Technische Universität München

Zentrum Mathematik

Bayesian Inference for Bivariate Compound Poisson Processes

Diplomarbeit

von

Philipp Gebhard

Themenstellerin:	Prof. Dr. Claudia Klüppelberg
Betreuer:	Dr. Gernot Müller
Abgabetermin:	01. Juni 2010

Hiermit erkläre ich, dass ich die Diplomarbeit selbstständig angefertigt und nur die angegebenen Quellen verwendet habe.

Garching, den 31. Mai 2010

Acknowledgments

Foremost, I would like to express my sincere gratitude to Prof. Dr. Claudia Klüppelberg for giving me the opportunity to write this interesting diploma thesis at her Chair of Mathematical Statistics. I am very thankful for her expertised and valuable suggestions.

I want to thank Dr. Gernot Müller for the encouragement and for sharing his profound knowledge with me. I appreciate the excellent mentoring very much.

Further I thank Dr. Klaus Böcker and Habib Esmaeili for the fruitful discussions and their experienced advices.

Last but not least I thank my family for their enduring support and encouragement.

Contents

Chapter 1

Introduction

Many problems in mathematical finance require multivariate modelling to capture the dependence between the components. An important field of application is, for example, the modelling of operational risks. According to Basel II (2004), banks are obliged to calculate capital charges for their operational risk. Moreover, they are required to distinguish between different cells (determined by business lines and loss event types) which all contribute to the total operational risk of the bank. Thus, the univariate risks of the single cells have to be modelled simultaneously and the dependencies between these cells must be taken into consideration.

By defining copulas and deriving their fundamental properties Sklar (1959) introduced a convenient way to model the marginal processes and the dependence structure between them separately. This classical concept is particularly useful if there are few sources of risk.

In the framework of multivariate Lévy processes the dependence may be described by Lévy copulas. They are defined on the domain of Lévy measures instead of probability measures. Kallsen and Tankov (2006) amongst others showed that many results from the ordinary copula theory can be extended to Lévy copulas. Cont and Tankov (2004) discussed the concept and employed it for financial modelling. Further important references for the theory of copulas are Nelsen (1997) and Joe (1997).

An application of Lévy copulas for modelling operational risks was presented by Böcker and Klüppelberg (2006; 2008; 2009). They used a multivariate compound Poisson process (CPP) to model the risk in the univariate cells. In two other publications Esmaeili and Klüppelberg (2010a; 2010b) introduced a maximum likelihood estimation procedure for bivariate compound Poisson processes and bivariate stable Lévy processes, respectively.

The aim of this thesis is to study the estimation of bivariate CPPs from a Bayesian

perspective. We develop a Markov chain Monte Carlo (MCMC) estimation procedure which, in particular, enables us to involve prior knowledge about the parameters of interest. Moreover, whereas a purely frequentist approach results in point estimates of the model parameters, the Bayesian method allows to derive additional knowledge about the posterior distribution and hence about the uncertainty of the estimates given the expert information contained in the priors.

This thesis is organized as follows. In Chapter 2 we summarize the theoretical background for the subsequent chapters. We first recall some definitions and properties of stochastic processes and review the concept of Lévy copulas to describe the dependence structure of a multivariate Lévy process. After presenting the most important notions of Bayesian inference we also recall the concept of Bayes factors for model selection. When outlining the main ideas of MCMC we discuss particularly the Metropolis-Hastings (MH) algorithm and the Gibbs sampler.

In Chapter 3 we develop a MCMC procedure for the estimation of the posterior distribution in bivariate compound Poisson models. Since the choice of the proposal distributions for the MH-steps has, as usual, a major impact on the behaviour of the sampling algorithm, we present some guidelines on how to calibrate the proposals to make the sampler highly efficient. Additionally, we discuss some other important issues, including the choice of initial values, burn-in period and subsampling.

In Chapter 4 we check the performance of the adapted sampler in a simulation study. First, we investigate the mixing and convergence behaviour of the produced chains. By repeating the analysis for several simulations we are able to asses the quality of the posterior mean estimates. Furthermore, we compare them to the maximum likelihood estimates which are calculated using the method in Esmaeili and Klüppelberg (2010a).

In Chapter 5 we finally apply the sampler to Danish fire insurance data. After describing the structure of the data we present two different approaches. In the first we consider the data in its original form whereas in the second we restructure the data and transform it with the logarithm. We discuss briefly the advantages and drawbacks of these approaches, before we analyze the two data sets separately. In each case, we fit several Clayton models – including sliced distribution models – to the data and select the best fitting one using Bayes factors and weighted Bayes factors. For the chosen model we investigate how sensitive the marginal posterior distributions are with respect to the choice of the prior distributions. Subsequently, we illustrate how the estimated distributions vary if less observations are included into the analysis and examine the impact of outliers. Moreover, we compare the prior distributions to the estimated marginal posteriors.

Chapter 2

Theoretical background

Here we want to summarize some background from stochastics and statistics which is important for the following chapters. First we recall some special stochastic processes and look at their Lévy measure. Then we focus on dependence concepts for Lévy processes. Introducing the notion of Lévy copula we are able to represent bivariate compound Poisson processes (CPP) in an intuitive way and derive their likelihood which is fundamental for our work. After some brief comments on Bayesian inference we deal with Markov chain Monte Carlo (MCMC) methods. We state some results concerning Markov chains and then discuss the two basic MCMC algorithms, the Metropolis-Hastings (MH) algorithm and the Gibbs sampler.

2.1 Stochastic processes

In the following we assume that the reader is familiar with the notions of random variable, probability space, measure, filtration, càdlàg function, Lévy process, Poisson process and some other basic concepts. When stating the upcoming definitions and properties we follow the explanations of Cont and Tankov (2004).

Since the *multivariate compound Poisson process* is very fundamental for this thesis we want to recall its definition first.

Definition 2.1.1 (Multivariate compound Poisson process (CPP)).
A d-dimensional compound Poisson process with intensity $\boldsymbol{\lambda} > 0$ *and jump size distributions* f_i, $i = 1, \dots, d$, *is a stochastic process* $\mathbf{S} = (\mathbf{S}_t)_{t \geq 0} := (S_t^1, \dots, S_t^d)_{t \geq 0}$ *on a filtered probability space* $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ *with values in* $\mathbb{R}^d$. *Each component* S_t^i, $i = 1, \dots, d$, *is defined by*

$$S_t^i = \sum_{k=1}^{N_t^i} X_k^i \,,$$

where, for all i, the jump sizes $(X_k^i)_{k\geq 1}$ are i.i.d. with distribution f_i and $(N_t^i)_{t\geq 0}$ is a Poisson process with intensity λ_i, independent from $(X_k^i)_{k\geq 1}$.

The CPP is one of the simplest examples of Lévy processes. A Poisson process itself can be seen as a CPP on $\mathbb{R}$ such that $X_k \equiv 1$ for all k. This explains where the term 'compound Poisson' comes from. CPPs are the only Lévy processes with piecewise constant trajectories, as shown by the following proposition.

Proposition 2.1.2.
$(\mathbf{S}_t)_{t\geq 0}$ is a CPP if and only if it is a Lévy process and its sample paths are piecewise constant functions.

For the proof we refer to Cont and Tankov (2004).

Since any càdlàg function may be approximated by a piecewise constant function, one may expect that general Lévy processes can be well approximated by compound Poisson ones and that by studying CPPs one can gain insight into the properties of Lévy processes. That is why we will focus mainly on CPPs.

The jump times $(T_n^i)_{n\geq 1}$, $i = 1, \ldots, d$, of the components of the CPP have the same law as the jump times of the underlying Poisson process $(N_t^i)_{t\geq 0}$. From the definition of a Poisson process we know that they can be expressed as partial sums of independent exponential random variables with parameter λ_i. We make use of this property in Section 2.2.2 in order to derive the likelihood of CPPs. These jump times together with the jump sizes constitute for every single CPP its *marked point process*. Knowledge of it determines the process uniquely.

Definition 2.1.3 (Marked point process).
A marked point process on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ is a sequence $(T_n, \boldsymbol{X}_n)_{n\geq 1}$ where

- *$(T_n)_{n\geq 1}$ is an increasing sequence of $\mathcal{F}_{T_n}$-adapted random times with $T_n \to \infty$ almost surely as $n \to \infty$.*
- *$(\boldsymbol{X}_n)_{n\geq 1}$ is a sequence of random variables taking values in $E \subseteq \mathbb{R}^d$.*
- *The value of $\boldsymbol{X}_n$ is revealed at T_n: $\boldsymbol{X}_n$ is $\mathcal{F}_{T_n}$-measurable.*

In the following we need the concept of Lévy measure of stochastic processes. Before introducing it we clarify some notions which lead us to the definition of the Lévy measure for Lévy processes.

Let us begin with a simple example of a measure, determined by any Poisson process $(N_t)_{t\geq 0}$. Here the jump times $T_1, T_2, \ldots$ form a random configuration of points on $[0, \infty[$

and the Poisson process $(N_t)_{t\geq 0}$ counts the number of such points in the interval $[0,t]$. This counting procedure defines a measure M on $[0,\infty[$: for any measurable set $A \subset \mathbb{R}^+$ let

$$M(\omega, A) = \#\left\{n \geq 1,\, T_n(\omega) \in A\right\} .$$

Then $M(\omega,\cdot)$ is a positive, integer valued measure and $M(\omega, A)$ is finite with probability 1 for any bounded set A. Since the measure $M(\omega,\cdot)$ depends on ω, it is called a *random measure*.

In the same manner as in the example above we can now associate a random measure on $[0,\infty[\times\mathbb{R}^d$ to every d-dimensional càdlàg process $(\mathbf{S}_t)_{t\geq 0}$ (in particular to every CPP). For every measurable set $B \subset [0,\infty[\times\mathbb{R}^d$ we define

$$J_S(B) = \#\left\{(t, \Delta\mathbf{S}_t) \in B\right\} .$$

$J_S([t_1,t_2]\times A)$ counts, for any measurable set $A \subset \mathbb{R}^d$, the number of jump times of $\mathbf{S}$ between t_1 and t_2 such that their jump sizes $\Delta\mathbf{S}_t := \mathbf{S}_t - \mathbf{S}_{t-}$ are in A. Hence, J_S is called the *jump measure* of the process $\mathbf{S}$.

From the notion of jump measure it is only a small step to the notion of *Lévy measure*. All we have to do is to normalize the time interval on $[0,1]$ and to consider the average numberinstead of the random number of jumps. This approach is much more general, since the resulting measure does not depend on the uncertainty of the realization ω anymore.

Definition 2.1.4 (Lévy measure).
Let $(\mathbf{S}_t)_{t\geq 0}$ be a Lévy process on $\mathbb{R}^d$. The measure Π on $\mathbb{R}^d$ defined by:

$$\Pi(A) = E\left[\#\left\{t \in [0,1] : \Delta\mathbf{S}(t) \neq 0,\, \Delta\mathbf{S}(t) \in A\right\}\right] , \qquad A \in \mathcal{B}(\mathbb{R}^d) ,$$

is called the Lévy measure of $\mathbf{S}$: $\Pi(A)$ is the expected number, per unit time, of jumps whose size belongs to A.

Basically, the Lévy measure controls the jump behaviour of a Lévy process. For instance, the Lévy measure Π of an one-dimensional CPP $(S_t)_{t\geq 0}$ can be written in terms of the (frequency) rate $\lambda > 0$ and the jump size distribution function F, namely $\Pi([0,x]) = \lambda\, P(\Delta S \leq x) = \lambda\, F(x)$ for $x \in [0,\infty)$. Hence, the Lévy measure of a univariate CPP gives the expected number of jumps per unit time with a jump size in a pre-specified interval.

For the multivariate case it is quite similar: the Lévy measure controls the single and joint jump behaviour (per unit time) of all components and contains all information of dependence between the univariate components.

2.2 Dependence concepts for Lévy processes

In modelling financial data with Lévy processess – applications are e.g. portfolios of insurance claims or operational risks – things get complicated when multi-dimensional processes have to be considered. Then the main difficulty is to model the dependence between the different marginal processes, in particular the dependence between the jumps of these components. If the jumps are assumed to be Gaussian, their associatioin can be described via correlation coefficients. If they are not, the classical approach is to use copulas. They open a convenient way to represent the dependence structure for random variables, since they contain all the dependence information which is, thus, clearly separated from the marginal behaviour of the components. If there are few sources of jump risk, this classical concept is very useful, because it allows to achieve a precise description of dependence within a simple model. But when there are several sources of jump risk, the model quickly becomes very complicated, because one has to introduce a separate copula for each jump risk source. Another inconvenience of this modelling approach is that it does not allow to couple components of different types.

To avoid these drawbacks, the Lévy copula is a sophisticated way to model the dependence between the components in the framework of Lévy processes. In contrast to distributional copulas which are defined on the domain of distribution functions, Lévy copulas are defined on a different domain. Lévy copulas can be used to construct a d-dimensional Lévy process by taking any set of one-dimensional Lévy processes and coupling them. They allow to model the whole range of possible dependence structures in a parametric fashion with a small number of parameters.

For ease of notation we present our Lévy copula concept for spectrally positive Lévy processes (i.e. processes with only non-negative jumps). This is no restriction of the theory, since the Lévy copula for general Lévy processes is specified for each quadrant separately. In this section we follow again Cont and Tankov (2004), Böcker and Klüppelberg (2008) and Kallsen and Tankov (2006).

2.2.1 The Lévy copula

The important dependence concept for Lévy copulas is the dependence of jumps. Hence, for parametrizing the dependence between jumps of Lévy processes, the Lévy measure plays the same role as the probability measure does for random variables. The principal difference from the ordinary copula case is that Lévy measures are not necessarily finite. Due to this fact, Lévy copulas are defined on infinite intervals rather than on $[0,1]^d$. The role of distribution function is now played by the *tail integral*.

Definition 2.2.1 (Tail integral).
Let Π be a Lévy measure on $\mathbb{R}_+^d$. The tail integral is a function $\overline{\Pi} : [0,\infty]^d \to [0,\infty]$ defined by

$$\overline{\Pi}(x_1,\dots,x_d) = \begin{cases} \Pi([x_1,\infty)\times\cdots\times[x_d,\infty))\,, & (x_1,\dots,x_d)\in[0,\infty)^d\,,\\ 0\,, & \text{if } x_i=\infty \text{ for at least one } i\,. \end{cases}$$

The marginal tail integrals are defined for $i = 1,\dots,d$ as $\overline{\Pi}_i(x) = \Pi_i([x,\infty))$ for $x \geq 0$.

Practically, an one-dimensional tail integral is simply the expected number of jumps per unit time that are above a given threshold x,

$$\overline{\Pi}_i(x) = \Pi_i([x,\infty)) = \lambda_i\, P(\Delta S^i > x) = \lambda_i\, \overline{F}_i(x)\,, \qquad x\in[0,\infty)\,.$$

In the multivariate case the tail integral is the expected number of joint jumps such that the marginal jump of each component is greater than x_i, $i = 1,\dots,d$. We see that the dependence of frequency and the dependence of jump size between different components are both encoded in the tail integral.

To define the Lévy copula we need the notions of *d-increasing function* and *grounded function.*

Definition 2.2.2 (Grounded function, d-increasing function).
Let F be a real d-dimensional function.

- *Suppose that the domain of F is $D_1\times\cdots\times D_d$ where each D_k has a smallest element a_k. F is said to be grounded, if $F(\boldsymbol{t}) = 0$ for all $\boldsymbol{t}$ in Dom F such that $t_k = a_k$ for at least one k.*

- *F is called d-increasing if $V_F(B) := \sum sgn(\boldsymbol{c})F(\boldsymbol{c}) \geq 0$ for all d-boxes $B = [\boldsymbol{a},\boldsymbol{b}]$, $\boldsymbol{a}\leq\boldsymbol{b}$, whose vertices $\boldsymbol{c}$ lie in Dom F. Here $sgn(\boldsymbol{c})$ is defined by*

$$sgn(\boldsymbol{c}) = \begin{cases} 1\,, & \text{if } c_k = a_k \text{ for an even number of vertices}\,,\\ -1\,, & \text{else}\,. \end{cases}$$

The notion d-increasing is thus nothing else than the multivariate extension of 'increasing'. Groundedness guarantees that the Lévy copula defines a measure on $[0,\infty]^d$.

Definition 2.2.3 (Positive Lévy copula).
A d-dimensional Lévy copula for Lévy processes with positive jumps, or, for short, a positive Lévy copula, is a d-increasing grounded function $\mathfrak{C} : [0,\infty]^d \to [0,\infty]$ *with margins* $\mathfrak{C}_k$, $k = 1,\dots,d$, *which satisfy* $\mathfrak{C}_k(u) = u$ *for all* u *in* $[0,\infty]$.

We want to stress that Lévy copulas allow for an intuitive and structural explanation of what 'dependence' actually can be thought of: dependence means that there are jumps in different components which occur at the same time. More precisely, independence means that jumps in different components never occur at the same time and that their jump size variables are also independent. Complete positive dependence, on the other hand, means that jumps always occur at the same points in time and that the jump size variables also have a perfect positive dependence structure (comonotonicity).

Let us further remark that Lévy copulas give a *dynamic* description of the dependence structure of a Lévy process $\mathbf{S}$, in contrast to static models where the *distributional* dependence between the margins of $\mathbf{S}$ for a predetermined and fixed $t \geq 0$ is considered. This is in particular an advantage when operational losses are modelled. Since operational losses occur in time, a statical dependence model can never reflect coincidence of losses in different cells, caused e.g. by the same catastrophic event. This is also acknowledged by the regulators who, by assuming a statical dependence model, demand that losses which affect different cells, but which are caused by one and the same event are not counted as several small losses (simultaneously happening in different cells), but rather as one single big loss impacting only a single cell. The reason is that a statical model would 'forget' that these losses actually have the same origin and it would falsely treat them as independent events instead of one single, perhaps disastrous, incident. Of course, in the framework of Lévy copulas as suggested here, this artificial correction is not necessary, because the observation of joint losses is properly reflected in the dependence model.

The theorem we can finally formulate is a reformulation of Sklar's (1959) theorem for tail integrals and Lévy copulas. It shows that Lévy copulas link multidimensional tail integrals to their margins in the same way as the distributional copulas link the multivariate distribution functions to their margins.

Theorem 2.2.4.
Let $\overline{\Pi}$ *be the tail integral of a d-dimensional Lévy process with positive jumps and let* $\overline{\Pi}_1,\dots,\overline{\Pi}_d$ *be the tail integrals of its components. Then there exists a d-dimensional positive Lévy copula* $\mathfrak{C}$ *such that for all vectors* $(x_1,\dots,x_d)$ *in* $\mathbb{R}^d_+$,

$$\overline{\Pi}(x_1,\dots,x_d) = \mathfrak{C}(\overline{\Pi}_1(x_1),\dots,\overline{\Pi}_d(x_d))\,.$$

If $\overline{\Pi}_1, \ldots, \overline{\Pi}_d$ are continuous then $\mathfrak{C}$ is unique, otherwise it is unique on $Ran\,\overline{\Pi}_1 \times \cdots \times Ran\,\overline{\Pi}_d$.

Conversely, if $\mathfrak{C}$ is a d-dimensional positive Lévy copula and $\overline{\Pi}_1, \ldots, \overline{\Pi}_d$ are the tail integrals of Lévy measures on $[0, \infty)$, then the function $\overline{\Pi}$ defined above is the tail integral of a d-dimensional Lévy process with positive jumps having marginal tail integrals $\overline{\Pi}_1, \ldots, \overline{\Pi}_d$.

A proof for the two-dimensional case is for example given in Cont and Tankov (2004). For the general multivariate case, we refer to the proof of Sklar's (1959) theorem.

The first part of this theorem states that all types of dependence of Lévy processes – including complete dependence and independence – can be represented with Lévy copulas. The second part provides a systematic way to construct multivariate Lévy processes by specifying separately a jump dependence structure and one-dimensional Lévy processes. These componenents can have very different structure, that is to say, one can couple different Lévy processes.

If the dependence is specified via a Lévy copula and both the copula and the one-dimensional tail integrals are sufficiently smooth, the Lévy density for bivariate Lévy processes can be computed by differentiation, cf. Cont and Tankov (2004).

Proposition 2.2.5.
Let $\mathfrak{C}$ be a two-dimensional Lévy copula, continuous on $[0, \infty]^2$, such that $\frac{\partial^2 \mathfrak{C}(u,v)}{\partial u \partial v}$ exists on $(0, \infty)^2$ and let $\overline{\Pi}_1$ and $\overline{\Pi}_2$ be one-dimensional tail integrals with densities ν_1 and ν_2. Then

$$\nu(x, y) = \frac{\partial^2 \mathfrak{C}(u, v)}{\partial u \partial v} \Big|_{u=\overline{\Pi}_1(x), v=\overline{\Pi}_2(y)} \nu_1(x)\nu_2(y) \tag{2.2.1}$$

is the Lévy density of a Lévy measure with marginal Lévy densities ν_1 and ν_2.

2.2.2 Representation of CPPs based on Lévy copulas

A bivariate model is particularly useful to illustrate how Bayesian statistics can be used to estimate the parameters of Lévy processes. Therefore, we now consider the two-dimensional case. Moreover, we focus on CPPs, since – as mentioned before – general Lévy processes can be well approximated by them. We refer to Esmaeili and Klüppelberg (2010a) for details about the subsequent results.

Let us assume that we observe a bivariate CPP $(\mathbf{S}(t))_{t\in[0,T]} = (S_1(t), S_2(t))_{t\in[0,T]}$ over

a fixed time interval $[0,T]$, $T>0$, where

$$S_1(t) = \sum_{i=1}^{N_1(t)} X_i\,, \quad 0 \le t \le T\,, \qquad \text{and} \qquad S_2(t) = \sum_{j=1}^{N_2(t)} Y_j, \quad 0 \le t \le T\,.$$

Our goal is to determine the likelihood function of this CCP based on the observed jump times and jump sizes in both components. As we will see, it is therefore very useful when we work with a representation of CPPs which is based on Lévy copulas instead of ordinary copulas.

The components S_i, $i=1,2$, (representing the aggregate processes) of a bivariate CPP $\mathbf{S}$ can always be split into *jump dependent parts* $S_i^{\|}$ and *independent parts* $S_i^{\perp}$, $i=1,2$, (this is a consequence of the Lévy-Itô decomposition),

$$S_1 = \sum_{i=1}^{N_1} X_i = S_1^{\perp} + S_1^{\|} = \sum_{k=1}^{N_1^{\perp}} X_k^{\perp} + \sum_{m=1}^{N^{\|}} X_m^{\|}\,, \tag{2.2.2}$$

$$S_2 = \sum_{j=1}^{N_2} Y_j = S_2^{\perp} + S_2^{\|} = \sum_{l=1}^{N_2^{\perp}} Y_l^{\perp} + \sum_{m=1}^{N^{\|}} Y_m^{\|}\,. \tag{2.2.3}$$

Here $S_1^{\perp}$ and $S_2^{\perp}$ are independent from each other (no joint jumps) and from the other two components, whereas $S_1^{\|}$ and $S_2^{\|}$ are dependent (the jumps of both components are caused by the same event and thus always happen together). It can be shown that all three processes $S_1^{\perp}$, $S_2^{\perp}$ and $(S_1^{\|}, S_2^{\|})$ are again compound Poisson and independent with Poisson processes $N_1^{\perp}$, $N_2^{\perp}$ and $N^{\|}$, respectively.

Fixing the model completely requires to specify different quantities. Let us compare these specifications for our two copula concepts: the ordinary copula and the Lévy copula. In the first case we have to determine

- the intensity and jump size distribution of $S_1^{\perp}$,
- the intensity and jump size distribution of $S_2^{\perp}$,
- the intensity of common jumps,
- jump size distributions of $S_1^{\|}$ and $S_2^{\|}$,
- the copula of the last two distributions.

We see that the distributional copula approach requires a lot of different quantities. When using the Lévy copula this is not the case. Here we only have to specify

- the margins via the intensity and jump size distribution of S_1 and S_2,
- the dependence structure via the Lévy copula of the process.

All other quantities can be derived from these ones, cf. Cont and Tankov (2004). This results in a model with comparably few parameters, making it particularly advantageous in case of rare data. Hence, the Lévy copula approach is obviously much more convenient for modelling CPPs.

With the specification of CPPs by Lévy copulas, we can now compute the full likelihood of two-dimensional CPPs. Due to Equations (2.2.2) and (2.2.3) and the fact that the single components are independent from each other, one can write the likelihood function of the bivariate process (S_1, S_2) as the product of the likelihoods of the processes $S_1^{\perp}$, $S_2^{\perp}$ and $(S_1^{\parallel}, S_2^{\parallel})$. Let us briefly describe the likelihood of an univariate CPP, before considering the bivariate case.

Assume that within the time interval $[0, T]$ one observes n jumps at times $T_1, \ldots, T_n$, each with jump size $x_i \sim f(\cdot|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector of the jump size distribution. Defining the inter-arrival times $Z_i := T_i - T_{i-1}$ for $i = 1, \ldots, n$ with $T_0 = 0$, and recalling that the Z_i are i.i.d. exponential random variables with parameter λ, we can write the likelihood as

$$\begin{aligned} f(\mathbf{x}, n|\lambda, \boldsymbol{\theta}) &= e^{-\lambda(T-T_n)} \prod_{i=1}^{n} \lambda e^{-\lambda Z_i} \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \\ &= \lambda^n e^{-\lambda T} \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \,. \end{aligned} \tag{2.2.4}$$

The last term in (2.2.4) is the likelihood of the observed jump sizes, the part in the middle is the likelihood of the observed inter-arrival times and the first factor is simply the probability that there is no jump within the interval $(T_n, T]$, that is $P(T_{n+1} > T) = P(Z_{n+1} > T - T_n) = e^{-\lambda(T-T_n)}$.

The following theorem deals with the bivariate equivalent to the likelihood from above. It is given in Esmaeili and Klüppelberg (2010a). Since it is a fundamental result for the upcoming explanations, we here want to give a more detailed proof. Assume that we observe a bivariate CPP $\mathbf{S} = (S_1, S_2)$ which is fully determined by the parameter set

$$\boldsymbol{\psi} := (\lambda_1, \boldsymbol{\theta_1}, \lambda_2, \boldsymbol{\theta_2}, \boldsymbol{\delta}) \,.$$

Here $\lambda_i > 0$, $i = 1, 2$, denotes the frequency parameter of component S_i which has jump size distribution F_i with parametrization $\boldsymbol{\theta}_i$, $i = 1, 2$. The Lévy copula $\mathfrak{C}$ is determined by the parameter vector $\boldsymbol{\delta}$.

Observing a CPP continuosly over a fixed time period $[0,T]$ is equivalent to observing all jump times and jump sizes in this time interval. Let $\tilde{\mathbf{x}}$, $\tilde{\mathbf{y}}$ denote the observed independent jump sizes of S_1, S_2, respectively, and $(\mathbf{x},\mathbf{y})$ the joint jump sizes. Furthermore, we write $n_1^{\perp} = N_1^{\perp}(T)$, $n_2^{\perp} = N_2^{\perp}(T)$ for the number of independent jumps in S_1, S_2, respectively, and $n^{\parallel} = N^{\parallel}(T)$ for the number of joint jumps. For notational convenience we set

$$\mathbf{z} := (\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathbf{x}, \mathbf{y}, n_1^{\perp}, n_2^{\perp}, n^{\parallel})\,.$$

Theorem 2.2.6.
Assume an observation scheme as above. Assume further that $\frac{\partial^2}{\partial u \partial v}\mathfrak{C}(u,v;\boldsymbol{\delta})$ exists for all $(u,v) \in (0,\lambda_1)\times(0,\lambda_2)$, which is the domain of $\mathfrak{C}$. Then the full likelihood of the bivariate CPP is given by

$$\begin{aligned} f(\mathbf{z} \mid \boldsymbol{\psi}) = {} & (\lambda_1)^{n_1^{\perp}} e^{-\lambda_1^{\perp} T} \prod_{i=1}^{n_1^{\perp}} \left[f_1(\tilde{x}_i;\boldsymbol{\theta_1}) \left(1 - \frac{\partial}{\partial u}\mathfrak{C}\left(u,\lambda_2;\boldsymbol{\delta}\right)\Big|_{u=\lambda_1 \overline{F}_1(\tilde{x}_i;\boldsymbol{\theta_1})} \right) \right] \\ & \times (\lambda_2)^{n_2^{\perp}} e^{-\lambda_2^{\perp} T} \prod_{i=1}^{n_2^{\perp}} \left[f_2(\tilde{y}_i;\boldsymbol{\theta_2}) \left(1 - \frac{\partial}{\partial v}\mathfrak{C}\left(\lambda_1,v;\boldsymbol{\delta}\right)\Big|_{u=\lambda_2 \overline{F}_2(\tilde{y}_i;\boldsymbol{\theta_2})} \right) \right] \\ & \times (\lambda_1\lambda_2)^{n^{\parallel}} e^{-\lambda^{\parallel} T} \prod_{i=1}^{n^{\parallel}} \left[f_1(x_i;\boldsymbol{\theta_1}) f_2(y_i;\boldsymbol{\theta_2}) \frac{\partial^2}{\partial u \partial v}\mathfrak{C}\left(u,v;\boldsymbol{\delta}\right)\Big|_{u=\lambda_1 \overline{F}_1(x_i;\boldsymbol{\theta_1}), v=\lambda_2 \overline{F}_2(y_i;\boldsymbol{\theta_2})} \right] \end{aligned} \tag{2.2.5}$$

with $\lambda^{\parallel} = \lambda^{\parallel}(\boldsymbol{\delta}) = \mathfrak{C}(\lambda_1,\lambda_2;\boldsymbol{\delta})$ and $\lambda_i^{\perp}(\boldsymbol{\delta}) = \lambda_i - \lambda^{\parallel}(\boldsymbol{\delta})$ for $i = 1,2$.

Proof: (cf. Esmaeili and Klüppelberg (2010a))
To calculate the likelihood function, we start with representations (2.2.2) and (2.2.3) of a bivariate CPP. Since the components $S_1^{\perp}$, $S_2^{\perp}$ and $(S_1^{\parallel},S_2^{\parallel})$ are independent from each other the tail integrals can be represented as $\overline{\Pi}_i = \overline{\Pi}_i^{\perp} + \overline{\Pi}_i^{\parallel}$, $i = 1,2$.

Here we have for $x > 0$

$$\overline{\Pi}_1(x) = \Pi_1([x,\infty)) = \lambda_1 P(\Delta S_1 > x) = \lambda_1 \overline{F}_1(x) \qquad \text{(marginal tail int.)}\,, \tag{2.2.6}$$

$$\begin{aligned} \overline{\Pi}_1^{\perp}(x) &= \Pi([x,\infty),\{0\}) = \overline{\Pi}(x,0) - \lim_{y\to 0^+} \overline{\Pi}(x,y) \\ &\overset{Sklar}{=} \overline{\Pi}_1(x) - \lim_{y\to 0^+} \mathfrak{C}(\overline{\Pi}_1(x), \overline{\Pi}_2(y);\boldsymbol{\delta}) \\ &\overset{\mathfrak{C}\ cont.}{=} \overline{\Pi}_1(x) - \mathfrak{C}(\overline{\Pi}_1(x), \lim_{y\to 0^+} \overline{\Pi}_2(y);\boldsymbol{\delta}) \\ &= \overline{\Pi}_1(x) - \mathfrak{C}(\overline{\Pi}_1(x), \lambda_2;\boldsymbol{\delta}) \qquad \text{(tail int. of the independent part)}\,, \end{aligned} \tag{2.2.7}$$

$$\begin{aligned} \overline{\Pi}_1^{\parallel}(x) &= \lim_{y\to 0^+} \overline{\Pi}^{\parallel}(x,y) \overset{x>0}{=} \lim_{y\to 0^+} \overline{\Pi}(x,y) \\ &= \mathfrak{C}(\overline{\Pi}_1(x), \lambda_2;\boldsymbol{\delta}) \qquad \text{(tail int. of the jump dependent part)}\,, \end{aligned}$$

and analogously for $i = 2$. Furthermore, we can write

$$\begin{aligned}
\lambda^{\parallel} &= \lim_{x,y\to 0^+} \overline{\Pi}(x,y) = \mathfrak{C}(\lambda_1,\lambda_2;\boldsymbol{\delta})\,,\\
\lambda_1^{\perp} &= \lim_{x\to 0^+} \overline{\Pi}_1(x) - \lim_{x,y\to 0^+} \overline{\Pi}(x,y) = \lambda_1 - \mathfrak{C}(\lambda_1,\lambda_2;\boldsymbol{\delta})\,,\\
\lambda_2^{\perp} &= \lambda_2 - \mathfrak{C}(\lambda_1,\lambda_2;\boldsymbol{\delta})\,.
\end{aligned}$$

Thus, we obtain for $x, y > 0$ the independent parts and the jump dependent part of (S_1, S_2) as

$$\begin{aligned}
\lambda_1^{\perp}\overline{F}_1^{\perp}(x) &= \overline{\Pi}_1^{\perp}(x) \overset{(2.2.7)}{=} \overline{\Pi}_1(x) - \mathfrak{C}(\overline{\Pi}_1(x),\lambda_2;\boldsymbol{\delta}) \overset{(2.2.6)}{=} \lambda_1\overline{F}_1(x) - \mathfrak{C}(\lambda_1\overline{F}_1(x),\lambda_2;\boldsymbol{\delta})\,,\\
\lambda_2^{\perp}\overline{F}_2^{\perp}(y) &= \lambda_2\overline{F}_2(y) - \mathfrak{C}(\lambda_1,\lambda_2\overline{F}_2(y);\boldsymbol{\delta})\,, \qquad (2.2.8)\\
\lambda^{\parallel}\overline{F}^{\parallel}(x,y) &= \overline{\Pi}^{\parallel}(x,y) \overset{x,y>0}{=} \overline{\Pi}(x,y) \overset{Sklar}{=} \mathfrak{C}(\overline{\Pi}_1(x),\overline{\Pi}_2(y);\boldsymbol{\delta}) \overset{(2.2.6)}{=} \mathfrak{C}(\lambda_1\overline{F}_1(x),\lambda_2\overline{F}_2(y);\boldsymbol{\delta})\,.
\end{aligned}$$

Let now $L_1(\tilde{\mathbf{x}}, n_1^{\perp} \mid \lambda_1^{\perp}, \boldsymbol{\theta_1})$ denote the marginal likelihood function based on the observations of the jump times and jump sizes of the first component $S_1^{\perp}$. To derive the function L_1 let $\tilde{T}_1,\ldots,\tilde{T}_{n_1^{\perp}}$ denote the jump times of $S_1^{\perp}$, and define the sequence of inter-arrival times $\tilde{Z}_i = \tilde{T}_i - \tilde{T}_{i-1}$ for $i = 1,\ldots,n_1^{\perp}$, where $\tilde{T}_0 = 0$. Then the $\tilde{Z}_i$ are i.i.d. exponentially distributed random variables with parameter $\lambda_1^{\perp}$ and they are independent of the observed jump sizes $\tilde{x}_1,\ldots,\tilde{x}_{n_1^{\perp}}$. The likelihood function of the observations concerning $S_1^{\perp}$ is then given by

$$\begin{aligned}
L_1(\tilde{\mathbf{x}}, n_1^{\perp} \mid \lambda_1^{\perp}, \boldsymbol{\theta_1}) &= e^{-\lambda_1^{\perp}(T-\tilde{T}_{n_1^{\perp}})} \prod_{i=1}^{n_1^{\perp}} \left(\lambda_1^{\perp} e^{-\lambda_1^{\perp}\tilde{Z}_i}\right) \prod_{i=1}^{n_1^{\perp}} f_1^{\perp}(\tilde{x}_i;\boldsymbol{\theta_1})\\
&= (\lambda_1^{\perp})^{n_1^{\perp}} e^{-\lambda_1^{\perp}T} \prod_{i=1}^{n_1^{\perp}} f_1^{\perp}(\tilde{x}_i;\boldsymbol{\theta_1})\,, \qquad (2.2.9)
\end{aligned}$$

as we know from Equation (2.2.4). The density $f_1^{\perp}$ is found by taking the derivative in the first equation of (2.2.8) which yields

$$\begin{aligned}
f_1^{\perp}(x) &= \frac{\partial F_1^{\perp}(x)}{\partial x} = -\frac{\partial \overline{F}_1^{\perp}(x)}{\partial x}\\
&= -\frac{1}{\lambda_1^{\perp}}\left(-\lambda_1 f_1(x) - \frac{\partial}{\partial u}\mathfrak{C}(u,\lambda_2;\boldsymbol{\delta})\,\Big|_{u=\lambda_1\overline{F}_1(x)}\,(-\lambda_1 f_1(x))\right)\\
&= \frac{\lambda_1 f_1(x)}{\lambda_1^{\perp}}\left(1 - \frac{\partial}{\partial u}\mathfrak{C}(u,\lambda_2;\boldsymbol{\delta})\,\Big|_{u=\lambda_1\overline{F}_1(x)}\right)\,.
\end{aligned}$$

Together with Equation (2.2.9) we get

$$\begin{aligned} L_1(\tilde{\mathbf{x}}, n_1^{\perp} \mid \lambda_1, \boldsymbol{\theta_1}, \boldsymbol{\delta}) &= (\lambda_1^{\perp})^{n_1^{\perp}} e^{-\lambda_1^{\perp} T} \prod_{i=1}^{n_1^{\perp}} \left[\frac{\lambda_1 f_1(\tilde{x}_i; \boldsymbol{\theta}_1)}{\lambda_1^{\perp}} \left(1 - \frac{\partial}{\partial u} \mathfrak{C}(u, \lambda_2; \boldsymbol{\delta}) \Big|_{u=\lambda_1 \overline{F}_1(\tilde{x}_i; \boldsymbol{\theta}_1)} \right) \right] \\ &= (\lambda_1)^{n_1^{\perp}} e^{-\lambda_1^{\perp} T} \prod_{i=1}^{n_1^{\perp}} \left[f_1(\tilde{x}_i; \boldsymbol{\theta_1}) \left(1 - \frac{\partial}{\partial u} \mathfrak{C}(u, \lambda_2; \boldsymbol{\delta}) \Big|_{u=\lambda_1 \overline{F}_1(\tilde{x}_i; \boldsymbol{\theta_1})} \right) \right] . \end{aligned}$$

The second part $S_2^{\perp}$ is treated analogously and we obtain $L_2(\tilde{\mathbf{y}}, n_2^{\perp} \mid \lambda_2, \boldsymbol{\theta_2}, \boldsymbol{\delta})$.

For the joint jump part of the process $(S_1^{\parallel}, S_2^{\parallel})$ we observe the number $n^{\parallel} = n_1 - n_1^{\perp} = n_2 - n_2^{\perp}$ of joint jumps with frequency $\lambda^{\parallel}$ at times $T_1, \ldots, T_{n^{\parallel}}$ with the observed bivariate jump sizes $(x_1, y_1), \ldots (x_{n^{\parallel}}, y_{n^{\parallel}})$. Denote $Z_i = T_i - T_{i-1}$ the inter-arrival times and $F^{\parallel}(x, y)$ the joint jump distribution of the jump sizes with joint density $f^{\parallel}(x, y)$. These are observations of a jump dependent CPP with frequency parameter $\lambda^{\parallel}$ and Lévy measure concentrated on $(0, \infty)^2$. Let us recall formula (2.2.1) for the Lévy density,

$$\Pi(dx, dy) = \frac{\partial^2}{\partial u \partial v} \mathfrak{C}(u, v; \boldsymbol{\delta}) \Big|_{u=\overline{\Pi}_1(x), v=\overline{\Pi}_2(y)} \Pi_1(dx) \Pi_2(dy) .$$

The derivative $\frac{\partial^2}{\partial u \partial v} \mathfrak{C}(u, v; \boldsymbol{\delta})$ exists by assumption and it is $\Pi_1(dx) = \lambda_1 f_1(x)$ and $\Pi_2(dy) = \lambda_2 f_2(x)$ and $\Pi(dx, dy) = \lambda^{\parallel} f^{\parallel}(x, y)$. Hence, for $(x, y) \in (0, \infty)^2$, the likelihood of the joint jump process is given by

$$\begin{aligned} &L^{\parallel}(\mathbf{x}, \mathbf{y}, n^{\parallel} \mid \lambda_1, \boldsymbol{\theta_1}, \lambda_2, \boldsymbol{\theta_2}, \boldsymbol{\delta}) \\ &= (\lambda^{\parallel})^{n^{\parallel}} e^{-\lambda^{\parallel} T} \prod_{i=1}^{n^{\parallel}} f^{\parallel}(x_i, y_i; \boldsymbol{\delta}) \\ &= (\lambda^{\parallel})^{n^{\parallel}} e^{-\lambda^{\parallel} T} \prod_{i=1}^{n^{\parallel}} \left[\frac{1}{\lambda^{\parallel}} \lambda_1 f_1(x_i; \boldsymbol{\theta_1}) \lambda_2 f_2(y_i; \boldsymbol{\theta_2}) \frac{\partial^2}{\partial u \partial v} \mathfrak{C}(u, v; \boldsymbol{\delta}) \Big|_{u=\lambda_1 \overline{F}_1(x_i; \boldsymbol{\theta_1}), v=\lambda_2 \overline{F}_2(y_i; \boldsymbol{\theta_2})} \right] \\ &= (\lambda_1 \lambda_2)^{n^{\parallel}} e^{-\lambda^{\parallel} T} \prod_{i=1}^{n^{\parallel}} \left[f_1(x_i; \boldsymbol{\theta_1}) f_2(y_i; \boldsymbol{\theta_2}) \frac{\partial^2}{\partial u \partial v} \mathfrak{C}(u, v; \boldsymbol{\delta}) \Big|_{u=\lambda_1 \overline{F}_1(x_i; \boldsymbol{\theta_1}), v=\lambda_2 \overline{F}_2(y_i; \boldsymbol{\theta_2})} \right] . \end{aligned}$$

Because the components $S_1^{\perp}$, $S_2^{\perp}$ and $(S_1^{\parallel}, S_2^{\parallel})$ are independent from each other, the likelihood of the bivariate CPP is the product of the likelihoods of these components which concludes the proof.

□

2.3 Bayesian inference

In this section we briefly present the basic mathematics and notations of Bayesian data analysis. Moreover, we introduce the Bayes factors which represent a Bayesian solution for model selection. For a more detailed treatment of Bayesian statistics we refer to Gelman et al. (2004) and Gilks (1996).

We understand under the notion *Bayesian inference* the process of fitting a probability model to a set of data and, with it, being able to make inferences of quantities about which we wish to learn, but which can not be observed directly. These results are summarized by a probability distribution on the parameters of the model. Hence, the essential characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis.

The unobserved quantities $\boldsymbol{\psi} \in \mathbb{R}^m$ about which we want to draw conclusions can be two kind: either they are potentially observable quantities, such as future observations of a process, or they are quantities that are not directly observable, that is, parameters that govern the hypothetical process leading to the observed data $\mathbf{z}$. Since in our MCMC application they are the latter we throughout talk about parameters in this context. From a Bayesian perspective there is no fundamental difference between random variables and parameters of a statistical model insofar as both are considered to be random. This uncertainty is reflected in the prior distribution which contains all the prior knowledge about the parameters $\boldsymbol{\psi}$. The observation $\mathbf{z}$ contains new information about these parameters and can therefore be used to update the knowledge. Let us explain how this is formally done.

In order to make probability statements about $\boldsymbol{\psi}$ given $\mathbf{z}$ we must begin with a statistical model providing a *joint distribution* for $\boldsymbol{\psi}$ and $\mathbf{z}$. The joint density function $f(\boldsymbol{\psi}, \mathbf{z})$ can be written as a product of the *prior distribution* $\pi(\boldsymbol{\psi})$ and the *likelihood function* $f(\mathbf{z}|\boldsymbol{\psi})$:

$$f(\boldsymbol{\psi}, \mathbf{z}) = f(\mathbf{z}|\boldsymbol{\psi}) \times \pi(\boldsymbol{\psi})\,.$$

Simply conditioning on the observation $\mathbf{z}$, using the basic property of conditional probability known as Bayes' rule, yields the *posterior distribution* $f(\boldsymbol{\psi}|\mathbf{z})$ which is the object of all Bayesian inference,

$$f(\boldsymbol{\psi}|\mathbf{z}) = \frac{f(\boldsymbol{\psi}, \mathbf{z})}{f(\mathbf{z})} = \frac{f(\mathbf{z}|\boldsymbol{\psi}) \times \pi(\boldsymbol{\psi})}{f(\mathbf{z})}\,. \tag{2.3.1}$$

Here $f(\mathbf{z}) = \int f(\boldsymbol{\psi}|\mathbf{z})\pi(\boldsymbol{\psi})d\boldsymbol{\psi}$ is called the *marginal distribution*. It does, for fixed $\mathbf{z}$, not depend on $\boldsymbol{\psi}$ and can hence be considered a normalizing constant.

Most of the practical difficulties in Bayesian statistics occur when trying to evaluate the integral above. For some applications, e.g. the MCMC methods (to be introduced in Section 2.4), this problem is irrelevant, since knowledge of the fundamental proportionality

$$f(\boldsymbol{\psi}|\mathbf{z}) \propto f(\mathbf{z}|\boldsymbol{\psi}) \times \pi(\boldsymbol{\psi}) \tag{2.3.2}$$

is sufficient. Hence, Equations (2.3.1) or (2.3.2), respectively, constitute the technical core of Bayesian inference.

Since we will make use of so-called noninformative prior distributions in upcoming chapters, let us define these distributions here. A prior is said to be *noninformative* or *uniform* on the support $\mathbf{S}$, if

$$\pi(\boldsymbol{\psi}) \propto 1_{\mathrm{S}}(\boldsymbol{\psi})\,.$$

Even though the name does not suggest it, we point out that such priors may contain some information about $\boldsymbol{\psi}$, reflected in the support $\mathbf{S}$. The noninformative priors have to be distinguished from the *improper* priors which integrate to ∞. A noniformative prior may be improper, but obviously it does not have to be.

We want to clarify another notion used in the context of Bayesian inference. Instead of confidence intervals considered in the classical approach one can determine *credible intervals* for the parameters $\boldsymbol{\psi}$. A $100(1-\alpha)\%$ credible interval for a parameter $\psi_k \in \mathbb{R}$, $k = 1, \ldots, m$, is an interval $I^k := \left[I^k_{\text{left}}, I^k_{\text{right}}\right]$ for which

$$\int_{I^k} f(\psi_k|\mathbf{z})\, d\psi_k = 1 - \alpha\,,$$

where $f(\psi_k|\mathbf{z})$ denotes the *marginal posterior distribution* for ψ_k, $k = 1, \ldots, m$. If we choose I^k to be symmetric, the calculation of the $100(1-\alpha)\%$ credible interval simplifies to computing the $\alpha/2$ and $1-\alpha/2$ quantiles of $f(\psi_k|\mathbf{z})$, $k = 1, \ldots, m$. If the marginal posterior distributions are not available in closed form, one has to use the empirical quantiles.

The notion of credible interval facilitates a common-sense interpretation of statistical conclusions: a credible interval for an unknown quantity of interest can be directly regarded as having a high probability of containing the unknown quantity, in contrast to a frequentist confidence interval which may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated practice.

We now want to consider the problem of comparing different models $\{M_1, \ldots, M_L\}$ reflecting competing hypotheses about the data. Exemplary, these different assumptions may be expressed by different marginal jump size distributions, as we will see in our

real data analysis in Chapter 5. The standard Bayesian solution to assess which of the competing models is best (for fitting to an empirical data set) is to employ the *Bayes factors*.

Each of the models M_l, $l = 1, \ldots, L$, is defined by the specification of a joint distribution $f(\boldsymbol{\psi}, \mathbf{z})$ for the observation (denoted by $\mathbf{z}$) and the unobservable parameters (denoted by $\boldsymbol{\psi}$). From a Bayesian perspective it is clear that inference proceeds from $f(\boldsymbol{\psi}|\mathbf{z})$. But since different models have different sets of parameters, the posterior distribution $f(\boldsymbol{\psi}|\mathbf{z})$ does not allow us to judge a model given the observed data nor does it permit comparison amongst models. Rather, it is $f(\mathbf{z})$ which can assess model performance. Regardless of the model – and hence regardless of the set of parameters – $f(\mathbf{z})$ is a density over the space of observables which can be compared with what was actually observed. The Bayes factors are based on these densities and thus provide a relative weight of evidence for one model against the other. We say relative, since they do not compare the models itself, but rather the models in the light of the observed data.

Bayesian model selection proceeds by pairwise comparison of the models $M_1, \ldots, M_L$. Let each model M_l, $l = 1, \ldots, L$, be described by a model-specific parameter vector $\boldsymbol{\psi}^l \in \Psi^l \subset \mathbb{R}^{d_l}$, $l = 1, \ldots, L$. If $f(\mathbf{z}|M_l)$ denotes the marginal density under model M_l, the Bayes factor for model i against model j, $i, j = 1, \ldots, M$, $i \neq j$ is defined by

$$B_{ij} = \frac{f(\mathbf{z}|M_i)}{f(\mathbf{z}|M_j)}\,. \tag{2.3.3}$$

Formally, the Bayes factor arises as the ratio of the posterior odds $P(M_i|\mathbf{z})/P(M_j|\mathbf{z})$ to the prior odds $P(M_i)/P(M_j)$,

$$B_{ij} = \frac{P(M_i|\mathbf{z})/P(M_j|\mathbf{z})}{P(M_i)/P(M_j)} = \frac{P(M_i|\mathbf{z})/P(M_i)}{P(M_j|\mathbf{z})/P(M_j)} = \frac{f(\mathbf{z}|M_i)}{f(\mathbf{z}|M_j)}\,.$$

When the prior odds on the models M_i and M_j is equal to one, the Bayes factor and the posterior odds are obviously equal.

The problem that often occurs in practice – and will occur in our example, see Chapter 5 – when trying to calculate the Bayes factor (2.3.3), is the evaluation of the marginal likelihood. It is equal to the normalizing constant of the posterior density and hence, $f(\mathbf{z}|M_l) = \int_{\Psi^l} f_l(\mathbf{z}|\boldsymbol{\psi}^l)\pi_l(\boldsymbol{\psi}^l)d\boldsymbol{\psi}^l$, with $f_l(\mathbf{z}|\boldsymbol{\psi}^l)$ denoting the posterior and $\pi_l(\boldsymbol{\psi}^l)$ the prior distribution for $\boldsymbol{\psi}^l$ in model M_l, $l = 1, \ldots, L$. The computation of the integral is not trivial to carry out in the general case. MCMC methods (reduced run) can be helpful, see Chib (1995) for details.

Another way to avoid these problems is to approximate the Bayes factors. Among

others Robert and Marin (2010) suggest to use the estimates

$$\tilde{B}_{ij}(\mathbf{z}) = \frac{n_i^{-1}\sum_{k=1}^{n_i} f_i(\mathbf{z}|\boldsymbol{\psi}^{i,k})}{n_j^{-1}\sum_{k=1}^{n_j} f_j(\mathbf{z}|\boldsymbol{\psi}^{j,k})}, \qquad i,j \in \{1,\ldots,L\}\,,\ i \neq j\,,$$

where $\boldsymbol{\psi}^{i,k}$, $k = 1, ..., n_i$, and $\boldsymbol{\psi}^{j,k}$, $k = 1, ..., n_j$, are two independent samples which are generated from the prior distributions π_i and π_j, respectively. When representing the Bayes factor as

$$\begin{aligned} B_{ij}(\mathbf{z}) &= \frac{f(\mathbf{z}|M_i)}{f(\mathbf{z}|M_j)} = \frac{\int_{\Psi^i} f_i(\mathbf{z}|\boldsymbol{\psi}^i)\pi_i(\boldsymbol{\psi}^i)d\boldsymbol{\psi}^i}{\int_{\Psi^j} f_j(\mathbf{z}|\boldsymbol{\psi}^j)\pi_j(\boldsymbol{\psi}^j)d\boldsymbol{\psi}^j} \\ &= \frac{E_{\pi_i}\left[f_i(\mathbf{z}|\psi^i)\right]}{E_{\pi_j}\left[f_j(\mathbf{z}|\psi^j)\right]}, \qquad i,j \in \{1,\ldots,L\}\,,\ i \neq j\,, \end{aligned}$$

it can be shown that $\tilde{B}_{ij}(\mathbf{z})$ is a strongly consistent estimate of $B_{ij}(\mathbf{z})$, that is to say $\tilde{B}_{ij}(\mathbf{z})$ converges almost surely to $B_{ij}(\mathbf{z})$.

No matter in which form the Bayes factors are applied in practice, one important use of the them is as a summary of the evidence for model i against model j provided by the data. The strength of evidence can be expressed according to the Bayes factor scale by Jeffreys[1] (1961), see Table 2.1.

B_{ij}	Evidence for model i vs. model j
< 1	Negative (supports model j)
$1 - 3.2$	Barely worth mentioning
$3.2 - 10$	Substantial
$10 - 100$	Strong
> 100	Decisive

Table 2.1: Calibration of the Bayes factor B_{ij} for model i against j according to Jeffreys' Bayes factor scale.

[1]This scale was proposed by the British mathematician, geophysicist and astronomer Sir Harold Jeffreys, 1891-1989.

2.4 Markov chain Monte Carlo methods

To conclude the chapter about fundamentals, we finally give an introduction to *Markov chain Monte Carlo*[2] *(MCMC)* methods. They are used to draw samples from a probability distribution which – in this context – is known as the *target density*. First, let us briefly discuss the question, where the advantage of MCMC methods is, compared to other sampling procedures.

The simplest way to draw samples from a probability distribution is to make use of the so-called probability integral transform. It says that any univariate random variable can be represented as a transform of a uniform random variable via the generalized inverse F^-. The problem here is oviously that we need a closed form distribution function F, that means e.g. proportional knowledge of the target density is not enough. Thus, this procedure only covers a small number of cases and we need a more general approach to generate samples from mathematically less convenient distributions. There are several alternative techniques – including *Accept-Reject*, *importance sampling* or MCMC strategies – which only require to know the functional form of the target density f. The key to these methods is to use a simpler density from which the simulation is actually done. However, the appeal to MCMC methods is that they allow for greater universality than the other two methods mentioned. They can be used to generate random samples from virtually any target distribution known up to a normalizing constant, regardless of its analytical complexity and its dimension.

Let us now state some results concerning Markov chains in general. Afterwards, we have a look at the basic MCMC algorithm, which is called the Metropolis-Hastings algorithm, and its special case, the Gibbs sampler. For a detailed introduction and backgrounds to MCMC theory we refer Robert and Casella (2000), Liu (2001) and Chen et al. (2000).

2.4.1 Markov chains

In this section we present some fundamental notions and results for Markov chains[3] that are needed to establish the convergence of the upcoming MCMC algorithms. Let us note that we do not deal here with Markov models in continuous time (called Markov processes) since the very nature of simulation leads us to consider only discrete-time stochastic processes.

A *Markov chain* is a sequence of random variables that can be thought of as evolving

[2]The name was given by John von Neumann in reference to the casinos of Monte Carlo.
[3]Named after Andrey Andreyevich Markov, Russian mathematician, 1856-1922.

over time, with transition probability depending on the particular set in which the chain is. Thus, for the definition of the Markov chain we need the *transition kernel* which governs its evolution on a space $\mathcal{X} \subseteq \mathbb{R}^d$.

Definition 2.4.1 (Transition kernel).
A transition kernel is a function K defined on $\mathcal{X} \times \mathcal{B}(\mathcal{X})$ such that

1. *$\forall \boldsymbol{x} \in \mathcal{X}$, $K(\boldsymbol{x}, \cdot)$ is a probability measure,*
2. *$\forall A \in \mathcal{B}(\mathcal{X})$, $K(\cdot, A)$ is measurable.*

In case of Markov chains, $\mathcal{X}$ is discrete and the transition kernel simply is a (transition) matrix K with elements

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\boldsymbol{X}_n = \boldsymbol{y} | \boldsymbol{X}_{n-1} = \boldsymbol{x}) , \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} .$$

In the following, we thus write $P(\cdot, \cdot)$ and $P(\boldsymbol{x}, A) := \sum_{\boldsymbol{y} \in A} P(\boldsymbol{x}, \boldsymbol{y})$ instead of $K(\cdot, \cdot)$ and $K(\boldsymbol{x}, A)$, respectively.

Definition 2.4.2 (Markov chain).
Given a transition kernel $P(\cdot, \cdot)$, a sequence $(\boldsymbol{X}_n)_{n \in \mathbb{N}}$ of random variables is a Markov chain if, for any t, the distribution of $\boldsymbol{X}_t$ given $\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots, \boldsymbol{x}_0$ is the same as the distribution of $\boldsymbol{X}_t$ given $\boldsymbol{x}_{t-1}$, that is,

$$P(\boldsymbol{x}_k, A) := P(\boldsymbol{X}_{k+1} \in A | \boldsymbol{x}_o, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) = P(\boldsymbol{X}_{k+1} \in A | \boldsymbol{x}_k) , \qquad \boldsymbol{x}_k \in \mathcal{X} , \, A \in \mathcal{B}(\mathcal{X}) .$$

So, if the initial distribution or the initial state is known, the construction of the Markov chain is entirely determined by its transition. The nth-step-ahead transition kernel is then given by

$$P^{(n)}(\boldsymbol{x}, A) = \sum_{\boldsymbol{y} \in \mathcal{X}} P^{(n-1)}(\boldsymbol{y}, A) P(\boldsymbol{x}, \boldsymbol{y})$$

where $P^{(1)}(\boldsymbol{x}, \boldsymbol{y}) = P(\boldsymbol{x}, \boldsymbol{y})$.

The two properties to be introduced next are important for a Markov chain's asymptotic behaviour. *Aperiodicity* of the chain ensures that the chain does not cycle through a finite number of sets. The notion of *α-irreducibility*, where α is a probability measure, is basically the requirement that the chain is able to visit all sets with positive probability under α from any starting point in $\mathcal{X}$. This feature is crucial in the setup of MCMC algorithms, because it leads to a guarantee of convergence as we will see.

Definition 2.4.3 (Irreducibility, Aperiodicity).

- *Given a probability measure α, the Markov chain $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ with transition kernel $P(\cdot,\cdot)$ is α-irreducible if, for every $A \in \mathcal{B}(\mathcal{X})$ with $\alpha(A) > 0$, there exists n such that $P^{(n)}(\boldsymbol{x}, A) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$.*

- *The Markov chain $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ with transition kernel $P(\cdot,\cdot)$ is said to be aperiodic, if, for all $\boldsymbol{x} \in \mathcal{X}$, the greatest divider of $\left\{n : P^{(n)}(\boldsymbol{x}, \boldsymbol{x}) > 0\right\}$ is 1.*

The Markov chains encountered in MCMC settings enjoy a very strong stability property, namely that an *invariant distribution* exists by construction. We prove this in Section 2.4.2. Therefore we need a property of Markov chains which is called *reversibility*. Let us define these two notions next.

Definition 2.4.4 (Invariant distribution).
A σ-finite measure α is invariant on $\mathcal{X}$, if, for all $\boldsymbol{x} \in \mathcal{X}$,

$$\alpha(A) = \sum_{\boldsymbol{x}\in\mathcal{X}} P(\boldsymbol{x}, A)\, \alpha(\boldsymbol{x})\,, \qquad \forall A \in \mathcal{B}(\mathcal{X})\,.$$

If α is a probability measure, the invariant distribution is also referred to as *stationary*, since $\boldsymbol{X}_n \sim \alpha$, for any n, implies that all subsequent elements of the chain are also distributed according to α. Thus, the chain is stationary in distribution.

Definition 2.4.5 (Reversibility).
The Markov chain $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ with transition kernel $P(\cdot,\cdot)$ is reversible, if, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

$$g(\boldsymbol{x})P(\boldsymbol{x}, \boldsymbol{y}) = g(\boldsymbol{y})P(\boldsymbol{y}, \boldsymbol{x}) \tag{2.4.1}$$

for a density g.

If the reversibility condition holds, then g is an invariant distribution, since

$$\sum_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x})P(\boldsymbol{x}, A) = \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{y}\in A} g(\boldsymbol{x})P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{y}\in A} g(\boldsymbol{y})P(\boldsymbol{y}, \boldsymbol{x}) = \sum_{\boldsymbol{y}\in A} g(\boldsymbol{y}) = g(A)\,.$$

These definitions allow us to state the following results which are fundamental for MCMC methods. The first theorem gives conditions under which a strong law of large numbers holds.

Theorem 2.4.6.
Suppose $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ is an α-irreducible, aperiodic Markov chain with transition kernel $P(\cdot,\cdot)$ and invariant distribution α. If $P(\boldsymbol{x},\cdot)$ is absolutely continuous with respect to α for all $\boldsymbol{x}\in\mathcal{X}$, then α is the unique invariant distribution of $P(\cdot,\cdot)$ and for all α-integrable real-valued functions h,

$$\frac{1}{M}\sum_{n=1}^{M} h(\boldsymbol{X}_n) \to \int h(\boldsymbol{x})\alpha(\boldsymbol{x})d\boldsymbol{x} \qquad \text{as } M\to\infty\,,\ a.s.$$

The second result gives conditions under which the probability density of the Mth iterate converges to its unique, invariant density. For the proofs of the theorems see e.g. Tierney (1994).

Theorem 2.4.7.
Suppose that $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ is an α-irreducible, aperiodic Markov chain which has transition kernel $P(\cdot,\cdot)$ and invariant distribution α. Then for α-almost every $\boldsymbol{x}\in\mathcal{X}$ and all sets $A\in\mathcal{B}(\mathcal{X})$

$$\lim_{M\to\infty} \left\| P^{(M)}(\boldsymbol{x},A) - \alpha(A) \right\| = 0\,,$$

where $\|\cdot\|$ denotes the total variation distance.

So, for almost every starting value the stationary distribution is a limiting distribution in the upper sense. However, in the context of Markov chain Monte Carlo theory, where we start the algorithm from some arbitrary point of measure zero, almost sure convergence is not enough.

We need to guarantee convergence from every starting point. In general, this can be attained when further ensuring a property called *Harris recurrence*. It basically says, that the probability of an infinite number of returns to any set A is equal to 1. Because Markov chains related to MCMC methods are, in fact, finite state-space Markov chains, the property of Harris recurrence follows directly from irreducibility, cf. Robert and Casella (2000).

In applications of MCMC methods, Theorem 2.4.6 and Theorem 2.4.7 play important roles. Since the existence of the invariant distribution is given by construction, we only have to check, if the produced chain is aperiodic and irreducible. We see in the following section that these properties can be satisfied easily. More convergence results can be found in Robert and Casella (2000).

2.4.2 Metropolis-Hastings algorithm

A general form of MCMC methods is the *Metropolis-Hastings*[4] *(MH)* algorithm which is used to draw samples from the multivariate target density. The fundamental idea of this method is to evolve a Markov chain so that the stationary distribution of it is the target distribution f. Given this, we have seen in Section 2.4.1 that quite general conditions on the Markov chain (e.g. aperiodicity and irreducibility) are sufficient to ensure that the drawn samples are coming from the target density in the limit. Some exemplary conditions which guarantee that the produced chain is aperiodic and irreducible are stated later on.

To produce values from the target density $f(\boldsymbol{\psi})$ – where $\boldsymbol{\psi}$ denotes the parameters of interest – the MH algorithm employs a *proposal density* $q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p)$. This density serves for two purposes: first, it supplies proposal values $\boldsymbol{\psi}^p$ of the parameters, given the current values $\boldsymbol{\psi}^c$, and second, it is the instrument to decide whether this values will be accepted or not. This is done with the help of the acceptance probability $\chi(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p)$. So, before stating the algorithm, we first need to define χ,

$$\chi(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p) := \begin{cases} \min\left[\frac{f(\boldsymbol{\psi}^p)\times q(\boldsymbol{\psi}^p, \boldsymbol{\psi}^c)}{f(\boldsymbol{\psi}^c)\times q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p)}, 1\right] & \text{if } f(\boldsymbol{\psi}^c) \times q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p) > 0\,, \\ 1 & \text{otherwise}\,. \end{cases} \tag{2.4.2}$$

We see that here only the ratio of the target density is involved. Hence, proportional knowledge of the target density is sufficient, the normalizing constant is not required. Denoting by I the total number of iterations, the MH sampling procedure is classical, cf. Chen et al. (2000):

Algorithm 2.4.8 (**Metropolis-Hastings algorithm).**

1. *Specify the initial value* $\boldsymbol{\psi}^{(0)}$.
2. *Repeat for i=1,...,I*
   - *Draw a proposal value* $\boldsymbol{\psi}^p$ *from* $q(\boldsymbol{\psi}^{(i-1)}, \cdot)$.
   - *Draw a sample* $u^{(i)}$ *from the uniform distribution* $U(0,1)$.
   - *Let*

$$\boldsymbol{\psi}^{(i)} := \begin{cases} \boldsymbol{\psi}^p & \textit{if } u^{(i)} \leq \chi(\boldsymbol{\psi}^{(i-1)}, \boldsymbol{\psi}^p)\,, \\ \boldsymbol{\psi}^{(i-1)} & \textit{otherwise}\,. \end{cases}$$

3. *Return the values* $\left\{\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, ..., \boldsymbol{\psi}^{(I)}\right\}$.

[4]The Greek American physicist Nicholas Constantine Metropolis (1915-1999), along with others, first proposed the algorithm for the specific case of the Boltzmann distribution. W. Keith Hastings, born 1930 in Toronto, Canada, extended the algorithm to the more general case in 1970.

We already mentioned that the MH algorithm – as any MCMC procedure – always produces chains whose stationary distribution is the target distribution by construction. Let us briefly verify that this is true.

Let $P(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p)$ be the actual transition function of the algorithm which differs from the proposal function $q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p)$: the probability that we actually make the move from $\boldsymbol{\psi}^c$ to $\boldsymbol{\psi}^p$ is equal to the proposal probability, $q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p)$, multiplied by the acceptance probability. That is,

$$P(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p) = q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p) \times \min\left\{ \frac{f(\boldsymbol{\psi}^p) \times q(\boldsymbol{\psi}^p, \boldsymbol{\psi}^c)}{f(\boldsymbol{\psi}^c) \times q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p)}, 1 \right\} .$$

Hence, for all $\boldsymbol{\psi}^c$, $\boldsymbol{\psi}^p$,

$$\begin{aligned} f(\boldsymbol{\psi}^c)P(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p) &= f(\boldsymbol{\psi}^c)q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p) \times \min\left\{ \frac{f(\boldsymbol{\psi}^p) \times q(\boldsymbol{\psi}^p, \boldsymbol{\psi}^c)}{f(\boldsymbol{\psi}^c) \times q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p)}, 1 \right\} \\ &= \min\left\{ f(\boldsymbol{\psi}^c)q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p), f(\boldsymbol{\psi}^p)q(\boldsymbol{\psi}^p, \boldsymbol{\psi}^c) \right\} , \end{aligned}$$

which is a symmetric function in $\boldsymbol{\psi}^c$ and $\boldsymbol{\psi}^p$. Thus, the reversibility condition (2.4.1) is satisfied with the target density f, which yields that f is a stationary distribution of the Markov chain, cf. Section 2.4.1.

To conclude this section, we state a proposition which ensures irreducibility and aperiodicity of the produced Markov chain. See Roberts and Tweedie (1996) for the proof.

Proposition 2.4.9.
Let f be the target density and q the proposal density. Assume that f is bounded and positive on every compact set of its support S. If there exist positive numbers ϵ and δ such that

$$q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p) > \epsilon \qquad if \qquad \|\boldsymbol{\psi}^c - \boldsymbol{\psi}^p\| < \delta ,$$

then the MH Markov chain is f-irreducible and aperiodic.

2.4.3 Gibbs sampler

The proposal transition in a MH sampler is often an arbitrary choice out of convenience. However, MH algorithms can achieve higher levels of efficiency, if they take the specifics of the target density f into account and, hence, enable to follow the local dynamics of it. A very simple and powerful conditional sampling technique to be discussed in this section is the *Gibbs sampler*[5]. It is a special case of the so-called multiple-block MH method: here

[5]The Gibbs sampler was given its name by Gelman and Gelman, who used it for analyzing Gibbs distributions on lattices. However, its applicability is not limited to Gibbs distributions, of course.

the parameter vector $\boldsymbol{\psi}$ is grouped into m blocks, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m)$, and each block is sampled separately by the MH algorithm, conditioned on the remaining blocks.

A distinctive feature of the Gibbs sampler is that it uses at each iteration the so-called *full conditional distributions*

$$f(\boldsymbol{\psi}_k|\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_{k-1}, \boldsymbol{\psi}_{k+1}, ..., \boldsymbol{\psi}_m)\,,$$

as proposal densities for $\boldsymbol{\psi}_k$, $k = 1, \ldots, m$. In doing so, high-dimensional problems can be reduced to lower dimensions which usually makes the sampling procedure more tractable. However, the main advantage of using full conditionals for construction of Markov chain moves is that no rejection is incurred at any of the sampling steps. To explain why this holds, we denote $\boldsymbol{\psi}_{-k}^{(i)} := (\boldsymbol{\psi}_1^{(i)}, \ldots, \boldsymbol{\psi}_{k-1}^{(i)}, \boldsymbol{\psi}_{k+1}^{(i-1)}, \ldots, \boldsymbol{\psi}_m^{(i-1)})$ for $k = 1, \ldots, m$, $i = 1, \ldots, I$, and write

$$f(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k}^{(i)}) = \frac{f(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k}^{(i)})}{f(\boldsymbol{\psi}_{-k}^{(i)})} \propto f(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k}^{(i)})\,.$$

The acceptance probability in Equation (2.4.2) is then given by

$$\begin{aligned}
\chi_k(\boldsymbol{\psi}_k^{(i-1)}, \boldsymbol{\psi}_k^p \mid \boldsymbol{\psi}_{-k}^{(i)}) &= \min\left[\frac{f(\boldsymbol{\psi}_k^p, \boldsymbol{\psi}_{-k}^{(i)}) \times f(\boldsymbol{\psi}_k^{(i-1)}|\boldsymbol{\psi}_{-k}^{(i)})}{f(\boldsymbol{\psi}_k^{(i-1)}, \boldsymbol{\psi}_{-k}^{(i)}) \times f(\boldsymbol{\psi}_k^p|\boldsymbol{\psi}_{-k}^{(i)})}, 1\right] \\
&= \min\left[\frac{f(\boldsymbol{\psi}_k^p, \boldsymbol{\psi}_{-k}^{(i)}) \times f(\boldsymbol{\psi}_k^{(i-1)}, \boldsymbol{\psi}_{-k}^{(i)})}{f(\boldsymbol{\psi}_k^{(i-1)}, \boldsymbol{\psi}_{-k}^{(i)}) \times f(\boldsymbol{\psi}_k^p, \boldsymbol{\psi}_{-k}^{(i)})}, 1\right] \\
&= 1\,,
\end{aligned}$$

for $k = 1, \ldots, m$, $i = 1, \ldots, I$. That is to say, every proposed value is accepted. Thus, Algorithm 2.4.8 simplifies to:

Algorithm 2.4.10 (Gibbs sampler).

1. *Specify the initial value* $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\psi}_1^{(0)}, \ldots, \boldsymbol{\psi}_m^{(0)})$.
2. *Repeat for* $i = 1, \ldots, I$
   - *Generate* $\boldsymbol{\psi}_1^{(i)}$ *from the full conditional* $f(\cdot \mid \boldsymbol{\psi}_{-1}^{(i)})$.
   - *Generate* $\boldsymbol{\psi}_2^{(i)}$ *from the full conditional* $f(\cdot \mid \boldsymbol{\psi}_{-2}^{(i)})$.
   - $\vdots$
   - *Generate* $\boldsymbol{\psi}_m^{(i)}$ *from the full conditional* $f(\cdot \mid \boldsymbol{\psi}_{-m}^{(i)})$.
3. *Return the values* $\left\{\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots, \boldsymbol{\psi}^{(I)}\right\}$.

If the full conditionals appearing above are standard distributions, sampling from them is trivial. We will see in Section 3.2 how the generation of values is actually done, if this is not the case.

Chapter 3

The MCMC sampler for Clayton models

The aim of this chapter is to develop a Markov chain Monte Carlo sampler for bivariate CPPs. We will concentrate on a special class of CPPs, called Clayton models. Note however, that other bivariate models can be treated analogously.

Proceeding from a Bayesian setting, we want to derive the posterior distribution of the parameters of a bivariate CPP for a given prior. In Section 2.2.2 we derived the full likelihood $f(\mathbf{z}|\boldsymbol{\psi})$ of two-dimensional CPPs based on Lévy copulas, see Equation (2.2.5). Together with the prior distribution $\pi(\boldsymbol{\psi})$ this likelihood determines the posterior distribution $f(\boldsymbol{\psi}|\mathbf{z})$, see Equation (2.3.1). The difficulty that occurs when trying to work with this term is to evaluate the marginal likelihood

$$f(\mathbf{z}) = \int f(\mathbf{z}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})\, d\boldsymbol{\psi}\,.$$

For general bivariate CPPs, analytical evaluation is impossible and numerical evaluation is very difficult and inaccurate. To avoid these problems we approximate the posterior distribution using MCMC methods for which proportional knowledge of the posterior is sufficient.

Sections 2.4.2 and 2.4.3 have shown that under fairly general conditions the chains produced by MCMC algorithms converge to their target density (which is the posterior distribution in our case). While such developments are obviously necessary, they are nonetheless insufficient from the point of view of the implementation of MCMC methods. We know that convergence can be 'assured', however it is not clear how fast the chains converge in practice. The problem is that the stated results do not directly result in methods of controlling the chain produced by an algorithm.

Therefore, we now want to address some practical issues of our MCMC algorithm – which is an application of the classical Gibbs sampler. As mentioned above we do this for the specific class of Clayton models. After introducing them, we discuss the important steps for adapting the Gibbs sampler to our application. In particular, we here have a closer look at the choice of proposal densities. Finally, we state our MCMC algorithm to be applied in detail.

3.1 Clayton models

In Section 2.2.2 we explained that the parametrization $\boldsymbol{\psi} = (\lambda_1, \lambda_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\delta})$ determines a bivariate CPP completely. Now we want to consider a certain family of CPP models, namely the *Clayton models*. Here the dependence structure between the two processes S_1 and S_2 is modelled by the one-parametric *Clayton Lévy copula* with parameter $\delta > 0$,

$$\mathfrak{C}(u, v) = (u^{-\delta} + v^{-\delta})^{-1/\delta}\,, \qquad u, v > 0\,. \tag{3.1.1}$$

This copula covers the whole range of positive dependence: for $\delta \to 0$ we obtain independence of the marginal processes given by $\mathfrak{C}_\perp(u, v) = u\,\mathbf{1}_{v=\infty} + v\,\mathbf{1}_{u=\infty}$, and jumps in different cells never occur at the same time. For $\delta \to \infty$ we get the complete positive dependence Lévy copula given by $\mathfrak{C}_\parallel(u, v) = \min(u, v)$, and jumps always occur at the same points in time. By varying δ the component dependence changes smoothly between these two extremes.

When the Clayton Lévy copula (3.1.1) is used to model the dependence between the two components we can calculate that

$$\begin{aligned} \frac{\partial}{\partial u}\mathfrak{C}(u, v) &= \left(1 + \left(\frac{u}{v}\right)^{\delta}\right)^{-1/\delta-1}\,, \qquad u, v > 0\,, \\ \frac{\partial^2}{\partial u \partial v}\mathfrak{C}(u, v) &= (1+\delta)\,(uv)^{\delta}\left(u^{\delta} + v^{\delta}\right)^{-1/\delta-2}\,, \qquad u, v > 0\,, \end{aligned}$$

which is used when considering the full likelihood of bivariate CPPs, see (2.2.5). Moreover, the parameters $\lambda_1^\perp$, $\lambda_2^\perp$ and $\lambda^\parallel$ can be calculated as

$$\lambda^\parallel = \mathfrak{C}(\lambda_1, \lambda_2) = (\lambda_1^{-\delta} + \lambda_2^{-\delta})^{-1/\delta} \qquad \text{and} \qquad \lambda_i^\perp = \lambda_i - \lambda^\parallel\,, \qquad i = 1, 2\,. \tag{3.1.2}$$

That is to say the frequency of simultaneous jumps is a simple function of the Clayton copula parameter δ.

Summarizing we list what is needed to specify a Clayton CPP uniquely:

- λ_1, λ_2, the frequency parameters of the underlying univariate Poisson processes,
- $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, the parameter vectors of the marginal jump size distributions F_1, F_2 and
- δ, the Clayton Lévy copula parameter.

We see that Clayton models only differ in the choice of the marginal jump size distributions F_1, F_2 of the components. We will write F_1 - F_2 - Clayton model to specify the model uniquely. For models where $F_1 = F_2 = F$, we introduce the notation $(F)^2$ - Clayton model.

In principle, any parametric distribution function that is considered as an appropriate choice for modelling jump sizes can be used. In operational risk, for example, these are typically (one-sided) heavy-tailed distributions that can handle big losses.

Motivated by Chapter 5 we now want to introduce two specific Clayton models which we will consider in the following.

Example 3.1.1 (The (Burr/GPD)^2 - Clayton model).

Under a (Burr/GPD)^2 - Clayton model we understand a bivariate Clayton model where the jump sizes in both components are modelled by a sliced (five-parametric) Burr/GPD distribution.

It is composed by a Burr[1] distribution with shape parameters $c_i, k_i > 0$ (location and scale parameters could be introduced easily) and a generalized Pareto distribution[2] (GPD) with parameters $h_i > 0$ (location parameter), $\beta_i > 0$ (scale parameter) and $\xi_i > 0$ (tail parameter), $i = 1, 2$. The transition between the two distributions is executed at the fixed thresholds $u_i > 0$, $i = 1, 2$. That is to say, we apply the Burr distribution for $z_i \leq u_i$, whereas for $z_i > u_i$ the GPD is used. For $z_i > 0$, the tail distribution is given by

$$\overline{F}_i(z_i) = A_i \left(\mathbf{1}_{(0,u_i]} \left(A_i^{-1} - 1 + (1 + z_i^{c_i})^{-k_i} \right) + \mathbf{1}_{(u_i,\infty)} \left(1 + \xi_i \frac{z_i + h_i - u_i}{\beta_i} \right)^{-1/\xi_i} \right).$$

The constant A_i is due to the truncation and is defined by

$$A_i := \left(1 - (1 + u_i^{c_i})^{-k_i} + \left(1 + \xi_i \frac{h_i}{\beta_i} \right)^{-1/\xi_i} \right)^{-1}.$$

[1]The Burr distribution – also known as the Singh-Maddala distribution – is named after the American Irving Wingate Burr, 1908-1989. It is most commonly used to model household income and insurance claims.

[2]The Pareto distribution is named after the Italian economist Vilfredo Pareto, 1848-1923. The distribution often describes social, scientific, geophysical, actuarial and many other types of observable phenomena very well.

The corresponding density for $z_i > 0$ can be written as

$$f_i(z_i) = A_i \left(\mathbf{1}_{(0,u_i]}\, c_i k_i z_i^{c_i-1} \left(1+z_i^{c_i}\right)^{-(k_i+1)} + \mathbf{1}_{(u_i,\infty)} \frac{1}{\beta_i} \left(1+\xi_i \frac{z_i+h_i-u_i}{\beta_i}\right)^{-1/\xi_i-1} \right).$$

For $\boldsymbol{\theta}_i = (\lambda_i, c_i, k_i, h_i, \beta_i, \xi_i)$, $i = 1, 2$, the likelihood function of a bivariate CPP in the (Burr/GPD)^2 - Clayton model is then given by

$$\begin{aligned}
&f\left(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathbf{x}, \mathbf{y}, n_1^{\perp}, n_2^{\perp}, n^{\parallel} \mid \lambda_1, \boldsymbol{\theta}_1, \lambda_2, \boldsymbol{\theta}_2, \delta\right) \\
&= (\lambda_1 A_1)^{n_1^{\perp}} (\lambda_2 A_2)^{n_2^{\perp}} \left((1+\delta)(\lambda_1\lambda_2 A_1 A_2)^{1+\delta}\right)^{n^{\parallel}} e^{-T\left(\lambda_1^{\perp}+\lambda_2^{\perp}+\lambda^{\parallel}\right)} \qquad (3.1.3) \\
&\times \prod_{i=1}^{n_1^{\perp}} \left[\left(\mathbf{1}_{(0,u_1]}\, c_1 k_1 \tilde{x}_i^{c_1-1} \left(1+\tilde{x}_i^{c_1}\right)^{-(k_1+1)} + \mathbf{1}_{(u_1,\infty)} \frac{1}{\beta_1} \left(1+\xi_1 \frac{\tilde{x}_i+h_1-u_1}{\beta_1}\right)^{-\frac{1}{\xi_1}-1} \right) \right. \\
&\left. \left(1-\left(1+\left(\frac{\lambda_1}{\lambda_2}A_1\right)^{\delta}\left(\mathbf{1}_{(0,u_1]}\left(A_1^{-1}-1+\left(1+\tilde{x}_i^{c_1}\right)^{-k_1}\right)+\mathbf{1}_{(u_1,\infty)}\left(1+\xi_1\frac{\tilde{x}_i+h_1-u_1}{\beta_1}\right)^{-\frac{1}{\xi_1}}\right)^{\delta}\right)^{-\frac{1}{\delta}-1}\right)\right] \\
&\times \prod_{i=1}^{n_2^{\perp}} \left[\left(\mathbf{1}_{(0,u_2]}\, c_2 k_2 \tilde{y}_i^{c_2-1} \left(1+\tilde{y}_i^{c_2}\right)^{-(k_2+1)} + \mathbf{1}_{(u_2,\infty)} \frac{1}{\beta_2} \left(1+\xi_2 \frac{\tilde{y}_i+h_2-u_2}{\beta_2}\right)^{-\frac{1}{\xi_2}-1} \right) \right. \\
&\left. \left(1-\left(1+\left(\frac{\lambda_2}{\lambda_1}A_2\right)^{\delta}\left(\mathbf{1}_{(0,u_2]}\left(A_2^{-1}-1+\left(1+\tilde{y}_i^{c_2}\right)^{-k_2}\right)+\mathbf{1}_{(u_2,\infty)}\left(1+\xi_2\frac{\tilde{y}_i+h_2-u_2}{\beta_2}\right)^{-\frac{1}{\xi_2}}\right)^{\delta}\right)^{-\frac{1}{\delta}-1}\right)\right] \\
&\times \prod_{i=1}^{n^{\parallel}} \left[\left(\mathbf{1}_{(0,u_1]}\, c_1 k_1 x_i^{c_1-1} \left(1+x_i^{c_1}\right)^{-(k_1+1)} + \mathbf{1}_{(u_1,\infty)} \frac{1}{\beta_1} \left(1+\xi_1 \frac{x_i+h_1-u_1}{\beta_1}\right)^{-\frac{1}{\xi_1}-1} \right) \right. \\
&\left(\mathbf{1}_{(0,u_2]}\, c_2 k_2 y_i^{c_2-1} \left(1+y_i^{c_2}\right)^{-(k_2+1)} + \mathbf{1}_{(u_2,\infty)} \frac{1}{\beta_2} \left(1+\xi_2 \frac{y_i+h_2-u_2}{\beta_2}\right)^{-\frac{1}{\xi_2}-1} \right) \\
&\left(\mathbf{1}_{(0,u_1]} \left(A_1^{-1}-1+\left(1+x_i^{c_1}\right)^{-k_1}\right) + \mathbf{1}_{(u_1,\infty)} \left(1+\xi_1\frac{x_i+h_1-u_1}{\beta_1}\right)^{-\frac{1}{\xi_1}} \right)^{\delta} \\
&\left(\mathbf{1}_{(0,u_2]} \left(A_2^{-1}-1+\left(1+y_i^{c_2}\right)^{-k_2}\right) + \mathbf{1}_{(u_2,\infty)} \left(1+\xi_2\frac{y_i+h_2-u_2}{\beta_2}\right)^{-\frac{1}{\xi_2}} \right)^{\delta} \\
&\left(\lambda_1^{\delta} A_1^{\delta} \left(\mathbf{1}_{(0,u_1]} \left(A_1^{-1}-1+\left(1+x_i^{c_1}\right)^{-k_1}\right) + \mathbf{1}_{(u_1,\infty)} \left(1+\xi_1\frac{x_i+h_1-u_1}{\beta_1}\right)^{-\frac{1}{\xi_1}} \right)^{\delta} \right. \\
&\left. \left. +\lambda_1^{\delta} A_2^{\delta} \left(\mathbf{1}_{(0,u_2]} \left(A_2^{-1}-1+\left(1+y_i^{c_2}\right)^{-k_2}\right) + \mathbf{1}_{(u_2,\infty)} \left(1+\xi_2\frac{y_i+h_2-u_2}{\beta_2}\right)^{-\frac{1}{\xi_2}} \right)^{\delta} \right)^{-\frac{1}{\delta}-2} \right].
\end{aligned}$$

Example 3.1.2 (The (Weibull)^2 - Clayton model).

In the (Weibull)^2 - Clayton model the jump sizes in both components of the bivariate compound Poisson process are modelled by a Weibull distribution with parameters $a_i > 0$ (scale parameter) and $b_i > 0$ (shape parameter), $i = 1, 2$. The tail distribution is then given by

$$\overline{F}_i\left(z_i; a_i, b_i\right) = e^{-\left(\frac{z_i}{a_i}\right)^{b_i}}, \qquad z_i \geq 0\,,$$

with density

$$f_i\left(z_i\right) = \frac{b_i}{a_i^{b_i}} z_i^{b_i-1} e^{-\left(\frac{z_i}{a_i}\right)^{b_i}}, \qquad z_i \geq 0\,.$$

With the notation as in (2.2.5) the likelihood function of a bivariate CPP in the (Weibull)^2 - Clayton model is thus given by

$$\begin{aligned}
&f\left(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathbf{x}, \mathbf{y}, n_1^{\perp}, n_2^{\perp}, n^{\parallel} \;\middle|\; \lambda_1, a_1, b_1, \lambda_2, a_2, b_2, \delta\right) \\
&= \left(\lambda_1 b_1 a_1^{-b_1}\right)^{n_1^{\perp}} e^{-\lambda_1^{\perp} T - \sum_{i=1}^{n_1^{\perp}} (\tilde{x}_i/a_1)^{b_1}} \prod_{i=1}^{n_1^{\perp}} \left[\tilde{x}_i^{b_1-1} \left(1 - \left(1 + \left(\frac{\lambda_1 e^{-(\tilde{x}_i/a_1)^{b_1}}}{\lambda_2}\right)^{\delta}\right)^{-1/\delta-1}\right)\right] \\
&\times \left(\lambda_2 b_2 a_2^{-b_2}\right)^{n_2^{\perp}} e^{-\lambda_2^{\perp} T - \sum_{i=1}^{n_2^{\perp}} (\tilde{y}_i/a_2)^{b_2}} \prod_{i=1}^{n_2^{\perp}} \left[\tilde{y}_i^{b_2-1} \left(1 - \left(1 + \left(\frac{\lambda_2 e^{-(\tilde{y}_i/a_2)^{b_2}}}{\lambda_1}\right)^{\delta}\right)^{-1/\delta-1}\right)\right] \\
&\times \left((1+\delta)\left(\lambda_1\lambda_2\right)^{1+\delta} b_1 b_2 a_1^{-b_1} a_2^{-b_2}\right)^{n^{\parallel}} e^{-\lambda^{\parallel} T - (1+\delta)\sum_{i=1}^{n^{\parallel}} \left((x_i/a_1)^{b_1} + (y_i/a_2)^{b_2}\right)} \qquad (3.1.4) \\
&\times \prod_{i=1}^{n_1^{\parallel}} \left[x_i^{b_1-1} y_i^{b_2-1} \left(\left(\lambda_1 e^{-(x_i/a_1)^{b_1}}\right)^{\delta} + \left(\lambda_2 e^{-(y_i/a_2)^{b_2}}\right)^{\delta}\right)^{-1/\delta-2}\right].
\end{aligned}$$

3.2 Adaption of the sampler for Clayton models

We approximate the posterior distributions of our models by an adapted Gibbs sampling procedure using the MATLAB software. Now we want to explain the most important features of our developed sampler. Thereby we follow the main steps that are required for implementing the Gibbs sampler which are:

- starting values must be provided;
- methods for sampling from the full conditional distributions must be determined;
- the output must be analyzed;
- the length of the burn-in period, the number of iterations and the subsampling procedure must be specified.

3.2.1 Initialization

The very first step in every sampling procedure is the choice of initial values. Since the convergence results from Sections 2.4.1 and 2.4.2 are independent from the starting values, theoretically any initial value can be taken. Also practically it is not necessary to expend much effort in choosing starting values, since their choice is unimportant if the Gibbs sampler (or any other MCMC sampler) is run long enough to 'forget' its initial states. To decide what is 'long enough' in the particular case, it is useful to perform a number of runs with widely dispersed starting values. A rapidly mixing chain will quickly find its way from extreme starting values. Initial values may need to be chosen more carefully for slow-mixing chains, to avoid lengthy burn-in.

The convergence behaviour of the sampler is very satisfying in our case. From several simulation studies we found that even if the initial values have been chosen quite far away from the values used for simulation, the chain converged fast. However, a sophisticated choice of starting values can avoid a very long burn-in. That is why we suggest maximum likelihood estimates (MLE). They can be calculated quite efficiently since the maximization of the likelihood for the Clayton models is only a relative small dimensional problem. Therefore, we will use MLEs as initial values throughout, if not stated otherwise.

3.2.2 Sampling from full conditional distributions

For reasons given later we use the Gibbs sampler for our application. When applying the Gibbs sampler (Algorithm 2.4.10), we have to draw serial values of the parameter blocks $\boldsymbol{\psi}_k$ from the corresponding full conditionals $f(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k}, \mathbf{z})$, $k = 1, \ldots, m$. In our case each parameter forms a block by itself, i.e. we sample from one-dimensional distributions. Considering that

$$f(\psi_k|\boldsymbol{\psi}_{-k}, \mathbf{z}) = \frac{f(\boldsymbol{\psi}, \mathbf{z})}{f(\boldsymbol{\psi}_{-k}, \mathbf{z})} = \frac{f(\mathbf{z}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{f(\boldsymbol{\psi}_{-k}, \mathbf{z})} \propto f(\mathbf{z}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})\,, \qquad k = 1, \ldots, m\,, \tag{3.2.1}$$

and looking at Equation (2.2.5) it gets clear immediately that all full conditional distributions appearing in our application are non-standard. This means that simulation from these functions is not trivial. Hence, (one-dimensional) Metropolis-Hastings steps have to be conducted to simulate from the full conditionals and our Gibbs sampler can be seen as a combination of MH steps applied to the different components. Note that in each sampling step only one iteration of MH is required, because if $\boldsymbol{\psi}^{(i)}$ is from the posterior distribution $f(\cdot|\mathbf{z})$, then so is $(\psi_k^{(i+1)}, \boldsymbol{\psi}_{-k}^{(i+1)})$, $k = 1, \ldots, m$.

It is essential that sampling from our full conditional distributions is highly efficient computationally. Thus, the question arises how to choose the required proposal densities. As seen in Section 2.4.2 proposal densities are needed to generate samples from non-standard target densities. Their function is double: first, they supply proposal values and, second, they help to decide whether these values are accepted as coming from the target density, or not. In particular the second purpose influences the performance of the MCMC sampler significantly. That is why we now want to emphasize the importance of good proposals, before determining the proposal densities of our algorithm for the Clayton models.

The need for good proposal densities

One of the most appealing aspects of MCMC algorithms is their universality. That is, the fact that an arbitrary proposal distribution q, which has the same support as the target density f, will ultimately deliver samples from f and, thus, will lead to the simulation of the target density. However, this universality may be only a formality if the proposal distribution q only rarely simulates points in the region where most of the mass of the target distribution f is located. Knowing that the probability density of the produced chain converges to the target distribution *at any time*, is not satisfying for practical applications.

That means, even though theoretically the choice of the proposal density does not matter (as long as the proposal has an adequate support), that is to say any density could be taken, in practice the selection of good proposal densities is the decisive point in the setup of a MCMC procedure. First, the rate of convergence to the stationary distribution depends crucially on the relationship between the proposal q and the target density f, cf. Gilks et al. (1996). But moreover, even when the chain has 'converged', it may still mix slowly (i.e. move slowly around the support of f). The consequence would be that the sampler has to be run much longer to obtain reliable estimates. On the other hand, the better a proposal density is adapted to the target density, the better the mixing of the produced Markov chains will be; as illustrated by Figure 3.1 it is obvious that there

will be – generally speaking – less MH rejections and, hence, the acceptance rate will be higher, if the proposal mimics the target density very well.

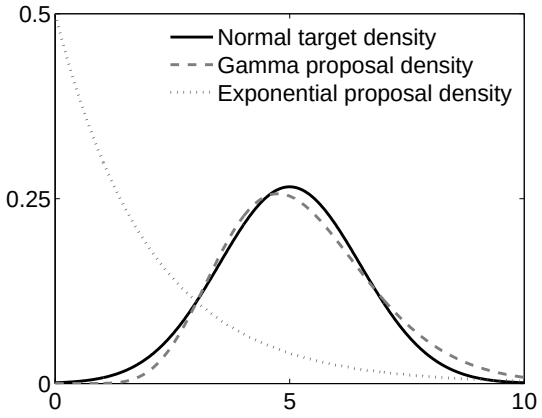


Figure 3.1: Comparison of proposal densities for a normal target density. The acceptance probability of the MCMC sampler depends on ratios of the target density to the proposal density.

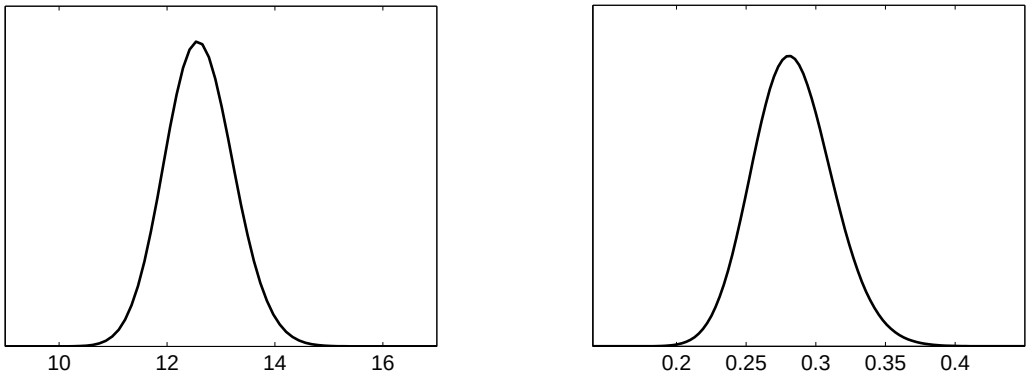
To meet these requirement it is often necessary, especially in high-dimensional problems, to perform exploratory analyses to determine roughly the shape of f. This will help in constructing a proposal q which leads to rapid mixing and fast convergence. Thereby, progress in practice often depends on experimentation.

Apart from these considerations, there is an other aspect we finally want to mention. During the sampling procedure we have to simulate many times from the proposal densities. So, for computational efficiency, the proposals should be chosen in such a way that sampling and evaluation can be conducted easily.

Choice of proposal densities for Clayton models

Since the Gibbs sampler in our case uses m univariate updates we have to select proposal densities for each of the full conditionals $f(\psi_k|\boldsymbol{\psi}_{-k}, \mathbf{z})$, $k = 1, \ldots, m$. To get an idea of the shape of the full conditionals it is helpful to make use of graphical analyzing tools. Hence, we simulate data sets from the models (see Chapter 4) and plot the full conditionals given the data, where all parameters except one are fixed to the known values from the simulation. Here we make use of the proportionality relation in Equation (3.2.1).

For several Clayton models and various parameter sets and simulations, these full conditionals looked quite symmetric and unimodal, see Figure 3.2. Thus, normal proposal distributions seem to be a promising choice. They are very convenient, since simulating from them as well as evaluating their densities can be carried out efficiently: implemen-



(a) λ_1 ((Weibull)^2 - Clayton model) (b) k_1 ((Burr/GPD)^2 - Clayton model)

Figure 3.2: Density estimates of full conditionals for two parameters of two different Clayton model simulations. The densities are given up to a normalizing constant.

tations are available in every common statistical package. Using normal proposals is a special case of the so-called *independence sampler*. Here the proposal does not depend on the current value $\boldsymbol{\psi}^c$,

$$q(\boldsymbol{\psi}^c, \boldsymbol{\psi}^p) = q(\boldsymbol{\psi}^p)\,.$$

In case of normally distributed proposal densities it can also be shown quite easily, by employing Proposition 2.4.9, that the produced chain is aperiodic and irreducible which ensures that the drawn samples can be regarded as coming from the posterior distribution, see Section 2.4.1.

To fully determine our proposal densities we have to specify the location parameters μ_k and the scale parameters σ_k, $k = 1, \ldots, m$. To achieve good acceptance rates for the MH steps, the expected value μ_k and the variance σ_k^2 are adapted several times during the sampling procedure to fit the full conditional distributions best possible. This is done by the following approach.

The parameters μ_k, $k = 1, \ldots, m$, are set to the values which maximize (for fixed current values of the other parameters) the full conditionals $f(\psi_k \mid \boldsymbol{\psi}_{-k}, \mathbf{z})$, each. Finding these maxima is a univariate optimization problem each and requires only few computation time. The standard deviations σ_k, $k = 1, \ldots, m$, are determined using a specific property of the normal distribution density. Denoting by $g(x; \mu, \sigma^2)$ the density of the univariate normal distribution with mean μ and variance σ^2 evaluated at x, one easily derives that the second derivative of g is

$$g''(x; \mu, \sigma^2) = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}(x-\mu)^2 - 1\right) g(x; \mu, \sigma^2)\,.$$

Hence, $\sigma^2 = -g(\mu;\mu,\sigma^2)/g''(\mu;\mu,\sigma^2)$, and a sensible choice for the variance of the proposal density is given by

$$\sigma_k^2 = -\frac{f(\mu_k|\boldsymbol{\psi}_{-k},\mathbf{z})}{f''(\psi_k|\boldsymbol{\psi}_{-k},\mathbf{z})|_{\mu_k}}\,, \qquad k = 1,\ldots,m\,,$$

which can be evaluated using numerical approximations of the second derivative of the full conditional $f(\cdot|\boldsymbol{\psi}_{-k},\mathbf{z})$ at μ_k.

Several simulation studies have shown that it is not necessary to adapt μ_k and σ_k^2, $k = 1,\ldots,m$, in each iteration of our MCMC sampling algorithm, except of the burn-in phase, where adaption in each iteration improves the mixing of the sampler significantly. Subsequently it is enough to redetermine the parameters of the proposal in every fiftieth iteration.

3.2.3 Analyzing the output

The values for the parameters of interest generated by the sampler must be graphically and statistically summarized to check mixing and convergence. Given the MCMC output $\left\{\psi_k^{(i)},\, i = 1,\ldots,I\right\}$ for every parameter ψ_k, $k = 1,\ldots,m$, there are several analyzing tools we make use of. Here we want to introduce them briefly.

Let B denote the length of burn-in (to be discussed in the next section), the *posterior mean estimates* are then given by

$$\overline{\psi}_k = \frac{1}{I-B}\sum_{i=B+1}^{I}\psi_k^{(i)}\,, \qquad k = 1,\ldots,m\,.$$

We will assess the performance of our sampler using these posterior means in the Sections 4.3 and 4.4.

Our estimation of the marginal posterior distributions is based on *kernel density estimations* with normal kernel functions, using a bandwidth h which is a function of the number of samples. That is, the approximated marginal posterior distributions are functions $\tilde{f}_k(\cdot|\mathbf{z}) : \mathbb{R} \to \mathbb{R}^+$,

$$\tilde{f}_k(\psi_k|\mathbf{z}) = \frac{1}{(I-B)h}\sum_{i=B+1}^{I}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{\sqrt{2\pi}}\left(\frac{\psi_k-\psi_k^{(i)}}{h}\right)^2\right)\,,$$

for $k = 1,\ldots,m$.

To assess the convergence and mixing behaviour of the produced chains we apply

two very simple but powerful visual methods. First, we plot the sample paths which are defined as the functions $s_k : \mathbb{N} \to \mathbb{R}$,

$$s_k(i) = \psi_k^{(i)}\,, \qquad k = 1, \ldots, m\,.$$

Second, we analyze the autocorrelation functions $\rho_k : \mathbb{N} \to [-1, 1]$,

$$\rho_k(z) = \frac{\sum_{i=B+1}^{I-z}(\psi_k^{(i+z)} - \overline{\psi}_k)(\psi_k^{(i)} - \overline{\psi}_k)}{\sum_{i=B+1}^{I}(\psi_k^{(i)} - \overline{\psi}_k)^2}\,, \qquad k = 1, \ldots, m\,.$$

3.2.4 Burn-in, number of iterations and subsampling

We have seen in the theoretical part of this thesis that under quite general conditions the produced Markov chain converges to its stationary distribution which is the target density. But still we have to be aware that this theoretically holds only in the limit: stationarity is only achieved asymptotically. Hence, the simulated values can be regarded as coming from the target density $f(\cdot|\mathbf{z})$ only when the number of iterations has become very large. To cope with this fact we conduct an initial burn-in phase, after which the chain is assumed to have converged. The drawn samples which are simulated during burn-in are not considered anymore and the subsequent values are supposed to be approximate draws from the posterior distribution.

The length of burn-in depends on the starting value and on the rate of convergence. We note, that the use of maximum likelihood estimates as initial values – and thus values which are close to the mode of $f(\cdot|\mathbf{z})$ – does not remove the need for burn-in, because we want to assure that the samples are 'independent' from the starting position. With the graphical methods presented in the previous section we found from several simulation studies that our produced chains converge quite fast. The good effective convergence behaviour allows us to use only 1000 iterations for burn-in.

Deciding when to stop the chain is an important practical matter. The most obvious informal method for determining the sample size I is again the monitoring of the sample paths for different chains. In our case we set $I = 21000$, including the burn-in period.

The last practical aspect we want to address is the subsampling of simulated values. As we will see in Chapter 4 the produced Markov chains for the parameters of interest show significant autocorrelations over several time lags. One reason for this is obviously the nature of the MH steps which allows to replicate the current value when the proposed value is rejected. Even though autocorrelations are not problematic (there are even applications for which autocorrelations are advantageous, e.g. the calculation of the posterior mean, cf. Robert and Casella (2004)), one often wants to obtain samples that can be considered to

be more or less 'independent'. Practically, to reduce the dependence between the samples from our MCMC algorithm, we subsample; that is, we use only every 10th iteration to approximate the posterior distribution. Subsampling in our application is possible, since the computation time for 1000 iterations is moderate in our examples. Therefore, in order to get 2000 samples from the posterior distribution we run the algorithm always for 21000 iterations and then use iterations $1010, 1020, \ldots, 21000$.

3.3 The sampling algorithm

To conclude the Chapter we want to state our developed MCMC sampler. As mentioned before it is an application of the Gibbs sampler using MH steps for simulation from the full conditionals. With the notation as in Section 2.4.2 and 2.4.3 it yields that the acceptance probability is given by

$$\chi_k(\psi_k^{(i-1)}, \psi_k^p \mid \boldsymbol{\psi}_{-k}^{(i)}) = \begin{cases} \min\left[\frac{f(\mathbf{z}|\psi_k^p, \boldsymbol{\psi}_{-k}^{(i)}) \times \pi(\psi_k^p, \boldsymbol{\psi}_{-k}^{(i)}) \times g(\psi_k^{(i-1)}; \mu_k, \sigma_k^2)}{f(\mathbf{z}|\psi_k^{(i-1)}, \boldsymbol{\psi}_{-k}^{(i)}) \times \pi(\psi_k^{(i-1)}, \boldsymbol{\psi}_{-k}^{(i)}) \times g(\psi_k^p; \mu_k, \sigma_k^2)}, 1\right] & \text{if denom. } > 0\,, \\ 1 & \text{otherwise}\,. \end{cases}$$

Algorithm 3.3.1.

1. *Compute the MLEs and use them as initial values:* $\boldsymbol{\psi}^{(0)} = (\psi_1^{(0)}, \ldots, \psi_m^{(0)})$.
2. *Repeat for i=1,. . . ,21000*

   Repeat for k=1,. . . ,m

   - *If* $\left(i \leq 1000 \quad or \quad \frac{i}{50} \in \mathbb{N}\right)$

     $\mu_k = \text{argmax}\left\{f(\psi_k|\boldsymbol{\psi}_{-k}^{(i)}, \mathbf{z})\right\}$,

     $\sigma_k^2 = -\,f(\mu_k|\boldsymbol{\psi}_{-k}^{(i)}, \mathbf{z}) \,/\, f''(\psi_k|\boldsymbol{\psi}_{-k}^{(i)}, \mathbf{z})|_{\mu_k}$.
   - *Draw a proposal value* ψ_k^p *from* $N(\mu_k, \sigma_k^2)$.
   - *Draw a sample* $u_k^{(i)}$ *from the uniform distribution* $U(0,1)$.
   - *Let*

$$\psi_k^{(i)} := \begin{cases} \psi_k^p & \text{if} \quad u_k^{(i)} \leq \chi_k\left(\psi_k^{(i-1)}, \psi_k^p \mid \boldsymbol{\psi}_{-k}^{(i)}\right), \\ \psi_k^{(i-1)} & \text{otherwise}, \end{cases}$$

   with $\boldsymbol{\psi}_{-k}^{(i)} := \left(\psi_1^{(i)}, \ldots, \psi_{k-1}^{(i)}, \psi_{k+1}^{(i-1)}, \ldots, \psi_m^{(i-1)}\right)$.

3. *Return the values* $\left\{\boldsymbol{\psi}^{(1010)}, \boldsymbol{\psi}^{(1020)}, \boldsymbol{\psi}^{(1030)}, \ldots, \boldsymbol{\psi}^{(21000)}\right\}$.

We apply this sampler for the Clayton models in Chapter 4 and Chapter 5 for simulation and real data studies, respectively.

Chapter 4

Simulation study

In this Chapter we assess mixing and convergence behaviour of the produced chains and the quality of the posterior mean estimates using simulated data. Furthermore, we compare the posterior mean estimates to the maximum likelihood estimates to check whether there are significant differences. Since we therefore need to simulate sample paths of bivariate CPPs we first discuss the simulation algorithm proposed in Esmaeili and Klüppelberg (2010a) which is used for the entire study. Motivated by the real data analysis in Chapter 5 we present the results for two different Clayton models, the (Burr/GPD)^2 - Clayton model and the (Weibull)^2 - Clayton model.

4.1 Simulation algorithm for bivariate CPPs

In the representation of bivariate CPPs based on Lévy copulas we have seen that the process is fully determined by $\boldsymbol{\psi} = (\lambda_1, \boldsymbol{\theta}_1, \lambda_2, \boldsymbol{\theta}_2, \boldsymbol{\delta})$, cf. Section 2.2.2. For simulation from this parametric model, we assume that we are given this parameter set and a time interval $[0, T]$ for prespecified $T > 0$.

We now state the simulation algorithm for bivariate CPPs which is an extension of Algorithm 6.2 of Cont and Tankov (2004) to two dimensions. It makes use of the decomposition of the two components of the process. Let us therefore reconsider Equations (2.2.2) and (2.2.3). Here $(S_1^{\parallel}, S_2^{\parallel})$ is the dependent part and $S_1^{\perp}$, $S_2^{\perp}$ are the independent parts with jump intensities $\lambda^{\parallel} = \mathfrak{C}(\lambda_1, \lambda_2; \boldsymbol{\delta})$ and $\lambda_1^{\perp} = \lambda_1 - \lambda^{\parallel}$, $\lambda_2^{\perp} = \lambda_2 - \lambda^{\parallel}$. The algorithm simulates the jump times and the jump sizes of these three components independently. The functions $F_1^{\perp}$, $F_2^{\perp}$ and $F^{\parallel}$ from Equation (2.2.8) describe the distribution functions

of $S_1^{\perp}$, $S_2^{\perp}$ and $(S_1^{\parallel}, S_2^{\parallel})$, respectively:

$$\begin{aligned}
\overline{F}_1^{\perp}(x) &= \frac{1}{\lambda_1^{\perp}}\left(\lambda_1\overline{F}_1(x) - \mathfrak{C}\left(\lambda_1\overline{F}_1(x), \lambda_2; \boldsymbol{\delta}\right)\right), && (4.1.1)\\
\overline{F}_2^{\perp}(y) &= \frac{1}{\lambda_2^{\perp}}\left(\lambda_2\overline{F}_2(y) - \mathfrak{C}\left(\lambda_1, \lambda_2\overline{F}_2(y); \boldsymbol{\delta}\right)\right), && (4.1.2)\\
\overline{F}^{\parallel}(x,y) &= \frac{1}{\lambda^{\parallel}}\,\mathfrak{C}\left(\lambda_1\overline{F}_1(x), \lambda_2\overline{F}_2(y); \boldsymbol{\delta}\right).
\end{aligned}$$

When we further assume that the jump distributions F_1 and F_2 have no atom at 0, it yields for the margins of the bivariate distribution function of the joint jumps that

$$\begin{aligned}
\overline{F}_1^{\parallel}(x) &= \lim_{y\to 0}\overline{F}^{\parallel}(x,y) = \lim_{y\to 0}\frac{1}{\lambda^{\parallel}}\,\mathfrak{C}\left(\lambda_1\overline{F}_1(x), \lambda_2\overline{F}_2(y); \boldsymbol{\delta}\right), && (4.1.3)\\
\overline{F}_2^{\parallel}(y) &= \lim_{x\to 0}\overline{F}^{\parallel}(x,y) = \lim_{x\to 0}\frac{1}{\lambda^{\parallel}}\,\mathfrak{C}\left(\lambda_1\overline{F}_1(x), \lambda_2\overline{F}_2(y); \boldsymbol{\delta}\right).
\end{aligned}$$

Let us denote by $\overline{C}(u,v)$ the survival copula of the joint jumps of $(S_1^{\parallel}, S_2^{\parallel})$ given by $\overline{C}\left(\overline{F}_1^{\parallel}(x), \overline{F}_1^{\parallel}(x)\right) = \overline{F}^{\parallel}(x,y)$. It follows that the distribution function of $S_2^{\parallel}$ given $S_1^{\parallel}$ equals

$$\overline{H}_x(y) := \frac{\partial}{\partial u}\overline{C}\left(u, \overline{F}_2^{\parallel}(y)\right)\Big|_{u=\overline{F}_1^{\parallel}(x)} = \left(1 + \left(\frac{\overline{F}_1^{\parallel}(x)}{\overline{F}_2^{\parallel}(y)}\right)^{\delta} - \left(\overline{F}_1^{\parallel}(x)\right)^{\delta}\right)^{-1/\delta-1}. \quad (4.1.4)$$

See Esmaeili and Klüppelberg (2010a) for details about the previous result.
Defining the generalized inverse $h^{\leftarrow}(u) := \inf\{s \in \mathbb{R} : h(s) \geq u\}$ for any increasing function h, we can finally state the algorithm.

Algorithm 4.1.1 (Simulation of a bivariate CPP).

1. *Generate random numbers $N_1(T)$, $N_2(T)$ and $N^{\parallel}(T)$ from Poisson distributions with parameters $\lambda_1 T$, $\lambda_2 T$ and $\lambda^{\parallel}T = \mathfrak{C}(\lambda_1, \lambda_2)\,T$, respectively. This implies then for the number of single jumps that $N_1^{\perp}(T) = N_1(T) - N^{\parallel}(T)$ and $N_2^{\perp}(T) = N_2(T) - N^{\parallel}(T)$.*

2. *Generate independent $[0,T]$-uniformly distributed random variables: $U_{1,i}^{\perp}$ for $i = 1, \ldots, N_1^{\perp}(T)$, $U_{2,i}^{\perp}$ for $i = 1, \ldots, N_2^{\perp}(T)$, and $U_i^{\parallel}$ for $i = 1, \ldots, N^{\parallel}(T)$. These are the Poisson points of single and joint jumps.*

3. *Generate independent standard uniform random variables: U_i for $i = 1, \ldots, N_1^{\perp}(T)$, and V_i for $i = 1, \ldots, N_2^{\perp}(T)$. Then the single jump sizes of both components are found by taking the inverse of $F_1^{\perp}$ and $F_2^{\perp}$, that is, $X_i^{\perp} \stackrel{d}{=} F_1^{\perp\leftarrow}(U_i)$, $i = 1, \ldots, N_1^{\perp}(T)$, and $Y_i^{\perp} \stackrel{d}{=} F_2^{\perp\leftarrow}(V_i)$, $i = 1, \ldots, N_2^{\perp}(T)$.*

4. *Generate new independent* $[0,1]$*-uniform random variables for the bivariate jump sizes:* U_i *for* $i = 1, \dots, N^{\parallel}(T)$*, and* V_i *for* $i = 1, \dots, N^{\parallel}(T)$*. Then* $X_i^{\parallel} \stackrel{d}{=} F_1^{\parallel \leftarrow}(U_i)$ *and, given* $X_i^{\parallel} = x$*,* $Y_i^{\parallel} \stackrel{d}{=} H_x^{\leftarrow}(V_i)$*,* $i = 1, \dots, N^{\parallel}(T)$*.*

5. *The bivariate trajectory is then given by*

$$\begin{pmatrix} S_1(t) \\ S_2(t) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N_1^{\perp}(T)} \mathbf{1}_{\left\{U_{1,i}^{\perp} \le t\right\}} X_i^{\perp} + \sum_{i=1}^{N^{\parallel}(T)} \mathbf{1}_{\left\{U_i^{\parallel} \le t\right\}} X_i^{\parallel} \\ \sum_{i=1}^{N_2^{\perp}(T)} \mathbf{1}_{\left\{U_{2,i}^{\perp} \le t\right\}} Y_i^{\perp} + \sum_{i=1}^{N^{\parallel}(T)} \mathbf{1}_{\left\{U_i^{\parallel} \le t\right\}} Y_i^{\parallel} \end{pmatrix}, \qquad 0 < t < T\,.$$

Since we conduct our simulation study for the (Burr/GPD)^2 - Clayton model and the (Weibull)^2 - Clayton model, we now explain briefly how to simulate from these specific models.

Example 4.1.2 (Continuation of Example 3.1.1).

Let us consider again the case of the (Burr/GPD)^2 - Clayton model, i.e. we have for $x, y > 0$

$$\begin{aligned} \overline{F}_1(x) &= A_1 \left(\mathbf{1}_{(0,u_1]} \left(A_1^{-1} - 1 + (1 + x^{c_1})^{-k_1} \right) + \mathbf{1}_{(u_1,\infty)} \left(1 + \xi_1 \frac{x + h_1 - u_1}{\beta_1} \right)^{-1/\xi_1} \right), \\ \overline{F}_2(y) &= A_2 \left(\mathbf{1}_{(0,u_2]} \left(A_2^{-1} - 1 + (1 + y^{c_2})^{-k_2} \right) + \mathbf{1}_{(u_2,\infty)} \left(1 + \xi_2 \frac{y + h_2 - u_2}{\beta_2} \right)^{-1/\xi_2} \right), \end{aligned}$$

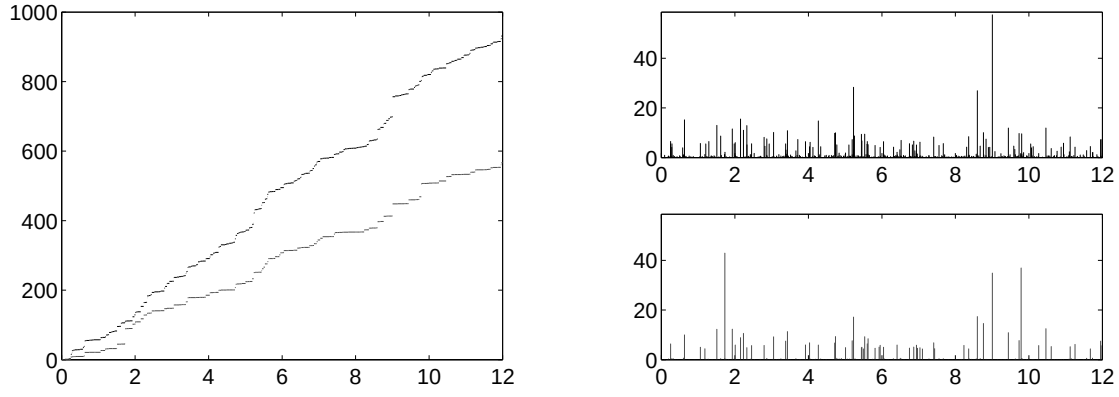
for $A_i = \left(1 - (1 + u_i^{c_i})^{-k_i} + \left(1 + \xi_i \frac{h_i}{\beta_i} \right)^{-1/\xi_i} \right)^{-1}$, $i = 1, 2$. The dependence structure is given by the Clayton Lévy copula $\mathfrak{C}(u, v) = (u^{-\delta} + v^{-\delta})^{-1/\delta}$, $u, v > 0$.

All we need for simulation from the (Burr/GPD)^2 - Clayton model are the functions $\overline{F}_1^{\perp}$, $\overline{F}_2^{\perp}$, $\overline{F}_1^{\parallel}$ and $\overline{H}_x$. Applying Equations (4.1.1), (4.1.2), (4.1.3) and (4.1.4) for the given model yields for $x, y > 0$,

$$\begin{aligned} \overline{F}_1^{\perp}(x) &= \frac{1}{\lambda_1^{\perp}} \Bigg[\lambda_1 A_1 \left(\mathbf{1}_{(0,u_1]} \left(A_1^{-1} - 1 + (1 + x^{c_1})^{-k_1} \right) + \mathbf{1}_{(u_1,\infty)} \left(1 + \xi_1 \frac{x + h_1 - u_1}{\beta_1} \right)^{-\frac{1}{\xi_1}} \right) \\ &\quad - \left((\lambda_1 A_1)^{-\delta} \left(\mathbf{1}_{(0,u_1]} \left(A_1^{-1} - 1 + (1 + x^{c_1})^{-k_1} \right) + \mathbf{1}_{(u_1,\infty)} \left(1 + \xi_1 \frac{x + h_1 - u_1}{\beta_1} \right)^{-\frac{1}{\xi_1}} \right)^{-\delta} + \lambda_2^{-\delta} \right)^{-\frac{1}{\delta}} \Bigg], \end{aligned}$$

$$\begin{aligned}
\overline{F}_2^{\perp}(y) &= \frac{1}{\lambda_2^{\perp}}\Bigg[\lambda_2 A_2\left(\mathbf{1}_{(0,u_2]}\left(A_2^{-1}-1+\left(1+y^{c_2}\right)^{-k_2}\right)+\mathbf{1}_{(u_2,\infty)}\left(1+\xi_2\frac{y+h_2-u_2}{\beta_2}\right)^{-\frac{1}{\xi_2}}\right)\\
&\quad-\left(\lambda_1^{-\delta}+(\lambda_2 A_2)^{-\delta}\left(\mathbf{1}_{(0,u_2]}\left(A_2^{-1}-1+\left(1+y^{c_2}\right)^{-k_2}\right)+\mathbf{1}_{(u_2,\infty)}\left(1+\xi_2\frac{y+h_2-u_2}{\beta_2}\right)^{-\frac{1}{\xi_2}}\right)^{-\delta}\right)^{-\frac{1}{\delta}}\Bigg],\\
\overline{F}_1^{\parallel}(x) &= \frac{1}{\lambda^{\parallel}}\left[(\lambda_1 A_1)^{-\delta}\left(\mathbf{1}_{(0,u_1]}\left(A_1^{-1}-1+\left(1+x^{c_1}\right)^{-k_1}\right)+\mathbf{1}_{(u_1,\infty)}\left(1+\xi_1\frac{x+h_1-u_1}{\beta_1}\right)^{-\frac{1}{\xi_1}}\right)^{-\delta}+\lambda_2^{-\delta}\right]^{-\frac{1}{\delta}},\\
\overline{H}_x(y) &= \Bigg[\frac{(\lambda_1 A_1)^{-\delta}\left(\mathbf{1}_{(0,u_1]}\left(A_1^{-1}-1+\left(1+x^{c_1}\right)^{-k_1}\right)+\mathbf{1}_{(u_1,\infty)}\left(1+\xi_1\frac{x+h_1-u_1}{\beta_1}\right)^{-\frac{1}{\xi_1}}\right)^{-\delta}}{(\lambda_1 A_1)^{-\delta}\left(\mathbf{1}_{(0,u_1]}\left(A_1^{-1}-1+\left(1+x^{c_1}\right)^{-k_1}\right)+\mathbf{1}_{(u_1,\infty)}\left(1+\xi_1\frac{x+h_1-u_1}{\beta_1}\right)^{-\frac{1}{\xi_1}}\right)^{-\delta}+\lambda_2^{-\delta}}\\
&\quad+\frac{(\lambda_2 A_2)^{-\delta}\left(\mathbf{1}_{(0,u_2]}\left(A_2^{-1}-1+\left(1+y^{c_2}\right)^{-k_2}\right)+\mathbf{1}_{(u_2,\infty)}\left(1+\xi_2\frac{y+h_2-u_2}{\beta_2}\right)^{-\frac{1}{\xi_2}}\right)^{-\delta}}{(\lambda_1 A_1)^{-\delta}\left(\mathbf{1}_{(0,u_1]}\left(A_1^{-1}-1+\left(1+x^{c_1}\right)^{-k_1}\right)+\mathbf{1}_{(u_1,\infty)}\left(1+\xi_1\frac{x+h_1-u_1}{\beta_1}\right)^{-\frac{1}{\xi_1}}\right)^{-\delta}+\lambda_2^{-\delta}}\Bigg]^{-\frac{1}{\delta}-1}.
\end{aligned}$$

Simulation of CPPs is then straighforward, employing Algorithm 4.1.1. In Figure 4.1 we see the sample paths and the marked point processes of one simulated CPP of the (Burr/GPD)^2 - Clayton model.



(a) Sample paths

(b) Same paths as marked point processes

Figure 4.1: Simulation of a bivariate CPP in the (Burr/GPD)^2 - Clayton model over a time interval of 12 months.

Example 4.1.3 (Continuation of Example 3.1.2).

For simulation from the the (Weibull)^2 - Clayton model we need for $x, y > 0$,

$$\begin{aligned}
\overline{F}_1^{\perp}(x) &= \frac{1}{\lambda_1^{\perp}}\left[\lambda_1 e^{-\left(\frac{x}{a_1}\right)^{b_1}} - \left(\lambda_1^{-\delta} e^{\delta\left(\frac{x}{a_1}\right)^{b_1}} + \lambda_2^{-\delta}\right)^{-1/\delta}\right], \\
\overline{F}_2^{\perp}(y) &= \frac{1}{\lambda_2^{\perp}}\left[\lambda_2 e^{-\left(\frac{y}{a_2}\right)^{b_2}} - \left(\lambda_1^{-\delta} + \lambda_2^{-\delta} e^{\delta\left(\frac{y}{a_2}\right)^{b_2}}\right)^{-1/\delta}\right], \\
\overline{F}_1^{\parallel}(x) &= \frac{1}{\lambda^{\parallel}}\left[\lambda_1^{-\delta} e^{\delta\left(\frac{x}{a_1}\right)^{b_1}} + \lambda_2^{-\delta}\right]^{-1/\delta}, \\
\overline{H}_x(y) &= \left(\frac{\lambda_1^{-\delta} e^{\delta\left(\frac{x}{a_1}\right)^{b_1}} + \lambda_2^{-\delta} e^{\delta\left(\frac{y}{a_2}\right)^{b_2}}}{\lambda_1^{-\delta} e^{\delta\left(\frac{x}{a_1}\right)^{b_1}} + \lambda_2^{-\delta}}\right)^{-1/\delta-1}.
\end{aligned}$$

4.2 Illustrative example

Let us now illustrate the MCMC algorithm in practice. In this section we present the results for fitting the (Burr/GPD)^2 - Clayton model to the simulated data set shown in Figure 4.1. We also conducted the same study for other Clayton models, in particular the (Weibull)^2 - Clayton model. However, we do not explain them separately since the results are pretty much the same. Nevertheless, we want to emphasize that our MCMC sampler works very well for all Clayton models we tried.

We used the parameter values $\lambda_1 = 34$, $c_1 = 4.1$, $k_1 = 0.42$, $h_1 = 7.3$, $\beta_1 = 3.8$, $\xi_1 = 0.42$, $\lambda_2 = 26$, $c_2 = 1.2$, $k_2 = 1.9$, $h_2 = 16$, $\beta_2 = 8.4$, $\xi_2 = 0.18$ and $\delta = 1.8$ for simulation. These values are motivated by the analysis of the Danish fire insurance data in the following chapter. Furthermore, we set $T = 12$, which is again motivated by Chapter 5, where a time unit corresponds to one month. Observable are the single jump sizes $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ of both components, the jump sizes $\mathbf{x}$ and $\mathbf{y}$ of the joint jumps, the numbers $n_1^{\perp}$ and $n_2^{\perp}$ of single jumps in both components and the number $n^{\parallel}$ of joint jumps. In general, the observed number of jumps in these simulations is random, but it is always around 480.

Before starting the sampler we need to specify a prior distribution. That is the point where, in practice, prior knowledge can be integrated. The hyperparameters[1] would typically be specified by means of expert elicitation. Hence, the prior knowledge influences

[1] In Bayesian statistics hyperparameters are parameters of the prior distribution; the term is used to distinguish them from parameters of the model.

the model indirectly.

We use for all thirteen parameters independent Gamma[2] prior distributions $\Gamma(p_k, b_k)$, $k = 1, \ldots, 13$. Note that we always follow the parametrization where the Gamma density, for $p, b > 0$, is given by

$$f_\Gamma(x; p, b) = \frac{b^p}{\Gamma(p)} x^{p-1} e^{-bx} \mathbf{1}_{[0,\infty)}(x)\,, \qquad x > 0\,.$$

Hence, the mean is given by p/b and the variance by p/b^2. The hyperparameters p_k and b_k, $k = 1, \ldots, 13$, are chosen in such a way that the means correspond to initial ML estimates and the standard deviations are about one fourth of the means. In particular, the means of the independent Gamma priors are calculated as 32, 4.2, 0.41, 4.7, 3.2, 0.72, 26, 1.3, 2.1, 16, 12, 0.17 and 2.3, respectively, for the parameters λ_1, c_1, k_1, h_1, β_1, ξ_1, λ_2, c_2, k_2, h_2, β_2, ξ_2 and δ. The standard deviations are chosen to be 8.0, 1.0, 0.10, 1.2, 0.80, 0.18, 6.5, 0.33, 0.53, 4.1, 3.0, 0.043 and 0.57, respectively.

Now we can apply the Bayesian method using MCMC to recover the marginal posterior distributions of all parameters used for simulation. As stated above the maximum likelihood estimates are a convenient choice for starting values, because they help to shorten the burn-in period. Figure 4.2 shows for each parameter of the (Burr/GPD)^2 - Clayton model the marginal posterior distribution together with the produced sample path.

The marginal posterior distributions are all unimodal with peaks which are very close to the MLEs and the distributions are all quite symmetric. Considering the sample paths we see that the mixing behaviour of our MCMC sampler is very satisfying. We notice that there are small differences in the evolution of the paths for the different parameters. Particularly for β_i, ξ_i, $i = 1, 2$, which determine the tails of the distributions, the chain does not mix as fast as it does for the other parameters; in consequence the autocorrelations are higher. Hence, we decided to use subsampling and to take only every 10th iteration in order to reduce the autocorrelations. The acceptance rates for the thirteen parameters of interest are all between 55% and 70%.

Repeating the analysis with different initial values has shown that the produced Markov chains converge very fast, usually within 500 iterations. Thus, a burn-in period of 1000 iterations is sufficient, and we use iterations $1010, 1020, 1030, \ldots$ to derive nearly independent samples from the posterior distribution.

[2]We want to stress that our explicit choice of prior distribution as well as the calibration of it, is exemplary. Any appropriate distribution could be used instead.

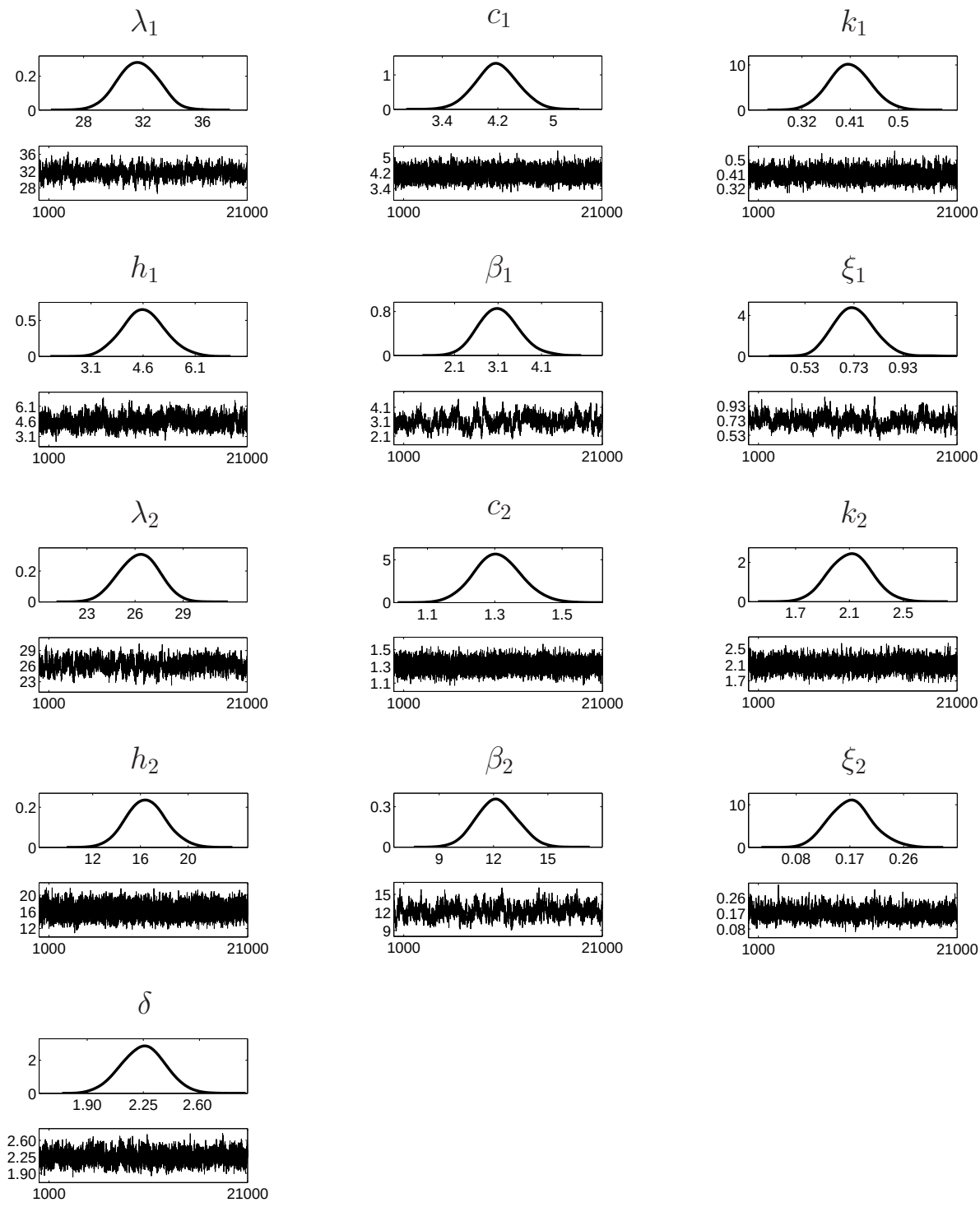


Figure 4.2: Density estimates of marginal posterior distributions and sample paths for the parameters of the (Burr/GPD)^2 - Clayton model. Simulated data and independent Gamma priors are used.

4.3 Quality of posterior mean estimates

To assess the quality of the posterior mean estimates (PME) we repeat the analysis from Section 4.2 100 times for uniform priors. More precisely, we simulate 100 data sets, using always the same fixed parameter set as before. We fit the (Burr/GPD)^2 - Clayton model to these 100 data sets by MCMC. Then we calculate the posterior mean estimates of the parameters for all 100 data sets.

	λ_1	c_1	k_1	h_1	β_1	ξ_1	
True value	34	4.1	0.42	7.3	3.8	0.42	
Mean	34.2	4.08	0.409	7.38	3.93	0.438	
Std	1.72	0.320	0.0313	1.84	1.08	0.143	
	λ_2	c_2	k_2	h_2	β_2	ξ_2	δ
True value	26	1.2	1.9	16	8.4	0.18	1.8
Mean	26.1	1.26	1.97	16.5	8.68	0.173	1.88
Std	1.79	0.0726	0.150	3.18	1.99	0.0475	0.211

Table 4.1: Means and standard deviations of the posterior mean estimates in the (Burr/GPD)^2 - Clayton model for 100 simulations. The simulation parameters are given in the first row of the table.

Table 4.1 assesses the means and the standard deviations of the posterior mean estimates for the 100 data sets coming from the (Burr/GPD)^2 - Clayton model, for all thirteen parameters. It shows that on average the posterior mean estimates match the simulation values very well and that the standard deviations are reasonably small. The ones for the GPD parameters (h_i, β_i, ξ_i, $i = 1, 2$) could be further reduced when taking more observations into consideration. We can conclude that the performance of our sampler is very satisfying.

We conduct the same procedure for the (Weibull)^2 - Clayton model. For simulation from this model we refer to Example 4.1.3. The corresponding results, which are given in Table 4.2, illustrate that also in case of this specific Clayton model the MCMC sampler works very well.

4.4 Comparison of posterior mean and maximum likelihood estimates

To conclude the simulation study, we finally want to examine if there are significant differences between the posterior mean estimates (PME) and the maximum likelihood estimates (MLE). Because the results for the two previously considered Clayton models are again very similar, we only show the results for the (Weibull)^2 - Clayton model.

In Section 4.3 we calculated for 100 different simulations the PMEs for the parameters λ_1, λ_2, a_1, b_1, a_2, b_2 and δ. We therefore used uniform priors to make the results 'independent' of the choice of the prior distribution (in particular of its mean). For the given simulated data sets we now compute the MLEs which is a seven-dimensional optimization problem, each.

Table 4.2 compares the means and the standard deviations of the MLEs and the PMEs. We see that the means of both estimates are very close to each other and also to the true simulation values. Neither the MLE nor the PME is systematically better than the other (compare e.g. the means of the parameters λ_1 and λ_2). Also the differences in the standard deviations are hardly worth mentioning.

	λ_1	λ_2	a_1	b_1	a_2	b_2	δ
True value	12	5.5	0.83	1.1	1.3	1.1	0.86
Mean of MLE	12.0	5.41	0.829	1.10	1.34	1.13	0.891
Mean of PME	12.1	5.46	0.835	1.10	1.35	1.13	0.900
Std of MLE	0.681	0.514	0.0530	0.0521	0.113	0.0784	0.131
Std of PME	0.683	0.517	0.0534	0.0570	0.113	0.0782	0.132

Table 4.2: Means and standard deviations of the maximum likelihood and the posterior mean estimates in the (Weibull)^2 - Clayton model. The simulation parameters are given in the first row of the table.

Figure 4.3 shows the residuals of the MLEs and the PMEs for the first 50 simulated data sets. Here we see that for the parameters λ_1, λ_2, a_1, a_2 and δ the MLE for every single simulation is greater than the corresponding PME (for b_1 and b_2 the estimates are almost the same). Although this behaviour is interesting, the difference is too small to judge which of the estimates is better in general.

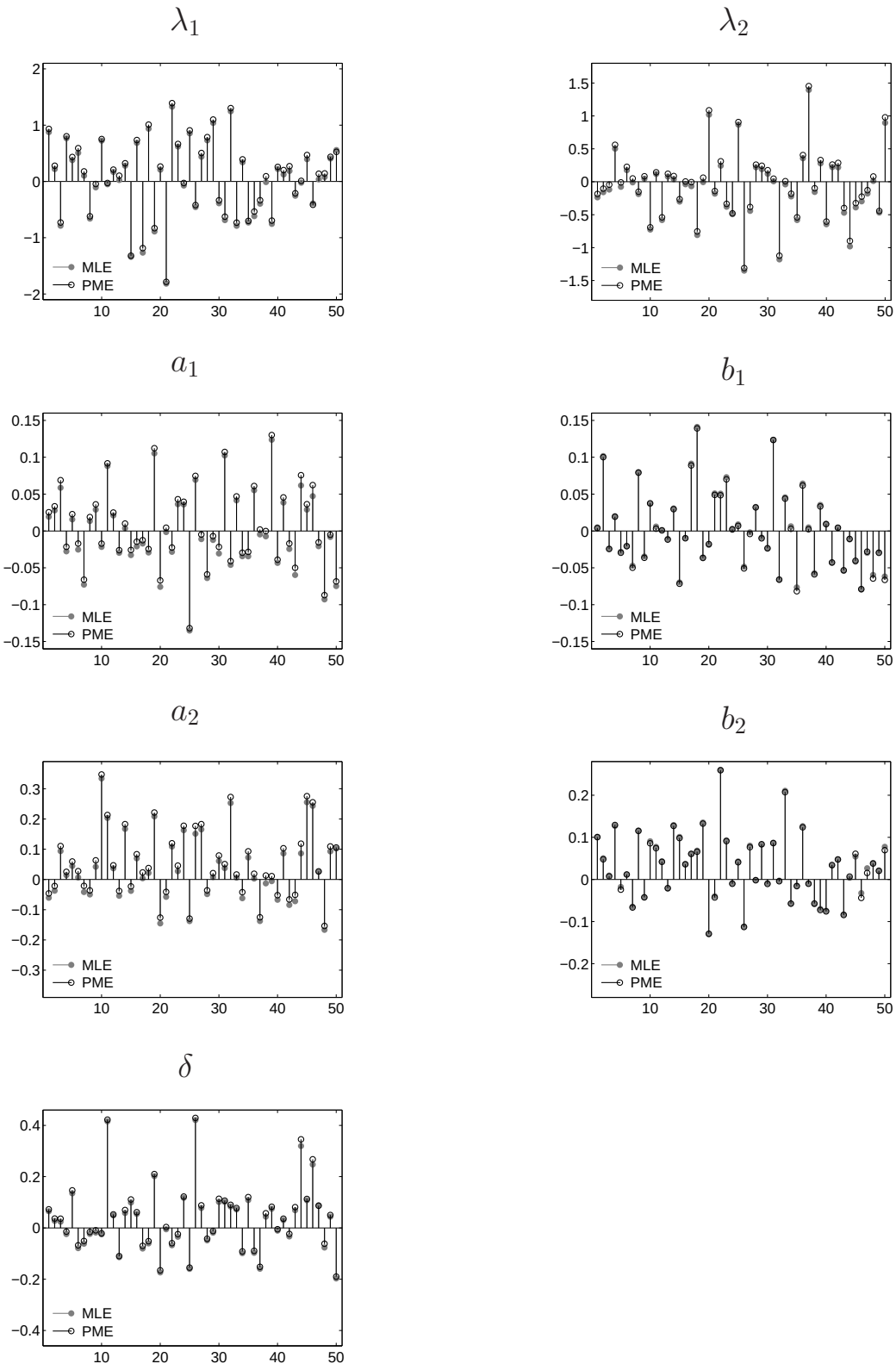


Figure 4.3: Comparison of residuals for maximum likelihood and posterior mean estimates for the parameters λ_1, λ_2, a_1, b_1, a_2, b_2 and δ in the (Weibull)^2 - Clayton model.

Chapter 5

Analysis of Danish fire insurance data

In Chapter 4 we have shown that the developed MCMC sampler works very well. Hence, we are now ready to apply it to an empirical data set which, in our case comes from Danish fire insurance. By employing our MCMC sampler developed in Chapter 3 we want to find a Clayton model which fits the data best and discuss some results implied by this specific model. We do this for two different approaches to the data. First we consider the data in its original form and second we transform it with the logarithm.

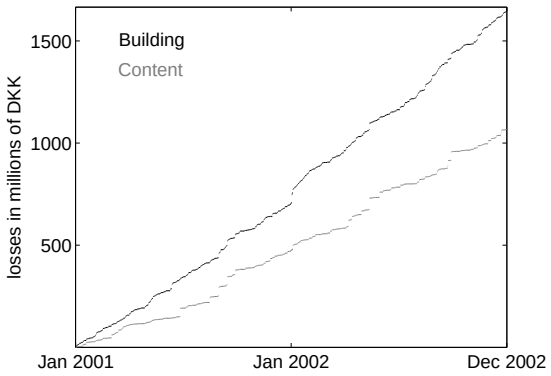
5.1 An exploratory analysis of the data set

The data were collected at Copenhagen Reinsurance, an aggregated form of them appears for example in Embrechts et al. (1997). Our analysis comprises in total 847 observations which were collected over the two-year period January 1, 2001 to December 31, 2002. The fire losses are reported in millions of Danish Kroner (DKK)[1]. Every total claim has been divided into loss of building, loss of content and loss of profit. However, we restrict the data set to the first two categories of claims, because the losses of profit have rarely non-zero values. Figure 5.1 shows the observed sample paths of the accumulated losses as well as their representation as marked point processes for the whole data set and for the year 2002.

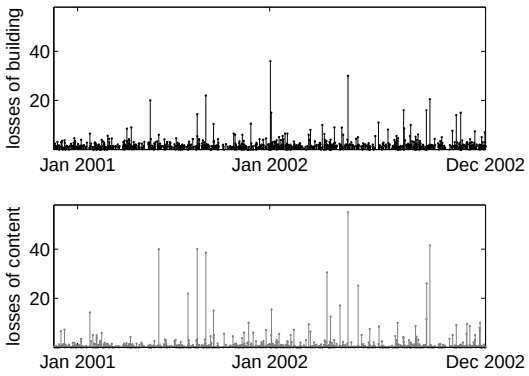
As stated above the data is coming from a reinsurance company. This company is only interested in claims that are in total above one million DKK, because the reinsurer only has to pay in these cases. That is why only losses that are in the sum greater than this threshold are reported. The consequence is that the data is incomplete, insofar as

[1]Given the exchange rate from May 25, 2010, 1 DKK is equal to 0.134409 EUR

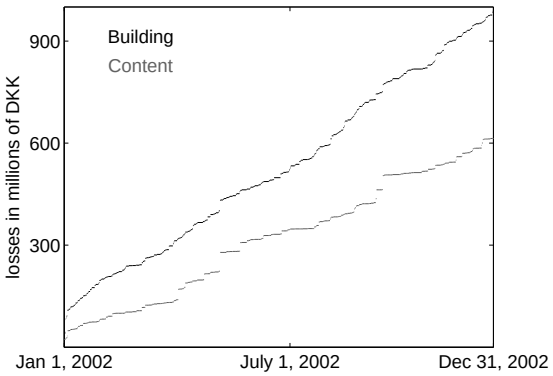
the present small losses are not representative; there are 'small' single losses in the data set, but only if the sum of losses is above one million DKK. Also other small losses might have occured in practice, but they are not reported since the total loss is smaller than one million DKK.



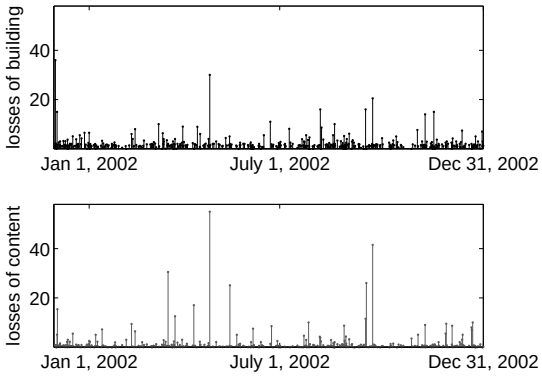
(a) Sample paths, two years



(b) Marked point processes, two years



(c) Sample paths, one year



(d) Marked point processes, one year

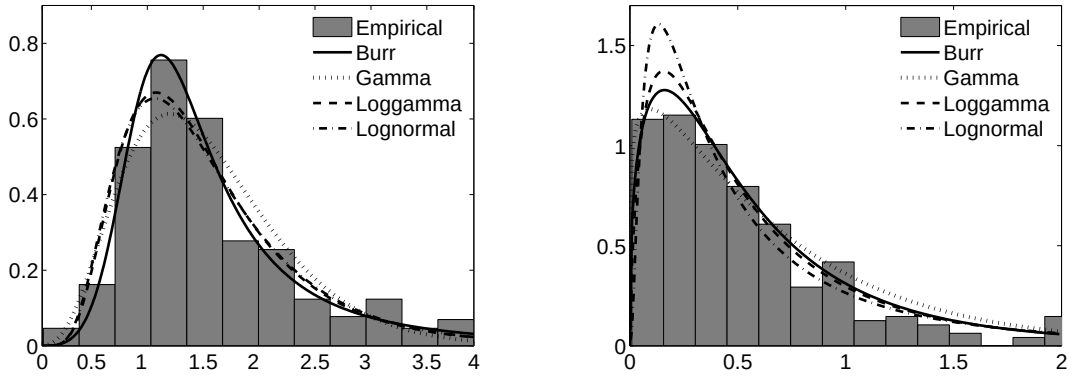
Figure 5.1: The Danish fire insurance data: observed sample paths of the accumulated losses for buildings and contents, and their representation as marked point processes.

When trying to fit a good model to the data we have to be aware of this incompleteness. Depending on the purpose we thus make different approaches to the Danish fire insurance data. We consider in particular the following approaches. In the first we analyze the data in its original form. The motivation for it is to cover also the 'small' losses, i.e. losses that are between one and two million DKK. These losses are important for the reinsurance company, because there are many of them. In the second approach we remove all losses which are smaller than one million DKK. Hence, we can guarantee that the remaining claims are all coming from the same distribution which allows appropriate modelling of large losses. In this approach we will also transform the data with the logarithm (as

Esmaeili and Klüppelberg (2010a) did).

Let us have a look at the histograms of small losses of the year 2002, given in Figure 5.2. Obviously the marginal claims in both components are naturally bounded below by zero. There are much more very small claims (below one million DKK) of content which goes together with the intuitive idea of fire losses: small damages happen often inside the house and when the building itself is affected it is, generally speaking, more expensive.

Note, however, that there are several big losses in both components which are not given in the histograms. Hence, when trying to find a Clayton model which describes the data very well, right-sided heavy-tailed distributions seem to be a promising choice and should be considered above all. Moreover, the structure of the data suggests to include sliced distributions into the analysis.



(a) Building (b) Content

Figure 5.2: Histograms of small losses together with different body distributions for the Danish fire insurance data, year 2002.

To get an idea which distributions are appropriate for modelling the body of the data, we compare some fits in Figure 5.2. We used in particular the following distributions:

$$\begin{aligned}
&\text{Burr:} && f(x) = ck\,x^{c-1}\,(1+x^c)^{-(k+1)}\,, && x>0\,,\ c,k>0.\\
&\text{Gamma:} && f(x) = \tfrac{x^{\alpha-1}}{\beta^\alpha\Gamma(\alpha)}\exp\left(-\tfrac{x}{\beta}\right), && x>0\,,\ \alpha,\beta>0.\\
&\text{Loggamma:} && f(x) = \tfrac{b^a}{\Gamma(a)}(x+1)^{-(b+1)}\,(\log(x+1))^{a-1}\,, && x>0\,,\ a,b>0.\\
&\text{Lognormal:} && f(x) = \tfrac{1}{\sqrt{2\pi}\sigma x}\exp\left(\tfrac{-(\log x-\mu)^2}{2\sigma^2}\right), && x>0\,,\ \mu\in\mathbb{R},\,\sigma>0.
\end{aligned}$$

As we can see, the fit of the Burr distribution as well as the fit of the Loggamma distribution is quite good. Hence, we particularly consider these two distributions in the following.

	single jumps	common jumps	minimal loss	maximal loss	average loss
			period January to December 2002		
building	131	308	70,000	36,000,000	2,076,335
content	35	308	9,000	55,000,000	1,286,908
			period July to December 2002		
building	71	147	70,000	20,500,000	2,257,215
content	18	147	11,000	41,502,000	1,708,442

Table 5.1: Summary of the Danish fire insurance data: number of single and common jumps and minimal, maximal and average losses in DKK for insured buildings and contents.

Let us now have a look at the summary given in Table 5.1. It is obvious, that especially the losses of content are scattered very widely, although the average loss is relative small. Furthermore, we see that about two out of three losses affect the building as well as the contents, i.e. happen together. One reason therefore is probably the fact that only claims above one million DKK are reported. Such high losses are caused particularly when there are damages inside and outside the building. This also explains why there are quite few losses which only occur in the content component: losses above one million DKK which do not affect the building itself are happening rarely.

For completeness we also give the summary of the period July to December 2002. In Section 5.2.4 we compare the estimates for these two time periods.

5.2 Analysis of the original data

In this section we analyze the Danish fire insurance data in its original form. That is to say, we do neither remove any claims nor do we transform the data. We here consider 474 losses which were reported over the year 2002. In the second row of Figure 5.1 we see the sample paths and the time series of the observations.

We begin our analysis by fitting several Clayton models to the data. Using the Bayes factors introduced in Section 2.3 we decide which of them fits the data best. For the selected model we finally present in detail the results coming from our MCMC sampler: we examine the role of the prior and check the parameter estimates for robustness.

5.2.1 Model selection via Bayes factors

We now want to find the Clayton model which describes the Danish fire insurance data best possible. Therefore we fit different models to the data and compare the results using the Bayes factors. Let us start with some remarks on the model selection procedure.

At first we consider only Clayton models in which the marginal jump sizes in both components are described by the same distribution family. Motivated by the explorative data analysis we compare in particular right-sided heavy-tailed distributions, including Burr, Exponential, (shifted) Loggamma, Lognormal, (truncated) Normal, Pareto and Weibull distributions. Assessing the competing models with the Bayes factors yields the results given in Table 5.2.

M_1	M_2					
	Exp	Logg	Logn	trNorm	Par	Wei
Burr	10^{72}	10^{10}	10^{12}	10^{49}	10^{250}	10^{37}
Exponential		10^{-62}	10^{-60}	10^{-23}	10^{178}	10^{-35}
(shifted) Loggamma			10^{2}	10^{39}	10^{240}	10^{27}
Lognormal				10^{37}	10^{238}	10^{25}
(truncated) Normal					10^{201}	10^{-12}
Pareto						10^{-213}

Table 5.2: Bayes factors of model M_1 vs. model M_2 for Clayton models.

The high values of these factors tell us that there are huge differences in the goodness of fit. For example, the (Burr)^2 - Clayton model has a Bayes factor of 10^{10} against the second best (Loggamma)^2 - Clayton model. Hence, according to Jeffreys' Bayes factor scale, see Table 2.1 in Section 2.3, we have an absolutely decisive evidence in favor of the (Burr)^2 - Clayton model versus the other. However, with the conclusions from our explorative data analysis in mind, there is the chance to improve our fit further if we allow combined models or sliced models. That is why we now consider additionaly combined Clayton models, where the marginal jump size distributions F_1, F_2 may be different from each other, and sliced Clayton models, where the marginal jump size distributions F_1, F_2 may be sliced distributions, each.

After having fitted several of these Clayton models – we hereby focused on combinations of the distributions from above – to the data set, it finally turned out that the

(Burr/GPD)^2 - Clayton model is the best choice (following again a Bayes factor analysis). In Table 5.3 we compare the best five models using the approximated Bayes factors. We see here that the choice of the (Burr/GPD)^2 - Clayton model is clearly justified by very decisive Bayes factors.

(Burr/GPD)^2 vs. (Loggamma/GPD)^2	2.1×10^3
(Burr/GPD)^2 vs. (Lognormal/GPD)^2	7.3×10^4
(Burr/GPD)^2 vs. (Burr)^2	1.4×10^8
(Burr/GPD)^2 vs. Burr - Loggamma	1.8×10^8

Table 5.3: Approximated Bayes factors for the (Burr/GPD)^2 -Clayton model, results for Γ_1 priors. 10,000 simulations are used for calculation.

5.2.2 Model selection via weighted Bayes factors

In some applications – for example in operational risk where modelling of large losses is crucial – one may like to increase the impact of the tails on the model selection procedure. Therefore, it can be advantageous to use a weighted Bayes factor instead of the common Bayes factor. Let us point out how this idea can be realized practically.

So far we made use of the Bayes factor approximation

$$\tilde{B}_{ij}(\mathbf{z}) = \frac{n_i^{-1} \sum_{k=1}^{n_i} f_i(\mathbf{z}|\boldsymbol{\psi}^{i,k})}{n_j^{-1} \sum_{k=1}^{n_j} f_j(\mathbf{z}|\boldsymbol{\psi}^{j,k})}, \qquad i,j \in \{1,\dots,L\}\,,\, i \neq j\,,$$

introduced in Section 2.3. We now heuristically define the weighted approximated Bayes factor as

$$\tilde{W}_{ij}^{\alpha}(\mathbf{z}) = \frac{n_i^{-1} \sum_{k=1}^{n_i} \left[\alpha\, f_i(\mathbf{z}_{[0\%-90\%)}|\boldsymbol{\psi}^{i,k}) + (1-\alpha)\, f_i(\mathbf{z}_{[90\%-100\%]}|\boldsymbol{\psi}^{i,k})\right]}{n_j^{-1} \sum_{k=1}^{n_j} \left[\alpha\, f_j(\mathbf{z}_{[0\%-90\%)}|\boldsymbol{\psi}^{j,k}) + (1-\alpha)\, f_j(\mathbf{z}_{[90\%-100\%]}|\boldsymbol{\psi}^{j,k})\right]}\,,$$

for $i,j \in \{1,\dots,L\}$, $i \neq j$, $0 \leq \alpha \leq 1$. Here $\mathbf{z}_{[0\%-90\%)}$ denotes the observations which are below the empirical 90% quantiles, and $\mathbf{z}_{[90\%-100\%]}$ the ones which are above these values.

Obviously, we get for the marginal value $\alpha = 0$ the classical Bayes factor for the tails only, whereas for $\alpha = 1$ we obtain the classical Bayes factor for the body. For $0 < \alpha < 1$ we get a combination of these two extremes.

In applications where the fit of the tails is of particular interest, one should therefore also consider the weighted factors. We illustrate this for the exemplary chosen value

$\alpha = 60\%$. That is to say, we assign 60% of the total weight to the lower 90% of the data and consequently 40% of it to the upper 10%. The results are given in Table 5.4. We see that using the weighted factors instead of the common Bayes factors does not have an impact on the final model choice in our case; the (Burr/GPD)^2 - Clayton model is still the best choice.

(Burr/GPD)^2 vs. (Loggamma/GPD)^2	1.3×10^3
(Burr/GPD)^2 vs. (Lognormal/GPD)^2	4.3×10^4
(Burr/GPD)^2 vs. (Burr)^2	5.7×10^8
(Burr/GPD)^2 vs. Burr - Loggamma	6.5×10^8

Table 5.4: Approximated weighted Bayes factors for the (Burr/GPD)^2 -Clayton model, results for Γ_1 priors. 10,000 simulations are used for calculation.

5.2.3 The impact of the prior distribution

We now examine how the choice of prior distribution affects the approximated posteriors. Since we have seen in the previous section that the (Burr/GPD)^2 - Clayton model actually allows the best fit to the data, we present the results for this specific model. Note that we had to fix the thresholds u_1, u_2 of this sliced model (cf. Example 3.1.1) before fitting the data to it. Exemplary, we set these values equal to the empirical 90% quantiles of the components, i.e. $u_1 = 3.96$, $u_2 = 3.90$. For a comprehensive theory how to model these thresholds and extremal events in general we refer to Embrechts et al. (1997), Section 6.5.

In Figure 5.3 we show approximated marginal posterior densities for the parameters $\lambda_1, \boldsymbol{\theta}_1, \lambda_2, \boldsymbol{\theta}_2, \delta$ of the (Burr/GPD)^2 - Clayton model for the original Danish fire insurance data of the year 2002. Furthermore, the last plot in this figure describes the relative frequency of joint jumps which is directly calculated from the drawn samples of the parameters. We see that on average the probability for a joint jump is about 50%.

Three different prior distributions for the parameters are used: uniform priors, independent Gamma priors with large standard deviations (denoted by Γ_1) and independent Gamma priors with smaller standard deviations (denoted by Γ_2). In the first case the standard deviations of the priors are taken as the corresponding mean divided by 4, in the second case as the corresponding mean divided by 20, hence smaller by factor 5.

We clearly see the impact of the prior distributions on the uncertainty in the parameters. Whereas for the uniform prior and the Γ_1 priors the results are quite similar, the

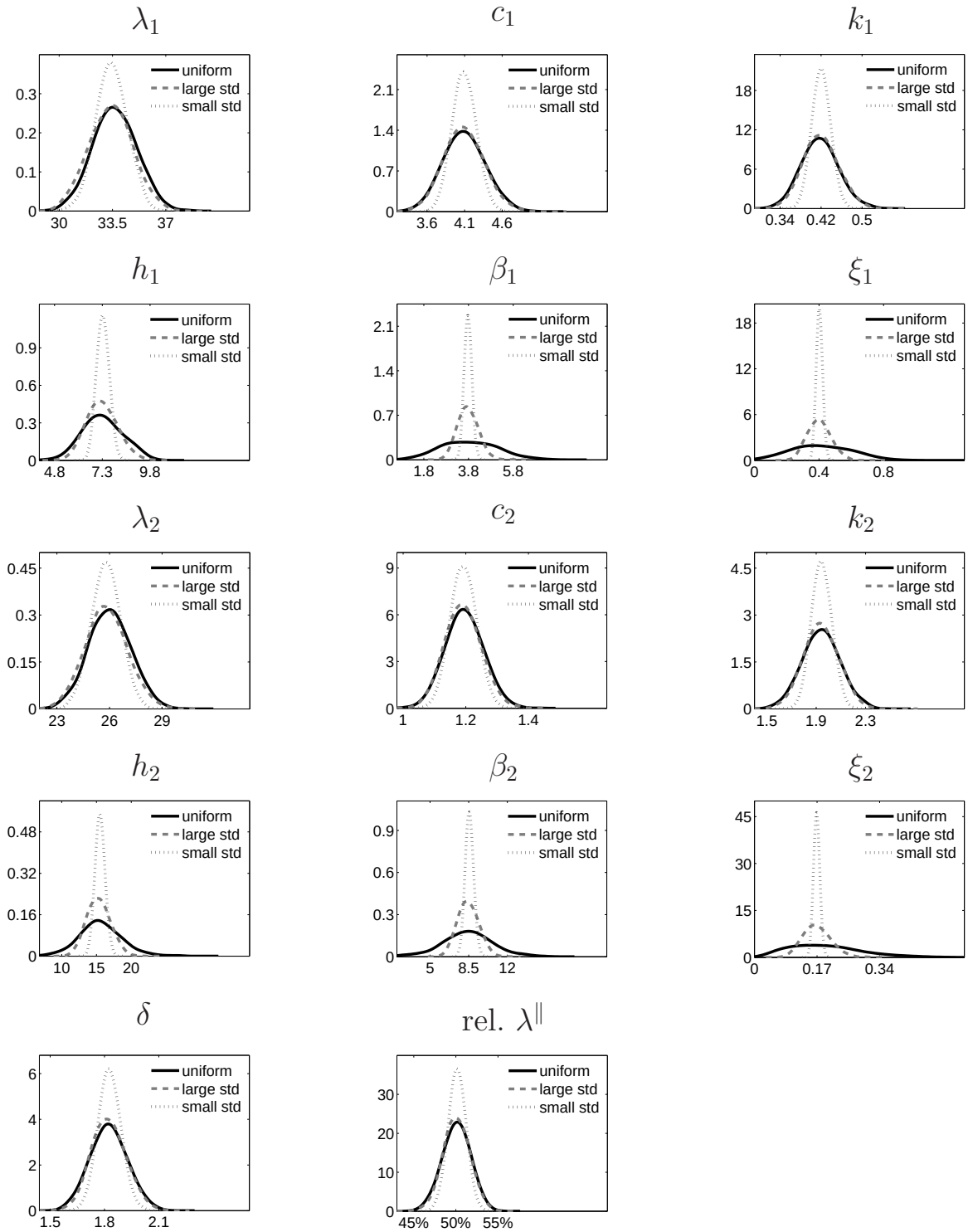


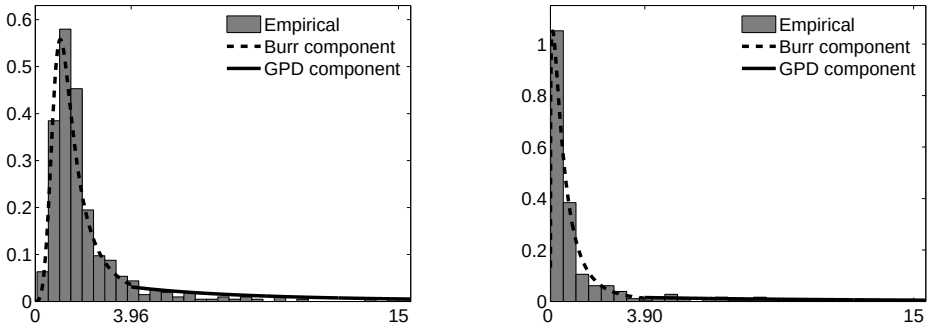
Figure 5.3: Marginal posterior distributions for the parameters of the (Burr/GPD)^2 - Clayton model together with the relative frequency of joint jumps. Three different prior distributions are used.

marginal densities are much more concentrated for the Γ_2 priors. Particularly the estimates for the parameters of the GPD distribution used to model the tail of the sliced distributions show significant uncertainties for the uniform and the Γ_1 prior. Remembering that we have only about 40 observations available to model both tails, this behaviour is not further remarkable and could have been expected.

As we can see in Table 5.5 the posterior mean estimate is, for each parameter, very close to the MLE. Moreover, the table reports the 95% credible intervals of the posterior means for all parameters and for the relative joint jump frequency, given the three different priors. Summarizing the plots and the credible intervals it becomes apparent that the choice of the standard deviations in the prior distributions has a major impact on the simulated values and the estimated posterior distributions. When choosing large values there is no significant difference compared to an uniform prior. Maybe in practice the experience of the user of the MCMC sampler allows to choose small standard deviations in the priors and, hence, to make them quite informative. From Figure 5.3, however, one can see that this comes along with the risk of underestimating the uncertainty about the parameter estimates.

Finally, to get a visual idea of the quality of the fit, we plot the density of the adapted (Burr/GPD)^2 - Clayton model together with the empirical histograms, see Figure 5.4. We do this only for the fit based on the uniform prior distributions, since the differences are minor. Using the posterior mean estimates as marginal jump size parameters in either case yields the explicit densities for $x, y > 0$:

$$
\begin{aligned}
f_1(x) &= \mathbf{1}_{(0,3.96]}\, 1.49\, x^{3.10} \left(1 + x^{4.10}\right)^{-1.42} + \mathbf{1}_{(3.96,\infty)}\, 0.226 \left(1 + 0.108 \quad (x + 3.36)\right)^{-3.40}, \\
f_2(y) &= \mathbf{1}_{(0,3.90]}\, 1.98\, y^{0.196} \left(1 + y^{1.20}\right)^{-2.94} + \mathbf{1}_{(3.90,\infty)}\, 0.101 \left(1 + 0.0217\, (y + 11.7)\right)^{-6.47}.
\end{aligned}
$$



(a) Building, small jumps (b) Content, small jumps

Figure 5.4: Fitted marginal jump size densities in the (Burr/GPD)^2 - Clayton model.

	λ_1	c_1	k_1
(ML estimate)	33.49	4.094	0.4202
uniform prior	33.66 [30.90, 36.47]	4.096 [3.578, 4.658]	0.4178 [0.3507, 0.4926]
Γ_1 prior	33.37 [30.52, 36.25]	4.095 [3.568, 4.637]	0.4202 [0.3541, 0.4895]
Γ_2 prior	33.44 [31.44, 35.39]	4.091 [3.791, 4.396]	0.4200 [0.3865, 0.4558]
	h_1	β_1	ξ_1
(ML estimate)	7.325	3.776	0.4027
uniform prior	7.319 [5.506, 9.235]	3.846 [1.597, 6.356]	0.4170 [0.0885, 0.7748]
Γ_1 prior	7.291 [5.830, 8.973]	3.776 [2.928, 4.744]	0.4007 [0.2718, 0.5489]
Γ_2 prior	7.338 [6.707, 7.984]	3.778 [3.457, 4.116]	0.4023 [0.3646, 0.4412]
	λ_2	c_2	k_2
(ML estimate)	25.85	1.193	1.946
uniform prior	26.02 [23.69, 28.39]	1.196 [1.081, 1.318]	1.940 [1.647, 2.251]
Γ_1 prior	25.82 [23.49, 28.16]	1.191 [1.084, 1.306]	1.941 [1.680, 2.209]
Γ_2 prior	25.80 [24.14, 27.39]	1.192 [1.113, 1.273]	1.943 [1.790, 2.102]
	h_2	β_2	ξ_2
(ML estimate)	15.51	8.566	0.1690
uniform prior	15.60 [9.67, 22.23]	8.424 [3.629, 13.32]	0.1830 [0.0499, 0.4152]
Γ_1 prior	15.29 [12.16, 18.73]	8.460 [6.616, 10.33]	0.1700 [0.1049, 0.2504]
Γ_2 prior	15.52 [14.22, 16.93]	8.578 [7.840, 9.32]	0.1689 [0.1528, 0.1863]
	δ	rel. $\lambda^{\|\|}$	
(ML estimate)	1.829	50.2%	
uniform prior	1.824 [1.630, 2.025]	50.1% [46.8%, 53.2%]	
Γ_1 prior	1.822 [1.650, 2.016]	50.1% [47.0%, 53.1%]	
Γ_2 prior	1.826 [1.704, 1.950]	50.2% [48.1%, 52.2%]	

Table 5.5: Posterior mean estimates together with 95% credible intervals, maximum likelihood estimates for comparison. The 95% credible intervals are based on the empirical 2.5% and 97.5% quantiles. Results for three different prior distributions.

5.2.4 Robustness of the parameter estimates

Let us now examine how the estimates change when we base our analysis on less observations. Therefore we now apply our MCMC sampler to a smaller data set, namely the data points of the Danish fire insurance data which were reported during the six-month period July 1, 2002, to December 31, 2002 (T=6), cf. Table 5.1. We then compare the estimates to the ones obtained from the previous analysis of the 12-month period (T=12). Table 5.6 contains the results of the fit of the (Burr/GPD)^2 - Clayton model to both data sets.

	λ_1	c_1	k_1
$T = 12$	33.37 [30.52, 36.25]	4.095 [3.568, 4.637]	0.4202 [0.3541, 0.4895]
$T = 6$	33.20 [29.55, 37.07]	4.113 [3.441, 4.799]	0.4092 [0.3258, 0.4957]
	h_1	β_1	ξ_1
$T = 12$	7.291 [5.830, 8.973]	3.776 [2.928, 4.744]	0.4007 [0.2718, 0.5489]
$T = 6$	7.210 [5.424, 9.277]	3.839 [2.838, 4.961]	0.3900 [0.2447, 0.5597]
	λ_2	c_2	k_2
$T = 12$	25.82 [23.49, 28.16]	1.191 [1.084 , 1.306]	1.941 [1.680, 2.209]
$T = 6$	25.17 [22.00, 28.47]	1.132 [0.9830, 1.288]	1.881 [1.504, 2.301]
	h_2	β_2	ξ_2
$T = 12$	15.29 [12.16, 18.73]	8.460 [6.616, 10.33]	0.1700 [0.1049 , 0.2504]
$T = 6$	14.40 [10.56, 18.56]	7.645 [5.689, 9.789]	0.1630 [0.09763, 0.2462]
	δ	rel. $\lambda^{\parallel}$	
$T = 12$	1.822 [1.650, 2.016]	50.1% [47.0%, 53.1%]	
$T = 6$	1.778 [1.529, 2.044]	48.1% [44.7%, 53.3%]	

Table 5.6: Comparison of posterior mean estimates (for Γ_1 prior) for the periods January to December 2002 to the estimates for the period July to December 2002. The estimates are given with 95% credible intervals which are based on the empirical 2.5% and 97.5% quantiles.

As to expect, the 95% credible intervals are bigger for the shorter time period, since our sampler works with less data points. However, from the credible intervals we can conclude that the changes are not significant.

We finally want to mention that we also checked our model fits for the impact of

outliers. We therefore removed the three largest claims of both components – the losses of building and the losses of content. We then applied our MCMC sampler to the remaining data set for different Clayton models. Using the Bayes factors it turned out that the order of the models did not change; the Bayes factors themselves only changed slightly. For the best Clayton model, i.e. $(\text{Burr}/\text{GPD})^2$ - Clayton, the posterior mean estimates and the credible intervals did not show significant differences compared to the full data set. Hence, we can conclude that the existence of outliers does not have decisive impact on the estimates.

5.3 Analysis of the log-transformed data

In this section we present our second approach to the Danish fire insurance data. First, we describe how we adapt and transform the data being considered. Afterwards we analyze this data set with the help of MCMC. Here we focus mainly on the results and do not discuss in detail the single steps of the analyzing procedure, since they are very similar to the ones of the first approach.

5.3.1 The transformed data

As explained in Section 5.1 the Danish fire insurance data is incomplete, because there are those small claims missing which are in the sum smaller than one million DKK. That is why we now base our analysis on the losses being larger than this threshold in both components. The remaining set is complete as far as only the total losses above two million DKK are considered. Hence we are able to model 'large' losses, that is to say losses which are in the sum above two million DKK, in an appropriate way. The drawback of this approach is obvious: total claims below two million DKK which are also of interest for the reinsurance company are not modelled at all.

	single losses of building	single losses of content	joint losses
whole data	215	70	562
restructured data	201	51	85
remaining losses	93.5%	72.9%	15.1%

Table 5.7: Danish fire insurance data: comparison of the number of losses before and after removing of 'small' claims.

Moreover, we transform the data with the logarithm before fitting the models to it. By doing so outlying claims can be modelled more exactly and the fit gets better.

We here consider the losses of the period January 1, 2001, to December 31, 2002, as given in the first row of Figure 5.1. After removing the 'small' claims we keep 336 observation points out of the original 847, which is about 39.8%. In Table 5.7 we see the number of single and joint losses before and after adapting the data set. As illustrated the removing of the claims changes the relation of single and joint jumps significantly. There are only 85 joint losses out of the initial 562 left which is a share of 15.1%. In comparison we keep 93.5% and 72.9% of single losses, respectively. Hence, due to the restructuring of the data we change the dependence structure of the claims. We state these results without intending to judge the different approaches.

5.3.2 Results

After having transformed the losses with the logarithm the resulting data set is not that much heavy-tailed as it is in its original form, see the empirical histograms in Figure 5.5. That also explains why the model selection procedure (which is again based on the Bayes factors) yields a Clayton model which is less heavy-tailed than the one for the untransformed data.

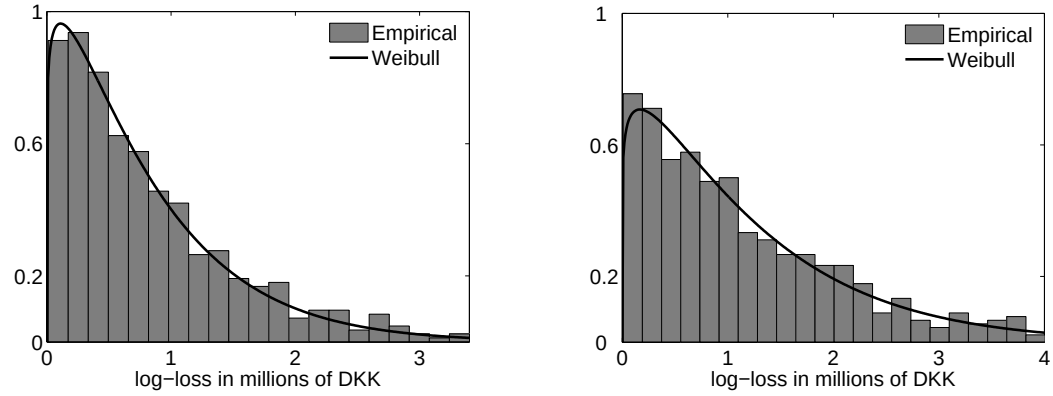
$(\text{Weibull})^2$ vs. $(\text{Gamma})^2$	1.13
$(\text{Weibull})^2$ vs. Gamma - Weibull	6.81
$(\text{Weibull})^2$ vs. $(\text{Exponential})^2$	69.8

Table 5.8: Approximated Bayes factors for the $(\text{Weibull})^2$ - Clayton model. 10000 simulations are used for simulation.

Table 5.8 compares the four best fitting models. According to Jeffreys' Bayes factor scale the evidence of the $(\text{Weibull})^2$ - Clayton model over the $(\text{Gamma})^2$ - Clayton model is barely worth mentioning. However, we decide to use the $(\text{Weibull})^2$ - Clayton model for the upcoming analysis. When using the posterior mean estimates as jump size parameters the marginal jump size densities in the $(\text{Weibull})^2$ - Clayton model are explicitly given by:

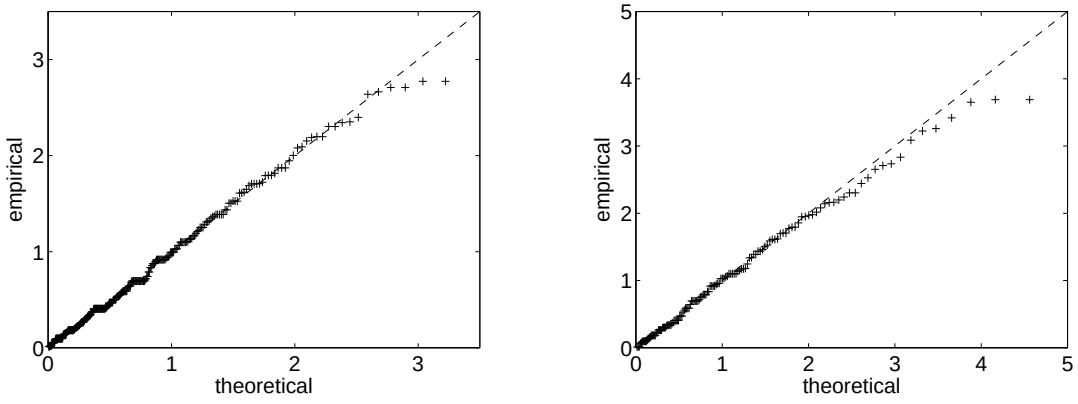
$$f_1(x) = 1.379 \ (\log x)^{0.115} \exp\left(-1.237 \ (\log x)^{1.115}\right), \qquad x > 1,$$
$$f_2(y) = 0.8696\,(\log y)^{0.110} \exp\left(-0.7834\,(\log y)^{1.110}\right), \qquad y > 1.$$

In Figure 5.5 we see the fit of the marginal jump size distributions to the transformed data and Figure 5.6 depicts the corresponding QQ-Plots.



(a) Building (b) Content

Figure 5.5: Histograms and fitted marginal jump size densities in the (Weibull)^2 - Clayton model.



(a) Building (b) Content

Figure 5.6: QQ-Plots for the marginal jump size distributions in the (Weibull)^2 - Clayton model.

The results for the fit by MCMC are given in Figure 5.7. Note that we used the prior Γ_1 which has relatively large standard deviation, cf. Section 5.2.3 for details. The corresponding credible intervals are illustrated in Table 5.9.

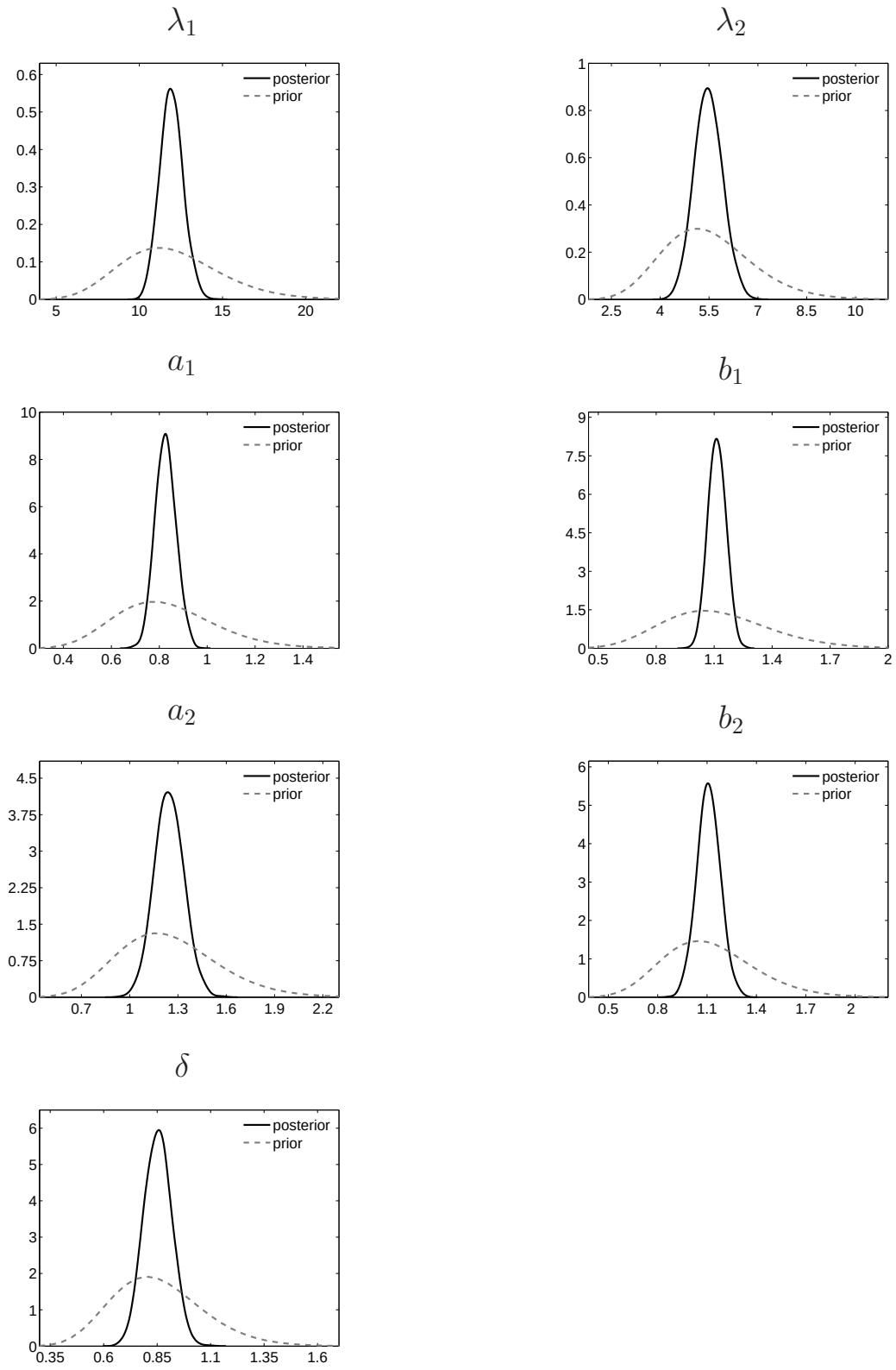


Figure 5.7: Comparison of marginal prior distributions and marginal posterior distributions for the parameters λ_1, λ_2, a_1, b_1, a_2, b_2 and δ in the (Weibull)^2 - Clayton model for the Danish fire insurance data.

	λ_1	λ_2
prior	11.94 [6.824, 18.46]	5.480 [3.132, 8.474]
posterior	11.93 [10.663, 13.30]	5.480 [4.682, 6.327]
	a_1	b_1
prior	0.8277 [0.4728, 1.279]	1.119 [0.6402, 1.732]
posterior	0.8262 [0.7478, 0.9122]	1.115 [1.0291, 1.203]
	a_2	b_2
prior	1.246 [0.7115, 1.925]	1.118 [0.6393, 1.729]
posterior	1.246 [1.0730, 1.433]	1.110 [0.9735, 1.247]
	δ	
prior	0.8562 [0.4893, 1.3239]	
posterior	0.8532 [0.7349, 0.9772]	

Table 5.9: Comparison of posterior mean estimates (for Γ_1 prior) to prior means. The estimates are given with 95% credible intervals which are based on the (empirical) 2.5% and 97.5% quantiles.

We strikingly see what happens when passing from the prior to the posterior distribution. The data contains lots of information and adds substantially to the knowledge about the parameters, which is expressed in a significant reduction of the uncertainty given by the 95% credible intervals. These are reduced to about one fourth of their original size. Hence, the posterior distributions are much less variable than the priors.

Finally, to conclude our analysis we have a look at the estimated number of single and joint losses in our two different approaches to the Danish fire insurance data. Table 5.10 states the estimated absolute numbers of losses per month together with the relative frequencies. These estimates are the means of the number of losses which were calculated for each of the 2000 samples via Equation (3.1.2).

Obviously, these estimates differ quite clearly for the two approaches. Whereas in our first approach every second loss affects both components, this is only the case for every fourth in the second approach. Instead, lots of losses occur only in the building component. Also the number of estimated absolute losses depends significantly on the approach which is, of course, explained by the removing of data points.

	single losses of building	single losses of content	joint losses
First approach	13.6 (34.5%)	6.1 (15.4%)	19.7 (50.1%)
Second approach	8.6 (61.0%)	2.1 (15.1%)	3.4 (23.9%)

Table 5.10: Comparison of estimated numbers of losses per month for the two different approaches. Absolute numbers together with relative frequencies.

Chapter 6

Final remarks

In this thesis we developed a Bayesian estimation procedure for the parameters of bivariate compound Poisson processes whose dependence is modelled by the Lévy Clayton copula. The Bayesian approach allows to take prior knowledge into consideration and permits to derive uncertainties about the parameters of interest.

Based on Sklar's theorem for Lévy copulas and the accordingly derived likelihood function of bivariate CPPs we developed a Markov chain Monte Carlo sampler for the specific class of Clayton models. By applying this sampler we were able to derive and analyze the posterior distributions.

A simulation study has proven that the sampler works very well. The elaborate dynamic adaption procedure for the proposal densities guarantees a satisfying mixing and convergence behaviour. Moreover, the posterior mean estimates match the simulation values very well. Significant differences between the maximum likelihood estimates and the posterior mean estimates were not detected. We have seen that the choice of the prior distributions plays an important role. One must be aware, that decisions based on the posterior mean estimates and the posterior distributions are affected also by the priors. Therefore, the choice of a certain informative prior must be well-founded, in particular when the data sets used for the analysis are small. Here the impact of the prior is, of course, higher than for large data sets.

Using the sampler we analyzed the Danish fire insurance data. We pursued two different approaches. First, we modelled the original data and illustrated the use of sliced distributions. In the second approach we transformed the data taking logarithms.

After an initial explorative data analysis we treated the two approaches separately. In each case we considered several Clayton models and selected the best among these models using Bayes factors. By introducing weighted Bayes factors we were able to analyze the fits of the models more detailed. The ultimately chosen models were the sliced

(Burr/GPD)^2 -Clayton model (thirteen parameters) and the (Weibull)^2 - Clayton model (seven parameters) for the two data sets, respectively. Finally, we investigated how the information about the parameters increased from the prior to the posterior distribution.

Let us emphasize that the intention of this thesis was to illustrate the procedure of model fitting and model selection for bivariate CPPs in a Bayesian context. The specific selection of the transformations and the prior distributions was exemplary. Practically, these choices have to be made carefully with respect to the application.

Moreover, we want to stress again that the presented Bayesian estimation procedure for the specific class of bivariate Clayton models can easily be carried over to other bivariate compound Poisson processes, as long as we know the likelihood function up to a normalizing constant. The likelihood for general (not necessarily Clayton) bivariate CPPs is given in Esmaeili and Klüppelberg (2010a). Thus, for any bivariate Lévy copula, which is chosen to model the dependence between the components of the CPP, our MCMC sampler can be adapted quite easily.

In many applications two-dimensional modelling is not sufficient and one has to consider a higher-dimensional case. One-parametric Lévy copula models would not be appropriate anymore, because one should be able to model the dependence between the different components more flexible. Combining several bivariate copulas might be a solution, but that increases the number of parameters considerably. Besides, new parameters for the additional margins have to be introduced. Thus, in those cases where it is possible to derive the likelihood function of the process, the large number of parameters could be problematic for frequentist approaches like maximum likelihood estimation. However, Markov chain Monte Carlo methods as introduced in this thesis should work fine also in such cases.

References

Basel Committee on Banking Supervision. (2004) International Convergence of Capital Measurement and Capital Standards. Basel.

Böcker, K. and Klüppelberg, C. (2006) Multivariate Models for Operational Risk. Submitted for publication. Preprint available at {*http://www-m4.ma.tum.de/Papers/*}

Böcker, K. and Klüppelberg, C. (2008) Modelling and Measuring Multivariate Operational Risk with Lévy Copulas. *Journal of Operational Risk* **3**, 3-27.

Böcker, K. and Klüppelberg, C. (2009) First Order Approximations to Operational Risk - Dependence and Consequences. In: Gregoriou, G.N. (Ed.), *Operational Risk Toward Basel III, Best Practices and Issues in Modeling, Management and Regulation.* Wiley, New York.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J.G. (2000) *Monte Carlo Methods in Bayesian Computation.* Springer, New York.

Chib, S. (1995) Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association* **90**, 1313-1321.

Cont, R. and Tankov, P. (2004) *Financial Modelling with Jump Processes.* Chapman & Hall/CRC, Boca Raton.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.

Esmaeili, H. and Klüppelberg, C. (2010) Parameter Estimation of a Bivariate Compound Poisson process. *Insurance: Mathematics and Economics*, to appear.

Esmaeili, H. and Klüppelberg, C. (2010) Parametric Estimation of a Bivariate Stable Lévy Process. Submitted for publication, Technische Universität München.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis.* Chapman & Hall/CRC, Boca Raton.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

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

Joe, H. (1997) *Multivariate Models and Dependence Concepts.* Chapman & Hall, London.

Kallsen, J. and Tankov, P. (2006) Characterization of Dependence of Multidimensional Lévy Processes Using Lévy Copulas. *Journal of Multivariate Analysis* **97**, 1551-1572.

Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing.* Springer, New York.

Nelsen, R. B. (1997) *An Introduction to Copulas.* Springer, New York.

Robert, C. P. and Casella, G. (2000). *Monte Carlo Statistical Methods.* Springer, New York.

Robert, C. P. and Marin, J.-M. (2010) On Computational Tools for Bayesian Data Analysis. In: Böcker, K. (Ed.), *Rethinking Risk Measurement and Reporting: Uncertainty, Bayesian Analysis and Expert Judgement.* Risk Books, London, to appear.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms, *Biometrika* **83**, 95-110.

Sklar, A. (1959) Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8**, 229-231.

Tierney, L. (1994) Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics* **22**, 1701-1728.

List of Figures

List of Tables