Aalborg University

Stochastic Channel Modelling

A Bayesian approach using reversible jump Markov chain Monte Carlo methods

by

Rasmus Froberg Brøndum
and Ege Rubak

June 2006

Department of Mathematical Sciences
Fredrik Bajers Vej 7G · 9220 Aalborg Ø · Denmark
Telephone: +45 96 35 88 02 · Fax: +45 98 15 81 29

AALBORG UNIVERSITET

Department of Mathematical Sciences
Fredrik Bajers Vej 7G
9220 Aalborg Ø
Telephone: +45 96 35 88 02
Fax: +45 98 15 81 29
http://www.math.aau.dk

Topic:
Stochastic Channel Modelling:
A Bayesian approach using reversible jump Markov chain Monte Carlo methods

Projectperiod:
January 20th - June 7th, 2006

Projectgroup:
G4-105

Groupmembers:
Rasmus Froberg Brøndum
Ege Rubak

Supervisor:
Martin Bøgsted Hansen

Semester:
MAT6, spring 2006

Number of copies: 6

Report – number of pages: 99

Deadline: June 7th, 2006.

Abstract:

This thesis presents a number of stochastic models and statistical methods for stochastic channel modeling. The purpose of this is to describe the statistical modeling and estimation of an impulse response function for an ultra wide band radio channel, in a rigorous mathematical context. Previous stochastic models are presented and reformulated using the theory of point processes.
The thesis provides brief descriptions of basic likelihood and Bayesian inference. The likelihood section introduces the expectation maximization algorithm, and an approach on how to implement this for inference in the statistical modeling of an impulse response function.
The theory of Markov chain Monte Carlo methods and the reversible jump Markov chain Monte Carlo algorithm is presented, for estimation purposes.
In the second part of the thesis an algorithm, based on the theory from the first part, is constructed, with the purpose of performing statistical inference on measurements of an impulse response function. The purpose is to fit a parametric model of an impulse response function to measured data.

Rasmus Froberg Brøndum

Ege Rubak

Preface

"obscurum per obscurius"

This report is the masters thesis of Rasmus Froberg Brøndum and the project work of Ege Rubak during the spring semester of 2006.

The project was proposed by Troels Pedersen, Department of Communication Technology, Aalborg University, and Martin Bøgsted Hansen, Department of Mathematical Sciences, Aalborg University. We would like to thank Troels Pedersen for supplying the Intel UWB database and answering our questions regarding the contextual part of the problem at hand. Furthermore we thank Martin Bøgsted Hansen for his thorough supervising and many useful ideas and comments through the entire process. We would also like to thank Bjarne Højgaard for supplying us with unpublished material on stochastic integration, and supervision on this part of the thesis. Finally we thank Kim Emil Andersen, for helping us with the implementation of reversible jump Markov Chain Monte Carlo (RJMCMC) methods in the programming language R.

The report gives a method for estimating the channel impulse response function of an UWB system using Bayesian inference. In particular this includes Markov Chain Monte Carlo(MCMC) methods for parameter estimation, which is extended to RJMCMC. This allows the algorithm to estimate the order of the model, thus giving an automated method to determine the number of parameters necessary to properly describe the data.

In order to give a broader approach to the parameter estimation, this thesis also supplies an introduction to the likelihood approach. This method is however not fully implemented.

Contents

Part I

Theory

Chapter 1

Introduction

In many modern applications, such as local area networks and cell phones, there is a rising demand for wireless communications. There is a special interest in ultra wide band (UWB) technologies, because of attractive properties such as high resistance towards interference and higher data rates.

Under certain assumptions for a UWB system it is possible to describe the wireless communication channel via the channel impulse response function. This function determines how the channel effects the input signal and produces an output signal. The impulse response thus plays a key role in the design of communication systems and it is desirable to determine this function. The wireless channel changes depending on the specific location of the antennas, which makes it stochastic and thus lends itself to statistical modeling. A satisfactory statistical model describing the general dynamics of the impulse response of e.g. a typical office environment would be useful in several ways. It would make it possible to simulate a wireless channel for testing equipment, which otherwise can be time demanding, expensive and non-repeatable. Besides that it might give a better understanding of the channel and factors affecting the characteristics of the channel.

In this chapter some general comments on communication systems are given as well as a description of various statistical models proposed in literature for the impulse response function of a wireless communication channel.

1.1 Communication systems

In general a communication system consists of a communication channel, which may be regarded as an operator $\mathscr{S} : \mathbb{R}^{\mathbb{R}} \to \mathbb{R}^{\mathbb{R}}$ which given an input signal x produces an output signal y, i.e. $y = \mathscr{S}(x)$. In many cases, including the one at hand, it is reasonable to model the channel as being linear and time invariant. If we denote the operator of the linear time invariant system by $\mathscr{L}$, this means that

$$\mathscr{L}(c_1 x_1 + c_2 x_2) = c_1 y_1 + c_2 y_2$$
$$\mathscr{L}(x_{t_o}) = y_{t_0}$$

where $\mathscr{L}(x_i) = y_i$ for $i = 1, 2$, $x_{t_0}(t) = x(t - t_0)$ and $y_{t_0}(t) = y(t - t_0)$ for all $t \in \mathbb{R}$.

Under these conditions the communication system can be summed up in a simple mathematical model (Papoulis (1962), chapter 5)

$$y = x * h$$

where h denotes the impulse response function, which is defined as $\mathscr{L}(\delta)$, where δ is the Dirac δ-function. The formal mathematical treatment of the δ-function belongs to the theory of distributions, and an introduction which covers the needed parts in this context is found in Richards and Youn (1990). From the theory it is known that the δ-function is the limit of real C^∞ functions, which covers the way it is used here. The impulse response h is the limiting behavior of the system when the input is a sequence of functions approximating the δ-function. The input and output represent real signals and usually these functions are modeled to have well defined Fourier transforms X and Y. If the Fourier transform of h exists it is called the transfer function or frequency response, and it is denoted H. This leads to the relation

$$Y = X \cdot H.$$

The importance of the impulse response function or equivalently the transfer function of a communication system is obvious, since knowing h or H makes it possible to calculate the output of the system for any given input and vice versa.

1.1.1 Complex baseband representation

The analysis of communication systems, is often done using complex baseband notation, and this concept will therefore be briefly introduced. A communication system often works within a specific range of frequencies, and information is sent in this frequency range when it is transmitted through the channel. Since any transmitted signal x is a real function the Fourier transform X is complex symmetric implying that the support of X is symmetric around zero. The signal is said to be passband around f_c if

$$0 < f_{min} < f_c < f_{max} < \infty, \quad \text{and} \quad \operatorname{supp}(X) \subset (-f_{max}, -f_{min}) \cup (f_{min}, f_{max}).$$

In order to give a unified framework for analyzing passband signals we wish to move the frequency content of the signal from the vicinity of f_c to the vicinity of 0. When the signal is transformed in this way it is called a baseband representation of the signal denoted by $\tilde{x}$. Since the baseband representation is given by a frequency translation it satisfies $\tilde{X}(f) = X(f - f_c)$ for all f. Then $\tilde{x}$ can be found by using the inverse Fourier transform on $\tilde{X}$ and it is given by

$$\tilde{x}(t) = \exp(-2\pi i f_c t)x(t), \quad \text{for } t \in \mathbb{R}.$$

To obtain this result the general translation formula for the inverse Fourier transform is used. Letting g_ω denote a translation of the function g by ω then it holds that $\mathcal{F}^{-1}(g_\omega)(t) = \exp(-2\pi i\omega t)\mathcal{F}^{-1}(g)(t)$ for all $t \in \mathbb{R}$.

In a similar manner it is possible to introduce baseband equivalents of the received signal y and the impulse response h. In the complex baseband context the relation between the input and output is (Haykin, 2001, chap. A2)

$$\tilde{y} = \tfrac{1}{2}\tilde{x} * \tilde{h}.$$

1.1.2 Multipath propagation

In wireless communication a common feature of the channel is multipath propagation (Proakis, 2001, chap. 14). This phenomenon is due to interactions with various objects between the transmitter and receiver. Whenever a time signal x is transmitted through the wireless channel, the expected output is an attenuated and delayed version of the signal

$$y = \sum_{n=0}^{\infty} \beta_n x_{\tau_n},$$

where $\beta_n \in (0, 1)$ is the attenuation factor or path gain of the n'th path, and the signal x_{τ_n} corresponds to a delayed version of the input signal such that $x_{\tau_n}(t) = x(t - \tau_n)$ for all t. This model leads to the impulse response function

$$h = \sum_{n=0}^{\infty} \beta_n \delta_{\tau_n}, \tag{1.1.1}$$

where δ_{τ_n} is the Dirac δ-function translated by τ_n. The corresponding frequency response for a model of this type is given as

$$H(f) = \sum_{n=0}^{\infty} \beta_n \exp(-2\pi i f \tau_n). \tag{1.1.2}$$

It is worth noting that for many practical purposes it is sufficient to model a finite number of signal components, which means that the sum in (1.1.1) is finite. This is due to the fact that the magnitude of the reflections will decay to below the noise level in the measuring equipment.

Since the δ-function causes the Fourier translation factor $\exp(2\pi i f_c t)$ only to be evaluated at the τ_n's the complex baseband equivalent of this impulse response is

$$\tilde{h} = \sum_{n=1}^{\infty} \beta_n \exp(-2\pi i f_c \tau_n)\delta_{\tau_n} = \sum_{n=1}^{\infty} \beta_n \exp(-2\pi i \theta_n)\delta_{\tau_n}.$$

This model is used in much of the literature on channel modeling and has the advantage of being very simple. The complex baseband notation is used in much of the literature and it is thus necessary to be familiar with this concept when working with models within this field, but for our purposes the complex baseband notation will not be necessary, and we will not mention it further in this thesis.

1.2 Stochastic modeling

This section introduces a number of stochastic models introduced in the literature for the impulse response function of a wireless channel.

1.2.1 The Turin model

In Turin et al. (1972) the model in (1.1.1) is used and it is assumed that the arrival times $\{\tau_i - \tau_0\}_{i\in\mathbb{N}}$ form a Poisson process on $\mathbb{R}_+$, when the time of the first arrival τ_0

is given. The Turin model assumes that the attenuation factors are log-normally distributed. The model is not specifically designed for UWB channels, but for wideband urban radio propagation, it is however used as a reference model in the literature. The authors conclude that although the model is "sufficiently refined to be useful, further refinement may be possible by a more elaborate analysis". In particular they notice an inhomogeneity in the arrival rate.

1.2.2 The Saleh-Valenzuela model

In Saleh and Valenzuela (1987) an extension of the Turin model is proposed. Saleh and Valenzuela noticed a temporal clustering in their observed data, and in order to better describe this they proposed an impulse response function given by

$$h = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \beta_{kj} \delta_{(\tau_{kj}+T_k)}, \tag{1.2.1}$$

where $\{T_k - T_0\}_{k\in\mathbb{N}}$ is a Poisson process on $\mathbb{R}_+$, describing the arrivals of the clusters, when the time of the first arrival, T_0, is given and $\{\tau_{kj} - T_k\}_{j\in\mathbb{N}}$, $\quad k = 0, 1, 2, \ldots$ are Poisson sequences on $\mathbb{R}_+$ given the cluster arrivals, describing the arrivals of rays within each cluster. The squares of the path gains $\{\beta_{kj}\}$ are assumed to have exponentially decaying mean, given by

$$\mathbb{E}(\beta_{kj}^2) = \mathbb{E}(\beta_{00}^2) \exp(-\Gamma(T_k - T_0) - \gamma(\tau_{kj} - T_k)), \quad k, j = 0, 1, 2, \ldots \tag{1.2.2}$$

Equation (1.2.2) states that the overall path gain of the clusters is exponentially decaying with a decay rate Γ, as shown by the dashed line in figure 1.2.1, and that the path gains within clusters are exponentially decaying with a decay rate γ as shown by the solid lines. In the original paper by Saleh and Valenzuela cluster arrivals were

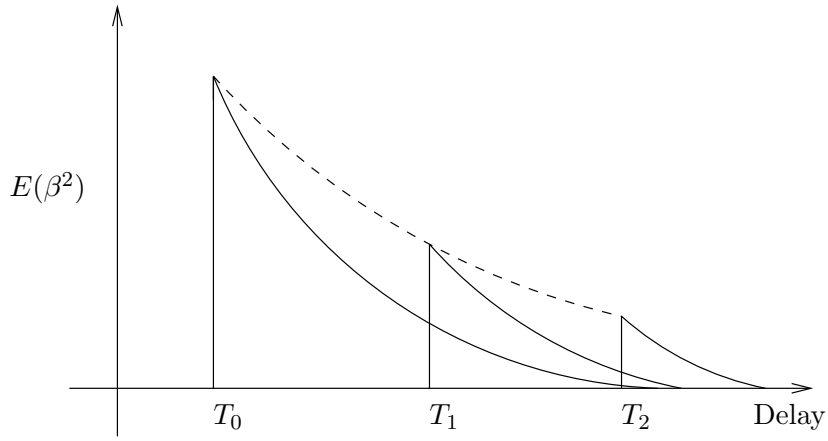


Figure 1.2.1: Illustration of the Saleh-Valenzuela model. The cluster arrivals form a Poisson process with exponentially decaying path gains (dashed line). Within the clusters the ray arrivals form another Poisson process with exponentially decaying path gains.

estimated, by superimposing graphs of data, and manually selecting a mean arrival rate, which in the original paper is found to be 200 ns to 300 ns. Arrivals of rays were estimated using a ray resolving algorithm, which the paper unfortunately does not elaborate on. In a later article by Spencer et al. (2000) which offers an extension of the Saleh Valenzuelah model, where spatial clustering is included, they however describe a ray resolving algorithm witch identifies the most significant arrival, and then subtracts the signal belonging to that arrival. This is done recursively until residuals are below a given noise threshold.

In Saleh and Valenzuela (1987) the ray arrival rate is found to be 5 ns to 10 ns, and it is furthermore found that the normalized path gains $\beta_{kl}^2/\mathbb{E}(\beta_{kl}^2)$ are approximately Rayleigh distributed (see appendix A.4).

The model proposed by Saleh and Valenzuela does not offer the same well-founded physical interpretation as the one proposed by Turin, since there is no clear connection between the environment in which data is collected and the number of clusters. The model does however fit the data well, and is therefore widely used.

1.2.3 The Molisch et al model

In Molisch et al. (2005) a further extension of the Saleh-Valenzuela model is presented. The model has been adopted by the IEEE 802.15.4a working group as a standard model for evaluating proposals of UWB systems. One of the major changes in the Molisch et al. model is the introduction of frequency dependent attenuation factors. The physical interpretation of this, is that different frequencies may have different interactions with objects in the channel. The model furthermore describes the number of cluster arrivals as a Poisson distributed random variable, and allows for a soft unset, so that the first arrival does not in general correspond to the greatest value of β. Finally the arrivals of the rays is extended to a mixture of Poisson processes with different arrival rates.

1.2.4 An autoregressive model

Morrison and Fattouche (1998) proposes an entirely different model for the frequency response of the indoor channel. When looking from the frequency domain, the time domain representation of the signal can be thought of as the spectrum of a time series. Due to the spiked look of the observed impulse responses, Morrison and Fattouche suggest modeling the frequency response as an autoregressive process of order p, i.e.

$$H(f_n) = \sum_{k=1}^{p} \alpha_k H(f_{n-p}) + n(f_n), \quad n = p+1, \ldots, N \tag{1.2.3}$$

where $n(\cdot)$ is a white noise process and f_n is the n'th observation of the frequency response.

The model parameters $(\alpha_1, \ldots, \alpha_k)$ were estimated by minimizing the sum of squares. Inference was done on data from 5 m and 30 m non line of sight scenarios. In both cases the model was found to poorly resemble observed data and the authors conclude that the model is unacceptable for the 30 m case.

1.2.5 Modeling the attenuation factor

The models mentioned above are mostly concerned with modeling the arrival times of the output. Schuster and Bölcskei (2006) provides a comparison between different models for the attenuation factors, using the Akaike Information Criterion. Their measurements were obtained using two different scenarios, with respectively static and dynamic environment, i.e. an environment with moving objects. For the static environment, they find that the best fit, is that of a Rayleigh distribution, closely followed by Nakagami, Weibull and Rice distributions. Tests furthermore show a consistently bad fit for the log-normal distribution. For the dynamic environment the best fit is found to be a combination of Rice and Rayleigh distributions, where the first part of the impulse response has Rician amplitude, and the latter part has Rayleigh distributed amplitude.

1.3 Summary

For physicists and engineers it is natural to treat impulse response functions on the form (1.1.1) which involves a weighted sum of δ-functions with both the weights and locations of the δ-functions being random. In order to make a formal theory for random impulse response functions in this setting we are forced to treat them as a random mapping from some measure space to the space of generalized functions equipped with a suitable σ-field. Although this is in principle possible by the theory of generalized random processes (see e.g. Gelfand and Vilenkin (1964)), we will use the (for statisticians more natural) approach and consider the impulse response as a random measure.

Working within this frame, the goal of this thesis is to do statistical inference on the Turin model, and a new model of our own design. In order to do this, we need to introduce the theory of point processes and random measures, and give an overview of Markov Chain Monte Carlo(MCMC) methods which will be used to do Bayesian inference on data. We furthermore supply a brief description on how the inference might be done in a maximum likelihood setting.

1.4 Overview of the thesis

The thesis is organized in three parts. Part I gives an introduction to the statistical theory which will be used in Part II which covers the actual statistical inference. Part III is the appendix, which contains miscellaneous mathematical results .

Chapter 2 Gives an introduction to the theory of point processes, with an emphasis on the Poisson point process. Basic theorems are proved, and some of the stochastic models for the impulse response given above are described with a random measures representation of point processes on $\mathbb{R}_+^2$.

Chapter 3 Introduces both likelihood and Bayesian inference. The likelihood theory deals with maximum likelihood estimation, and presents the EM algorithm and an example on how it could be used to estimate parameters in e.g. the Turin model. The Bayesian section is a short introduction which covers the needed concepts for the rest

of the thesis.

Chapter 4 Covers Markov Chain Monte Carlo Methods, i.e. numerical methods used in Bayesian inference. The chapter describes the basic theory, the Metropolis-Hastings algorithm and the Reversible Jump MCMC algorithm which is used to determine an unknown number of parameters.

Chapter 5 Contains a description of the dataset used in the thesis, that is how the data is collected, and a descriptive data analysis with a short summary of characteristics of the data.

Chapter 6 Describes how the methods from chapter 4 are used to design algorithms and to do actual statistical inference on data. The methods are implemented on the Turin model, and a new model which deviates from the delta train setting.

Chapter 7 Contains conclusive remarks on the inference, and an outlook on further developments of statistical inference of the problem at hand.

Chapter 2

Point processes

A mathematical framework for treating the problem introduced in this report is the theory of point processes. In this way it is possible to base the mathematics and statistics on a rigorous foundation. The problem at hand can be summarized as modeling a number of arrivals $\tau_1, \tau_2, \ldots$ on the positive real line along with a sequence of positive real numbers $\beta_1, \beta_2, \ldots$ associated with each of the arrivals, that describes the attenuation of the arrivals. The chapter is based on Daley and Vere-Jones (1988) and Møller and Waagepetersen (2004).

2.1 General Point processes

2.1.1 Basic definition

We consider a metric space (S, d) which we for technical reasons assume to be a Polish space (see appendix A.1). For $x \subseteq S$ let $n(x)$ denote the cardinality of x, and let $x_B = x \cap B$ for $B \subseteq S$. The space of locally finite subsets of S is given by

$$N = \{x \subseteq S \mid n(x_B) < \infty \text{ for all } B \in \mathbb{B}_0(S)\},$$

where $\mathbb{B}_0(S)$ denotes the class of bounded Borel sets in S.

In order to define a point process, a σ-algebra on N is needed, and here the following is used

$$\mathcal{N} = \sigma\big(\{x \in N \mid n(x_B) = m\} \mid B \in \mathbb{B}_0(S), m \in \mathbb{N}_0\big). \tag{2.1.1}$$

Definition 2.1.1
A point process $X : \Omega \to N$ is a measurable mapping defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ taking values in the measurable space $(N, \mathcal{N})$. The probability distribution P_X on $(N, \mathcal{N})$ induced by X is called the distribution of the point process X.
Unless otherwise stated the probability measure $\mathcal{P}$ is assumed to be complete (definition A.2.2). □

The existence of such a measurable mapping is discussed in section 2.1.2.

The mean outcome of a point process, is determined by its intensity measure.

Definition 2.1.2 (Intensity measure)
The intensity measure of a point process X is a measure μ defined on $\mathbb{B}(S)$, such that $\mu(B) = \mathbb{E}(n(X_B))$. In some cases this measure has a density with respect to some

measure ν on S, i.e. a function ρ such that

$$\mu(B) = \int_B \rho(\xi) d\nu(\xi), \text{ for all } B \in \mathbb{B}(S).$$

This density is called the intensity function (with respect to ν). □

For a point process we define the class of void events by

$$N^0 = \{\{x \in N \mid n(x_B) = 0\} \mid B \in \mathbb{B}_0(S)\}. \tag{2.1.2}$$

Lemma 2.1.3
The class of void events is a generating class for $\mathcal{N}$, i.e.

$$\mathcal{N} = \sigma(N^0). \tag{2.1.3}$$

PROOF:
Since $N^0 \subseteq N$, it is obvious that $\sigma(N^0) \subseteq \mathcal{N}$. We thus need to prove that $\{x \in \mathcal{N} \mid n(x_B) = m\} \in \sigma(N^0)$ for all $B \in \mathcal{B}_0(S)$ and all $m \in \mathbb{N}$.

According to proposition A.1.4 S contains a dissecting system, i.e. a sequence $\mathcal{A}_n = \{A_{n,i} \in \mathbb{B}(S) \mid i = 1, \ldots, k_n\}$, $n \in \mathbb{N}$ of partitions of S, such that S is a disjoint union of the sets in each $\mathcal{A}_n$, and with the property that $A_{n-1,i} \cap A_{n,i}$ is either $\emptyset$ or $A_{n,i}$. As a consequence we have that for separate points $\xi, \eta \in S$, $\xi \in A_{n,i}$ implies $\eta \notin A_{n,i}$ for sufficiently large n.

This means that $n(x_B) = m$ if and only if there exists an $n_0 \in \mathbb{N}$ with $k_{n_0} \geq m$ such that for all $n > n_0$ there exists $\{j_1, \ldots, j_m\} \subseteq \{1, 2, \ldots, k_n\}$ such that $n(x \cap B \cap A_{n,j_i}) > 0, \quad i = 1, \ldots, m$ and $n(x \cap (B \setminus \cup_{i=1}^m A_{n,j_i})) = 0$. We then have

$$\{x \in \mathcal{N} \mid n(x_B) = m\} = \bigcup_{n_0 | k_{n_0} \geq m} \bigcap_{n \geq n_0} \bigcup_{j_i, i=1,\ldots,k_n} A(n, m, B, \{j_1, \ldots, j_m\},$$

where each of the events

$$A(n, m, B, \{j_1, \ldots, j_m\}) = \bigcap_{i=1}^m \{x \in \mathcal{N} \mid n(x \cap B \cap A_{n,j_i}) > 0\} \bigcap \{x \in \mathcal{N} \mid n(x \cap (B \setminus \cup_{i=1}^m A_{n,j_i})) = 0\}$$

belongs to $\sigma(N^0)$, since $\{x \in \mathcal{N} \mid n(x \cap (B \setminus \cup_{i=1}^m A_{n,j_i})) = 0\} \in N^0$ and $\{x \in \mathcal{N} \mid n(x \cap B \cap A_{n,j_i}) > 0\} \in \sigma(N^0)$ as the complementary set of an event from N^0. As a consequence $\{x \in \mathcal{N} \mid n(x_B) = m\} \in \sigma(\mathcal{N}^0)$. ■

Using the result above, we now have the following theorem for point processes.

Theorem 2.1.4
A point process is uniquely determined by its void probabilities.

PROOF:
By lemma 2.1.3 the set of void events is a generating set for $\mathcal{N}$. Since N^0 is closed under intersection it then follows from lemma A.2.1, that the probability measure for the configurations of a point process is uniquely determined by its values on N^0. ■

2.1.2 Existence of point processes

To show the existence of point processes we turn briefly to the theory of random measures.

Let $\mathscr{M}_S$ denote the space of all locally finite Borel measures on the polish space S, and let the mapping $T_A : \mathscr{M}_S \to \mathbb{R}_+$ be given by $\mu \mapsto \mu(A)$ for all $A \in \mathbb{B}_0(S)$. Then the Borel σ-algebra on $\mathscr{M}_S$ is given by

$$\mathbb{B}(\mathscr{M}_S) = \sigma\big(\{T_A^{-1}(B) \mid B \in \mathbb{B}(\mathbb{R}_+), A \in \mathbb{B}_0(S)\}\big),$$

which is the smallest σ-algebra such that all T_A are $\big(\mathbb{B}(\mathscr{M}_S), \mathbb{B}(\mathbb{R}_+)\big)$ measurable.

Definition 2.1.5 (Random Measure)
A random measure ξ with phase or state space S, is a measurable mapping from a probability space $(\Omega, \mathcal{F}, P)$ into $(\mathscr{M}_S, \mathbb{B}(\mathscr{M}_S))$. The distribution of a random measure is the probability measure it induces on $(\mathscr{M}_S, \mathbb{B}(\mathscr{M}_S))$. □

Thus for each event $\omega \in \Omega$, we have a measure $\xi(\cdot, \omega)$. We may also consider the function $\xi(A, \cdot)$, which maps Ω to $\mathbb{R}_+$ for each fixed $A \in \mathbb{B}_0(S)$. A general result for random measures states that $\xi(\cdot, \omega)$ is a random measure if and only if $\xi(A, \cdot)$ is a random variable, for each $A \in \mathbb{B}_0(S)$. Another way of regarding a random measure, is thus as a family of random variables, indexed by the Borel sets of S satisfying the properties of a measure. A special type of random measures are the random counting measures

Definition 2.1.6
Let S be a Polish space. The space of counting measures on S, $\mathscr{N}_S$, consists of all boundedly finite, integer-valued measures ν defined on $\mathbb{B}(S)$. A counting measure on S is called simple if

$$\nu(\{x\}) = 0 \text{ or } \nu(\{x\}) = 1 \quad \text{for all } x \in S.$$

The space of simple counting measures on S is denoted $\mathscr{N}_S^*$. □

For the space of counting measures we similarly define a mapping, $S_A : \mathscr{N}_S \to \mathbb{N}_0$ given by $\nu \to \nu(A)$ for all $A \in \mathbb{B}_0(S)$, and let the Borel σ-algebra on $\mathscr{N}_S$ be given by

$$\mathbb{B}(\mathscr{N}_S) = \sigma\left(\{S_A^{-1}(B) \mid B \in \mathbb{B}(\mathbb{N}_0), A \in \mathbb{B}_0(S)\}\right). \tag{2.1.4}$$

In proposition 7.1.II in Daley and Vere-Jones (1988) it is shown that $N \in \mathscr{N}_S$ if and only if it is expressible as

$$N(A) = \sum_{i=1}^{\infty} k_i \mathbf{I}(x_i \in A), \quad A \in \mathbb{B}(S),$$

where each k_i is a positive integer, and $\mathbf{I}(\cdot \in A)$ is an indicator function for the set A. With this at hand it is possible to give an alternative definition of a point process

Definition 2.1.7
Let S be a Polish space. A point process on S is a random counting measure on S, i.e. a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ into $(\mathscr{N}_S, \mathbb{B}(\mathscr{N}_S))$. □

Clearly there is an analogy to definition 2.1.1, but it might seem less intuitive to think of a point process as a measure. The two definitions are equivalent and the advantage of the latter is that existence properties of random measures are well studied and these can thus be used to ensure the existence of point processes. To investigate the connection between the two definitions consider the following mapping from $(N, \mathcal{N})$ to $(\mathcal{N}_S, \mathbb{B}(\mathcal{N}_S))$

$$f : \{x_1, x_2, \dots\} \mapsto \sum_{i=1}^{\infty} \mathbf{I}_{x_i}$$

and correspondingly the inverse mapping from $(\mathcal{N}_S, \mathbb{B}(\mathcal{N}_S))$ to $(N, \mathcal{N})$ given by

$$g : \sum_{i=1}^{\infty} \mathbf{I}_{x_i} \mapsto \{x_1, x_2, \dots\},$$

where $\mathbf{I}_{x_i}(A) = \mathbf{I}(x_i \in A)$ for all $A \in \mathbb{B}(S)$.

Proposition 2.1.8
The mappings $f : N \to \mathcal{N}_s$ and $g : \mathcal{N}_S \to N$ are measurable.

Proof:
The proof follows as a consequence of proposition A.2.5.

(i) First we consider the mapping g. Since $\mathcal{N} = \sigma(N^0)$ it suffices to show that for any set $C \in N^0$ it holds that $g^{-1}(C) \in \mathbb{B}(\mathcal{N}_S)$. Since $C \in N^0$ there exists a set $B \in \mathbb{B}_0(S)$ such that

$$C = \{x \subset S \mid n(x \cap B) = 0\}.$$

Then $g^{-1}(C)$ consists of all measures of the form $\sum \mathbf{I}_{x_i}$, where $x \cap B = \emptyset$, i.e.

$$\begin{aligned} g^{-1}(C) &= \{\mu = \sum \mathbf{I}_{x_i} \mid x \cap B = \emptyset\} \\ &= \{\mu \in \mathcal{N}_S \mid \mu(B) = 0\} \\ &= \{\mu \in \mathcal{N}_S \mid S_B(\mu) = 0\} \\ &= S_B^{-1}(0). \end{aligned}$$

This is an element of $\mathbb{B}(\mathcal{N}_S)$ since $B \in \mathbb{B}_0(S)$ and $0 \in \mathbb{B}(\mathbb{N}_0)$.

(ii) We now consider the mapping f. Using (2.1.4) we let $A \in \mathbb{B}_0(S)$ and $B \in \mathbb{B}(\mathbb{N}_0)$, and set

$$C' = S_A^{-1}(B) = \{\nu \in \mathcal{N}_S \mid \nu(A) \in B\}.$$

We then have

$$\begin{aligned} f^{-1}(C') &= \{x \in N \mid n(x \cap A) \in B\} \\ &= \bigcup_{b_i \in B} \{x \in N \mid n(x \cap A) = b_i\}. \end{aligned} \tag{2.1.5}$$

The last equation follows since $\mathbb{B}(\mathbb{N}_0) = 2^{\mathbb{N}_0}$, i.e. the power set of $\mathbb{N}_0$, and every element is thus a countable union of natural numbers. The set (2.1.5) is an element of $\mathcal{N}$, since it is a countable union of elements from the generator given by (2.1.1).

■

Since these mappings are measurable a point process which is a measurable mapping from $(\Omega, \mathcal{F}, \mathcal{P})$ into either $(\mathcal{N}_S, \mathbb{B}(\mathcal{N}_S))$ or $(N, \mathcal{N})$ can thus be composed with one of these to obtain a measurable mapping to the other space. This ensures that which ever definition we use leads to the same results.

The most common way of assessing the properties of the distribution of a random measure, is through its finite dimensional distributions.

Definition 2.1.9
The finite dimensional distributions of a random measure ξ are the joint distributions, for all finite families of bounded Borel sets $A_1, \ldots, A_k$ of the random variables $\xi(A_1), \ldots, \xi(A_k)$, that is the family of proper distribution functions.

$$F_k(A_1, \ldots, A_k; x_1, \ldots, x_k) = P(\xi(A_i) \leq x_i, i = 1, \ldots, k). \tag{2.1.6}$$

□

It can in fact be proven that the distribution of a random measure is completely determined by its finite dimensional distributions. The existence of a random measure can be ensured by a number of necessary and sufficient conditions. The following theorem is a special case of an existence theorem for random measures, aimed at point processes.

Theorem 2.1.10 (Existence Theorem For Point Processes)
In order that a family $P_k(A_1, \ldots, A_k; n_1, \ldots, n_k)$ of discrete finite dimensional distributions defined on bounded Borel sets be the finite dimensional distributions of a point process, it is necessary and sufficient that

(i) For any permutation $i_1, \ldots, i_k$ of the indices $1, \ldots, k$,

$$P_k(A_1, \ldots, A_k; n_1, \ldots, n_k) = P_k(A_{i_1}, \ldots, A_{i_k}; n_{i_1}, \ldots, n_{i_k}).$$

(ii) $\sum_{r=0}^{\infty} P_k(A_1, \ldots, A_k; n_1, \ldots, n_{k-1}, r) = P_{k-1}(A_1, \ldots, A_{k-1}; n_1, \ldots, n_{k-1})$.

(iii) For each disjoint pair $A_1, A_2 \in \mathbb{B}_0(S)$, $P_3(A_1, A_2, A_1 \cup A_2; n_1, n_2, n_3)$ has zero mass outside the set where $n_1 + n_2 = n_3$.

(iv) For sequences $\{A_n\}$ of bounded Borel sets with $A_n \downarrow \emptyset$, $P_1(A_n, 0) \to 1$.

The first two conditions, ensure the existence of the distribution as a consequence of the Kolmogorov extension theorem, th. A.1.2, and the two latter conditions ensure that we are in fact dealing with a measure. This ensures that if we specify finite dimensional distributions satisfying the conditions in theorem 2.1.10 then there exists a measurable mapping from the probability space to $(\mathcal{N}_S, \mathbb{B}(\mathcal{N}_S))$ which induces a probability distribution with the given finite dimensional distributions. As discussed earlier this measurable mapping can be extended to a measurable mapping to the measure space $(N, \mathcal{N}, P_X)$.

2.1.3 Poisson point process

The Poisson point process is a fundamental point process, that serves as the basis of other more complex types of point processes. We only consider the case of a Poisson point process on $S \subseteq \mathbb{R}^n$ with a locally integrable intensity function $\rho : S \to [0, \infty)$ since it covers the problem at hand and avoids some measure theoretic details. Local integrability means, that $\int_B \rho(\xi) d\xi < \infty$ for all $B \in \mathbb{B}_0(S)$.
The definition of a binomial point process is needed to define the Poisson point process later.

Definition 2.1.11
Let f be a density on a set $B \in \mathbb{B}_0(S)$ and let $n \in \mathbb{N}$. A point process X consisting of n independent identically distributed points with density f is called a binomial point process of n points in B with density f. This is noted $X \sim \text{binomial}(B, n, f)$. □

Definition 2.1.12
A Poisson point process with intensity function ρ is a point process X on S, which satisfies, that

1. for any $B \in \mathbb{B}(S)$ with $\mu(B) < \infty$, $n(X_B) \sim \text{Poisson}(\mu(B))$, where Poisson(0) is taken as a distribution with all probability mass at 0.

2. for any $n \in \mathbb{N}$ and $B \in \mathbb{B}(S)$ with $0 < \mu(B) < \infty$, given $n(X_B) = n$, $X_B \sim \text{binomial}(B, n, f)$ with $f(\xi) = \rho(\xi)/\mu(B)$.

□

In the special case, where ρ is a constant Poisson(S, ρ) is referred to as a homogeneous Poisson Point process on S with rate or intensity ρ. Poisson$(S, 1)$ is called the standard or unit rate Poisson point process.

Proposition 2.1.13
(i) $X \sim \text{Poisson}(S, \rho)$ if and only if for all $B \in \mathbb{B}(S)$ with $\mu(B) < \infty$ and all $F \in \mathcal{N}$,

$$P(X_B \in F) = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B 1_F(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \rho(x_i) dx_1 \cdots dx_n.$$

(ii) If $X \sim \text{Poisson}(S, \rho)$, then for measurable functions $h : N \to [0, \infty)$ and $B \in \mathbb{B}(S)$ with $\mu(B) < \infty$,

$$\mathbb{E}(h(X_B)) = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B h(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \rho(x_i) dx_1 \cdots dx_n.$$

PROOF:
Starting with part (i). Let $X \sim \text{Poisson}(S, \rho)$ so that definition 2.1.12 holds for X.

Then for $B \in \mathbb{B}(S)$ with $\mu(B) < \infty$ and $F \in \mathcal{N}$,

$$\begin{aligned} P(X_B \in F) &= \sum_{n=0}^{\infty} P(n(X_B) = n) P(X_B \in F | n(X_B) = n) \\ &= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))\mu(B)^n}{n!} \int_B \cdots \int_B 1_F(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \frac{\rho(x_i)}{\mu(B)} dx_1 \cdots dx_n \\ &= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B 1_F(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \rho(x_i) dx_1 \cdots dx_n. \end{aligned}$$

To prove part (ii) the monotone class argument measure theory is used. If h is an indicator function for an $F \in \mathcal{N}$, then

$$\mathbb{E}(h(X_B)) = \mathbb{E}(1_F(X_B)) = P(X_B \in F),$$

and the result is given by (i). By linearity this extends to simple functions, that is functions on the form $f = \sum_i a_i \mathbf{I}_{A_i}$, where $a_i \in \mathbb{R}$ and $\mathbf{I}_{A_i}$ is an indicator function for a set A_i. The result then follows by the monotone convergence theorem. ■

Theorem 2.1.14
The Poisson point process $X \sim \text{Poisson}(S, \rho)$ exists and is uniquely determined by its void probabilities

$$v(B) = \exp(-\mu(B)), \quad \text{for all bounded } B \in \mathbb{B}(S)$$

PROOF:
Let $\xi \in S$ be an arbitrary point and set $B_i = \{\eta \in S \mid i - 1 \leq \|\eta - \xi\| < i\}$ for $i \in \mathbb{N}$. It follows that S is a disjoint union of the bounded B_i. Let $X = \cup_1^\infty X_i$ where $X_i \sim \text{Poisson}(B_i, \rho_i), \quad i = 1, 2, \ldots$, are independent, and where ρ_i is the restriction of ρ to B_i. Then for bounded $B \subseteq S$,

$$\begin{aligned} P(X \cap B = \emptyset) &= \prod_{i=1}^{\infty} P(X_i \cap B = \emptyset) = \prod_{i=1}^{\infty} \exp(-\mu(B \cap B_i)) \\ &= \exp\left(-\sum_{i=1}^{\infty} \mu(B \cap B_i)\right) = \exp(-\mu(B)), \end{aligned}$$

which is the void probability for a Poisson process with intensity measure μ. The existence and uniqueness now follow from theorem 2.1.4. ■

Proposition 2.1.15
If X is a Poisson process on S, then $X_{B_1}, X_{B_2}, \ldots$ are independent for disjoint sets $B_1, B_2, \ldots \subseteq S$

A proof of this proposition may be found in Møller and Waagepetersen (2004).

For the one-dimensional Poisson point process, we have the following very useful result (Stirzaker (2003)), which gives a distribution on the inter arrival times.

Proposition 2.1.16
Let $\{\xi_1, \xi_2, \ldots\} \sim \text{Poisson}(\mathbb{R}^+, \lambda)$ for $\lambda \in \mathbb{R}^+$, $\xi_0 = 0$ and $T_i = \xi_{i+1} - \xi_i$, then

(i) $T_i \sim \exp(\lambda), \quad i = 1, 2, \ldots$

(ii) T_i is independent of T_j for $i \neq j$

(iii) $\xi_{i+m} - \xi_i \sim \text{Gamma}(m, \lambda), \quad i = 1, 2, \ldots, \quad m \in \mathbb{N}$

PROOF:
The T_i's correspond to times between arrivals in the Poisson process. This means that

$$P(T_1 > t) = P(n(0, t) = 0) = \exp(-\lambda t),$$

which means that T_1 has the exponential distribution. Now conditional on T_1

$$p(T_2 > t | T_1 = t_1) = P(n(t_1, t_1 + t) = 0 \mid T_1 = t_1).$$

By proposition 2.1.15 events in disjoint sets are independent. This gives

$$= P(n(t_1, t_1 + t) = 0) = \exp(-\lambda t).$$

Similarly we have

$$P(T_{n+1} > t \mid T_1 = t_1, \ldots, T_n = t_n) = P(n(t_n, t_n + t) = 0) = \exp(-\lambda t).$$

Part (iii) follows since the convolution of m independent $\exp(\lambda)$ distributions is a $\text{Gamma}(m, \lambda)$ distribution. ■

It may in fact be proven that exponentially distributed inter arrival times is equivalent with a Poisson point process, and in certain cases this is used to characterize the Poisson process (see eg. Billingsley (1995)).

Proposition 2.1.16 is now used to derive a another property of the Poisson process (Stirzaker (2003)) which will be used to simplify some calculations in part II

Proposition 2.1.17
Let $\xi_1, \xi_2, \ldots$ be arrival times in a Poisson process $X \sim \text{Poisson}(\mathbb{R}_+, \lambda)$, $\xi_0 = 0$, then

$$f_{\Xi_1, \ldots, \Xi_k | n(X_{[0,t]}) = k}(\xi_1, \ldots, \xi_k) = \frac{k!}{t^k}, \quad 0 < \xi_1 < \cdots < \xi_k < t, \tag{2.1.7}$$

which is the density function for k uniformly distributed order variables.

PROOF:
By proposition 2.1.16

$$T_i = \xi_i - \xi_{i-1}, \quad i = 1, 2, \ldots,$$

are independent exponentially distributed random variables, which means that

$$f_{T_1, \ldots, T_{k+1}}(t_1, \ldots, t_{k+1}) = \lambda^{k+1} \exp\left(-\lambda \sum_{i=1}^{k+1} t_i\right).$$

By the change of variable formula Stirzaker (2003) the linear transformation

$$\Xi_i = \sum_{j=1}^{i} T_j, \quad i = 1, \ldots, k+1$$

may be used to show that

$$f_{\Xi_1,\ldots,\Xi_{k1+}}(\xi_1,\ldots,\xi_{k+1}) = \lambda^{k+1}\exp(-\lambda\xi_{k+1}).$$

This then yields

$$\begin{aligned}
&P(0<\Xi_1<\xi_1<\Xi_2<\cdots<\Xi_k<\xi_k, n(X_{[0,t]}=k))\\
&= P(0<\Xi_1<\xi_1<\Xi_2<\cdots<\Xi_k<\xi_k<t<\Xi_{k+1})\\
&= \int_0^{\xi_1}\cdots\int_t^{\infty} f_{\Xi_1,\ldots,\Xi_{k1+}}(\xi_1,\ldots,\xi_{k+1})d\xi_{k+1}\cdots d\xi_1\\
&= \lambda^k\xi_1(\xi_2-\xi_1)\cdots(\xi_k-\xi_{k-1})\exp(-\lambda t).
\end{aligned}$$

The conditional distribution is then

$$\begin{aligned}
P(\Xi_1<\xi_1<\cdots<\Xi_k<\xi_k|n(X_{[0,t]}=k))&\\
&= \frac{\lambda^k\xi_1(\xi_2-\xi_1)\cdots(\xi_k-\xi_{k-1})\exp(-\lambda t)}{t^k(k!)^{-1}\exp(-\lambda t)}\\
&= \xi_1(\xi_2-\xi_1)\cdots(\xi_k-\xi_{k-1})\frac{k!}{t^k}.
\end{aligned}$$

The density given by (2.1.7) is then obtained as

$$\frac{\partial}{\partial\xi_1\ldots\partial\xi_k}\left[\xi_1(\xi_2-\xi_1)\cdots(\xi_k-\xi_{k-1})\frac{k!}{t^k}\right] = \frac{k!}{t^k}.$$

■

The following standard result (see e.g. Kingman (2006)) is used to show that the superposition of Poisson point processes is a Poisson process.

Lemma 2.1.18 (Disjointness Lemma)
Let $X_i \sim \text{Poisson}(S,\rho_i)$ with $\mu_i = \int_S \rho_i(\xi)d\xi < \infty$, $i = 1,2$. If μ_1 and μ_2 both have the bisection property (definition A.2.3) then X_1 and X_2 are disjoint with probability 1:

$$P(X_1 \cap X_2 = \emptyset) = 1.$$

PROOF:
Let $n = 2^k$ be any power of 2. By the bisection property there exists a disjoint union $S = \bigcup_{i=1}^n S_i$ of measurable sets S_i with $\mu_1(S_i) = n^{-1}\mu_1(S)$, $i = 1,\ldots,n$. For each S_i the bisection property again ensures that there exists a disjoint union $S_i = \bigcup_{j=1}^n S_{ij}$ of measurable sets S_{ij} with $\mu_2(S_{ij}) = n^{-1}\mu_2(S_i)$, $j = 1,\ldots,n$. Define

$$E_n = \bigcup_{i,j=1}^{n}\{\omega\in\Omega \mid n(X_1\cap S_{ij})\geq 1, n(X_2\cap S_{ij})\geq 1\}$$

The event that the superposition of X_1 and X_2 is not disjoint is a subset of E_n:

$$\{\omega\in\Omega \mid X_1\cap X_2\neq\emptyset\}\subseteq E_n.$$

Since the cardinality function $n(\cdot)$ is a measurable mapping then E_n is measurable and

$$\begin{aligned} P(E_n) &\leq \sum_{i,j=1}^{n} P(n(X_1 \cap S_{ij}) \geq 1, n(X_2 \cap S_{ij}) \geq 1) \\ &= \sum_{i,j=1}^{n} P(n(X_1 \cap S_{ij}) \geq 1) P(n(X_2 \cap S_{ij}) \geq 1). \end{aligned}$$

Using that for any point process X with intensity measure μ and measurable $A \in S$ it holds that $P(n(X \cap A) \geq 1) \leq P(n(X \cap A) = 1) + P(n(X \cap A) = 2) + \cdots \leq P(n(X \cap A) = 1) + 2P(n(X \cap A) = 2) + \cdots = \mu(A)$ we get

$$\begin{aligned} &\leq \sum_{i,j=1}^{n} \mu_1(S_{ij})\mu_2(S_{ij}) \\ &= \tfrac{1}{n} \sum_{i,j=1}^{n} \mu_1(S_{ij})\mu_2(S_i) \\ &= \tfrac{1}{n} \sum_{i=1}^{n} \mu_1(S_i)\mu_2(S_i) \\ &= \tfrac{1}{n^2} \sum_{i=1}^{n} \mu_1(S)\mu_2(S_i) \\ &= \tfrac{1}{n^2} \mu_1(S)\mu_2(S). \end{aligned}$$

Since each E_n is measurable the intersection $\bigcap_{n=1}^{\infty} E_n$ is measurable and it is seen that

$$P(\bigcap_{n=1}^{\infty} E_n) = 0.$$

Since P is a complete probability measure $\{\omega \in \Omega \mid X_1 \cap X_2 \neq \emptyset\} \subseteq E_n$ is measurable and $P(\{\omega \in \Omega \mid X_1 \cap X_2 \neq \emptyset\}) = 0$, which proves that the union is disjoint with probability one. ■

Proposition 2.1.19
If $X_i \sim \text{Poisson}(S, \rho_i)$, $i = 1, 2, \ldots$ are mutually independent and $\rho = \sum \rho_i$ is locally integrable, then with probability one, $X = \bigcup_i X_i$ is a disjoint union and $X \sim \text{Poisson}(S, \rho)$.

PROOF:
By considering a bounded ball in S the conditions in the disjointness lemma are fulfilled and by induction the countable union is disjoint with probability one. This expands to all of S and we need only to verify the last part of the proposition. Let $B \in \mathbb{B}_0(S)$ then

$$P(X_B = \emptyset) = \prod_{i=1}^{\infty} P(X_i \cap B = \emptyset) = \prod_{i=1}^{\infty} \exp(-\mu_i(B)) = \exp(-\mu(B)),$$

and the result follows from theorem 2.1.14. ■

2.1.4 Cox processes

Often a Poisson point process is not adequate to model a given problem and more complex models are needed. In many cases Cox processes turn out to be more appropriate than the Poisson case. The extension to Cox processes consists of considering the intensity function of the Poisson process as a stochastic process.

Definition 2.1.20
Suppose that Z is a nonnegative stochastic process on S, so that with probability one, $\xi \mapsto Z(\xi)$ is a locally integrable function. If the conditional distribution of X given Z is a Poisson process on S with intensity function Z, then X is said to be a Cox process driven by by Z. □

2.1.5 Marked point processes

In some cases it can be convenient to label the points of a point process with marks of a certain type. This may e.g. be an integer to label the type of point in the case of multiple point types, but the marks may be of much more general type. Processes of this kind are called marked point processes.

Definition 2.1.21
A marked point process X with positions in the Polish space T and marks in the Polish space M, is a point process on $S = T \times M$ with the additional property that the marginal process of locations is itself a point process on T. This marginal process may be denoted by X_T, and the marked process can then be denoted $X = \{(t, m_t) \mid t \in X_T\}$. □

As in the case for ordinary point processes the Poisson case plays a fundamental role for marked point processes, and it is defined as follows.

Definition 2.1.22
If $X_T \sim \text{Poisson}(T, \phi)$, where ϕ is a locally integrable intensity function, and given X_T the marks $\{m_t \mid t \in X_T\}$ are mutually independent, then X is a marked Poisson process. □

For a marked Poisson point process we have the following useful proposition.

Proposition 2.1.23
Let X be a marked Poisson point process with locations in a measure space $(T, \mathbb{B}(T), \mu_1)$ and marks in a measure space $(M, \mathbb{B}(M), \mu_2)$. If each mark m_t conditional on X_T has a density p_t with respect to μ_2 which does not depend on $X_T \backslash \{t\}$. Then $X \sim \text{Poisson}(S, \rho)$, with $S = T \times M$ and $\rho(t, m) = \phi(t)p_t(m)$.

For a proof of this proposition we refer to Møller and Waagepetersen (2004).

The following definition is based on Brémaud and Massoulié (2002)

Definition 2.1.24 (Shot Noise Process)
Let X be a marked point process on $\mathbb{R} \times M$ as given by definition 2.1.21 and h be a measurable function such that for all $t \in \mathbb{R}$ the sum

$$Y(t) = \sum_{n \in \mathbb{Z}} h(t - t_n, m_{t_n})$$

is well defined. The stochastic process Y is then called a shot noise process with independent random excitation. □

2.2 Examples

It is the goal to describe impulse response function models of e.g. Turin et al. (1972) and Saleh and Valenzuela (1987) as marked point processes. In both cases the impulse response consists of a sequence of arrivals $\tau_0, \tau_1, \tau_2, \cdots \in \mathbb{R}_+$ and a corresponding attenuation factor of each arrival $\beta_0, \beta_1, \beta_2, \cdots \in \mathbb{R}_+$. This can be considered as a marked point process X on $\mathbb{R}_+ \times \mathbb{R}_+$ with location process $X_T = \{\tau_0, \tau_1, \tau_2, \dots\}$ and marks $\{\beta_0, \beta_1, \beta_2, \dots\}$.

Before treating each model in detail the concept of an impulse response measure is introduced.

Definition 2.2.1
Let X be a marked point process with location process $X_T = \{\tau_0, \tau_1, \dots\}$ and marks $\{\beta_0, \beta_1, \dots\}$, where both are in $\mathbb{R}_+$. The impulse response measure is defined as

$$\mu_X(B) = \sum_{i=0}^{\infty} \beta_i \mathbf{I}[\tau_i \in B], \quad B \in \mathbb{B}(\mathbb{R}_+). \tag{2.2.1}$$

Furthermore

$$\mu_X(\mathbb{R}_+) = \sum_{i=0}^{\infty} \beta_i \tag{2.2.2}$$

is referred to as the total impulse response. □

In some cases, depending on the statistical properties of X, the impulse response measure will be finite almost surely and the Fourier-Stieltjes transform of the measure exists (see appendix A.2.1). In this case it is given by

$$\hat{\mu}_X(f) = \int \exp(-2\pi i f t) d\mu_X(t) = \sum_{i=0}^{\infty} \beta_i \exp(-2\pi i f \tau_i).$$

This is seen to coincide with the traditional Fourier transform of the weighted δ-train (1.1.2). In the physical interpretation of the model, it is also important that we do not expect to send a signal with finite energy and receive infinite energy from the channel.

2.2.1 The Turin model

In the model described in section 1.2.1 the marginal process X_T is assumed to be a Poisson point process on (τ_0, ∞) assuming that τ_0 is known. That is if we define $t_i = \tau_i - \tau_0$

for $i = 1, 2, \dots$ then $\{t_1, t_2, \dots\}$ are assumed to be Poisson$(\mathbb{R}_+, \lambda)$ for some constant parameter λ, which is equivalent to assuming $X_T \sim \text{Poisson}(\mathbb{R}_+, \lambda 1_{(\tau_0,\infty)}(t))$. Since we always condition on τ_0, knowing $\{\tau_1, \tau_2, \dots\}$ is equivalent to knowing $\{t_1, t_2, \dots\}$. Thus when we condition on the location process $X_T = \{\tau_1, \tau_2, \dots\}$ we also know $\{t_1, t_2, \dots\}$ and we will use either alternative as convenient.

Besides the Poisson assumption the Turin model assumes that given the location process X_T and the amplitude of the first arrival β_0 the marks $\{\beta_1, \beta_2, \dots\}$ are independent and $\beta_i \sim P_{t_i}$ where P_{t_i} is a probability distribution which does not depend on $X_T \backslash \{\tau_i\}$. For the rest of this section we will also let β_0 be given. If P_{t_i} has density p_{t_i} with respect to the Lebesgue measure then proposition 2.1.23 states, that $X \sim \text{Poisson}(S, \rho)$, where $S = \mathbb{R}_+^2$ and $\rho(t, m_t) = \lambda \mathbf{I}[t > \tau_0] p_t(m_t)$. The points of this Poisson point process on $\mathbb{R}_+^2$ will typically be denoted by $\xi_i = (\tau_i, \beta_i)$, and the entire process is written as

$$X = \{\xi_i \mid \xi_i \in \mathbb{R}_+^2, i = 1, 2, \dots\} \sim \text{Poisson}(\mathbb{R}_+^2, \lambda \mathbf{I}[t > \tau_0] p_t(m_t)).$$

The actual distribution P_{t_i} to be used when modeling the impulse response of a wireless channel has been widely discussed in the literature, but the general convention is that $\mathbb{E}(\beta_i|\tau_i)$ should be a decaying function of τ_i. In the Turin model it is assumed that $\beta_i|\tau_i$ is log-normally distributed. The conditional mean is not assumed to have a specific functional relation to the arrival time in the Turin model. It just states that the mean should be fitted to the data at hand, but in order to investigate the properties of the model from a theoretical point of view we treat two separate cases. First a model with polynomial decaying marks is discussed and afterwards an exponential decay model.

These models will be used to ensure that the Turin model leads to an almost surely finite total impulse response given by (2.2.2), which is equivalent to requiring that the sum of the marks is finite almost surely. Since each mark is finite it is sufficient to show that $\sum_{i=n}^{\infty} \beta_i$ is finite almost surely for some $n \in \mathbb{N}$. In the case of polynomial decay we assume that $\mathbb{E}(\beta_i \mid \tau_i, \beta_0) = \beta_0 t_i^{-a}$ for an $a > 0$. A sufficient condition for a random variable to be finite a.s. is that its mean exists and therefore we consider the following expression letting $n > a$ and using that $T_i \sim \text{Gamma}(i, \lambda)$ according to proposition 2.1.16.

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=n}^{\infty} \beta_i\right) &= \sum_{i=n}^{\infty} \mathbb{E}(\mathbb{E}(\beta_i|\tau_i)) \\
&= \sum_{i=n}^{\infty} \mathbb{E}(\beta_0 t_i^{-a}) \\
&= \sum_{i=n}^{\infty} \beta_0 \int_{\mathbb{R}_+} t^{-a} \frac{\lambda^i \exp(-t\lambda)}{\Gamma(i)} t^{i-1} dt \\
&= \sum_{i=n}^{\infty} \frac{\beta_0 \lambda^a \Gamma(i-a)}{\Gamma(i)} \int_{\mathbb{R}_+} \frac{\lambda^{(i-a)} \exp(-t\lambda)}{\Gamma(i-a)} t^{(i-a)-1} dt \\
&= \beta_0 \lambda^a \sum_{i=n}^{\infty} \frac{\Gamma(i-a)}{\Gamma(i)}.
\end{aligned}$$

This expression diverges if $a \leq 1$ and converges for $a > 1$, and thus, depending on the speed of the decay, the polynomial decay leads to an almost surely finite impulse response.

Now we turn to the case of exponential decay of the attenuation factors, such that

$$\mathbb{E}(\beta_i|\tau_i) = \beta_0 \exp(-\gamma t_i).$$

Since exponential decay is faster than polynomial decay, we have that for all degrees of polynomial decay, a, an $N(a) \in \mathbb{N}$ exists such that for all $b > N(a)$ the mean $\mathbb{E}\left(\sum_{i=b}^{\infty} \beta_i\right)$ is dominated by the corresponding mean for the polynomial decay, which means that the exponential decay impulse response is almost surely finite. Since we are able to obtain a closed form expression for the sum, we however choose to carry out the calculations

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=0}^{\infty} \beta_i\right) &= \beta_0 + \sum_{i=1}^{\infty} \mathbb{E}(\mathbb{E}(\beta_i|\tau_i)) \\
&= \beta_0 + \sum_{i=1}^{\infty} \mathbb{E}(\beta_0 \exp(-\gamma t_i)) \\
&= \beta_0 + \sum_{i=1}^{\infty} \beta_0 \int_{\mathbb{R}_+} \exp(-\gamma t_i) \frac{\lambda^i \exp(-t\lambda)}{\Gamma(i)} t^{i-1} dt \\
&= \beta_0 + \sum_{i=1}^{\infty} \frac{\beta_0 \lambda^i}{(\lambda+\gamma)^i} \int_{\mathbb{R}_+} \frac{(\lambda+\gamma)^i \exp(-t(\lambda+\gamma))}{\Gamma(i)} t^{i-1} dt \\
&= \beta_0 \Big(1 + \sum_{i=1}^{\infty} \big(\frac{\lambda}{\lambda+\gamma}\big)^i\Big) \\
&= \frac{\beta_0}{1 - \frac{\lambda}{\lambda+\gamma}} \\
&= \frac{\beta_0(\lambda+\gamma)}{\gamma}.
\end{aligned}$$

We conclude that both models for the attenuation factor, leads to an almost surely finite impulse response measure, and therefore has well-defined Fourier-Stieltjes Transform. Furthermore if we want $\mathbb{E}\left(\sum_i \beta_i\right) \leq 1$, we may control this by selecting β_0, λ and γ accordingly.

2.2.2 The Saleh-Valenzuela model

To describe the more complex model proposed by Saleh and Valenzuela (1987) we start with the following process from the Turin model conditional on (τ_0, β_0)

$$X' = \{\xi_i \mid \xi_i \in \mathbb{R}_+^2, i = 1, 2, \dots\} \sim \text{Poisson}(\mathbb{R}_+^2, \Lambda \mathbf{I}[t > \tau_0] p'_t(m_t)),$$

where $p'_t(m_t)$ is the conditional distribution of the marks given the location process. This will be denoted the mother process and at each point of this process an offspring process is started, and the process started by the i'th point of the mother process is

$$X^i|X' = \{\xi_{ij} \mid \xi_{ij} \in \mathbb{R}_+^2, j = 1, 2, \dots\} \sim \text{Poisson}(\mathbb{R}_+^2, \lambda \mathbf{I}[t > \tau_i] p_t(m_t)) \quad i \in \mathbb{N}_0$$

where $p_t(m_t)$ is the conditional distribution of the marks in the offspring process, given the locations. The total offspring process is then

$$X^* = \bigcup_{i=0}^{\infty} X^i.$$

The model assumes that the arrival times and attenuation factors of an impulse response are determined by the process

$$X = X' \cup X^*.$$

Conditional on (τ_0, β_0), it is assumed that

$$\mathbb{E}(\beta_i|\tau_i) = \beta_0 \exp(-\Gamma(\tau_i - \tau_0)).$$

Furthermore, conditional on the mother process, it is assumed that

$$\mathbb{E}(\beta_{ij}|\tau_{ij}) = \beta_i \exp(-\gamma(\tau_{ij} - \tau_i)).$$

An impulse response measure on this forms leads to a point process on $\mathbb{R}_+^2$ where the points are scattered around a pattern as displayed by the solid lines in figure 1.2.1.

We recall the notation $T_i = \tau_i - \tau_0$ from the previous section and similarly we define $T_{ij} = \tau_{ij} - \tau_i$, and note that $T_i \sim \text{Gamma}(i, \Lambda)$ and $T_{ij} \sim \text{Gamma}(j, \lambda)$. The total impulse response measure is given by the sum of the marks

$$\mu_X(\mathbb{R}_+) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \beta_{ij},$$

where $\beta_{i0} = \beta_i$, $i = 0, 1, 2, \ldots$ are the attenuation factors of the mother process. To check that this is finite almost surely we calculate the mean using the corresponding result from the Turin model in the previous section.

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\beta_{ij}\right) &= \mathbb{E}\left(\sum_{i=0}^{\infty}\beta_{i0}\right) + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty}\mathbb{E}(\mathbb{E}(\beta_{ij}|\tau_{ij})) \\
&= \sum_{i=0}^{\infty}\mathbb{E}(\beta_i) + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty}\mathbb{E}(\beta_i \exp(-\gamma t_{ij})) \\
&= \frac{\beta_0(\Lambda+\Gamma)}{\Gamma} + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty}\mathbb{E}(\mathbb{E}(\beta_i \exp(-\gamma t_{ij})|\beta_i)) \\
&= \frac{\beta_0(\Lambda+\Gamma)}{\Gamma} + \sum_{i=0}^{\infty}\mathbb{E}(\beta_i)\left(\frac{\lambda+\gamma}{\gamma} - 1\right) \\
&= \frac{\beta_0(\Lambda+\Gamma)}{\Gamma}\frac{\lambda+\gamma}{\gamma}. \qquad (2.2.3)
\end{aligned}$$

Before describing this process further we concentrate on the offspring process X^* given the mother process X' . Since

$$\rho^*(t) = \sum_{i=0}^{\infty} \lambda \mathbf{I}[t > \tau_i] p_t(m_t)$$

is locally integrable we may use proposition 2.1.19 to verify $X^*|X' \sim \text{Poisson}(\mathbb{R}_+^2, \rho^*)$. Furthermore the marginal offspring process

$$X_T^*|X' = \bigcup_{i=1}^{\infty} X_T^i|X'$$

is a superposition of Poisson point processes $X_T^i \sim \text{Poisson}(\mathbb{R}_+, \lambda \mathbf{I}[t > \tau_i])$, and thus

$$X_T^*|X' \sim \text{Poisson}\left(\mathbb{R}_+, \lambda \sum_{i=0}^{\infty} \mathbf{I}[t > \tau_i]\right).$$

It is seen that both X^* and X_T^* are Cox processes since they have a stochastic intensity function and conditional on the intensity function they are Poisson point processes. In the following we wish to calculate the mean intensity function and measure for both. First we recall that $T_i = \tau_i - \tau_0$ is a homogeneous Poisson point process with intensity Λ, and the corresponding counting process $N_{[0,t]} \sim \text{Poisson}(\Lambda t)$. Assuming that τ_0 is known we calculate the mean intensity function for the marginal process X_T^*.

$$\begin{aligned}
\mathbb{E}(\rho_T^*(t)) &= \mathbb{E}\big(\lambda \sum_{i=0}^{\infty} \mathbf{I}[t > \tau_i]\big) \\
&= \lambda \mathbf{I}[t > \tau_0] + \lambda \mathbb{E}\left(\sum_{i=1}^{\infty} \mathbf{I}[t - \tau_0 > \tau_i - \tau_0]\right) \\
&= \lambda \mathbf{I}[t > \tau_0] + \lambda \mathbb{E}(N_{[0,t-\tau_0]}) \\
&= \big(\lambda + \lambda\Lambda(t - \tau_0)\big)\mathbf{I}[t > \tau_0]
\end{aligned}$$

This leads to the mean intensity measure of the marginal offspring process.

$$\begin{aligned}
\mathbb{E}(\mu_T^*([0,t])) &= \mathbb{E}\left(\int_0^t \rho_T^*(s) ds\right) \\
&= \int_0^t \mathbb{E}(\rho_T^*(s)) ds \\
&= \lambda t \mathbf{I}[t > \tau_0] + \lambda\Lambda \int_0^t (s - \tau_0)\mathbf{I}[s > \tau_0] ds \\
&= \big(\lambda t + \frac{\lambda\Lambda}{2}(t - \tau_0)^2\big)\mathbf{I}[t > \tau_0]
\end{aligned}$$

Turning to the case of the entire marginal process X_T we obtain

$$\begin{aligned}
\mathbb{E}(\mu_T([0,t])) &= \mathbb{E}\big(\mu_T'([0,t]) + \mu_T^*([0,t])\big) \\
&= \mu_T'([0,t]) + \big(\lambda t + \frac{\lambda\Lambda}{2}(t - \tau_0)^2\big)\mathbf{I}[t > \tau_0] \\
&= \big(\Lambda(t - \tau_0) + \lambda t + \frac{\lambda\Lambda}{2}(t - \tau_0)^2\big)\mathbf{I}[t > \tau_0].
\end{aligned}$$

The mean intensity function of the entire marginal process is then

$$\mathbb{E}(\rho_T(t)) = \big(\Lambda + \lambda + \lambda\Lambda(t-\tau_0)\big)\mathbf{I}[t > \tau_0]$$

The above results for the marginal process X_T shows that the number of arrivals grows rather rapidly, but as shown in (2.2.3) the attenuation factors decay fast enough to ensure an almost surely finite impulse response measure.

Modifications

The model proposed by Saleh and Valenzuela (1987) has a couple of assumptions that might be unrealistic, which are straightforward to change. First of all it is assumed that each offspring process X^i has intensity λ for $i = 0, 1, 2, \ldots$ This is not an obvious assumption and it is easily dealt with by giving an index to λ corresponding to the offspring process such that $X_T^i \sim \text{Poisson}((\tau_i, \infty), \lambda_i)$, $i = 0, 1, 2, \ldots$. The effects on the above calculations are not overwhelming, but e.g. ρ_T becomes a piecewise linear function with slopes λ_i in the intervals (τ_i, τ_{i+1}). Furthermore it is possible to introduce individual decay rates for each offspring process such that $\mathbb{E}(\beta_{ij}\tau_{ij}) = \beta_i \exp(-\gamma_i(\tau_{ij} - \tau_i))$. In the following section the cluster idea of the model is developed in another direction.

2.2.3 A shot noise model

The models treated so far all assume that the impulse response of a system is given by discrete arrivals of attenuated impulses located at single points in time. This model however might be too simplistic and another approach is to assume that it is a shot noise process. The purpose of this model is to describe the clustering effect described in the Saleh Valenzuelah model in section 1.2.2 without using a weighted delta train.

Following definition 2.1.24 it is assumed that the impulse response function of channel is given by a stochastic process h, with

$$h(t) = \sum_{j=0}^{\infty} \mathbf{I}[t \geq \tau_j]\beta_j \exp(-\gamma_j(t - \tau_j)), \quad \text{for all } t \in \mathbb{R}, \tag{2.2.4}$$

where $(\tau_j, \beta_j, \gamma_j)$, $j = 1, 2, \ldots$ is assumed to be a marked point process on $\mathbb{R}_+ \times \mathbb{R}_+^2$. The arrival times and attenuation factors τ and β are assumed to be as in the Turin model in section 2.2.1. The decay rates γ are assumed to be iid. independent of τ, β and with $\mathbb{E}(\gamma_j^{-1}) < \infty$ for $j = 1, 2, \ldots$.

We wish to Fourier transform (2.2.4) and thus need to ensure that it is integrable almost surely, which is done by considering the mean of the integral

$$\begin{aligned}\mathbb{E}\left[\int h(t)dt\right] &= \sum_{j=0}^{\infty} \mathbb{E}\left[\int_{\tau_j}^{\infty} \beta_j \exp(-\gamma_j(t-\tau_j))dt\right] \\ &= \sum_{j=0}^{\infty} \mathbb{E}\left[\frac{\beta_j}{\gamma_j}\right] = \mathbb{E}[\gamma_0^{-1}] \sum_{j=0}^{n} \mathbb{E}[\beta_j].\end{aligned}$$

The integrability then reduces to summability of the β_j's. This has already been treated in the case of polynomial or exponential decay in section 2.2.1, and it is then meaningful to find the Fourier transform of (2.2.4)

$$\begin{aligned}
H(f) &= \int_{-\infty}^{\infty} \sum_{j=1}^{\infty} 1_{[\tau_j,\infty)} \beta_j \exp(-\gamma_j(t-\tau_j)) \exp(-2\pi i f) dt \\
&= \sum_{j=1}^{\infty} \beta_j \exp(\gamma_j \tau_j) \int_{\tau_j}^{\infty} \exp(-(\gamma_j + 2\pi i f)t) dt \\
&= \sum_{j=1}^{\infty} \beta_j \exp(\gamma_j \tau_j) \lim_{n\to\infty} \left[\frac{-1}{\gamma_j + 2\pi i f} \exp(-(\gamma_j + 2\pi i f)t) \right]_{\tau_j}^{n} \\
&= \sum_{j=1}^{\infty} \frac{\beta_j}{\gamma_j + 2\pi i f} \exp(-2\pi i f \tau_j) \\
&= \sum_{j=1}^{\infty} \frac{\beta_j}{\gamma_j^2 + (2\pi f)^2} \big(\gamma_j \cos(2\pi f \tau_j) - 2\pi f \sin(2\pi f \tau_j)\big) - \qquad (2.2.5) \\
&\quad i \sum_{j=1}^{\infty} \frac{\beta_j}{\gamma_j^2 + (2\pi f)^2} \big(2\pi f \cos(2\pi f \tau_j) + \gamma_j \sin(2\pi f \tau_j)\big). \qquad (2.2.6)
\end{aligned}$$

Summary

Inference for the Turin and Shot-noise model is done in chapter 6, using simulation based inference, which is described in chapter 4. The Saleh-Valenzuela model is not implemented for inference, but the clustering effect is to some degree described by the decaying exponential functions in the shot-noise model.

Chapter 3

Statistical inference

3.1 Maximum likelihood estimation

This section is based on Azzalini (1996) and Jensen (2006). The purpose of this section is to provide a brief insight into why the maximum likelihood method is the method of choice in most cases in likelihood statistics.

In likelihood theory we consider a parametric model for data, and given an underlying "true" parameter θ^* we assume that data x is sampled from a stochastic variable $X : (\Omega, \mathcal{F}, \mathcal{P}) \to (\mathcal{X}, \mathscr{F}, P_{\theta^*})$. The goal of the statistical inference, is to determine the value of the parameter, which has "produced" the data. We consider the statistical model $(\mathcal{X}, \mathscr{F}, \mathscr{P})$, where $\mathcal{X}$ is the state space, $\mathscr{F}$ is a σ-algebra, and $\mathscr{P} = \{P_\theta | \theta \in \Theta\}$ is a parametrized class of probability measures on $(\mathcal{X}, \mathscr{F})$. The parameter θ cannot be directly observed, we have only indirect knowledge through the fact, that data follows a distribution parametrized by θ.

If P_θ is absolutely continuous with respect to some measure μ for all $\theta \in \Theta$, we define the likelihood function as

Definition 3.1.1
The Likelihood function $L(\theta)$ or $L(\theta; x)$ is a function of θ which for each $x \in \mathcal{X}$ is given by

$$L(\theta) = L(\theta; x) = \frac{dP_\theta}{d\mu}(x), \quad \theta \in \Theta. \tag{3.1.1}$$

The log-likelihood function is defined as

$$l(\theta) = l(\theta; x) = \log\left[L(\theta; x)\right].$$

□

The definition given by (3.1.1) is the Radon-Nikodym derivative of the probability distribution with respect to μ, as described in appendix A.2. For most practical purposes μ is the Lebesgue measure on $\mathcal{X}$, and the likelihood is simply the density function for P_θ parametrized by θ.

An important principle in likelihood theory is the likelihood principle, which states that if $L(\theta; x) \propto L(\theta; y)$ for $x, y \in \mathcal{X}$ then the inference must lead to the same conclusions about θ.

Often we want to determine a point estimate for the underlying parameter θ^*. An estimate is a "guess" on the underlying parameter given a sample. An estimator is

defined as a measurable function $\tilde{\theta} : \mathcal{X} \to \Theta$, and an estimate is the value of the estimator evaluated at a sample point x. The typical approach is to maximize the likelihood function with respect to θ.

Definition 3.1.2
If $\hat{\theta} = \hat{\theta}(x)$ is such that

$$L(\theta) \leq L(\hat{\theta}), \quad \text{for all } \theta \in \Theta,$$

then $\hat{\theta}$ is called a maximum likelihood estimate(MLE). □

The MLE can be obtained by maximizing the likelihood with respect to θ or equivalently the log-likelihood, since the logarithm is a strictly increasing function. Assuming that $L(\theta)$ is differentiable and Θ is an open subset of $\mathbb{R}^k$ we consider

$$\frac{\partial l}{\partial \theta}(\theta) = 0. \tag{3.1.2}$$

This is a key element in maximum likelihood estimation, and (3.1.2) is called the likelihood equation. An MLE must of course satisfy (3.1.2), but to make sure a solution is a global maximum, one needs to establish further conditions, e.g. concavity of the likelihood function.

The MLE is not necessarily unique, and may in some cases not even exist. A more common problem is however that the MLE cannot always be derived analytically, and one must turn to numerical methods, such as the EM algorithm, which will be explained in detail in section 3.1.1.

A point estimate alone is not very informative, and it is desirable to have some information on the variation of this estimate, and on how close it approximates the true parameter. First we consider the mean of the estimate.

Definition 3.1.3
An estimate $\tilde{\theta} : \mathcal{X} \to \Theta$ is called unbiased if

$$\mathbb{E}_{\theta^*}\left[\tilde{\theta}(X)\right] = \theta^*, \quad \text{for all } \theta \in \Theta.$$

Otherwise the estimate is called biased. □

That an estimate is unbiased does of course not ensure a correct approximation of the true parameter, but it does ensure that the estimate is correct on average, and that there is no systematic error.

To gain some idea of the variation of the MLE, one might consider the probability that the true parameter is contained in some subset of Θ

Definition 3.1.4
A $(1-\alpha)$-confidence region is a mapping $K : \mathcal{X} \to 2^{\Theta}$, where 2^{Θ} denotes the set of all subsets of Θ, such that

$$P_{\theta^*}(\theta^* \in K(X)) = 1 - \alpha$$

□

The probability above holds a priori, i.e. before the experiment is conducted. Once a sample has been taken, it is not appropriate to say that the true parameter is contained in $K(x)$ with probability $1-\alpha$, it either is or is not.

Some other quantities which hint at the variation of the MLE are given in the definition below.

Definition 3.1.5
The stochastic variable

$$U(\theta) = \frac{\partial l(\theta; X)}{\partial \theta}$$

is called the Fisher score function, and the matrix

$$j(\theta) = -\frac{\partial^2 l(\theta; X)}{\partial \theta \partial \theta^\top}$$

is called the observed Fisher information. The mean of $j(\theta)$

$$I(\theta) = E_\theta \left[j(\theta) \right]$$

is called the expected Fisher information. □

For a one-dimensional parameter space the Fisher information is used to measure the curvature of the likelihood function around e.g. $\hat{\theta}$. Large values of $I(\hat{\theta})$ thus suggests that the variation around the MLE is small, and that it is a good estimate. For the multivariate case, one may consider e.g. the curvature parallel to the coordinate axes, or the size of the eigenvalues of the matrix evaluated at the MLE, which will provide some idea of the curvature of the entire function. This does however not give a complete picture, and there is as far as we know, no standard on this subject. The observed Fisher information provides an approximation for the expected Fisher information.

For samples x_1 and x_2 from independent random variables X_1 and X_2 we have that $L(\theta; x_1, x_2) = L(\theta; x_1)L(\theta; x_2)$ which means that $l(\theta; x_1, x_2) = l(\theta; x_1) + l(\theta; x_2)$. Using this we obtain that the expected Fisher information is additive, that is

$$I(\theta; x_1, x_2) = I(\theta; x_1) + I_2(\theta; x_2).$$

This means that for an iid. sample of length n the expected Fisher information may be calculated as

$$I(\theta) = ni(\theta),$$

where $i(\theta)$ denotes the information for a single observation, and the Fisher information thus increases with the number of observations.

For an unbiased estimator $\tilde{\theta}$ a theoretical lower bound on the variance exist. It is given as

$$\mathrm{Var}(\tilde{\theta}; \theta) \geq \frac{1}{I(\theta)}. \tag{3.1.3}$$

Equation (3.1.3) is known as the Cramér-Rao inequality and may be used to check the efficiency of a given estimator. The expected Fisher information thus represents an index of the maximal mean precision for an unbiased estimator. In some cases the MLE

can be proven to be unbiased and attain the Cramér-Rao lower bound, which means that it cannot be improved. A more general version of the Cramér-Rao inequality and a proof of this can be found in Azzalini (1996) page 73.

Under suitable regularity conditions (see e.g. Azzalini (1996) page 82), it is possible to gain an asymptotic result for the distribution of the MLE. For a k-dimensional θ, the general result is given as

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N_k(0, I(\theta^*)^{-1}),$$

where $I(\theta^*)^{-1}$ is the inverse matrix for $I(\theta^*)$. For large n the distribution of the MLE, may then be approximated as

$$\hat{\theta} \sim N_k\left(\theta^*, \frac{1}{n} I(\theta^*)^{-1}\right).$$

One may then calculate approximate confidence region, by using the above distribution. It is also seen that the precision of the distribution increases with the sample size, and that for large samples the MLE closely approximates the true parameter.

3.1.1 The EM algorithm

This section provides an introduction to the EM algorithm based on Bilmes (1998) and Ng et al. (2002). As mentioned above, the Likelihood cannot always be maximized analytically, and one must turn to numerical methods. The EM algorithm is one such method developed to estimate the MLE for datasets with missing values, but it can however be successfully implemented in other settings by augmenting the data. The algorithm works by iteratively maximizing the likelihood function, where each iteration consists of two steps: An expectation(E) step, and a maximization(M) step.

We assume that the observed data x has a probability density function $f_X(x;\theta) = L(\theta;x)$, where θ is a vector containing the unknown parameters for which we wish to obtain an MLE. The basic setup for the EM algorithm is as follows.
Assume that the data at hand x has missing values. We will call this the incomplete data. Assume further that a complete data set $z = (x, y)$ exists and that the complete data has a joint density function

$$f_Z(z|\theta) = f_{X,Y}(x, y|\theta) = f_{Y|X}(y|x, \theta) f_X(x|\theta) \tag{3.1.4}$$

As usual the likelihood function for the complete data corresponds to the density function such that $L_c(\theta) = L_c(\theta; z) = f_Z(z|\theta)$.

At the $(k+1)$'th iteration of the algorithm, given the current parameter estimate $\theta^{(k)}$, the algorithm then performs the two following steps:

E-Step
Calculate

$$Q(\theta; \theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}(l_c(\theta)|x)$$

M-Step
Choose $\theta^{(k+1)}$ such that

$$Q(\theta^{(k+1)};\theta^{(k)}) \geq Q(\theta;\theta^{(k)}), \quad \text{for all } \theta \in \Theta. \tag{3.1.5}$$

The E-step is the expected value of the missing data, given the observed data, using the current parameter $\theta^{(k)}$. This gives a function of θ which is then maximized in the M-step. The two steps are then repeated until some criteria of convergence, e.g. $L(\theta^{(k+1)}) - L(\theta^{(k)}) < \epsilon$ for some $\epsilon > 0$, is fulfilled.

In some cases it is not possible to maximize $Q(\theta, \theta^{(k)})$ analytically, and one can instead use a generalized version of the EM algorithm where $\theta^{(k+1)}$ is chosen such that $Q(\theta^{(k+1)}, \theta_k) \geq Q(\theta^{(k)}, \theta^{(k)})$.

Convergence of the EM algorithm

Using (3.1.4) the complete data log likelihood can be expressed as

$$l_c(\theta) = \log(f_{Y|X}(y|x,\theta) + l(\theta)$$

The expected value at the $(k+1)$'th iteration, given data x and the current parameter estimate θ_k then yields

$$\begin{aligned} \mathbb{E}_{\theta^{(k)}}(l_c(\theta)|x) &= \mathbb{E}_{\theta^{(k)}}(\log(f_{Y|X}(y|x,\theta))|x) + E_{\theta^{(k)}}(l(\theta)|x) \\ Q(\theta,\theta^{(k)}) &= H(\theta,\theta^{(k)}) + l(\theta), \end{aligned} \tag{3.1.6}$$

where $H(\theta,\theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}(\log(f_{Y|X}(y|x,\theta))|x)$. Using (3.1.6) it then follows that

$$l(\theta^{(k+1)}) - l(\theta^{(k)}) = \left[Q(\theta^{(k+1)},\theta^{(k)}) - Q(\theta^{(k)},\theta^{(k)})\right] - \left[H(\theta^{(k+1)},\theta) - H(\theta_k,\theta^{(k)})\right].$$

By (3.1.5) the expression in the first brackets on the right hand side is nonnegative. By Jensen's inequality we have

$$\begin{aligned} -\left[H(\theta^{(k+1)},\theta) - H(\theta^{(k)},\theta^{(k)})\right] &= \mathbb{E}_{\theta^{(k)}}\left[-\log\left(\frac{f_{Y|X}(y|x,\theta^{(k+1)})}{f_{Y|X}(y|x,\theta^{(k)})}\right)|X\right] \\ &\geq -\log\left(\mathbb{E}_{\theta_k}\left[\frac{f_{Y|X}(y|x,\theta^{(k+1)})}{f_{Y|X}(y|x,\theta^{(k)})}\right]|X\right) \\ &= -\log\left(\int \frac{f_{Y|X}(y|x,\theta^{(k+1)})}{f_{Y|X}(y|x,\theta^{(k)})} f_{Y|X}(y|x,\theta^{(k)})dy\right) \\ &= -\log(1) = 0. \end{aligned}$$

This means that the entire right side is nonnegative, and that the value of the likelihood function increases at each step of the algorithm. That the likelihood increases at each step of the algorithm does however not ensure convergence to the MLE, it does in fact not even ensure convergence to a local maximum of a multimodal likelihood. To ensure convergence Cappé et al. (2005) shows that if all points of the sequence $\{\theta^{(k)}\}_{k\in\mathbb{N}_0}$ are contained in a compact subset of Θ regardless of the choice of the initial point $\theta^{(0)}$,

and if Q is continuous in both arguments, then the algorithm converges to some local extremum of the likelihood, although this may only be a saddle point. For a more thorough discussion of criteria of convergence we refer to Cappé et al. (2005) and Wu (1983).
The point is however that given the right conditions the convergence is ensured, and for practical purposes one may need to start the algorithm from a number of initial points, unless the likelihood is unimodal.

3.1.2 Implementation of the EM algorithm in signal estimation

Feder and Weinstein (1988) gives an elegant approach on how to implement the EM algorithm to determine parameter values, when the data consists of superimposed signals, that is models on the form

$$y(t) = \sum_{k=1}^{K} s_k(t;\theta_k) + n(t),$$

where θ_k denotes the unknown parameters, associated with the k'th signal component s_k which conditional on θ_k is a known deterministic real or complex function. Correspondingly $n(t)$ denotes either real or complex Gaussian white noise with known variance σ^2. Assuming that the signal is observed at discrete times $t_i, \quad i = 1, \ldots, N$ the log likelihood for the unknown parameter vector $\theta = (\theta_1, \ldots, \theta_K)$ takes the form

$$l(\theta) = c + \frac{\lambda}{2}\left[\sum_{i=1}^{N} \|y(t_i) - \sum_{k=1}^{K} s_k(t_i;\theta_k)\|^2\right], \tag{3.1.7}$$

where c is a normalizing constant, and $\lambda = 1$ if $s(t)$ is real and $\lambda = 2$ if $s(t)$ is complex (see appendix A.3). Maximizing (3.1.7) with respect to θ, is not necessarily an easy task, and may require numerical methods such as the Newton-Raphson method. In order to simplify the calculation we here choose to implement the EM algorithm.
We choose the complete data as the decomposition of the observed signal into the different signal components, i.e.

$$x(t_i) = [x_1(t_i), \ldots, x_k(t_i)]^\top,$$

where $x_k(t_i) = s_k(t_i;\theta_k) + n_k(t_i)$. The random variable n_k is the k'th element in an arbitrary decomposition of the noise, such that $n = \sum_k n_k$ and are assumed to be independent of n_j for $k \neq j$ and normal distributed with zero mean and variance $\beta_k\sigma^2$ where $\beta_k > 0$ and $\sum_k \beta_k = 1$.

The log likelihood for the complete data, can then be written as

$$l_c(\theta) = c' - \frac{\lambda}{2}\left(\sum_{i=1}^{N} [x(t_i) - s(t_i;\theta)]^\top \Sigma^{-1} [x(t_i) - s(t_i;\theta)]\right)$$

where the covariance matrix Σ is a diagonal matrix with entries $\beta_1\sigma^2, \ldots, \beta_K\sigma^2$. The mean vector is defined as $s(t_i,\theta) = (s_1(t_i,\theta_1), \ldots, s_K(t_i,\theta_K))^\top$ and c' is a constant

independent of θ.

Applying the EM algorithm to this setting, the E-step yields

$$Q(\theta,\theta') = c' - \frac{\lambda}{2}\Bigg(\sum_{i=1}^{N} \left[\mathbb{E}_{\theta'}[X(t_i)|y(t_i)] - s(t_i;\theta)\right]^\top \cdot \Sigma^{-1}\left[\mathbb{E}_{\theta'}\left[X(t_i)|y(t_i)\right] - s(t_i;\theta)\right]\Bigg). \tag{3.1.8}$$

In order to calculate the conditional mean in the expression above, a few results for the multivariate normal distribution is applied. Since X and Y are related by a linear transformation, that is $Y = HX$, where $H = [1,\ldots,1]$, the joint distribution of (X,Y) is a degenerate multivariate real or complex normal distribution with mean and covariance given by

$$\mathbb{E}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} s(\theta) \\ Hs(\theta) \end{pmatrix}, \quad \mathrm{Cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{bmatrix} \Sigma & \Sigma H^\top \\ H\Sigma & H\Sigma H^\top \end{bmatrix}$$

By proposition A.3.3 the conditional mean is then given as

$$\mathbb{E}_{\theta'}\left[X(t_i)|y(t_i)\right] = s(t_i,\theta') + \Sigma H^\top [H\Sigma H^\top]^- \left(y(t_i) - Hs(t_i,\theta')\right)$$

The result in the appendix refers to the complex case, but it also holds for the real multivariate normal distribution (see e.g. appendix C.1 in Lauritzen (1996)).

An easy calculation shows that

$$\Sigma H^\top [H\Sigma H^\top]^{-1} = [\beta_1,\ldots,\beta_k]^\top,$$

which means that

$$\mathbb{E}_{\theta'}\left[X_k(t_i)|y(t_i)\right] = s_k(t_i,\theta') + \beta_k\Big[y(t_i) - \sum_{l=1}^{K} s_l(t_i,\theta'_k)\Big].$$

Using this and the fact that Σ is a diagonal matrix, (3.1.8) may be rewritten as

$$Q(\theta,\theta') = c' - \frac{\lambda}{2}\left(\sum_{k=1}^{K}\sum_{i=1}^{N} \beta_k \|\mathbb{E}_{\theta'}\left[X_k(t_i)|y(t_i)\right] - s_k(t_i,\theta_k)\|^2\right).$$

It follows that Q is maximized by minimizing each of the terms in the sum over k, and the EM algorithm at step $j+1$ given the current estimate $\theta^{(j)}$ is then given as

E-step
For $k = 1,\ldots,K$ calculate

$$\mathbb{E}_{\theta^{(j)}}\left[X_k(t_i)|y(t_i)\right] = s_k(t_i,\theta^{(j)}) + \beta_k\Big[y(t_i) - \sum_{l=1}^{K} s_l(t_i,\theta_k^{(j)})\Big]. \tag{3.1.9}$$

M-step
For $k = 1, \ldots, K$ obtain $\theta_k^{(j+1)}$ as the value which minimizes (3.1.9)

In Feder and Weinstein (1988) it is mentioned that the β_k's can be used to control the rate of convergence of the algorithm, and possibly to avoid convergence to an unwanted stationary point. The authors furthermore state that the RMS error performance of the algorithm is the minimum attainable by the Cramér-Rao lower bound.

Summary

The setting above may be used to estimate e.g. arrival times and attenuation factors in the Turin Model as described in section 1.2.1. Due to the unified frame for real and complex signals, the algorithm can be used on measurements of either the impulse or frequency response. For measurements of the impulse response $s_k(t, \theta_k) = \beta_k \delta(t - \tau_k)$ and for measurements of the frequency response $s_k(\omega, \theta_k) = \beta_k \exp(-2\pi i \omega \tau_k)$, where the unknown parameter in both cases is given as $\theta_k = (\tau_k, \beta_k)$.

Although the algorithm is computationally efficient, it does however have two major problems. One is the inability to work with unknown noise. The algorithm could have been developed for this setting, with this extra parameter it is however not possible to maximize the likelihood for each signal separately. The other problem is the fact that the number of signals has to be known. Although methods for estimating the order of the model exists (see e.g. Cappé et al. (2005), chapter 15), it would be preferable to treat both the number of parameters and their values as unknowns in the same statistical inference. An algorithm which treats this issue is described in section 4.3 and implemented on the Turin model and the shot noise model in section 6.2 and 6.3.

3.2 Bayesian inference

This section gives a short introduction to Bayesian inference and it is mainly based on Lee (2004) and Gelman et al. (2003), and we refer especially to the latter for a more thorough treatment of Bayesian inference.

3.2.1 Basic definitions

In likelihood based inference as described in section 3.1 we wish obtain information about the unknown parameter vector θ based on observed data $\boldsymbol{x}$. Data is linked to the parameter through the statistical model used in the specific scenario, and the inference is based solely on the model and the given data. In Bayesian statistics the parameters are considered as unknown stochastic variables, for which we wish to determine the distribution. It is assumed that the parameters have a known a priori distribution before we have knowledge of the data, which is formally defined as

Definition 3.2.1 (Prior)
Let the parameter space $(\Theta, \mathcal{F}_\Theta)$ be a measurable space, and let θ be distributed according to some probability measure $\mathcal{P}_\theta$ on $(\Theta, \mathcal{F}_\Theta)$ prior to obtaining the data $\boldsymbol{x}$.

Then $\mathcal{P}_\theta$ is the a priori distribution of θ, and if it has density p_θ with respect to a measure ξ, this will be denoted the a priori density for θ. □

Often the term "a priori" will be replaced with the more common "prior", and both the distribution and the density will sometimes be referred to simply as the prior. The concept of a prior is discussed widely in the Bayesian literature, and we will treat this later, but for now we omit this discussion. The observation model is a statistical model for how the data is distributed given the parameters of the model, which is defined as

Definition 3.2.2 (Observation model)
Let data $\boldsymbol{X}$ be a stochastic variable on the measurable space $(\mathcal{X}, \mathcal{F}_\mathcal{X})$, which given a parameter value θ is distributed according to $\mathcal{P}_{\boldsymbol{X}|\theta}$. Then $\mathcal{P}_{\boldsymbol{X}|\theta}$ is the observation model, and it is assumed to have density $p_{\boldsymbol{X}|\theta}$ with respect to some measure μ on $(\mathcal{X}, \mathcal{F}_\mathcal{X})$. □

Note that the density $p_{\boldsymbol{X}|\theta}$ is actually the likelihood function $L(\theta; \boldsymbol{X})$. With these concepts at hand we can show the basic proposition in Bayesian statistics

Proposition 3.2.3 (Bayes formula)
Let the prior and observation model be given as in definition 3.2.1 and 3.2.2, and let $\mathcal{P}_{\theta|\boldsymbol{x}}$ be the distribution of the parameters θ given the observed data $\boldsymbol{X} = \boldsymbol{x}$. Then $\mathcal{P}_{\theta|\boldsymbol{x}}$ is called the a posteriori distribution of θ, and it has density

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta)p(\boldsymbol{x}|\theta)}{\int_\Theta p(\tilde{\theta})p(\boldsymbol{x}|\tilde{\theta})d\xi(\tilde{\theta})}$$

with respect to ξ.

PROOF:
The simultaneous density for $(\boldsymbol{X}, \theta)$ on $(\mathcal{X} \times \Theta, \mathcal{F}_\mathcal{X} \otimes \mathcal{F}_\theta, \mu \otimes \xi)$ is $p(\theta)p(\boldsymbol{x}|\theta)$, which yields the result. ■

In Bayesian statistics the following version of Bayes formula is often used

$$p(\theta|\boldsymbol{x}) \propto p(\boldsymbol{x}|\theta)p(\theta),$$

since it in some applications is sufficient to know the posterior up to proportionality, and if the exact distribution is needed the normalizing constant can be found by integration.

The basic feature of Bayesian inference is that when data $\boldsymbol{X} = \boldsymbol{x}$ is observed we update our prior beliefs by multiplying the prior and the likelihood (and normalizing if it is necessary) to obtain a new distribution describing our beliefs about the parameter of interest θ. Now if we have some new data $\boldsymbol{Y} = \boldsymbol{y}$, which is independent of $\boldsymbol{X}$, it is not necessary to restart the inference. Rather we can use the posterior based on $\boldsymbol{x}$ as the new prior for θ, and update our beliefs using the likelihood $L(\theta; \boldsymbol{y})$. This is easily seen to give the same result as starting inference from scratch

$$p(\theta|\boldsymbol{x}, \boldsymbol{y}) \propto p(\theta)L(\theta; \boldsymbol{x}, \boldsymbol{y}) \propto p(\theta)L(\theta; \boldsymbol{x})L(\theta; \boldsymbol{y}) \propto p(\theta|\boldsymbol{x})L(\theta; \boldsymbol{y}).$$

3.2.2 Prior distributions

Now we return to a discussion of the prior and some difficulties concerning this. The prior should reflect the knowledge available before obtaining data, which can be a combination of former experiences, expert knowledge, etc. We consider a simple example, where it is assumed, that we are going to measure the height of a group of Danish men and we are interested in estimating the mean height. If we denote this mean by θ we might expect that $\theta \sim N(\theta_0, \sigma_0^2)$ with $\theta_0 = 178$cm and $\sigma_0 = 5$cm, illustrating that prior to the experiment we already have an idea of what the parameter is, but with some uncertainty of course. This illustrates a key issue in Bayesian inference; the probabilities are subjective. If somebody else was to do inference for this dataset their prior beliefs might be different from the ones stated here leading to different results. It is however often possible to show that the influence of the prior vanishes as the number of observations grows, so in the case of large data sets different priors often lead to similar results. This can be illustrated by assuming that the height of Danish men in general follow a Gaussian distribution with the unknown mean θ and variance σ^2. In order to make the example very simple we unrealistically assume that σ^2 is known, such that we only need to make inference for the mean. Then if we have n independent measurements of the height it can be shown that the posterior is Gaussian with variance σ_1^2 given by the relation

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + n\frac{1}{\sigma^2} \tag{3.2.1}$$

and mean θ_1, given by

$$\theta_1 = \frac{\sigma_1^2}{\sigma_0^2}\theta_0 + n\frac{\sigma_1^2}{\sigma^2}\bar{x}. \tag{3.2.2}$$

It is seen that the precision (reciprocal value of variance) of the posterior grows with the number of observations, and the mean value is a weighted sum of the prior and observed means with the weight of the observed mean growing with the number of observations. This illustrates that a different Gaussian prior would not change the posterior considerably if the number of observations is large, which seems reassuring.

There are however other problematic issues concerning the prior, which we will treat briefly here. Again this is most easily illustrated by an example. If we are to make inference of some probability parameter π, which we have no prior knowledge of, it would seem reasonable to use a uniform prior on $[0, 1]$, but we might as well argue that we know nothing of $\sqrt{\pi}$ and choose this to be uniform on $[0, 1]$. This leads to different results, and it is not obvious which one to use. In general the problem of expressing lack of prior knowledge is discussed widely in the Bayesian literature, and there is no obvious way to do this. In the one-dimensional case Jeffreys (1961) suggests using a prior given by

$$p(\theta) \propto \sqrt{I(\theta)},$$

where $I(\theta)$ denotes the Fisher information (see definition 3.1.5). The reason for this choice is that it is invariant under certain transformations of the parameter. Consider the transformation $\psi = \psi(\theta)$, where ψ is assumed to be a monotone c^2-function. Then using that $I(\psi) = I(\theta)|d\theta/d\psi|^2$ and the transformation theorem for stochastic

variables, we obtain

$$\begin{aligned} p(\psi) &= p(\theta)|d\theta/d\psi| \\ &\propto \sqrt{I(\theta)}|d\theta/d\psi| \\ &= \sqrt{I(\psi)/|d\theta/d\psi|^2}|d\theta/d\psi| \\ &= \sqrt{I(\psi)}, \end{aligned}$$

which is the same as we would have specified using Jeffreys' prior directly on the transformed variable ψ. It is possible to extend this rule to the multi-dimensional case, but this leads to some controversial results, and a discussion of this is outside the scope of the current introduction to Bayesian inference.

In the discussion of priors representing no knowledge (sometimes referred to as reference priors) the concept of improper priors is introduced. This is used when we approximate our prior beliefs by an infinite measure (e.g. the Lebesgue measure on the real line), and not a proper probability measure. The line of thought can be illustrated by returning to the example of a Gaussian prior and a Gaussian likelihood with known variance. If we have a very vague prior knowledge we could express this by using a large variance on the prior, leading to a very flat prior distribution. We might argue that the limiting case of the prior variance $\sigma_0^2 = \infty$ expresses total lack of knowledge, but this is not well defined of course. It corresponds to a prior $p(\theta) \propto 1$ on the whole real line, which is not a probability density with respect to the Lebesgue measure. We could however use this improper prior as an approximation to our prior knowledge, and it is seen from (3.2.1) and (3.2.2) that it leads to the posterior distribution $N(\bar{x}, \sigma^2/n)$. In this case the improper prior is combined with the likelihood to give a proper posterior, but this is not always the case and improper priors should be treated with care.

In some cases it is possible to select a prior such that the prior and posterior belongs to the same class of distributions. If $p(\boldsymbol{x}|\theta)$ is an observation model, then a class Π of prior distributions is said to form a conjugate family if the posterior density

$$p(\theta|\boldsymbol{x}) \propto p(\theta)p(\boldsymbol{x}|\theta)$$

is in the class Π for all $\boldsymbol{x}$ whenever the prior density $p(\theta) \in \Pi$. If it is possible to express ones prior knowledge about θ as a conjugate prior, which is not always the case, it is often easier to analytically derive the wanted results.

3.2.3 Posterior summaries

Once the posterior distribution is calculated we wish to give a summary of its properties, such as point and interval estimates. As a point estimate either the posterior mean, median or mode is usually presented, and unless stated otherwise we will use the posterior mean $\mathbb{E}(\theta|\boldsymbol{x})$. Regarding interval estimates we are usually concerned with an interval containing a specific amount of the total probability mass (typically 90%, 95% or 99%, but there is no canonical choice), and there are several different criteria to choose this interval by. A highest density region (HDR) is the smallest set that includes the chosen amount of the probability mass. This set is not necessarily

connected if the posterior is multi-modal, and another possibility is to use a central posterior region (CPR), which is the smallest connected set containing the chosen probability mass. These interval estimates can be rather difficult to derive when the posterior is not on a well known form, and often it will be more convenient simply to work with an interval given by an upper and lower quantile q_u and q_l. That is if we are interested in a 95% posterior interval we use the $2,5\%$ quantile $q_{2,5\%}$ and the $97,5\%$ quantile $q_{97,5\%}$ to form the interval estimate $[q_{2,5\%}, q_{97,5\%}]$. Intervals of this type are denoted posterior confidence intervals (PCI).

For many practical purposes, an analytical derivation of the posterior distribution is a very difficult task, and point and interval estimates are therefore calculated using simulations. The method for simulation from the unnormalized distributions that commonly arise in Bayesian inference relies heavily on the theory of Markov chains which is treated before actual simulation schemes such as the Metropolis-Hastings algorithm are presented.

Chapter 4

Simulation based inference

4.1 Markov chains

This section deals with discrete time Markov chains with continuous state space, i.e. a sequence of stochastic variables $\{X(n)\}_{n\in\mathbb{N}_0}$, where $X(n) : \Omega \to \mathcal{X}, \quad \mathcal{X} \subseteq \mathbb{R}$ for all $n \in \mathbb{N}_0$. This section is based on Robert and Casella (1999). A Markov chain is defined via. its transition kernel.

Definition 4.1.1 (Transition Kernel)
A transition kernel is a function K defined on $\mathcal{X} \times \mathbb{B}(\mathcal{X})$, where $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$ is a measurable space, such that

- For all $x \in \mathcal{X}$, $K(x, \cdot)$ is a probability measure
- For all $A \in \mathbb{B}(\mathcal{X})$, $K(\cdot, A)$ is measurable.

□

Definition 4.1.2 (Markovchain)
A stochastic process $\{X(n)\}_{n\in\mathbb{N}}$ is a Markov chain with transition kernel K, if for all $n \geq 0$, all $A \in \mathbb{B}(\mathcal{X})$ and all $x_0, \ldots, x_n \in \mathcal{X}$

$$\begin{aligned} P(X(n+1) \in A | X(0) = x_0, \ldots, X(n) = x_n) &= P(X(n+1) \in A | X(n) = x_n) \\ &= \int_A K(x_n, dy) \end{aligned} \tag{4.1.1}$$

□

If the initial value $X(0) \sim \mu$, the distribution of $\{X(n)\}_{n\in\mathbb{N}_0}$ is denoted by P_μ, or P_{x_0}, if μ is a degenerate distribution. If the initial distribution μ is known, the distribution of the Markov chain is completely determined by the transition kernel, since

$$\begin{aligned} P_\mu(X(1) \in A_1) &= \int_{\mathcal{X}} K(y_0, A_1)\mu(dy_0) \\ P_\mu((X(1), X(2)) \in A_1 \times A_2) &= P_\mu(X(2) \in A_2 | X(1) \in A_1) P_\mu(X(1) \in A_1) \\ &= \int_{\mathcal{X}} \int_{A_1} K(y, A_2) K(y_0, dy)\mu(dy_0) \\ &\vdots \\ P_\mu((X(1), \ldots, X(n)) \in \times_{i=1}^n A_i) &= \int_{\mathcal{X}} \int_{A_1} \cdots \int_{A_{n-1}} K(y_{n-1}, A_n) \\ &\quad \times K(y_{n-2}, dy_{n-1}) \cdots K(y_0, dy_1)\mu(dy_0) \end{aligned}$$

If $K^1(x, A) = K(x, A)$, the n-step transition kernel is recursively defined by

$$K^n(x, A) = \int_{\mathcal{X}} K^{n-1}(y, A)K(x, dy).$$

Markov chains are often used as estimation tools in Bayesian inference, to obtain samples from a given posterior distribution. This will be described in further detail in section 4.2. In order to be useful in this regard, the Markov chain is required to meet a number of conditions, to ensure convergence of the ergodic mean to the mean of the given posterior distribution, and convergence in total variation norm.

Definition 4.1.3 (φ-irreducibility)
Given a measure φ, a Markov chain with transition kernel K is φ-irreducible, if for every $A \in \mathbb{B}(\mathcal{X})$, with $\varphi(A) > 0$ there exists an $n \in \mathbb{N}$ such that $K^n(x, A) > 0$ for all $x \in \mathcal{X}$. The chain is strongly irreducible if $n = 1$ for all measurable A, with $\phi(A) > 0$. □

The irreducibility, ensures that the Markov chain can move from any position in the state space, to any Borel element with positive measure, within a finite number of steps. If the Markov chain $\{X(n)\}_{n\in\mathbb{N}_0}$ is φ-irreducible, it can be shown that there exists a probability measure ψ, such that $\{X(n)\}_{n\in\mathbb{N}_0}$ is ψ-irreducible.

Although irreducibility ensures that the Markov chain visits every element of the Borel algebra, this condition is not sufficient to guarantee the wanted convergence. We need to impose a further demand, on how often the chain visits every element. A Markov chain is recurrent if the mean number of visits in every element of the Borel algebra is infinite. In this section we however restrict our attention to a stronger property called Harris recurrence.

Definition 4.1.4 (Harris recurrence)
Let $\eta_A = \sum_{n=0}^{\infty} 1_A\{X(n)\}_{n\in\mathbb{N}_0}$ and $A \in \mathbb{B}(\mathcal{X})$. If $P_x(\eta_A = \infty) = 1$ for all $x \in A$, then A is Harris recurrent. The Markov chain $\{X(n)\}_{n\in\mathbb{N}_0}$ is Harris recurrent, if there exists a probability measure ψ such that the chain is ψ-irreducible, and if for all $x \in \mathcal{X}$ and all $A \in \mathbb{B}(\mathcal{X})$ with $\psi(A) > 0$, A is Harris recurrent. □

For a ψ-irreducible Markov chain Meyn and Tweedie (1993) shows the existence of a disjoint partitioning of the state space $\mathcal{X} = A_0 \cup \cdots \cup A_{d-1} \cup A_d$, with $\psi(A_d) = 0$, such that $x \in A_0 \Rightarrow K(x, A_1) = 1, x \in A_1 \Rightarrow K(x, A_2) = 1, \cdots, x \in A_{d-1} \Rightarrow K(x, A_0) = 1$.

Definition 4.1.5
A ψ-irreducible Markov chain $\{X(n)\}_{n\in\mathbb{N}_0}$ is periodic if $d > 1$, and otherwise aperiodic. □

If $\{X(n)\}_{n\in\mathbb{N}_0}$ is strongly irreducible it is also aperiodic, since there is positive probability of moving to every element with positive measure within a single step.

Definition 4.1.6
A σ-finite measure Π is invariant for the transition kernel K, and the associated Markov chain, if

$$\Pi(A) = \int_{\mathcal{X}} K(x, A)\Pi(dx), \quad \forall A \in \mathbb{B}(\mathcal{X}).$$

□

If Π is a probability measure, it describes a stationary distribution, which means that if $X(n) \sim \Pi$, then

$$P(X(n+1) \in A) = \int_{\mathcal{X}} K(x_n, A)\Pi(dx_n) = \Pi(A), \tag{4.1.2}$$

which means that $X(n+1) \sim \Pi$. By induction this is valid for all $X(m), \quad m = n+1, n+2, \ldots$, which means that if the Markov chain is sampling from the stationary distribution, it will continue to do so.

Proposition 4.1.7
If $\{X(n)\}_{n\in\mathbb{N}_0}$ is a Harris recurrent Markov chain, there exists an invariant measure Π, which is unique op to a multiplicative factor. □

If Π is an invariant measure, for the chain $\{X(n)\}_{n\in\mathbb{N}_0}$, then $\{X(n)\}_{n\in\mathbb{N}_0}$ is Π-irreducible.

If we define the total variation norm by

$$\|\mu_1 - \mu_2\|_{TV} = \sup_{A\in\mathcal{X}} |\mu_1(A) - \mu_2(A)|,$$

then the stationary distribution can be used to describe limiting results for Markov chains under certain regularity conditions.

Theorem 4.1.8 (Convergence Theorem for Markov chains)
If $\{X(n)\}_{n\in\mathbb{N}_0}$ is Harris recurrent, aperiodic and has Π as stationary distribution, then

$$\lim_{n\to\infty} \left\| \int K^n(x, \cdot)\mu(dx) - \Pi \right\|_{TV}$$

for every initial distribution μ □

Theorem 4.1.8 states that regardless of how the Markov chain is initialized it converges to the same distribution. For practical purposes the chain is run for a certain amount of time, called the burn in, after which it is assumed to sample from the stationary distribution.

The ergodic average of a discrete time Markov chain is defined by

$$S_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X(i)).$$

Using this we have

Theorem 4.1.9 (Ergodic Theorem)
If $\{X(n)\}_{n\in\mathbb{N}_0}$ has a σ-finite invariant measure Π, the following statements are equivalent

1. If $f, g \in L^1(\Pi)$ with $\int g(x)\Pi(dx) \neq 0$, then
$$\lim_{n\to\infty} \frac{S_n(f)}{S_n(g)} = \frac{\int f(x)\Pi(dx)}{\int g(x)\Pi(dx)}, \quad \Pi - \text{a.e.}$$
2. The Markov chain $\{X(n)\}_{n\in\mathbb{N}_0}$ is Harris recurrent.

□

4.2 The Metropolis-Hastings algorithm

This section provides a method to obtain empirical estimates of a posterior distribution via. Markov Chain Monte Carlo methodology. The results are based on Robert and Casella (1999) and Berthelsen and Møller (2004). A Markov Chain Monte Carlo (MCMC) method for simulation of a distribution P is any method that generates an ergodic Markov chain $\{X_n\}_{n\in\mathbb{N}_0}$ with stationary distribution P. The main focus in this section lies on the Metropolis-Hastings algorithm, which is a very general MCMC method which can be applied to a broad variety of problems.

The Metropolis-Hastings algorithm is defined via its target distribution Π, which is the distribution we wish to sample from, and a conditional density called the proposal density.

Definition 4.2.1 (The Metropolis-Hastings algorithm)
Let $X(n) = x_n$.

Generate $U \sim \text{Unif}(0,1)$ and $Y \sim q(x_n, \cdot)$

$$X(n+1) = \begin{cases} Y & \text{if } U \leq \alpha(X(n), Y) \\ X(n) & \text{otherwise.} \end{cases}$$

The acceptance probability α is defined by

$$\alpha(x,y) = \min\left\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right\}. \tag{4.2.1}$$

□

It is easily seen from (4.2.1) that it is sufficient to know the target density op to a multiplicative factor, and that the acceptance probability is simplified if the proposal density is symmetrical, which then gives

$$\alpha(x,y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

To prove the wanted convergence results for the Markov chain generated by the Metropolis-Hastings algorithm we need to impose certain regularity conditions on the target- and proposal density. First of the results are more easily proven if $\text{supp}(\pi) = S$ is a connected set, which we will assume. Furthermore the algorithm does not work if a measurable set A, such that

$$\int_A \pi(x)dx > 0 \quad \text{and} \quad \int_A q(x,y)dy = 0, \quad \forall x \in S,$$

exists. This is caused by the fact that the chain never visits the set A, which has positive probability in the stationary distribution, if $X_0 \notin A$. A necessary condition is thus

$$\text{supp}(\pi) \subset \bigcup_{x\in\text{supp}(\pi)} \text{supp}(q(x,\cdot)).$$

To prove that the Markov chain produced by the Metropolis-Hastings algorithm has the target distribution as stationary distribution, we need to introduce the kernels associated density function, and a property called the detailed balance condition.

Definition 4.2.2
If a measurable space $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$ has a σ-finite measure ν such that $K(x, \cdot)$ has a density function with respect to ν, this density function is denoted $k(x, y)$. □

In the following it is assumed that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a measurable function.

Definition 4.2.3
A Markov chain with transition kernel K satisfies the detailed balance condition(DBC) if there exists a function π satisfying

$$k(y, x)\pi(y) = k(x, y)\pi(x), \quad \forall(x, y).$$

□

Proposition 4.2.4
If a Markov chain with transition kernel K satisfies DBC for a probability density function π, then Π is the chains stationary distribution.

PROOF:
The strategy is to prove that if $X(n) \sim \Pi$, then $X(n+1) \sim \Pi$, which means that Π is the stationary distribution for $\{X(n)\}_{n\in\mathbb{N}_0}$.

For every measurable set B, we have

$$\begin{aligned}\int_{\mathcal{Y}} K(y, B)\pi(y)dy &= \int_{\mathcal{Y}}\int_{B} k(y, x)\pi(y)dxdy \\ &= \int_{\mathcal{Y}}\int_{B} k(x, y)\pi(x)dxdy \\ &= \int_{B} \pi(x)dx\end{aligned}$$

■

Proposition 4.2.5
For every proposal density q with $\text{supp}(\pi) \subseteq \text{supp}(q)$, the Markov chain produced by the Metropolis-Hastings algorithm has Π as stationary distribution.

PROOF:
The density function for the Metropolis-Hastings algorithms transition kernel is

$$k(x, y) = \alpha(x, y)q(x, y) + \left(1 - \int \alpha(x, y)q(x, y)dy\right)\delta(y - x), \tag{4.2.2}$$

where the first part corresponds to proposing and accepting a move from x to y and the second part corresponds to rejecting a move and $x = y$.

To prove that the transition kernel for the Metropolis-Hastings algorithm satisfies DBC, we first prove that the first part of (4.2.2) satisfies

$$\alpha(x, y)q(x, y)\pi(x) = \alpha(y, x)q(y, x)\pi(y).$$

This is verified by considering either $\pi(x)q(x,y) > \pi(y)q(y,x)$ or $\pi(y)q(y,x) > \pi(x)q(x,y)$.

The last part of the proof, is to show that

$$\pi(x)\left(1 - \int \alpha(x,y)q(x,y)dy\right)\delta(y-x) = \pi(y)\left(1 - \int \alpha(y,x)q(y,x)dx\right)\delta(x-y).$$

Since the weights on the delta function are equal for $x = y$, and $\delta(y-x) = \delta(x-y) = 0$ for $x \neq y$, the two expressions are equal.
This shows that the transition kernel satisfies DBC, and the result follows from proposition 4.2.4. ■

The results above says that the Metropolis-Hastings algorithm has the wanted target distribution as stationary distribution. We however need to ensure that the chain converges to this distribution. To do this it is sufficient to show that the chain is aperiodic and Harris recurrent, since the convergence then follows from the results in section 4.1.

Lemma 4.2.6
If the Metropolis-Hastings algorithm generates a Π-irreducible Markov chain $\{X(n)\}_{n\in\mathbb{N}_0}$, then the chain is Harris recurrent. □

As a consequence of this, we now have the convergence theorem for the Metropolis Hastings algorithm

Theorem 4.2.7
Assume that chain $\{X(n)\}_{n\in\mathbb{N}_0}$ generated by the Metropolis-Hastings algorithm is Π-irreducible.

1. If $h \in L(\Pi)$, then

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} h(X(n)) = \int h(x)\pi(x)dx, \quad \Pi - a.e.$$

2. If, $\{X(n)\}_{n\in\mathbb{N}_0}$ furthermore is aperiodic, then

$$\lim_{m\to\infty} \left|\left| \int K^m(x,\cdot)\mu(dx) - \Pi \right|\right|_{TV} = 0$$

for every initial distribution μ.

Proof:
Since $\{X(n)\}_{n\in\mathbb{N}_0}$ is assumed Π-irreducible,then by lemma 4.2.6 it is also Harris recurrent. The theorem is now a direct consequence of theorem 4.1.9 and theorem 4.1.8. ■

When designing a Metropolis-Hastings algorithm the requirements for the results above can be checked by the following two sufficient conditions.

A Markov chain $\{X(n)\}_{n\in\mathbb{N}_0}$ cannot be periodic if $P(X(n+1) = X(n)) > 0$, which according to definition 4.2.1 is the same as

$$P(\alpha(X(n), Y) \geq 1) < 1.$$

This means that the chain is aperiodic, if

$$P[\pi(X(n))q(X(n), Y) \leq \pi(Y)q(Y, X(n))] < 1.$$

A sufficient condition for irreducibility is, that the proposal density is positive for all $(x, y) \in S \times S$, since it is then possible to move to any measurable set with positive measure within a single step.

4.3 Reversible jump MCMC

In some cases in statistical inference one of the unknowns is the order of the model. There is a number of different methods of estimating the model order, and the method described below is the Bayesian approach. As described in section 4.2 computations in Bayesian inference can be handled via. MCMC methods. Reversible jump MCMC is an extension of the Metropolis Hastings algorithm that not only proposes new values of a number of parameters, but also proposes a jump in dimension which is accepted with some probability. The algorithm was first suggested for spatial point processes by Geyer and Møller (1994) and later generalized by Green (1995). The algorithm has a wide variety of usages, e.g. determining the number of change points for a Poisson process, or determining the number of components in a mixture density.

Let $\{\mathscr{M}_k, k \in \mathscr{K}\}$ be a countable collection of models, where model $\mathscr{M}_k$ has a parameter vector $\theta_k \in \mathbb{R}^{n_k}$. Let x denote the pair (k, θ_k), then for a given k, $x \in \mathscr{C}_k = k \times \mathbb{R}^{n_k}$, and in general $x \in \mathscr{C} = \bigcup_k \mathscr{C}_k$. The general state space $\mathscr{C}$ is equipped with the sigma algebra given by $\mathbb{B}(\mathscr{C}) = \sigma\{(k, \theta_k) | \theta_k \in \mathbb{B}(\mathscr{C}_k), k \in \mathscr{K}\}$. The goal of this section is to establish a Markov Chain on the space given above, with a transition kernel P that satisfies the detailed balance condition, that is

$$\int_A \int_B \pi(dx)P(x, dx') = \int_B \int_A \pi(dx')P(x', dx). \tag{4.3.1}$$

The probability of moving from state x to dx' with a move of type m, e.g. a shift up in dimension, is given by $q_m(x, dx')$, and the probability of no proposal of change is thus $1 - \sum_m q_m(x, \mathscr{C})$. The probability of accepting a move of type m is $\alpha_m(x, x')$, and we wish to derive an expression for α such that the chain satisfies (4.3.1). The transition kernel for the chain is given by

$$P(x, B) = \sum_m \int_B [q_m(x, dx')\alpha_m(x, x')] + s(x)1_B(x), \tag{4.3.2}$$

where $B \in \mathbb{B}(\mathscr{C})$, and

$$s(x) = \sum_m \int_{\mathscr{C}} [q_m(x, dx')(1 - \alpha_m(x, x'))] + 1 - \sum_m q_m(x, \mathscr{C}),$$

is the probability of rejection of a proposal or no proposal. If we define a measure $\mu_x(B) = 1_B(x)$, and substitute (4.3.2) into the left side of (4.3.1), we get

$$\begin{aligned}
&\int_A \int_B \pi(dx) \left[\sum_m q_m(x, dx')\alpha_m(x, x') + s(x)\mu_x(dx') \right] \\
&= \sum_m \int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') + \int_A \int_B s(x)\mu_x(dx')\pi(dx) \\
&= \sum_m \int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') + \int_A s(x)\pi(dx)1_B(x) \\
&= \sum_m \int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') + \int_{A\cap B} s(x)\pi(dx)
\end{aligned}$$

The right side gives

$$\sum_m \int_B \pi(dx') \int_A q_m(x', dx)\alpha_m(x', x) + \int_{B\cap A} \pi(dx')s(x')$$

Since the last terms are equal it is sufficient for (4.3.1) to hold that

$$\begin{aligned}
\int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') &= \int_B \pi(dx') \int_A q_m(x', dx)\alpha_m(x', x) \\
\int_A q_m(x, B)\pi(dx) &= \int_B q_m(x', A)\pi(dx') \qquad (4.3.3)
\end{aligned}$$

for all m and all $A, B \in \mathbb{B}(\mathscr{C})$.

The idea given by Green (1995) is to decompose the model jumping into jumps between two submodels, and then establish a bijection between the two subspaces $\mathscr{C}_k$ and $\mathscr{C}_{k'}$. More precisely we generate two continuous random auxiliary variables u_k and $u_{k'}$ of length m_k and $m_{k'}$, such that $n_k + m_k = n_{k'} + m_{k'}$, and set $(\theta_{k'}, u_{k'}) = T(\theta_k, u_k) = (T_1(\theta_k, u_k), T_2(\theta_k, u_k))$. The auxiliary variables are assumed to have density functions q_k and $q_{k'}$ with respect to the m_k and $m_{k'}$ dimensional Lebesgue measures. If the dimension jumping probability measure $q_m(\cdot)\pi(\cdot)$ also is assumed to have a density function, f, with respect to the proper dimensional Lebesgue measure, and the probability of choosing a jump is denoted $j(\cdot)$, the left side of (4.3.3) can be rewritten as

$$\begin{aligned}
&\int_{\mathbb{R}^{n_k}} 1_A(\theta_k) \int_{\mathbb{R}^{m_k}} 1_B(\theta_{k'})j(k \to k')\alpha(x, x')f(x)q_k(u_k)du_k d\theta_k \\
&= \int_{\mathbb{R}^{n_k}} \int_{\mathbb{R}^{m_k}} 1_A(\theta_k)1_B(\theta_{k'})j(k \to k')\alpha(x, x')f(x)q_k(u_k)du_k d\theta_k. \qquad (4.3.4)
\end{aligned}$$

The right side of (4.3.3) is rewritten as

$$\int_{\mathbb{R}^{n_{k'}}} \int_{\mathbb{R}^{m_{k'}}} 1_B(\theta_{k'})1_A(\theta_k)j(k' \to k)\alpha(x', x)f(x')q_{k'}(u_{k'})du_{k'} d\theta_{k'}.$$

Since T is a bijection, we can now apply the transformation theorem for integrals to the equation above, and we get

$$\int_{\mathbb{R}^{n_k}} \int_{\mathbb{R}^{m_k}} 1_B(T_1(\theta_k, u_k)) 1_A(\theta_k) j(k' \to k) \alpha((k', T_1(\theta_k, u_k)), (k, \theta_k)) \cdot f(k, T_1(\theta_k, u_k)) q_{k'}(T_2(u_k)) \left| \frac{\partial T}{\partial(\theta_k, u_k)} \right| du_k d\theta_k. \tag{4.3.5}$$

Comparing (4.3.4) with (4.3.5) we see that (4.3.3) is satisfied if

$$j(k \to k') f(k, \theta_k) q_k(u_k) \alpha(x, x') = j(k' \to k) f(k', \theta_{k'}) q_{k'}(u_{k'}) \left| \frac{\partial T}{\partial(\theta_k, u_k)} \right| \alpha(x', x).$$

This holds if the accept probability is chosen as

$$\alpha(x, x') = \min \left\{ 1, \frac{j(k' \to k) f(x') q_{k'}(u_{k'})}{j(k \to k') f(x) q_k(u_k)} \left| \frac{\partial T}{\partial(\theta_k, u_k)} \right| \right\}.$$

Part II

Implementation

Chapter 5

Data description

5.1 Data acquisition

In order to model the impulse response function of an UWB wireless channel, actual measurements are needed. There are several different ways of approaching the problem of measuring the impulse response. Since the Dirac-impulse is a limit of real functions one way of approximating the impulse response is by sending a very short electrical pulse to the transmitting antenna and then measure the received signal at the receiving antenna. An alternative way is to measure the transfer function, which was the method used for the data at hand. Since the transfer function is a complex valued function the method of measurement is not straight forward and the following describes how it is possible to evaluate the transfer function at some given frequency f_0.

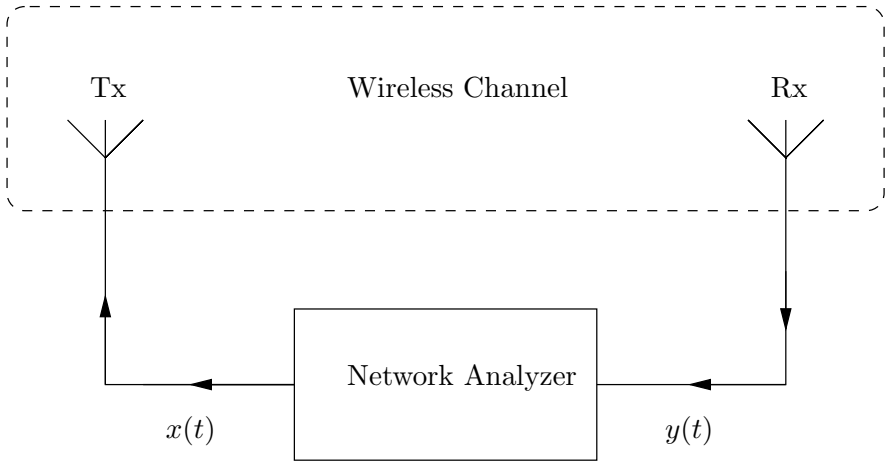


Figure 5.1.1: Schematic representation of the measurement.

Figure 5.1.1 is a schematic representation of the measurement technique. The network analyzer is connected to a transmitter and a receiver antenna positioned in the environment of interest. It sends a given signal x to the transmitter and measures the received signal y. The measurement of the transfer function evaluated at f_0 is done by sending the input signal $x = \cos_{2\pi f_0}$, where $\cos_{2\pi f_0}(t) = \cos(2\pi f_0 t)$ for all t, and then process the output y as illustrated in figure 5.1.2. This results in the values $\mathrm{Re}(H(f_0))$ and $\mathrm{Im}(H(f_0))$ as shown in the following. First we define the signal z_c by $z_c(t) = y(t)\cos(2\pi f_0 t)$ for all t and consider the Fourier transform Z_c, using

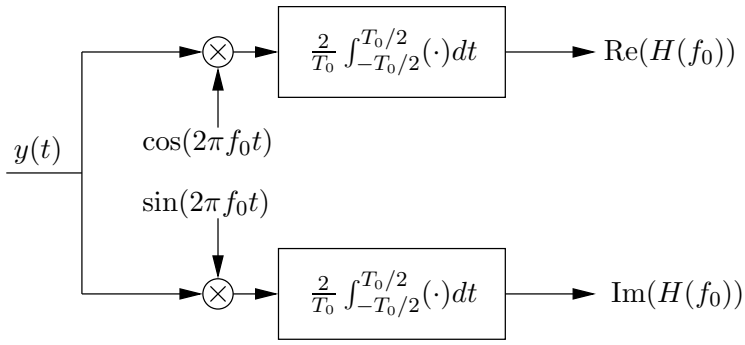


Figure 5.1.2: Network analyzer.

$$\begin{aligned}
\mathscr{F}(\cos_{2\pi f_0})(f) &= \tfrac{1}{2}(\delta(f-f_0)+\delta(f+f_0)) \\
Z_c(f) &= (Y * \mathscr{F}(\cos_{2\pi f_0}))(f) \\
&= \frac{1}{2}\int_{-\infty}^{\infty} Y(\xi)\left[\delta(f-f_0-\xi)+\delta(f+f_0-\xi)\right]d\xi \\
&= \frac{1}{2}\left[Y(f-f_0)+Y(f+f_0)\right]. \qquad (5.1.1)
\end{aligned}$$

Since $Y(f) = H(f)X(f) = H(f)\mathscr{F}(\cos_{2\pi f_0})(f)$ equation (5.1.1) leads to

$$\begin{aligned}
Z_c(f) &= \frac{1}{2}\Big[H(f-f_0)X(f-f_0)+H(f+f_0)X(f+f_0)\Big] \\
&= \frac{1}{4}\Big[H(f-f_0)\delta(f-2f_0)+[H(f-f_0) \\
&\qquad\qquad +H(f+f_0)]\delta(f)+H(f+f_0)\delta(f+2f_0)\Big] \\
&= \frac{1}{4}\Big[H(f_0)\delta(f-2f_0)+2\mathrm{Re}(H(f_0))\delta(f)+H(-f_0)\delta(f+2f_0)\Big],
\end{aligned}$$

where the final result follows from the fact that the impulse response is a real function, and thus has a complex symmetric Fourier transform, which leads to $H(f_0)+H(-f_0) = 2\mathrm{Re}(H(f_0))$. Finally Parseval's identity is used to see that integration over a period $T_0 = \frac{1}{f_0}$ as indicated in figure 5.1.2 gives the right result

$$\begin{aligned}
\frac{2}{T_0}\int_{-T_0/2}^{T_0/2} z_c(t)dt &= \frac{2}{T_0}\int_{-\infty}^{\infty} z_c(t)1_{[-T_0/2,T_0/2]}dt \\
&= \frac{2}{T_0}\int_{-\infty}^{\infty} Z_c(f)\frac{\sin(\pi f T_0)}{\pi f}df \\
&= 2\int_{-\infty}^{\infty} Z_c(f)\mathrm{sinc}(\pi f T_0)df \\
&= \frac{1}{2}H(f_0)\mathrm{sinc}(\pi 2 f_0 T_0)+\mathrm{Re}(H(f_0))\mathrm{sinc}(0) \\
&\qquad\qquad +\frac{1}{2}H(-f_0)\mathrm{sinc}(-\pi 2 f_0 T_0) \\
&= \mathrm{Re}(H(f_0)).
\end{aligned}$$

The calculations can be performed the same way for the signal z_s given by $z_s(t) = y(t)\sin(2\pi f_0 t)$ for all t leading to $\mathrm{Im}(H(f_0))$.

5.2 The noise model

In a realistic channel model some kind of noise model is needed, and a very commonly used model is the additive white Gaussian noise channel. This simply states that the receiver does not receive the theoretically correct signal y but rather a version corrupted by noise given by $y + W$. The process W is then assumed to be a white Gaussian process with autocorrelation function $R_W = \sigma^2\delta$. This implies that in our measurements the actual values are

$$\frac{2}{T_0}\int_{-T_0/2}^{T_0/2}(y(t)+W(t))\cos(2\pi f_0 t)dt = \mathrm{Re}(H(f_0)) + \frac{2}{T_0}\int_{-T_0/2}^{T_0/2}W(t)\cos(2\pi f_0 t)dt,$$

and correspondingly

$$\mathrm{Im}(H(f_0)) + \frac{2}{T_0}\int_{-T_0/2}^{T_0/2}W(t)\sin(2\pi f_0 t)dt.$$

These quantities are not well defined however, which is due to the chaotic behavior of white noise. It is actually not a stochastic process but rather a generalized stochastic process which is related to the theory of generalized functions (also referred to as distributions). A common solution to this problem is to restate the problem as integration with respect to Brownian motion. For the moment we will ignore these problems and show the desired results by some informal calculations. Afterwards a comment regarding the mathematical formalism will be given. The informal calculations below are inspired by Land and Fleury (2006).

We start by introducing the notation

$$W_c = \frac{2}{T_0}\int_{-T_0/2}^{T_0/2}W(t)\cos(2\pi f_0 t)dt,$$

and

$$W_s = \frac{2}{T_0}\int_{-T_0/2}^{T_0/2}W(t)\sin(2\pi f_0 t)dt.$$

We are interested in the statistical properties of the noise quantities W_c and W_s. Since they are linear transforms of Gaussian processes they follow a Gaussian distribution and we just have to determine the mean, variance and covariance.

$$\begin{aligned}\mathbb{E}(W_c) &= \mathbb{E}\left[\frac{2}{T_0}\int_{-T_0/2}^{T_0/2}W(t)\cos(2\pi f_0 t)dt\right]\\ &= \frac{2}{T_0}\int_{-T_0/2}^{T_0/2}\mathbb{E}(W(t))\cos(2\pi f_0 t)dt\\ &= 0,\end{aligned}$$

and in the same way $\mathbb{E}(W_s) = 0$. The variances are calculated as follows

$$\begin{aligned}
\mathbb{E}[(W_c - \mathbb{E}(W_c))^2] &= \mathbb{E}(W_c^2) \\
&= \frac{4}{T_0^2}\mathbb{E}[(\int_{-T_0/2}^{T_0/2} W(t)\cos(2\pi f_0 t)dt)(\int_{-T_0/2}^{T_0/2} W(s)\cos(2\pi f_0 s)ds)] \\
&= \frac{4}{T_0^2}\int_{-T_0/2}^{T_0/2}\int_{-T_0/2}^{T_0/2} \mathbb{E}[W(t)W(s)]\cos(2\pi f_0 t)\cos(2\pi f_0 s)dtds \\
&= \frac{4\sigma^2}{T_0^2}\int_{-T_0/2}^{T_0/2}\int_{-T_0/2}^{T_0/2} \delta(t-s)\cos(2\pi f_0 s)ds\cos(2\pi f_0 t)dt \\
&= \frac{4\sigma^2}{T_0^2}\int_{-T_0/2}^{T_0/2} \cos(2\pi f_0 t)\cos(2\pi f_0 t)dt \\
&= \frac{4\pi\sigma^2}{T_0^2}.
\end{aligned}$$

And the same type of calculations leads to $\mathrm{Var}(W_s) = \frac{4\pi\sigma^2}{T_0^2}$. To calculate the covariance of W_c and W_s a couple of similar calculations are done

$$\begin{aligned}
\mathrm{Cov}(W_c, W_s) &= \mathbb{E}(W_c W_s) \\
&= \frac{4}{T_0^2}\int_{-T_0/2}^{T_0/2}\int_{-T_0/2}^{T_0/2} \mathbb{E}[W(t)W(s)]\cos(2\pi f_0 t)\sin(2\pi f_0 s)dtds \\
&= \frac{4\sigma^2}{T_0^2}\int_{-T_0/2}^{T_0/2}\int_{-T_0/2}^{T_0/2} \delta(t-s)\sin(2\pi f_0 s)ds\cos(2\pi f_0 t)dt \\
&= \frac{4\sigma^2}{T_0^2}\int_{-T_0/2}^{T_0/2} \sin(2\pi f_0 t)\cos(2\pi f_0 t)dt \\
&= 0.
\end{aligned}$$

This leads to the conclusion that the we observe $H(f_0) + \hat{W}(f_0)$, where $\hat{W}(f_0) \sim \mathbb{C}N(0, \frac{8\pi\sigma^2}{T_0^2})$.

As mentioned earlier the formal mathematics of the above needs to be clarified, which involves the theory of stochastic integration and an introduction is found in Øksendal (2003). Here it is argued that the white noise could be replaced by a Brownian motion which has independent Gaussian distributed increments and is a proper stochastic process with continuous realizations. So we replace the meaningless

$$\text{"}\int f(t)W(t)dt\text{"}$$

with

$$\text{"}\int f(t)\sigma dB_t\text{"},$$

where σB_t is a Brownian motion with $\mathrm{Var}(\sigma B_t) = \sigma^2 t$. This integral is developed much like the usual Riemann-Stieltjes integral as the limit of the sum

$$\sum_{j=0}^{n-1} f(t_j^*)\sigma(B_{t_{j+1}} - B_{t_j}), \tag{5.2.1}$$

but due to the infinite variation of B_t this sum has different limits dependent on the choice of t_j^*. The two common choices are the midpoint of the interval leading to the Stratonovich integral and the left endpoint leading to the Itô integral, which is used here. Due to the Gaussian distributed increments of a Brownian motion the sum in (5.2.1) is Gaussian distributed and so is the limit, which justifies that the integral of the noise process follows a Gaussian distribution. A basic property of the Itô integral is that $\mathbb{E}\left(\int_S^T f(t)\sigma dB_t\right) = 0$, which also coincides with the result in our informal calculations. Furthermore a generalization of the Itô isometry (Taksar and Højgaard (2006), chapter 1)

$$\mathbb{E}\Big[\int_S^T f(t)\sigma dB_t \int_S^T g(t)\sigma dB_t\Big] = \mathbb{E}\big[\int_S^T f(t)g(t)\sigma^2 dt\big]$$

can be used to verify

$$\begin{aligned}\mathrm{Var}\Big[\frac{2}{T_0}\big(\int_{-T_0/2}^{T_0/2}\cos(2\pi f_0 t)\sigma dB_t\big)^2\Big] &= \frac{4}{T_0^2}\mathbb{E}\Big[\big(\int_{-T_0/2}^{T_0/2}\cos(2\pi f_0 t)\sigma dB_t\big)^2\Big] \\ &= \frac{4}{T_0^2}\mathbb{E}\big[\int_{-T_0/2}^{T_0/2}\cos^2(2\pi f_0 t)\sigma^2 dt\big] \\ &= \frac{4\pi\sigma^2}{T_0^2}.\end{aligned}$$

And in the same way we obtain

$$\mathrm{Var}\left[\left(\int_{-T_0/2}^{T_0/2}\sin(2\pi f_0 t)\sigma dB_t\right)^2\right] = \frac{4\pi\sigma^2}{T_0^2}.$$

This shows that the variance of the two noise components is the same and it does not depend on the frequency f_0.

Along with independence of the two integrals this ensures that the noise is complex Gaussian as concluded in the informal calculations. The independence follows from the covariance, which is calculated using the Itô isometry

$$\begin{aligned}\mathrm{Cov}(W_c, W_s) &= \mathbb{E}\Big[\int_{-T_0/2}^{T_0/2}\cos(2\pi f_0 t)\sigma dB_t \int_{-T_0/2}^{T_0/2}\sin(2\pi f_0 t)\sigma dB_t\Big] \\ &= \mathbb{E}\Big[\int_{-T_0/2}^{T_0/2}\cos(2\pi f_0 t)\sin(2\pi f_0 t)\sigma^2 dt\Big] \\ &= 0.\end{aligned}$$

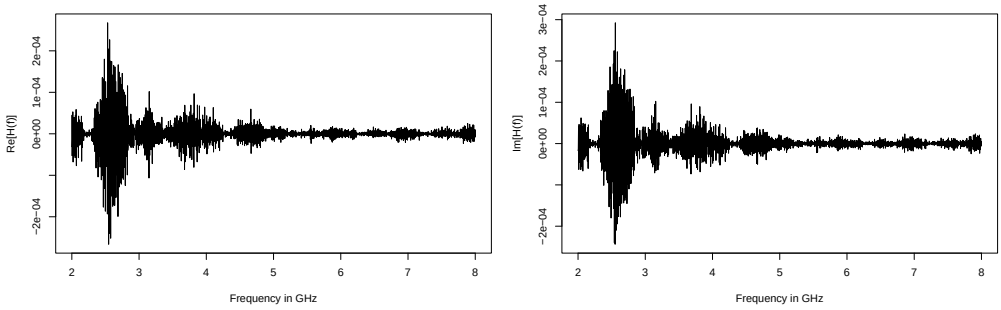
5.3 Descriptive Data Analysis

In this section we will give a short description of the actual observed data. A database of measurements and instructions for usage supplied by the Intel Corporation may be found at Scholtz (2006). The data is given as an array with 1847 observations of 1601-dimensional complex vectors. Each of these vectors is a discrete sampled frequency

response function with equally spaced samples in the interval $[2, 8]$ GHz. The data was observed in three different scenarios: An office, a townhouse and an anechoic chamber. We will use four different graphical representations of the signal.

Real and imaginary part of the frequency response

Figure 5.3.1 shows an example of the data in its unprocessed form. The figure shows a plot of the real and imaginary parts of the complex frequency response, as a function of the frequency. The two signals parts are seen to be very similar, there is however a small shift in phase between the two. This similarity is verified by the empirical cross correlation, as shown in figure 5.3.2. The figure shows the correlation between $\mathrm{Re}[H(f_i)]$ and $\mathrm{Im}[H(f_{i+k})]$. The cross correllation is actually only defined for stationary stochastic processes, but even though the data does not fulfill this requirement, the empirical cross correlation still hints at the covariation between real and imaginary parts of data.



(a) Real part of frequency response.

(b) Imaginary part of frequency response.

Figure 5.3.1: Example of data.

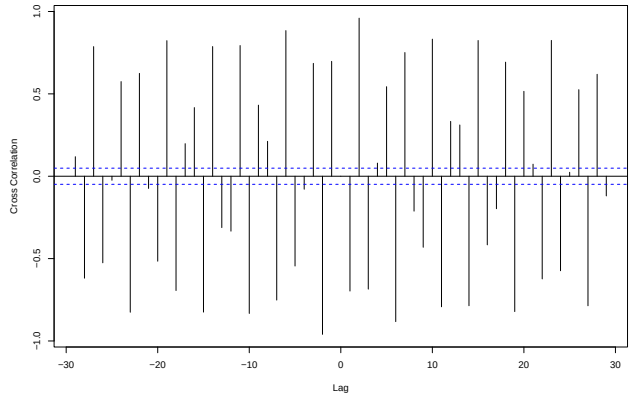


Figure 5.3.2: Cross Correlation between real and imaginary parts.

Impulse response

Figure 5.3.3 shows a filtered version of the impulse response function. The IFFT used to get a time domain version of the observed signal, works under the assumption that the time domain signal is periodic with period T and the samples in the frequency domain are equally spaced with distance $1/T$ (see Hazewinkel (1995), page 648). In this case the samples are spaced with 3.75 MHz, and the assumed period of the impulse response, in which we have observations, can thus be obtained as $[0, 266]$ ns.

Since the frequency response is observed on a closed interval $[f_m, f_M]$, the observed function can be thought of as the frequency response multiplied by an indicator function for this closed interval. Since the impulse response is real, the frequency response is complex symmetric, and we can also assume knowledge of the frequency response on $[-f_M, -f_m]$. The time domain representation for the signal can then be obtained as

$$\mathcal{F}^{-1}\Big[H \cdot \mathbf{I}\big(\cdot \in [-f_M, -f_m] \cup [f_m, f_M]\big)\Big] = h * g$$

where the filter g is given by $g(t) = \big(\sin(2\pi f_M t) - \sin(2\pi f_m t)\big)(\pi t)^{-1}$ for all $t \in \mathbb{R}$.

Thus when we use the IFFT the output is the impulse response function blurred by a sinc-like filter, and the time domain representation of the signal is therefore unfit for statistical inference. The graph can however still be used to gain some idea of the arrival times and possible clusters.

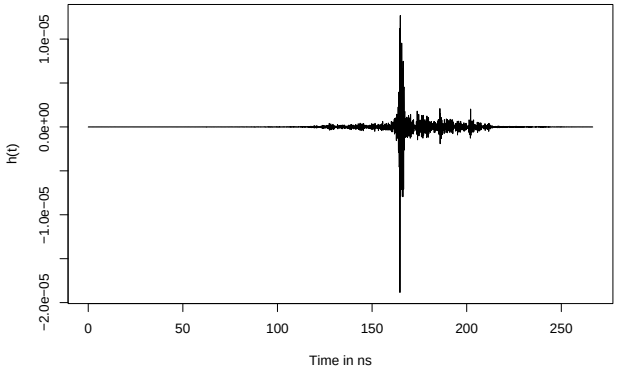


Figure 5.3.3: Filtered impulse response.

Absolute value of the frequency response

Figure 5.3.4 shows the absolute value of an observed impulse response. In chapter 6 the actual inference on data, will be done in the frequency domain. For this purpose the absolute value of the signal, can be used to get an idea of the size of the attenuation factors and noise levels, in the different data sets. The maximum values of the absolute value of the frequency response ranges from $4 \cdot 10^{-8}$ to $1.5 \cdot 10^{-3}$, and for the simulation based inference initial values and prior distributions should be chosen accordingly.

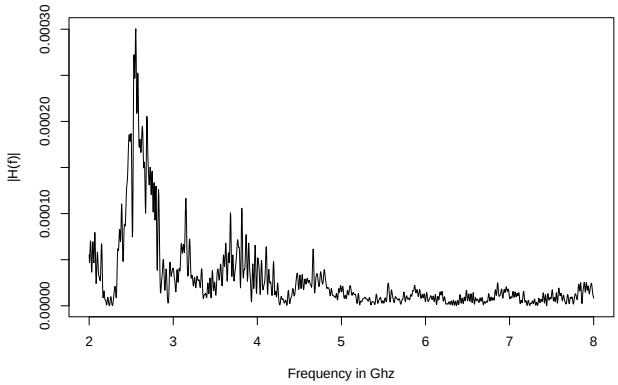


Figure 5.3.4: Absolute value of frequency response.

Example of a simulated data set

Figure 5.3.5 shows the impulse response for a model with $\tau = (100, 102, \ldots, 120)$ ns and $\beta = (11, 10, \ldots, 1)$. The impulse response model is assumed to be a weighted delta train with delays and weights given by τ and β. The model is then simulated as the corresponding frequency response, and sampled at the same frequencies as the measured data. As seen in the figure the weights are blurred by the sinc filter, but it is however possible to make out the arrivals.

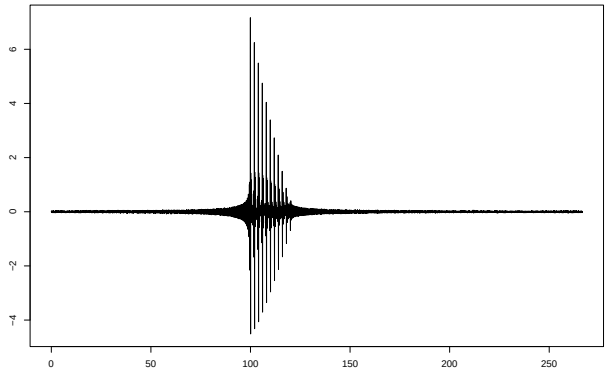


Figure 5.3.5: Simulated impulse response.

Chapter 6

Inference on data

In this chapter statistical inference based on the data described in chapter 5 is conducted using a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. The basic assumption is that data is sampled from a frequency response function given by

$$H(f) = \sum_{i=0}^{n-1} \beta_i \exp(-2\pi i \tau_i f), \quad f \in \mathbb{R}, \tag{6.0.1}$$

which is corrupted by additive complex Gaussian noise. In section 6.1 we develop an RJMCMC algorithm to estimate the standard deviation of the noise σ, the number of arrivals n, the arrival times $\tau = (\tau_0, \ldots, \tau_{n-1})$ and the corresponding attenuation factors $\beta = (\beta_0, \ldots, \beta_{n-1})$. In section 6.2 the algorithm is further developed to make statistical inference for the Turin model, and in section 6.3 the shot noise model described in section 2.2.3 is implemented.

6.1 The basic algorithm

As mentioned the channel noise is modeled as additive complex Gaussian, such that the observed data $\boldsymbol{y} = (y_1, \ldots, y_N)$ is given by

$$y_i = H(f_i) + n_i, \quad n_i \sim \mathbb{C}N(0, \sigma^2), \quad i = 1, \ldots, N,$$

where $\mathbb{C}N(\cdot, \cdot)$ is the univariate complex Gaussian distribution described in appendix A.3 and it is assumed that $n_1, \ldots, n_N$ are mutually independent.

Following the Bayesian approach we wish to calculate the posterior density of the parameters $\theta = (\tau, \beta, \sigma, n)$ given data y

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

The prior is assumed to factorize as

$$p(\theta) = p(\beta|n)p(\tau|n)p(n)p(\sigma) = p(n)p(\sigma)\prod_{i=0}^{n-1} p(\beta_i)p(\tau_i),$$

where it is implicitly assumed that β is independent of τ and that β and τ are iid. Since we in this section wish to make inference based only on the assumptions mentioned above, the priors are chosen non-informative as

$$\begin{aligned} \tau_i &\sim \text{Unif}(0, 266) \\ \beta_i &\sim \text{Unif}(0, 1) \\ \sigma &\sim \text{Unif}(0, 1) \\ n &\sim \text{Unif}(1, 1000). \end{aligned}$$

The frequency range of the the data corresponds to a time interval of 266ns which is the reason for choosing the prior for τ and since β describes attenuation the interval $(0, 1)$ seems reasonable. The measurement data typically has a magnitude of 10^{-5}, so the interval $(0, 1)$ should by far cover the needed range of the standard deviation of the noise. Finally the number of arrivals is limited to 1000, which again should be more than adequate for the number of arrivals in 266ns. Combining this prior with the complex Gaussian observation model yields the posterior

$$p(\theta|\boldsymbol{y}) \propto \sigma^{-2n} \exp\left(-\frac{1}{\sigma}\sum_{i=1}^{N} \|y_i - H(f_i)\|^2\right) \times \mathbf{I}(\tau \in [0, 266]^n)\mathbf{I}(\beta \in [0, 1]^n)\mathbf{I}(\sigma \in [0, 1])\mathbf{I}(1 \leq n \leq 1000).$$

This is simply the likelihood function truncated at the support of the prior. This implies that if the likelihood attains its maximum within this area then the MLE corresponds to the mode of the posterior.

6.1.1 RJMCMC algorithm

In this section an RJMCMC algorithm, having the posterior distribution of the parameters given data as invariant distribution is described based on the theory of section 4.3. A typical parameter configuration is denoted $\theta = (\tau, \beta, \sigma, n)$, where τ and β are vectors of length n, and thus the length of θ is $2n + 2$.

The algorithm consists of two parts:

1. A dimension change, which updates the number of arrivals n.
2. A sequential update of all other parameter values.

Dimension Change

The dimension change is chosen with probability p_{dim} and the parameter update is thus chosen with probability $1 - p_{\text{dim}}$. The dimension changing update consist of two moves, which are the birth or death of a parameter pair (τ', β'). Birth is chosen with probability p_b and death is chosen with probability $p_d = 1 - p_b$. The probability of choosing the birth move is thus $p_{\text{dim}}p_b$. Following the approach described in section 4.3 we draw the new parameter pair from a uniform distribution on $[0, 266] \times [0, 1]$, and set

$$\theta' = T(\theta, \tau', \beta') = \big((\tau, \tau'), (\beta, \beta'), \sigma, n + 1\big),$$

This satisfies the dimension matching criteria for the RJMCMC algorithm, and an easy calculation shows that $\left|\frac{\partial T}{\partial(\theta,\tau',\beta')}\right| = 1$.

The reverse move from θ' to θ is chosen with probability $p_{\text{dim}}p_d\frac{1}{n+1}$, since the parameter pair to be killed is chosen uniformly. The accept probability for the birth move is thus given by

$$\alpha_b(\theta, \theta') = \min\left\{1, \frac{p(\theta'|\boldsymbol{y})p_d\frac{1}{n+1}}{p(\theta|\boldsymbol{y})p_b p(\tau', \beta')}\right\}.$$

Similar arguments shows that the accept probability of a death move is given by

$$\alpha_d(\theta, \theta') = \min\left\{1, \frac{p(\theta'|\boldsymbol{y})p_b p(\tau', \beta')}{p(\theta|\boldsymbol{y})p_d \frac{1}{n}}\right\}.$$

Parameter updates

The updates of the parameter values is done by sequentially updating each entry of the parameter vector θ except n. The new values are proposed from a Gaussian distribution with the current parameter value as mean and a chosen variance. This update is then accepted with probability

$$\alpha_p(\theta, \theta') = \min\left\{1, \frac{p(\theta'|\boldsymbol{y})}{p(\theta|\boldsymbol{y})}\right\}.$$

Due to the choice of uniform priors all the prior densities cancel out in the calculations of the various accept probabilities, which means that they only depend on the likelihood ratio between the two parameter values. This is of course only true if both values are within the range where the priors are non zero, and any proposal outside this range is automatically rejected.

In the practical construction of MCMC algorithms the choice of proposal distributions is important in order to ensure proper convergence and mixing properties of the resulting Markov Chain. In this regard the variance of the Gaussian proposal is crucial for the performance of the algorithm. If the variance is chosen too small the accept rate is usually high and the chain will sample from a small area around the current value. On the other hand a large variance can result in proposals far away from the current value which are rarely accepted and the chain will have many samples with the same value. In Roberts et al. (1997) it is shown that given certain regularity conditions the optimal accept rate is 0.234. This has been accepted as a general guideline, even for algorithms where the conditions are not met, as is the case here. In Berthelsen and Møller (2004) it is advised to choose the proposal variance such that the accept rate is between 0.2 and 0.4. The latter has been used as a guideline for the algorithm.

This algorithm has been further developed to incorporate the Turin model for the impulse response function which is treated in the next section.

6.2 The Turin Model

To simplify the notation we now let $\tau = (\tau_1, \ldots, \tau_{n-1})$ and correspondingly $\beta = (\beta_1, \ldots, \beta_{n-1})$. In the Turin model it is assumed that the frequency response takes the form (6.0.1), where (τ_0, β_0) is the offset, that has to be estimated, and τ conditional on τ_0 consists of the points from a Poisson point process on (τ_0, ∞) contained in the observation interval $[0, 266]$ ns. Conditional on τ the attenuation factors are assumed to be independently Rayleigh distributed (see appendix A.4) with mean

$$\mathbb{E}(\beta_i|\tau_i) = \beta_0 \exp(-\alpha(\tau_i - \tau_0)), \quad i = 1, \ldots n-1.$$

This differs from the log-normal assumption in the original model proposed by Turin. The Rayleigh assumption is motivated by the results in Schuster and Bölcskei (2006),

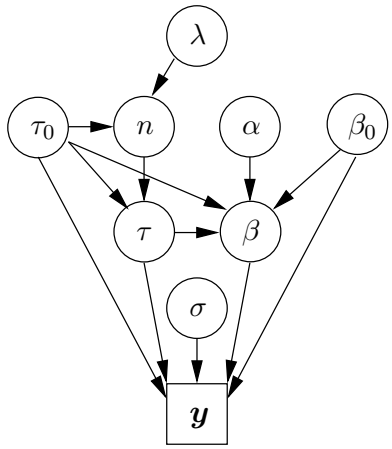


Figure 6.2.1: Graphical model representing the assumed conditional independence structure in the Turin model.

mentioned in section 1.2.5.

This model introduces two new parameters α and λ and new assumptions regarding the prior are made. The new parameter vector for which we wish to calculate the posterior distribution conditional on data, $\boldsymbol{y}$, is $\theta = (\tau_0, \tau, \beta_0, \beta, \lambda, \alpha, \sigma, n)$. In order to specify the prior, the assumed conditional independence relations of the model is illustrated in the graphical model in figure 6.2.1. The graph represents that conditional on the parents of a given variable X we assume that X is independent of its non-descendants. This is known as the directed local Markov property, and in section 3.2.2 of Lauritzen (1996) it is shown that this property holds if and only if the joint probability distribution allows a recursive factorization. This means that the joint probability distribution can be expressed as the product of distributions of the variables given their parents, and we obtain the following factorization of the prior

$$p(\theta) = p(\beta|\tau_0, \tau, \beta_0, \alpha)p(\tau|\tau_0, n)p(n|\tau_0, \lambda)p(\lambda)p(\alpha)p(\sigma)p(\tau_0)p(\beta_0). \tag{6.2.1}$$

Furthermore the assumed conditional independence implies that

$$p(\boldsymbol{y}|\theta) = p(\boldsymbol{y}|\tau_0, \tau, \beta_0, \beta, \sigma).$$

The assumption of a priori independence between τ_0 and β_0 might seem too simplistic, but due to lack of any prior knowledge of the structure of this dependence it is left out of the model. In general it is attempted to use non-informative priors for the parameters that are not affected by the Poisson and Rayleigh assumptions of the model, whereas the priors of the rest of the parameters are chosen according to the model assumptions. This leads to the following list of priors (letting $\text{Order}_n([a, b])$

denote the distribution of the order statistic for n uniform random variables on $[a, b]$):

$$\begin{aligned}
\alpha &\sim \text{Unif}(0, 1000) \\
\lambda &\sim \text{Unif}(0, 1000) \\
\tau_0 &\sim \text{Unif}(0, 266) \\
\beta_0 &\sim \text{Unif}(0, 1) \\
\sigma &\sim \text{Unif}(0, 1) \\
n|\lambda, \tau_0 &\sim \text{Poisson}\Big(\lambda(266 - \tau_0)\Big) \\
\tau|\tau_0, n &\sim \text{Order}_{n-1}\Big([\tau_0, 266]\Big) \\
\beta_i|\tau_0, \tau, \beta_0, \alpha &\sim \text{Rayleigh}\Big((\pi/2)^{-\frac{1}{2}}\beta_0 \exp(-\alpha(\tau_i - \tau_0))\Big), \quad i = 1, \ldots, n-1, \quad \text{iid.}
\end{aligned}$$

To specify these priors it has been used that the number of points of a homogeneous Poisson point process with intensity λ on an interval (a, b) is $\text{Poisson}\big(\lambda(b-a)\big)$ and conditional on the number of arrivals n in the interval the points follow the distribution of the order statistic of n uniformly distributed stochastic variables as shown in proposition 2.1.17.

In order to simplify the expression for the prior we split the parameter vector in two parts $\theta = (\theta_1, \theta_2)$, where $\theta_1 = (\alpha, \lambda, \tau_0, \beta_0, \sigma)$ and $\theta_2 = (\tau, \beta, n)$. Using that if $X \sim \text{Unif}(0, a)$ then $\frac{X}{a} \sim \text{Unif}(0, 1)$, we have

$$p(\theta_1) \propto \mathbf{I}\big((\tfrac{\alpha}{1000}, \tfrac{\lambda}{1000}, \tfrac{\tau_0}{266}, \beta_0, \sigma) \in (0, 1)^5\big).$$

Furthermore we have that

$$\begin{aligned}
p(\theta_2|\theta_1) &= \frac{\big(\lambda(266 - \tau_0)\big)^n}{n!} \exp(-\lambda(266 - \tau_0)) \frac{n!}{(266 - \tau_0)^n} \prod_{i=1}^{n-1} \frac{\beta_i}{s_i^2} \exp\left(-\frac{\beta_i^2}{2s_i^2}\right) \\
&= \lambda^n \exp\left(-\lambda(266 - \tau_0) - \sum_{i=1}^{n-1} \frac{\beta_i^2}{2s_i^2}\right) \prod_{i=1}^{n-1} \frac{\beta_i}{s_i^2},
\end{aligned}$$

with the Rayleigh parameter being given by

$$s_i = (\pi/2)^{-\frac{1}{2}} \mathbb{E}(\beta_i|\tau_i) = (\pi/2)^{-\frac{1}{2}} \beta_0 \exp(-\alpha(\tau_i - \tau_0)), \quad i = 1, \ldots, n-1.$$

Using this prior with the complex Gaussian observation model leads to the posterior distribution

$$\begin{aligned}
p(\theta|\boldsymbol{y}) \propto \left(\frac{\lambda}{\sigma^2}\right)^n \exp\left(-\frac{1}{\sigma} \sum_{i=1}^{N} \|y_i - H(f_i)\|^2 - \lambda(266 - \tau_0) - \sum_{i=1}^{n-1} \frac{\beta_i^2}{2s_i^2} \right) \\
\times \prod_{i=1}^{n-1} \frac{\beta_i}{s_i^2} \mathbf{I}\big[(\tfrac{\alpha}{1000}, \tfrac{\lambda}{1000}, \tfrac{\tau_0}{266}, \beta_0, \sigma) \in (0, 1)^5\big].
\end{aligned}$$

6.2.1 RJMCMC algorithm

In order to describe the properties of the posterior and calculate estimates for the parameters in the Turin model the RJMCMC algorithm is modified according to the

above assumptions. The two new parameters are updated along with the rest of the parameters in the sequential update, but it is important to notice that due to the new model assumptions this is no longer just a likelihood ratio. Furthermore the factorization according to the graph in figure 6.2.1 should be exploited to limit calculation time. E.g. when the current parameter to be updated is λ, then λ' is drawn from $N(\lambda, \phi^2)$, and the fraction in the accept probability reduces to

$$\frac{p(\boldsymbol{y}|\theta')p(\theta')}{p(\boldsymbol{y}|\theta)p(\theta)} = \frac{p(\theta')}{p(\theta)} = \left(\frac{\lambda'}{\lambda}\right)^n \exp\Big((\lambda - \lambda')(266 - \tau_0)\Big).$$

This requires practically no calculation time compared to the original fraction of the posteriors, where the likelihood has to be evaluated. Evaluating the likelihood at every update is very time consuming in our case, since this requires the function (6.0.1) to be evaluated at 1601 data points.

Finally before presenting the inference based on these algorithms we modify the algorithm to incorporate another model.

6.3 Shot noise model

In this section the statistical inference for the shot noise model described in section 2.2.3 is considered. The impulse response is assumed to be given by (2.2.4) and the frequency response is given by (2.2.6). We wish to use this model on the observed data in order to model clustering effects in the impulse response function. The inference based on this model only requires minor adjustments to the Turin model. The unknown parameter vector is now given as

$$\theta = (a, b, \alpha, \lambda, \gamma, \tau_0, \beta_0, \sigma, \tau, \beta, n),$$

where the only new parameter with direct influence on the model is the decay of the clusters γ. As mentioned in section 2.2.3 these must satisfy $\mathbb{E}(\gamma_i^{-1}) < \infty$ for $i = 0, \ldots, n$. This is done by selecting a gamma prior with shape parameter $a > 1$ and rate $b > 0$, since we then have

$$\begin{aligned}
\mathbb{E}(\gamma^{-1}) &= \int_{\mathbb{R}_+} \gamma^{-1} \frac{b^a \exp(-b\gamma)}{\Gamma(a)} \gamma^{a-1} d\gamma \\
&= \frac{\Gamma(a-1)b^a}{\Gamma(a)b^{a-1}} \int_{\mathbb{R}_+} \frac{b^{a-1}\exp(-b\gamma)}{\Gamma(a-1)} \gamma^{a-2} d\gamma \\
&= \frac{\Gamma(a-1)b^a}{\Gamma(a)b^{a-1}} < \infty.
\end{aligned}$$

Since we have no information on the decay rates we give uninformative improper hyperpriors on the shape and rate parameters such that

$$\begin{aligned}
\gamma_i &\sim \text{Gamma}(a, b), \quad i = 0, \ldots, n, \quad \text{iid.} \\
p(a) &\propto \mathbf{I}(a > 1) \\
p(b) &\propto \mathbf{I}(b > 0).
\end{aligned}$$

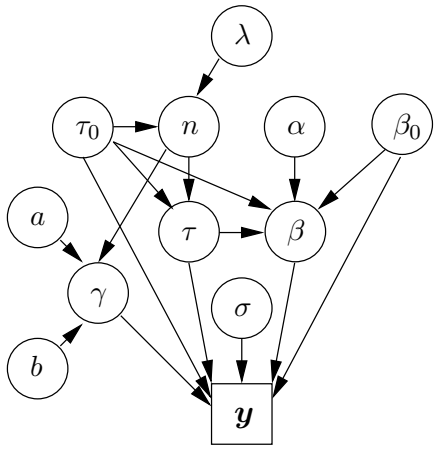


Figure 6.3.1: Graphical model representing the assumed conditional independence structure in the shot noise model.

We let the rest of the priors be given as in section 6.2. The conditional independence structure for this model is displayed in figure 6.3.1.

Combining this information with the complex Gaussian observation model and the results from section 6.2 leads to the posterior distribution

$$p(\theta|\boldsymbol{y}) \propto \left(\frac{\lambda b^a \gamma^{a-1}}{\sigma^2}\right)^n \exp\left(-\frac{1}{\sigma}\sum_{i=1}^{N}\|y_i - H(f_i)\|^2 - \lambda(266-\tau_0) - nb\gamma - \sum_{i=1}^{n-1}\frac{\beta_i^2}{2s_i^2}\right) \times \prod_{i=1}^{n-1}\frac{\beta_i}{s_i^2}\mathbf{I}\big[(\tfrac{\alpha}{1000}, \tfrac{\lambda}{1000}, \tfrac{\tau_0}{266}, \beta_0, \sigma) \in (0,1)^5\big].$$

6.3.1 RJMCMC algorithm

The RJMCMC algorithm for this model is quite similar to the one for the Turin model. The major difference is the addition of the decay rate γ to the dimension change. A birth move thus consists of drawing a parameter vector (τ', β', γ') uniformly on $[0, 266] \times [0, 1] \times [0, 5]$ and setting

$$\theta' = T(\theta, \tau', \beta', \gamma') = (a, b, \alpha, \lambda, (\gamma, \gamma'), \tau_0, \beta_0, \sigma, (\tau, \tau'), (\beta, \beta'), n).$$

This mapping also has $\left|\frac{\partial T}{\partial(\theta,\tau',\beta',\gamma')}\right| = 1$, and the acceptance probabilities for the dimension changes follow as in section 6.1.1. Finally the new parameters a, b, γ should be included in the sequential update, and factorizations according to figure 6.3.1 should be taken into consideration in order to reduce calculation time.

6.4 Artificial Data Analysis

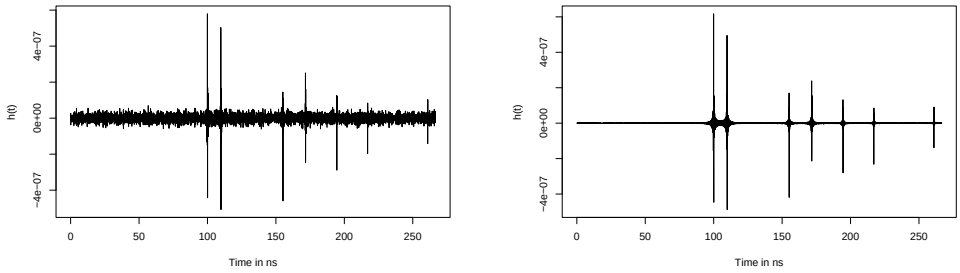
To test and calibrate the algorithms described in sections 6.1, 6.2 and 6.3, a collection of artificial datasets based on the models in question were created. In the following

sections this is described for the basic and Turin model.

Since the basic algorithm only identifies arrivals and attenuation factors it also works on data constructed by the Turin model and a separate data set was not necessary. To simulate two artificial data sets from the Turin model the following parameters were chosen:

$$\alpha = 0.01, \quad \lambda = 0.05, \quad \tau_0 = 100, \quad \beta_0 = 10^{-6}.$$

From these values a single frequency response function as given by the Turin model was simulated, and two data sets were obtained by adding complex Gaussian noise with respectively $\sigma_L = 1.414 \cdot 10^{-6}$ and $\sigma_S = 0.71 \cdot 10^{-7}$. The simulation resulted in 7 arrivals on the entire interval and the two data sets are shown in the time domain in figure 6.4.1. Since the actual data was simulated in the frequency domain, the IFFT was used to obtain the time domain data.



(a) $\sigma_L = 1.141 \cdot 10^{-6}$ (b) $\sigma_S = 0.71 \cdot 10^{-7}$

Figure 6.4.1: Simulated data sets.

The artificial data was implemented for inference in the RJMCMC algorithms developed in section 6.1 and 6.2. Since the parameter space is multi-dimensional and of varying dimension, determining the burn-in time of the chains is a difficult task. As a measure of how well the estimated signal fits the data we choose to trace the residual sum of squares $\{RSS\}_{j \in \mathbb{N}_0}$, where

$$RSS_j = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{H}^{(j)}(f_i))^2,$$

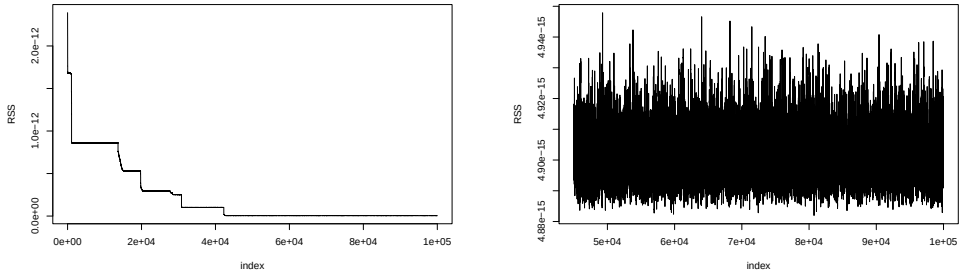
and $\hat{H}^{(j)}$ is the estimated frequency response at the j'th iteration, based on $\tau^{(j)}$ and $\beta^{(j)}$. The reason for this choice is that the RSS provides a measure of how well the estimated model fits the data. Combined with traces of the number of parameters, n, we will then use the convergence of this sequence as an indicator of the burn-in of the underlying Markov chain.

6.4.1 The basic algorithm

The simulated datasets were implemented in the algorithm for the basic model, and burn-in was estimated by visual inspection of traces of RSS and n, as displayed in figure 6.4.2, 6.4.3 and 6.4.4. It is seen the burn-in time for the dataset with σ_L was longer, indicating that larger noise, makes it harder for the underlying chain to converge.

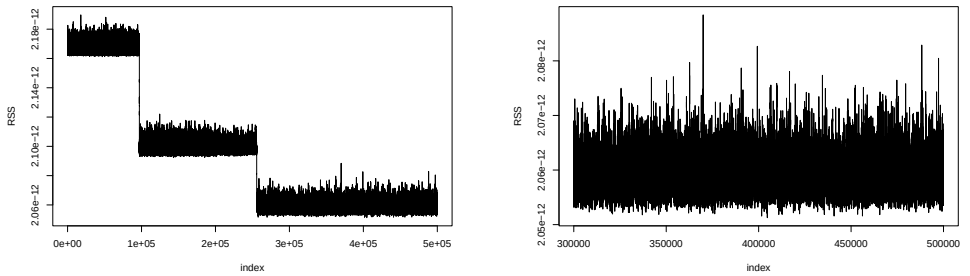
The results for the basic algorithm are displayed in table 6.4.1 and 6.4.2. It is seen that the estimated values closely approximate the true values, and these were all contained in the PCI indicating that the algorithm works as intended. It is worth noting that RSS is an MLE for the variance, and as seen from the results, this coincides with the posterior mean of the standard deviation.

As expected there is a higher uncertainty for the dataset with σ_L, and a few times the algorithm has accepted an extra arrival, the length of the PCI's are however still small compared to the mean. It is also noteworthy that the basic algorithm gives a very precise estimate on the number of parameters, regardless of the noise level.



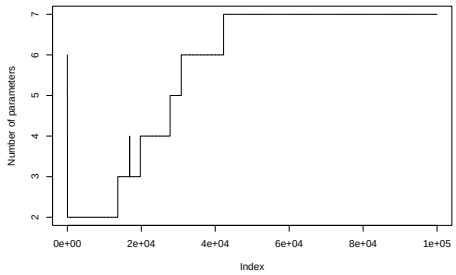
(a) With burn-in. (b) Without Burn-in.

Figure 6.4.2: Traces of RSS for the basic algorithm on data with σ_S.

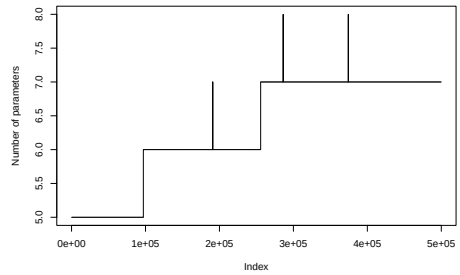


(a) With burn-in. (b) Without Burn-in.

Figure 6.4.3: Traces of RSS for the basic algorithm on data with σ_L.



(a) Number of parameters for the basic algorithm on data with σ_S.



(b) Number of parameters for the basic algorithm on data with σ_L.

Figure 6.4.4: Traces of n.

Data with σ_S				
	True value	Mean	$q_{2,5\%}$	$q_{97,5\%}$
τ_0	100.0000	100.0000	99.9999	100.0001
τ_1	109.8178	109.8178	109.8177	109.8179
τ_2	155.1446	155.1447	155.1445	155.1448
τ_3	171.6717	171.6719	171.6717	171.6721
τ_4	194.5243	194.5243	194.5241	194.5245
τ_5	217.0126	217.0125	217.0123	217.0128
τ_6	260.9388	260.9388	260.9384	260.9392
β_0	10	10.032	10.01	10.06
β_1	9.064	9.055	9.03	9.08
β_2	5.761	5.733	5.71	5.76
β_3	4.883	4.884	4.86	4.91
β_4	3.885	3.874	3.85	3.90
β_5	3.103	3.100	3.08	3.13
β_6	2.000	2.005	1.98	2.02
σ_S	0.71	0.70	0.68	0.72
n	7	7	7	7
RSS	N/A	$4.902 \cdot 10^{-15}$	$4.886 \cdot 10^{-15}$	$4.917 \cdot 10^{-15}$

Table 6.4.1: Results after burn-in from the basic RJMCMC algorithm on the artificial dataset. Scale on β and σ_S is 10^{-7}.

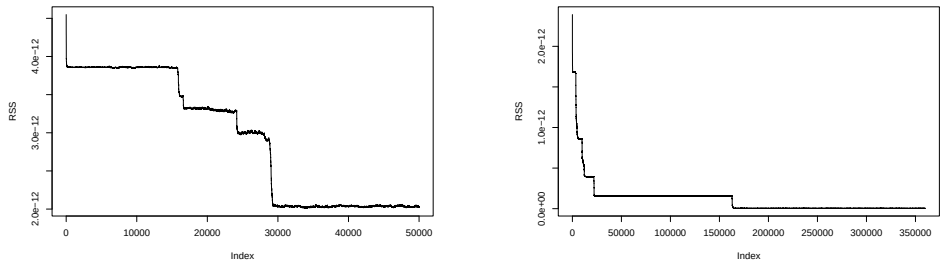
Data with σ_L				
	True value	Mean	$q_{2,5\%}$	$q_{97,5\%}$
τ_0	100.0000	100.0008	99.9992	100.0023
τ_1	109.8178	109.8180	109.8163	109.8196
τ_2	155.1446	155.1435	155.1410	155.1459
τ_3	171.6717	171.6708	171.6680	171.6736
τ_4	194.5243	194.5245	194.5208	194.5283
τ_5	217.0126	217.0133	217.0078	217.0189
τ_6	260.9388	260.9420	260.9348	260.9490
β_0	10	9.693	9.2	10.18
β_1	9.064	9.307	8.81	9.80
β_2	5.761	6.165	5.67	6.66
β_3	4.883	5.357	4.86	5.86
β_4	3.885	3.977	3.48	4.48
β_5	3.103	2.614	2.10	3.12
β_6	2.000	2.022	1.51	2.51
σ_L	14.14	14.35	14	14.7
n	7	7	7	7
RSS	N/A	$2.059 \cdot 10^{-12}$	$2.054 \cdot 10^{-12}$	$2.067 \cdot 10^{-12}$

Table 6.4.2: Results after burn-in from the basic RJMCMC algorithm on the artificial dataset. Scale on β and σ_L is 10^{-7}.

6.4.2 The Turin algorithm

The algorithm from section 6.2 was also implemented for statistical inference on the two simulated datasets. Burn-in was again estimated by visual inspection of the trace of RSS, as displayed in figure 6.4.5. Results for the parameters of interest are summarized in table 6.4.3. As seen the algorithm closely approximates the true values for the dataset with σ_S. For the dataset with σ_L, the algorithm however gives a large overestimation on the number of parameters. We suspect this to be a consequence of the model assumption with exponentially decaying attenuation factors, which makes it more likely to accept small values of β for late arrival times. Inspection of the estimated values of τ and β shows that the actual values are estimated properly, but a number of extra arrivals are added. The amplitudes of these arrivals are however so small that they are hidden by the noise, but due their nice fit to the model assumption the algorithm is not inclined to remove them. This is also seen on figure 6.4.6(b) which shows a higher concentration of arrivals later in the interval. As seen from table 6.4.3 this also causes an overestimation on the arrival rate λ in the case of σ_L.

By comparing RSS for the basic and Turin model it is seen, that they provide a similar fit for the case with σ_S. In the case of σ_L the Turin model however gives a large overestimation on the number of arrivals, without giving a significant improvement to the fit, which indicates a problem with this model.



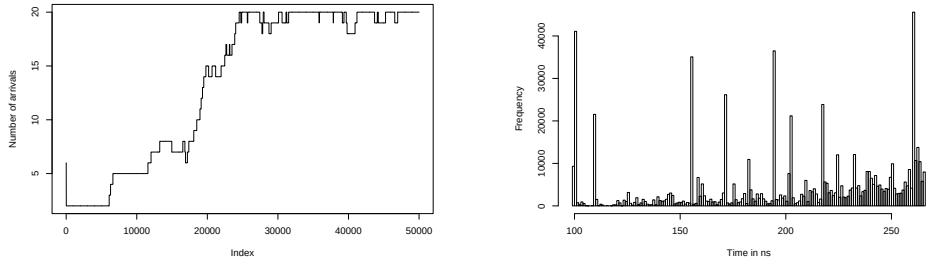
(a) Trace of RSS for the algorithm on data with σ_L.

(b) Trace of RSS for the algorithm on data with σ_S

Figure 6.4.5: Traces of RSS.

	True Value	Mean	$q_{2,5\%}$	$q_{97,5\%}$
Data with σ_S				
τ_0	100	100	99.9999	100.0001
β_0	10^{-6}	$1.003 \cdot 10^{-6}$	$1.001 \cdot 10^{-6}$	$1.006 \cdot 10^{-6}$
n	7	7	7	7
α	0.01	0.0107	0.0063	0.0141
λ	0.05	0.0481	0.0207	0.0875
σ_S	$0.71 \cdot 10^{-7}$	$0.7003 \cdot 10^{-7}$	$6.83 \cdot 10^{-7}$	$7.18 \cdot 10^{-7}$
RSS	N/A	$4.901 \cdot 10^{-15}$	$4.888 \cdot 10^{-15}$	$4.921 \cdot 10^{-15}$
Data with σ_L				
τ_0	100	100.0008	99.9993	100.0023
β_0	10^{-6}	$0.968 \cdot 10^{-6}$	$0.92 \cdot 10^{-6}$	$1.02 \cdot 10^{-6}$
n	7	19.71	18	20
α	0.01	0.0165	0.014	0.019
λ	0.05	0.124	0.071	0.178
σ_L	$1.414 \cdot 10^{-6}$	$1.425 \cdot 10^{-6}$	$1.399 \cdot 10^{-6}$	$1.450 \cdot 10^{-6}$
RSS	N/A	$2.035 \cdot 10^{-12}$	$2.019 \cdot 10^{-12}$	$2.05 \cdot 10^{-12}$

Table 6.4.3: Results from the Turin model on artificial data.



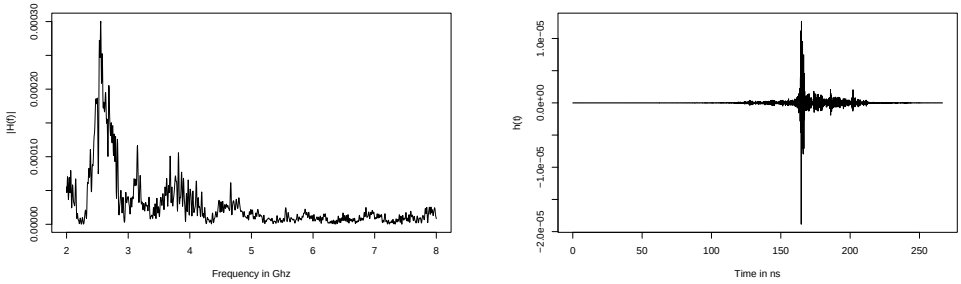
(a) Trace of the number of arrivals for the Turin algorithm on data σ_L.

(b) Histogram of arrival times for the Turin algorithm with σ_L.

Figure 6.4.6: Plots for the Turin algorithm on data with σ_L.

6.5 Analysis of real data

The inference for real data is based on observations from a townhouse with a transmitter-receiver distance of 2.44 m and free line of sight passage. Both the absolute value of the frequency response and the time domain data obtained using IFFT is shown in figure 6.5.1.



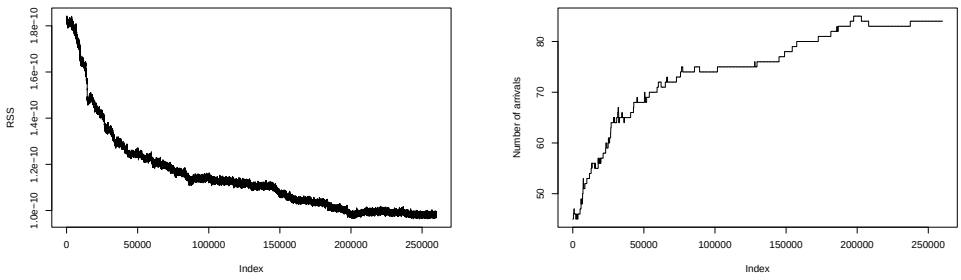
(a) Observed frequency response.

(b) Observed impulse response.

Figure 6.5.1: Plots of data.

6.5.1 The basic algorithm

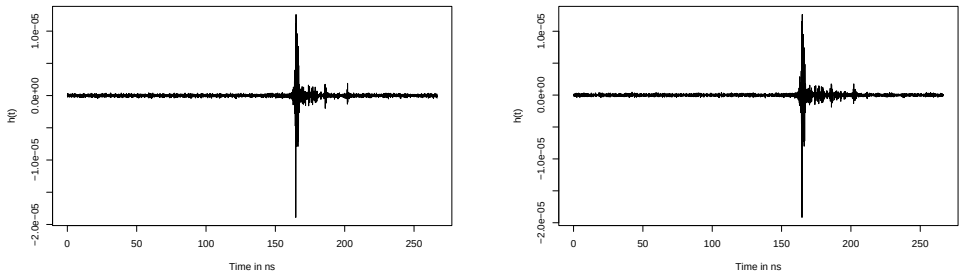
The dataset was implemented in the basic algorithm, and burn-in was estimated by visual inspection of the traces of RSS and n, as displayed in figure 6.5.2. Results from the algorithm are summarized in table 6.5.1. Comparing with the Turin model, we see that the basic algorithm estimates fewer arrivals and has a higher RSS. As seen on figure 6.5.3(a), the estimated values are however sufficient to recreate the observed signal.



(a) Trace of RSS for the basic algorithm.

(b) Trace of n for the basic algorithm.

Figure 6.5.2: Traces of RSS and n from the basic algorithm.



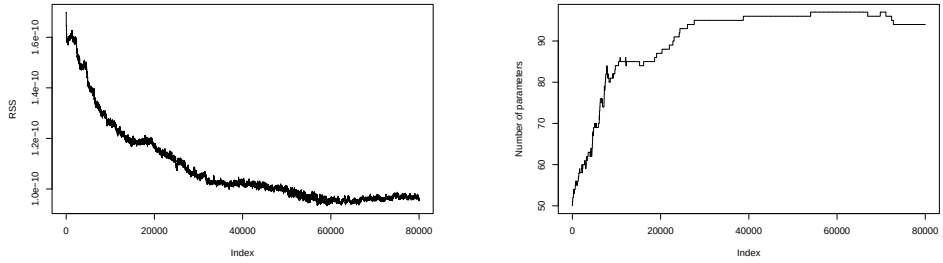
(a) Impulse response generated from values estimated by the basic algorithm.

(b) Impulse response generated from values estimated by the Turin algorithm.

Figure 6.5.3: Simulated impulse responses.

6.5.2 The Turin model

The dataset was implemented in the algorithm for the Turin model, and the burn-in was estimated by visual inspection of the traces of RSS and n as shown in figure 6.5.4. The results from the algorithm are summarized in table 6.5.1.



(a) Trace of RSS for the Turin algorithm on real data.

(b) Trace of n for the Turin algorithm on real data.

Figure 6.5.4: Traces of RSS and n

To compare the model with the observed data, we simulate 9 datasets based on the estimated values. These are displayed in figure 6.5.5 and 6.5.6. It is clear that the plots in the frequency domain in no way resembles the observed frequency response in figure 6.5.1(a), and that the plots for the time domain only to a small degree resembles the observed impulse response in figure 6.5.1(b). This implies that the model does not efficiently describe the physical phenomenon. If we simulate another 1000 frequency responses and consider their sums of squares, as displayed in figure 6.5.7 and compare this with the observed sum of squares $SS(y) = 4.122 \cdot 10^{-6}$ indicated by the solid circle, we see that the overall magnitude of the signal also is not properly described by the model. The estimated values of τ and β which for this model may be considered auxiliary variables, do however provide a qualitative similar reconstruction of data as seen in figure 6.5.3(b)

Turin model			
	Mean	$q_{2.5\%}$	$q_{97.5\%}$
τ_0	163.09108	163.0871	163.0948
β_0	$6.937 \cdot 10^{-6}$	$6.628 \cdot 10^{-7}$	$7.217 \cdot 10^{-7}$
n	94.62	94	97
RSS	$9.666 \cdot 10^{-11}$	$9.543 \cdot 10^{-11}$	$9.8 \cdot 10^{-11}$
α	0.036	0.0298	0.0407
λ	0.93	0.7514	1.116
σ	$9.827 \cdot 10^{-6}$	$9.607 \cdot 10^{-6}$	$10.09 \cdot 10^{-6}$
Basic Model			
τ_0	162.7954	162.7838	162.8068
β_0	$1.099 \cdot 10^{-6}$	$0.729 \cdot 10^{-6}$	$1.45 \cdot 10^{-6}$
n	83.57	83	84
RSS	$9.87 \cdot 10^{-11}$	$9.72 \cdot 10^{-11}$	$10.0 \cdot 10^{-11}$
σ	$9.93 \cdot 10^{-6}$	$9.66 \cdot 10^{-6}$	$10.20 \cdot 10^{-6}$

Table 6.5.1: Results for the Turin and basic model on real data.

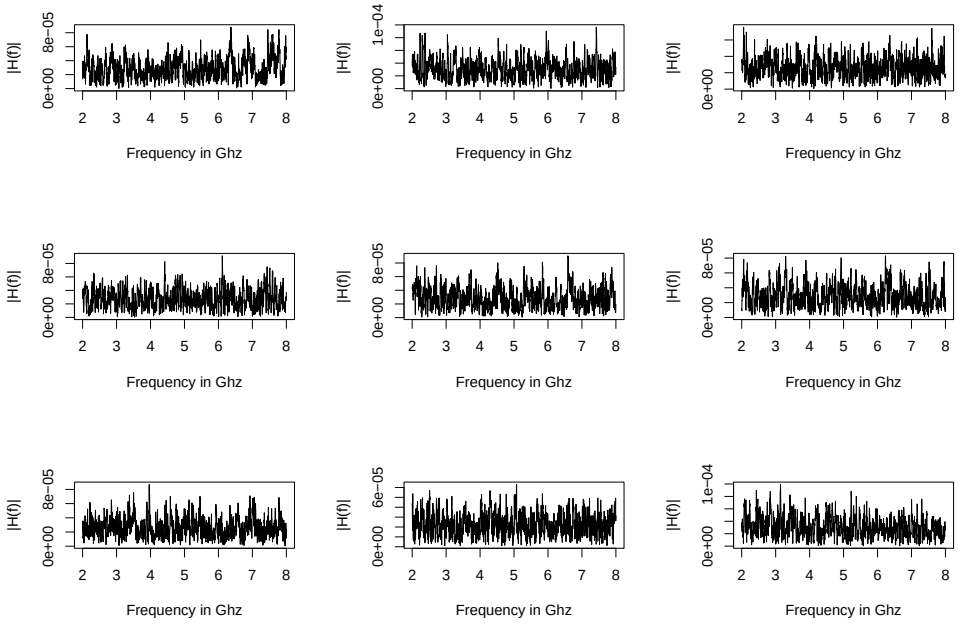


Figure 6.5.5: Simulated frequency responses.

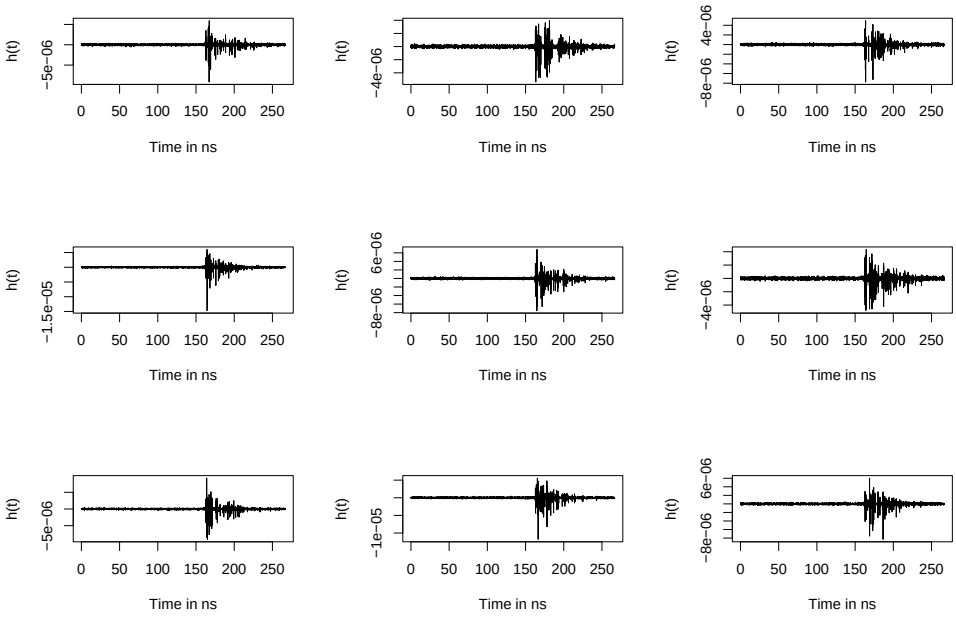
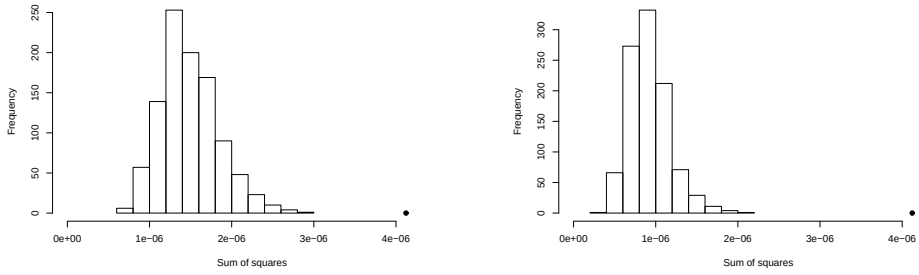


Figure 6.5.6: Simulated impulse responses.



(a) Histogram of sum of squares for simulated frequency responses from the Turin model.

(b) Histogram of sum of squares for simulated frequency responses from the shot noise model.

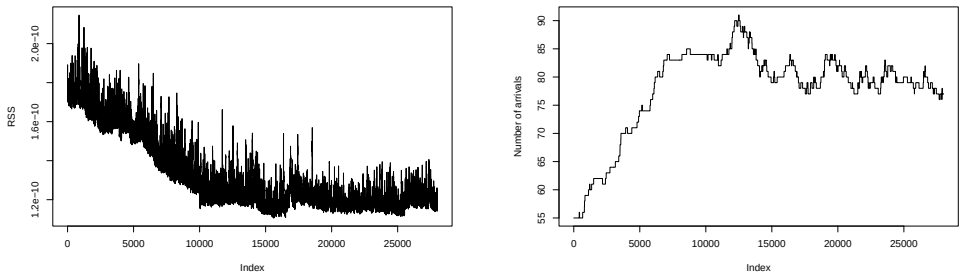
Figure 6.5.7: Histograms of sum of squares.

Parameter	Mean	$q_{2.5\%}$	$q_{97.5\%}$
τ_0	163.81	163.805	163.815
β_0	$1.61 \cdot 10^{-4}$	$1.50 \cdot 10^{-4}$	$1.72 \cdot 10^{-4}$
n	79.59	77	84
RSS	$1.197 \cdot 10^{-10}$	$1.132 \cdot 10^{-10}$	$1.323 \cdot 10^{-10}$
α	0.049	0.043	0.054
λ	0.788	0.625	0.969
σ	$1.09 \cdot 10^{-5}$	$1.06 \cdot 10^{-5}$	$1.15 \cdot 10^{-5}$
a/b	0.34	0.24	0.44
a/b^2	0.0036	0.0022	0.0051

Table 6.5.2: Results for the shot noise model on real data.

6.5.3 The shot noise model

The data was also implemented in the algorithm for the shot noise model. Burn-in time is once again obtained from visual inspection of the RSS as displayed in figure 6.5.8(a). Results for the algorithm are summarized in table 6.5.2. The number of



(a) Trace of RSS for shotnoise algorithm on real data

(b) Trace of n for shotnoise algorithm on real data.

Figure 6.5.8: Traces of RSS and n for the shotnoise algorithm on real data.

arrivals, n, is estimated at about 80 which is in between the estimate of the basic and the Turin model. The RSS of the fitted signal is however higher than in these models. Futhermore the model has a decay rate parameter γ for each arrival, but as mentioned these extra parameters do not result in a better fit. The mean a/b and variance a/b^2 of the prior gamma distribution indicate that only small values of γ fit the model and this is confirmed by the histogram of all the accepted values of γ after burn-in shown in figure figure 6.5.9.

The purpose of this model was to capture some of the clustering effect in the data and in this way reduce the number of arrivals needed to fit the data. This did not succeed. As in the Turin case the estimated parameters were used to simulate frequency responses and these realizations were similar to the ones from the Turin model displayed in figure 6.5.5. The predictive capability of the shot noise model is checked as in the Turin model by plotting the observed sum of squares $SS(y) = 4.122 \cdot 10^{-6}$

in the histogram of the sum of squares for 1000 realizations from the estimated shot noise model. As seen in figure 6.5.7(b) the observed sum of squares is considerably larger than the all the simulated, leading to the conclusion that simulation from the estimated model does not produce data sets of the same magnitude as the observed.

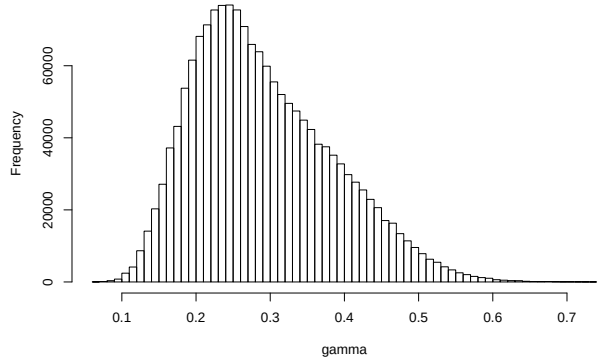


Figure 6.5.9: Histogram of accepted values of γ.

6.6 Summary

The basic algorithm does a good job of fitting a weighted delta train, to the data. Using these estimated values, makes it possible to recreate the observed signal. By implementing the Turin model, the algorithm still gives good estimates on τ and β, from which it is possible to recreate the signal. By fitting a marked Poisson process, with exponentially decaying marks, the model however fails to properly describe data. This is most significant for the attenuation factors, since the model fails to generate large values at the start of the signal as seen in figure 6.5.1(b). We suspect this to be because the model on average requires the first arrival to have the largest amplitude, which is not the case in the data at hand. A possible correction of this is by modelling the mean of the attenuation factors as $\mathbb{E}(\beta_i|\tau, B, \alpha) = B\exp\big(-\alpha(\tau_i - \tau_0)\big)$ for $i \in \mathbb{N}_0$, where B is a parameter to be determined from measured data.

The shot noise model does not provide any improvement compared to the former models, since the high number of parameters do not provide the desired model reduction. Furthermore the higher number of parameters, do not supply a better reconstruction of data, measured by the residual sum of squares.

Chapter 7

Conclusion and Further developments

The main problem in this thesis was to describe previously suggested stochastic models of an impulse response function in a strict mathematical context, and implement these for statistical inference using numerical Bayesian methods.

Using the theory of point processes, described as random measures, we reformulated the models from Saleh and Valenzuela (1987) and Turin et al. (1972), and showed that for exponentially or polynomially decaying amplitudes, these leads to almost surely finite impulse responses. For proper choices of the parameters this guarantees that the physical interpretation of the models is well founded, since the channel doesn't increase the strength of the signal.

The statistical inference in section 6.2 showed that the Turin model does not sufficiently describe the characteristics of measured data, which leads to the conclusion that a Poisson arrival model is insufficient to describe multipath propagation. This is consistent with the conclusion of Turin et al. (1972) and the work in Saleh and Valenzuela (1987) and Suzuki (1977), where extensions to the homogeneous Poisson arrival assumption are investigated. A future prospect of the results in this thesis, is thus implementation of e.g. the Saleh Valenzuelah or Molisch et al. model in an RJMCMC algorithm.

In section 6.3 we investigated the possibility of describing the cluster effect suggested by Saleh and Valenzuela (1987) as a sum of decaying exponential functions instead of using discrete arrivals. Inference showed that this model did not provide a satisfactory fit for the measured data.

Results from section 6.1 are however promising, since the basic algorithm gives an automated method of estimating the model order for the impulse response function. The results are of course insufficient in regards of simulating new impulse response functions, since they only give a reconstruction of the measured data. By using the RJMCMC algorithm to estimate arrival times and attenuation factors, it may however be possible to extract a new dataset, which could be used to do statistical inference in e.g. a pure point process setup. An introduction to how this might be done is found in Møller and Waagepetersen (2004).

Part III

Appendix

Appendix A

Miscellaneous results

A.1 Polish spaces

Most of the results in this thesis are confined to Polish spaces which are defined as

Definition A.1.1 (Polish space)
A metric space (S, d) is called a Polish space if it is complete and separable. Meaning respectively that all Cauchy sequences converge in S and S has a countable dense subset. □

A classical example of a Polish space is $\mathbb{R}^n$ with the metric being the Euclidean distance $\|\cdot\|$ since all Cauchy sequences converge in $\mathbb{R}^n$ with this metric and $\mathbb{Q}^n$ is a countable dense subset ($\mathbb{Q}$ denotes the rationals). A famous result by Kolmogorov shown for Polish spaces is the main reason for restricting the treatment of point processes to these spaces (Daley and Vere-Jones (1988)).

Theorem A.1.2 (Kolmogorov Extension Theorem)
Let $\mathscr{T}$ be any arbitrary index set, and for $t \in \mathscr{T}$ suppose $(S_t, \mathscr{F}_t)$ is a polish space with its associated σ-algebra. Suppose further that for each finite subfamily $(\sigma) = \{t_1, \ldots, t_n\}$ of indices from $\mathscr{T}$, there is given a probability measure $\pi_{(\sigma)}$ on $\mathscr{F}_{(\sigma)} = \mathscr{F}_{t_1} \otimes \cdots \otimes \mathscr{F}_{t_n}$. In order that there exist a measure π on $\mathscr{F}_\infty$ such that for all (σ),$\pi_{(\sigma)}$ is the projection of π onto $\mathscr{F}_{(\sigma)}$, it is necessary and sufficient that for all $(\sigma), (\sigma_1), (\sigma_2)$

(i) $\pi_{(\sigma)}$ depends only on the choice of indices in (σ), not on the order in which they are written down

(ii) if $(\sigma_1) \subseteq (\sigma_2)$, then $\pi_{(\sigma_1)}$ is the projection of $\pi_{(\sigma_2)}$ onto (σ_1).

□

For a proof of this theorem we refer to Billingsley (1995).

Another result for Polish spaces used in this thesis is given below.

Definition A.1.3 (Dissecting System)
The sequence $\mathcal{A} = \{\mathcal{A}_n\}$ of finite partitions $\mathcal{A}_n = \{A_{ni} \mid i = 1, \ldots, k_n\}, n = 1, 2, \ldots$ consisting of Borel sets in the space S is a dissecting system for S when

(i) $A_{ni} \cap A_{nj} = \emptyset$ for $i \neq j$ and $A_{n1} \cup \cdots \cup A_{nk_n} = S$

(ii) $A_{n-1,i} \cap A_{nj} = A_{nj}$ or $\emptyset$

(iii) Given distinct $x, y \in S$ there exists an integer $n = n(x, y)$ such that $x \in A_{ni}$ implies $y \notin A_{ni}$.

□

Proposition A.1.4
Any Polish space (S, d) contains a dissecting system. □

For the proof of this proposition we refer to Daley and Vere-Jones (1988).

A.2 Measure theoretical results

A basic knowledge of measure theory is presumed, but some of the results used throughout the thesis are summarized here. First a lemma regarding the uniqueness of a measure once it is know on a generator which is closed under intersection.

Lemma A.2.1
Let μ_1 and μ_2 be two measures defined on a space Ω equipped with a σ-algebra $\mathcal{F} = \sigma(\mathcal{A})$, for a paving $\mathcal{A}$ which is closed under intersection. If $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ and $\mu_1 = \mu_2$ on $\mathcal{A}$, then $\mu_1 = \mu_2$ on $\mathcal{F}$.

Usually the measures are assumed to be complete, which is defined below.

Definition A.2.2 (Complete measure)
Let $(S, \mathcal{F}, \mu)$ be a measure space, then μ is said to be complete if for all $A \in \mathcal{F}$ with $\mu(A) = 0$ all subsets $B \subseteq A$ are measurable. By the properties of a measure this implies that any $B \subseteq A$ has measure $\mu(B) = 0$. □

In the preprint Kingman (2006) it is argued that some proofs regarding point processes are easier and more transparent when the so called bisection property is fulfilled. The rest of this section follows Kingman (2006) very closely.

Definition A.2.3 (Bisection property)
Let $(S, \mathcal{F}, \mu)$ be a measure space. The measure μ is said to have the bisection property if for any $A \in \mathcal{F}$ with $\mu(A) < \infty$, there exists a measurable $B \subseteq A$ with

$$\mu(B) = \tfrac{1}{2}\mu(A).$$

□

One way of checking if a measure has the bisection property is to construct a so called cheesewire, which is defined as

Definition A.2.4 (Cheesewire)
A cheesewire on the measure space $(S, \mathcal{F}, \mu)$ is a measurable function $f : S \to \mathbb{R}$ satisfying that for any $\xi \in \mathbb{R}$, the measurable set

$$f^{-1}(\xi) = \{x \in S \mid f(x) = \xi\}$$

has measure

$$\mu(f^{-1}(\xi)) = 0 \tag{A.2.1}$$

□

The existence of a cheesewire ensures μ has the bisection property, which is seen in the following.
Let $A \subseteq S$ be measurable with $\mu(A) < \infty$, and define $g : \mathbb{R} \to \mathbb{R}$ by

$$g(\xi) = \mu(\{x \in A \mid f(x) \leq \xi\}).$$

Then g is monotone increasing, with

$$\lim_{\xi \to -\infty} g(\xi) = 0 \quad \text{and} \quad \lim_{\xi \to \infty} g(\xi) = \mu(A).$$

Since g is monotone the only possible discontinuities are jumps, but this requires $\mu(\{x \in A \mid f(x) < \xi\}) < \mu(\{x \in A \mid f(x) \leq \xi\})$, for some $\xi \in \mathbb{R}$, which contradicts (A.2.1), which states that $\mu(\{x \in A \mid f(x) = \xi\}) = 0$. This means that a $\xi \in \mathbb{R}$ exists such that $g(\xi) = \frac{1}{2}\mu(A)$, and that μ has the bisection property.

For a given measure space the bisection property is checked by constructing a cheesewire, and the most important example for our purposes is $(\mathbb{R}^n, \mathbb{B}(\mathbb{R}^n), \lambda^n)$, where λ^n is the n-dimensional Lebesgue measure. In this case any coordinate function is a cheesewire, and therefore the Lebesgue measure has the bisection property.

For a mapping between to measurable spaces $(S, \mathscr{F})$ and $(S', \mathscr{F}')$ we have the following useful proposition for classifying measurable functions (Billingsley (1995), page 182).

Proposition A.2.5
Let $T : S \to S'$, and $\mathscr{A}'$ be a generator for $\mathscr{F}'$. If $T^{-1}(A) \in \mathscr{F}$ for each $A \in \mathscr{A}'$ then T is $(\mathscr{F}, \mathscr{F}')$ measurable.

A.2.1 Fourier-Stieltjes transform

An important mathematical tool in the analysis of functions is the Fourier transform, and in measure theory a very similar transform of finite measures is used.

Definition A.2.6 (Fourier-Stieltjes Transform)
For a totally finite measure μ the Fourier Stieltjes Transform is defined as the bounded uniformly continuous function

$$\hat{\mu}(\omega) = \int \exp(-2\pi i \omega t)\mu(t).$$

If μ is a probability measure $\hat{\mu}$ is its characteristic function. □

It is seen that if μ has density p with respect to the Lebesgue measure then the Fourier-Stieltjes transform of μ is simply the usual Fourier transform of the density, i.e. $\hat{\mu} = \hat{p}$.

A.2.2 Radon Nikodym

Definition A.2.7
A measure ν is absolutely continuous with respect to a positive measure μ if $\nu(E) = 0$ for every set with $\mu(E) = 0$. □

Theorem A.2.8 (Radon Nikodym Theorem)
Let ν be absolutely continuous with respect to a measure μ, then

$$\nu(E) = \int_E f d\mu,$$

for some $L^1(\mu)$-function f. This may also be written as

$$f = \frac{d\nu}{d\mu},$$

and f is called the Radon Nikodym derivative of ν with respect to μ.

A.3 Complex Gaussian distribution

If we let U, V be random variables, then the unique random variable $X = U + iV$ is said to be a complex random variable. This section is dedicated to a special case of complex random variables based on Andersen et al. (1995).

The complex normal distribution is defined via. the two dimensional normal distribution. If we let $[\cdot] : \mathbb{C}^p \to \mathbb{R}^{2p}$ denote the natural bijection, between complex and real vectors, given by

$$[x] = \begin{pmatrix} \mathrm{Re}(x) \\ \mathrm{Im}(x) \end{pmatrix}, \quad x \in \mathbb{C}^p,$$

then the univariate complex normal distribution is defined as

Definition A.3.1
A complex random variable X is univariate complex normal distributed, with mean $\theta \in \mathbb{C}$ and variance $\sigma^2 \in \mathbb{R}^+$ if $[X] \sim N_2\Big([\theta], \frac{\sigma^2}{2} I_2\Big)$. This is denoted as $X \sim \mathbb{C}N(\theta, \sigma^2)$

The density function with respect to the Lebesgue measure on $\mathbb{C}$ is given as

$$f_X(x) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{1}{\sigma}\|x - \theta\|^2\right), \quad x \in \mathbb{C}.$$

□

Following a similar approach the multivariate complex normal distribution may be defined via. the multivariate normal distribution. Let A, B be $n \times p$ matrices with real entries, and let $C = A + iB$, then we define a matrix relation by

$$\{C\} = \begin{bmatrix} A & -B \\ B & A \end{bmatrix}.$$

The multivariate complex normal distribution is then defined as

Definition A.3.2
A p-dimensional complex random vector X is said to be multivariate complex normal distributed with mean $\theta \in \mathbb{C}^p$ and covariance matrix $\Sigma \in \mathbb{C}_+^{p\times p}$, where $\mathbb{C}_+^{p\times p}$ is the space of positive definite complex valued $p \times p$ matrices, if $[X] \sim N_{2p}\Big([\theta], \frac{1}{2}\{\Sigma\}\Big)$. This

is denoted as $X \sim \mathbb{C}N_p(\theta, \Sigma)$.
The density function with respect to the Lebesgue measure on $\mathbb{C}^p$ is given as

$$f_X(x) = \frac{1}{|\Sigma|\pi^p} \exp\left(-(x-\theta)\Sigma^{-1}(x-\theta)^\top\right), \quad x \in \mathbb{C}^p.$$

□

For the multivariate complex normal distribution, the following result for the conditional distribution holds.

Proposition A.3.3
Let $\sim \mathbb{C}N_p(\theta, \Sigma)$, and let X, θ and Σ be factorized as

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where X_j, θ_j are p_j dimensional and Σ_{ij} is $p_i \times p_j$, where $p = p_1 + p_2$. The conditional distribution of X_1 given X_2 is then given as

$$X_1|X_2 \sim \mathbb{C}N_{p_1}\left(\theta_1 - \Sigma_{12}\Sigma_{22}^{-}(X_2 - \theta_2), \Sigma_1 - \Sigma_{12}\Sigma_{22}^{-}\Sigma_{21}\right),$$

where Σ_{22}^{-} denotes a generalized matrix inverse. □

A.4 The Rayleigh Distribution

If we consider a random variable $(X_1, X_2)^\top \sim N_2(0, \sigma^2 I_2)$ then this has probability density

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(x_1^2 + x_2^2)}{2\sigma^2}\right].$$

Transforming this density into polar coordinates we obtain

$$f_{R,\Psi} = \frac{r}{2\pi\sigma^2} \exp\left[-\frac{r^2}{2\sigma^2}\right].$$

This is seen to be independent of the phase ψ which means that $p(r, \psi) = p(r)p(\psi)$. We furthermore have

$$f_\Psi(\psi) = \int_0^\infty \frac{r}{2\pi\sigma^2} \exp\left[-\frac{r^2}{2\sigma^2}\right] dr = \frac{1}{2\pi}, \quad 0 \le \psi \le 2\pi \tag{A.4.1}$$

$$f_R(r) = \int_0^{2\pi} \frac{r}{2\pi\sigma^2} \exp\left[-\frac{r^2}{2\sigma^2}\right] d\psi = \frac{r}{\sigma^2} \exp\left[-\frac{r}{2\sigma^2}\right], \quad 0 \le r, \tag{A.4.2}$$

where (A.4.1) follows by substituting $r^2 = x$. Thus if $X \sim N_2(0, \sigma^2 I_2)$ or equivalently $Z \sim \mathbb{C}N(0, 2\sigma^2)$, then the distribution of $R = \|X\|$ is independent of the phase and has probability density given by (A.4.2). A distribution with this density is called a Rayleigh distribution. This is denoted $R \sim \text{Rayleigh}(\sigma)$, and the mean and variance are given by

$$\mathbb{E}(R) = \sigma\sqrt{\frac{\pi}{2}}$$

$$\text{Var}(R) = \frac{(4-\pi)\sigma^2}{2}$$

Bibliography

Andersen, H., M. Højbjerre, D. Sørensen, and P. Eriksen (1995). *Linear and Graphical Models for the Multivariate Complex Normal Distribution*. Springer-Verlag.

Azzalini, A. (1996). *Statistical Inference - Based on the likelihood*. Chapman & Hall/CRC.

Berthelsen, K. and J. Møller (2004). *A short diversion into the theory of Markov chains, with a view to Markov chain Monte Carlo Methods*. Department of Mathematical Sciences, Aalborg University.

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). John Wiley and Sons.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, International Computer Science Institute.

Brémaud, P. and L. Massoulié (2002). Power spectra of general shot noises and hawkes point processes with a random excitation. *Advanced Applied Probability, vol. 34, pp. 205-222*.

Cappé, O., E. Moulines, and T. Rydén (2005). *Inference in Hidden Markov Models*. Springer.

Daley, D. and D. Vere-Jones (1988). *An introduction to the theory of Point Processes*. Springer-Verlag.

Feder, M. and E. Weinstein (1988). Parameter estimation of superimposed signals using the em algorithm. *IEEE Transactions on acoustics, speech and signal processing, vol. 36, no. 4 pp. 477-489*.

Gelfand, I. M. and N. Y. Vilenkin (1964). *Generalized Functions, Volume 4: Applications of Harmonic Analysis*. New York: Academic Press.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2003). *Bayesian Data Analysis* (2nd ed.). Chapman & Hall/CRC.

Geyer, C. and J. Møller (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics, 21, pp. 359-373*.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika, 82, pp. 711-732*.

Haykin, S. (2001). *Communication Systems*. John Wiley and sons Inc.

Hazewinkel, M. (Ed.) (1995). *Encyclopedia of Mathematics*, Volume 2. Kluwer Academic Publishers.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford University Press.

Jensen, J. L. (2006). Et første kursus i teoretisk statistik. Afdeling for teoretisk statistik Aarhus Universitet.

Kingman, J. (2006). Poisson processes revisited
http://www.newton.cam.ac.uk/preprints/NI06001.pdf.

Land, I. and B. Fleury (2006). Digital modulation 1 - lecture notes. http://kom.aau.dk/project/sipcom/SIPCom06/sites/sipcom8/courses/SIPCom8-2/notes.pdf.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Lee, P. M. (2004). *Baysian Statistics - an introduction* (3rd ed.). Arnold.

Meyn, S. and R. Tweedie (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag.

Molisch, A., K. Balakrishnan, D. Cassioli, C. Chong, S. Emami, A. Fort, J. Karedal, J. Kunisch, H. Schantz, and K. Siwiak (2005). A comprehensice model for ultrawideband propagation channels. *IEEE GLOBECOM 2005 proceedings*.

Morrison, G. and M. Fattouche (1998). Super-resolution modeling of the indoor radio propagation channel. *IEEE transactions on vehicular technology, vol. 47, no. 2, pp. 649-657*.

Møller, J. and R. P. Waagepetersen (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC.

Ng, S. K., T. Krishnan, and G. J. McLachlan (2002). The EM algorithm. http://www.quantlet.com/mdstat/scripts/csa/html/node42.html.

Papoulis, A. (1962). *The Fourier integral and its applications*. McGraw-Hill Book Company Inc.

Proakis, J. G. (2001). *Digital Communications*. McGraw-Hill Compainies, Inc.

Richards, J. and H. Youn (1990). *Theory of Distributions. A non technical introduction*. Cambridge University Press.

Robert, C. and G. Casella (1999). *Monte Carlo Statistical Methods*. Springer-Verlag.

Roberts, G., A. Gelman, and W. Wilks (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied Probability, vol. 7, pp. 110-120*.

Saleh, A. and R. Valenzuela (1987). A statistical model for indoor multipath propagation. *IEEE Journal on Selected Areas in Communications, vol. SAC-5, no. 2, pp. 128-137*.

Scholtz, R. (2006, June). Intel UWB and MIMO database. Web page. http://impulse.usc.edu/.

Schuster, U. G. and H. Bölcskei (2006). Ultrawideband channel modeling on the basis of information-theoretic criteria. *IEEE Transactions on Wireless Communications*.

Spencer, Q., B. Jeffs, M. Jensen, and A. Swindlehurst (2000). Modeling the statistical time and angle of arrival characteristics of an indoor multipath channel. *IEEE Journal on Selected Areas in Communications, vol. 18, no. 3, pp. 347-359*.

Stirzaker, D. (2003). *Elementary probability.* Cambridge university press.

Suzuki, H. (1977). A statistical model for urban radio propagation. *IEEE Transactions on Communications vol. COM-25, no. 7, pp. 673-680*.

Taksar, M. and B. Højgaard (2006). Diffusion optimization models in insurance and finance. Incomplete Book Manuscript.

Turin, G., F. Clapp, T. Johnston, S. Fine, and D. Lavry (1972). A statistical model of urban multipath propagation channels. *IEEE Transactions on Vehicular Technology, vol. 21, pp. 1-9*.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics, vol. 11, pp. 95-103*.

Øksendal, B. (2003). *Stochastic Differential Equations* (6th ed.). Springer-Verlag.

Summary

This thesis deals with the stochastic modeling of impulse response functions, which are functions that completely determine the behavior of a given radio channel. A short description of the problem and previous models is given in the first chapter.

In order to fit the modeling into a strict mathematical context, we introduce the theory of point processes. Using results from the theory of random measures we establish the existence of point processes as random integer valued measures as a consequence of the Kolmogorov extension theorem. The Poisson process is defined, and a number of relevant propositions and theorems are proved. To describe the problem at hand, as a point process we introduce marked point processes, which in short terms consists of marking each point in a point process, with e.g. a number or a function. Under certain assumptions a marked point process, may be shown to be equivalent to a point process on the product space of the point space and the mark space, and using this we describe some of the previous stochastic models for an impulse response function, as point processes on $\mathbb{R}^2$.

Chapter 3 gives an introduction to maximum likelihood estimation, and introduces the EM algorithm with an example on how this may be implemented to estimate parameters in an impulse response function. Chapter 3 also provides the necessary concepts from Bayesian inference, which we will use in connection with the MCMC methods in chapter 4. The MCMC section consist of basic concepts regarding Markov chains, such as criteria of convergence. The second half of the chapter shows how these concepts may be applied to design Markov chains which converge to a chosen distribution. In particular we describe the Metropolis Hastings algorithm, and show how this is extended to a reversible jump MCMC algorithm, which is a Bayesian inference tool designed to estimate an unknown number of parameters.

Using the theory from chapter 4, we design a reversible jump MCMC algorithm and use this to estimate parameter values and model orders, in stochastic models of the impulse response function for the data at hand.