representation in the desired way:

$$\begin{aligned} y[k] &= \sum_{i=1}^{N} a_i \, \mathrm{e}^{-(\frac{k-k_i}{\alpha})^2} \cos(\omega_0(k-\phi_i)) + n[k] \\ &= \Re\left\{ \sum_{i=1}^{N} a_i \, \mathrm{e}^{-(\frac{k-k_i}{\alpha})^2} \mathrm{e}^{j\omega_0(k-\phi_i)} \right\} + n[k] \\ &= \Re\left\{ \left(\sum_{i=1}^{N} \underbrace{a_i \mathrm{e}^{-j\omega_0\phi_i}}_{a_i^{(\mathcal{C})}} \underbrace{\mathrm{e}^{-(\frac{k-k_i}{\alpha})^2}}_{f^{(\mathcal{C})}[k-k_i]} + n^{(\mathcal{C})}[k] \right) \mathrm{e}^{j\omega_0 k} \right\} \\ &= \Re\left\{ y^{(\mathcal{C})}[k] \; \mathrm{e}^{j\omega_0 k} \right\}. \end{aligned} \tag{4.14}$$

The system can thus be described conveniently in a complex baseband representation:

$$y^{(\mathcal{C})}[k] = \sum_{i=1}^{N} a_i^{(\mathcal{C})} \, f^{(\mathcal{C})}[k-k_i] + n^{(\mathcal{C})}[k]$$

The Fringe. In this complex baseband representation, the fringe appears shifted by its center frequency and is therefore represented as a Gaussian function around 0:

$$f^{(\mathcal{C})}[k] = \mathrm{e}^{-(\frac{k}{\alpha})^2}.$$

This definition can be seen as a direct consequence of (4.14). However, the baseband fringe can alternatively be derived from the original fringe by a passband-baseband transformation, which is most conveniently described in the frequency domain:

$$F^{(\mathcal{C})}(\mathrm{e}^{j\vartheta}) = F(\mathrm{e}^{j(\vartheta+\omega_0)}) \; u\left(\Im\{\mathrm{e}^{j(\vartheta+\omega_0)}\}\right),$$

where the discrete Fourier transform is denoted as

$$F(\mathrm{e}^{j\vartheta}) = \sum_{k=-\infty}^{\infty} f[k] \; \mathrm{e}^{-j\vartheta k},$$

$u(\cdot)$ denotes the unit step function, and ω_0 is dimensionless, as throughout this survey.

Resulting System Model. For simplicity of notation, the retina function will from here on not be specifically marked as complex. Returning to the multiplicative notation introduced in (3.4), the system is described as:

$$y^{(\mathcal{C})}[k] = (f^{(\mathcal{C})} * (a \cdot r))[k] + n^{(\mathcal{C})}[k],$$

or, in vector/matrix notation:

$$\mathbf{y}^{(\mathcal{C})} = \mathbf{F}^{(\mathcal{C})} \, \mathrm{diag}(\mathbf{r}) \, \mathbf{a} + \mathbf{n}^{(\mathcal{C})}.$$

4.3.3 Likelihood Function

The noise, denoted as $\mathbf{n}^{(\mathcal{C})}$ in the refined complex model, is modeled as complex Gaussian distributed, zero-mean and white:

$$\mathbf{n}^{(\mathcal{C})} \sim \mathcal{CN}(\mathbf{0}\,,\, \sigma_n^2 \mathbf{I}_K).$$

Evidently, this determines the shape of the complex likelihood function, which is given as follows:

$$\begin{aligned} f_{\mathbf{y}^{(\mathcal{C})}|\mathbf{r},\mathbf{a},\sigma_n^2}(\mathbf{y}^{(\mathcal{C})}|\mathbf{r},\mathbf{a},\sigma_n^2) &= f_{\mathbf{n}^{(\mathcal{C})}|\sigma_n^2}(\mathbf{y}^{(\mathcal{C})} - \mathbf{F}^{(\mathcal{C})}\,\mathrm{diag}(\mathbf{r})\,\mathbf{a}\,|\,\sigma_n^2) \\ &= \mathcal{CN}(\mathbf{F}^{(\mathcal{C})}\,\mathrm{diag}(\mathbf{r})\,\mathbf{a}\,,\, \sigma_n^2 \mathbf{I}_K). \end{aligned} \tag{4.15}$$

4.3.4 Prior Distributions

The Position Function. The retina function's prior distribution that was defined in the basic model, as depicted in Fig. 4.2, was chosen to comply with our model assumptions: a sparse signal (ensured by the peak at zero) with nonzero values that are on average above the noise level (small amplitudes are suppressed, while the symmetry around 0 is retained).

These fundamental properties of the prior distribution (sparseness, suppression of small amplitudes, and symmetry) are still desirable in the complex case, however, symmetry now means circular symmetry in the complex plane.

As explained in section 4.1.2, the way the retina function is split into $r[k]$ and $a[k]$ is arbitrary and serves the purpose of simplifying the calculations. The choices made for real values are not optimal in the complex domain. More specifically, information about the sign, previously included in the position function $r[k]$, is meaningless now. Therefore $r[k]$ is redefined to be yet again a binary function. Apart from this minor adjustment the considerations of section 4.2.2 are still valid, since the concept of the minimum interval length is maintained. The resulting distribution of the position vector $\mathbf{r}$ is therefore

$$p(\mathbf{r}) \propto \begin{cases} (1-p_0)^L p_0^{K-L} & \text{if } \forall l \in [1, L-1] : d_l \geq d_{\min} \\ 0 & \text{else,} \end{cases}$$

where p_0 is again a predefined constant, L denotes the number of pulses contained in the signal, and K is the length of $\mathbf{r}$.

The Amplitude Function. While the above definition ensures sparseness, the prior to be chosen for a_k should achieve the other two desired properties. A simple and effective solution is a

zero-mean complex Gaussian distribution:

$$a_k \sim \mathcal{CN}(0\,,\,\sigma_a^2). \tag{4.16}$$

The circular symmetry of this distribution is evident, and the suppression of smaller amplitudes can be verified by inspecting the resulting distribution of the magnitude $|a_k|$, which is a Rayleigh distribution [17] (as illustrated in Fig. 4.9):

$$|a_k| \sim \frac{|a_k|}{\sigma_a^2}\, \mathrm{e}^{-\frac{|a_k|^2}{2\sigma_a^2}}\, u(|a_k|).$$

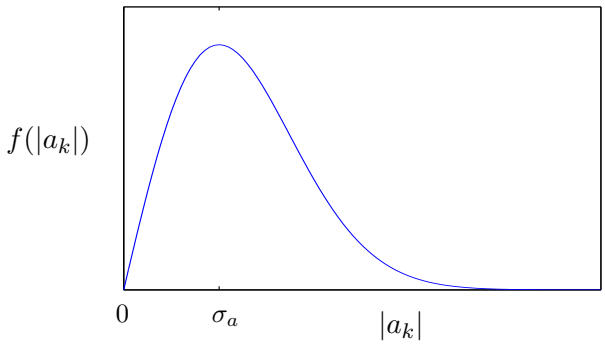


Figure 4.9: Prior distribution of $|a_k|$.

The Noise Variance. The prior distribution of σ_n^2 remains the same as in the basic model, since the extension of the likelihood function into the complex domain leaves the property of the Inverse Gamma distribution as a conjugate prior intact.

4.3.5 Posterior Distributions

The Position Function. As described in section 4.2.3, the posterior distribution of r_k is obtained in two steps: First, an expression proportional to the posterior distribution of $\mathbf{r}_{\mathrm{vic},k}$ is derived, using (4.6). This expression can then be normalized and marginalized to yield the desired distribution of r_k.

Once again, $\mathbf{a}_{\mathrm{vic},k}$ is a nuisance parameter and will be eliminated by integration. For the sake of shorter equations we will write $p(\mathbf{r}_{\mathrm{vic},k}|\ldots)$ for $p(\mathbf{r}_{\mathrm{vic},k}|\mathbf{y}^{(\mathcal{C})}, \mathbf{r}_{\mathrm{rest},k}, \mathbf{a}_{\mathrm{rest},k}, \sigma_n^2)$. The derivation is mostly analogous to that in the previous models:

$$p(\mathbf{r}_{\mathrm{vic},k}|\ldots) \propto \int f(\mathbf{y}^{(\mathcal{C})}|\mathbf{r}_{\mathrm{vic},k}, \mathbf{r}_{\mathrm{rest},k}, \mathbf{a}_{\mathrm{vic},k}, \mathbf{a}_{\mathrm{rest},k}, \sigma_n^2) f(\mathbf{a}_{\mathrm{vic},k})\, \mathrm{d}\mathbf{a}_{\mathrm{vic},k}\, p(\mathbf{r}_{\mathrm{vic},k})$$

$$\propto \int \exp\left(- \frac{\|\mathbf{y}^{(\mathcal{C})} - \mathbf{F}^{(\mathcal{C})}_{\mathrm{rest},k}\, \mathbf{h}_{\mathrm{rest},k} - \mathbf{F}^{(\mathcal{C})}_{\mathrm{vic},k}\, \mathrm{diag}(\mathbf{r}_{\mathrm{vic},k})\, \mathbf{a}_{\mathrm{vic},k}\|^2}{2\sigma_n^2} \right) \exp\left(- \frac{\|\mathbf{a}_{\mathrm{vic},k}\|^2}{2\sigma_a^2} \right) \mathrm{d}\mathbf{a}_{\mathrm{vic},k}\, p(\mathbf{r}_{\mathrm{vic},k}),$$

where all constants have been dropped. The two exponents can be rearranged to yield a complete square of the integrand and an error term independent of $\mathbf{a}_{\mathrm{vic},k}$:

$$p(\mathbf{r}_{\mathrm{vic},k}|\ldots) \propto \int \exp\left(-\frac{\|\mathbf{a}_{\mathrm{vic},k}-\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})\|^2}{2\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})}\right)\mathrm{d}\mathbf{a}_{\mathrm{vic},k}$$
$$\exp\left(\frac{\|\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})\|^2}{2\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})}-\frac{(\mathbf{y}^{(\mathcal{C})}-\mathbf{F}^{(\mathcal{C})}_{\mathrm{rest},k}\,\mathbf{h}_{\mathrm{rest},k})^{\mathrm{H}}(\mathbf{y}-\mathbf{F}^{(\mathcal{C})}_{\mathrm{rest},k}\,\mathbf{h}_{\mathrm{rest},k})}{2\sigma_n^2}\right)p(\mathbf{r}_{\mathrm{vic},k})$$

using $\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})$ and $\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})$ as defined below. These are again constants with respect to the integrand, but not with respect to $\mathbf{r}_{\mathrm{vic},k}$. Integration yields

$$p(\mathbf{r}_{\mathrm{vic},k}|\ldots) \propto \sqrt{2\pi}\tilde{\sigma}(\mathbf{r}_{\mathrm{vic},k})\,\exp\left(\frac{\|\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})\|^2}{2\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})}-\frac{\|\mathbf{y}^{(\mathcal{C})}-\mathbf{F}^{(\mathcal{C})}_{\mathrm{rest},k}\,\mathbf{h}_{\mathrm{rest},k}\|^2}{2\sigma^2}\right)p(\mathbf{r}_{\mathrm{vic},k})$$
$$\propto \tilde{\sigma}(\mathbf{r}_{\mathrm{vic},k})\,\exp\left(\frac{\|\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k})\|^2}{2\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})}\right)p(\mathbf{r}_{\mathrm{vic},k})$$

with the two auxiliary functions of $\mathbf{r}_{\mathrm{vic},k}$

$$\tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k}) = \left(\frac{\mathrm{diag}(\mathbf{r}_{\mathrm{vic},k})(\mathbf{F}^{(\mathcal{C})}_{\mathrm{vic},k})^{\mathrm{H}}\mathbf{F}^{(\mathcal{C})}_{\mathrm{vic},k}\mathrm{diag}(\mathbf{r}_{\mathrm{vic},k})}{\sigma_n^2}+\frac{1}{\sigma_a^2}\right)^{-1}$$
$$\tilde{\boldsymbol{\mu}}(\mathbf{r}_{\mathrm{vic},k}) = \tilde{\sigma}^2(\mathbf{r}_{\mathrm{vic},k})\left(\frac{\mathrm{diag}(\mathbf{r}_{\mathrm{vic},k})(\mathbf{F}^{(\mathcal{C})}_{\mathrm{vic},k})^{\mathrm{H}}(\mathbf{y}^{(\mathcal{C})}-\mathbf{F}^{(\mathcal{C})}_{\mathrm{rest},k}\,\mathbf{h}_{\mathrm{rest},k})}{\sigma_n^2}\right).$$

Since r_k can only assume the values 0 or 1 in this model and the length of $\mathbf{r}_{\mathrm{vic},k}$ is $d_{\min}$, the number of possible realizations of $\mathbf{r}_{\mathrm{vic},k}$ is $d_{\min}+1$, one of which contains only zero entries, while the others contain one nonzero entry each. Evaluating the probabilities of these realizations and adding them up appropriately yields the posterior distribution of r_k.

The Amplitude Function. With a complex Gaussian likelihood function and a complex Gaussian prior distribution for a_k, the convenience of a conjugate prior can once again be exploited, and the posterior distribution of a_k is of the same shape as its prior, with only the mean and the variance altered. It is obtained using (4.5):

$$f(a_k|\mathbf{y}^{(\mathcal{C})},\mathbf{r},\mathbf{a}_{\sim k},\sigma_n^2) \propto f(\mathbf{y}^{(\mathcal{C})}|r_k,\mathbf{r}_{\sim k},a_k,\mathbf{a}_{\sim k},\sigma_n^2)\,f(a_k)$$
$$\propto \exp\left(-\frac{(\mathbf{y}^{(\mathcal{C})}-\mathbf{F}^{(\mathcal{C})}_{\sim k}\,\mathbf{h}_{\sim k}-\mathbf{f}^{(\mathcal{C})}_k\,r_k\,a_k)^{\mathrm{H}}(\mathbf{y}^{(\mathcal{C})}-\mathbf{F}^{(\mathcal{C})}_{\sim k}\,\mathbf{h}_{\sim k}-\mathbf{f}^{(\mathcal{C})}_k\,r_k\,a_k)}{2\sigma_n^2}\right)\exp\left(-\frac{|a_k|^2}{2\sigma_a^2}\right)$$
$$= \exp\left(-\frac{|a_k-\tilde{\mu}|^2}{2\tilde{\sigma}^2}\right)$$
$$f(a_k|\mathbf{y}^{(\mathcal{C})},\mathbf{r},\mathbf{a}_{\sim k},\sigma_n^2) = \mathcal{CN}(\tilde{\mu},\tilde{\sigma}^2) \tag{4.17}$$

$$\text{with}\quad \tilde{\sigma}^2 = \left(\frac{r_k^2\|\mathbf{f}^{(\mathcal{C})}_k\|^2}{\sigma_n^2}+\frac{1}{\sigma_a^2}\right)^{-1}\quad\text{and}\quad \tilde{\mu} = \tilde{\sigma}^2\left(\frac{r_k(\mathbf{f}^{(\mathcal{C})}_k)^{\mathrm{H}}(\mathbf{y}^{(\mathcal{C})}-\mathbf{F}^{(\mathcal{C})}_{\sim k}\,\mathbf{h}_{\sim k})}{\sigma_n^2}\right).$$

The Noise Variance. The posterior distribution of σ_n^2 can be derived in analogy to (4.9), using the complex baseband representations:

$$
\begin{aligned}
f(\sigma_n^2|\mathbf{y}^{(\mathcal{C})},\mathbf{r},\mathbf{a}) &\propto f(\mathbf{y}^{(\mathcal{C})}|\mathbf{r},\mathbf{a},\sigma_n^2)\, f(\sigma_n^2) \\
&\propto \frac{1}{(2\pi\sigma_n^2)^{K/2}} \quad \exp\left(-\frac{\|\mathbf{y}^{(\mathcal{C})}-\mathbf{F}^{(\mathcal{C})}\,\mathbf{h}\|^2}{2\sigma_n^2}\right) \quad \frac{\eta^\xi}{\Gamma(\xi)}\,(\sigma_n^2)^{-\xi-1} \quad \exp\left(-\frac{\eta}{\sigma_n^2}\right) \quad u(\sigma_n^2) \\
&\propto (\sigma_n^2)^{-\xi-1-K/2} \quad \exp\left(-\frac{\eta+\frac{1}{2}\|\mathbf{y}^{(\mathcal{C})}-\mathbf{F}^{(\mathcal{C})}\,\mathbf{h}\|^2}{\sigma_n^2}\right) \quad u(\sigma_n^2) \\
f(\sigma_n^2|\mathbf{y}^{(\mathcal{C})},\mathbf{r},\mathbf{a}) &= \mathcal{IG}\left(\xi+\frac{K}{2}\,, \quad \eta+\frac{1}{2}\|\mathbf{y}^{(\mathcal{C})}-\mathbf{F}^{(\mathcal{C})}\,\mathbf{h}\|^2\right).
\end{aligned}
$$

4.3.6 Gibbs Sampler Algorithm

The implementation of the Gibbs sampler algorithm using this refined model is illustrated in Fig. 4.10.

4.3.7 Detection

Detecting the position sequence from the sample average becomes somewhat more complicated in this model (compared to section 4.1.5). Fig. 4.11 shows an example of such a sample average. The proposed positions are still discernible, since they have intervals of zeros between them. Furthermore, most nonzero intervals can be summed up to a value close to 1, which means that almost all samples contain a pulse in the region. However, their maxima generally do not exceed 0.5 as they do in the case of the other models. Simply lowering the threshold would lead to multiple pulses at adjacent positions.

This leads to the question why detection was in fact so easy in the models with real amplitudes. It can be explained as follows: The Gibbs sampler, in using the posterior distribution (which is narrow around its peaks since we assume a low level of observation noise), basically places a pulse at a position where the signal contains the maximum of a fringe (superposed with other fringes). If the fringe's phase and envelope have a fixed relation, the second most likely position is an entire period away from the maximum, which means that the Gaussian envelope is considerably smaller at this point. This gives the detector of such a model a great advantage. In the model with complex amplitudes, however, with the phase being independent from the fringe's envelope, the most plausible pulse positions are not restricted to multiples of the period around the true maximum, but cumulated directly around it, where the likelihood slowly decreases with the distance.

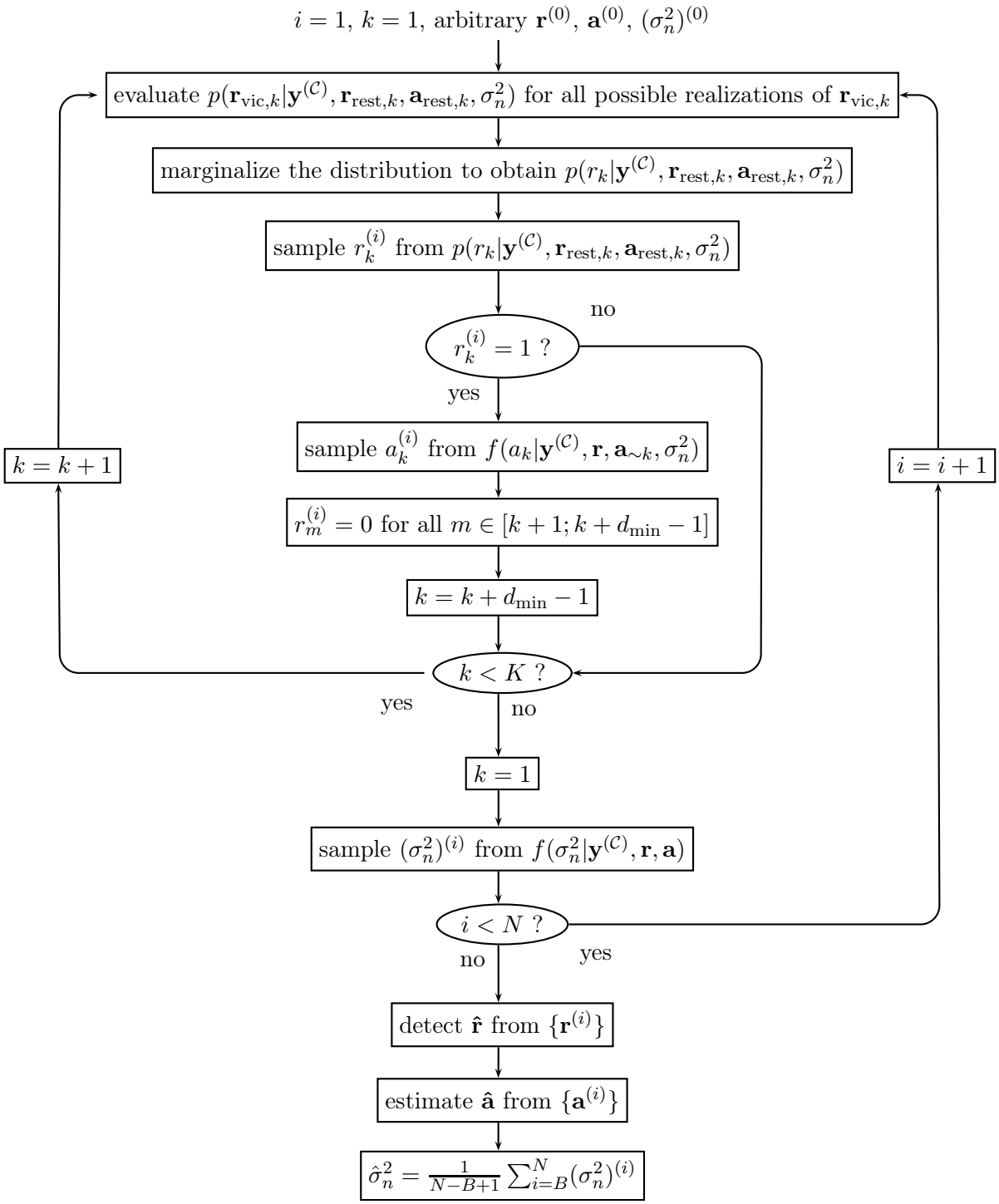


Figure 4.10: Flow chart of a MMSE estimator for the parameters $\mathbf{r}$, $\mathbf{a}$, and σ_n^2 using a Gibbs sampler.

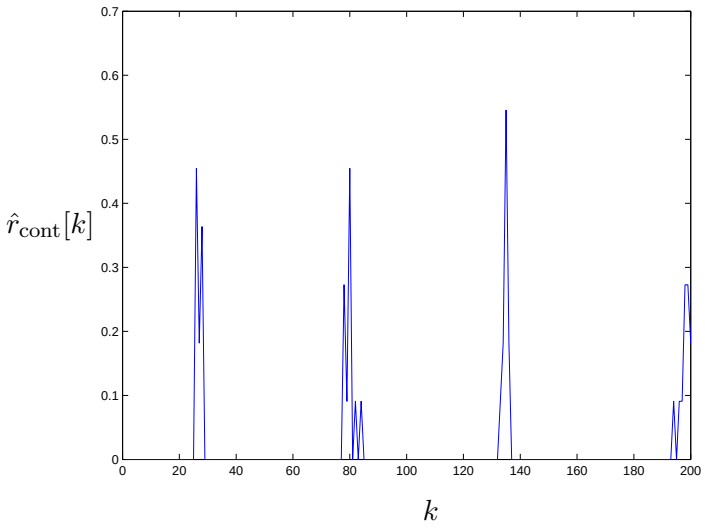


Figure 4.11: Exemplary sample average $\hat{\mathbf{r}}_{\text{cont}}$. The discrete-value function $\hat{\mathbf{r}}$ is to be obtained from this.

Various strategies seem viable for choosing the detected positions from the continuous function $\hat{\mathbf{r}}_{\text{cont}}$. At any rate, it seems plausible to proceed in two steps: defining intervals within which one pulse should be detected and choosing the exact position within that interval. The second step is considerably simple, we can use the mean of each interval, or the maximum. The latter seems plausible in our case since the maxima are quite distinct. For the first step we may rely on the fact that $\hat{\mathbf{r}}_{\text{cont}}$ contains zero intervals between all pulses and claim that each non-zero interval should contain a pulse. Alternatively we may rely on the fact that an interval where all samples contain a pulse sums up to 1 and place the interval borders based on the cumulative sum of $\hat{\mathbf{r}}_{\text{cont}}$. Both assumptions are usually justified in the case of OCT signals, which are characterized by low noise levels. In a combination of the two approaches we may choose to place one pulse in each nonzero interval with a sum greater than 0.5.

Chapter 5

Single Most Likely Replacement (SMLR) Detection

5.1 Introduction

SMLR [18] is an efficient detection method which has been used successfully in signal processing with OCT signals [4]. Since it is explicitly a detection method, it works with signal models which allow for a separation of detection and estimation by splitting the parameters into discrete and continuous variables, like the one introduced in chapter 3 and used in chapter 4.

Detection is based on the idea of maximizing the posterior probability. However, if the parameter is discrete and has a nonlinear relation with the observation, a global maximum can only be found by an exhaustive search, by evaluating the probability for all possible realizations of the vector [18]. Since this is highly inefficient or even impossible for longer vectors, the SMLR detector avoids it by aiming for a local maximum of the posterior probability. It is therefore a suboptimal but very efficient detection method.

The local maximum of the posterior probability is found in an iterative process, in which a given estimate is used to generate a number of hypotheses. For each of these, the posterior probability is evaluated, and the one with the largest probability is passed on to the next iteration step.

The algorithm is, therefore, not self-starting [18]. Its result strongly depends on the quality of the initial estimation. Obtaining such an initial estimation is discussed in section 5.5.

Like the Bayesian sampling methods, SMLR, too, uses the posterior probability distribution to obtain its results. However, in contrast to the samplers, its values at every step and substep are fully determined by the initial guess. Sampling methods, on the other hand, are independent

of initialization (after a sufficiently long burn-in period), and since they rely on drawing random samples, their outcome (e.g. the sample average) only becomes deterministic as soon as the law of large numbers applies (making the sample average equivalent to the mean of the distribution, in our example).

5.2 General Derivation

In a Bayesian estimation context, SMLR detection works with the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ of a binary parameter vector $\boldsymbol{\theta}$. Since the distribution is only used in a maximization, normalization can be neglected, which allows us to use the simple product of likelihood function and prior distribution instead:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})\ p(\boldsymbol{\theta}).$$

One step in the iterative process,

$$\boldsymbol{\theta}^{(i)} \longrightarrow \boldsymbol{\theta}^{(i+1)},$$

is performed as follows: the binary vector $\boldsymbol{\theta}^{(i)}$ is used to create M hypotheses $\boldsymbol{\zeta}_{m,i}$ (with $m = 1, \ldots, M$). Each of these hypotheses is a vector that differs from $\boldsymbol{\theta}^{(i)}$ in only one element. Therefore, the number of hypotheses equals the length of the parameter vector. Each of these $\boldsymbol{\zeta}_{m,i}$ (at one fixed i) is inserted as $\boldsymbol{\theta}$ in the posterior probability distribution. The hypothesis which achieves the largest probability is the candidate $\mathbf{w}$ for becoming the next guess $\boldsymbol{\theta}^{(i+1)}$:

$$\mathbf{w} = \arg\max_{\boldsymbol{\zeta}_m} p(\boldsymbol{\zeta}_{m,i}|\mathbf{y}) \qquad \text{for} \quad m = \{1, \ldots, M\}.$$

However, it is only accepted if its probability is larger than that of the current guess $\boldsymbol{\theta}^{(i)}$:

$$\boldsymbol{\theta}^{(i+1)} = \begin{cases} \mathbf{w} & \text{if } \Lambda(\mathbf{w}, \boldsymbol{\theta}^{(i)}) > 1 \\ \boldsymbol{\theta}^{(i)} & \text{else,} \end{cases}$$

using the likelihood ratio

$$\Lambda(\mathbf{w}, \boldsymbol{\theta}^{(i)}) = \frac{p(\mathbf{w}|\mathbf{y})}{p(\boldsymbol{\theta}^{(i)}|\mathbf{y})}.$$

Once a candidate $\mathbf{w}$ is not accepted, the iteration has obviously come to an end, and $\boldsymbol{\theta}^{(i)}$ is the detected sequence $\hat{\boldsymbol{\theta}}_{\mathrm{SMLR}}$, since it achieves the local maximum of the posterior probability distribution.

5.3 OCT Model Framework for SMLR Detection

SMLR detection has been used on OCT signals in [4], using a Bernoulli-Gaussian model similar to the one defined in section 4.1. Starting from a slightly modified version of the latter, we will introduce some refinements until we reach the model of section 4.3, which will allow us to compare the results obtained with the different estimators.

Our system model is, once again, the one defined in (3.6):

$$\begin{aligned} \mathbf{y} &= \mathbf{F}\,\mathbf{h} + \mathbf{n} \\ &= \mathbf{F}\,\mathrm{diag}(\mathbf{r})\,\mathbf{a} + \mathbf{n}, \end{aligned} \tag{5.1}$$

with i.i.d. parameter vectors $\mathbf{r}$ and $\mathbf{a}$ and an i.i.d. Gaussian noise vector $\mathbf{n}$. This leads to the likelihood function

$$f_{\mathbf{y}|\mathbf{r},\mathbf{a},\sigma_n^2}(\mathbf{y}|\mathbf{r},\mathbf{a},\sigma_n^2) = \mathcal{N}(\mathbf{F}\,\mathrm{diag}(\mathbf{r})\,\mathbf{a}\,,\ \sigma_n^2\mathbf{I}_K).$$

A fundamental difference between the procedures described in this chapter and the previous ones lies in their treatment of the parameters. The Bayesian samplers see $\mathbf{r}$, $\mathbf{a}$, and σ_n^2 as entirely equal and use the same estimation method for them. SMLR, on the other hand, is a detection method and thus concentrates on $\mathbf{r}$. The amplitudes $\mathbf{a}$ are estimated in a seperate intermediate step using a different method, while the noise variance σ_n^2 is not estimated at all but assumed to be known or guessed.

The posterior distribution we use for assessing the hypotheses is the joint posterior distribution $f(\mathbf{r},\mathbf{a}|\mathbf{y},\sigma_n^2)$. Since normalization is irrelevant, we can easily obtain it as a product of the above:

$$f(\mathbf{r},\mathbf{a}|\mathbf{y},\sigma_n^2) \propto f(\mathbf{y}|\mathbf{r},\mathbf{a},\sigma_n^2)p(\mathbf{r})f(\mathbf{a}).$$

This less strict treatment of the normalization - compared to the Gibbs sampler - is a direct consequence of the fact that the SMLR detector uses the distribution not for drawing samples, but for evaluating and comparing the probability of given samples. Another consequence of this difference, however, is a restriction that was in turn irrelevant for the Gibbs sampler: when evaluating the probability of a sequence, we must be careful to eliminate all nuisance parameters, since they would otherwise influence the probabilities we intend to compare. A consideration of our system model shows that the vector $\mathbf{a}$ contains such nuisance parameters. Only the entries at positions k where r_k is different from zero are relevant. The rest of $\mathbf{a}$ is filled with arbitrary values. This is already reflected by the likelihood function which ignores these values. However, the posterior product above also contains the prior distribution, which depends on all entries of $\mathbf{a}$. It must therefore

be marginalized to eliminate the arbitrary entries. This does not imply any computational effort, since the amplitudes are generally defined as statistically independent of each other. Denoting the non-zero entries of $\mathbf{a}$ by $\mathbf{a_r}$, we can therefore define (and use) a modified posterior distribution:

$$f_{\mathrm{SMLR}}(\mathbf{r}, \mathbf{a_r}|\mathbf{y}, \sigma_n^2) = f(\mathbf{y}|\mathbf{r}, \mathbf{a_r}, \sigma_n^2) p(\mathbf{r}) f(\mathbf{a_r}). \tag{5.2}$$

5.4 Amplitude Estimation

Using a joint distribution of $\mathbf{r}$ and $\mathbf{a}$ means that for every sequence $\mathbf{r}$ we want to assess we must also supply an adequate amplitude vector $\mathbf{a}$. Therefore, we have to estimate the amplitudes for every hypothesis we create. This estimation is performed using a least squares estimator, which exploits the linear relation of $\mathbf{y}$ and $\mathbf{a}$ (apart from the noise). However, the relation (5.1) is not invertible, since the values a_k for all k where $r_k = 0$ do not influence $\mathbf{y}$. We can avoid the problem by using the vector $\mathbf{a_r}$ again, along with the matrix $\mathbf{F_r}$, which is defined as $\mathbf{F}$ with all columns k removed where $r_k = 0$. The system model (5.1) can then be rewritten as

$$\mathbf{y} = \mathbf{F_r}\,\mathbf{a_r} + \mathbf{n},$$

which leads to the following LS estimate:

$$\hat{\mathbf{a}}_\mathbf{r} = \mathbf{F_r}^\sharp\,\mathbf{y} = \mathbf{a_r} + \mathbf{F_r}^\sharp\,\mathbf{n},$$

where $\mathbf{A}^\sharp$ denotes the left pseudoinverse of $\mathbf{A}$:

$$\mathbf{A}^\sharp = (\mathbf{A}^\mathrm{H}\,\mathbf{A})^{-1}\,\mathbf{A}^\mathrm{H}.$$

This estimate can then be inserted into (5.2).

5.5 Initialization of the Detector

Following the course of [4], we derive the initial estimate using the Nulling-And-Canceling (NAC) algorithm. This algorithm exploits the structure of our system model, more specifically the convolution of a known sequence (the fringe) with a sparse parameter vector. The observed signal can be seen as a superposition of a number of fringes, with the respective weights and time lags determined by the pulses of the retina function. Using a matched filter derived from the fringe, the NAC detector finds the position and value of the largest pulse. A fringe with the respective weight is then subtracted from the observed signal at this position. The procedure is repeated several times, until a limit is reached by setting a lower boundary for the significance of the detected pulse. As soon as the energy of the detected pulse does not exceed a certain proportion of the observation, the algorithm stops.

The results of this algorithm are good enough for the SMLR algorithm to find the "right" local maximum of the posterior distribution in most cases.

5.6 Prior Distributions and Model Refinements

Since the SMLR algorithm requires the position vector $\mathbf{r}$ to be binary, it is assigned a Bernoulli distribution with a fixed value of p_0:

$$p(r_k) = \begin{cases} p_0 & \text{for } r_k = 0 \\ 1-p_0 & \text{for } r_k = 1. \end{cases}$$

This definition evidently differs from one used for the basic model in section in 4.1.2, where r_k was allowed to assume the value -1. To obtain an estimator for the same basic model, i.e. with the prior distribution of $\mathbf{h}$ as shown in Fig. 4.2, the prior of $\mathbf{a}$ has to be redefined accordingly. The desired result is achieved by a Gaussian mixture distribution:

$$a_k \sim \frac{1}{2}\mathcal{N}(\mu_a, \sigma_a^2) + \frac{1}{2}\mathcal{N}(-\mu_a, \sigma_a^2).$$

Using these priors, we can directly compare the SMLR detector with the Bayesian sampler that works on the basic model.

Minimum Distance. Section 4.2 introduces a minimum distance between nonzero entries of $\mathbf{r}$. For the Bayesian sampler, this is implemented by changing the prior distribution - by defining a joint distribution of the vector rather than modeling its entries as i.i.d.. The same can be done to modify the SMLR detector, simply by inserting the new joint prior given in (4.12) into the joint posterior used for assessing the hypotheses.

However, the problem of correlation discussed in section 4.2.3 is even more severe for the SMLR detector than for the Gibbs sampler: If the detected sequence at one stage of the iterative process contains a pulse within the minimum distance from the true position of a pulse, the hypothesis adding a pulse at the true position will invariably have a probability of 0. Shifting the pulse from the wrong to the right position is only possible within two steps, by canceling the wrong pulse first, then adding the right pulse. However, the intermediate step with neither of the two would in most cases have a lower probability, therefore making the shift impossible altogether.

This problem can be solved by an amendment of the SMLR algorithm, which strictly speaking contradicts the term "single replacement", while however maintaining the basic idea of the procedure. We will simply alter the set of hypotheses assessed in each iteration. The pure SMLR

algorithm uses all realizations which differ from the previous sequence in one entry. Out of these, we can exclude a set of hypotheses which will a priori have a zero probability: all those which propose a new pulse within the minimum distance of a pulse from the previous iteration. On the other hand we want to include a set of new hypotheses, each with one of the previously detected pulses shifted within its own minimum distance vicinity. This way we avert the necessity of the intermediate step described above. Both goals can easily be combined by applying the following rule: starting from the set used by SMLR, we modify all hypotheses which propose additional pulses by setting the minimum distance interval around the new pulse to zero. The number of hypotheses is still the length of the vector. Therefore the problem of correlation is solved, while the key to the efficiency of the algorithm, its small number of hypotheses, is preserved.

Complex Amplitudes. As explained in section 4.3, the OCT system can be described more appropriately by allowing for phase changes at each reflection. The complex baseband model which reflects this extension can be used for SMLR detection. We therefore recall the system model,

$$\mathbf{y}^{(\mathcal{C})} = \mathbf{F}^{(\mathcal{C})} \operatorname{diag}(\mathbf{r})\, \mathbf{a} + \mathbf{n}^{(\mathcal{C})},$$

with the Töplitz matrix $\mathbf{F}^{(\mathcal{C})}$ containing the baseband fringe,

$$F_{ij}^{(\mathcal{C})} = f^{(\mathcal{C})}[i-j], \quad \text{where} \quad f^{(\mathcal{C})}[k] = \mathrm{e}^{-(\frac{k}{\alpha})^2},$$

and the noise distributed according to

$$\mathbf{n}^{(\mathcal{C})} \sim \mathcal{CN}(\mathbf{0}\,,\, \sigma_n^2 \mathbf{I}_K).$$

The posterior distribution used in the detector algorithm is still (5.2), however we insert the complex likelihood function (4.15)

$$f_{\mathbf{y}^{(\mathcal{C})}|\mathbf{r},\mathbf{a},\sigma_n^2}(\mathbf{y}^{(\mathcal{C})}|\mathbf{r},\mathbf{a},\sigma_n^2) = \mathcal{CN}(\mathbf{F}^{(\mathcal{C})} \operatorname{diag}(\mathbf{r})\, \mathbf{a}\,,\, \sigma_n^2 \mathbf{I}_K)$$

and the complex amplitude prior (4.16)

$$a_k \sim \mathcal{CN}(0\,,\, \sigma_a^2).$$

5.7 Application of the Algorithm

The application of the entire procedure comprising NAC initialization and SMLR pulse detection with intermediate steps of LS amplitude estimation is depicted in flow charts in Fig. 5.1 and Fig. 5.2. In these diagrams, $f_{\mathrm{MF}}[k] = \left(f^{(\mathcal{C})}\right)^*[-k]$, $\mathbf{0}_K$ denotes a vector of K zeros, and λ denotes the stopping threshold of the NAC algorithm.

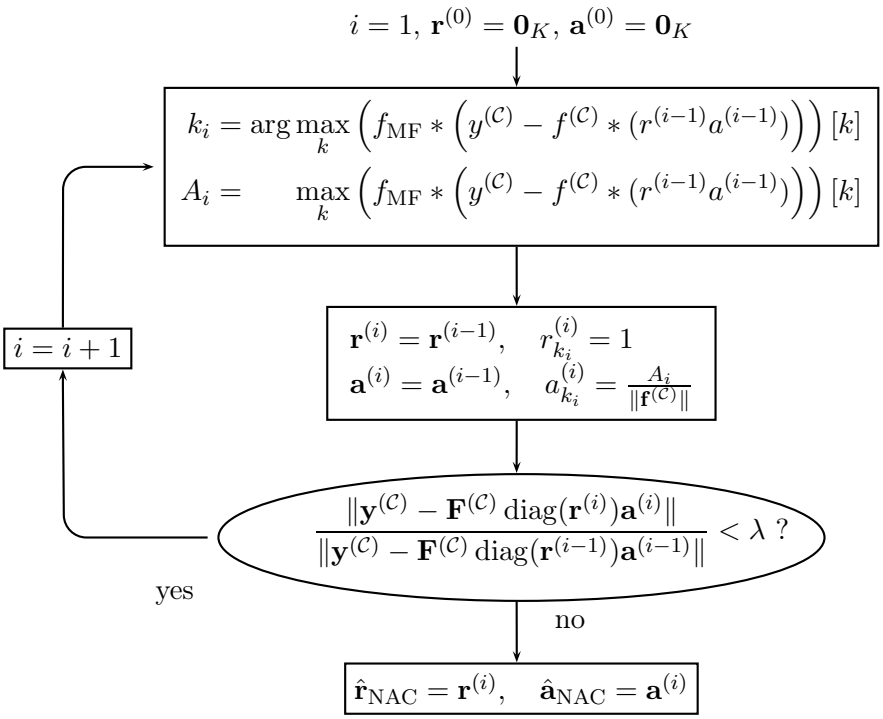


Figure 5.1: Flow chart of the NAC algorithm applied to OCT.

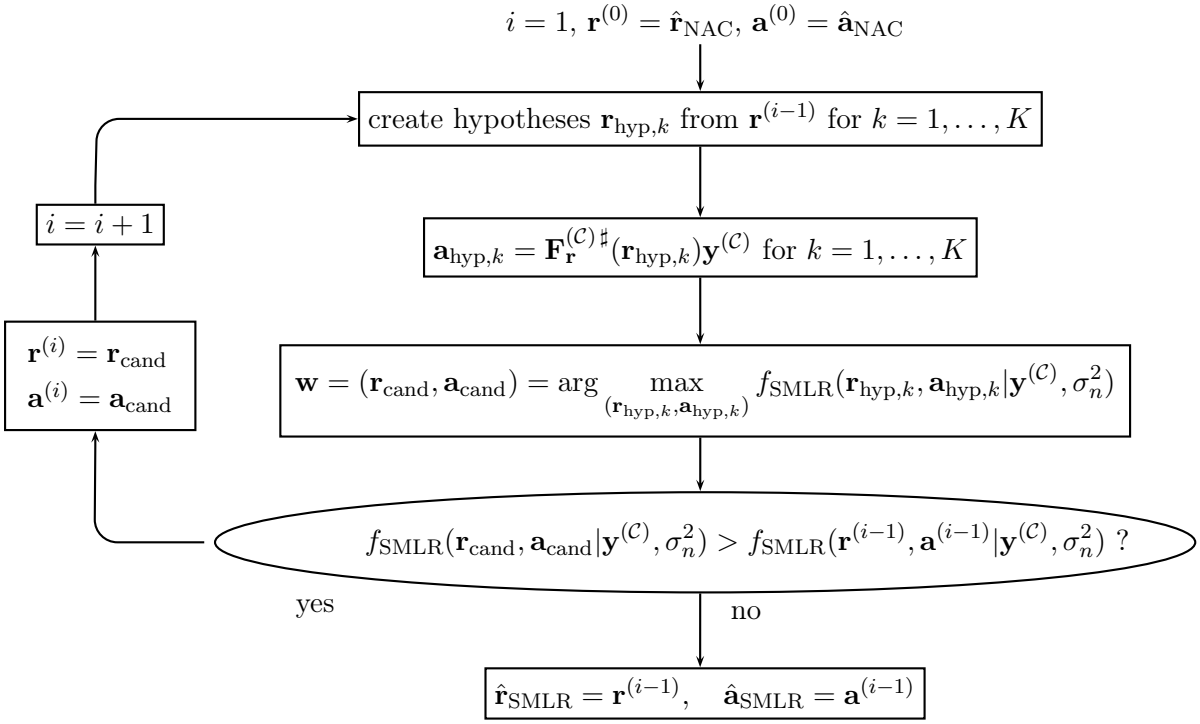


Figure 5.2: Flow chart of the SMLR detector applied to OCT.

Chapter 6

Numerical Results

To evaluate and compare the performance of the different estimators, synthetic OCT signals are generated based on the models and prior distributions described in chapter 4. The results obtained with these synthetic signals are discussed in sections 6.1 to 6.5, while real OCT signals are used in section 6.6.

6.1 Measuring Estimation Errors

The Mean Square Error. The essential question of how to rate the quality of an estimate turns out to be non-trivial in this case of a joint detection and estimation problem. A standard error measure for estimation is given in the mean square error (MSE). We can use it to assess how accurately an estimated signal $\hat{\mathbf{y}} = \mathbf{F}\hat{\mathbf{h}}$ approximates the true signal:

$$\mathrm{MSE} = \frac{\|\hat{\mathbf{y}} - \mathbf{F}\mathbf{h}\|^2}{\|\mathbf{F}\mathbf{h}\|^2}.$$

However, very close approximations may be achieved with retina functions $\hat{\mathbf{h}}$ that differ strongly from the true retina function, especially if the number of pulses (or the distance between pulses) is not limited. Therefore, estimates must be assessed directly, without reconstructing a signal $\hat{\mathbf{y}}$. Calculating a MSE between the sparse vectors $\mathbf{h}$ and $\hat{\mathbf{h}}$ is useless, because it disregards the distance between a true pulse position and a detected pulse position.

Custom Error Measure. The most important information within the retina function is contained in the pulse positions. The quality of the detected position function $\hat{\mathbf{r}}$ is therefore most decisive for the performance of the estimator. Furthermore, the quality of the estimated amplitudes is strongly correlated with the quality of the detected pulse positions. We therefore define

an error measure that solely regards the pulse positions, comparing $\mathbf{r}$ and $\hat{\mathbf{r}}$. Let R and $\hat{R}$ denote the number of nonzero elements in $\mathbf{r}$ and $\hat{\mathbf{r}}$:

$$R = \sum_{k=1}^{K} |r_k| \qquad \hat{R} = \sum_{k=1}^{K} |\hat{r}_k|,$$

and $q_{\mathbf{r},i}$ with $i = 1, \ldots, R$ the position of the i-th nonzero element of $\mathbf{r}$, i.e. the position the i-th pulse in the retina function. Then the error measure is defined as

$$\epsilon_{\mathbf{r}} = \frac{1}{2R} \left(\sum_i \min_j |q_{\mathbf{r},i} - q_{\hat{\mathbf{r}},j}| + \sum_j \min_i |q_{\mathbf{r},i} - q_{\hat{\mathbf{r}},j}| \right).$$

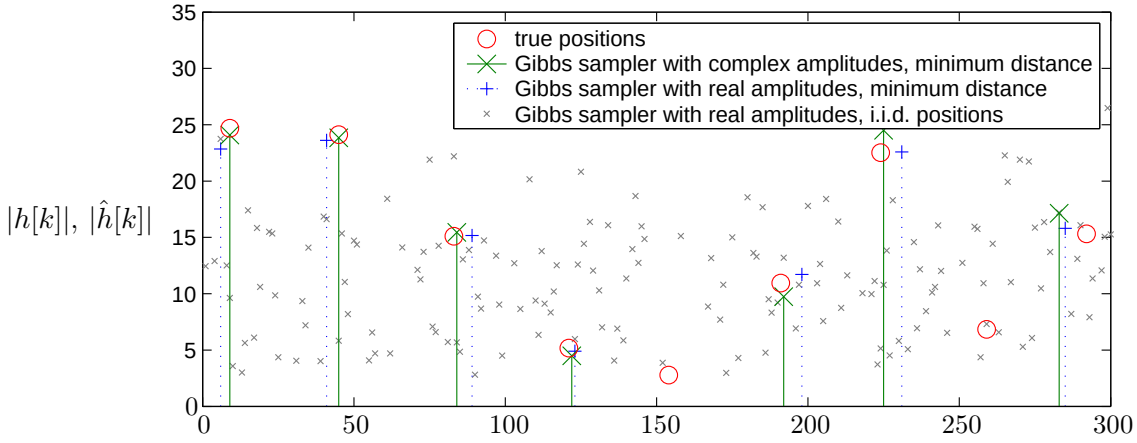
It can be interpreted as follows: If the detected sequence contains the right number of pulses and each detected pulse is in the vicinity of one true pulse (i.e. within a distance that is small compared to the distances between the true pulses), then the error measure is equal to the average distance between a true position and the detected position in its vicinity. For each true position that is missed completely, a bigger penalty term is added (equal to half the distance to the nearest detected pulse). Conversely, for each additional pulse in the detected sequence, a similar penalty term is added (equal to half the distance to the nearest true pulse).

One important special case has to be defined separately: if he detected sequence does not contain any pulses (but the true sequence does), the error measure is set to equal the vector length K.

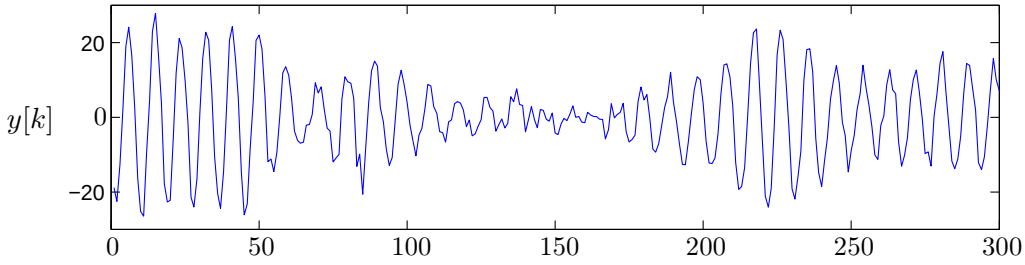
6.2 Performance of Gibbs Samplers with Different Models

Chapter 4 describes three different signal models and the implementation of the respective Gibbs sampler estimators. Since the three estimators are based on different assumptions, they are optimized for different working conditions. For simulations, however, we generate all test signals according to the model which is closest to reality, i.e. the refined model with complex amplitudes and a minimum distance.

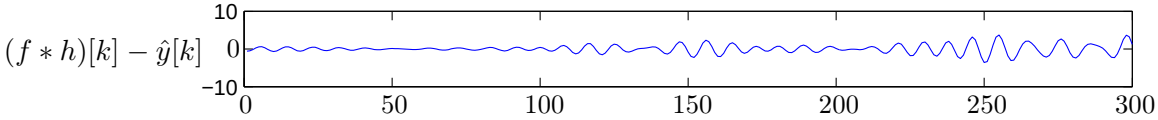
Fig. 6.1 shows an example of a synthetic OCT signal and the respective estimates. In this example, the SNR is 14dB, which is still considerably lower than in typical OCT signals. The parameters of the estimators' prior distribution of σ_n^2 are set such that the mean equals the true value. The estimates are obtained each from a chain with 256 iterations and a burn-in period of length 6. The fringe is defined to have a fringe width $\alpha = 23$ and a period $2\pi/\omega_0 = 9$. The minimum distance is set to 33, with a probability $p_0 = 0.8$ for zeros in the position function, which results in an average pulse distance of 37.



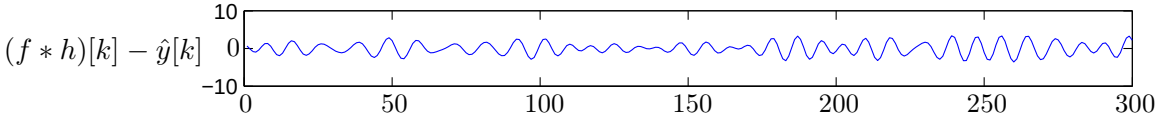
(a) True and estimated retina function.



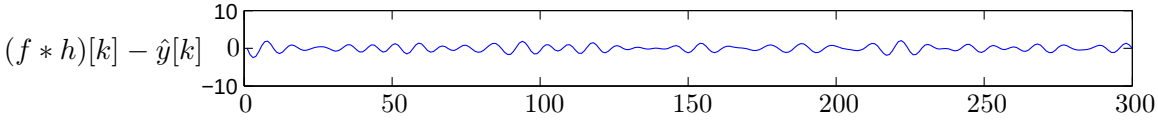
(b) Observed signal.



(c) Error signal of the Gibbs sampler with complex amplitudes, minimum distance.



(d) Error signal of the Gibbs sampler with real amplitudes, minimum distance.



(e) Error signal of the Gibbs sampler with real amplitudes, i.i.d. positions.

Figure 6.1: Comparison of different Gibbs sampler implementations.

The error measures in this example are as follows:

	$\epsilon_{\mathbf{r}}$	MSE
Gibbs sampler with complex amplitudes, minimum distance	4.56	−19.78dB
Gibbs sampler with real amplitudes, minimum distance	4.17	−16.89dB
Gibbs sampler with real amplitudes, i.i.d. positions	75.22	−22.66dB

The example illustrates what was mentioned in the previous section: the MSE cannot be considered a useful error measure. The lowest MSE is achieved by the estimator using the i.i.d. model, while its estimate is obviously useless. The numerous excess pulses it detects lead to the high error measure $\epsilon_{\mathbf{r}}$, but their superposition optimizes the approximation of the signal. This may be surprising when we take into account that the prior was set such that the average distance between pulses equals the true average distance. We can conclude that the likelihood function, due to its small variance, has much stronger influence on the posterior distribution than the prior does. It takes a strict precondition like the enforced minimum distance to visibly overcome this dominance of the likelihood function.

More extensive simulations in the following sections will offer further and more representative comparison between the estimators under varying conditions.

6.3 Comparison of Gibbs Sampler and SMLR Detector

Fig. 6.2 uses the same synthetic retina function as an example to compare the outcome of the Gibbs sampler and the SMLR detector, both with the most realistic signal model. This example should give an insight as to how differences in the resulting performance are caused, while more representative simulations comparing the performance itself under varying circumstances are discussed later.

The SMLR detector is provided with the true value of the SNR and initialized with a NAC-algorithm using a threshold parameter of $\lambda = 0.5$.

The error measures in this example are as follows:

	$\epsilon_{\mathbf{r}}$	MSE
Gibbs sampler with complex amplitudes, minimum distance	4.56	−19.78dB
SMLR detector with complex amplitudes, minimum distance	7.33	−9.95dB

One detail that illustrates the different working procedures of the two algorithms is the SMLR detector's error at position 30. The position may have been suggested by the NAC algorithm because it covers a good part of the energy that is really caused by the two true neighboring pulses. While placing two pulses at (or near) the true positions would evidently achieve a higher

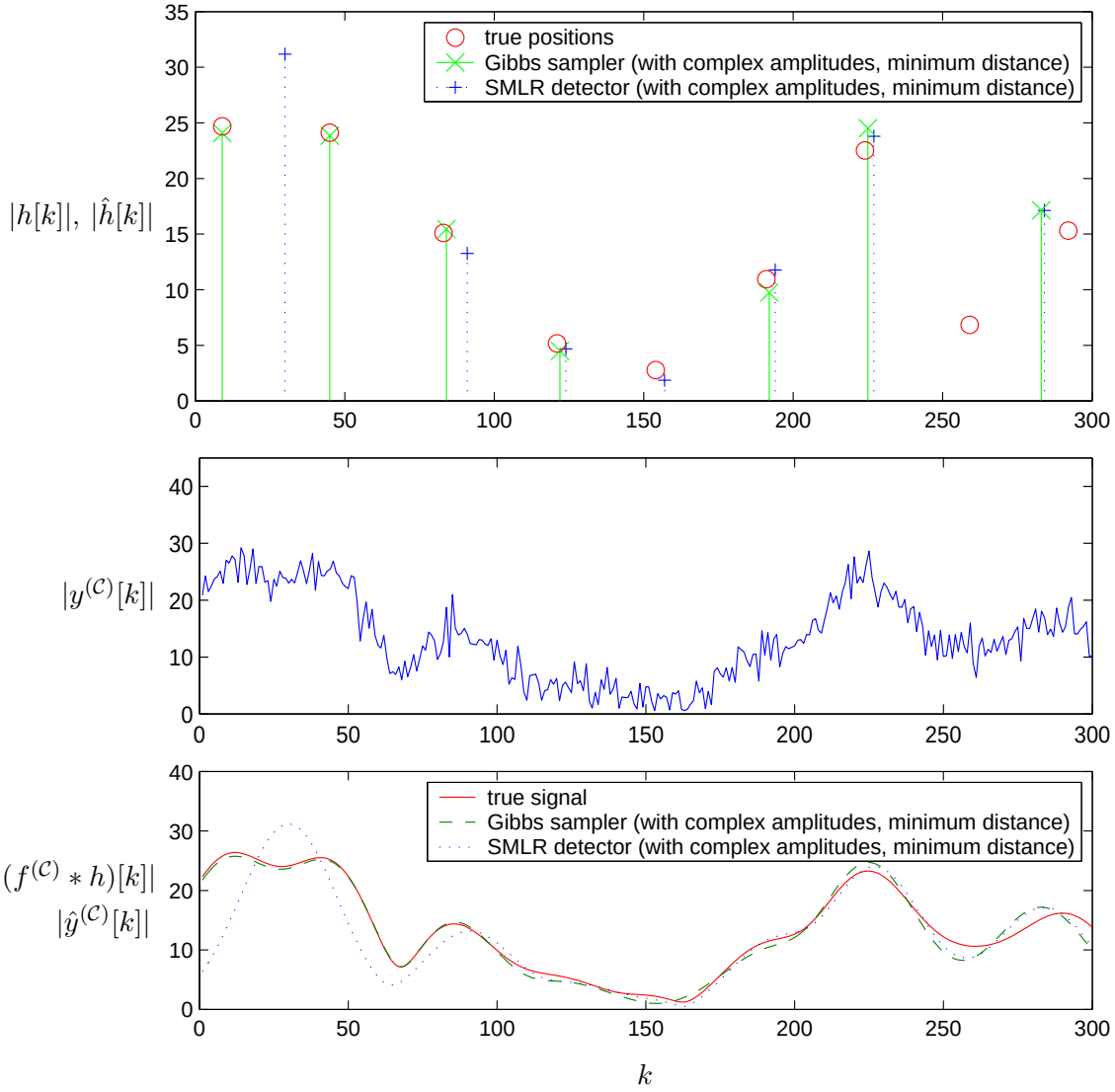


Figure 6.2: Example comparing Gibbs sampler and SMLR detector.

posterior probability, it cannot be done by the SMLR algorithm, since it uses only hypotheses with one replacement, not two. It would take two steps, moving the pulse to one of the true positions first, then adding the second pulse. This intermediate step, however, would have a lower posterior probability, therefore the detector remains at this local maximum of the posterior distribution. While this behavior is a disadvantage of the SMLR detector, its performance appears not to be generally lower than that of the Gibbs sampler, as can be seen in the simulation results following Fig. 6.3.

The following simulation results (Fig. 6.3-6.8) were obtained by averaging over the results of

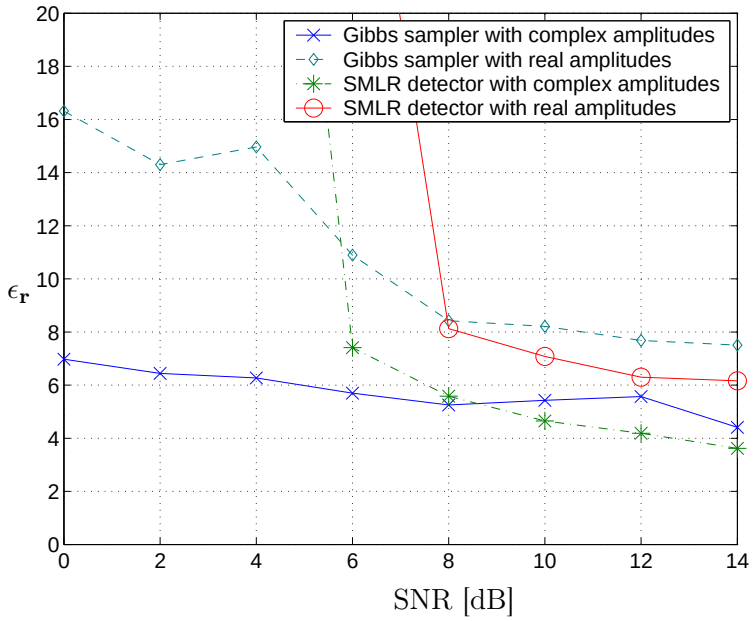


Figure 6.3: Error measure $\epsilon_{\mathbf{r}}$ for varying SNR.

150 experiments with synthetic OCT signals generated independently from the prior distributions given in chapter 4. Unless otherwise specified, all parameters are kept equal to the values mentioned in the section above.

Fig. 6.3 and Fig. 6.4 show the performance of the Gibbs sampler and the SMLR detector, both with minimum distance and complex amplitudes, for varying SNR.

Since the average distance between pulses is 37, an error measure $\epsilon_{\mathbf{r}}$ that comes close to half that value means that the estimate is useless. Much higher values of $\epsilon_{\mathbf{r}}$ can occur if several estimates are empty, which is the case when the SNR becomes too low for the SMLR detector.

Fig. 6.5 and Fig. 6.6 show the performance of the different estimators for a varying minimum distance. With a constant $p_0 = 0.8$, the average distance between pulses is always $d_{\min} + 4$ (as explained in (6.1) in the next section). This should be kept in mind when looking at the error measure $\epsilon_{\mathbf{r}}$, since an error measure close to half the average distance means the estimate is useless, as mentioned above. In this simulation, this is the case for all estimators at $d_{\min} = 10$ but for none at $d_{\min} \geq 20$. A minimum distance of 1 denotes an i.i.d. position function.

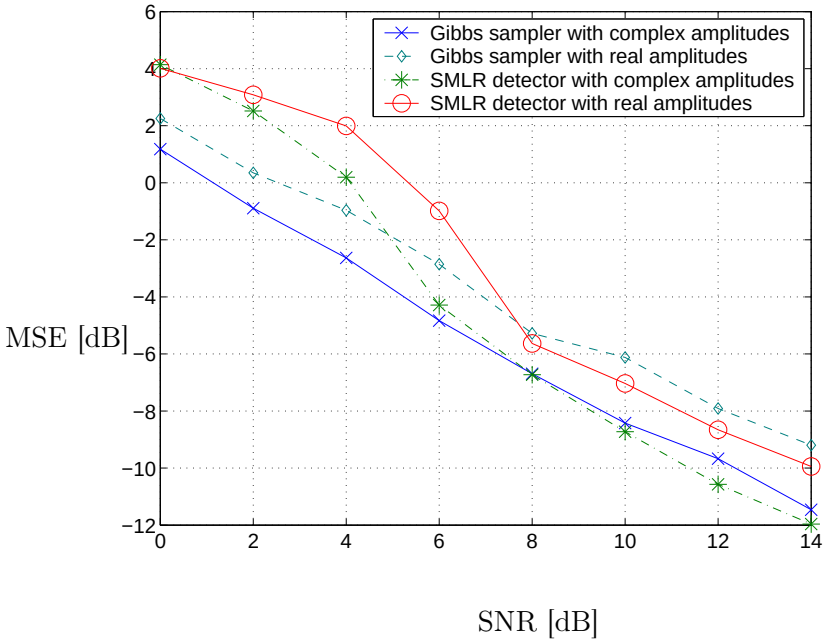


Figure 6.4: MSE for varying SNR.

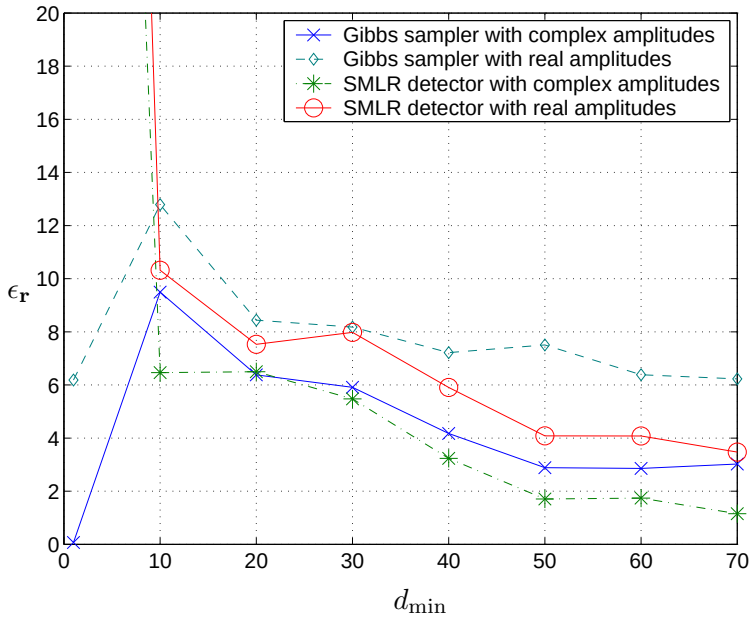


Figure 6.5: Error measure $\epsilon_{\mathbf{r}}$ for varying minimum distance $d_{\min}$.

6.4 Variation of Parameters

The signal model described in chapter 4 contains a number of constants which are predefined. These are the minimum distance $d_{\min}$ and the zero probability p_0 in the prior distribution of $\mathbf{r}$,

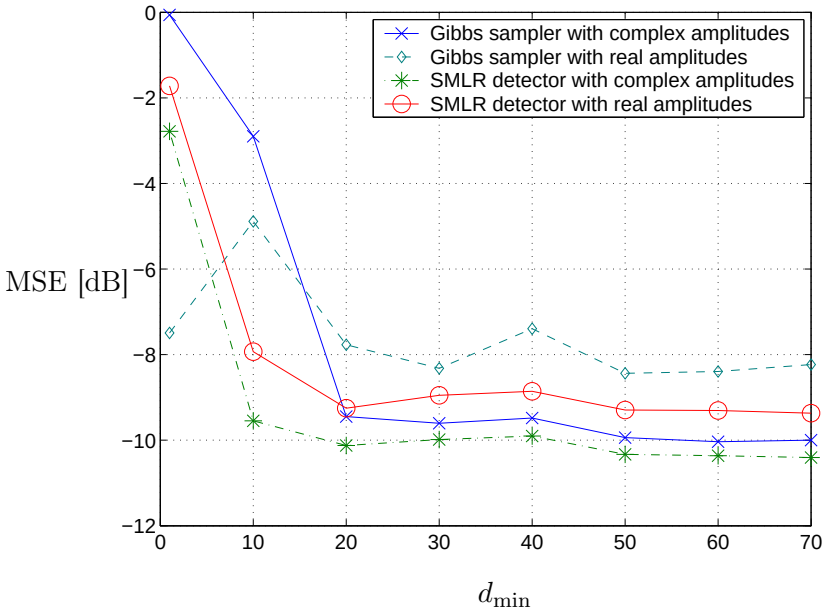


Figure 6.6: MSE for varying minimum distance $d_{\min}$.

the mean μ_a and variance σ_a^2 in the prior distribution of $\mathbf{a}$ and the shape parameters ξ and η in the prior distribution of σ_n^2. The constants in the priors of $\mathbf{a}$ and σ_n^2 influence the performance only insofar as they alter the SNR. After illustrating the dependence of the performance on the SNR and $d_{\min}$, the role of p_0 remains to be discussed.

The zero probability p_0 has a simple interpretation in the i.i.d. model, where it denotes the probability that $r_k = 0$ at any position k. In the refined models with a joint probability distribution of $\mathbf{r}$ it loses this meaning. However, it still has a direct influence on the average distance between pulses, which is given as

$$\bar{d} = d_{\min} - 1 + \frac{1}{1 - p_0}. \tag{6.1}$$

Note that $d_{\min}$ can never be 0, in the i.i.d. case it is 1.

The influence of p_0 on the estimators' performance is shown in Fig. 6.7 and Fig. 6.8.

6.5 Fringe Mismatch

When working with real OCT signals, the fringe parameters are not perfectly known but have to be estimated from the signals. Therefore the influence of a mismatch of the fringe parameters on the performance seems worth discussion.

A frequency mismatch appears to have no discernible influence on the error measure $\epsilon_{\mathbf{r}}$, i.e.

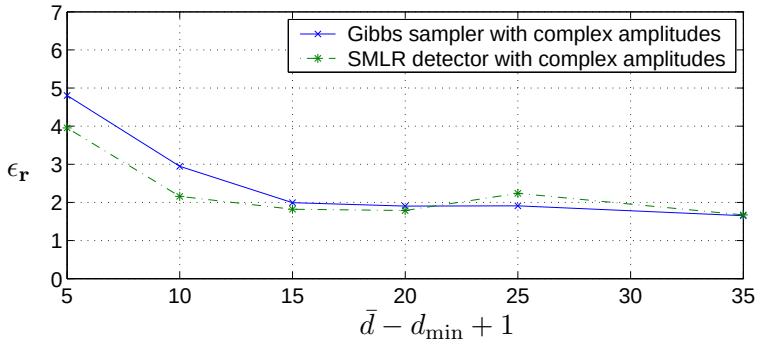


Figure 6.7: Error measure $\epsilon_{\mathbf{r}}$ for varying average distance.

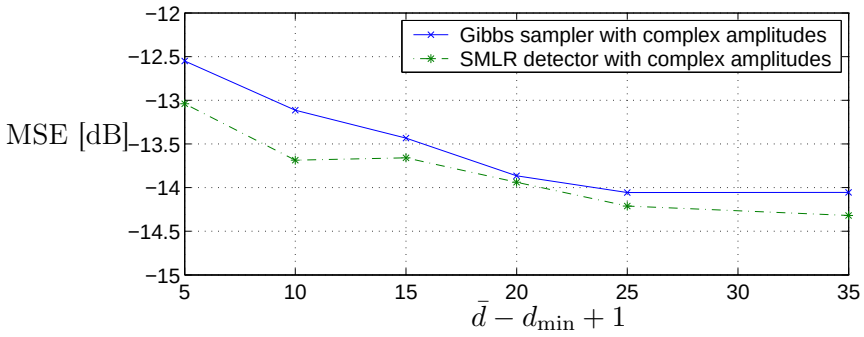


Figure 6.8: MSE for varying average distance.

detection of the pulse positions (Fig. 6.9). At the same time, the MSE, which represents how accurately the signal is approximated, is greatly affected by it (Fig. 6.10). This result can be explained as follows: The detectors basically find the positions where the fringe maxima are located in the signal, which does not depend on the frequency. On the other hand it is plausible that the signal cannot be approximated closely with a wrong set of basic functions. Both effects are direct consequences of introducing a minimum distance: an estimator with an i.i.d. model would achieve much better approximation, while deteriorating the estimate with numerous excess pulses.

In the case of a fringe width mismatch, the influence on estimation results is again quite small (Fig. 6.11 and Fig. 6.12). In analogy to the above considerations, this can be explained by the fact that the deviant fringe width does not prevent the estimator from finding the maximum.

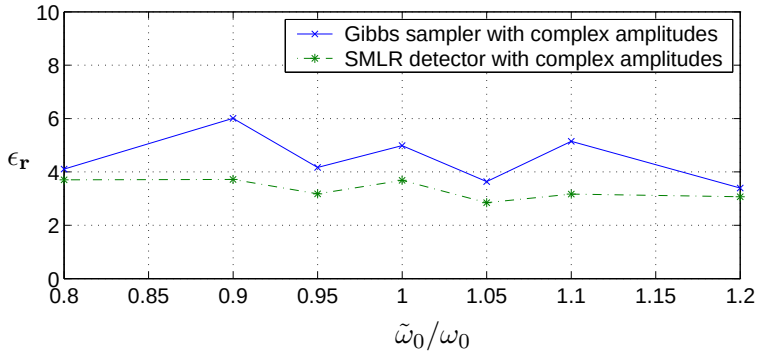


Figure 6.9: Error measure $\epsilon_{\mathbf{r}}$ for varying frequency mismatch.

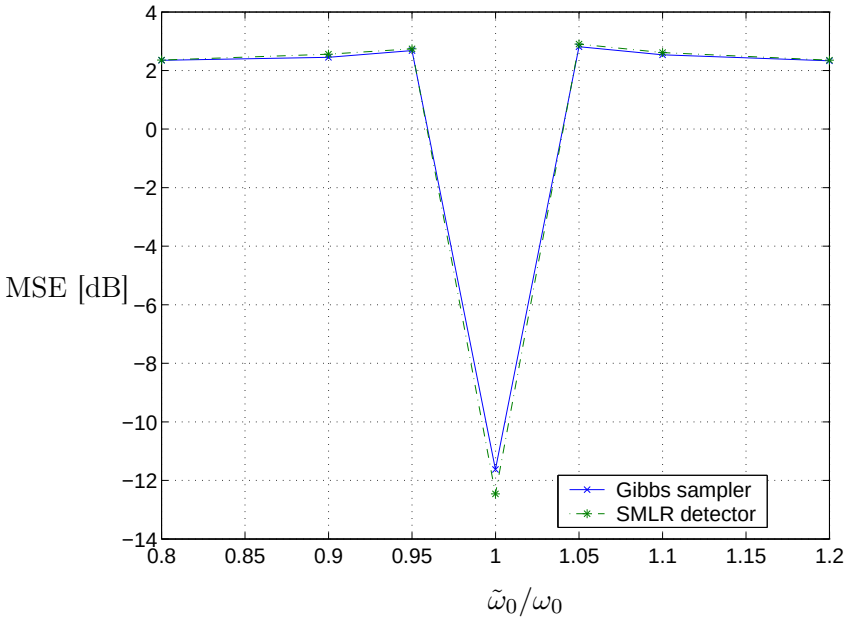


Figure 6.10: MSE for varying frequency mismatch.

6.6 Application to Real OCT Signals

The ultimate goal of developing OCT signal models and estimators is their application to real OCT signals. Fig. 6.13 and Fig. 6.14 show the result of estimating retina functions from such a signal. However, since the true retina function is not known, it is impossible to judge the accuracy of the estimates. The MSE, now redefined as

$$\text{MSE} = \frac{\|\hat{\mathbf{y}} - \mathbf{y}\|^2}{\|\mathbf{y}\|^2},$$

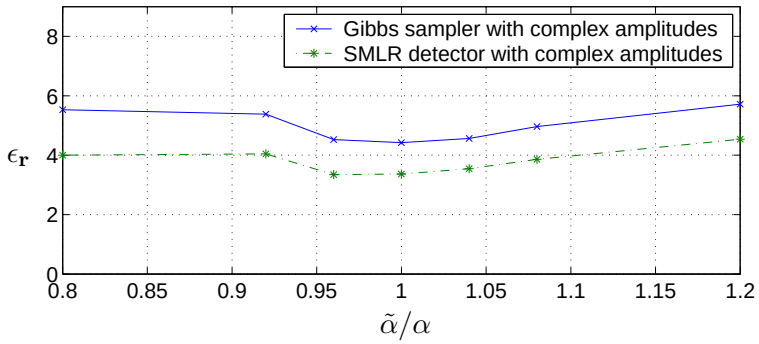


Figure 6.11: Error measure $\epsilon_{\mathbf{r}}$ for varying fringe width mismatch.

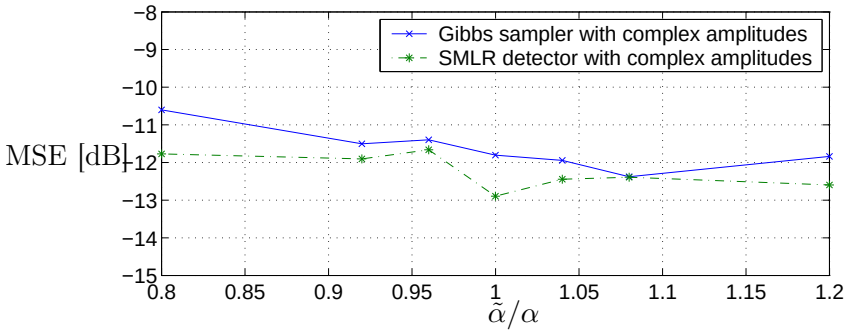
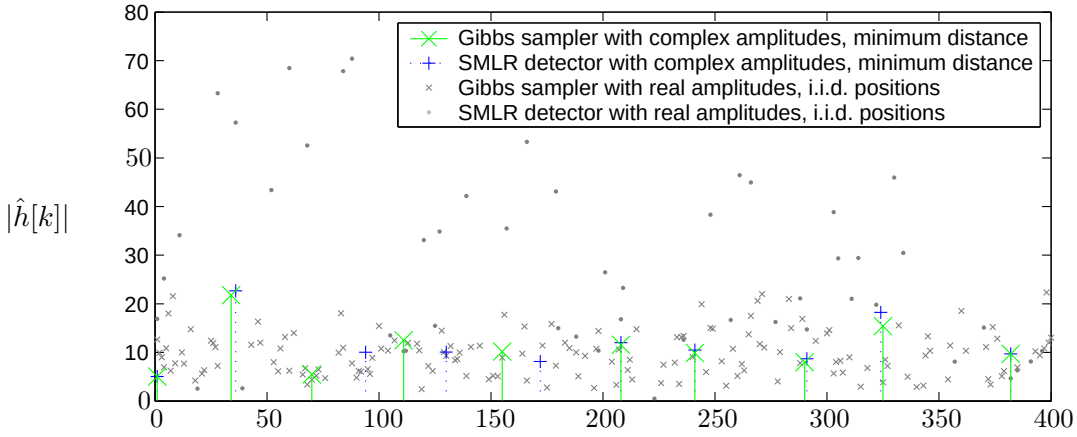


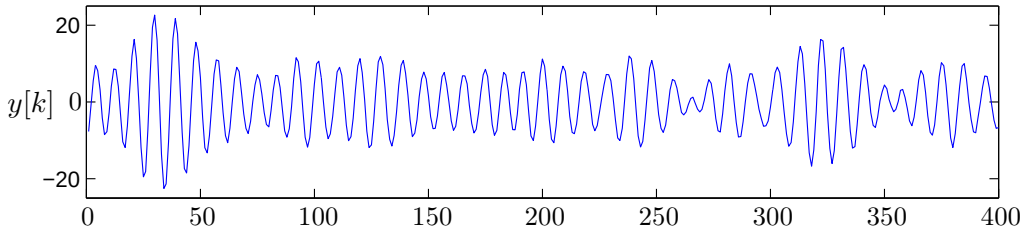
Figure 6.12: MSE for varying fringe width mismatch.

can be used to assess the approximation of the observed signal. However, the relevance of this measure is questionable, as mentioned in section 6.1, which becomes obvious even in the example given in Fig. 6.13. The estimator with the least realistic model achieves the best approximation, while its estimate is useless, with pulses at nearly all possible positions. Fig. 6.14 shows a comparison of the estimates obtained using the models with complex and with real amplitudes.

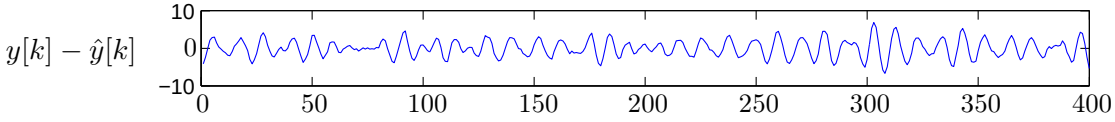
One interesting detail visible in Fig. 6.13 is the influence of the amplitude prior. The two estimators with a minimum distance are forced to approximate the signal by detecting pulse positions at distances that rule out strong interference of the superposed fringes. The i.i.d. estimators try to optimize the approximation by superposing numerous fringes and exploiting constructive and destructive interference. Therefore, the range of the typical estimated amplitude values is basically predefined for the minimum distance estimators, but not for the i.i.d. estimators. The best signal approximation in this example is apparently achieved by estimating amplitudes in a range between 10 and 70, as does the i.i.d. SMLR detector. The i.i.d. Gibbs sampler, on the other hand, constrains itself to values in the range between 0 and 20 as a tribute to the posterior distribution



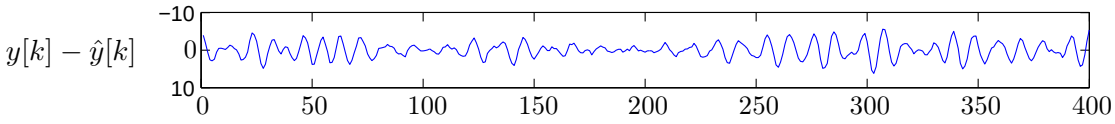
(a) Estimated retina function.



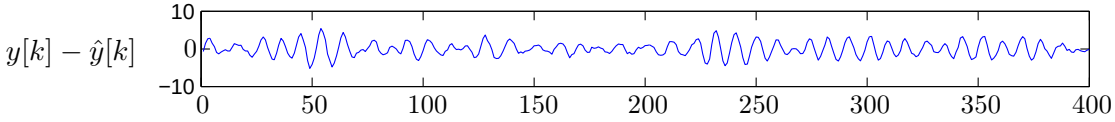
(b) Observed signal.



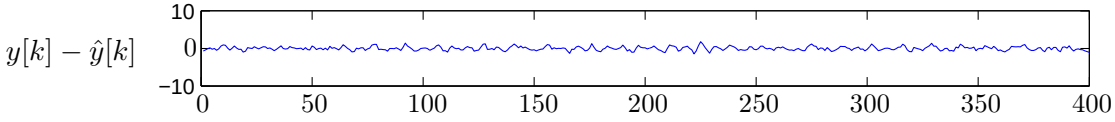
(c) Error signal of the Gibbs sampler with complex amplitudes, minimum distance.



(d) Error signal of the SMLR detector with complex amplitudes, minimum distance.

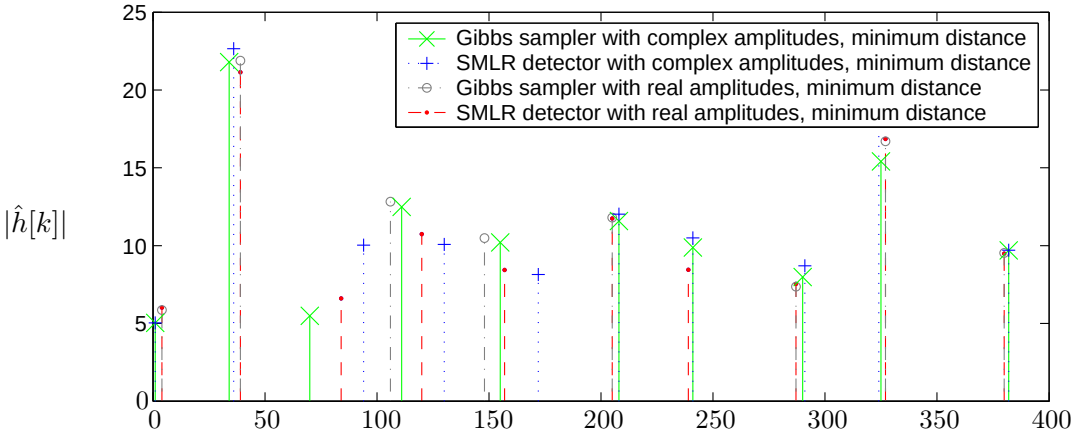


(e) Error signal of the Gibbs sampler with real amplitudes, i.i.d. positions.

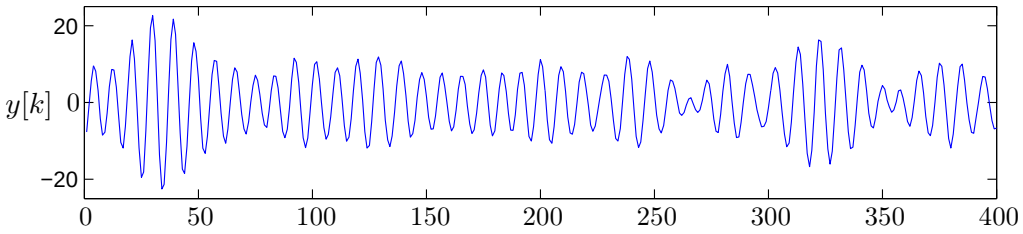


(f) Error signal of the SMLR detector with real amplitudes, i.i.d. positions.

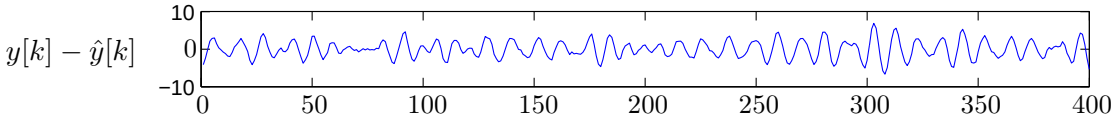
Figure 6.13: Results of different estimators applied to a real OCT signal.



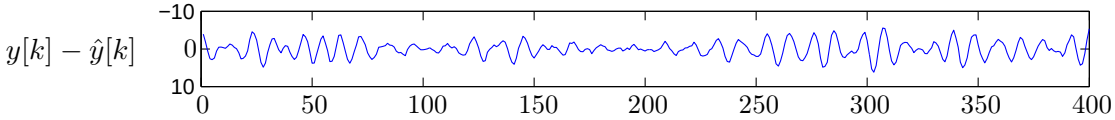
(a) Estimated retina function.



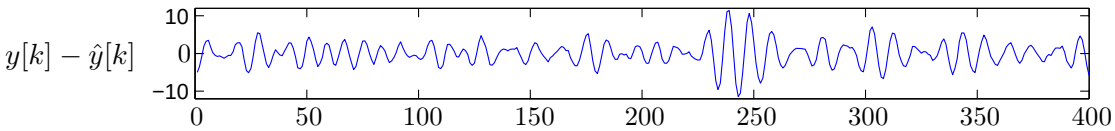
(b) Observed signal.



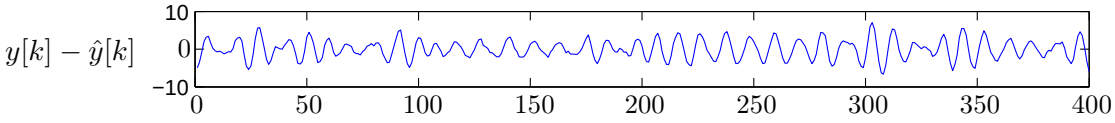
(c) Error signal of the Gibbs sampler with complex amplitudes, minimum distance.



(d) Error signal of the SMLR detector with complex amplitudes, minimum distance.



(e) Error signal of the Gibbs sampler with real amplitudes, minimum distance.



(f) Error signal of the SMLR detector with real amplitudes, minimum distance.

Figure 6.14: Results of different estimators applied to a real OCT signal.

from which the samples are drawn. The SMLR detector uses the posterior distribution to choose one hypothesis in each step, but the hypotheses themselves are generated in a deterministic procedure. The amplitudes are obtained from a LS estimator, which means they optimize the signal approximation but ignore the prior.

While the SNR in this signal appears to be much higher than in our synthetic signals (by well over 10dB, according to visual judgement), the MSE is in fact bigger. A possible explanation is that the fringe parameters estimated from the signal are not perfectly accurate. The estimated fringe parameters are $\alpha = 23$ and $2\pi/\omega_0 = 9$.

The errors in this example are:

	MSE
Gibbs sampler with complex amplitudes, minimum distance	−10.56dB
SMLR detector with complex amplitudes, minimum distance	−11.32dB
Gibbs sampler with real amplitudes, minimum distance	−8.09dB
SMLR detector with real amplitudes, minimum distance	−9.75dB
Gibbs sampler with real amplitudes, i.i.d. positions	−12.43dB
SMLR detector with real amplitudes, i.i.d. positions	−23.50dB

Chapter 7

Outlook

The possibilities offered by the concept of a Bayesian estimation framework suggest some extensions of the method described in this survey, which appear to be promising and feasible at the same time.

Joint estimation of fringe and retina function. In practice, the fringe that corresponds to an OCT signal is not known, as mentioned in section 6.6. It is instead estimated from the signal itself. This has been the most successful approach so far, outperforming efforts of experimental retrieval of the fringe. However, it seems plausible that combining the two separate steps of estimation into a joint estimation would offer some additional gain of performance.

The Bayesian estimation framework is well suited for such an extension. The list of parameters can be augmented by parameters characterizing the fringe, which would each be assigned its own prior. From this, a posterior distribution can be derived and used for sampling.

There are basically two different concepts of characterizing the fringe:

- In the approach chosen in this survey, the shape of the fringe is assumed to be known, and only some shape and scale parameters are adapted to the specific problem. This seems an elegant way of describing the fringe, since it implements all our prior understanding of it and limits the degrees of freedom of the estimator. However, it may be more restrictive than desired. Another problem with this way of modeling is that the fringe parameters, due to their complex relation with the observed signal, generally have posterior distributions that cannot be derived analytically, let alone sampled from. This would require applying another approximation method within the MCMC algorithm.
- An alternative approach is to describe the fringe in a more general way, by means of a series expansion. The series may be chosen such that it is capable of yielding an accurate

approximation of the expected shape of the fringe with a limited number of coefficients. These coefficients can be included in the MCMC algorithm's list of parameters. Their relation with the observed signal is a linear one, avoiding the problem mentioned above.

Laterally adjacent depth scans. Another promising extension seems plausible when considering the structure which our algorithm intends to detect. Layer boundaries can be expected to be at similar depths in adjacent depth scans [19], leading to a strong lateral correlation between neighboring scans. This can be exploited by implementing a joint detection algorithm for a two-dimensional image rather than single depth scans. Again, the Bayesian framework seems inviting for such an extension, since the problem of dimensionality is avoided by the Gibbs sampler. However, more thorough consideration is needed for an appropriate definition of the joint prior, and especially for measures to maintain the efficiency to the algorithm despite correlations (as in section 4.2.3).

Bibliography

[1] C. P. Robert and G. Casella, *Monte Carlo statistical methods.* New York: Springer, 2004.

[2] B. Walsh, "Markov chain Monte Carlo and Gibbs sampling," *Lecture Notes for EEB 581*, Apr. 2004.

[3] J. M. Schmitt, "Optical Coherence Tomography (OCT): A review," *IEEE J. Select. Topics Quantum Electron.*, vol. 5, pp. 1205–1215, July/August 1999.

[4] C. Novak, *Optical Coherence Tomography: Signal Modeling and Processing.* Wien: Institut für Nachrichtentechnik und Hochfrequenztechnik, Vienna University of Technology, 2006.

[5] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Trans. Signal Processing*, vol. 50, pp. 747–758, Mar. 2002.

[6] N. Dobigeon, J.-Y. Tourneret, and M. Davy, "Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach," *IEEE Trans. Signal Processing*, vol. 55, pp. 1251–1263, Apr. 2007.

[7] Q. Cheng, R. Chen, and T.-H. Li, "Simultaneous wavelet estimation and deconvolution of reflection seismic signals," *IEEE Trans. Geoscience and Remote Sensing*, vol. 34, pp. 377–384, Mar. 1996.

[8] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis.* Boca Raton: Chapman and Hall/CRC, 2004.

[9] J. O. Berger, *Statistical decision theory.* New York: Springer, 1980.

[10] C. P. Robert, *The Bayesian Choice.* New York: Springer, 1996.

[11] J. C. Spall, "Estimation via Markov chain Monte Carlo," *IEEE Control Systems Magazine*, pp. 34–45, Apr. 2003.

[12] G. C. Orsak and B. Aazhang, "On the theory of importance sampling applied to the analysis of detection systems," *IEEE Trans. Comm.*, vol. 37, pp. 332–339, Apr. 1989.

[13] W. R. Gilks, *Markov chain Monte Carlo in Practice*. Boca Raton: Chapman and Hall/CRC, 1998.

[14] A. Dubois, L. Vabre, A.-C. Boccara, and E. Beaurepaire, "High-resolution full-field Optical Coherence Tomography with a Linnik microscope," *Appl. Opt.*, vol. 41, pp. 805–812, July/August 2002.

[15] F. L. Pedrotti, *Optik für Ingenieure*. Berlin: Springer, 2005.

[16] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 59, pp. 731–792, 1997.

[17] J. G. Proakis, *Digital Communications*. Boston: McGraw-Hill, 2001.

[18] J. J. Kormylo and J. M. Mendel, "Maximum Likelihood detection and estimation of Bernoulli-Gaussian processes," *IEEE Trans. Inf. Theory*, vol. IT-28, pp. 482–488, May 1982.

[19] D. Koozekanani, K. Boyer, and C. Roberts, "Retinal thickness measurements from Optical Coherence Tomography using a Markov boundary model," *IEEE Trans. Medical Imaging*, vol. 20, pp. 900–916, September 2001.