Deterministic posterior approximations for a copula-based MCMC sampler

Jakob Krause

München, 13. März 2014

Deterministic posterior approximations for a copula-based MCMC sampler

Jakob Krause

Masterarbeit
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Jakob Krause
aus Halle (Saale)

München, den 13. März 2014

Erstbetreuer: Prof. Dr. Konstantinos Panagiotou
Zweitbetreuer: Prof. Dr. Dr. Fabian Theis

Abstract

Based on a question raised in [Girolami, Mira], this work investigates if one can increase the efficiency of the copula-based Markov Chain Monte Carlo sampler introduced in [Schmidl, Czado, Hug, Theis] by employing deterministic posterior approximations or methods based on direct evaluations of the posterior on a grid ('grid methods').

Contents

1 Introduction **3**

2 Preliminaries **5**
2.1 Probability Theory 5
2.2 Markov Chains 8
2.3 Vine - Copulas 12
2.3.1 Analytic solutions 15
2.4 Bayesian inference 16
2.5 Markov Chain Monte Carlo samplers 18
2.6 Deterministic posterior approximations 21
2.6.1 Laplace Approximation 22
2.6.2 Variational Inference 24
2.7 Profile Likelihoods 25

3 The examples **27**
3.1 A highly correlated normal distribution 27
3.2 Banana shaped distribution 29
3.3 A biokinetic model for the processing of Zirconium in the Human Body 29
3.4 A small compartment model 33

4 Results - Laplace Approximation **35**
4.1 Results - Laplace Approximation 40
4.1.1 Normal distribution 41
4.1.2 Banana shaped distribution 42
4.1.3 Compartment model 43
4.1.4 Zirconium model 44

5 Results - Grid methods **46**
5.1 Cube 47
5.1.1 Normal distribution 47
5.1.2 Banana shaped distribution 48
5.1.3 Compartment model 48
5.1.4 Zirconium model 49
5.2 Profile Likelihoods 49
5.2.1 Normal distribution 49
5.2.2 Banana shaped distribution 50

5.2.3 Compartment model 53
5.2.4 Zirconium model 53

6 Conclusion **55**

7 Bibliography **59**

1 Introduction

The problem of this work is located within the field of Bayesian parameter estimation via Markov Chain Monte Carlo(MCMC) simulation. MCMC algorithms can be used to sample from complicated distributions for which classical sampling procedures break down. See [Brooks et al] for an introduction to the topic. The downside is that they are computationally expensive and one has to put effort into maximizing their efficiency.

This work investigates such an efficiency question for the vine-copula based MCMC sampler introduced in [Schmidl, Czado, Hug, Theis]. This algorithm has been developed for Bayesian parameter estimation in dynamical systems modeling biological processes. It incorporates the dependence structure of the underlying problem by means of a vine-copula in the sampling process. By exploiting the problemspecific dependence structure this algorithm can outperform classical MCMC samplers (see [Schmidl, Czado, Hug, Theis]). It has been succesfully applied in the inference of dynamical models for biological processes in [Schmidl] and [Schmidl, Hug et al]. However, the algorithm requires several computationally costly preperation steps, prior to the actual sampling, to estimate the vine-copula encoding the dependence structure. So far this has been done by means of a classical MCMC prerun. For high dimensional problems this prerun and the subsequent copula estimation are very time-consuming and in terms of efficiency one can argue in favor of less customized samplers: *'... in high-dimensional systems, it is our experience that simpler sampling algorithms are often more efficient.'* ([Hug, Raue et al]).
When the copula-based sampler was introduced, in a comment to the article, it was suggested to circumvent the prerun by using deterministic posterior approximations for a rough estimate of the dependene structure to increase the efficiency (see [Girolami, Mira]). The investigation of this question is the objective of this work. In the progress we use the biological examples for which the sampler has been introduced for. They possess a rich and very non-standard dependence structure and one can expect that, by exploiting the dependence, one can get to a boost in efficiency when compared to a less customized sampler. In addition we investigate the question for 'simpler' examples, chosen to highlight specifical features and problems.

The copula-based sampler in question has been developed for the inference of biological dynamical systems. The distinctive feature of this algorithms is that it brings together for the first time the two highly popular tools for dependence modelling, copulas, and inference, MCMC samplers. On their own these tools have been extensively used in the modelling and inference of multidimensional systems. Copula models are applied in such diverse scientific disciplines as Geostatistics, Ecology ,and Finance (see, e.g., [McNeil, Frey, Embrechts]). MCMC samplers, on the other hand, can be found, e.g., in the fields of Computational Physics (see, e.g., [Stickler, Schachinger]) and Econometrics (see, e.g., [Greenberg]).

The structure of this thesis is as follows: The first goal is, to (heuristically) present the original copula-based sampler and compare it to the sampler based

on a deterministic posterior approximation or a grid method to illustrate the problem. Afterwards we introduce the necessary notions and notations from probability theory as well as the posterior approximations. This is followed up by the introduction of the examples, we use to carry out the sampling. We conclude with the results.

The problem

Before presenting the necessary mathematical background, we want to give a brief account of the problem we are dealing with.
The objective of this work is to refine an algorithm which is used for Bayesian parameter inference by Markov Chain Monte Carlo sampling. The sampler in question exploits the dependence structure of the underlying problem by incorporating it in a problemspecific proposal function. Before starting into the parameter estimation one needs to have an estimate of the underlying dependence structure. So far, this has been accomplished by using a computationally rather expensive MCMC prerun, a uniformization step using knowledge on the type of priors, and a subsequent copula estimation.

The idea, formulated in [Girolami, Mira], is now to substitute the mentioned prerun by a presumably less costly deterministic procedure and proceed as above by uniforization, estimation, and sampling. This deterministic posterior approximations will give us a standardized distribution approximating the posterior distribution in some 'optimal' way, depending on the used type of approximation. With a standardized distribution and an assumption on the type of marginals one now can either hope to find a formula giving us the dependence structure directly or estimate the copula based on samples from the approximated distribution. The question we want to investigate is: By not having to do the prerun and replacing it by a deterministic procedure we (presumably) save some computational costs. However, by carrying out an approximation and a subsequent estimation of the dependence structure one can assume that one looses some information in the process and that the quality of the copula estimated from the prerun will be better. This will (presumably) slow down the main algorithm of the approximation based sampler. The question we want to investigate is **What is faster overall?**
Since our investigation is based on a sampler specifically introduced to exploit the dependence structure we have to monitor what the different approximations do to this efficiency driver.

The copula based sampler has been introduced for the inference of parameters in dynamical systems. We will briefly touch upon this topic and recover some results presented in [Schmidl, Czado, Hug, Theis] to prove that the sampling results from the approximation based samplers can be used to make statements on dynamical systems. The efficiency question, however, will be treated seperately without a connection to the inference in dynamical systems. MCMC samplers are used in a variety of scientific disciplines and the question of efficient sampling procedures is not restricted to dynamical systems alone.

2 Preliminaries

The aim of this chapter is to introduce the basic notions and notations from probability theory, the theory of Markov Chains, Markov Chain Monte Carlo algorithms and pair-copula decompositions we need to rigorously understand the problem. It also contains an introduction to Bayesian inference, deterministic posterior approximations and profile likelihoods. The presentation follows Chapter 2 in [Schmidl]. In addition, we give a brief summary of the copula-based sampling algorithm.

2.1 Probability Theory

In this section basic definition and notations from probability theory will be stated. A thorough exposition to the theory of probability can be found in [Durret] or [Klenke].

Definition 2.1.1
A *probability space* is a triplet $(\Omega, \mathcal{A}, \mathbb{P})$ consisting of a set $\Omega \neq \emptyset$, a σ-Algebra $\mathcal{A}$ on Ω and a σ-additive measure $\mathbb{P} : \mathcal{A} \to [0,1]$ with $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$, called a *probability measure*.

Elements of $\mathcal{A}$ are called *events*. For $A \in \mathcal{A}$ with $\mathbb{P}(A) = 1$ we say that 'A occurs *almost surely*' often abbreviated as *a.s.*.

Definition 2.1.2
Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a measurable space $(E, \mathcal{E})$ we call a map $X : \Omega \to E$ a *random variable* if it is $(\mathcal{A}, \mathcal{E})$-measurable, i.e. it fulfills the property that

$$\forall A \in \mathcal{E} : X^{-1}(A) \in \mathcal{A} \, ,$$

where $X^{-1}(\cdot)$ denotes the preimage.

We can identify a random variable by its preimage in the following way: For $A \in \mathcal{E}$ write

$$\mathbb{P}(X^{-1}(A)) := \mathbb{P}(X \in A) := \mathbb{P}_X(A) \, .$$

The function $\mathbb{P}_X$ is called the *distribution* of X with respect to $\mathbb{P}$. It is a probability measure and we write $X \approx \mathbb{P}_X$ to denote that 'X is $\mathbb{P}_X$ distributed'.

Definition 2.1.3
For a real-valued random variable $X : \Omega \to \mathbb{R}^n$ on $(\Omega, \mathcal{A}, \mathbb{P})$ we call the function

$$F_X : \mathbb{R}^n \to [0,1], x \in \mathbb{R}^n \mapsto \mathbb{P}_X((-\infty, x_1] \times ... \times (-\infty, x_n])$$

the *(cumulative) distribution function* of X with respect to $\mathbb{P}$.

Definition 2.1.4
If $F_X : \mathbb{R}^n \to [0,1]$ can be represented in terms of a function $f_X : \mathbb{R}^n \to \mathbb{R}_0^+$ by means of the non-negative Lebesgue integral

$$F_X(x_1, ..., x_n) = \int_{-\infty}^{x_n} ... \int_{-\infty}^{x_1} f_X(x_1, ..., x_n) dx_1 ... dx_n \, ,$$

we call f_X the *density function* of X with respect to $\mathbb{P}$. If there is no ambiguity we will repress the dependence on X and write f instead of f_X and F instead of F_X.

If a distribution function F is sufficiently regular (allowing for the interchange of limit (or differenciation) and integral) at $x = (x_1, ..., x_n) \in \mathbb{R}^n$, the according probability density function f is given by

$$f(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 ... \partial x_n} F(x_1, ..., x_n) \ .$$

The next definitions are made, to acchieve control over partial information within a multidimensional random variable. Hence, we are interested in definitions of the notions introduced so far, for subvectors of random variables.

Definition 2.1.5
Consider a random variable $X = (X_1, ..., X_n) : \Omega \to \mathbb{R}^n$ with density function $f_X : \mathbb{R}^n \to \mathbb{R}$. Given a subset $\{n_1, ..., n_k\} \subset \{1, ..., n\}$ and defining the random variable $Y := (X_{n_1}, .., X_{n_k})$, and $\{n_1^{'}, ..., n_{n-k}^{'}\} := \{1, ..., n\} - \{n_1, ..., n_k\}$ the associated distribution function F_Y is given by

$$F_Y(x_{n_1}, ..., x_{n_k}) = \int_{-\infty}^{x_{n_1}} ... \int_{-\infty}^{x_{n_k}} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_X(u_1, .., u_n) du_{n_1^{'}} ... du_{n_{n-k}^{'}} du_{n_k} ... du_{n_1} \ .$$

F_Y is said to be the *marginal distribution function* of Y. Assuming w.l.o.g. that $\{n_1, ..., n_k\} = \{1, ..., k\}$ the associated density function f_Y is called *marginal density* and is given by

$$f_Y(x_1, ..., x_k) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_X(x_1, ..., x_k, u_{k+1}.., u_n) du_{k+1} ... du_n \ .$$

Definition 2.1.6
We now turn to the question of the distribution within a random variable $X = (Y, Z)$ given partial information.

More precisely, let $Y = (X_1, ..., X_k)$ and $Z = (X_{k+1}, ..., X_n)$ be two random variables. The *conditional distribution* of Y, given the realization $Z = (x_{k+1}, ..., x_n)$ for some $x_{k+1}, ..., x_n \in \mathbb{R}$, is defined as

$$F_{Y|Z}(x_1, ..., x_k | x_{k+1}, ..., x_n) := \int_{-\infty}^{x_k} ... \int_{-\infty}^{x_1} \frac{f_X(u_1, ..., u_k, x_{k+1}, ..., x_n)}{f_Z(x_{k+1}, ..., x_n)} du_1 ... du_k \ .$$

As before the corresponding *conditional density* function is given by

$$f_{Y|Z}(x_1, ..., x_k | x_{k+1}, ..., x_n) = \frac{f_X(x_1, ..., x_n)}{f_Z(x_{k+1}, ..., x_n)} \ .$$

For conditional probability density functions we also introduce the following result.

Bayes' Theorem

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let X and Y be two random variables. Then, the following equation holds

$$f_X(x|Y=y) = \frac{f_Y(y|X=x) \cdot f_X(x)}{f_Y(y)} \ .$$

The next step is the introduction of 'stability results', i.e. a notion of an equilibrium state for a stochastic system. In this context we need the notions of independence and expectation.

Definition 2.1.7

Given a finite sequence of random variables $X^1, ..., X^N$ on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Let $F_{X^1 \otimes ... \otimes X^N}(\cdot)$ denote the distribution function, called *joint distribution function*, corresponding to the random variable $(X^1, ..., X^N)$. The random variables $X^1, ..., X^N$ are said to be *independent* if the distribution function factorizes into the distribution functions of the individual random variables, i.e. for $x^i \in \mathbb{R}^d$

$$F_{X^1 \otimes ... \otimes X^N}(x^1, ..., x^N) = \prod_{i=1}^{N} F_{X^i}(x^i) \ .$$

Definition 2.1.8

Let $X : \Omega \to \mathbb{R}^n$ be an integrable random variable.

$$\mathbb{E}^{\mathbb{P}_X}[X] := \int_{\mathbb{R}^n} X d\mathbb{P}_X$$

is referred to as the *expectation* of X. If there is no ambiguity we write $\mathbb{E}[X]$.

Definition 2.1.9

A sequence of random variables $X^1, X^2, ...$ is said to be *identically distributed*, if $\mathbb{P}_{X^1} = \mathbb{P}_{X^i}$ for all $i \in \mathbb{N}$.

Let x^i denote a realization of X^i.

Theorem - Strong law of large numbers

Let $X, X^1, X^2, ...,$ be a sequence of identically distributed, independent random variables. Then

$$\mathbb{E}[X] = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} x^i \quad (a.s.).$$

This theorem is at the heart of every *Monte Carlo method*. If one is able to draw independent, identically distributed samples from a distribution $\mathbb{P}_X$ this theorem ensures that one can approximate the integral

$$\int_\Omega f(X) d\mathbb{P}_X = \mathbb{E}[f(X)]$$

as the limit of

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{n} f(x^i) \ .$$

However, to apply this result we have to be able to sample from the distribution $\mathbb{P}_X$ in an efficient way. Since inference problems ask questions on the expectation of certain random quantities this result is extremely useful and applied extensively (see, e.g., [Glasserman]). In the models we have in mind, however, we only know the distribution up to an (incomputable) constant. Therefore we have to rely on more evolved sampling techniques (here: Markov Chain Monte Carlo sampling) violating the assumptions of the theorem above. We have to ensure that we have an analogue of this theorem in the more general setup. To introduce this (so called 'ergodic theory') result is the motivation of the presentation of Markov Chains in the next subsection.

We furthermore define the following statistical quantities:

Definition 2.1.10

The *variance* is a measure of how much a random variable scatters. If X is square-integrable it is defined as

$$\mathbb{V}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

The *covariance* of two random variables is a weak measure of dependence, i.e how two random variables act together. For one dimensional random variables X, Y it is defined as

$$\mathbb{C}ov(X, Y) := \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] .$$

Note, that there are several examples of uncorrelated but dependent random variables.

For one dimensional random variables X, Y the *Pearson correlation* is defined as

$$\rho(X, Y) := \frac{\mathbb{C}ov(X, Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} .$$

It is a normalized version of the covariance.

For n-dimensional random variables $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_n)$ the according notions are defined via the following matrices

$$\mathbb{C}ov(X, Y) := (\mathbb{C}ov(X_i, Y_j))_{i,j=1,...,n} ,$$

$$\rho(X, Y) := (\rho(X_i, Y_j))_{i,j=1,...,n} .$$

2.2 Markov Chains

This section has the objective of introducing Markov Chains (in discrete time). This important subclass of stochastic processes has the distinctive feature of a 'limited memory'. The stochastic properties of an increment of the process is only dependent on the current state of the process. The technical goal of this section is, to introduce easy criterions that will give us valid Markov Chain

Monte Carlo samplers in the sense that the conditions are sufficient for the 'stability result' mentioned above.

Markov Chains are of great importance, because their structure allows the introduction of a so called kernel comprising their stochastic properties. The presentation follows Chapter 2 in [Schmidl] and, starting from the general concept of a stochastic process, only introduces the concepts necessary for applications in MCMC algorithms. For a thorough introduction to the topic see, e.g., [Meyn, Tweedie].

Definition 2.2.1 (Stochastic Process)
Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ a measurable space. A *stochastic process* $(X_t)_{i\in I}$ on some index set $I \subset \mathbb{N}$ is a function

$$X : \Omega \times I \to E$$

$$(\omega, t) \mapsto X_t(\omega)$$

such that the functions X_t are random variables (i.e., they are assumed to be $(\mathcal{A}, \mathcal{E})$ measurable).

For a fixed $\omega \in \Omega$ we call the map $t \mapsto X_t(\omega)$ a *trajectory* (or *sample path* or *realization*) of $(X_t)_{t\in I}$.

Definition 2.2.2 (Markov Chain)
Let $(\Omega, F, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ a measurable space. A stochastic process with values in E is called a *Markov Chain*, if for any measurable set $A \in \mathcal{E}$, any $T \in \mathbb{N}, T \geq 2$, and any realization $(x_0, ..., x_T)$ of $(X_t)_{t\in\{0,...,T\}}$ the random variable X_{T+1} is depending only on x_T, i.e.

$$\mathbb{P}_{X_{T+1}|X_0\otimes...\otimes X_T}(X_{T+1} \in A|x_0, ..., x_T) = \mathbb{P}_{X_{T+1}|X_T}(X_{T+1} \in A|x_T) \ .$$

Definition 2.2.3 (Stationary Process)
A stochastic process $(X_t)_{t\in I}$ is said to be *stationary*, if for all $\{t_1, ..., t_k\} \subset I$ and $\tau \in I$ the joint distributions of

$$X_{t_1+\tau}, ..., X_{t_k+\tau} \text{ and } X_{t_1}, ..., X_{t_k}$$

are equal, i.e.

$$\mathbb{P}_{X_{t_1+\tau}\otimes...\otimes X_{t_k+\tau}}(X_{t_1+\tau}, ..., X_{t_k+\tau}) = \mathbb{P}_{X_{t_1}\otimes...\otimes X_{t_k}}(X_{t_1}, ..., X_{t_k})$$

Assumption! From now on we will restrict ourselve to stationary (or *homogeneous*) Markov Chains!

Definition 2.2.4 (Kernel)
For a Markov Chain $(X_t)_{t\in I}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in E and a set $A \in \mathcal{E}$ the distribution

$$k(A|x) := \mathbb{P}_{X_{t+1}|X_t}(X_{t+1} \in A|X_t = x) = \int_A \mathbb{P}_{X_{t+1}|X_t}(dy|x)$$

is called the *(transition) kernel* from x to A.

Proposition 2.2.1
For the class of Markov Chains we can recover the joint distribution from the transition kernel, i.e. the transition kernel fully determines the stochastic properties of the respective Markov Chain, given the chain starts in $x^0 \in E$.

Proof A proof of this result can be found in [Schmidl, Prop 2.3].

Properties of Markov Chains

Definition 2.2.5
We call a distribution π *invariant* (or *stationary*) for the transition kernel k, if for any $A \in \mathcal{E}$

$$\pi(A) = \int_E k(A|x)\pi(dx) \; .$$

Definition 2.2.6
A stationary Markov Chain is called *reversible*, if for $A \in \mathcal{E}$

$$\mathbb{P}(X_{t+1} \in A | X_{t+2} = x) = \mathbb{P}(X_{t+1} \in A | X_t = x) \; .$$

Definition 2.2.7
A stationary Markov Chain with transition kernel $k(dy|x) = p(y|x)dy + r(x)\chi_x(dy)$ is said to satisfy the *detailed balance condition*, if there exists a probability density function π_d, such that

$$p(x|y)\pi_d(y) = p(y|x)\pi_d(x) \; .$$

The next theorem states that the detailed balance condition is a sufficient conidition for the existence of an invariant distribution and reversibility of the respective Markov Chain.

Theorem 2.2.1
If the detailed balance condition holds for a Markov Chain $(X_t)_{t \in I}$ with transition kernel k, then

1) the associated distribution π is invariant with respect to k
2) the Markov Chain is reversible.

Proof The proof of this result is given in [Schmidl, Thm 2.4].

Since we are interested in the inference of a distribution independent from the starting value we have to impose further conditions to ensure that the invariant distribution is unique. This gives raise to the notion of an *equilibrium distribution.*

Definition 2.2.8
If every Markov Chain governed by the transition kernel k is converging to the same invariant distribution π, independent of the starting value $x_0 \in E$, we call π an *equilibrium distribution.*

Further properties of Markov Chains we need to ensure the existence of an equilibrium distribution are the following

Definitions 2.2.9

A Markov Chain $(X_t)_{t\in I}$ with transition kernel k is said to be π-irreducible for a σ-finite measure π, if for any $x \in E$ and $A \in \mathcal{E}$ with $\pi(A) > 0$ there exists an $m \in \mathbb{N}$ such that the *m-step transition kernel*

$$k^m(A|x) := \int_E k^{m-1}(A|y)k(dy|x)$$

is positive, i.e,

$$k^m(A|x) > 0 \ .$$

If $m = 1$ we call a Markov Chain *strongly π-irreducible.*

Interpretation: An irreducible Markov Chain can reach any point in the state space in a finite number of steps with a positive probability.

A π-irreducible Markov Chain with transition kernel k is periodic, if for some integer $n \geq 2$ there exists a sequence $(E_0, E_1, ..., E_{s-1})$ of pairwise disjoint non-empty sets $E_i \in \mathcal{E}$ such that for all $i = 0, ..., s-1$ and all $x \in E_i$

$$k(E_j|x) = 1 \ for j = i + 1 \bmod s \ .$$

A π-irreducible Markov Chain is aperiodic if it is not periodic. Interpretation: An aperiodic Markov Chain does not allow deterministic circles.

Let $\mathbb{P}_x(A)$ be the probability that, starting at $x \in E$ we obtain for the number $c_t(A) := |x_s \in A|0 \leq t|$ of visits to some subset $A \in \mathcal{E}$ that $c_t(A) \to \infty$ for $t \to \infty$.

A Markov Chain is *Harris recurrent*, if there exists an invariant distribution π such that for every $A \in \mathcal{E}$ with $\pi(A) > 0$

$$\mathbb{P}_x(A) = 1 \ \forall x \in E \ .$$

Interpretation: The probability that the Markov Chain visits every point of the state space infinitely often is 1.

Given the definitions above we can formulate a theorem providing us with the existence of the (unique) equilibrium distribution.

Theorem 2.2.2

Let $(X_t)_{t\in I}$ be a π-irreducible, aperiodic and Harris recurrent Markov Chain with transition kernel k and invariant distribution π. Then

1) k is positive Harris recurrent

2) π is the equilibrium distribution

3) k is ergodic for π, i.e. $(X_t)_{t\in I}$ converges regardless of its starting value $x_0 \in E$.

This theorem states the conditions we have to check, in order ensure that a given MCMC method is 'valid'.
The following result (Theorem 2.6 in [Schmidl]) is the corresponding stability result for Markov Chains mentioned in context of the Law of large numbers.

Theorem 2.2.3
Suppose $(X_t)_{t\in I}$ is a positive Harris recurrent and aperiodic Markov Chain with invariant distribution π. Suppose furthermore that $f : E \to \mathbb{R}$ is π-integrable, i.e. $\int_E |f(x)|\pi(dx) < \infty$. Then for a realization $(x_t)_{t\in I}$ we have

$$\frac{1}{m+1}\sum_{t=0}^{m} f(x_t) \to \int_E f(x)\pi(dx) = \mathbb{E}_\pi\left[f(E)\right] \ ,$$

where $f(E)$ denotes the image of E under f.

2.3 Vine - Copulas

The notion of a copula is used to factorize a continous n-variate distribution function $F(\cdot)$ into a part containing information only on the marginals of the distribution and a second part containing the dependence structure.

Definition 2.3.1
A function $C : [0,1]^n \to [0,1]$ is called n-dimensional copula, if it fulfills the following properties

1.) $C(u) = 0$ for all $u \in [0,1]^n$ with $u_i = 0$ for some i

2.) $C(u) = u_i$ for all $u \in [0,1]^n$ with $u_j = 1$ for $j \neq i$.

3.) $C(u)$ satisfies the rectangle inequality, i.e. for each hyperrectangle $B = \prod_{i=1}^{d}[x_i, y_i] \subset [0,1]^d$ the volume of B measured by C is non-negative, i.e.

$$\int_B dC(u) = \sum_{z\in\times_{i=1}^{d}\{x_i,y_i\}} (-1)^{|k:z_k=x_k|}C(z) \geq 0 \ .$$

The following theorem shows that the class of copulas is sufficiently rich to describe joint distributions. It states that for any joint distribution with given marginals there exists a copula and vice versa, i.e. any set of marginals together with a copula will give a well-defined joint distribution.

Although we will use Sklar's theorem repeatedly in the construction of pair copula decompositions, we will not give a proof here. (see, e.g., [McNeil, Frey, Embrechts] or [Nelsen]).

Theorem (Sklar)
Suppose F is an n-dimensional distribution function with continous univariate

marginals $F_1, ..., F_n$. Then there exists a unique copula C, such that for all $x = (x_1, ..., x_n)' \in \mathbb{R}^n$

$$F(x) = C(F_1(x_1), ..., F_n(x_n)) \ .$$

Conversely, for any copula C and univariate distributions $F_1, ..., F_n$ the function F defined by $F(x) = C(F_1(x_1), ..., F_n(x_n))$ is a multivariate distribution function with margins $F_1, ..., F_n$.

In our MCMC application we will use special classes of density functions. From this perspective it is desirable to immediately draw a connection between a distribution function, its marginals and the role of the copula therein.

Suppose C and F are sufficiently regular. Then, by the chain rule, we get

$$f(x) = \frac{\partial C(F_1(x_1), ..., F_n(x_n))}{\partial F_1(x_1)...\partial F_n(x_n)} \prod_{i=1}^{n} f_i(x_i)$$

at some point $x = (x_1, .., x_n) \in \mathbb{R}^n$. This factorization suggests

$$c(F_1(x_1), ..., F_n(x_n)) := \frac{\partial C(F_1(x_1), ..., F_n(x_n))}{\partial F_1(x_1)...\partial F_n(x_n)} \ .$$

Hence, every probability density function f can be decomposed as a product

$$f(x) = c(F_1(x_1), ..., F_n(x_n)) \prod_{i=1}^{n} f_i(x_i) \ .$$

As mentioned before, the copula part models the dependence structure between the components of a random vector. Classically (as in the case of Archimedian or Elliptical copulas - see [McNeil, Frey, Embrechts]), one only had the possibility to choose one model of dependence for all the relationships between the components of a random vector.

The class of useful copula models has been considerably extended by the usage of pair copula decompositions introduced in [Bedford,Cooke 1,2] allowing for much more flexibility in the dependence model. 'Useful' here refers to efficient sampling and estimation procedures one needs for practical applications (see [Mai, Scherer]). We follow [Aas et al] for the introduction of these pair copula decompositions.

Suppose $X = (X_1, ..., X_n)'$ is a random vector with distribution function $F(x_1, ..., x_n)$ and a probability density function $f(x_1, ..., x_n)$. Then, given the definitions from the section on probability theory, we can write

$$f(x_1, ..., x_n) = f(x_n) \cdot f(x_{n-1}|x_n) \cdot f(x_{n-2}|x_{n-1}, x_n) \cdot ... \cdot f(x_1|x_2, ..., x_n) \ . \quad (1)$$

By using Sklars Theorem repeatedly we derive a pair copula decomposition iteratively in the following way.

$n = 2$ By Sklars theorem we know that we can represent the joint distribution $f(x_1, x_2)$ as

$$f(x_1, x_2) = c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2) \ . \quad (2)$$

In addition we know that we can represent $f(x_1, x_2)$ as

$$f(x_1, x_2) = f(x_2) \cdot f(x_1|x_2) \ . \tag{3}$$

This yields

$$f(x_1|x_2) = c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \ . \tag{4}$$

$n = 3$ We consider the conditional joint distribution $f(x_1, x_3|x_2)$ and the two representations

$$f(x_1, x_3|x_2) = c_{13|2}(F_1(x_1|x_2), F_3(x_3|x_2)) \cdot f(x_1|x_2) \cdot f(x_3|x_2) \tag{5}$$

and

$$f(x_1, x_2|x_3) = f(x_3|x_2) \cdot f(x_1|x_2, x_3) \ , \tag{6}$$

i.e.

$$f(x_1|x_2, x_3) = c_{13|2}(F_1(x_1|x_2), F_3(x_3|x_2)) \cdot f(x_1|x_2) \ . \tag{7}$$

By using the case $n = 2$ we can continue to

$$f(x_1|x_2, x_3) = c_{13|2}(F_1(x_1|x_2), F_3(x_3|x_2)) \cdot c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \tag{8}$$

implying

$$f(x_1, x_2, x_3) = f_3(x_3) \ \cdot \ c_{23}(F_1(x_1), F_2(x_2)) f_2(x_2)$$

$$\cdot \, c_{13|2}(F_1(x_1|x_2), F_3(x_3|x_2)) c_{12}(F_1(x_1), F_2(x_2)) f_1(x_1) \tag{9}$$

for the decomposition of the joint distribution.

general n For general $n \in \mathbb{N}$ we get that the decomposition of the conditional density function $f(x_t|x_{t+1}, ..., x_n)$ is given by

$$f(x|v) = c_{xv_j|v_{-j}}(F(x|v_{-j}), F(v_j|v_{-j})) \ \cdot \ f(x|v_{-j}), \tag{10}$$

where $v = (v_1, ..., v_n)$ is an n-dimensional vector and v_j is its j'th component and $v_{-j} = (v_1, ..., v_{j-1}, v_{j+1}, ..., v_n)$ is the $n - 1$ dimensional vector missing the $j_t h$-component of v.

Note that in general this conditional pair copula densities depend on the conditional values v_{-j}. We will assume that this is not the case, i.e. that the dependence on the conditioning values is captured by the values of $F(x_t|v_{-j})$ and $F(v_j|v_{-j})$. [Hoebaek Haff et al] argue that this is not a severe assumption.

Hence we can iteratively decompose an n-dimensional joint distribution into its components.

The decomposition given above is by no means unique. For example, in the case $n = 3$ one could have used a similar argument to decompose for $f(x_1, x_2|x_3)$ instead of $f(x_1, x_3|x_2)$. To develop efficient estimation and sampling algorithms,

it is advisable to introduce some kind of 'standardized layout' for the decomposition. This gives raise to the notion of 'vines' and their associated 'vine copulas'. For a thorough introduction to vine copulas see [Kurowicka, Joe]. The decomposition given above for $n = 3$ corresponds to the so called *D-vines*. To formally introduce this notion we need some definitions from graph theory

Definition 2.3.2

A *graph* is a pair $\mathcal{G} = (V, E)$ of sets V and $E \subset V \times V$, where for $k, l \in \mathbb{N}$ $V = \{v_1, ..., v_k\}$ is a set of *vertices* and $E \subset (V \times V) = \{e_1, ..., e_l\}$ is a set of *edges*. In the following we equip E with an equivalence relation by identifying (v_i, v_j) and (v_j, v_i) for all $i, j \in \{1, ..., k\}$. This corresponds to the notion of an *undirected graph*.

We call $P = (V^*, E^*)$ a *path* if for a pairwise disjoint set of vertices $\{v_1^*, ..., v_{m-1}^*\}$ and an additional vertex v_m^* it holds that $V^* = \{v_1^*, ..., v_m^*\} \subset V$, and $E^* = \{(v_1^*, v_2^*), (v_2^*, v_3^*), ..., (v_{m-1}^*, v_m^*)\} \subset V \times V$. P is said to be *cyclic*, if $v_m^* = v_1^*$.

A graph is called *acyclic* if it does not contain any cyclic paths. An acyclic graph $T = (V, E)$ is called a *tree*. For more notions on graph theory, see [Diestel].

With these definitions we can follow up by introducing 'vines':

Definition 2.3.3
A *regular vine* on $n \in \mathbb{N}$ vertices is a collection of $(n-1)$ trees $\mathcal{V} = (T_1, ..., T_{n-1})$ such that:

1.) $T_1 = (V_1, E_1)$ has the set of vertices $V_1 = \{v_1, ..., v_n\}$.

2.) For $i = 2, ..., n-1$ the set of vertices $T_i = (V_i, E_i)$ is given by $V_i = E_{i-1}$.

3.) For $i = 2, ..., n-1$ every element $(v_i, v_i^{'}) \in E_i$ consists of two elements $(v_{i-1}, v_{i-1}^{'})$ and $(w_{i-1}, w_{i-1}^{'}) \in E_{i-1}$ where exactly one of the v's coincides with one the w's.

Definition 2.3.4: A regular vine is called *D-vine*, if the degree of each vertex v in T_1 is at most two, i.e. v is contained in at most two edges of E_1.

An illustration of a D-vine copula can be found in Figure 1.

Now we want to illustrate a situation in which we can identify the pair-copula decomposition of a jointly normally distributed random variable under the assumption of standard-normal marginals. For more general situations, similar results seem not to be known.

2.3.1 Analytic solutions

Under the assumption that the marginals of our model are standard-normal and the posterior is a joint normal distribution the corresponding copula is a gaussian copula. This copula can be (analytically) indentified with the corresponding

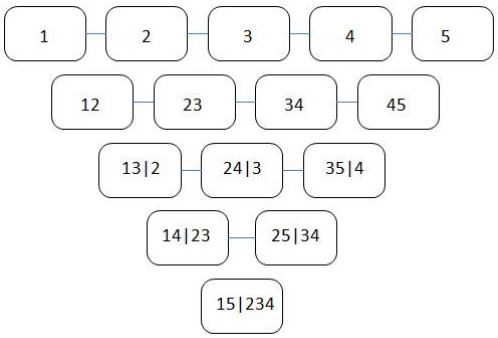


Figure 1: **D-Vine structure for 5 random variables** The nodes of each of the five trees are enumerated by conditioned and conditioning set of the copula assigned to it. For an n-dimensional model the number of copulas to be estimated is $\frac{n*(n-1)}{2}$.

pair copulas by choosing the pair copulas to be also gaussian copulas with coefficients according to the partial correlation. The following 3-dimensional example is taken from [Aas et al] (Section 2.6): Given a multidimensional normal distribution with correlation matrix

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$$

, and standard normal marginal distributions, the pair copula decomposition of f is given by: c_{12} Gaussian copula with parameter ρ_{12}, c_{23} Gaussian copula with parameter ρ_{23} and $c_{13|2}$ Gaussian copula with parameter

$$\rho_{13|2} = \frac{\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{23}^2} + \rho_{12} \cdot \rho_{23}} \ ,$$

the so called *partial correlation*. Similar results (i.e. a clear analytic identification of the copula) for non-normal joint distributions or normal-distributions with non-normal marginals seem not to be known. In addition this result seems not useful in the context of simulation, since there exist several streamlined algorithms to sample from multivariate normal distributions (see, e.g., [McNeil, Frey, Embrechts] or [Glasserman]). In terms of efficiency it is not feasible to sample from a mutlivariate normal distribution by an MCMC sampler.

2.4 Bayesian inference

In this section we want to give a brief account of Bayesian Inference. The presentation follows Chapter 3 of [Schmidl] and is occasionally supported by comments from [Lawrence et al].

The main advantages of employing Bayesian techniques are the following:

Since we can expect our data to contain a considerable amount of noise we are not just interested in some 'optimal' value for the parameters of our system but also in the stability of those parameters with respect to the noise. In addition one might not have access to a sufficient amount of data. The Bayesian approach allows for a control of the uncertainties caused by those effects.

The Bayesian framework will allow us to incorporate opinions we might have on the distribution of parameters into the statistical procedure. This allows us to incorporate more information on the problem than classical statistical procedures, like e.g. Maximum Likelihood Estimation, leading - presumably - to better results.

The metholodical difference of classical, or 'frequentistic', approaches in contrast to the Bayesian ones lies in the interpretation of the parameters of the underlying system. Classically one views observations $y = \{y_1, ..., y_n\}$ as realizations of a model given a parameter vector $\xi = (\xi_1, ..., \xi_m)$, i.e. samples from the random variable $X \approx f(y|\xi)$ for some density function f. In the Bayesian approach, however, we will view the parameter vector $\xi = (\xi_1, ..., \xi_m)$ as a realization of the observations $y = \{y_1, ..., y_n\}$, i.e. we want a distribution of the type $\pi(\xi|y)$. This distribution is called the *posterior distribution* and the introduction of the object $\pi(\xi|y)$ is the goal of this section.

Given the density function $f(\cdot|\xi)$ we define the *likelihood function* or simply *likelihood* for $y = (y_1, ..., y_n)$ by

$$L(y|\xi) = L(y_1, ..., y_n|\xi) := \prod_{i=1}^{n} f(y_i|\xi) \ .$$

The likelihood contains the information on how well a given set of observations is explained by the model, given the model is parametrized by $\xi = (\xi_1, ..., \xi_n)$. In addition we introduce the *prior density* $\pi(\xi)$, a function containing information on the distribution of the paramters before the observation of y. We can now define the *posterior distribution* $\pi(\xi|y)$ by

$$\pi(\xi|y) := \frac{L(y|\xi) \cdot \pi(\xi)}{p(y)} \ ,$$

where $p(y)$ is the so called *model evidence* (or *marginal likelihood*) and is defined by

$$p(y) := \int_{\mathbb{R}^n} L(y|\xi) \cdot \pi(\xi) d\xi \ .$$

The posterior distribution $\pi(\xi|y)$ is well defined by Bayes' theorem. The model evidence is needed to ensure that π is indeed a probability distribution. Note that for a fixed vector of observations y the model evidence is just a constant.

In general, the model evidence is analytically and numerically intractable. However, it plays a crucial role in the field of model selection and hence there

are methods that deal with its approximation (see, e.g., [Schmidl, Hug et al] or [Schmidl, Ch. 3.4]).

However, in the following we will introduce a set of algorithms that does not depend on the knowledge of the model evidence. Hence, we assume that the constant is unknown.

2.5 Markov Chain Monte Carlo samplers

In this section we want to adress the question of constructing a Markov Chain with a given equilibrium distribution π. If one is able to construct such a Markov Chain this will provide us with a method to sample from the target distribution π.

This section is entirely devoted to the discussion of one particular MCMC sampler, called the 'Metropolis-Hastings algorithm'. This type of sampler allows us, in addition to sampling from the distribution π, to circumvent the problem of an incomputable 'model evidence', i.e. to sample from π by only having access to $\frac{1}{c}\pi$ for an unknown constant c. In addition this allows for more general prior choices, since they do not have to be normalized.

The presentation of this chapter is based on [Liang et al] and [Del Moral] with occasional supplementary comments from [Lawrence et al].

For a historic overview, see [Brooks et al].

Construction of the Markov Chain

Given a target distribution π with density function π_d, we want to construct a Markov Chain X with equilibrium distribution π. As pointed out earlier (Proposition 2.2.1), it is sufficient to construct the kernel of the Markov Chain. For our goal we need to be able to express the kernel K in terms of the target distribution π (or its density function π_d). To do this, we use precisely the properties of Markov Chains defined in section 2.2. From this point of view we note that the detailed balance condition already provides a connection between a part of the kernel $k(x, y) = p(x, y) + r(x)$ and π_d:

$$p(x|y)\pi_d(y) = p(y|x)\pi_d(x)\,.$$

The following scheme, called the *Metropolis-Hastings algorithm* provides us with a solution to our problem:

1.) Given that the Markov Chain $(X)_{n\in\mathbb{N}}$ is in state X_n, draw a proposal y from a proposal function $q(y|X_n)$.

2.) Compute the acceptance ratio

$$\alpha(X_n, y) = \min\left\{1, \frac{\pi_d(y)q(X_n|y)}{\pi_d(X_n)q(y|X_n)}\right\}\,.$$

and set $X_{n+1} = y$ with probability $\alpha(x, y)$ or $X_{n+1} = X_n$.

For one run of this scheme we note the following facts:

1.) Although there is a dependence on the density function π_d (containing the incomputable constant) there is no dependence on the 'model evidence' since it factors out.

2.) For one run of this scheme we only need to know the current state of the process X_n and a proposal function $q(\cdot,\cdot)$ of our choice. I.e. the stochastic transition from X_n to X_{n+1} fulfills the Markov Chain definition.

3.) The $\alpha(\cdot,\cdot)$ in step 2 is chosen such that the detailed balance condition holds:

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

If $\alpha(x,y) = 1$, then $\alpha(y,x) = \frac{\pi_d(x)q(y,x)}{\pi_d(y)q(x,y)}$, implying the detailed balance condition. The other case follows similar.

To be more precise, the scheme provides us with the following kernel

$$k(x,y) = q(x,y) \cdot \alpha(x,y) + \delta_x \cdot (1 - \alpha(x,y)) \ .$$

We already saw that the kernel fulfills the detailed balance condition. By Theorem 2.2.1 this implies that there exists an invariant distribution, independent of the choice of the proposal function $q(\cdot)$. However, the performance of this scheme is highly dependent on the choice of the proposal function! In their paper, [Schmidl, Czado, Hug, Theis] made this algorithm problem specific in the sense that they incorporated the dependence structure of the parameters. To obtain a valid sampling procedure one has to check the existence of an equilibrium distribution. To apply the respective theorem one has to check the properties π-irreducibility, aperiodicity and Harris recurrence. Those properties are dependent on the choice of q. For the problemspecific copula-based proposal function this has been done in [Schmidl].

Problems of the Metropolis-Hastings sampler

Although Metropolis-Hastings samplers allow to sample from a large set of distributions we conclude this section by some remarks on their drawbacks. Essentially they are related to the question of efficiency. We want to mention the following three problems

1.) Correlation within the chain: Although in the long run the samples from a Metropolis-Hastings sampler follow the equilibrium distribution π, samples that are close to each other in the Markov Chain are dependent on each other. This means that if one wants to have a set of independent samples from the posterior distribution, one has to discard a number of samples from the chain to achieve this goal. If the dependence, e.g. measured by the autocorrelation, within the samples is high this can make the Metropolis-Hastings sampler very slow.

2.) Convergence to the equilibrium distribution: When one starts to draw samples via a Metroplis-Hastings algorithm one has to wait until the Markov Chain converges to its stationary distribution π. This might take a significant amount of time. Classical estimates on this convergence topic can be found in [Meyn, Tweedie]. More general results (for so called Feynman-Kac semigroups) can be found in Chapters 4 and 5 of [Del Moral]. Note that this more general setup still allows for Metropolis-Hastings type algorithms extensively used in various scientific disciplines.

3.) Getting stuck: If the Markov Chain is in a local optimum of the posterior distribution, the corresponding α might be very low. Consequently it is possible that the Markov Chains stays in the local optimum for a very long time.

Copula based MCMC sampling

We can now give a short summary of the copula based sampling procedure based on a prerun. This section follows Section 4 in [Schmidl, Czado, Hug, Theis]. The sampling scheme consists of 4 steps:

1.) A prerun: Via any given Markov Chain sampler an initial Markov Chain $(X_i)_{i\in\{1,...,T\}}$ for some $T > 1$ is sampled. We can view $(X_i)_{i\in\{1,...,T\}}$ as a matrix in $\mathbb{R}^{n\times T}$ where n denotes the dimension of the state space, i.e. for every $i \in \{1, ..., T\}$ we have $X_i = (X_{i,1}, ..., X_{i,n})'$.

2.) A uniformization step of the prerun samples: Since copulas are defined on $[0, 1]^n$ we have to transform the samples of step 1 to the copula domain. This is done by fitting the samples in the respective dimension by the class of distribution given by the prior, i.e. $(X_{i,1})_{i\in\{1,...,T\}}$ are used to fit a distribution function $G_{\theta_1,1}(\cdot)$ of the type the first prior has, and similarly for the other dimensions. By using the fitted marginals $(G_{\hat{\theta}_j,j}(\cdot))_{j\in\{1,...,n\}}$ we can transform $X_i = (X_{i,1}, ..., X_{i,n})'$ to $Y_i = (G_{\hat{\theta}_1,1}(X_{i,1}), ..., G_{\hat{\theta}_n,n}(X_{i,n})) \in [0, 1]^n$ in the copula domain.

3.) A D-vine copula estimation based on the prerun samples: Using $(Y_i)_{i\in\{1,...,T\}}$ an η-parametrized D-vine copula density

$$c_{1...n}(Y|\eta) = \prod_{j=1}^{n-1}\prod_{i=1}^{n-j} c_{j,j+i|j+1...j+i-1}(F(y_{j+1...j+i-1}, \eta), F(y_{j+i}|y_{j+1...j+i-1}, \eta)|\eta)$$

is fitted by the AIC approach from section 3 in [Schmidl, Czado, Hug, Theis]. The vector $\eta = (\eta_{i|j+i|(j+1)...(j-1+i)})_{j\in\{1,...,n-1\},i\in\{1,...,n-j\}}$ contains the types and parameters of the copulas.

4.) The copula-based Markov Chain Monte Carlo sampling: First, a sample $\tilde{Y} = (\tilde{Y}_1, ..., \tilde{Y}_n) \approx c_{1...n}(y|\hat{\eta})$ is drawn from the estimated copula

$c_{1...n}(y|\hat{\eta})$. By the marginals from step 2) the sample is transformed to $\tilde{X} = (G^{-1}_{\hat{\theta}_1,1}(\tilde{Y}_{i,1}), ..., G^{-1}_{\hat{\theta}_1,n}(\tilde{Y}_n)$. This yields a proposal from the density

$$q_1(X|\hat{\theta},\hat{\eta}) := c_{1...n}(G_{\hat{\theta}_1,1}(X_1), ..., G_{\hat{\theta}_n,n}(X_n)|\hat{\eta}) \cdot \prod_{i=1}^{n} g_{\hat{\theta}_i,i}(X_i) \ .$$

To ensure convergence of the proposed sampling scheme, a heavy tailed distribution is added to make the proposal function uniformly heavier in the tails than the posterior distribution. With this, the strong Doeblin condition is satisfied (see [Holden et al]). In addition, the Metropolis Hastings acceptance probability satisfies the detailed balance condition. In [Holden] it was shown that these two conditions imply the convergence of the respective Markov Chain.

In addition, a third density function is added to allow for regular Random Walk Metropolis Hastings steps. Concluding, the proposal density function looks as follows

$$q^{cop}(X_{n+1}|X_n,\hat{\theta},\hat{\eta}) := r_1 q_1(X_{n+1}|\hat{\theta},\hat{\eta}) + r_2 q_2(X_{n+1}|X_n) + (1-r_1-r_2) q_3(\theta), \tag{11}$$

where $r_1, r_2 \in [0,1)$ are constants with $r_1 + r_2 < 1$.

The problem revisited

We see that substituting the prerun in step 1) of the sampling procedure above has no logical implication on the sampling scheme. Any other way, to come up with samples from the posterior can be used to start the sampling scheme. Below we will investigate if the efficiency of the copula-based sampler can be increased, by using samples generated by other means than a MCMC prerun. To this end deterministic posterior approximations and profile likelihoods are introduced.

2.6 Deterministic posterior approximations

As pointed out in the section on Bayesian inference one main obstacle for Bayesian methods is the computation of the 'model evidence' resulting in an unknown scaling factor of the posterior distribution. This section introduces techniques to find standard distributions that approximate the posterior by means of properties that do not require knowledge on the model evidence constant.

In context of our problem the idea is the following: Given these techniques we approximate the posterior by a structurally easier distribution preserving as much weight of the true distribution as possible. Since the approximated distributions are structurally easier one can hope to find deterministic estimators for the parameters of the copula or at least draw samples from them, which one can use to fit the copula to. One question we have to monitor is whether the concept 'inclusion of the dependence structure in the sampling process' for which

the given sampler has been developed for in the first place is compatible with the assumptions and construction principle of the posterior approximations. We claim that this will not be the case for methods relying on variational inference. The calculations below do not rely on access to the model evidence.

2.6.1 Laplace Approximation

The posterior is approximated by a normal distribution identified by the Maximum Posterior Estimate and the curvature at this point. Note that the location of the Maximum Posterior Estimate is not influenced by monotone transformations (i.e. in particular one does not have to know the model evidence constant and one can transform the posterior (modulo constant) by monotone transformation like the logarithm to make it analytically more tractable).

For reasons of readability we will give the derivation of Laplaces method only in the 1-dimensional case. Higher dimensions follow analogously by substitution of one-time derivatives by gradients and second-order derivatives by Hessians.

Derivation

Initially, the Laplace approximation was introduced, to give approximations for integrals of the form

$$\int e^{S(x)} dx \ .$$

To this end, the function we will use in the derivation will be the logarithm of the posterior, i.e.

$$\int e^{S(x)} dx = \int e^{\ln p(x)} dx = \int p(x) dx \ ,$$

where one identifies the integral on the right as the model evidence.

Let $p(x) := \frac{f(x)}{\alpha}$ be the posterior with the unknown model evidence α. The Taylor serios expansion of $\ln f(x)$ is

$$\ln f(x) = \ln f(x_0) + \frac{\partial \ln f(x)}{\partial x}(x_0) \cdot (x - x_0) + \frac{1}{2} \frac{\partial^2 \ln f(x)}{\partial x^2}(x_0) \cdot (x - x_0)^2 + \text{higher terms} \ .$$

Consider the first order term

$$\frac{\partial \ln f(x)}{\partial x}(x_0) \cdot (x - x_0) = \frac{1}{f(x)} \frac{\partial f(x)}{\partial x}(x_0) \cdot (x - x_0) \ .$$

This term is zero in a maximum (local or global) x_{max} of $f(x)$.

This means that at a maximum the Taylor expansion simplifies to

$$\ln f(x) = \ln f(x_{max}) + \frac{1}{2} \frac{\partial^2 ln f(x)}{\partial x^2}(x_{max}) \cdot (x - x_{max})^2 + \text{higher terms} \ .$$

After an exponentiation and assuming that the higher order terms are negligable we reach

$$\exp(\ln f(x)) = \exp\left[\ln(f(x_{max})) + \frac{1}{2} \frac{\partial^2 \ln f(x)}{\partial x^2}(x_{max}) \cdot (x - x_{max})^2 \right] \ .$$

which corresponds to a gaussian density with mean x_{max} and volatility $-\frac{1}{f''(x_{max})}$.

In conclusion: The steps we need to do in order to apply the Laplace Approximation on a posterior distribution $p(\cdot)$ are

1.) Find the Maximum Posterior estimate $\hat{\theta}$.

2.) Calculate the Hessian at $\hat{\theta}$: $H_{\ln(p(\cdot))}(\hat{\theta})$.

3.) The Laplace Approximation of p is given by $\mathcal{N}(\hat{\theta}, -H_{\ln(p(\cdot))}(\hat{\theta})^{-1})$.

Discussion: Based on the steps of the approximation one immediately sees that this approximation is very sensitive with respect to the local shape of the posterior at the mode and in addition requires the posterior to be a C^2-function, at least for the point at which the Hessian is calculated. This last point will turn out to be a problem for an example with a triangle distributed prior. At the maximum of a triangle distribution the Hessian is not defined. In addition, since the dependence structure is a very non-local property, and this approximation only considers the curvature at one point, we cannot expect this approximation to work very well. In fact, given a fixed set of marginals, every type of dependence structure in the original posterior distribution is simply substituted to the dependence structure compatible with the priors and a joint normal distribution.

> By Sklar's Theorem, and given a fixed set of marginals, we can decompose the posterior distribution $p(\theta)$ into a copula and a set of marginals $(f_i(\cdot))_{i\in\{1,\dots,n\}}$
>
> $$p(\theta) = c_p(\cdot)f_1(\theta_1)\cdot\ldots\cdot f_n(\theta_n)\ .$$
>
> If we now approximate the left hand side by a normal distribution and stay with the same marginals the copula is (on the basis of the evaluation of the curvature of one point of the true posterior) substituted by the one that is compatible with the marginals and the approximated posterior distribution
>
> $$p_{\text{app}}(\theta) = c_{p_{\text{app}}}(\cdot)f_1(\theta_1)\cdot\ldots\cdot f_n(\theta_n)\ .$$
>
> The existence of this copula is again guaranteed by Sklar's Theorem.

For standard-normal priors, for example, this would correspond to the gaussian copula. However, the dependence structure after the approximation is solely based on this compatibility and is, in general, chosen out of very limited information on the original problem.

In context of the copula-based sampler, situation is slightly different. We only have an opinion on the type of marginals. The estimation of the marginals, used for the uniformization und subsequent copula estimation, is carried out after the approximation (in step 2 of the sampling scheme). Hence, the marginals we will estimate from the (normally distributed) approximated posterior are those distribution that explain best the given marginal of this posterior, i.e. a normal distribution, or a truncated normal distribution in case the priors are

not defined in all of $\mathbb{R}$. This corresponds to fitting, if we assume that the prior is student distributed, a student distribution to normally distributed data. For the case of a log-normal prior we would fit the marginal to data from a truncated normal distribution. This allows for more flexibility than for a fixed set of marginals.

Coming back to the fitting properties of the laplace approximation, we note that for multimodal distribution it might not conserve much probability weight of the original distribution. This however, can be limited to some extend by localizing local minima and perform laplace approximations for each of the modes. Those laplace approximation can be merged into a gaussian mixture model weighted with the value of the posterior at the mode. We will call this approximation *gaussian mixture laplace approximation*. Futhermore, one encounters some computational obstacles in the process of the implementation: For the application of the original, unimodal Laplace Approxmation one has to note that the posteriors of applications are not necessarily unimodal, i.e. one has to ensure that the maximum one uses in the approximation is the global one and not only a local one. For unknown posteriors this might turn out to be computationally expensive. In the implementation this is carried out via latin hyper cube sampling. In terms of efficiency this question would highly benefit from a parallel implementation (e.g. by means of Nvidias CUDA-architecture). Esspecially when minimizing the loss of information in the approximation process by using the multimodal gaussian mixture laplace approximation, one has to calculate the Hessian at the respective modes. Esspecially in higher dimensions this task is computationally challenging. In the implementation we use 'Adaptive Robust Numerical Differentiation'. A pack of MATLAB-functions made available by John D'Errico. MATLABs own Hessian-computations provided by *fmincon* and *fminunc* do not show satisfactory results. One also has to note that the posterior has to be sufficiently regular to calculate the Hessian, i.e. a C^2-function.

On the other hand it is very easy to generate samples from this normally distributed approximation (see, e.g., [McNeil, Frey, Embrechts]) or the respective gaussian mixture model that one can use for the estimation of the copula. Under very restrictive assumptions on both, joint distribution and marginals, one can also write down the pair-copula decomposition directly, as noted in the section on pair-copula decomposition (see also [Aas et al], Section 2.6).

2.6.2 Variational Inference

In the introduction to this chapter it was claimed that copula structures are incompatible with Variational Methods: The variation argument relies on the assumption that the joint distribution factors over a subset of the marginal distributions (see, e.g., [Lawrence et al])

$$f(x) = \prod_{i \in I} f_i(x_i) \tag{12}$$

and minimizes, e.g. the *Kullback-Leibler divergence* (a distance measure between distributions - see, e.g., [Lawrence et al]) of the object above to the true posterior by an argument based on methods from the *calculus of variation*. Since the Bayesian framework and the copula estimation procedure rely on prior information and a uniformization in every component, the only option is to use the factorization

$$\prod_{i=1}^{n} f_i(x_i) \ , \tag{13}$$

where we can assume that f_i is in the class of the priors. Although it is very simple to sample from this distribution, from a probability perspective this factorization corresponds exactly to the definition of independence. If we would continue from step 2.) in the copula estimation procedure we would, naturally, recover the independence copula, i.e. a complete negligation of the dependence of the parameters. To challenge the assumption on the decomposition makes the corresponding calculus of variation problem inapproachable: Given a variational problem of the type

$$p(x) = \inf_{c \in \mathbb{C}op, f_i} c(x) \prod_{i=1}^{n} f_i(x_i) \tag{14}$$

one has to define a distance measure on the copula space $\mathbb{C}op$ (i.e. the set of objects whose existence one gets by Sklar's theorem projected on the span of the copulas that are representable via vine copulas) that is either compatible with the distance measure one uses to optimize the f_i's (in case of a simultaneous optimization) or orthogonal to the optimzation procedure over the f_i's (in case of an iterative scheme). Considering one has knowledge on the posterior $p(\cdot)$ and knowledge on the type of the marginals f_i, one could argue that one can find a set of 'suitable' copulas (by Sklar's theorem one for every possible parametrization of the marginals). Now one could try to find the best combination of the marginals and the copula explaining the posterior. If there is an analytical method to approach this problem we consider it beyond the scope of this thesis.

[Lawrence et al] note that there are situations where variational inference *'leads to poor approximations because it does not capture important statistical relationships between parameters in the posterior distribution.'*.

2.7 Profile Likelihoods

The last two subsections showed that one cannot expect a deterministic posterior approximation to work very well. One therefore can ask whether one can synthesize a manageable but representative amount of data from the posterior based on direct evaluations of the posterior and perform the copula estimation based on those samples. The first guess here is to set up a set of hyperplanes and use the intersection of the planes as points in the grid (an option dicussed later). However, since the pratical problems we will be dealing with are highdimensional this approach is unfeasible and we instead have to rely on a method

taking into account the shape of the posterior in all of space and giving out a limited amount of representative data. To this end we introduce profile likelihoods, a tool initially designed to answer questions concerning the identifiability of parameters (see [Raue et al]).

Definition Given a parameter space $\Omega \subset \mathbb{R}^d$ and a likelihood function $L : \Omega \to \mathbb{R}$ the *profile likelihood function* $L^i_{PL} : \mathbb{R} \supset A \to \mathbb{R}$ for the parameter $\theta_i \in A$ is defined as

$$L^i_{PL}(\theta_i) = \sup_{\theta \in \Omega_{\theta_i}} \{L(\theta)\} \ ,$$

where $\Omega_{\theta_i} \subset \Omega$ denotes the set

$$\{\theta = (\theta_1, ..., \theta_n) \in \Omega : \theta_i \text{ fixed } \}.$$

For more information see [Murphy, van der Vaart] and [Venzon, Moolgavkar]. Figure 2 depicts information on profile likelihoods for a two dimensional normal distribution with correlation parameter 0.7.

Since our goal is to infer a dependence structure out of the samples we draw from this profile likelihood we have to ask if the shape of the profile likelihood is representative for the underlying copula. Figure 3 shows that this can be a problematic question: Given the same profile likelihoods, the underlying distributions can have different shapes, i.e. there is variety of relevant information in the posterior that is not represented in the profile likelihood. The picture on the left shows the niveau sets and profile likelihoods of a two-dimensional normal distribution. For the normal distribution the profile likelihoods follow the principal axis and if we consider a singular niveau set the points in the profile likelihood are given by the vertices of the ellipse (as depicted in the plot on the bottom right of the figure on normal distributions). If we consider the normal distribution and assume standard normally distributed marginals we know from the section on pair copulas that the corresponding dependence structure for this model is given by a gaussian copula. Let us focus on one niveau set. For an uncorrelated (i.e. independent) joint normal distribution, the points of the niveau set that are in the profile likelihood are the ones where the ellipse intersects with the axis. If we fix those points and consider the definition of the profile likelihood we see that we can change the niveau-set within the boundaries of the hyperplanes running through the vertices of the ellipse. For a fixed set of given marginals this correponds to very different choices of copulas having, e.g., different symmetry properties (Figure 3). As long as we do this same kind of change for every niveauset without changing the area of the niveauset we know by that we still have a probability distribution (Cavalierie's principle). However, the respective copulas of all the three examples have to be different and are not elliptic anymore, i.e. belong to a different copula family. Although the profile likelihoods will turn out to be very useful in the sampling process this argument shows that one cannot expect them to work for every example.

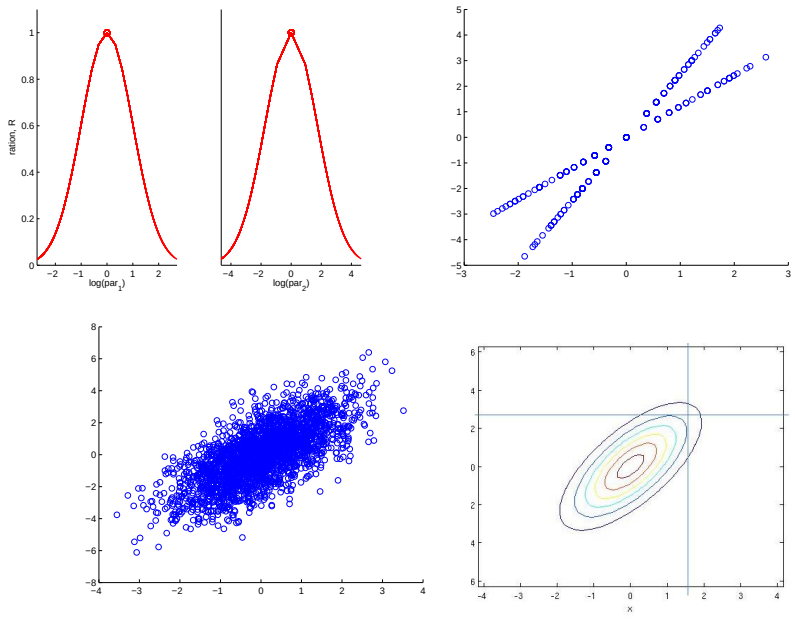


Figure 2: **Profile likelihoods for a two dimensional normal distribution with correlation of 0.7** On the top left the profiles are shown. The figure on the top right shows the location of the profiles. The figure at the bottom right shows the niveau sets of the normal distribution. The horizontal and vertical line represent a fixed value for either x or y. Along this line the point with the highest value, i.e. the point in the profile likelihood, is the single intersection between the niveau set and the line.

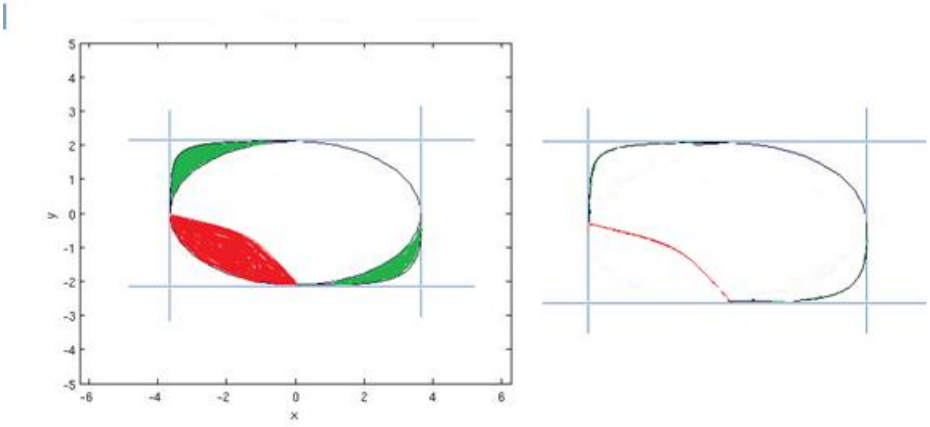


Figure 3: **Variability of the posterior for fixed profile likelihoods** We considered a niveau set and asked how much of the weight in this set is fixed by the location of the points in the profile likelihood. The figure depicts one possibility to 'push around' the weight in the distribution without changing the profile likelihoods. The green areas (top left and bottom right) are added whereas the red area (bottom left) is taken away to reach a distribution with the same profile likelihoods but different symmetry properties. Since the marginals are fixed ex ante this would correspond to a different copula by Sklar's theorem. The corners were chosen arbitrarily. Hence, we can construct a variety of distributions with the same profile likelihoods but different dependence structures.

3 The examples

This section is devoted to a brief presentation of the models for which the efficiency analysis of the sampler is carried out later

3.1 A highly correlated normal distribution

In the first example we draw samples from a 3-dimensional normal distribution $\mathcal{N}(0, \Sigma)$ with covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.95 * \sqrt{3} & 0 \\ 0.95 * \sqrt{3} & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix} .$$

This example is chosen to illustrate that the Laplace Approximation works and can give good results. The example has been chosen to be three dimensional to fulfill the input assumptions of the copula estimation procedure. The marginals for this model are assumed to be normally distributed. In addition, we note that the high correlation was chosen to show difficulties for profile likelihoods applied to normal distributions.

3.2 Banana shaped distribution

This problem has been chosen to illustrate that we can not expect that the approximation methods we use can be applied to every kind of distribution. In fact, we will see that the copula estimation procedure does not cover the dependence between the parameters of this distribution very well and therefore the acceptance rates are relatively low, considering that it is a two dimensional example. In addition, we immediately see that the likelihoods introduced before are not well defined for this example. For every negative for x there are two points (x, y_1) and (x, y_2) with $p(x, y_1) = p(x, y_2) = L^1_{PL}(x)$.

All those problems are caused by the symmetry properties of this distribution. Instead of being symmetric with respect to a point, it is symmetric with respect to an axis. This already excludes the well-understood elliptical copulas. In addition, the weight is mainly distributed in the vicinity of a parabola; a shape which seems to be difficult to capture by the pair copula decomposition. In other words, we observed that the dependence structure for this model can, in general, be not covered very well by the vine copula decomposition. If we assume this hypothesis we can expect that the samplers based on an approximation or grid method perform reasonably well against the prerun based sampler because the copula estimation is not able to utilize the (presumably) better information in the prerun. We encountered this distribution in [Girolami, Calderhead]. The marginals for this model are assumed to be normally distributed.

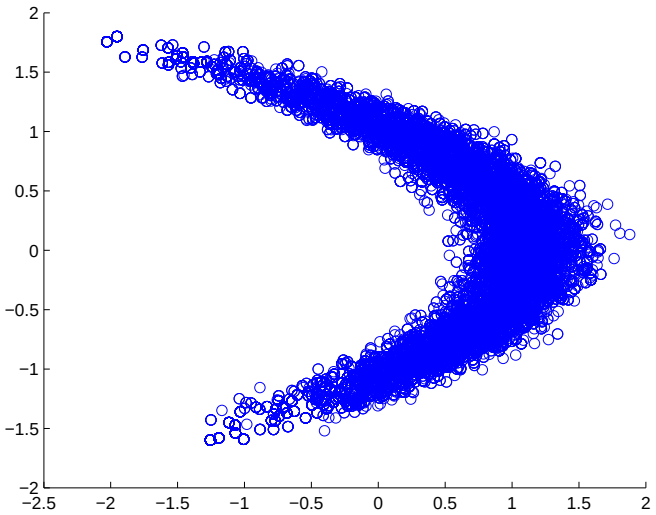


Figure 4: **Scatterplot for a prerunbased copula MCMC run of length 25000 for the bananashaped distribution**
We note that for most of the runs we did not see a symmetric exploration of the upper and the lower tail of this distribution. This could be one of the reasons that the copula for this model is estimated rather badly and only yields acceptance rates between 0.2 and 0.3.

3.3 A biokinetic model for the processing of Zirconium in the Human Body

This subsection introduces a biologically motivated dynamical system. We start with a brief discussion of biokinetic (or compartment) models and give an illustration for parameter estimation for such systems. In addition we introduce a model for the processing of zirconium in the human body. This section summarizes several ideas presented in [Schmidl] and [Schmidl, Hug et al].

A compartment model is a collection of mutually exclusive compartments together with a set of transition equations. In the context of biology the compartments can, e.g., correspond to different major parts of an eco system or, as in the model we will be interested in, major organs and tissues in the human body. Compartment models are, e.g., used to describe the density of a chemical element within different organs. The transition equations determine the relations between the compartments and are frequently modelled by Ordinary Differential Equations. Hence, compartment models can be seen as dynamical systems. We will assume that the compartments are homogeneous, i.e. that the density of ,e.g. the chemical element, is constant within a compartment. In addition it is assumed that the volume of the compartment is constant over time so that one can identify the amount of the chemical element in the compartment by its density.

In the context of radiation protection biokinetic models are frequently used to model the processing of radioactive substances with the objective of, e.g., providing limiting values for detrimental effects. The application of the posterior approximations and grid methods will be carried out for a model build to understand the processing of Zirconium in the human body. Radioactive Zirconium isotopes are produced in large amounts in nuclear fission reactors. The model discussed here has been introduced in [Greiter et al] based on human measurement data. In [Schmidl, Hug et al] the model was compared by model selection procedures to a different compartment model that is used by the International Commission on Radiological Protection and is based on animal data. The model build upon human measurement data proved superior. As indicated, in those models the human body, or better, its major organs and tissues are viewed as a set of different compartments representing kinetically homogeneous amounts of radionuclides. The connections between the compartments are described by transfer rates which in itself are governed by the law of mass balance. Mathematically, the dynamics of the model is described by a system of linear, coupled first-order Ordinary Differential Equations one can easily deduct from the connections of the compartments (see Figure 5)

$$\frac{d}{dt}y_j = \sum_{\alpha \in \mathcal{A}_{y_j}^{+}} x_\alpha y_{[x_\alpha]}(t) - \sum_{\beta \in \mathcal{A}_{y_j}^{-}} x_\beta y_j(t) \,, \tag{15}$$

where $\mathcal{A}_{y_j}^{+}$ denote the indices of rates flowing into y_j and $\mathcal{A}_{y_j}^{-}$ denotes the indices flowing out of y_j. If we, e.g., consider the bone-compartment y_2, we see that we only have connections to the transfer compartment y_7 (x_1 incoming and x_{11}

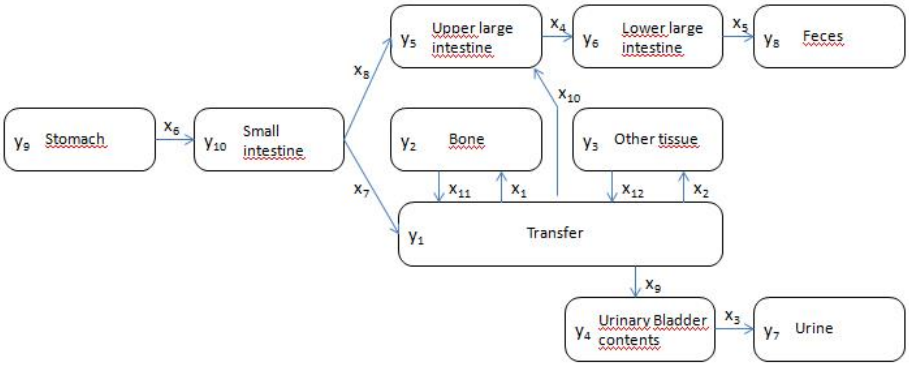


Figure 5: **Model for the biokinetic processing of zirconium** The model consists of 10 compartments $y_1, ..., y_{10}$ and twelve transfer rates $x_1, ..., x_{12}$. The zirconium enters the body in the stomach compartment y_9. The transfer compartment y_1 and the urine compartment y_7 have been measured directly during the study.

outgoing) and the corresponding differential equation for y_2 is given by

$$\frac{d}{dt}y_2 = x_1 y_1(t) - x_{11} y_2(t) \ .$$

The model parameters here are the rates of the transition equations $x_1, ..., x_{12}$. They contain a rich variety of non-standard dependence structures (see supplementary material to [Schmidl, Hug et al]). In contrast to the other examples, where the standardized dependence structures we can estimate after the approximation might still contain enough information of the original one, this example is chosen to assess what happens for a model with a rich dependence structure. The priors for this model are lognormal distributions for the rates $x_1, x_2, x_7, ..., x_{12}$ and triangle distributions for the rates $x_3, ..., x_6$. We are only interested in the type of the prior to set up the copula estimation. The full prior information can be found in the supplementary material to [Schmidl, Hug et al].

Parameter estimation in biological systems

Here we want to give a brief account of how to estimate the rate constants in the model introduced above by means of frequentistic approaches. Given the transition equations

$$\frac{d}{dt}y_j = \sum_{\alpha \in \mathcal{A}^+_{y_j}} x_\alpha y_{[x_\alpha]}(t) - \sum_{\beta \in \mathcal{A}^-_{y_j}} x_\beta y_j(t) \ ,$$

we are interested in the inference of the vector $x = (x_1, ..., x_12)$. This vector holds the parameters defining the system of equations for $(y_j)_{j=1,...,10}$. In addition it is assumed that the whole radioactive nuclides enter the body over the stomach-compartment y_9. The goal is to infer the parameters in x from a set of given observations $\{z_1, ..., z_m\}$, where $z_i = (z_{1,i}, ..., z_{l_i,i})$ corresponds to the data measured at time $t_i \in [0, T]$. Since we can assume that the data contains measurement errors we model the measurement as

$$z_{i,j} = h_x^{i,j}(y(t_i)) + \epsilon_{i,j} \ ,$$

where $\epsilon_{i,j}$ are independent normally distributed random variables. The measurements for this model have been carried out in a stable tracer study executed at the Helmholtz Zentrum. 12 healthy humans were ingested an investigation-specific amount of isotopically enriched stable zirconium with subsequent measurements and analysis of blood plasma and urine (see [Greiter et al 1] and [Greiter et al 2] for the details).

An estimator for x is now, e.g., given by minimizing the squared loss function

$$\chi^2(x) = \sum_{i=1}^{m} \sum_{j=1}^{l_i} \frac{\left(z_{i,j} - h_x^{i,j}(y(t_i))\right)^2}{\sigma_{i,j}^2}$$

over x. For the numerical background to solving this ODE system see [Schmidl, Hug et al - supplementary material].

3.4 A small compartment model

Since we will see that we encounter a variety of computational problems with the biokinetic model introduced above, we carry out our analysis with a very simplified version of it to reduce the complexity. However, we assume that the dependence between the parameters still remains 'non-standard' to some extend. This will allow us to assess whether the approximation can capture the dependence reasonably well.

The small compartment model is described via the following set of coupled first-linear ODEs

$$\frac{dc_1(t)}{dt} = -k_2 c_1(t) - k_3 c_1(t)$$
$$\frac{dc_2(t)}{dt} = k_2 c_1(t) - k_1 c_2(t)$$

Figure 2 visualizes this system.

This toy model has been used in the comparison of classical samplers with the copula-based Metropolis-Hastings algorithm in [Schmidl, Czado, Hug, Theis]. It assumes that the 'small intestine' compartment is unobservable. The data which is used to infer the rates is given by the concentration in the 'transfer compartment' at eleven time points $t_i = i \cdot 0.1, \ i = 0, ..., 10$ perturbed by a standard normally distributed error, i.e.

$$y_i = c_2(t_i) + \epsilon_i \ ,$$

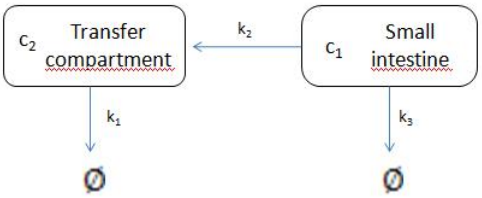


Figure 6: **Small compartment model** The model consists of two compartments c_1 and c_2. The Zirconium enters the system via the compartment c_1 and leaves the system either directly or via the transfer compartment c_2. The measurements are given by the concentrations in the transfer compartment c_2 perturbed with a normally distributed measurement error.

where $\epsilon_i \approx \mathcal{N}(0, 1)$. The priors for this model were set to $k_1, k_2 \approx \mathcal{N}_{[0,1000]}(1, 1)$ and $k_3 \approx \mathcal{N}_{[0,1000]}(20, 400)$, where $\mathcal{N}_{[0,1000]}$ denotes the $[0, 1000]$-truncated normal distribution.

Inference of dynamical systems

The two dynamical systems above have been developed to understand how high the concentration of a radioactive substance in a given organ is. Given that the model parameters of this system are the rate constants, we have to draw a connection between them and the actual concentration. To find the concentration within a single compartment, we have to find the solution to the corresponding ODE-system given by the transition equations and an estimator for the rates. We saw that we can construct a classical estimator via minimizing the loss function. A different approach is to sample from the joint distribution of $x_1, ..., x_{12}$ and use the samples to solve the ODE system. In figure 7 we show a discretization of the corresponding *posterior median solutions*, i.e. the function given by the pointwise median of the solutions (see [Schmidl, Czado, Hug, Theis] for more details), for the two models based on rate samples from the laplace approximation-based and the profile likelihood-based sampler. Note that the posterior median solution is not a solution to the equation system since it is a pointwise median value.

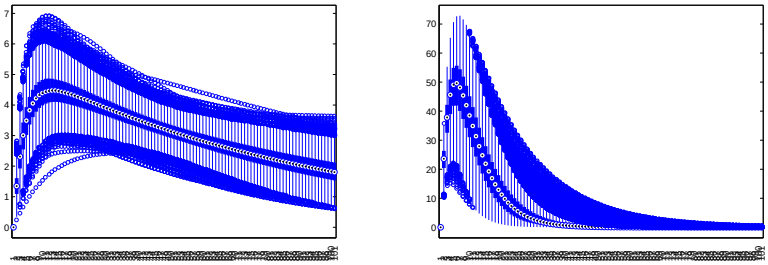


Figure 7: **Posterior Median solutions for the small compartment and the zirconium model** The figures depict a pointwise boxplot of the solutions to one of the equations for the small compartment model and the zirconium model based on samples from the laplace approximation based sampler for the compartment model and for samples from the adaptive profile likelihood based sampler for the zirconium model. The x axis depicts the time from 0 to 1, whereas the y axis depicts the concentration of the compartment c_2 (small intestine) for the small compartment model (left) and the compartment y_{10} (small intestine) for the zirconium model (right). The black line in the middle corresponds to the posterior median solution, whereas the white shaded part around corresponds to the 25 and 75 percent percentiles. Note that the posterior median solutions is not a solution to the equation system! The shape of the posterior median solution for the small compartment model in [Schmidl, Czado, Hug, Theis] is nicely recovered.

4 Results - Laplace Approximation

The following two sections are devoted to the comparison of the different samplers. To this end, we first have to find criteria that assess the quality of an MCMC sampler. Since we are interested in a question of efficiency we have to assess how much independent samples we get per time interval. For MCMC samplers this means that we also have to take into account the dependence (here measured via the autocorrelation) within the Markov Chain. This leaves us with the criterion of 'Effective Sampling Size per second' indicating that one discards 'ineffective', i.e. correlated, samples. The following definitions are made to get to the desired benchmark

> The *estimated effective sampling size (ESS)* is given by the fraction of samples in the Markov Chain and the highest autocorrelation within its components. It is defined as
>
> $$\text{ESS} := \max_{i\in\{1,\dots,d\}} \left\{ \frac{K+1}{1+2\sum_{\tau=1}^{K_c^i}\left(1-\frac{\tau}{K+1}\right)\hat{\rho}_i(\tau)} \right\}$$
>
> for the MCMC sampling length K and the *estimated autocorrelation functions*
>
> $$\hat{\rho}_i = \frac{1}{\hat{\sigma}^2(K+1-\tau)}\sum_{j=\tau}^{K}(X_i^{(j)}-\hat{\mu}_i)(X_i^{j-\tau}-\hat{\mu}_i)$$
>
> of lag τ and dimension i of a Markov Chain X_t, where K is chosen so that $K = \text{argmin}_\tau \hat{\rho}(\tau) < 0.05$.
>
> The run times naturally include the times of the prerun or data synthesis.

We end up with the following benchmarks

> For $x \in \{\{ds\},\{ls\},\{ps\},\{as\}\}$ we define the *effective sampling size per second* for the data-based sampler (ds), laplace-approximation based sampler (ls), prerun-based sampler (ps) and adaptive sampler (as)
>
> $$ESSps_x := \frac{ESS_x}{t_x}\ .$$

This gives us a benchmark that measures the samplers against each other.

In addition, we need a benchmark to measure the 'quality of information' contained in the samples from the preperation steps. To this end, we propose to use the acceptance rate of the corresponding copula sampler. For an independence sampler the acceptance rate is close to one for a proposal function that coincides with the posterior. Hence we can see the difference in the acceptance rate for a proposal function as a distance measure and a benchmark for the loss of information with respect to the true posterior. One could argue to start the samplers from 'equal ground', i.e. use either the time to acchieve a certain

'quality of information' as a benchmark (by, e.g., cutting down the prerun so that the acceptance rate of the approximation based sampler and the prerun sampler are similar) or use the information extracted in a given time frame. We feel that both results will favor the approximation based sampler. The copula sampler has been introduced to outperform other samplers based on more refined information. Therefore we will try to acchieve the best possible results for both samplers and compare those, i.e. tune the parameters for optimal results.

Note that this procedure is targeted at the question of 'Given 'optimal' parameters, which sampler is more efficient?', not the question to propose a more efficient sampler for an unknown problem. For an answer to this problem we would have to have access to procedures giving us the optimal parameters ex ante.

In addition, note that the initialization for different seeds might turn out to be significantly different. Since we use a variety of 'toy models' whose observations are based on pseudo-random numbers already the initialization of the posterior is highly influenced by the underlying random numbers. To illustrate this fact, we show the Maximum Posterior Estimates, located by latin hyper cube sampling, of the compartment model for 5 different initializations of the posterior. The true parameters of $(1, 1, 20)$ are siginificantly perturbed. In the sampling process we fix the seeds of the random number generator so that we will work with a fixed posterior function for one example. For the compartment model our seeds will initialize the posterior with the Maximum Posterior Estimate $(1.0766, 1.26066, 22.8445)$.

Initialization	x_1	x_2	x_3
1	1.0766	1.2606	22.8445
2	1.2015	0.8611	13.6689
3	1.3855	1.0212	14.5556
4	1.0558	1.3366	22.3901
5	1.4358	0.9824	16.2223

We also note that for the adaptive version the information cut out by the use of the approximation is recovered to some extend by the copula-based prerun. Hence we can assume that the acceptance rate of the adaptive approximation based sampler is better than the one based on the approximation alone. However, we also note that the approximation cuts down the information. Hence we can also assume that the convergence to the equilibrium distribution takes longer for the approximation based samplers than for the prerun based sampler.

For the different approximations we always show 4 figures presenting the ESSps, the acceptance rate, the thinning values and the overall computational time of the different samplers. The computational times correspond to the run time on either, one node of a 24-core 1.8 GHz Intel Xeon cluster, or one node of 24-core 2.27 GHz Intel Xeon cluster, or one node of a 4-core 2.9 Ghz Intel i5 Workstation depending on available ressources. Occasionally the runs had to be stopped and continued later. We put attention to run one comparison on the same machine.

Convergence

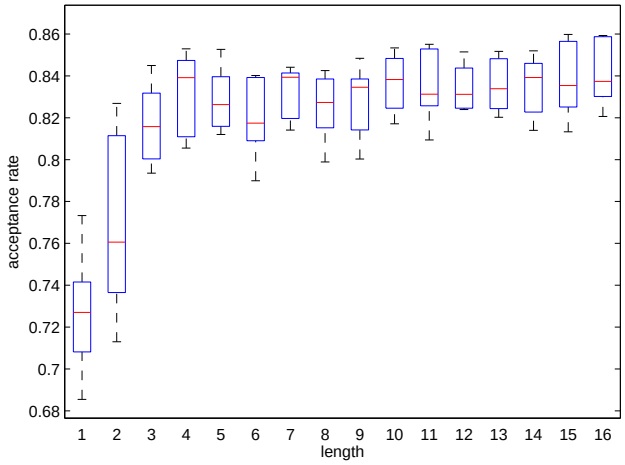
In this section we want to adress the question of 'How much samples are needed to represent the approximated posterior?'. We experienced that one does not need many samples from the approximated posterior to recover the information left after the approximation and, by default, use 100 samples to keep the time for the copula estimation low. This seems to be a very low number. However, even for the high dimensional models we experienced no increase in the acceptance rate by using more samples.

Prerun

Here we adress the question of how long a prerun should be. This question is of high importance since the estimation of the copula is based on this prerun and can take extremely long if the prerun is to long. Hence we are interested in a limited number of samples containing as much information on the dependence structure as possible. Below we show how the acceptance rate behaves with an increasing size of samples from the prerun. We assume that a burn-in of 2000 samples for the low dimensional examples is enough to limit the information on the starting distribution. The figures below show the acceptance rate of the copula based sampler for an increasing number of samples. Length 1 corresponds to the samples 2000 to 2500. Length 2 uses the samples 2000 to 3000. We use 5 different preruns of length 10000 and evaluate the acceptance rates.

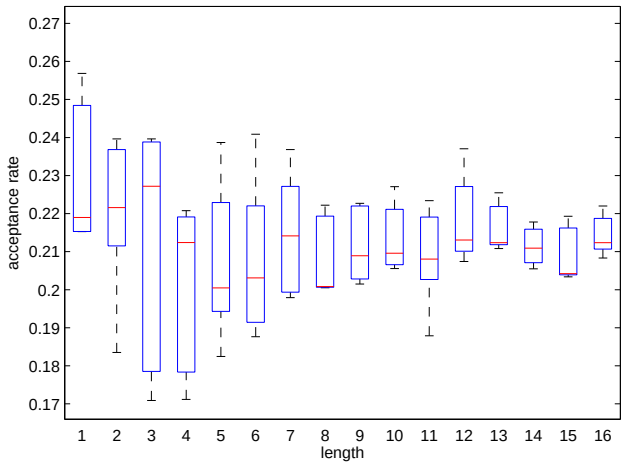
Normal distribution

The figure below depicts the acceptance rate of the copula based sampler based on the prerun samples 2000 to 2500 (length 1), 2000 to 3000 (length 2), ..., 2000 to 10000 (length 16). We see that the acceptance rate reaches the level of 0.8 relatively fast and stabilizes at approximately 0.84. We will use a prerunsize of 2500, i.e. the samples 2000 to 4500, to reach a compromise between the stability of the results and a short prerun.



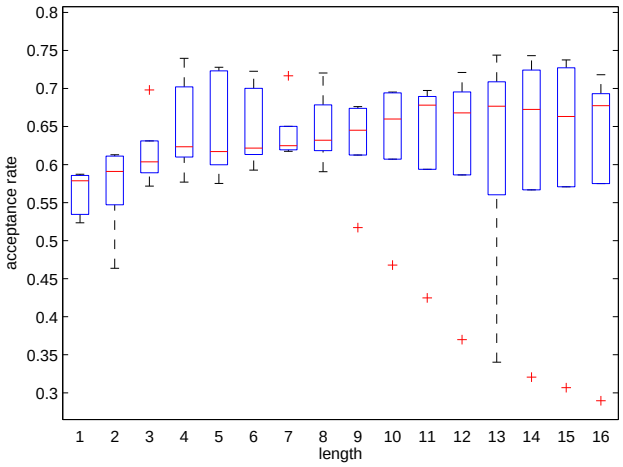
Bananashaped distribution

In the figure below we see that the accpetance rate for the bananashaped distribution stables approximately at the level of 0.21. We will use the prerun samples 2000 to 7000.

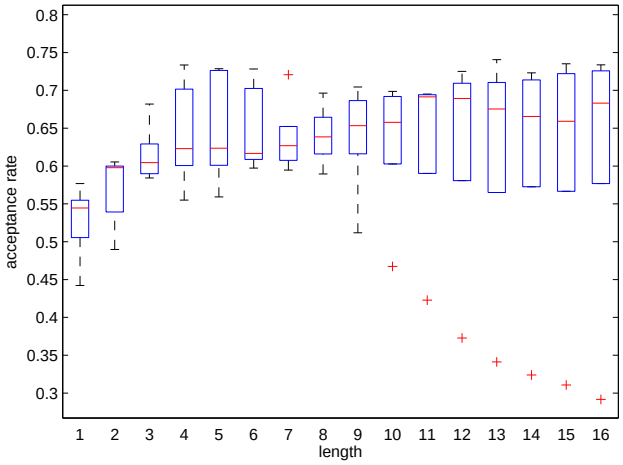


Small compartment model

Based on the results depicted in the figure below we use the prerun samples 2000 to 5000.



We note that one of the runs showed an 'odd' behaviour, a decreasing acceptance rate with increasing samples used to infer the vine copula. Reproducing the results showed a similar behaviour depicted in the figure below. The reason is that the sampler got stuck in one location.



Zirconium model

For this model we chose to use a prerun of 15000 samples and discard 5000 'burn-in' samples. Shorter preruns made the copula estimation very unstable. Longer preruns, on the other hand, took very much time and the results in terms of acceptance rate and ESSps were not better.

4.1 Results - Laplace Approximation

We compare

1.) The original prerun based sampler. In the figures below the corresponding sampler is denoted by 'MCMC prerun'.

2.) The sampler based on Laplace Approximation and subsequent sampling of 'synthetical data' and copula estimation. In the figures this sampler is abbreviated by 'data' and referred to as approximation-based sampler.

3.) The sampler based on 2.) as a prerun (adaptive version). In the figures below this sampler is abbreviated as 'data-CIMH'.

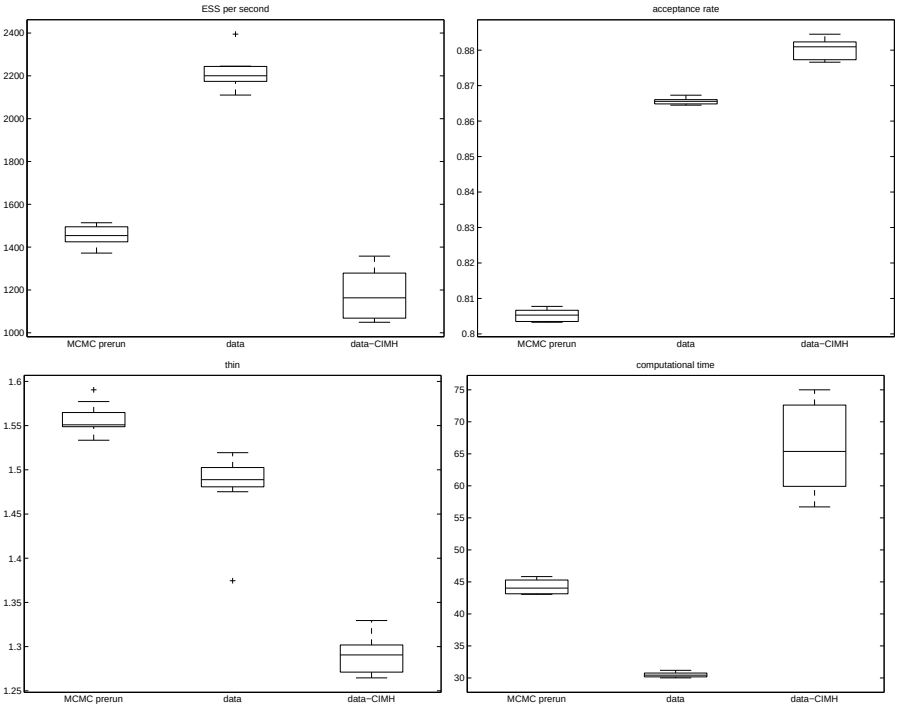
for the examples introduced in Chapter 4. If not states otherwise, the figures are based on data from 10 sampling runs of 50000 samples each. The autocorrelations are calculated from the last 40000 samples of those runs by comparison to the estimated mean and variance from a long (500000 samples) Random Walk Metropolis-Hastings run made before starting the copula runs. The fraction of random walk and heavy tailed samples are both set to 0.02. For the bananashaped distribution we can not include random walk samples due to an incompatibility with the vine structure. Hence, the respective parameter is set

to 0. For the heavy tailed distribution we used the default, i.e., a uniform distribution on $[0, 10^9]^n$, where n is the dimension of the respective sampling space and changed it for the examples with support in $\mathbb{R}^n$. For an implementation a uniform distribution on this domain seems feasible.

Note that, although we set seeds for the sake of reproducibility, the results for the autocorrelation and acceptance rates for the prerun based sampler in different sections are not expected to be equal due to different amounts of random numbers used by the approximations. The posteriors are identical.

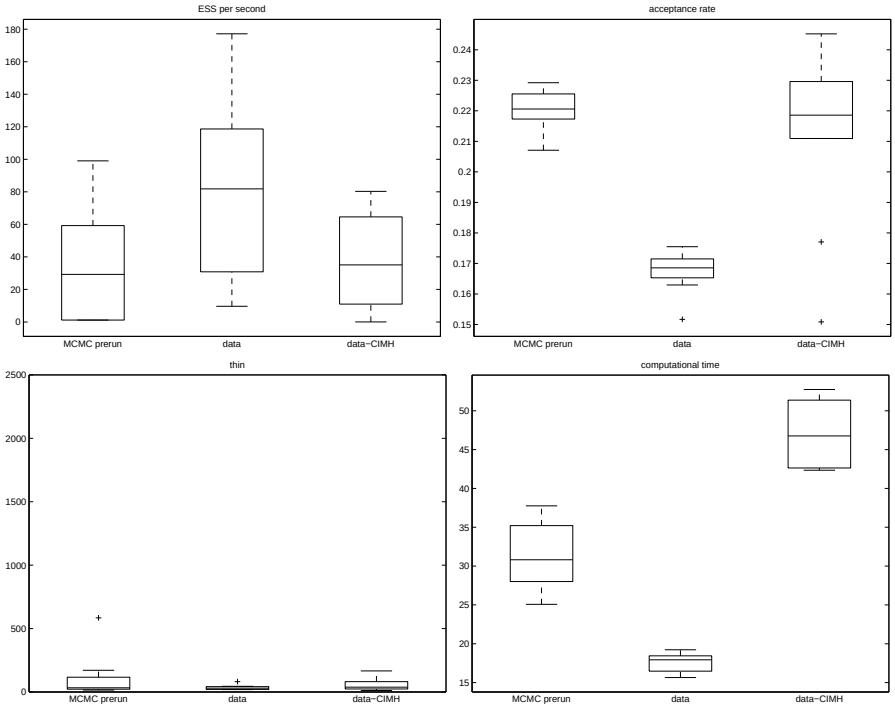
4.1.1 Normal distribution

Naturally, we expect that we can recover a normal distribution when applying the Laplace Approximation. Indeed, the Covariance structure is covered perfectly (up to the fifth post comma digit for every entry of covariance matrix). The time for the latin hyper cube sampling, laplace approximation and sampling took 0.82 seconds, compared to the 2.65 seconds it took to sample the prerun. Since the prerun also consists of more samples the copula estimation takes longer. In addition, the autocorellation in the chain is slightly lower for the data-based sampler (1.56 for the prerun based sampler compared to 1.48 for the approximation-based sampler, and 1.28 for the adaptive version). In terms of ESSps the approximation based sampler outperforms its contenders.

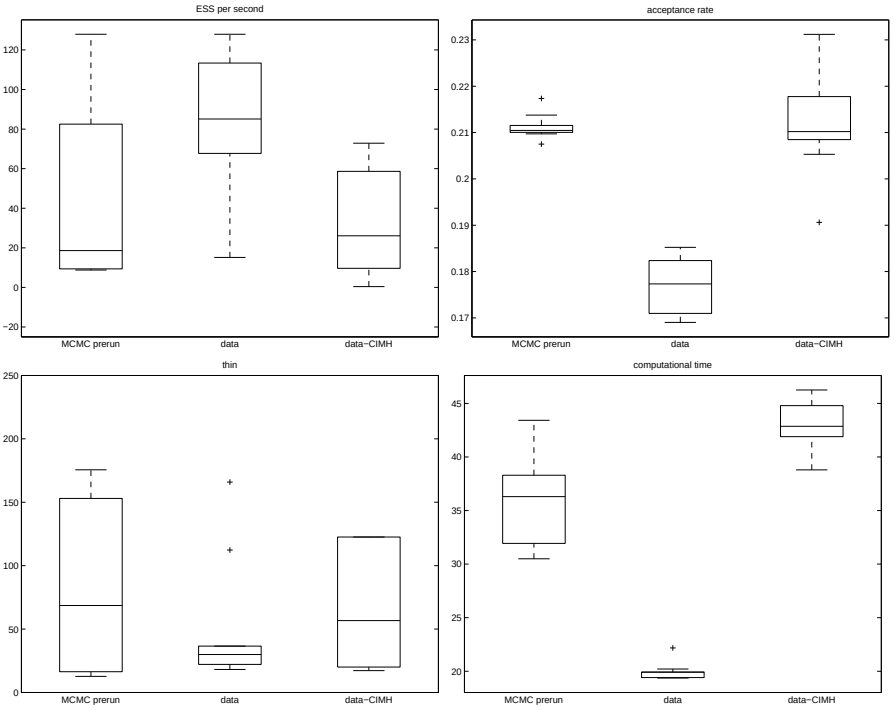


4.1.2 Banana shaped distribution

In this example the time for the prerun was shorter than the time for the laplace approximation (2.26 seconds compared to 4.46 seconds). The reason is that the latin hyper cube sampling that locates the minima of the posterior got stuck in one of its runs. The thinning values contain outliers for the runs of the adaptive approximation based sampler. For the prerun-based sampler the median for the number of samples to be discarded in order to acchieve independence is at 58. For the approximation based sampler the values are at 35 for the non-adaptive and 30 for the adpative approximation based sampler. Note, that the adaptive sampler had an outlier at 73000. In terms of ESSps the approximation based samplers outperform the prerun based sampler.

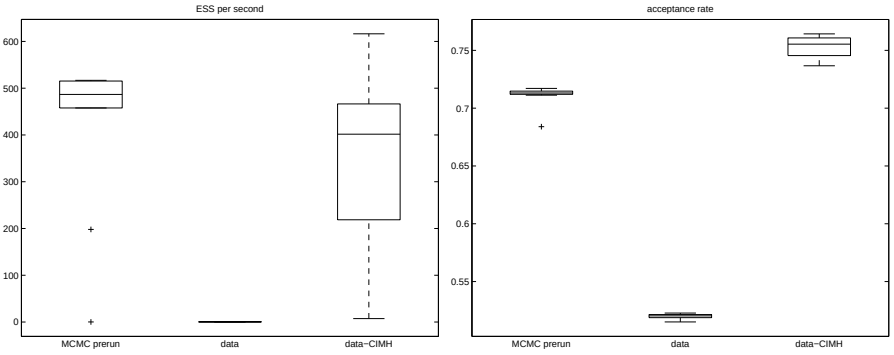


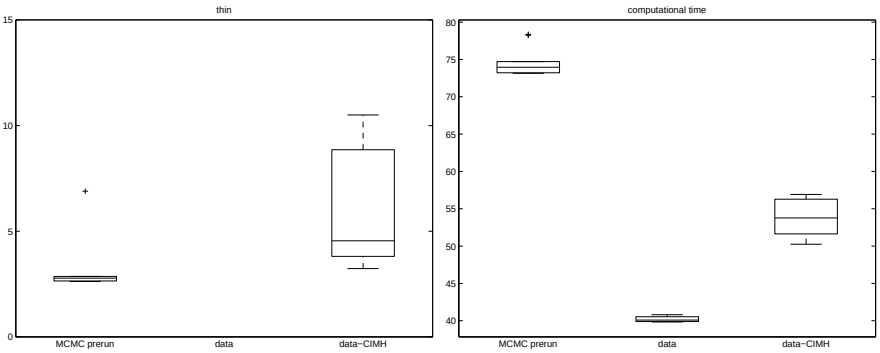
Compared to the other results for the autocorrelation of the prerun based sampler the median thinning value of 58 is higher. In addition there were several thinning values higher than 1000. Below we show another run based on a prerun of 10000 samples of which the samples 2000 to 10000 were used for the copula estimation. We again encounter outliers in the thinning values for the adpative approximation based sampler. For the prerun based sampler 68 samples had to be discarded in order to acchieve one independent sample from the posterior. We also included this example to illustrate that the longer prerun raised the time for the estimation of the copula by 5 seconds and the acceptance rate was, in fact, slightly lower than for the first run.



4.1.3 Compartment model

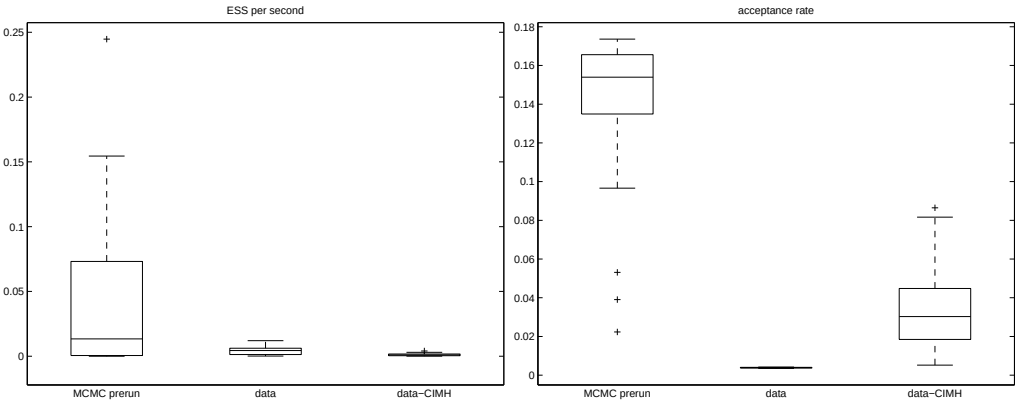
We oberserve a very high autocorrelation in the chain of the approximation based sampler (median thinning value of 6932 - not shown). However, for the adaptive sampler the autocorrelation is only slightly higher than for the prerun based sampler (2.77 compared to 4.54). Rerunning the sampler showed similar results. In addition to a higher acceptance rate, the adaptive approximation-based sampler also outperforms the prerun-based sampler in terms of computational time. In terms of ESSps, however, the prerun-based sampler leads with a median value of 494 to 413.

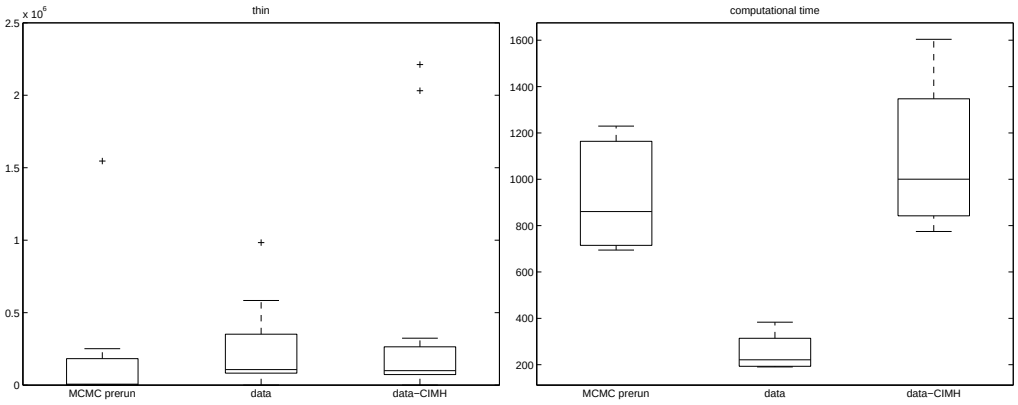




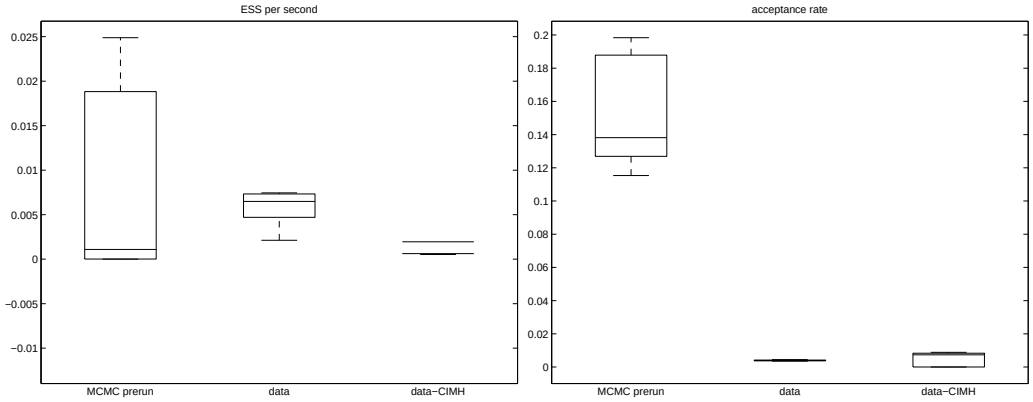
4.1.4 Zirconium model

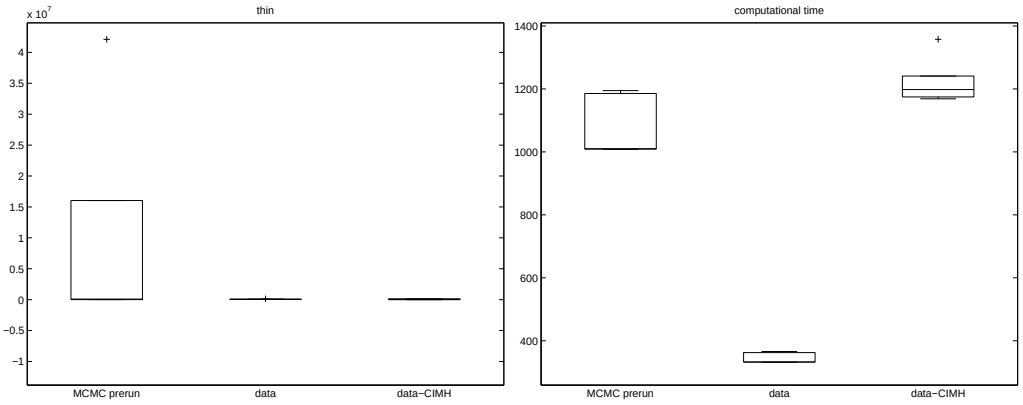
While simulating this example one encounters the problem of calculating a Hessian for a high-dimensional problem. In addition this model has triangle priors, i.e. from a mathematical perspective it is not well-defined to calculate the Hessian since the posterior is not a global C^2-function. Indeed, the Latin Hyper Cube sampling with a subsequent application of MATLABs *fminsearch* always gave back a value for which one of the triangle priors were on its maximum. I.e. technically we were not allowed to take the Hessian at the mode. However, in some cases one was still able to calculate a symmetric and negatively definite Hessian at the mode. To assess whether the Laplace Approximation might still be useful with this violated assumption we continued with sample generation from the approximated posterior and subsequent copula estimation. For the laplace approximation we encountered unfeasibly high thinning values and, for the approximation based sampler, very low acceptance rates of the order 0.001. We note however, that the prerun for the adaptive sampler nicely recovered some information and showed siginificantly better acceptance rates (median 0.03).





A possible explanation for the latin hyper cube sampling to stop at the vertex of the triangle is that the minimzation procedure itself is based on a derivative based method and gets stuck at this location. One could argue in favor of a derivative free method to find a minimum. We implemented a scheme starting from latin hypercube sampling followed up by simulated annealing (see [Kirkpatrick et al], a derivate free minimzation procedure) with a subsequent construction of a gaussian mixture model. From this model we drew 500 samples and proceded with the copula estimation and subsequent sampling. However, since the Hessians for this model are badly conditioned and normal distributions have support in all of $\mathbb{R}$ we discarded the samples incompatible with the priors, i.e. negative samples for the log-normal priors and samples not contained in the priors of the triangle. The results for this implementation are not satisfactory either. In 2 of the 5 runs the copula estimation for the adaptive approximation based sampler broke down. For the remaining 3 runs we note that, in contrast to the run based on the derivatebased minimzation and a subsequent Laplace Approximation (i.e. one normal distribution), the acceptance rate did only increase by the factor 2 instead of the factor 30 as for the laplace approximation.





Note: The Hessians for this model have been calculated by a set of functions made publicly available by John D'Errico.

5 Results - Grid methods

By 'Grid Methods' we refer to methods exploring the posterior on a grid and a subsequent generation of synthetic data which is used as input for the estimation of the copula. First, we will investigate a grid consisting of the intersection of hyperplanes. In two dimensions this could, e.g., consist of the points $[i,j]_{i\in\{1,...,N\},j\in\{1,...,M\}}$. The problem here lies in the high dimensionality of the problems we want our samplers to use for. We will see that for 'low dimensions' this method might outperform the prerun-based sampler. In high dimensions, however, it is not feasible to introduce a grid splitting up the space dimensions in equidistant intervals and a subsequent evaluation of the posterior, data generation and copula estimation. For this reason the analysis in this section is only carried out for the low dimensional models only. For higher dimensions we use the grid given by the profile likelihoods introduced before. The results can be found in the next subsection.

Discrete distribution

After a given procedure provided us with a set of points $s_1, ..., s_m$ in the domain of the posterior $p(\cdot)$ we can build a discrete distribution to sample from the points $s_1, ..., s_m$ according to the weight they are given in the posterior. The corresponding random variable is defined as

$$X = \sum_{i=1}^{m} \tilde{p}(s_i) 1_{s_i} \ ,$$

where

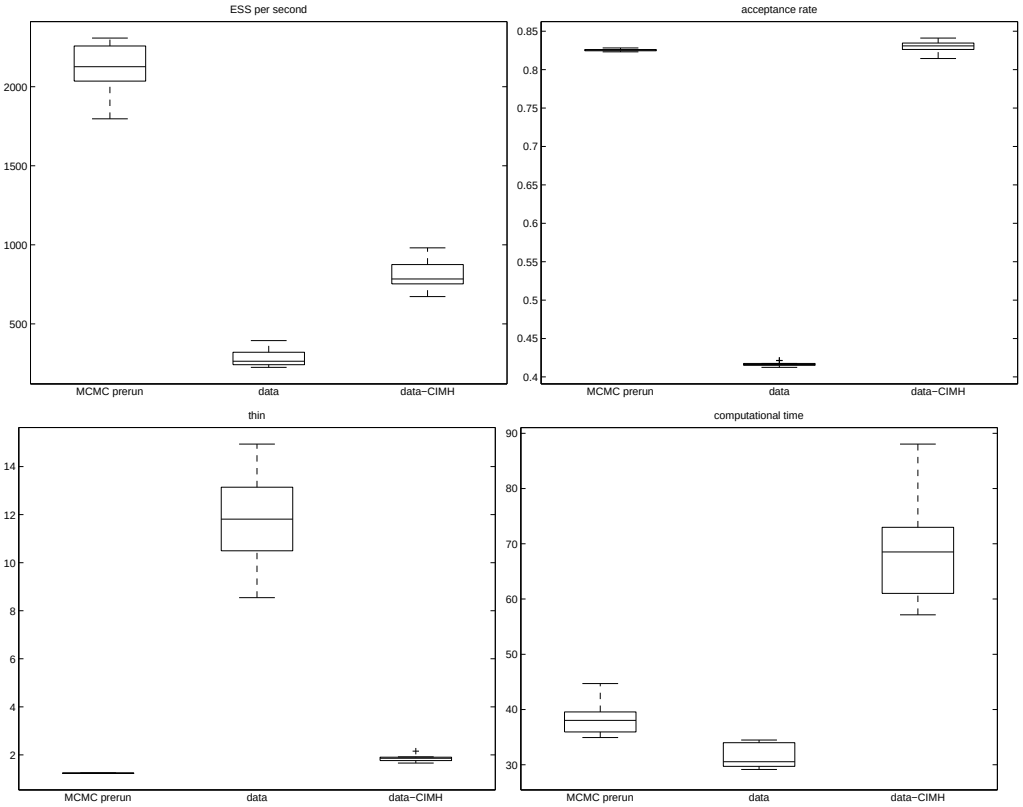
$$\tilde{p}(s_i) := \frac{p(s_i)}{\sum_{j=1}^{n} p(s_j)} \ ,$$

and 1. is the indicator function. We now implemented a sampling procedure based on the inversion method, i.e. we transformed a random number from a uniform distribution on $[0, 1]$ to the corresponding s_i via the inverse distribution function (see [Glasserman]).

5.1 Cube

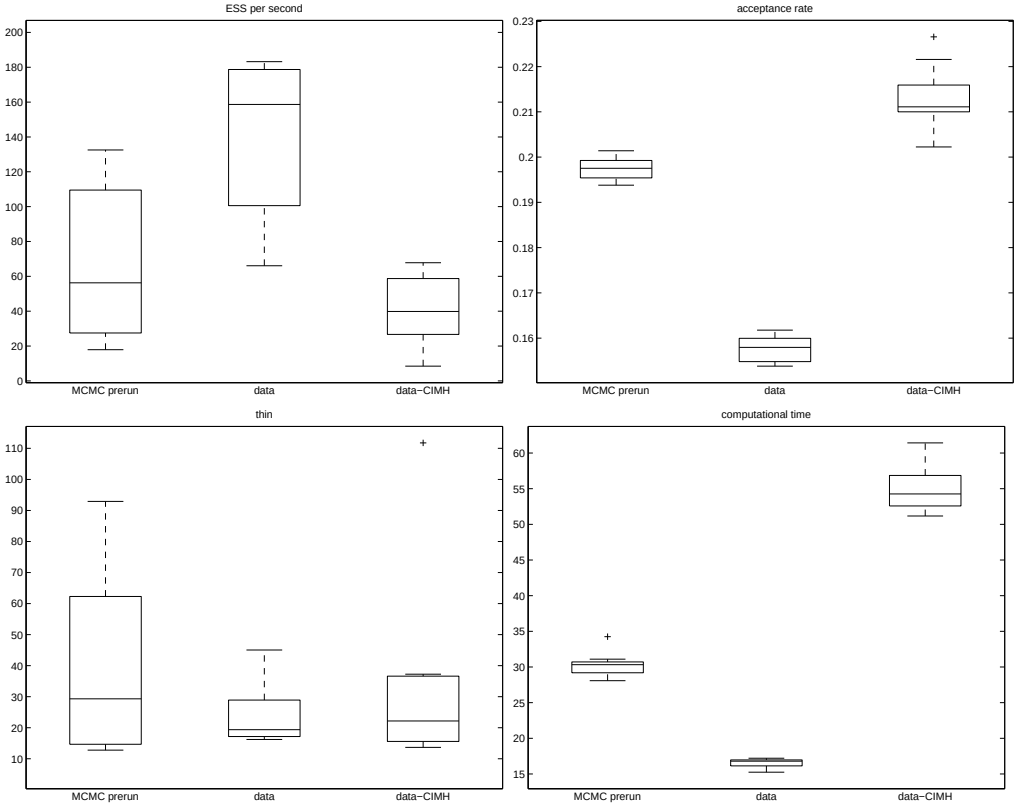
5.1.1 Normal distribution

For the normal distribution we again encounter the situation that the autocorrelation within the Markov Chain is relatively low (with a median value of 1.24 compared to 11.81 for the approxiation based and 1.86 for the adaptive approximation based sampler). Setting up the grid and subsequent sampling took 0.9 seconds compared to a prerun of 2.59 seconds. Due to the rather low acceptance rate (with a median value of 0.41) and high autocorrelation the approximation based sampler is outperformed by the prerun based sampler. The adaptive sampler suffers from a long time to estimate the copula. The acceptance rate on the other hand is essentially comparable to the one of the prerun sampler.



5.1.2 Banana shaped distribution

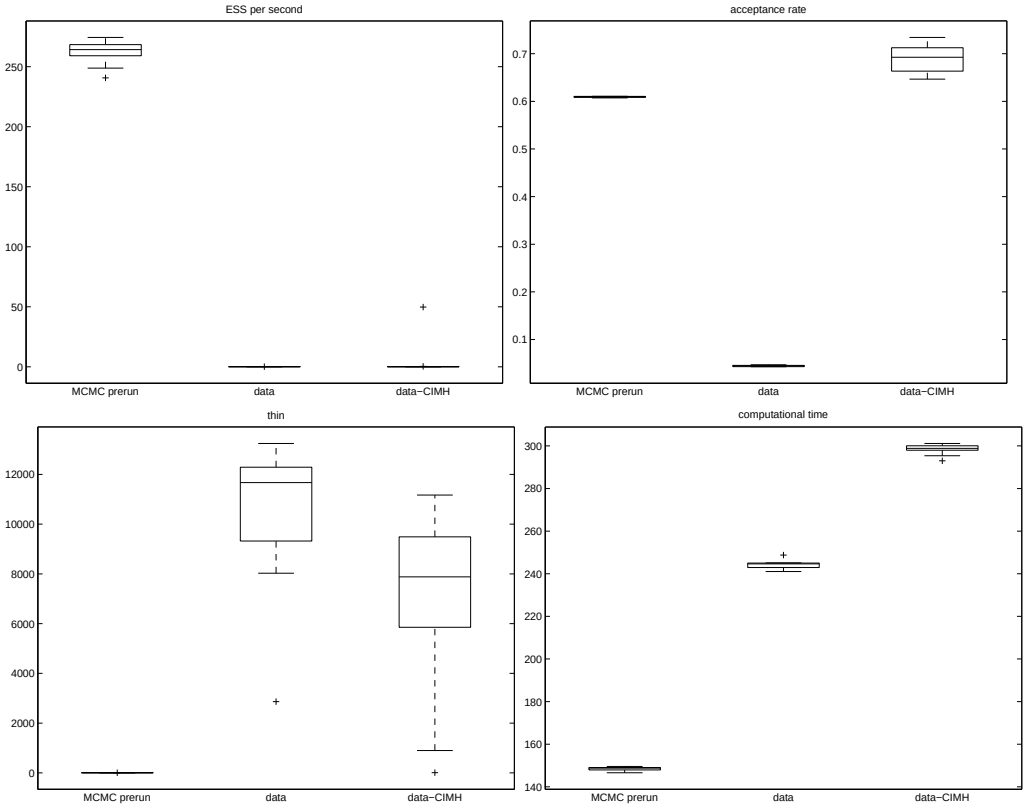
For this example we see that the autocorrelation within the chains based on the discrete distribution are lower than the ones for the prerun based sampler. The median values are at 29 samples for the prerun based and 19 and 22 for the grid based samplers. In terms of ESSps the prerun based sampler is outperformed by the approximation based sampler. The acceptance rate of 0.15 for the grid based sampler indicates that the dependence structure that is representable in the context of the vine copula estimation procedure is already covered well by the discrete distribution.



5.1.3 Compartment model

For the compartment model we see that the autocorrelation within the Markov Chains of the grid based samplers were unfeasibly high to yield acceptable results (median values of 11671 for the approximation based and 7881 for the adaptive sampler compared to 2.54 for the prerun based sampler). In addition, setting up a representative grid and sampling from it took significantly longer then the prerun and the copula estimation. Consequently, the overall computational time was higher even for the approximation based sampler. We note that the size of the prerun used for this run was at 10000 for the prerun based

sampler. Hence, the overall computational time is longer when compared to the results of the other sections.



The results for this sample show

5.1.4 Zirconium model

The results for the compartment model shows that setting up a representative grid is a daunting task. Even with a relatively fine structure we were not able to recover more information than corresponding to a 4 percent acceptance rate. Hence, we consider a twelve dimensional grid unfeasible for data generation.

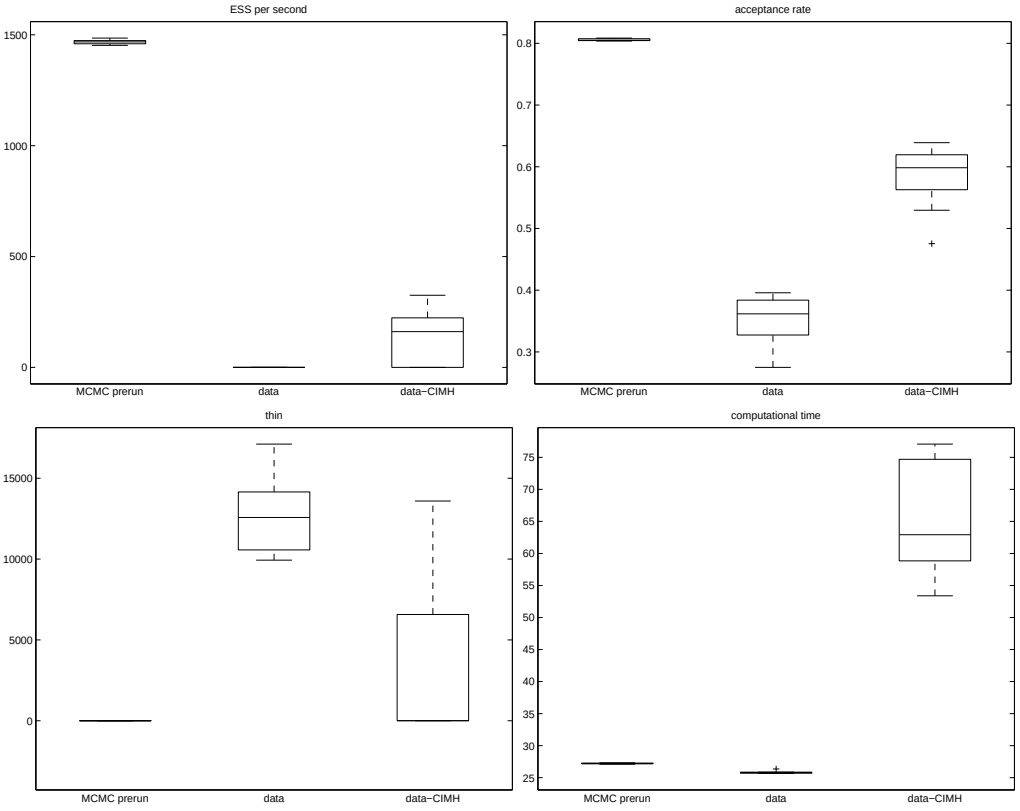
5.2 Profile Likelihoods

In this section we want to investigate to use profile likelihoods for the synthetic data generation.

5.2.1 Normal distribution

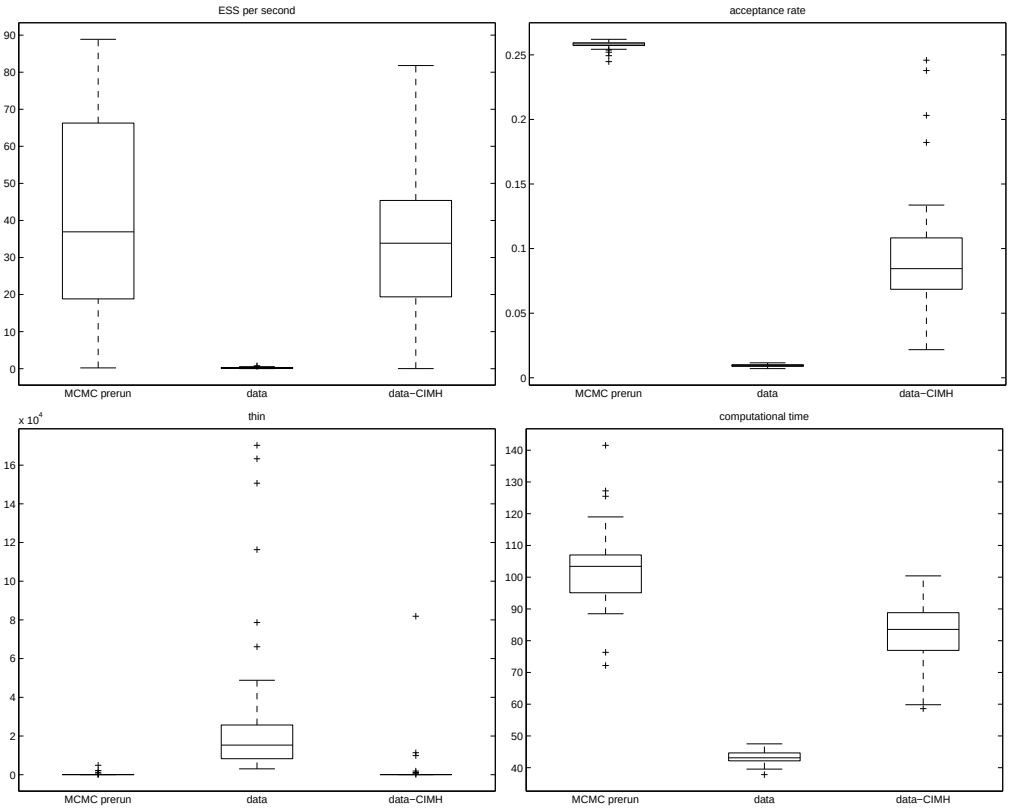
We see that the acceptance rate for the approximation based sampler is rather low when compared to the results for the other normal distributions from the previous sections and when compared against the prerun based sampler (0.39

against 0.85 for the prerun based sampler). In addition, the autocorrelation within the chain for the approximation based sampler is very high (at 12500). For the adaptive sampler we note that the median value for thinning is at 5 samples. However, in 40 percent of the runs the thinning was at more than 3800. For the prerun based sampler was again very low at 1.25. In terms of ESSps the prerun based sampler clearly outperforms the profile likelihood based samplers.

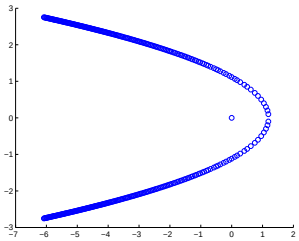


5.2.2 Banana shaped distribution

For this example we see that the adaptive approximation based sampler performs similarly to the prerun based sampler. The approximation based sampler, however, suffers from low acceptance rates and high thinning values (0.01 and 15720). From this point of view it is surprising that the adaptive version performs almost similar to the prerun based sampler in terms of ESS (median results: 36 ESSps for the prerun based sampler and 33 for the adaptive approximation based sampler). We already pointed out structural weaknesses of the profile likelihoods for the banana shaped distribution.



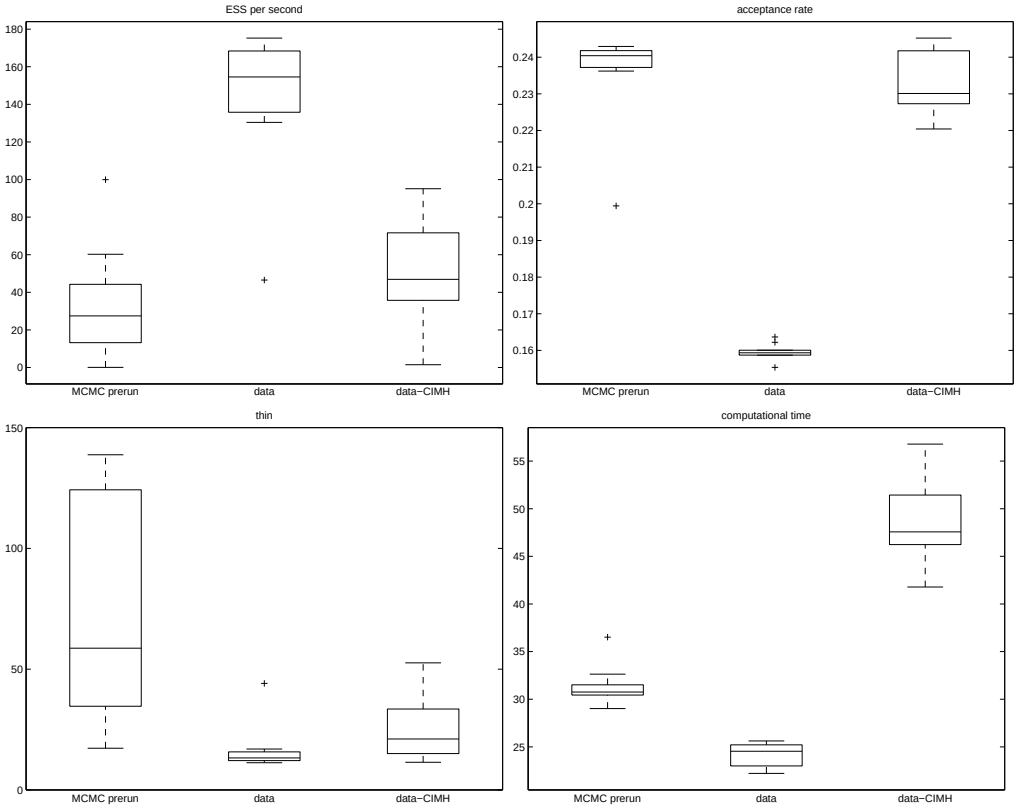
For this example we also include a run based on the profile likelihood for just one parameter. The reason for this is apparent when we consider the shape of the density and the shape of the profile in the second parameter. The majority of the mass is located around the profile of this parameter.



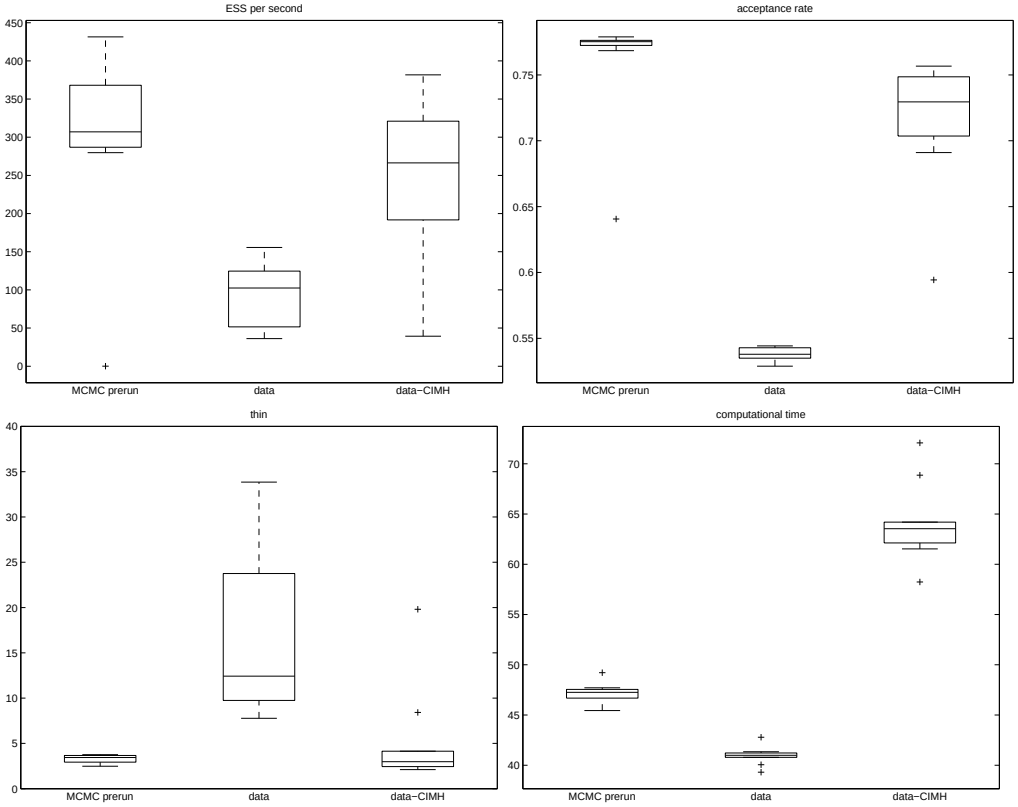
The figure depicts the profile likelihood of the parameter on the y-axis. Based on this profile we set up a discrete distribution and proceed with the sampling procedure.

For the results we see that, when compared to the other runs of the prerun based sampler, the thinning values are rather high at 58 samples. The auto-correlation for the chains of the approximation based samplers were at 13 and

21. In contrast to the very low acceptance rate of the profile likelihood based sampler using two parameters, the acceptance rate of the approximation based sampler nicely recovered to 0.18. For the adaptive version, the acceptance rate are just slightly lower than the ones for the prerun based sampler. In terms of ESSps the approximation based samplers outperform the prerun based with 261 for the approximation based and 59 for the adaptive version to 30 for the prerun based sampler.

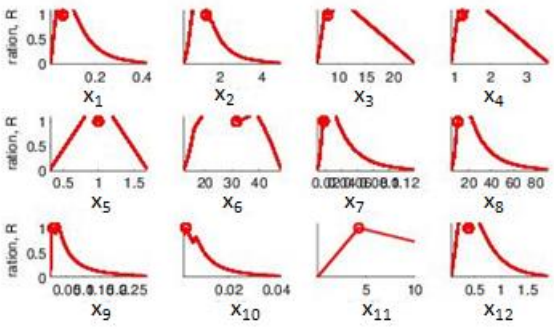


5.2.3 Compartment model



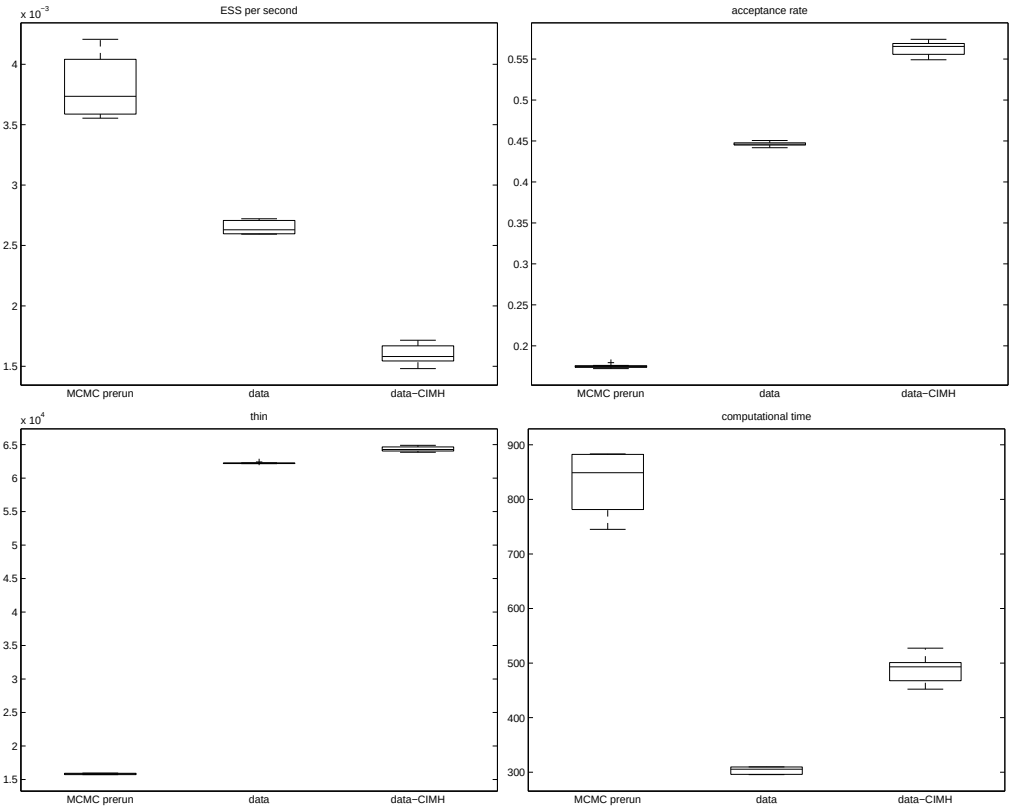
5.2.4 Zirconium model

For the zirconium model we first note that the profile for x_{11} depicted in the figure below consists only of three points. This is due to a chosen minimal stepsize of 10^{-4} in the profile likelihood calculation. Since the order of this parameter is so small, we would have to choose a significantly lower stepsize to get a reliable profile for this parameter. However, this would increase both the data from which the discrete distribution is build as well as the time to compute the likelihoods since the minimal step size would be used for all parameters. In addition, the acceptance rates for the sampler with the smaller stepsize were lower (results not shown).



For the run depicted below the most surprising fact is that the acceptance rate of the approximation based and the adaptive approximation based sampler were at 0.45 and 0.57, respectively. This is more than for the prerun based sampler and, in fact, better than the acceptance rate the profile likelihood-based sampler acchieved for the normal distribution. The thinning values for the approximation-based samplers were on average 4 times higher than for the prerun-based sampler (with the prerun sampler at around 16000, the approximation based sampler at 62000, and the adaptive sampler at 64000). The calculations of the profiles and subsequent sampling took 217 seconds. The prerun took 32 seconds. In terms of ESS the prerun-based sampler outperforms the profile likelihood-based samplers by 0.0037 ESSps to 0.0026 and 0.0016, respectively. We note, however, that in contrast to the laplace approximation the results are of the same order and we experienced runs in which the likelihood-based samplers were more competetive.

We note that the results for this sampler have been acchieved by using 99.8 percent copula samples and 0.2 percent heavy tailed samples and the results were comparable also for higher fractions of heavy tailed samples. Any inclusion of random walk samples, however, drastically decreased the acceptance rate to orders of 0.001! Since the argument in [Holden et al] and [Holden] relies entirely on the strong Doeblin condition and, in addition, allows for the inclusion of proposals depending on the current state of the process we still know that the Markov Chain converges. We were not able to detect any kind of muster in pairwise scatter plots or similar things that could explain why this behaviour occurs. If one considers profile likelihood-based copula sampling a promising sampling technique one should follow up this question.



6 Conclusion

We investigated the question if one can increase the efficiency of a copula-based sampling procedure by employing deterministic posterior approximations and grid methods. In their paper [Schmidl, Czado, Hug, Theis] showed that the inclusion of the dependence structure in the proposal function of a Metropolis-Hastings sampler can lead to significant computional gains when compared to classical samplers.

In principle, it is reasonable to question whether the preparation steps of the copula-based sampler are necessary or can be skipped or circumvented to some extend. However, the deterministic posterior approximations which are available to circumvent those steps show incompatibility with the given question. They ignore the dependence structure which made the copula sampler fast in the first place and an adaption of the methods to allow for dependence seems not immediately possible. For the complicated results even more refined methods like the gaussian mixture laplace approximation did not show acceptable results.

To offer an alternative we used 'Grid Methods': Direct evaluations of the

posterior on a grid and subsequent sampling from the corresponding discrete distribution. When setting up a grid on hyperplanes the problem lies in the highdimensionality. Sampling from points in such a 'hyperplane grid' and a subsequent copula estimation becomes unfeasible fast with increasing dimension. Profile Likelihoods offer an opportunity to investigate the space on a grid which covers weight in all parameters. For the inference of the zirconium model the sampler based on profile likelihoods significantly outperformed the prerun based sampler. However, the results for the normal distribution show that there are examples in which the profile likelihoods can not be used to capture the dependence structure of the posterior reasonably well. For the bananashaped distribution we already started a further investigation by excluding parameters to acchieve better results. In principle, we did the same for the zirconium model when setting the step size so high that one parameter was, effectively, not part of the calculation. One could follow up with the application to other examples and try to find criteria for posteriors that might be covered sufficiently well. Niveau set seem to be of use for this investigation.

Nevertheless, we saw that the data-based sampler might outperform the original sampler for examples that do not show very complicated dependence structures. Esspecially the correlated multivariate normal distribution can, naturally, be covered very well by the Laplace Approximation and we saw an outperformance of the data-based sampler for this example. The same holds for the example were the copula cannot be estimated very well in general. Here the better information on the dependence structure we can expect in the prerun cannot be exploited by the prerun based sampler. Hence we saw similar low acceptance rates as well as an outperformance of the data based and adaptive data based sampler due to less preperation steps. However, for the models with rich dependence only the profile likelihood based sampler turned out to be useful.

Acknowledgements

I want to thank Prof. Konstantinos Panagiotou and Prof. Fabian Theis for an uncomplicated administration that enabled me to get an insight in the extremely interesting field of computational and mathematical biology. A sincere 'Thank you!' to Jan Hasenauer for introducing me to profile likelihoods after I already lost hope that something would work out. At last I want to thank Sabine Hug for putting up with me in my current state.

Eigenständigkeitserklärung

Hier mit bestätige ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilsmittel benutzt habe.

Jakob Krause, 14.03.2014

7 Bibliography

[Aas et al] Aas K.; Czado, C.; Frigessi, A.; Bakken H. (2009). Pair-copula constructions of multiple dependence. Insurance, Mathematics and Economics 44

[Del Moral] Del Moral, P. (2004). Feynman-Kac Formulae. Springer Series: Probability and Its Applications

[Bedford, Cooke 1] Bedford T; Cooke R. M. . Probability density decomposition for conditionally dependent random variables modeled by vines. Annals of Mathematics and Artificial Intelligence 32, 245268.

[Bedford, Cooke 2] Bedford T; Cooke R. M. . Vines - a new graphical model for dependent random variables. Annals of Statistics 30 (4), 10311068.

[Brooks et al] Brooks S; Gelman A; Jones G. L.; Meng X. . (2010). Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC.

[Diestel] Reinhard Diestel. Graph Theory. Electronic Edition 2005 c Springer-Verlag Heidelberg, New York 1997, 2000, 2005.

[Durret] Durret R. . (2010). Probability Theory and Examples (4th Edition) Cambridge Series in Statistical and Probabilistic Mathematics

[Greenberg] Greenberg E. . Introduction to Bayesian Econometrics (2nd Edition). (2013). Cambridge University Press.

[Greiter et al 1] Greiter M, Giussani A, Höllriegl V, Li W, Oeh U: Human biokinetic data and a new compartmental model of zirconium A tracer study with enriched stable isotopes. Sci Total Environ 2011, 409:37013710.

[Greiter et al 2] Greiter M, Höllriegl V, Oeh U: Method development for thermal ionization mass spectrometry in the frame of a biokinetic tracer study with enriched stable isotopes of zirconium. Int J Mass Spectrom 2011, 304:18.

[Girolami, Calderhead] Girolami M.; Calderhead B.; (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology),73

[Girolami, Mira] Girolami M.; Mira A. . Comment on Article by Schmidl et al. Bayesian Analysis 8 (2013), no. 1, 27–32. doi:10.1214/13-BA801B. http://projecteuclid.org/euclid.ba/1362406649.

[Glasserman] Glasserman P. . (2003) Monte Carlo Methods in Financial Engineering, Springer-Verlag

[He et al] He Y.; Hodges J.; Carlin B. .(2007). Re-considering the variance parameterization in multiple precision models. Bayesian Analysis, 2(3): 529556

[Hobæk Haff et al] Hobæk Haff I.; Aas K.; Frigessi A. (2010). On the simplified pair-copula construction simply useful or too simplistic? Journal of Multivariate Analysis, 101

[Holden] Holden L. .(2000). Convergence of Markov chains in the relative supremum norm. Journal of Applied Probability, 37, 1074-1083.

[Holden et al] Holden L.; Hauge R.; Holden, M. .(2009). Adaptive independent Metropolis-Hastings. The Annals of Applied Probability, 19, 395-413.

[Hug, Raue et al] Hug S.; Raue A.; Hasenauer J.; Bachmann J.; Klingmller U.; Timmer J.; Theis F. J. . High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. Mathematical Biosciences, April 2013.

[Kirkpatrick et al] Kirkpatrick S.; Gelatt C. D.; Vecchi M. P. . Optimization by Simulated Annealing Science, New Series, Vol. 220, No. 4598. (May 13, 1983), pp. 671-680.

[Klenke] Klenke A. .(2013)., Wahrscheinlichkeitstheorie (3. Auflage), Springer-Verlag

[Kurowicka, Joe] Kurowicka D.; Joe H. (editors). Dependence Modelling (Vine Copula Handbook). (2011). World Scientific.

[Lawrence et al] Lawrence D.; Girolami M.; Rattray M.; Sanguinetti G. . (2010). Learning and Inference in Computational Systems Biology, MIT Press

[[Liang et al] Liang F.; Liu C.; Carroll R. (2010)., Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples (Wiley Series in Computational Statistics)

[Mai, Scherer] Mai J.; Scherer M. . (2012). Simulating Copulas: Stochastic Models, Sampling Algorithms and Applications. World Scientific, Series in Quantitative Finance: Volume 4.

[McNeil, Frey, Embrechts] McNeil A., Frey R., Embrechts P. (2005). Quantitative Risk Management. Princeton University Press.

[Meyn, Tweedie] S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-Verlag, London.

[Murphy, van der Vaart] S. A. Murphy; A. W. van der Vaart (2000). On Profile Likelihood Journal of the American Statistical Association, Vol. 95, No. 450. (Jun., 2000), pp. 449-465.

[Nelsen] Nelsen R.B. (2006), An Introduction to Copulas, Springer Series in Statistics

[Raue et al] Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmller U and Timmer J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics 25(15), 1923-1929, 2009

[Schmidl] Schmidl D. (2012). Bayesian model inference in dynamic biological systems using Markov Chain Monte Carlo methods. PhD Thesis.

[Schmidl, Czado, Hug, Theis] Schmidl D., Czado C., Hug S., Theis F. J. (2013) A Vine-copula Based Adaptive MCMC Sampler for Efficient Inference of Dynamical Systems. Volume 8, Number 1 (2013), 1-268 Bayesian Analysis

[Schmidl, Hug et al] Schmidl D., Hug S., Li W. B., Greiter M. B., Theis F. J. (2012). Bayesian model selection validates a biokinetic model for zirconium processing in humans. BMC Systems Biology 2012, 6:95

[Sethuraman et al] Sethuraman J., Athreya K., Ddoss H.. (1992). A Proof of convergence of the Markov Chain simulation Method. Technical Report 868

[Stickler, Schachinger] Stickler B. A., Schachinger E., Basic Concepts in Computational Physics (2014), Springer-Verlag

[Venzon, Moolgavkar] Venzon, D. J. and Moolgavkar, S. H. (1988), A Method for Computing Profile-Likelihood Based Confidence Intervals, Applied Statistics, 37, 8794.