Reversible Jump Markov Chain Monte Carlo: Some Theory and Applications

Master's Thesis in Statistics

Financial Theory and Insurance Mathematics

Hannu Lyyjynen

UNIVERSITY OF BERGEN
Faculty of Mathematics and Natural Sciences
Department of Mathematics

March 2014

Abstract

The history of MCMC, theories of Bayesian thinking and model choice, the Accept-Reject-algorithm, Markov chains, the Metropolis-Hastings-algorithm and the reversible jump MCMC are explained. Then the reversible jump MCMC as change-point analysis is applied to the coal mine disaster example, familiar from [Green, 1995], and to the examples of counting terrorism attacks (worldwide, in Iraq and in Afghanistan). The novel part is estimating the change points of the hazard rate of terrorism attacks in Afghanistan during the last 35 years.

Acknowledgements

First of all, I want to thank Professor Dag Tjøstheim for excellent supervision, vast patience and making me to really believe in that my program works by suggesting a run with the original coal mine accident data, which I still perceive as **harder** than the data in my own futile attempts.

> “You can do whatever you want, because that is what you are going to do anyway!” – prof. Tjøstheim

Secondly, I want to thank Associate Professor Bård Støve for his course Monte Carlo Methods in Statistics’ introducing me to the topic of MCMC-simulation.

A special thanks goes to Lars Jordanger for his expert advice on the LaTeX-typesetting system.

I also want to thank all fellow students both earlier in Kroepeliens hus and in the present location of Statistics in Realfagbygget for making my study time in Bergen so pleasurable.

Finally I want to thank my wife Yinru for her constant support, love, and spellchecking the thesis (the remaining mistakes are my own).

Contents

To my wife Yinru,

困难是石头,
决心是鎯头.
鎯头 打石头，
困难 就低头！

Chapter 1

Historical Review

Statistical sampling had been known for centuries, but it was really the advent of computers that made this approach feasible for attacking many problems of physics. The Monte Carlo method was part of the picture from the very beginning, thanks to Nicholas Metropolis, who was the leader of the team that designed and built one of the very first electric computing machines ENIAC and MANIAC in the Manhattan project in Los Alamos. Metropolis was also involved in improving the method by introducing a technique, which is today known by the name "importance sampling".

The Monte Carlo method applies the laws of probability by calculating samples from the modeled outcomes of real physical phenomena. For example approximating the real diffusion rate in neutron diffusion or estimating other physical quantities such as energy or density [Anderson, 1986] becomes possible.

The goal of the Markov Chain Monte Carlo (MCMC) is the opposite from analyzing the stationary distribution of the chain. One begins with the stationary distribution and constructs a reversible Markov chain (under some relatively mild regularity conditions) possessing this distribution as the stationary distribution. Simply sampling from the chain produces correlated data from the desired distribution.

However, in spite of the early dawn of the method right after the World War II, it still took decades until it revolutionized the statistical calculations. There were two main obstacles: even if the exceptional group of physicist working in Los Alamos computed on the fastest computers in the world of the time, the computing power of today is on a quite different level than is was at 1950's or even 1970's. Also, the Bayesian paradigm of thinking about statistical problems, which is most often used in simulations, was yet to become fully developed.

Nicolas Metropolis had come to Los Alamos in 1943 and was leading a team of physicists who were working hard on designing the hydrogen (H) bomb. Espesially Teller was obsessed with the bomb and with the better computing capacity offered by the second computer ever in existence called MANIAC the project was finally successful in the early 1950's The results of energy levels of a N-particle system were published in 1953 in Journal of Chemical Physics [P.Robert and Casella, 2011].

In the 1970's, the Metropolis method [N.Metropolis et al., 1953] was generalized by [Hastings, 1970] and his student [P.H.Peskun, 1973] in order to overcome the curse of dimensionality met by regular Monte Carlo methods. This difficulty had already been acknowledged by Metropolis' team.

Gibbs sampling was brought into the arena of statistical application by Geman and Geman [Geman and Geman, 1984] around mid-eighties. However, the real explosion in the use of the MCMC could only begin after the 1990's , as Gelfand and Smith published an influential paper [E.Gelfand and F.M.Smith, 1990] establishing a "genuine starting point for an intensive use of MCMC methods by the mainstream statistical community“[P.Robert and Casella, 2011].

In 1995 Peter Green generalized the MCMC method [Green, 1995] from a model parameter estimator into a model choice tool by making jumps between models of different dimensionality possible. The reversible jumping (RJ) still essentially applies the algorithm of Metropolis and Hastings. The moves needed for jumping between the different submodels are implemented in a more sophisticated fashion involving the use of the Jacobian determinant of the transformation between the submodel spaces.

In the following chapter 2 we discuss the Bayesian Paradigm and the relationship to the Choice of Model. In chapter 3 we present the Accept-Reject- sampling algorithm. Chapter 4 is an introduction to the cornerstones of the Markov Chain Monte Carlo-algorithm: the Markov Chain and the Metropolis-Hastings algorithm. Chapter 5 discusses the reversible jump MCMC and two change-point analysis applications: coal mining disasters in the UK and counting terrorism attacks worldwide, in Iraq and in Afghanistan. Finally, in chapter 6 we discuss applications, pros and cons and the future trends of the (RJ)MCMC.

Chapter 2

The Bayesian Paradigm and the Choice of a Model

"Essentially, all models are wrong, but some are useful."
– George E. P. Box

The advent of more powerful computing capacity during the last few decades has made simulation increasingly important as a method of performing statistical inference for systems of virtually unlimited complexity.

A model M has the purpose to capture and formalize the unknown dependency between some unknown y and some known quantities x, and to describe a phenomenon or a system. The model consists of structural assumptions S and the model parameters θ. Usually most if not all of the inferential attention is given to the analysis of the parameters, even though the model structure chosen also carries a profound importance. The usual procedure is to pick the ”best” model structure S^* for S by examining the data, and after identifying S^* to proceed as if S^* is acknowledged to give the correct inferences and predictions [Draper, 1995].

In the Bayesian approach, the parameters θ are not considered as being fixed to having certain values, but are also allowed to behave as stochastic variables having densities. Hence there is not much difference on the conceptual level on measured data and the model parameters. The likelihood function of parameters given the data x is identical to the density function of the variable(s) conditional on θ, only the point of view is reversed:

$$l(\theta) = l(\theta|x) = f(x|\theta). \tag{2.1}$$

Also the philosophy becomes reversed: the function will no longer be considered as the density of variables which are sampled from a distribution with certain parameter value, but the probability of obtaining a value of the parameter θ given the data x.

The experience and beliefs of the researcher on the phenomena at hand are incorporated into the model by introducing *a prior* distribution $\pi(\theta)$ for the parameter. This happens before measuring any data.

After the measurement of N points of data, X_N, is performed, the Bayes' theorem gives the *a posterior* distribution,

$$\pi(\theta|X_N) = \frac{f(X_N|\theta)\pi(\theta)}{f(X_N)} \propto l(\theta|X_N)\pi(\theta) \tag{2.2}$$

updating the beliefs of the parameter. Here the normalization constant

$$f(X_N) = \int f(X_N|\theta)\pi(\theta)\, d\theta \tag{2.3}$$

is usually unknown in practice, but as will be seen it won't be needed in the simulation. Calculating it is often as hard a problem as sampling from the original distribution. After all, if the distribution was fully known there would not be much need for using the Markov chain for sampling from it in the first place.

The prior and posterior are relative concepts: if still more data is collected, then the same principle can be re-applied as the current posterior replaces the prior in the new calculations.

As all distributions from the exponential family[1] have conjugate priors, some computational benefit and algebraic convenience can be acquired by using an appropriate conjugate prior distribution. This means that for a certain type of a prior distribution $\pi(\theta)$ the posterior calculated from

$$\pi(\theta|x) \propto l(\theta|x)\pi(\theta) \tag{2.4}$$

will have the same functional form as the prior, but with different, updated, parameter values. Well-known examples of this within the discrete distributions are prior-posterior pairs beta-binomial, gamma-Poisson and Dirichlet-multinomial. In the context of single-variable continuous distributions, for example, normal-normal make a conjugate pair.

[1] A distribution in the exponential family, can *(canonically)* be expressed as $f(x, \theta) = \exp(xb(\theta) + c(x) + d(\theta))$.

In the multivariate setting, if the covariance matrix Σ is known, the multivariate normal distribution is a conjugate of itselfs. However, such an approach always has a trade-off between eliminating any structural model uncertainty by once and for all fixing the shape of the model function and really modeling the data well.

The Bayesian approach to hypothesis testing also leads to an alternative for choosing a model between two or more possible candidates by using the so called Bayes factors [E.Kass and E.Raftery, 1995]. If D is the data and M_1 and M_2 are two alternative models (or hypotheses) such that $P(M_2) = 1 - P(M_1)$, then a direct application of the Bayes' theorem gives

$$P(M_j|D) = \frac{P(D|M_j)P(M_j)}{P(D|M_1)P(M_1) + P(D|M_2)P(M_2)} \quad (j = 1, 2), \tag{2.5}$$

and hence

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)} \cdot \frac{P(M_1)}{P(M_2)}. \tag{2.6}$$

Here it could also be observed, that in the case of equiprobable models, $P(M_1) = P(M_2) = 0.5$, the Bayes factor equals the posterior odds in favour of model M_1. The formula resulting from this case,

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)}, \tag{2.7}$$

is conceptually important turning point in the history of statistics for the first time inverting conditional probabilities [P.Robert, 2007].

As can be seen from (2.6), the Bayes factor,

$$B_{12} = \frac{P(D|M_1)}{P(D|M_2)}, \tag{2.8}$$

is the ratio of posterior odds and prior odds of the models. It can also be interpreted as the transformation factor from the prior opinion to the posterior opinion representing the evidence provided by the data. If the models contain no parameters, the Bayes factor is simply the likelihood ratio of the two models, and with parameters the ratio still looks like the likelihood ratio, but in contrast to maximizing the likelihoods, the conditional densities

$$P(D|M_j) = \int P(D|\theta_j, M_j)\pi(\theta_j|M_j)d\theta_j \tag{2.9}$$

need to be integrated. Here θ_j is a parameter (vector) of model M_j and $\pi(\theta_j|M_j)$ is its prior density.

Bayesian posteori distributions

$$\pi(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)\, d\theta} \propto l(\theta|x)p(\theta) \tag{2.10}$$

can be simulated from in order to produce a sample $\theta_1, \theta_2, \ldots, \theta_m$. This sample can then be used to approximate the moments of the posterior distribution or any other quantities characterizing the posterior needed.

There are several methods for measuring the complexity of a model, both in its own right and as a tool for the model choice. To mention few such so called *information criteria*, there are AIC [Akaike, 1974], BIC [Schwarz, 1978], TIC [Takeuchi, 1976] and NIC [N.Murata et al., 1978] which all trade off model fit against the effective number of model parameters. In these criteria, the fit of a model is guaranteed by maximizing the likelihood and overfitting is avoided by penalizing by the number of parameters.

In [J.Spiegelhalter et al., 2002] the authors develop a Bayesian approach, the deviance information criteria (DIC) which is directly computable from the MCMC posteori samples.

Chapter 3

The Accept-Reject Algorithm/Sampling

> "Accept-Reject- sampling is much harder than you think!"
> –Geir Drage Berentsen

The Accept-Reject algorithm for stochastic sampling was originally an idea of John von Neumann's. He received a letter in 1946 from Stanislaw Ulam who was interested in estimating the passing rate in the game of Canfield solitaire, which he had been playing while being sick. Von Neumann was also eager to expand the "games" to some other calculations of neutron fission energies and wanted to implement those on the very first calculating machine ENIAC [Eckhardt, 1987].

The Metropolis-Hastings algorithm always contains an Accept-Reject- step, while a componentwise variant called Gibbs sampling does accept every proposal move with probability one.

The goal is to sample from $\pi(x)$, which can be approximated by an *instrumental* density $g(x)$ such that $\pi(x) \leq Mg(x)$ and $\text{Supp}(\pi) \subset \text{Supp}(g)$. The constant $M \geq 1$ is not necessarily known. This is achieved by first generating a stochastic variable $Y \sim g$ and another uniformly distributed variable $U \sim \text{Unif}[0,1]$. The variable Y is then accepted as X if $U \leq \pi(Y)/Mg(Y)$.

Algorithmically speaking,

1. Generate $Y \sim g$ and $U \sim \text{Unif}[0,1]$;

2. Accept $X = Y$, if $U \leq \pi(Y)/Mg(Y)$;

3. If not, return to the 1. step.

Indeed, the distribution becomes π :

$$P(X \leq x) = P(Y \leq x \mid Y \text{was accepted}) = P(Y \leq x \mid U \leq \frac{\pi(Y)}{Mg(Y)}) =$$

$$\frac{P(Y \leq x \,\&\, U \leq \frac{\pi(Y)}{Mg(Y)})}{P(U \leq \frac{\pi(Y)}{Mg(Y)})} = \frac{\int_{-\infty}^{x} \int_{0}^{\frac{\pi(y)}{Mg(y)}} du\, g(y)\, dy}{\int_{-\infty}^{\infty} \int_{0}^{\frac{\pi(y)}{Mg(y)}} du\, g(y)\, dy} = \frac{\int_{-\infty}^{x} \frac{\pi(y)}{M}\, dy}{\frac{1}{M} \int_{-\infty}^{\infty} \pi(y)\, dy} = \int_{-\infty}^{x} \pi(y)\, dy.$$

Setting $M = \sup_y \frac{\pi(y)}{g(y)} < \infty$ gives the acceptance probability as

$$P(Y \text{ accepted}) = P(U \leq \frac{\frac{\pi(y)}{g(y)}}{\sup_y \frac{\pi(y)}{g(y)}}) = \int_{-\infty}^{\infty} \int_{0}^{\frac{\pi(y)}{Mg(y)}} du\, g(y)\ dy = \frac{1}{M},$$

which means that the waiting time of acceptance is geom$(1/M)-$ distributed and expected number of trials for accepting a variable is M.

For a different point of view [Mikusheva, 2007], denote by ρ the probability that a single draw was not accepted:

$$\rho = P[\text{a single draw was rejected}]. \tag{3.1}$$

We want to simulate from $X \sim \pi(x) = \frac{f(x)}{k}$ where the constant k is unknown. We have a candidate pdf $g(x)$ to simulate from and a known constant M such that

$$f(x) \leq Mg(x). \tag{3.2}$$

Use of the iterated expectations $EX = E_Y[E_{X|Y}(X|Y)]$ and observing that

$$M(1 - \rho) = k,$$

give

$$\begin{aligned} P(X \leq x) = P(y \leq x \ \& \ U \leq \frac{y}{Mg(y)})(1 + \rho + \rho^2 + \ldots) = \\ \frac{1}{1-\rho} P(y \leq x \ \& \ U \leq \frac{y}{Mg(y)}) = \\ \frac{1}{1-\rho} E_y[P(U \leq \frac{y}{Mg(y)} |y) \ \mathbb{1}_{\{y \leq x\}}] = \\ \frac{1}{1-\rho} \int_{-\infty}^{x} \frac{f(y)}{Mg(y)} g(y) \, dy = \int_{-\infty}^{x} \frac{f(y)}{M(1-\rho)} \, dy = \\ \int_{-\infty}^{x} \pi(y) \, dy. \end{aligned}$$

The waiting time is geometric, as the probability of acceptance becomes $1/M$.

In the Accept-Reject method only an instrumental density $g(x)$ such that the target distribution is approximated $\pi(x) \leq Mg(x)$ with some constant $M > 0$ is needed.

Chapter 4

Markov Chain Monte Carlo

4.1 Markov Chain

To give the discussions some substance, we first collect some essential definitions and theorems on Markov Chains in this section mainly from [Feller, 1968].

Let us consider a probability space $(S, \mathcal{F}_n, P)$, where $S \subset \Re^d$ is a subset of an Euclidian space (for a simpler setting), $\mathcal{F}_n$ is a filtration and P is a probability measure. The filtration is the $\sigma-$algebra generated by the sample points X, $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$.

Definition 4.1. A Markov transition kernel/matrix, $K(x, A)$, is defined by a transition probability function $K : S \times B \to \Re$ satisfying

- $\forall x \in S : A \to K(x, A)$ is a probability measure on (S, B).
- $\forall A \in B : x \to K(x, A)$ is a measurable function.

Here $x \in \Re^d$, $A \in B$, where B is the Borel $\sigma-$algebra on $\Re^d$. If the space of possible states is finite, the kernel becomes simply a matrix of transition probabilities. In a more abstract setting an integral operator is needed.

Hence the probability of the chain moving from sample point X_n into the Borel set B can be expressed as

$$P(X_{n+1} \in B \mid \mathcal{F}_n) = K(X_n, B). \tag{4.1}$$

Definition 4.2. A discrete-time homogeneous Markov chain with a countable state-space $S = \{X_k\}_{k=0,1,\dots}$ is a stochastic process X_n with the property that the next state only depends on the current state of the system but not on history preceding it. That is, the probability of moving from one state to another is

$$P[X_{n+1} = j|\mathcal{F}_n] = P[X_{n+1} = j|X_0 = x_0, X_1 = x_1, \dots, X_n = i] = \tag{4.2}$$

$$P[X_{n+1} = j|X_n = i] = p_{ij}. \tag{4.3}$$

Thus each step on the chain depends **only** on the previous one. Of course, $\sum_j p_{ij} = 1$. Denote the probability of reaching from state i to state j in n steps $p_{ij}^{(n)}$ and that this happens first time in n steps $f_{ij}^{(n)}$ and ever $f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$.

Definition 4.3. A state j in a Markov Chain is **accessible** from state i, if $p_{ij}^{(n)} > 0$ for some integer n. Two states i and j are said to **communicate** ($i \leftrightarrow j$) if they are accessible from each other.

This creates an equivalence relation which can be used to partition all states into equivalence classes.

Definition 4.4. The Markov chain is **irreducible** if it cannot be decomposed into distinct subsets of communicating states. A criterion for irreducibility is: Every state can be accessed from every other state.

Definition 4.5. A state i is **transient** if $f_{ii} < 1$. If $f_{ii} = 1$, the state is **positive recurrent**.

Definition 4.6. The **Markov chain is positive recurrent** or (persistent) if $f_{ii} = 1 \forall i$.

Definition 4.7. The Markov chain has **period** $k > 1$, if $p_{jj}^{(n)} = 0$, unless $n = \nu k$ is a multiple of k, and k is the largest integer with this property. If this is not the case, the Markov chain is called **aperiodic**.

Definition 4.8. A probability measure μ representing a possible equilibrium for the chain is called a **stationary** or **(invariant) distribution** if

$$\sum_x \mu(x)K(x,y) = \mu(y).$$

Theorem 4.9. *A positive recurrent aperiodic equivalence subclass C of a Markov chain has an invariant stationary distribution π, uniquely determined by equations*

$$\sum_{i \in C} \pi_i p_{ij} = \pi_j, \quad \sum_{i=0}^{\infty} \pi_i = 1, \quad \pi_i \geq 0 \quad \forall j \in C. \tag{4.4}$$

In a more general setting, such as [Tierney, 1994] and [S.P.Meyn and R.L.Tweedie, 2005], we define a posterior target distribution μ, which has a density (also denoted by μ) with respect to a $\sigma-$finite measure ν,

$$\mu(\mathrm{dx}) = \mu(x)\nu(\mathrm{dx}). \tag{4.5}$$

Definition 4.10. A **time-homogeneous Markov chain with invariant distribution** μ is a sequence of random variables $\{X_n\}_{n=0,1,\dots}$ such that the transition kernel

$$K(X_n, B) = P[X_{n+1} \in B | \mathcal{F}_n], \tag{4.6}$$

where $\mathcal{F}_n$ is the filtration[1] induced by the chain, satisfies

$$\mu(B) = \int \mu(dx) K(x, B) \tag{4.7}$$

for all measurable sets B.

4.1.1 Monte Carlo Integration

If π is a probability density, from which sampling is possible, it is easy to estimate the integral

$$E_\pi(h(X)) = \int_X h(x)\pi(x)\, dx \tag{4.8}$$

with

$$\bar{h}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i), \tag{4.9}$$

where $X_i \sim \pi(x)$. The estimate converges almost surely to $E_\pi(h(X))$ by the Strong Law of Large Numbers. Also, if

$$\int_X |h(x)|^2 \pi(x)\, dx < \infty, \tag{4.10}$$

[1] A filtration is a sequence of $\sigma-$algebras $\{F_t\}_{t \geq 0}$, such that $t_1 \leq t_2 \Rightarrow F_{t_1} \subseteq F_{t_2}$ and $F_t \subseteq F \forall t$, where (Ω, F) is a given measurable space. The concept of a filtration can be interpreted as the information of the system available at time t.

then assessing the convergence speed can be done by estimating

$$Var(\bar{h}_m) = \frac{1}{m}\int_X (h(x) - E_\pi(h(X)))^2\, \pi(x)\, dx \tag{4.11}$$

with

$$v_m = \frac{1}{m^2}\sum_{j=1}^{m}[h(X_j) - \bar{h}_m]^2. \tag{4.12}$$

The expression

$$\frac{\bar{h}_m - E_\pi(h(X))}{\sqrt{v_m}} = \sqrt{m}\frac{\bar{h}_m - E_\pi(h(X))}{\sqrt{\frac{1}{m}\sum_{j=1}^{m}[h(X_j) - \bar{h}_m]^2}} \tag{4.13}$$

will by the Central Limit Theorem be asymptotically standard normal $N(0,1)-$distributed and the probability $(1-\alpha)-$confidence bounds can hence be constructed as

$$[\bar{h}_m - z_{\frac{\alpha}{2}} m^{-\frac{1}{2}}\sqrt{mv_m}, \bar{h}_m + z_{\frac{\alpha}{2}} m^{-\frac{1}{2}}\sqrt{mv_m}], \tag{4.14}$$

where $z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}-$quantile of the standard Gaussian distribution. This technique is well-known as the classical Monte Carlo- integration.

Remark 4.11. It should be noted that in lower dimensions, for smooth integrand functions, the Gauss quadrature totally outperforms Monte Carlo as a numerical integration method. The convergence of Monte Carlo integration is slow. In practice, the rate $\frac{1}{\sqrt{m}}$ implies that an extra digit requires approximately 100 as many replications, but a remarkable fact is that the factor $\frac{1}{\sqrt{m}}$ remains the same, no matter how high the dimension of the underlying space X is [Cappé et al., 2005]. This is the reason Monte Carlo methods become attractive for simulating and doing inference in high-dimensional practical settings.

4.2 Metropolis-Hastings Algorithm

While one of the main concerns of the theory of Markov Chains is to confirm the existence and uniqueness of a stationary distribution for iterations of a given transition kernel, the algorithm of Metropolis [N.Metropolis et al., 1953] and Hastings [Hastings, 1970] employs the opposite strategy.

The purpose of the Metropolis-Hastings algorithm (MH) is drawing samples from a target probability distribution $\pi(x)$ by generating a Markov Chain whose stationary distribution is $\pi(x)$. Only a function proportional to the density (such as in the Bayesian setting) needs to be calculable.

The generated Markov Chain needs to be irreducible, positive recurrent, aperiodic and reversible. These relatively mild conditions guarantee the existence of the invariant stationary distribution.

There are many excellent accounts on the MH-algorithm, see for example
[N.Metropolis et al., 1953], [Hastings, 1970], [Tierney, 1994],
[Chib and Greenberg, 1995], [S.Liu, 2001], [Nummelin, 2002],
[Gamerman and Lopes, 2006], [P.Robert and Casella, 2004] or
[P.Robert and Casella, 2010].

4.2.1 Implementing the Metropolis-Hastings Algorithm

We are following [Tierney, 1994], [Nummelin, 2002], and [Chib and Greenberg, 1995].

The state space of the Markov chain $X_0, X_1, \ldots$ can be finite, countable, a subset of an Euclidian space: $E \subset R^k$ or even a more general measure space. The more general the space, the more abstract shape will the operator for the transition kernel that is governing the movement of the chain take.

We assume the Markov property on the chain entering a set $A \subset E$ (in an Euclidian space),

$$P(X_{n+1} \in A | X_0 = x_0, \ \ldots, X_n = x_n) = P(X_{n+1} \in A | X_n = x_n), \tag{4.15}$$

and that the chain is time-homogeneous.

Hence the next step of the chain depends **only** on the present state and the probability law, not on the history of reaching the present state. Time-homogeneity means that the stochastic transition mechanism does not change with the time index n and if X_n has a density $\lambda(x)$, it will be independent of time and

$$P(X_n \in A) = \int_A \lambda(x)\, dx. \tag{4.16}$$

Once the Markov chain has reached a point x, in the next step it either stays put in x or moves according to the probabilistic rule of the chain.

Let $0 \leq r(x) < 1$ be the probability of preserving the current state x,

$$r(x) = P(X_{n+1} = x | X_n = x),$$

and $p(x,y)$ a (sub)probability density[2] for the chain moving from a state x to a new state y.

That means, that the probability of the chain entering (or staying in) a set $A \subset E$ from a point x can be written

$$P(x,A) = P(X_{n+1} \in A | X_n = x) = \int_A p(x,y)dy + r(x)\delta_x(A), \tag{4.17}$$

where $\delta_x(A) = \begin{cases} 1, \text{ if } x \in A, \\ 0 \text{ otherwise.} \end{cases}$

The probability that $X_{n+1} \in A$ is given by

$$P(X_{n+1} \in A) = \int_E \lambda(x) P(x,A)dx =$$

$$\int_E \lambda(x)[\int_A p(x,y)dy + r(x)\delta_x(A)]dx = \int_A [\int_E \lambda(x)p(x,y)dx + \lambda(y)r(y)]dy, \tag{4.18}$$

which also defines a *Markov operator* P, mapping the probability density function $\lambda \mapsto \lambda P$. It can easily be iterated for $n = 2, 3, \ldots$ by defining $\lambda P^n = (\lambda P^{n-1})P$ and $\lambda P^0 \equiv \lambda$.

Definition 4.12. If $\pi P = \pi$, the probability density π is called the **stationary (invariant)** probability density π of the Markov chain.

It is such a chain we wish to construct from a given distribution π and draw random samples from.

Define the support of the invariant probability density function π as

$$E^+ = \{x \in E : \pi(x) > 0\},$$

and assume that it is closed in the sense of $P(x, E^+) = 1 \ \forall x \in E^+$ and that $\pi(x)$ is not a unit mass concentrated on a single point. In practice most often $E^+ = E$, but even if

[2] A subprobability density is a function $g(z) \geq 0$, such that $\int g(z)dz \leq 1$.

there would exist points z outside of E^+, so that $z \notin E^+$, in view of

$$\int_E \pi(x)P(x, E^+)dx = \int_{E^+} \pi(x)dx = 1 \tag{4.19}$$

set of such points would have measure zero as $P(x, E^+) = 1$ for almost every initial state $X_0 = x \in E^+$.

A reversed Markov chain is obtained by letting the time run backwards and studying the chain in reversed time order. Assume a homogeneous Markov chain has transition probability $p(x, y)$ and a stationary distribution π.

Then it can be shown, that

$$P(X_n = x_n | X_{n+1} = x_{n+1}, X_{n+2} = x_{n+2}, \ldots) = P(X_n = x_n | X_{n+1} = x_{n+1}) \tag{4.20}$$

and hence the time reversed chain also defines a Markov chain.

Its transition probabilities at step n are

$$p_n^{\leftarrow}(x, y) = P(X_n = y | X_{n+1} = x) = \\ \frac{P(X_{n+1} = x | X_n = y)P(X_n = y)}{P(X_{n+1} = x)} = \frac{p(y, x)\pi^{(n)}(y)}{\pi^{(n+1)}(x)}.$$

If the reversed chain also has the stationary distribution $\pi(\cdot)$, as $n \to -\infty$,

$$p_n^{\leftarrow}(x, y) \to p^{\leftarrow}(x, y) = \frac{p(y, x)\pi(y)}{\pi(x)}.$$

and the chain becomes time-homogeneous [Gamerman and Lopes, 2006].

If the transition probabilities of the reversed chain are the same as in the original chain $p^{\leftarrow}(x, y) = p(x, y)$, then

$$\pi(x)p(x, y) = \pi(y)p(y, x), \tag{4.21}$$

i.e. the rate at which the chain moves from x to y in equilibrium is the same as vice versa. As Besag puts it: "... if a stationary Markov chain $\ldots, X_{-1}, X_0, X_1, \ldots$ satisfies detailed balance (as in 4.21), then it is time reversible, which means that it is impossible to tell whether a film of a sample path is being shown forwards or backwards." [Besag, 2001].

Definition 4.13. A reversible chain has a probability density λ, such that

$$\lambda(x)p(x,y) = \lambda(y)p(y,x). \tag{4.22}$$

Remark 4.14. An alternative expression for the reversibility property is to say that the chain satisfies the detailed balance.

Theorem 4.15. *Reversibility of the chain is sufficient (not necessary) property to guarantee the existence of the stationary distribution.*

Proof. If the reversibility prevails,

$$\begin{align}\int_E \lambda(x)P(x,A)dx = \int_E \lambda(x)[\int_A p(x,y)dy + r(x)\delta_x(A)]dx = \tag{4.23}\\ \int_A [\int_E \lambda(x)p(x,y)dx]\,dy + \int_A \lambda(x)r(x)dx = \tag{4.24}\\ \int_A [\int_E \lambda(y)p(y,x)dx]\,dy + \int_A \lambda(x)r(x)dx = \tag{4.25}\\ \int_A [\lambda(y)\int_E p(y,x)dx]\,dy + \int_A \lambda(x)r(x)dx = \tag{4.26}\\ \int_A [\lambda(y)(1-r(y))]\,dy + \int_A \lambda(x)r(x)dx = \int_A \lambda(y)dy, \tag{4.27}\end{align}$$

then indeed, λ itself is a stationary distribution. □

In the Metropolis-Hastings algorithm for simulating from a stationary density π, a proposal (instrumental) density $q(x,y)$ for moving the chain from one point x in the state space to another y is needed.

If π and $q(x,y)$ satisfy the detailed balance condition (4.22), fine, we have the stationary distribution in π. If not, say

$$\pi(x)q(x,y) > \pi(y)q(y,x), \tag{4.28}$$

which means that the chain traverses from x to y too often (relative to moving in the opposite direction). Introduce $\alpha(x,y)$, the probability of a move from x to y getting accepted, as an additional accept-reject step (in this context "rejecting" the proposed value is regarded as keeping the previously obtained value of the chain also as the new sampled value),

$$\pi(x)q(x,y)\alpha(x,y) = \pi(y)q(y,x)\alpha(y,x), \tag{4.29}$$

and require that the chain always moves from y to x by setting $\alpha(y,x) = 1$.

Then, choosing

$$\alpha(x,y) = \begin{cases} \min\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\}, & \text{if } \pi(x)\,q(x,y) > 0, \\ 1, & \text{if } \pi(x)\,q(x,y) = 0, \end{cases} \qquad \text{if } \pi(x)\,q(x,y) > 0,$$

creates a reversible chain with the stationary distribution π, as will be seen below. Note that the normalizing constant of $\pi(\cdot)$ cancels. Also, it is not needed (as e.g. in the Bayesian setting) for calculating the acceptance probability $\alpha(x,y)$.

The above claimed result follows, as e.g. still assuming (4.28) and recalling

$$r(y) = 1 - \int_E q(y,x)\alpha(y,x)dx$$

would give,

$$\begin{aligned} \int_E \pi(x)P(x,A)dx = \int_E \pi(x)[\int_A q(x,y) \overbrace{\alpha(x,y)}^{=\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}} dy + r(x)\delta_x(A)]dx = \\ \int_A [\int_E \pi(y)q(y,x)dx]dy + \int_A \pi(x)r(x)dx = \\ \int_A \pi(y)[\int_E q(y,x)\overbrace{\alpha(y,x)}^{=1}dx]dy + \int_A \pi(x)r(x)dx = \\ \int_A \pi(y)[1 - r(y)]dy + \int_A \pi(x)r(x)dx = \\ \int_A \pi(y)dy. \end{aligned}$$

The transition kernel of the MH-chain, reversible by construction, can hence be expressed as

$$P_{MH}(x,dy) = q(x,y)\alpha(x,y)dy + (1 - \int_E q(x,y)\alpha(x,y)dy)\delta_x(dy). \tag{4.30}$$

At the time of the generalization of the Metropolis method [Hastings, 1970], there was an alternative sampling scheme by [Barker, 1965].

Namely, if $s(x,y) = s(y,x)$ is any symmetric function, then setting

$$\alpha(x,y) = \frac{s(x,y)}{1 + \frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}} \tag{4.31}$$

works in getting the reversibility condition (4.29) fulfilled. Barker advocated the choice $s(x,y) \equiv 1$. It was only shown by Hastings' student Peskun [P.H.Peskun, 1973], that the choice (4.30) is optimal in terms of reducing the autocorrelation of the chain.

Algorithmically speaking, the Metropolis-Hastings algorithm consists of the steps

1. Generate stochastic variables $Y \sim q(X_n, y) = q(\cdot|X_n)$ and $U \sim \text{Unif}[0,1]$;
2. Accept $X_{n+1} = Y$, if $U \leq \alpha(X_n, Y)$;
3. Else, keep $X_{n+1} = X_n$;
4. Return the (correlated) sample $X_0, X_1, \ldots, X_m$.

The construction is analoguous to the Accept-Reject algorithm in chapter 3. In the MCMC-context "rejecting" just means keeping the previously sampled point.

The MH-algorithm is usually used for parameter estimation. The posterior distribution $\pi(\theta|x)$ contains all information there is about the parameter(s) θ, given the data x. As mentioned after (2.10), the sample from the posterior can be considered as a sample from the real distribution of the parameter. Therefore estimating the parameter, or any statistics of it, can be performed using this sample.

Most of the estimates are typically integrals of the type (4.8). In the parameter estimation context

$$E(h(\theta)) = \int_\Theta h(\theta)\pi(\theta|x)d\theta, \tag{4.32}$$

different functions $h(\cdot)$ give different estimates of θ, which could be hard or even impossible to obtain with other methods. For example, setting $h(\theta) = \theta$, gives the mean estimate of theta.

In the next section we give an example of applying the Gibbs sampler to a problem of parameter estimation.

4.2.2 Gibbs sampling

An important special case of the MH-algorithm introduced by [Geman and Geman, 1984] is called Gibbs sampling. On a Gibbs sampler the coordinates of the sample points in the Markov Chain are sampled one at a time in turn. This requires all the densities of the coordinates conditional on all the other coordinates to be derived in advance.

A special feature of Gibbs' is that there is no accept-reject-step at all. Instead all proposed points are automatically accepted. The Markov Chain traverses through the state space with steps parallel to the coordinate axes.

In the seminal papers [E.Gelfand et al., 1990] and [E.Gelfand and F.M.Smith, 1990] the power of Gibbs sampling was illustrated. It performed the numerical Bayesian inference with ease on normal data problems with complications in awkward posterior distributions, distributional complexity introduced by order constraints on model parameters, dimensionality problems, messy and intractable distribution arising from missing data, general functions on model parameters and awkward predictive inference. If other solution methods were available at all, they often required sophisticated numerical or analytic expertise. In fact, these successful results led to Gibbs sampling becoming extremely popular since the 1990's for posterior simulation in a wide class of important problems.

For an example of Gibbs we look at a random effects model [E.Gelfand et al., 1990] where calculation of the marginal posteriors of variance components had previously proven a challenging technical problem,

$$Y_{ij} = \theta_i + e_{ij},\ i = 1, \dots, k,\ \ j = 1, \dots ., J. \tag{4.33}$$

Assume independence thoroughout, and assume

$$(\theta_i|\mu, \sigma_\theta^2) \sim N(\mu, \sigma_\theta^2) \tag{4.34}$$

$$(e_{ij}|\sigma_e^2) \sim N(0, \sigma_e^2). \tag{4.35}$$

Let $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, $Y = (Y_{11}, \dots, Y_{kJ})$ and assune that the parameters have priors

$$\begin{cases} \mu \sim N(\mu_0, \sigma_0^2) \\ \sigma_\theta^2 \sim IG(a_1, b_1) \\ \sigma_e^2 \sim IG(a_2, b_2). \end{cases} \tag{4.36}$$

where IG denotes the inverse gamma distribution and all hyperparameters $\mu_0, \sigma_0, a_1, b_1, a_2$ and b_2 are assumed to be known.

Then, the conditional distributions

$$\begin{cases} (\mu|Y,\theta,\sigma_\theta^2,\sigma_e^2) \sim N(\frac{\sigma_\theta^2\mu_0+\sigma_\theta^2\sum\theta_i}{\sigma_\theta^2+k\sigma_\theta^2}, \frac{\sigma_\theta^2\sigma_\theta^2}{\sigma_\theta^2+k\sigma_\theta^2}) \\ (\sigma_\theta^2|Y,\mu,\theta,\sigma_e^2) \sim IG(a_1+\frac{1}{2}k, b_1+\frac{1}{2}\sum_i(\theta_i-\mu)^2) \\ (\sigma_e^2|Y,\mu,\theta,\sigma_\theta^2) \sim IG(a_2+\frac{1}{2}kJ, b_1+\frac{1}{2}\sum_i\sum_j(Y_{ij}-\theta_i)^2) \\ (\theta|Y,\mu,\sigma_\theta^2,\sigma_e^2) \sim N(\frac{J\sigma_\theta^2}{J\sigma_\theta^2+\sigma_e^2}\bar{Y}+\frac{\sigma_e^2}{J\sigma_\theta^2+\sigma_e^2}\mu\bar{\mathbb{1}}, \frac{\sigma_\theta^2\sigma_e^2}{J\sigma_\theta^2+\sigma_e^2}I), \end{cases} \tag{4.37}$$

where $\bar{Y} = (\bar{Y_1},\ldots,\bar{Y_k})$, $\bar{Y_i} = \sum_j Y_{ij}/J$, $\bar{\mathbb{1}}$ is a $k \times 1$ comlumn vector of $1'$s and I is a $k \times k$ identity matrix, specify a Gibbs sampler for estimating the hyperparameters.

In (4.37) a_i or b_i can be set equal to 0 to represent improper[3] priors.

While performing the Gibbs sampling, the calculated quantities will be exploited to the maximum extent, which makes the updating scheme in this case look like

$$\begin{cases} \mu^{(i+1)} = (\mu|Y,\theta^{(i)},\sigma_\theta^{2^{(i)}},\sigma_e^{2^{(i)}}) \\ \sigma_\theta^{2^{(i+1)}} = (\sigma_\theta^2|Y,\mu^{(i+1)},\theta^{(i)},\sigma_e^{2^{(i)}}) \\ \sigma_e^{2^{(i+1)}} = (\sigma_e^2|Y,\mu^{(i+1)},\theta^{(i)},\sigma_\theta^{2^{(i+1)}}) \\ \theta^{(i+1)} = (\theta|Y,\mu^{(i+1)},\sigma_\theta^{2^{(i+1)}},\sigma_e^{2^{(i+1)}}). \end{cases} \tag{4.38}$$

While the original tailored analyses of this problem had suffered from "badly behaving" data with extreme posteriori skewness and standard ANOVA-methods resulting in a negative variance estimate of σ_θ^2, rendering inference of it difficult, the Gibbs sampling was easy to implement and solved the problem with ease. Also, when shifting the inferential interest e.g. from σ_θ^2 and σ_e^2 to $\frac{\sigma_\theta^2}{\sigma_e^2}$ or $\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_e^2}$ some of the other methods required substantial effort or even beginning the analysis anew while the sample-based Gibbs accomplished the shift of focus with essentially no further computational effort. See [E.Gelfand et al., 1990] for the details.

Large part of the popularity of Gibbs' sampling lends itself to the fact that in such hierarchical models it is often possible to express the conditional densities with well-known distributions, leading to efficient random variable generation. If this is not the case, other MCMC-methods, such as the MH-algorithm may be more appropriate for use.

[3] in many cases the sum or integral of an prior distribution need not even be finite, yet it can give sensible posteriori probabilities. This is called an *improper prior*.

There is a Windows-software called WinBUGS free of charge, http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml, for the task of Gibbs sampling.

A good introduction on Gibbs can be found in [Casella and I.George, 1992]. We will in the sequel only focus on the MH-sampling.

4.2.3 Convergence Diagnostics

While the convergence of the MCMC-chain usually is expected, verifying it in practice could be easier said than done.

An important result from Tierney says:

Theorem 4.16. *[Tierney, 1994] An irreducible Markov chain P with stationary distribution π, such that*

$$\pi P = \pi$$

is positive recurrent and the unique stationary distribution of P is π. If P is also aperiodic, then the convergence result

$$\delta(P^n(x_0, \cdot), \pi) \overset{def}{=} \sup_{B \in F} |P^n(x_0, B) - \pi(B)| \to 0, \tag{4.39}$$

for the total variation distance δ holds. Here B is any Borel set in filtration F of the probability space for $\pi-$almost all x_0. If P is Harris recurrent[4], then the convergence holds for all starting points x_0.

The sample obtained from a MH-simulation is usually not independent, but the successively accepted sample points from the Markov-chain are autocorrelated. The effect of this can be reduced by *thinning* i.e. subsampling from the chain. Theoretical considerations such as the strong law of large numbers and central limit theorem guarantee the ergodicity and convergence when sampled to infinity, but an annoying practical problem is to decide how large a finite sample of points is large enough for a reasonably reliable estimation of the desired quantities. The same goes for deciding the length of the initial *burn-in* period: how many iterations are needed before the chain can be considered having stabilized itself well enough in order to commence sampling "for real"?

[4] A chain is *Harris recurrent*, if for each Borel set B with $\pi(B) > 0$: $P(X_n \in B \quad \text{i.o.}) = 1 \quad \forall x_0$, where $X_n = P^n(x_0, B)$. Harris recurrency essentially eliminates any measure-theoretic pathologies.

Assuring convergence requires performing some statistical analysis. When analyzing a sampler, there are two aspects to consider [Castelloe and Zimmerman, 2002]:

- Are the sample points generated coming from the correct distribution?
- Has the entire parameter space been traversed?

A generally accepted strategy has been to run several chains from overdispersed starting values. If at some point the samples seem to have been generated approximately from the same distribution, then this distribution would be accepted as the presumably correct one. This assumption is of course justified for a properly designed Markov chain. If the starting values have been properly overdispersed, then it is also likely that the parameter space has become thoroughly traversed as well.

A review of the practical implementation of 13 convergence analysis methods is given in [Cowles and Carlin, 1996]. The mathematics of such methods is studied in [Brooks and Roberts, 1998]. A key thing to remember, as with all statistical procedures, is that any method cannot give a guarantee for successfully having diagnosed convergence. If in particular the chain *mixes* slowly the diagnostics are very likely to be unreliable since all convergence conclusions would be based on only a small region of the state space having been examined. Therefore it is always strongly recommendable to try to confirm oneself of the convergence by using additional methods for different diagnostics.

Multiple simulated sequences starting from an overdispersed estimate of the target distribution are used in [Gelman and Rubin, 1992] and the analysis resembles the regular ANOVA. At convergence the chain should originate from the same distribution. Convergence is assessed by using a conservative Student $t-$ distribution with a scale parameter containing both the between-chain (B) and within-chain (W) variances and then observing the factor by which the scale parameter shrinks [Cowles and Carlin, 1996],

$$\sqrt{(\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W})\frac{df}{df-2}}, \tag{4.40}$$

if sampling would be continued infinitely. Here m is number of parallel chains, n is number of iterations (of last $2n$) and df is the degrees of freedom of the t-distribution. Slowly mixing samples will initially have much larger B than W since the starting points are overdispersed relative to the target density. Gelman and Rubin base their convergence criteria on monitoring when the shrink factor has come close to 1.

Another method based on spectral analysis is the one of Geweke's [Geweke, 1991]. If the mean of some function g of parameter θ is estimated after each Gibbs iteration (as usual) with

$$E[g(\theta)] = \bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(\theta^{(i)})$$

and the spectral density $S_g(\omega)$ exists with no discontinuities at zero, then the asymptotic variance is $\frac{1}{n}S_g(0)$. Geweke monitors $\bar{g}(\theta^{(i)})_n^A$ based on the first n_A and $\bar{g}(\theta^{(i)})_n^B$ based on the last n_B iterations and if for example $\frac{n_A}{n} = 0.1$ and $\frac{n_B}{n} = 0.5$, then

$$(\bar{g}_n^A - \bar{g}_n^B)/(\frac{1}{n_A}S_g(0)^A + \frac{1}{n_B}S_g(0)^B)$$

should follow the standard normal distribution according to the central limit theorem.

For still 11 more alternatives of convergence diagnostics the interested reader is referred to [Cowles and Carlin, 1996] and the references therein.

4.2.4 Mixing and Adapting Proposals

Another potential problem related to convergence speed for a successful MCMC simulation might be caused by a lack of the chain to *mix* well. Mixing of the chain means the ability for it to explore the actual state space rapidly enough to produce meaningful results. If the posterior density function has a lot of multimodality, lack of effective mixing may become a real problem (see figure (4.2)) especially in a higher dimensional space. The chain must be able to jump efficiently out from areas close to local maxima of probability mass [O.Talton et al., 2011].

Adapting the proposal distribution to produce well-mixing proposal points relative to the target distribution is an art in itself. In [S.P. Brooks and Roberts, 2003] the authors study finding good proposals in the even more challenging reversible jump setting, where the connection between the different submodel spaces may lack obvious geometric intuitive relation such as Euclidian structure.

A robust method for creating a MCMC-sampler is the random walk MH. The new proposed steps can be drawn e.g. from a normal distribution and the acceptance procedure performed according to (5.11). For achieving a good mixing it is essential that the proposal step length agrees well with the shape and dimensionality of the target distribution (see figures 4.1 and 4.2 for what can go wrong even in the simplest setting and with multimodality). The random walk MCMC usually works and is easy to implement, but the convergence can be very slow with increasing complexity of the target distribution.

There are many variants where information on target derivatives is taken into account with harder-to-implement sampler and much faster convergence. [S.P. Brooks and Roberts, 2003] provides an analysis of Taylor-expanding the acceptance probabilities around certain canonical jumps, which turns out to have close connections to Langevin algorithms. Langevin algorithms use gradient information about the target distribution in proposing candidate moves which are more likely to be accepted. Such a more sophisticated approach allowing more ambitious moves to be proposed and accepted can improve the efficiency of the algorithm drastically.

4.2.5 Historical Background and Simulated Annealing

One heuristics to try and obtain better mixing is related to simulated annealing and called simulated tempering [E.Marinari and G.Parisi, 1992]. The idea is to use the temperature as a dynamic variable and flatten (“melt“) the high peaks of probability mass in the density by first “warming up” the system and then carefully annealing until the system “freezes”, while always keeping it in an energy equilibrium. Since the gross features of the eventual state of the system appear in higher temperatures and fine details develop in lower temperatures, the result will be kind of an adaptive divide-and-conquer algorithm.

The original implementation in [N.Metropolis et al., 1953] was in a setting of statistical physics and involved simulating N particles on a square of R^2 in a periodic structure by calculating energy integrals such as

$$I = \int F(\theta) \exp(-\frac{E(\theta)}{kT})\, d\theta \Big/ \int \exp(-\frac{E(\theta)}{kT})\, d\theta, \tag{4.41}$$

where θ represents the state of the particles, T is the temperature, k is the Boltzmann constant,

$$E = \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} V(d_{ij}) \tag{4.42}$$

is the potential energy for a potential function V and the weighting distribution $\exp(-\frac{E(\theta)}{kT})$ is the Boltzmann distribution. The numerical difficulty caused by $\exp(-\frac{E(\theta)}{kT})$ being very small for most particle configurations was solved by a random walk modification of the earlier Monte Carlo- method. The particle coordinates were uniformly perturbed, energy difference ΔE between old and new configuration calculated and the new state was accepted, if it was on a higher level of energy. If it was on a lower level, then it would get accepted only with probability $\exp(-\frac{\Delta E}{kT})$.

The given temperature parameter T can be modified to let the system *anneal* to the desired distribution. By "heating" the system up, the local maxima in probability density flatten out and the chain may proceed easier from one part of the space (energy well) to another. Then one can let the effective temperature carefully to "cool down", while preserving the "thermal energy equilibrium".

Since there is a strong analogy to the crystallization of annealing material, this variation of the MH-algorithm is called the simulated annealing. Using a cost function in place of energy and defining configurations by the set of parameters it is straightforward to apply the MH-procedure to combinatorial optimization problems [S.Kirkpatrick and M.P.Vecchi, 1983].

In more concrete terms, if an objective function $f(\theta)$ is to be minimized over a parameter vector $\theta \in \Theta$, the corresponding Bolzmann distribution admits a density

$$b_T \propto \exp(-f(\theta)/T). \tag{4.43}$$

Then the usual MH-procedure can be run for an initial temperature T_0 until an equilibrium state has been reached. After that one can lower the temperature and repeat the MH, until some stopping criteria has been met. The final configuration will approximate the minimum of $f(\theta)$ [S.P. Brooks and R.King, 2003].

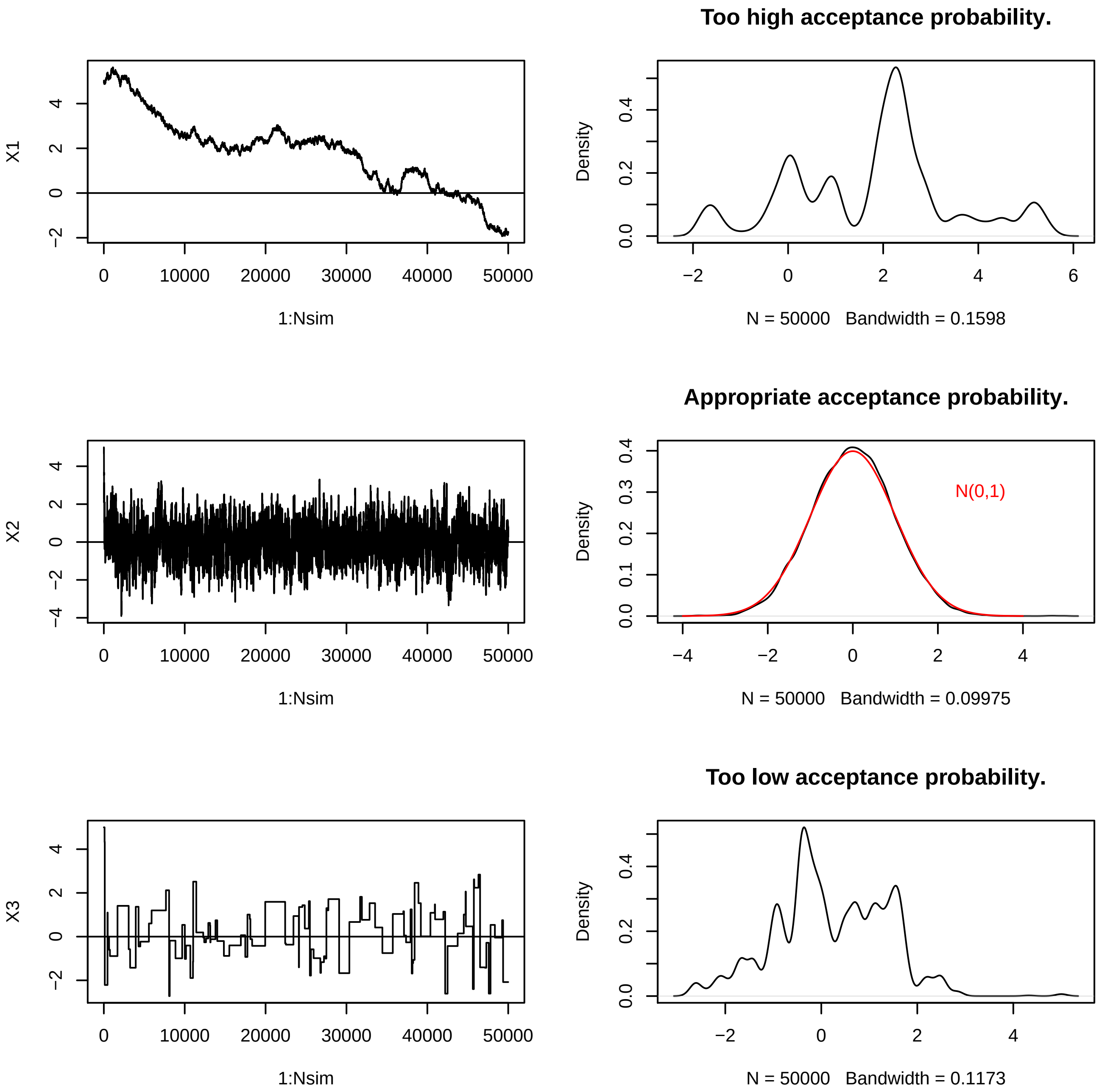

FIGURE 4.1: Three MCMC-chains for simulating the standard normal $N(0,1)-$ distribution from initial value $X = 5.0$ with the plain random walk MCMC. The proposal distributions centered at the present sample point are also normal, but with variances $\sigma_1^2 = 0.01, \sigma_2^2 = 0.25$ and $\sigma_3^2 = 600$. Hence the first chain $X1$ takes too small steps and practically every step gets accepted. The last chain $X3$ attempts to jump too far away and consequently extremely few steps become accepted whence convergence is being very slow. The chain in the middle $X2$ with appropriate proposal step lengths relative to the target distribution is mixing well.

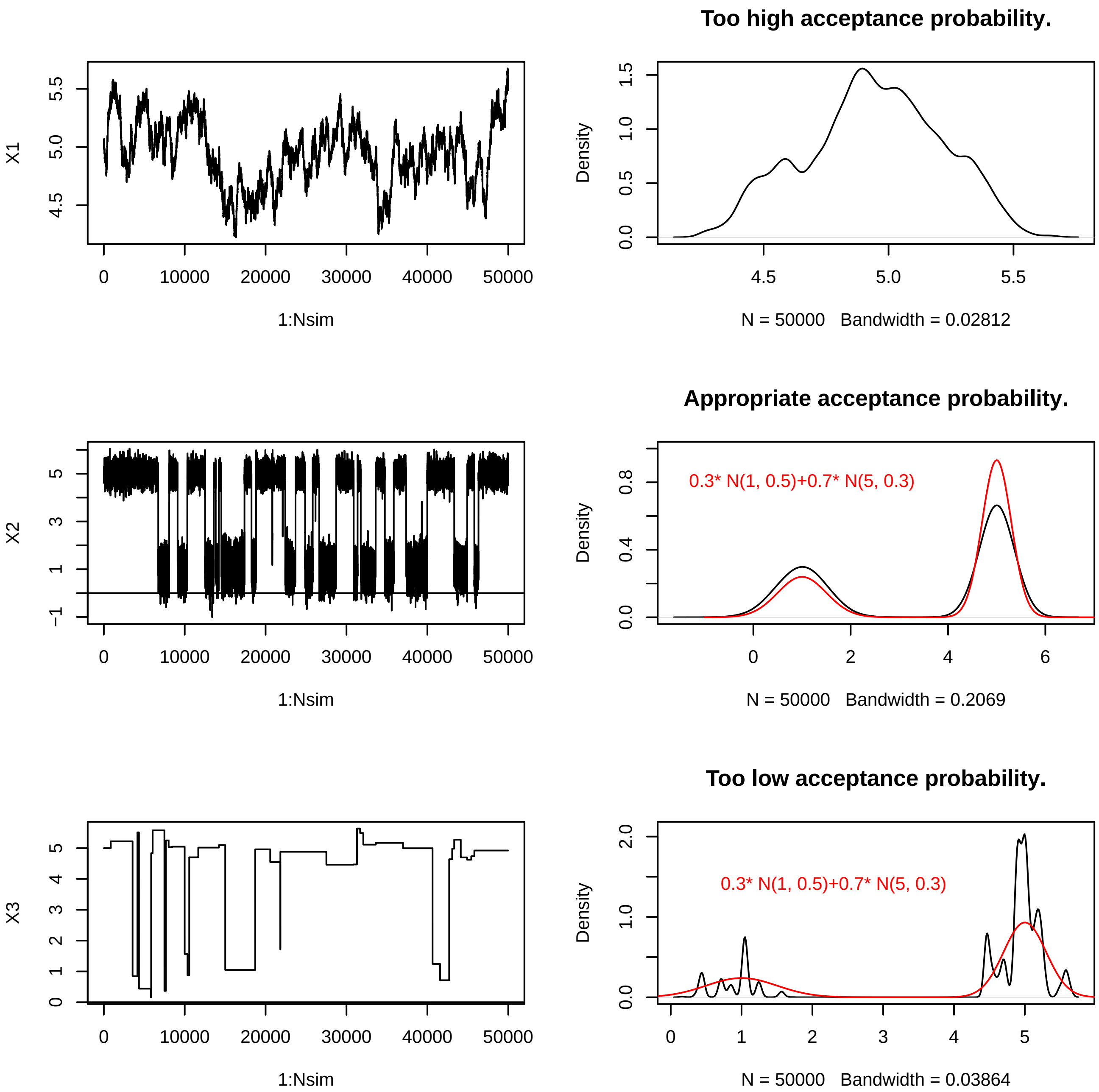

FIGURE 4.2: Additional problems in tuning the chain are caused by possible multimodality. In these pictures the target distribution is the normal mixture $0.3N(1, 0.5)+0.7N(5, 0.3)$ and we are still simulating from initial value $X = 5.0$ with the plain random walk MCMC. The proposal distributions centered at the present sample point are normal, with variances $\sigma_1^2 = 0.01, \sigma_2^2 = 0.95$ and $\sigma_3^2 = 1100$. The first chain $X1$ takes too small steps and has not yet found its way to the other mode at all in 50000 simulations. The last chain $X3$ attempts steps much too far away and the convergence is being very slow. The chain in the middle $X2$ with appropriate proposal step lengths is mixing relatively well.

Chapter 5

The Reversible Jump MCMC

> "The number of things you don't know is one of the things you don't know."
> –Peter J. Green

5.1 General Reversible Jump Theory

The celebrated paper [Green, 1995] of Peter Green's extends the MH-algorithm from model parameter fitting into a model choosing methodology. The setup of the Reversible Jump Markov Chain Monte Carlo (RJMCMC) is ideally suited for comparing different dimensional parametric models for the data.

In the reversible jump implementation the dimension of the parameter space is allowed to vary. Jumps between different models from one subspace to another are made possible as MH-moves. Hence the number of parameters in the model becomes a subject of inference itself. The posteori probabilities of the concurring submodels can be simply estimated by running the RJMCMC-simulation and checking what proportion of time the chain spends in each different submodel.

While the other model selection methodologies, such as the many different information criteria [Akaike, 1974, Schwarz, 1978, Takeuchi, 1976, N.Murata et al., 1978, J.Spiegelhalter et al., 2002], only choose the most appropriate submodel usually in view of maximized likelihood and minimized number of model parameters, the RJMCMC is able to output directly the *a posteriori* probabilities of the submodels and the *a posteriori* distributions of the parameters in the submodels.

The main technical difficulty of jumping between the models is finding a bijective diffeomorphism[1] and attaining the detailed balance between model spaces of possibly different dimensions. To accomplish this, either the spaces can be augmented into a larger-dimensional space, or continuous random variable vectors can be generated to fill in the missing dimensions.

Let us index the model space $M = \bigcup M_k$ with k. In practice it may often be hard to find "natural" proposal moves between two different subspaces. Let us denote the n_k-dimensional $(k = 1, 2)$ model parameter vectors with $\theta^{(k)}$ and continuous m_k-dimensional stochastic variable vectors with $u^{(k)}$. A bijection is needed between the two vectors $M_1 : (\theta^{(1)}, u^{(1)})$ and $M_2 : (\theta^{(2)}, u^{(2)})$.

Obviously, the dimension matching requirement

$$n_1 + m_1 = n_2 + m_2 \tag{5.1}$$

needs to get fulfilled.

In practical implementations (assuming for the moment $n_1 < n_2$) the simplification $m_2 = 0$ can often be employed. If the chain happens to be in state $(1, \theta^{(1)})$, in order to jump from M_1 to M_2 we need to generate a stochastic vector $u^{(1)}$ of dimension m_1 and establish a function $g : R^{n_1} \times R^{m_1} \to R^{n_2}$ such that $\theta^{(2)} = g(\theta^{(1)}, u^{(1)})$. For the inverse jump back from M_2 to M_1 we only need to find the deterministic inverse move function $h(\cdot) = g^{-1}(\cdot)$ for solving $\theta^{(1)} = h(\theta^{(2)})$.

From now on we denote the submodel M_k or the corresponding state of the Markov chain with $(k, \theta^{(k)})$.

Example In spite of the notation getting slightly awkward, we use the letter η for more concrete parameter values. Let us say for simplicity that there are only two submodels of dimensions one and two: $M_1 : (1, \theta^{(1)}) = (1, \eta)$ and $M_2 : (2, \theta^{(2)}) = (2, (\eta_1, \eta_2))$. To make the jump $M_1 \to M_2$ we could for example draw a stochastic variable u and set

$$\begin{cases} \eta_1 = \eta - u \\ \eta_2 = \eta + u. \end{cases} \tag{5.2}$$

[1] *A diffeomorphism* is a bijective differentiable mapping from a manifold to another such that the inverse mapping is differentiable as well. Of course, this requires that the two manifolds are of coinciding dimensions.

In order to make the jump back $M_2 \to M_1$ there is no need for randomness, just solve

$$\eta = \frac{1}{2}(\eta_1 + \eta_2). \tag{5.3}$$

Thus the missing dimensions were filled in with random elements as the reversible jump mapping was applied when going into the higher dimension and a plain deterministic function worked fine as the inverse move when returning back to the lower dimension. A thing to remember in designing the jump-reverse jump- pair to be consistent, is that the proposal probability must have a density with respect to a singular measure in $\Re \times \Re^2$ placing all probability mass on $\{(\eta, \eta_1, \eta_2) : \eta = \frac{1}{2}(\eta_1 + \eta_2)\}$ instead of a Lebesgue measure on $\Re^3$.

However, since a probability transformation is being made while jumping, the usual Jacobian factor, e.g.

$$J = \left|\frac{\partial(\theta^{(2)}, u^{(2)})}{\partial(\theta^{(1)}, u^{(1)})}\right| \overset{m_2=0}{=} \left|\frac{\partial(\theta^{(2)})}{\partial(\theta^{(1)}, u^{(1)})}\right| \tag{5.4}$$

becomes necessary as a factor in the expression for the MH-move acceptance probability.

Letting the chain jump between the different submodels improves the mixing properties of it.

At each state, all kind of moves are not necessarily available and available moves are assumed to exist in such invertible pairs where the detailed balance conditions are accomplished. For example, if a “birth“- move creates a new object in the state space that sends the chain into a higher dimensional space, there needs to exist an opposite, corresponding ”death”- move to balance it out.

A RJMCMC-chain can evolve in two different ways: either the chain undergoes a normal MCMC-move in the present submodel space or it can make a jump move into another subspace with a differing dimension. Thus the reversible jumping can be seen as a natural extension of the MH-algorithm but the jump moves are really still MH-moves. As discussed, jumping requires two things. The Jacobian factor J of the probability transformation in the expression for move acceptance probability needs to be taken into account. Also, the jump moves must appear in appropriate pairs in order to preserve the detailed balance.

We index the set of all available moves (including the move which preserves the present submodel) with m. Let $q_m(x, dx')$ be a sub-probability measure for a move of type m to take the chain from a current state x to dx'. Also, $\sum_m q_m(x, M) \leq 1$ and $1 -$

$\sum_m q_m(x, M)$ is the probability that no change to the present state is proposed. The probability of acceptance for this move is denoted $\alpha_m(x, x')$.

The transition kernel describing the probability for the chain to move from point x into a Borel set B is

$$P(x, B) = \sum_m \int_B q_m(x, dx')\alpha_m(x, x') + s(x)I(x \in B), \tag{5.5}$$

where $I(\cdot)$ is the indicator function and

$$s(x) = \sum_m \int_M q_m(x, dx')[1 - \alpha_m(x, x')] + 1 - \sum_m q_m(x, M) \tag{5.6}$$

is the probability of not moving from x, either due to a proposed move not being accepted or due to no move being proposed in the first place. This generalizes (4.17) by including the effect of jump moves.

Then the detailed balance equation for a move between two Borel sets $A \in M$ and $B \in M$

$$\sum_m \int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') + \int_{A \cap B} \pi(dx)s(x) = \tag{5.7}$$

$$\sum_m \int_B \pi(dx') \int_A q_m(x', dx)\alpha_m(x', x) + \int_{B \cap A} \pi(dx')s(x'). \tag{5.8}$$

Since the “no move”- integrals cancel, the condition

$$\int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') = \int_B \pi(dx') \int_A q_m(x', dx)\alpha_m(x', x) \quad \forall m, A, B, \tag{5.9}$$

sufficiently guarantees the detailed balance.

If $\pi(dx)q_m(x, dx')$ has a finite density $f_m(x, x')$ with respect to a symmetric measure ξ on $M \times M$, then

$$\int_A \pi(dx) \int_B q_m(x, dx')\alpha_m(x, x') = \int_A \int_B \xi(dx, dx')f_m(x, x')\alpha_m(x, x') =$$
$$\int_B \int_A \xi(dx', dx)f_m(x', x)\alpha_m(x', x) = \int_B \pi(dx') \int_A q_m(x', dx)\alpha_m(x', x)$$

holds provided that

$$\alpha_m(x, x')f_m(x, x') = \alpha_m(x', x)f_m(x', x). \tag{5.10}$$

Making the acceptance probability as large as possible (in the very same manner as within the plain MH-chain (4.30)), while still retaining the detailed balance

$$\alpha_m(x, x') = \min\{1, \frac{f(x', x)}{f(x, x')}\} := \min\{1, A\}, \tag{5.11}$$

reduces the autocorrelation of the realized chain as the results of [P.H.Peskun, 1973] show.

Remark 5.1. If the acceptance probability for a certain jump move is

$$\alpha(x, x') = \min\{1, A\},$$

then the acceptance probability for the opposing move retaining the detailed balance will be

$$\alpha(x', x) = \min\{1, \frac{1}{A}\}. \tag{5.12}$$

For a concrete formula of the acceptance probability between to different models consider just two subspaces $C_1 = \{1\} \times \Re$ and $C_2 = \{2\} \times \Re^2$ of the whole space $C = C_1 \cup C_2$ and a move of type q_m that always shifts the subspace.

If $A \subset C_1$ and $B \subset C_2$, set

$$\xi(A \times B) = \xi(B \times A) = \tag{5.13}$$

$$\lambda\{(\theta^{(1)}, u_1) : \theta^{(1)} \in A, \theta^{(2)}(\theta^{(1)}, u_1) = g(\theta^{(1)}, u_1) \in B\} \tag{5.14}$$

where λ is $(n_1 + m_1)-$dimensional Lebesgue measure. To obtain the required symmetric measure for any Borel sets $A, B \subset C$, put

$$\xi(A \times B) = \xi\{(A \times C_1) \times (B \times C_2)\} + \xi\{(A \times C_2) \times (B \times C_1)\}. \tag{5.15}$$

The context may suggest that a move from model two, $M_2 : (2, (\theta_1, \theta_2))$ into model one, $M_1 : (1, \theta)$ with $\theta = \frac{1}{2}(\theta_1 + \theta_2)$ might be a good idea. Then the equilibrium joint proposal probability

$$\int_B \pi(dx) \int_A q_m(x, dx')$$

must have a density with respect to a singular probability measure with all its probability mass concentrated on the set $\{(\eta, \eta_1, \eta_2) : \eta = \frac{1}{2}(\eta_1 + \eta_2)\}$.

Assume that the parameters $\theta^{(1)} = \eta$ and $\theta^{(2)} = (\eta_1, \eta_2)$ have proper densities $p(1, \theta^{(1)})$ in $\Re^{n_1} = \Re$ and $p(2, \theta^{(2)})$ in $\Re^{n_2} = \Re^2$. The probability of choosing a move to the other subspace is denoted by $j(\cdot)$ and the densities of random vectors u by $q(\cdot)$.

For points $x = (1, \theta^{(1)}) \in C_1$, $x' = (2, \theta^{(2)}) \in C_2$ and data y set

$$f(x, x') = p(1, \theta^{(1)}|y)j(1, \theta^{(1)})q_1(u^{(1)}) \tag{5.16}$$

and

$$f(x', x) = p(2, \theta^{(2)}|y)j(2, \theta^{(2)})q_2(u^{(2)}) \left|\frac{\partial(\theta^{(2)}, u^{(2)})}{\partial(\theta^{(1)}, u^{(1)})}\right|. \tag{5.17}$$

Then $f(x, x')$ is the density with respect to ξ of the equilibrium joint proposal distribution $\pi(dx)q(x, dx')$.

The acceptance probability (5.11) for a move from $x = (1, \theta^{(1)})$ to
$x = (2, \theta^{(2)})$ is given by

$$\alpha_m = \min\{1, \frac{p(2, \theta^{(2)}|y)j(2, \theta^{(2)})q_2(u^{(2)})}{p(1, \theta^{(1)}|y)j(1, \theta^{(1)})q_1(u^{(1)})} \left|\frac{\partial(\theta^{(2)}, u^{(2)})}{\partial(\theta^{(1)}, u^{(1)})}\right|\} \tag{5.18}$$

and if we set $m_2 = 0$,

$$\alpha_m = \min\{1, \frac{p(2, \theta^{(2)}|y)j(2, \theta^{(2)})}{p(1, \theta^{(1)}|y)j(1, \theta^{(1)})q_1(u^{(1)})} \left|\frac{\partial(\theta^{(2)})}{\partial(\theta^{(1)}, u^{(1)})}\right|\}. \tag{5.19}$$

5.1.1 Convergence Diagnostics in RJ

While diagnosing the convergence of a fixed-dimensional Markov chain is a hard problem, as discussed in chapter 4, only more complications are to be expected when one begins jumping between submodels.

Reversible jumping, that involves changing the dimension of the parameter space from one dimension to another, does certainly not make diagnosing the accomplished convergence to the stationary distribution any easier. However, Peter Green says: “The degree of confidence that convergence has been achieved provided by ’passing’ a diagnostic convergence test declines very rapidly as the dimension of the state space increases. In a more than, say, a dozen dimensions, it is difficult to believe that a few, even well-chosen, scalar statistics give an adequate picture of convergence of the multivariate distribution. It is high, rather than variable, dimensions that are the problem” [Green, 2001].

Few diagnostics methods for reversible jumping are available. Castelloe and Zimmermann

[Castelloe and Zimmerman, 2002] provide a method by extending the method of [Gelman and Rubin, 1992] into the reversible jumping (to encompass all model/ parameter spaces) and multivariate setup (to monitor several parameters simultaneously) for doing this. Their method detects

1. variation between chains (like in the Gelman Rubin diagnostic: non-homogeneous variation across the chains)
2. interaction between models and chains (between-model variation different from chain to chain)
3. significant differences in the frequencies of changing model (jumping)

The somewhat technical details are left out to be found in the original paper.

Also in [S.A.Sisson and Y.Fan, 2007] the convergence of a trans-dimensional (reversible jump) chain is examined by a distance-based method.

5.2 Coal Mining Disasters

A Poisson counting process data is well suited material to be modeled with the RJMCMC-algorithm. In the following example we model the arrival rate of coal mine accidents with a piecewise constant step function, where the number of change points is a priori unknown.

The number of days between 191 explosive coal mine accidents in Great Britain resulting in 10 or more casualties recorded during 112 years (15 March 1851 -22 March 1962) are given in the data corrected from the original [B.A.Maguire and A.H.A.Wynn, 1952] by [R.G.Jarrett, 1979]. Later it was noticed that the early year of 1851 and rest of 1962 were free from accidents

[A.E.Raftery and V.E.Akman, 1986] and two points were added to the data set. The data set “coal“ is available free of charge (in a slightly different format: accident times are given as decimal numbers representing years) in library “boot“ on software R [version 3.0.1 (2013-05-16) [Core Team, 2013]].

Following [Green, 1995] and also trying to fill in some more details, we model the rate of accidents as a Poisson process with a piecewise constant intensity function. In other words, the accident rate is assumed to be a step function with k change points and $k+1$ Poisson intensity values in time interval $t \in [0, L = 40907]$ (days). The dimensional parameter k is *a priori* unknown, hence also subject to inference. In Green's words: "The number of the things you don't know is one of the things you don't know."

In (figure 5.1) the occurrence of accidents has been presented in different ways: as a jitter plot, by count and by rate.

The rate points have been estimated from averages of 14 successive data points for the visual presentation only, not out of the necessity from the RJMCMC-sampler.

Definition 5.2. The counting process $N(t)$ with intensity $\lambda(t) > 0$, $(t > 0)$ is called a **non-homogeneous Poisson process** if

- $N(0) = 0$
- $\{N(t)\}_{t \geq 0}$ has independent increments
- $P(N(t+h) - N(t) = 1) = \lambda(t)h + o(h)$
- $P(N(t+h) - N(t) > 1) = o(h)$.

Define the *cumulative event rate* of the process in interval $[t, t+s)$ as

$$\Lambda(t, s) = \int_t^{t+s} \lambda(u)\, du.$$

Then

$$P(N(t+s) - N(t) = k) = \exp(-\Lambda(t,s)) \frac{\Lambda(t,s)^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{5.20}$$

If an event happens at time t and τ_t is the waiting time for the next event, then the cumulative distribution function for τ_t is

$$F_t(x) = P(\tau_t \leq x) = 1 - P(\tau_t > x) = \tag{5.21}$$

$$1 - P(N(t+x) - N(t) = 0) = \tag{5.22}$$

$$1 - \exp(-\int_t^{t+x} \lambda(u)\, du) = 1 - \exp(-\Lambda(t, t+x)). \tag{5.23}$$

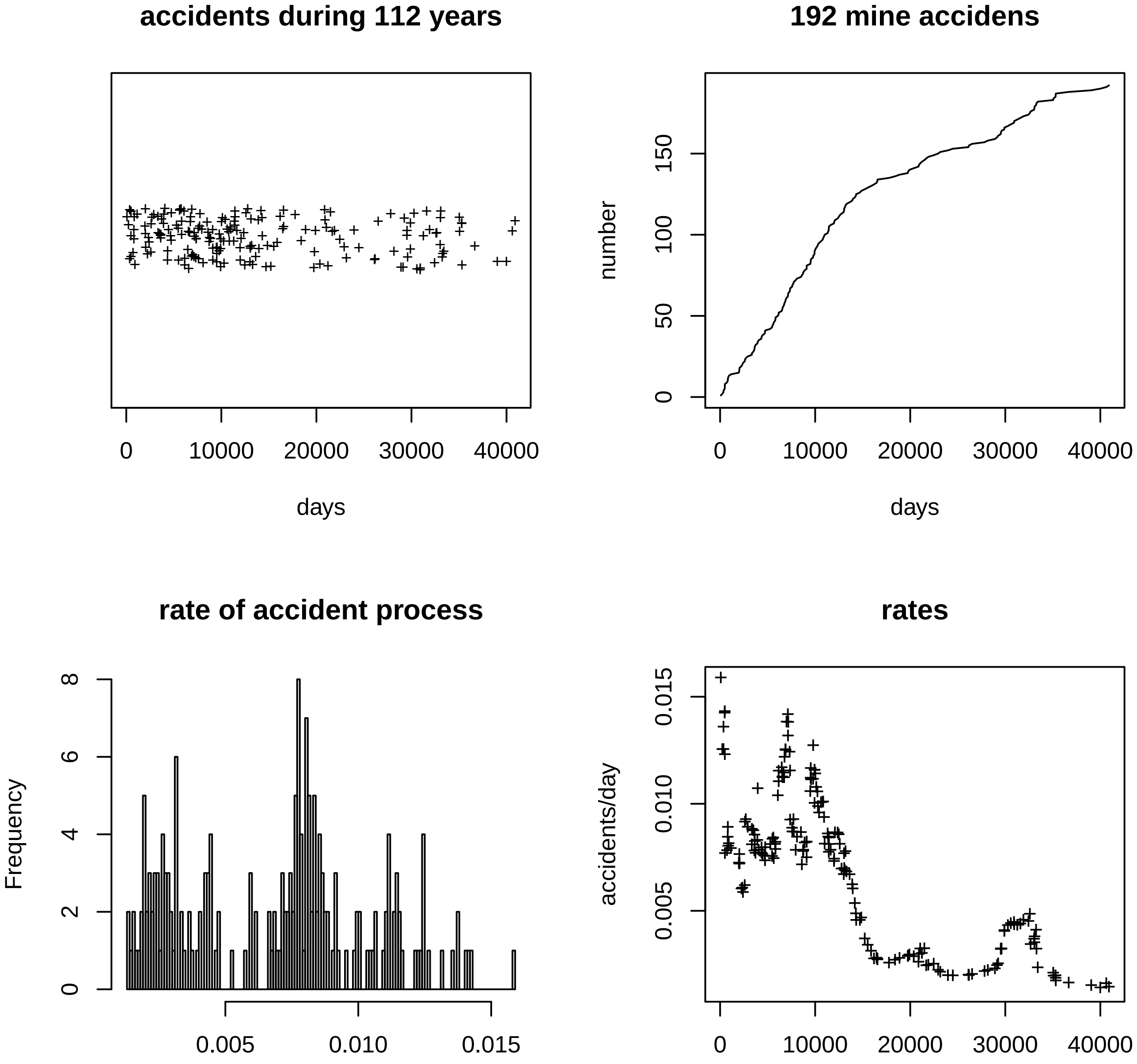

FIGURE 5.1: Fatal coal mine accidents in the UK during years 1851-1962.

The density function for the waiting time will be

$$f_t(x) = \frac{dF_t(x)}{dx} = \lambda(t+x)\exp(-\int_t^{t+x} \lambda(u)\,du). \tag{5.24}$$

Considering the joint density for observing data $D = \{y_1, y_2, \ldots, y_n\}$ in the interval $[0, L]$,

$$f(y_1,\ldots,y_n) = f_0(y_1)f_{y_1}(y_2-y_1)\cdot\ldots\cdot f_{y_{n-1}}(y_n-y_{n-1})[1-F_{y_n}(L-y_n)] \tag{5.25}$$

gives the likelihood for observing times $y_1, \ldots, y_n$ with intensity $\lambda(\cdot)$ as

$$L(y_1,\ldots,y_n|\lambda(\cdot)) = \prod_{k=1}^{n} \lambda(y_k)\exp(-\int_0^L \lambda(\xi)\,d\xi). \tag{5.26}$$

Hence the log-likelihood (in a constant dimension k) will become

$$\sum_{i=1}^{n} \log(\lambda(y_i)) - \int_0^L \lambda(\xi)d\xi. \tag{5.27}$$

By (5.20)

$$P(N(L)=n) = \frac{\Lambda(0,L)^n}{n!}\exp(-\Lambda(0,L)). \tag{5.28}$$

If the joint density is conditioned on $N(L) = n$, the conditioned density is

$$f_n(y_1,\ldots,y_n) = f(y_1,\ldots,y_n|N(L)=n) = \tag{5.29}$$

$$\frac{f(y_1,\ldots,y_n)}{P(N(T)=n)} = \frac{n!}{\Lambda(0,L)^n}\prod_{i=1}^{n}\lambda(y_i), \tag{5.30}$$

which is the same expression as one gets for the order statistics of n event points with intensity $\lambda(\cdot)$ distributed in the interval $[0, L]$.

If there is no accident in a short interval $[y_{j-1}, y_j)$, there is practically no penalty in the RJMCMC-algorithm for adding change points in the step function model of $\lambda(\cdot)$ in the interval. This does not really reflect the behaviour of data and, to avoid this unrealistic behaviour, we set a penalty for entering into a higher dimension by dividing with factor $n!$ in the likelihood expression (5.26). Consequently, the log-likelihood to be used in the

R-program [Core Team, 2013] implementation will look like

$$\sum_{i=1}^{n} \log(\lambda(y_i)) - \int_0^L \lambda(\xi) d\xi - \sum_{j=1}^{n} j. \tag{5.31}$$

This should be reasonable since the likelihoods in differing dimensions cannot be directly comparable to each other. Also, the idea of giving penalty for higher number of model parameters is in full agreement with different information criteria (such as Akaike's AIC [Akaike, 1974]).

Green only gives the log-likelihood (5.27), and does not comment on comparing likelihoods between different dimensions k. However, the last term in (5.31) was essential in making the R-program run smoothly also while reversible jumping.

In calculating the *a priori* likelihood it is assumed that the number of steps in the step function describing $\lambda(t)$ is Poisson distributed with $\mu = 3.0$. Step function heights are gamma distributed with parameters $\alpha = 1$ and $\beta = 200$. Hence the prior density function for the step height becomes

$$f(h) = \frac{\beta^\alpha}{\Gamma(\alpha)} h^{\alpha-1} \exp(-\beta h). \tag{5.32}$$

The step positions could be assumed to be uniformly distributed, a priori. However, in order to avoid too short intervals possibly containing no accident data and hence not really supporting the model a posteori, the k actual points are chosen as the ones with an even index from the order statistics of $(2k + 1)$ uniformly distributed points in the interval $[0, L]$. As Green points out, this has the effect of probabilistically spacing out the step positions.

The likelihood for step point positions thus becomes

$$\frac{(1+2k)!}{L^{1+2k}} \prod_{j=1}^{k+1} (s_j - s_{j-1}) \mathbb{1}_{\{0=s_0<s_1<s_2<\ldots<s_k<s_{k+1}=L\}}, \tag{5.33}$$

here $\mathbb{1}$ is the indicator function taking care of the change point ordering.

5.2.1 Implementation

In the sampling there are four different types of proposed moves available (figure 5.2). The step function can be altered in four different ways. First the height in one interval can be adjusted (H). Secondly the position of a change point can be shifted along the time axis (P). These moves do not influence the number of change points - the dimension of the model. The simplest approach is to restrict oneself to H- and P-moves only in a fixed dimension k. The result of this will be a regular MCMC-sampler without reversible jumping.

The two other types of moves, the birth- and death-moves, to be denoted by the present number of change points, change the number of change points into the adjacent dimension: a birth-move introduces a new change point and the dimension increases by one while a death move has the opposite effect by removing a change point and the dimension gets reduced by one.

The countable set $\{H, P, 0, 1, 2, \ldots\}$ denotes change of height or position move or a birth-death move pair between dimensions $m \leftrightarrow (m+1)$. At each step from state k with k change points, one of the (at most) four available moves $(H, P, k, k-1)$ representing height or position change, birth or death move is attempted with corresponding probability η_k, π_k, b_k, d_k. Of course, $b_{k_{max}} = 0$ if an upper limit k_{max} to the number of steps is to be preassigned, $d_0 = \pi_0 = 0$ and also $\eta_k + \pi_k + b_k + d_k = 1$. If $k \neq 0$, then set $\eta_k = \pi_k$.

The probabilities of birth and death are chosen from

$$b_k = c\min[1, \frac{p(k+1)}{p(k)}] \quad \text{and} \quad d_{k+1} = c\min[1, \frac{p(k)}{p(k+1)}] \tag{5.34}$$

with c as large as possible, subject to $b_k + d_k \leqslant 0.9$ for $k = 0, \ldots, k_{max}$.

Since the a priori model dimension is assumed to be Poisson distributed,

$$p(k) = \exp(-\mu)\frac{\mu^k}{k!}, \tag{5.35}$$

the ratios in birth and death probabilities b_k, d_k simplify to

$$\frac{p(k+1)}{p(k)} = \frac{\mu}{k+1} \text{ and } \frac{p(k-1)}{p(k)} = \frac{k}{\mu}. \tag{5.36}$$

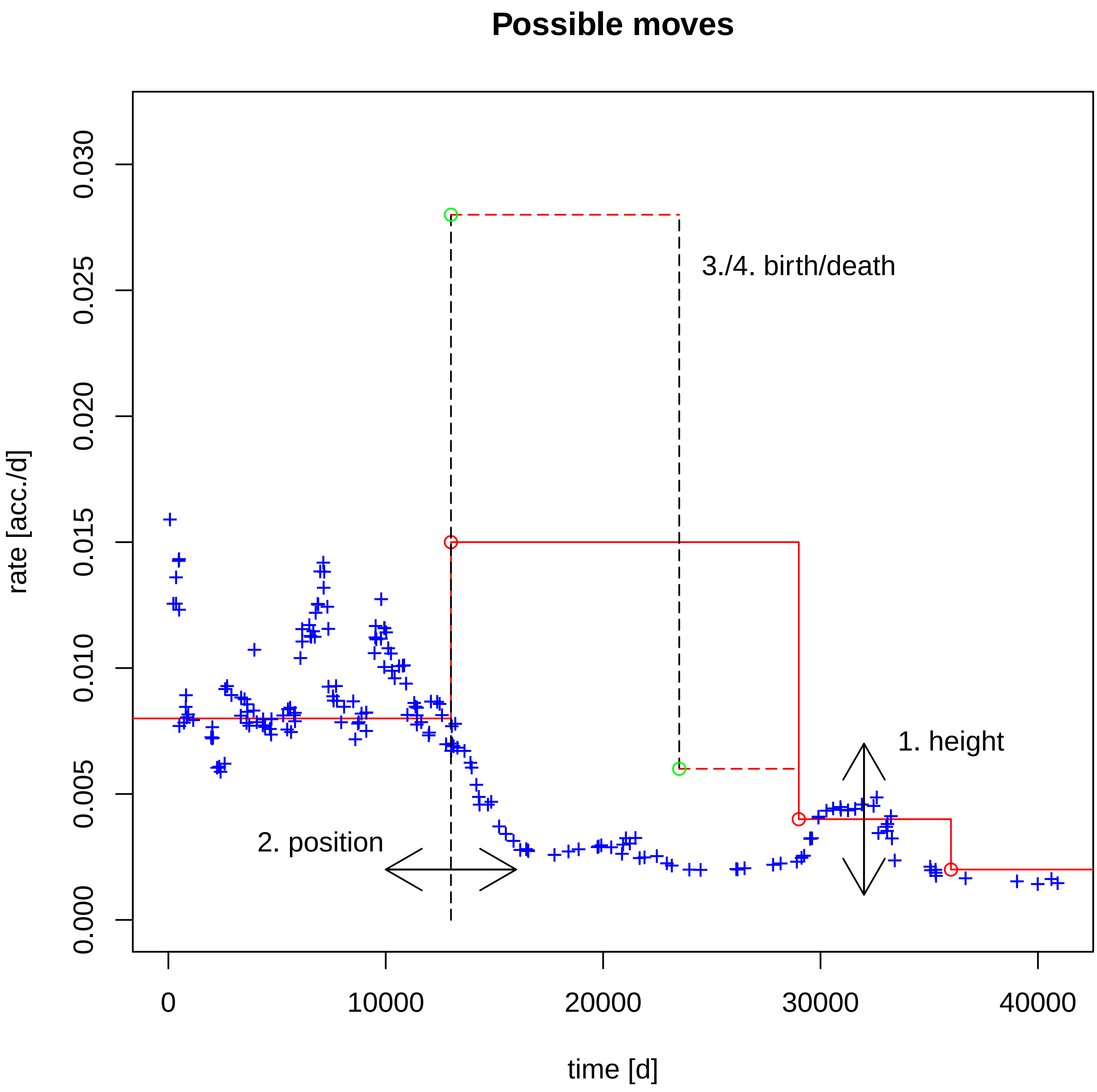

Figure 5.2: The four different step function move types. In the original step function candidate (red line), the number of change points $k = 3$. If a birth step would become accepted (red broken line), a new change point would appear and consequently $k = 4$ in the new model. The blue points represent the real data from the coal mine disaster problem.

These conditions for the probabilities should guarantee quite good an interdimensional mixing. Recall that even though the condition $b_k + d_k \leqslant 0.9$ causes a lot of attempts for a dimension change (Table 5.1), quite few of them will actually succeed in view of getting rejected in the subsequent MH-step.

TABLE 5.1: Height, position, birth and death move probabilities.

k	0	1	2	3	4	5	6	7	8	9	10
η_k	0.486	0.157	0.0714	0.05	0.0886	0.114	0.133	0.146	0.157	0.166	0.173
π_k	0	0.157	0.0714	0.05	0.0886	0.114	0.133	0.146	0.157	0.166	0.173
b_k	0.514	0.514	0.514	0.386	0.309	0.257	0.220	0.193	0.171	0.154	0.140
d_k	0	0.171	0.343	0.514	0.514	0.514	0.514	0.514	0.514	0.514	0.514
$b_k + d_k$	0.514	0.686	0.857	0.9	0.823	0.771	0.735	0.707	0.686	0.669	0.655

5.2.1.1 Height Step

First an interval of the existing $k+1$ heights is chosen at random. Then the new height $h_j^{'}$ is chosen so that $\log(\frac{h_j^{'}}{h_j}) = u \sim \text{Unif}\,(-\frac{1}{2}, \frac{1}{2})$, or

$$h_j^{'} = h_j \exp(u) \approx h_j \times \; [0.61; 1.65], \tag{5.37}$$

the bracket notation representing the interval of the random number $\exp(u)$. The acceptance probability for a height move becomes

$$\alpha_{height} = \min[1, \frac{p(y_1, y_2, \ldots, y_k|\lambda^{'})}{p(y_1, y_2, \ldots, y_k|\lambda)} \frac{h_j^{'}}{h_j} \exp(-\beta(h_j^{'} - h_j)], \tag{5.38}$$

where the first factor within the expression is the likelihood ratio of the data for the models with proposed new $(\lambda^{'}(\cdot))$ and current $(\lambda(\cdot))$ values of all parameters. The rest of the expression is the ratio of gamma-priors (5.32). The first factor is also the Bayes factor for one model relative to the other.

5.2.1.2 Position Step

A position move changes the location of a change point uniformly between the two neighbouring change points:

$$s^{\star} = u \sim \text{Unif}(s_{j-1}, s_{j+1}), \quad (j = 1, \ldots, k). \tag{5.39}$$

This influences both terms in the posteori log-likelihood (5.27). The number of accidents and the area in the integral need to be adjusted corresponding to the possible new location of the change point.

The acceptance probability for a position move due to the change of posteori likelihood becomes

$$\alpha_{position} = \min[1, \frac{p(y_1, y_2, \ldots, y_k|\lambda^{'})}{p(y_1, y_2, \ldots, y_k|\lambda)} \frac{(s_{j+1} - s^{\star})(s^{\star} - s_{j-1})}{(s_{j+1} - s_j)(s_j - s_{j-1})}]. \tag{5.40}$$

The first factor is the likelihood ratio with proposed and old models, the rest comes from the prior ratio (5.33).

5.2.1.3 Birth Step

The joint distribution of $(k, \lambda^{(k)}, y)$ can be factorized naturally into the product of model probability, prior and likelihood:

$$p(k, \lambda^{(k)}, y) = p(k)p(\lambda^{(k)}|k)p(y|k, \lambda^{(k)}). \tag{5.41}$$

Since the prior model dimension is Poisson distributed,

$$p(k) = \frac{\mu^k}{k!} \exp{(-\mu)}, \tag{5.42}$$

the prior likelihood for a particular step function becomes

$$\frac{\mu^k}{k!} \exp{(-\mu)} \frac{(1+2k)!}{L^{1+2k}} \prod_{j=1}^{k+1} (s_j - s_{j-1}) \frac{\beta^{\alpha}}{\Gamma(\alpha)} h^{\alpha-1} \exp(-\beta h). \tag{5.43}$$

The likelihood (5.43) consists of dimensional, positional and height likelihood factors corresponding to the model probability and prior in (5.41).

Here, using Bayesian calculus,

$$p(\lambda|y_1, y_2, \ldots, y_k) = \frac{p(y_1, y_2, \ldots, y_k|\lambda)p(\lambda)}{p(y_1, y_2, \ldots, y_k)} \tag{5.44}$$

and consequently

$$\frac{p(\lambda'|y_1, y_2, \ldots, y_k)}{p(\lambda|y_1, y_2, \ldots, y_k)} = \frac{p(y_1, y_2, \ldots, y_k|\lambda')}{p(y_1, y_2, \ldots, y_k|\lambda)} \frac{p(\lambda')}{p(\lambda)}, \tag{5.45}$$

it is helpful to re-write the birth move acceptance probability (5.19) as

$$\alpha_{birth} = \min\{1, (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio}) \times (\text{Jacobian})\}. \tag{5.46}$$

If a birth step gets chosen, and a new step change point $s^\star \in (s_j, s_{j+1})$ is generated from $s^\star = u_1 \in \text{Unif}(s_j, s_{j+1})$, with new heights $h_j^{'}$ and $h_{j+1}^{'}$ not completely discarding the old height h_j such that the weighted geometric average

$$(h_j^{'})^{s^\star - s_j}(h_{j+1}^{'})^{s_{j+1}-s^\star} = h_j{}^{s_{j+1}-s_j} \tag{5.47}$$

is preserved within the perturbation

$$\frac{h_{j+1}^{'}}{h_j^{'}} = \frac{1-u_2}{u_2}, \quad (\text{where} \ \ u_2 \in \text{Unif}(0,1)), \tag{5.48}$$

then the prior likelihood ratio of the new and old model becomes

$$\frac{p(\lambda')}{p(\lambda)} = \frac{p(k+1)}{p(k)} \frac{(2k+2)(2k+3)}{L^2} \frac{(s_{j+1}-s^\star)(s^\star - s_j)}{s_{j+1}-s_j} \frac{f(h_j^{'})f(h_{j+1}^{'})}{f(h_j)} =$$

$$\frac{2\mu(2k+3)}{L^2} \frac{(s_{j+1}-s^\star)(s^\star - s_j)}{s_{j+1}-s_j} \frac{\beta^\alpha}{\Gamma(\alpha)} (\frac{h_j^{'}h_{j+1}^{'}}{h_j})^{\alpha-1} \exp(-\beta(h_j^{'} + h_{j+1}^{'} - h_j)). \tag{5.49}$$

A model with k change points needs $k+1$ step function heigths. If a birth move occurs, it modifies an existing height into two new ones and a new change point (see figure (5.2)) is created.

The fact that the dimension increases by 2 in a birth move, from $2k+1$ to $2k+3$, is accounted for by the two random variables drawn from the uniform distributions.

Since the uniformly distributed stochastic variable u_1 can be thought to be drawn from the whole interval $[0, L]$ with a constant density $\frac{1}{L}$ and since a death move corresponding to d_{k+1} just removes one of the the existing $k+1$ change points with probability $\frac{1}{k+1}$, the proposal ratio from (5.19) can be written

$$\frac{j(2,\theta^{(2)})}{j(1,\theta^{(1)})q_1(u^{(1)})} = \frac{d_{k+1}L}{b_k(k+1)}. \tag{5.50}$$

Since a transformation,

$$\begin{cases} s^\star = u_1 \\ h_j^{'} = (\frac{1-u_2}{u_2})^{\frac{-w_+}{w_+ + w_-}} h_j \\ h_{j+1}^{'} = (\frac{1-u_2}{u_2})^{\frac{w_-}{w_+ + w_-}} h_j \end{cases} \tag{5.51}$$

(with notation $w_+ := s_{j+1} - s^\star$ and $w_- := s^\star - s_j$) is being made while performing a birth step, a Jacobian determinant is required as a factor in the acceptance probability α_{birth}. In order to check the correctfulness of Green's Jacobian, we calculate

$$\begin{aligned} J = \left|\frac{\partial(h_j^{'}, h_{j+1}^{'}, s^\star)}{\partial(h_j, u_1, u_2)}\right| = \begin{vmatrix} \frac{\partial h_j^{'}}{\partial h_j} & 0 & \frac{\partial h_j^{'}}{\partial u_2} \\ \frac{\partial h_{j+1}^{'}}{\partial h_j} & 0 & \frac{\partial h_{j+1}^{'}}{\partial u_2} \\ 0 & 1 & 0 \end{vmatrix} = \\ = - \begin{vmatrix} (\frac{1-u_2}{u_2})^{\frac{-w_+}{w_+ + w_-}} & h_j \frac{-w_+}{w_+ + w_-} (\frac{1-u_2}{u_2})^{\frac{-w_+}{w_+ + w_-} - 1} (\frac{-1}{u_2^2}) \\ (\frac{1-u_2}{u_2})^{\frac{w_-}{w_+ + w_-}} & h_j \frac{w_-}{w_+ + w_-} (\frac{1-u_2}{u_2})^{\frac{w_-}{w_+ + w_-} - 1} (\frac{-1}{u_2^2}) \end{vmatrix} = \\ = \frac{h_j}{u_2^2} [\frac{w_-}{w_+ + w_-} (\frac{1-u_2}{u_2})^{\frac{w_- - w_+}{w_+ + w_-} - 1} + \frac{w_+}{w_+ + w_-} (\frac{1-u_2}{u_2})^{\frac{w_- - w_+}{w_+ + w_-} - 1}] = \\ = \frac{h_j}{u_2^2} (\frac{1-u_2}{u_2})^{\frac{-2w_+}{w_+ + w_-}} = h_j (\frac{h_j^{'} + h_{j+1}^{'}}{h_j^{'}})^2 (\frac{h_{j+1}^{'}}{h_j^{'}})^{\frac{-2w_+}{w_+ + w_-}} = \\ = h_j (h_j^{'} + h_{j+1}^{'})^2 \frac{(h_{j+1}^{'})^{\frac{-2w_+}{w_+ + w_-}}}{(h_j^{'})^{2 - \frac{2w_+}{w_+ + w_-}}} = h_j (h_j^{'} + h_{j+1}^{'})^2 \frac{(h_{j+1}^{'})^{\frac{-2w_+}{w_+ + w_-}}}{(h_j^{'})^{\frac{2w_-}{w_+ + w_-}}} = \\ = h_j (h_j^{'} + h_{j+1}^{'})^2 (\frac{1}{h_j^{'\,w_-} {h_{j+1}^{'}}^{w_+}})^{\frac{2}{w_+ + w_-}} = \frac{(h_j^{'} + h_{j+1}^{'})^2}{h_j}. \end{aligned} \tag{5.52}$$

Surely there was never doubt of Green being correct.

Hence we get the A in (5.11) for a birth move as

$$\begin{aligned} A = \frac{p(y|\lambda')}{p(y|\lambda)} \frac{2\mu(2k+3)}{L^2} \frac{(s_{j+1} - s^\star)(s^\star - s_j)}{s_{j+1} - s_j} \frac{\beta^\alpha}{\Gamma(\alpha)} (\frac{h_j^{'} h_{j+1}^{'}}{h_j})^{\alpha - 1} \times \\ \exp(-\beta(h_j^{'} + h_{j+1}^{'} - h_j)) \frac{d_{k+1} L}{b_k (k+1)} \frac{(h_j^{'} + h_{j+1}^{'})^2}{h_j}. \end{aligned} \tag{5.53}$$

5.2.1.4 Death Step

In a death step one randomly chosen step position $s^\dagger = s_{j+1} \in (s_j, s_{j+2})$ will become removed and hence the dimension drops by two from $2k+1$ to $2k-1$. Two successive heights h_j and h_{j+1} are joined into a single one $h_j^{'}$, however preserving the weighted geometric average, so that

$$(s_{j+2}-s^{\dagger})\log(h_{j+1})+(s^{\dagger}-s_j)\log(h_j)=(s_{j+2}-s_j)\log(h_j^{'}). \tag{5.54}$$

This move needs to be in *detailed balance* with the corresponding birth move. The dimension matching will also hold by reversing the calculations from the previously defined birth move. If the probability for accepting a birth move is $\alpha_{birth}=\min(1,A)$, then the acceptance probability for the corresponding death move becomes

$$\alpha_{death}=\min(1,\frac{1}{A}). \tag{5.55}$$

We obtain (with an appropriate re-labeling) the expression $\frac{1}{A}$ within the acceptance probability as

$$\begin{aligned}\frac{1}{A}=\frac{p(y|\lambda')}{p(y|\lambda)}\frac{p(k-1)}{p(k)}\frac{L^2}{2k(2k+1)}\frac{s_{j+2}-s_j}{(s_{j+1}-s^{\dagger})(s^{\dagger}-s_j)}\frac{f(h_j^{'})}{f(h_j)f(h_{j+1})}\frac{j(1,\theta^{(1)})q_1(u^{(1)})}{j(2,\theta^{(2)})}\frac{1}{J}=\\ \frac{p(y|\lambda')}{p(y|\lambda)}\frac{L^2}{2\mu(2k+1)}\frac{s_{j+2}-s_j}{(s_{j+1}-s^{\dagger})(s^{\dagger}-s_j)}\\ \frac{\Gamma(\alpha)}{\beta^{\alpha}}(\frac{h_j^{'}}{h_jh_{j+1}})^{\alpha-1}\times\exp(\beta(h_j+h_{j+1}-h_j^{'}))\frac{b_{k-1}k}{d_kL}\frac{h_j^{'}}{(h_j+h_{j+1})^2}.\end{aligned} \tag{5.56}$$

5.2.1.5 The Simulation Results

We have run both a regular MCMC-program in dimensions $k=3$ and $k=4$ (one million iterations) and a RJMCMC-program (half million iterations). The results of the simulation runs are reassuringly in general in very good agreement with the results read out of the figures Green reports.

In accordance with Green we find (table 5.2, figure 5.5) the most likely number of change points to equal three.

We find clear change point at times 14400, 28700 and 35600 days. For the last change point a 95 % confidence interval was calculated to $[35062, 36130]$ days (figure 5.4). Assuming three change points, the accident rates are approximated to $0.0084, 0.0025, 0.003$ and 0.0009 accs./day. in the corresponding intervals. As can be seen with bare eye in (figures 5.3 and 5.4), the standard deviation for all of the height result would be about 0.001 accs./day.

Table 5.2: Results for *a posteriori* probabilities in a long RJ-simulation of 500 000 iterations. The most likely submodel is the one with three change points ($k = 3$).

k	0	1	2	3	4	5	6	7	8	9
frequency	0	53299	91097	160805	116559	52946	20769	3809	716	0
a posteori probability [%]	0	10.7	18.2	32.2	23.3	10.6	4.1	0.8	0.1	0

All simulations performed give a hint of an alternative change point to the last one at around 33500 days, but so does indeed Fig. 1 of Green's also. However, tha data is quite sparse in this area. Perhaps the sparsity of data leads to sharper contrasts and a clearer result, as the figures show.

The reversible jump- routine indicates potential change points around 700 and 5700 days, which Green does not report at all. However, this is actually quite well in accordance with the manually calculated hazard rate points in figures 5.1 and 5.2.

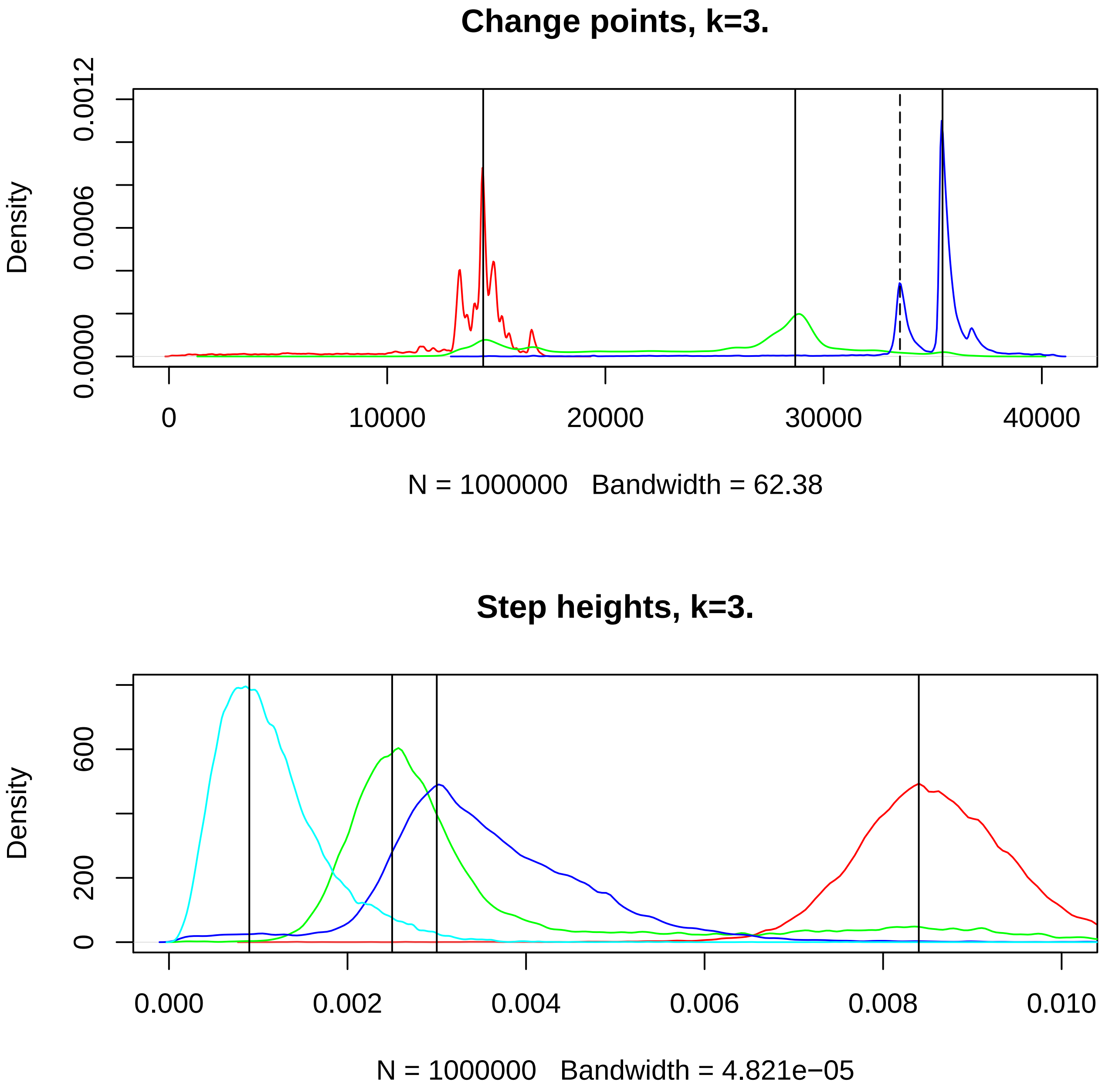

FIGURE 5.3: Location of the change points and step heights assuming constant $k = 3$. The change points and heights match very well with Green's figures. The second change point (the green curve) seems to have lost some probability mass for the first one. This is related to the labeling problem as reported in [S.Richardson and P.J.Green, 1997]. It is not always easy to identify which result has been found. The bandwith is a technical parameter from the density estimates of R.

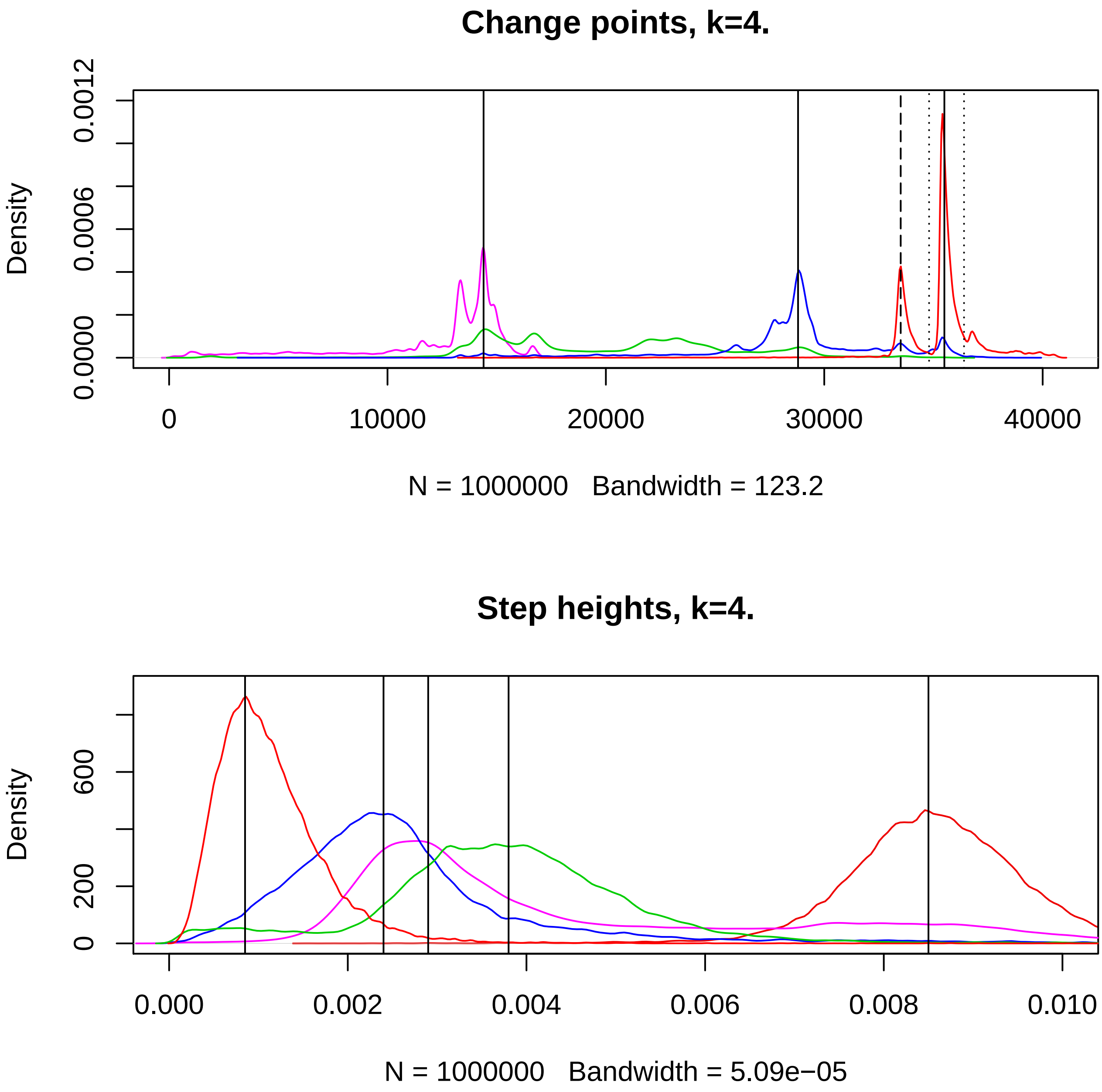

Figure 5.4: Heights and location of change points assuming constant $k = 4$. The green density function in the upper panel seems to fail in finding another change point. The posterior probabilities suggest (see table 5.2) that the most likely model has $k = 3$. There appears to be two strong candidates for the latest change point. The dotted lines indicate how to manually isolate a relevant area for error estimation. This example resulted in an estimate of 35600 (days) and a 95 % confidence interval of $[35062, 36130]$(days) for the last change point.

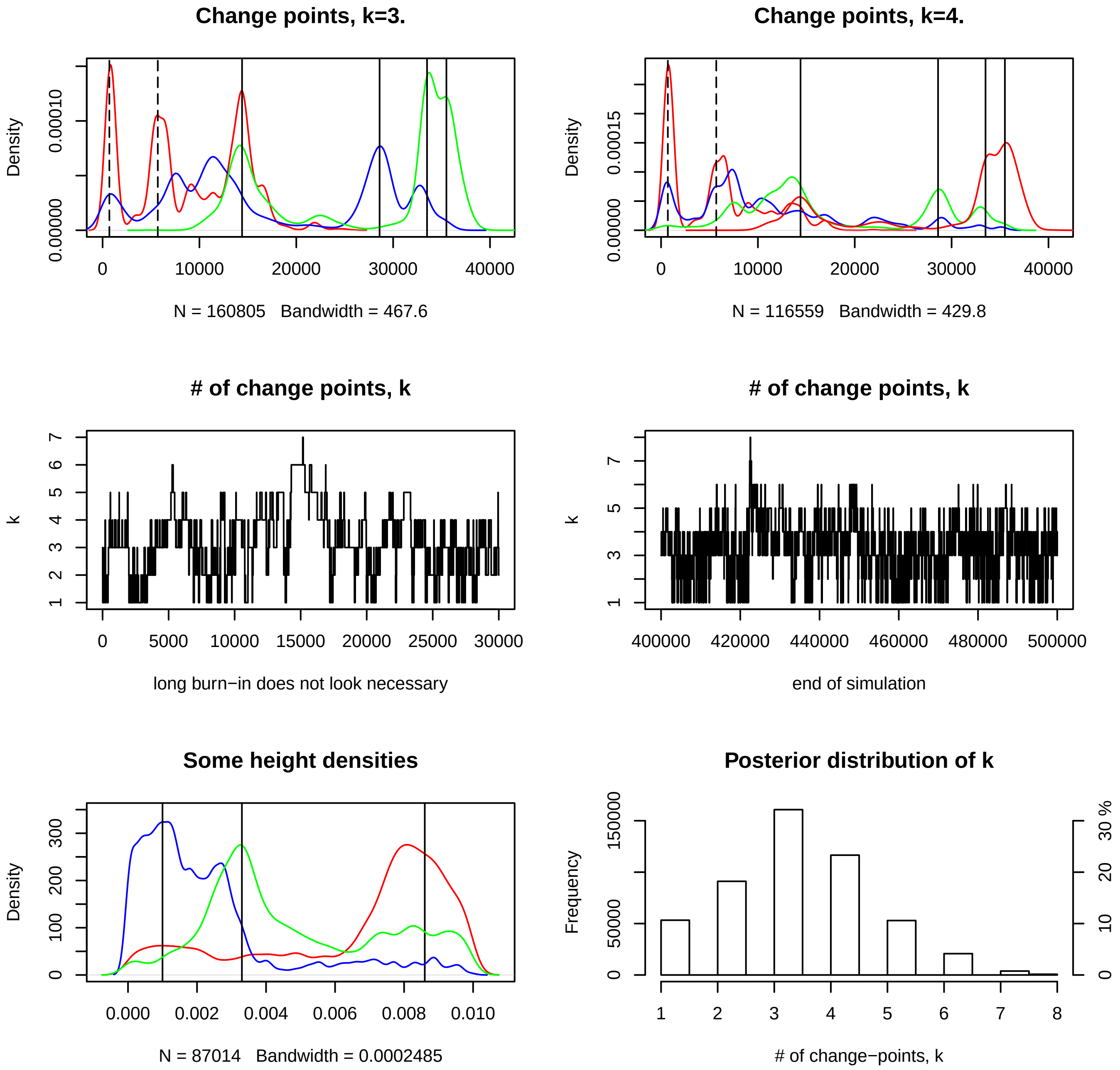

FIGURE 5.5: The upmost panels present location of change points conditioned on $k = 3$ and $k = 4$. The middle panels show the development of the model dimension while reversible jumping. The lowest two panels show some heights and the posterior model probabilities with reversible jumping.

5.3 Terrorism Attacks

The Bayesian approach and MCMC methods have become more popular also within the framework of political science. We focus on count data of terrorism attacks in different areas in the world.

In [Charlinda Santifort and Brandt, 2013] the RJMCMC is used for investigating the evolution of diversity in target choice and attack modes among domestic and transnational terrorists over the past 40 years. The changepoints are driven by changes in homeland security practice and changes in the dominant terrorist influence at the global level affecting the marginal benefits of target-attack combinations.

The study based on the Global Terror Database (to be discussed in detail in section 5.3.2) focuses on four target types (private parties, official, business and military) and four attack modes (bombings, hostage events, assassinations, armed attacks) and makes a RJMCMC run of each combination. The conclusion is that the hardest-to-defend target-attack pairing, the bombing of the private parties, has experienced the largest increase in violence both domestically and transnationally. This can be seen as a natural tendency since taking certain counter-terrorism measures may make some target-attack combinations more costly. This can change the utility function of the terrorists and shift the terrorist activity to different combinations less protected by these security enhancements. For example, implementing metal detectors in airports at the start of 1973 the terrorists' marginal costs for skyjackings increased and number of other hostage events (e.g. kidnappings) increased for all target types.

5.3.1 The Iraq Conflict

The Unites States and allied forces attacked Iraq with aerial bombardments followed by a land invasion on March 20, 2003. By mid-April Baghdad and Tikrit were under allied control practically ending the war. In [Spirling, 2007] the civilian casualty count data is analyzed from the official cessation of hostilities, May 2003, until May 2007 with the RJMCMC. The author finds evidence of four change points approximately coinciding with important events such as

- the capture of Saddam Hussein (late 2003 to spring 2004)

- the installation of Iraqi Interim Government and subsequent handover of power to the Iraqi Transitional government (summer 2004 to early 2005)
- the legistlative elections for, and negotiations to form the first full term Iraqi government (early 2006)
- the assumption of security and some military responsibilities by the Iraqi government (Aug.-Sep. 2006)

In each case the frequency of terror incidents has increased. At 2007 there had been approximately 60000 casualties, http://www.iraqbodycount.org/, of which some 3000 coalition force members and 57000 civilian fatalities (other studies report more than 10 times higher number). The data records civilian deaths caused by coalition military action and by military or paramilitary responses to the coalition presence. If there is conflict on figures, [Spirling, 2007] uses the minimum and defines a "casualty incident" as involving five deaths or more. The study focuses on how often there were incidents (the count of which is 1682) rather than how many casualties (assuming ≥ 5).

5.3.2 Terrorism in Afghanistan

The data in this section comes from the Global Terrorism Database (GTD), http://www.start.umd.edu/gtd/. We are investigating the possible increase of the rate of terrorism attacks in Afghanistan during the period of approximately last 35 years.

The Global Terrorism Database [Database, 2013] contains systematic information on terrorist events around the world from 1970 through 2012 (and annual updates are planned). It is currently the most comprehensive database on terrorist events with over 113 000 events, more than 52000 bombings, 14400 assassinations and 5600 kidnappings. There are at least 45 variables on all recorded events and more than 120 variables on more recent incidents.

The task of classifying a terrorist act is far from trivial, since the incidence could become extended in time, space or both. For an example case, say a group of hijackers divert a plane to Senegal and while at the Senegalese airport shoot two Senegalese policemen. This would still count as one incident since the hijacking was still in progress at the time of shooting and hence the two events occurred at the same time in the same place. Also, often there occur multiple attacks at the same time e.g. a suicide bombing simultaneously in five different parts of a major city. This needs bookkeeping of five incidents in the GTD.

A Terrorism act is defined by the GTD as the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious or social goal through fear, coercion or intimidation.

In practice GTD requires that the three following criteria are met in order to be considered as an incident for inclusion in the database.

1. **The incident must be intentional** - the result of a conscious calculation on the part of a perpetrator.
2. **The incident must entail some level of violence or threat of violence** - including property violence, as well as violence against people.
3. **The perpetrators of the incidents must be sub-national actors.** The database does not include acts of state terrorism.

In addition, at least two of the following three criteria are to be present for an inclusion into the database.

1. The act must be aimed at attaining a political, economic, religious, or social goal.
2. There must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) than the immediate victims.
3. The action must be outside the context of legitimate warfare activities.

The incidents are classified in 9 classes: assassination, hijacking, kidnapping, barricade incident, bombing/explosion, unknown, armed assault, unarmed assault, facility/infrastructure attack.

Of course, the data also contains exact geographichal coordinates of the incident location. We chose to study Afghanistan, one of the most dangerous countries in the world, and found 4511 events (also including 315 unsuccessful attacks) in Afghanistan in the GTD on 2082 unique dates. The "successfulness" (of the most serious attack type) depends on the type of the attack. The essential question is whether or not this attack type took place.

Table 5.3: Results for *a posteriori* probabilities in a long RJ-simulation of 500 000 iterations. The most likely submodel is the one with eight change points ($k = 8$), but in this particular run the model with one less change points ($k = 7$) comes practically as good. The other submodels have negligibly small probabilities.

k	5	6	7	8	9	10
frequency	9406	66031	194397	195511	31278	1688
a posteori probability [%]	1.8	13.2	38.9	39.1	6.3	0.3

The Taliban, an islamic fundamentalist political movement captured the capital Kabul in September 1996, spread throughout Afghanistan and ruled the Islamic Emirate of Afhganistan (diplomatically recognized only by Pakistan, Saudi Arabia and United Arab Emirates) until December 2001.

Since the 9/11-attacks, the following US invasion, the hunt of Osama bin Laden and war in Afghanistan the country has been in a mess and it is hard to draw conclusions on political turnpoints only based on the change of the hazard rate of terrorism. Yet we give it a try.

The RJMCMC-program identifies eight change points (table 5.3, figure 5.6) with posterior probability of 39.1 % in the rate of terrorism incidents in Afghanistan. The posterior probability for a model with seven change points is only slightly lower 38.9 % and with six change points 13.2 %.

Another run of 1/2 million iterations gave posterior probabilities 33.2, 41.3 and 12.9 % for submodels $k = 7, k = 8$ and $k = 9$. Hence it can be seen that the inference on the choice of a model can depend for example on the initial conditions and still more simulation runs could be conducted. Yet it seems well justified to claim that the best number of change points in this model is eight.

Reassuringly, the very same times of eight changepoints could also be calculated (figure (5.7)) with an another program, which has been used to produce the results in [Charlinda Santifort and Brandt, 2013].

This program completely outperformed the one of the author's and was approximately 400-500 times faster taking only about 20 seconds for the whole run.

While comparing the simulation results to the timeline of Afghan history[2], we find some interesting coincidences with the turning points of politics.

[2] http://en.wikipedia.org/wiki/Timeline_of_Afghan_history

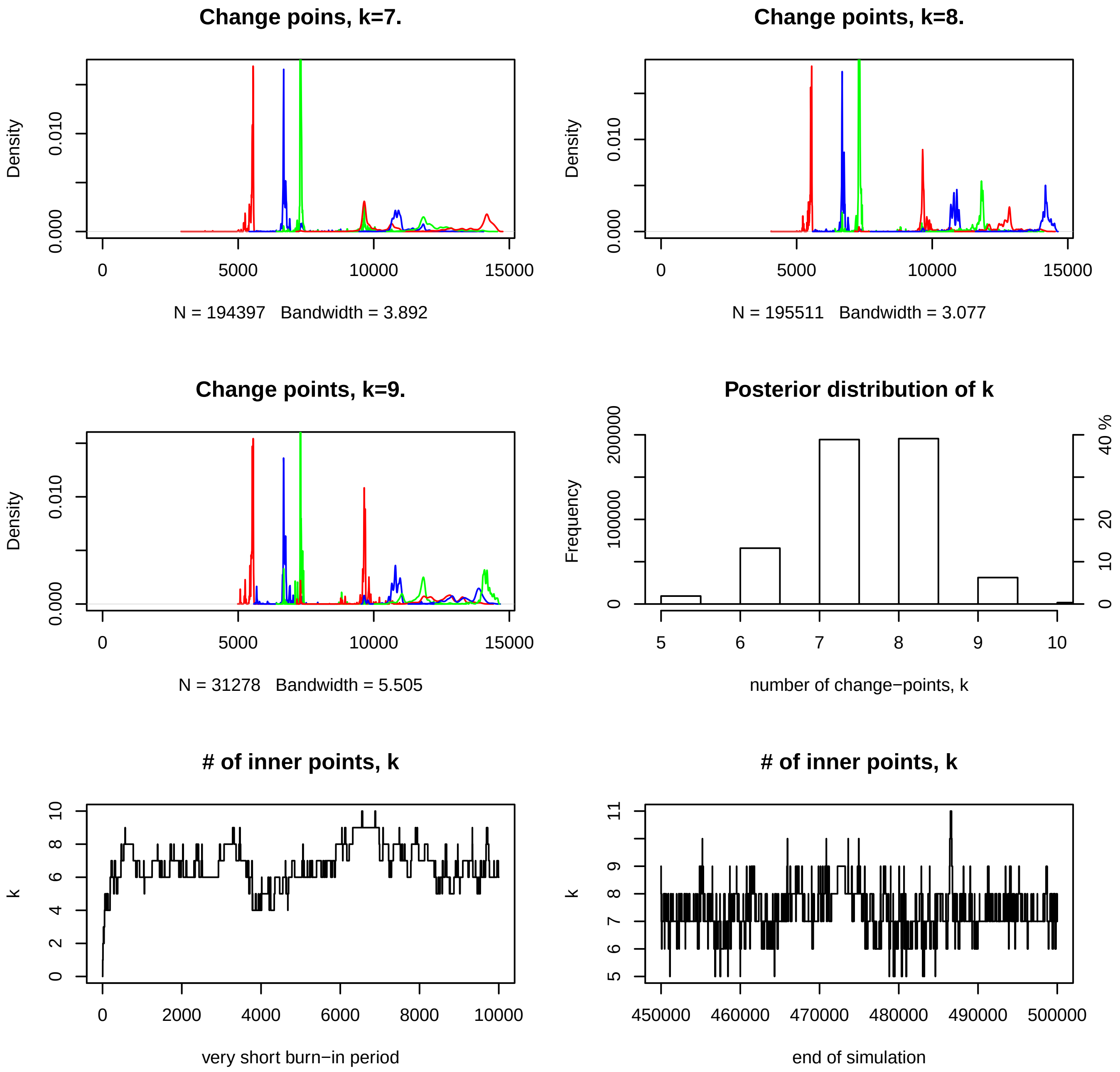

FIGURE 5.6: Densities for change points with three different models, $k = 7, k = 8$ and $k = 9$, and *a posteriori* probabilities of the submodels. The two last pictures indicate how the RJMCMC-sampler traverses in the model space. As it can be seen from the fifth picture, there should be no special need for a burn-in period longer than 1000 iterations (which is practically negligible since we did $1/2$ a million iterations in total).

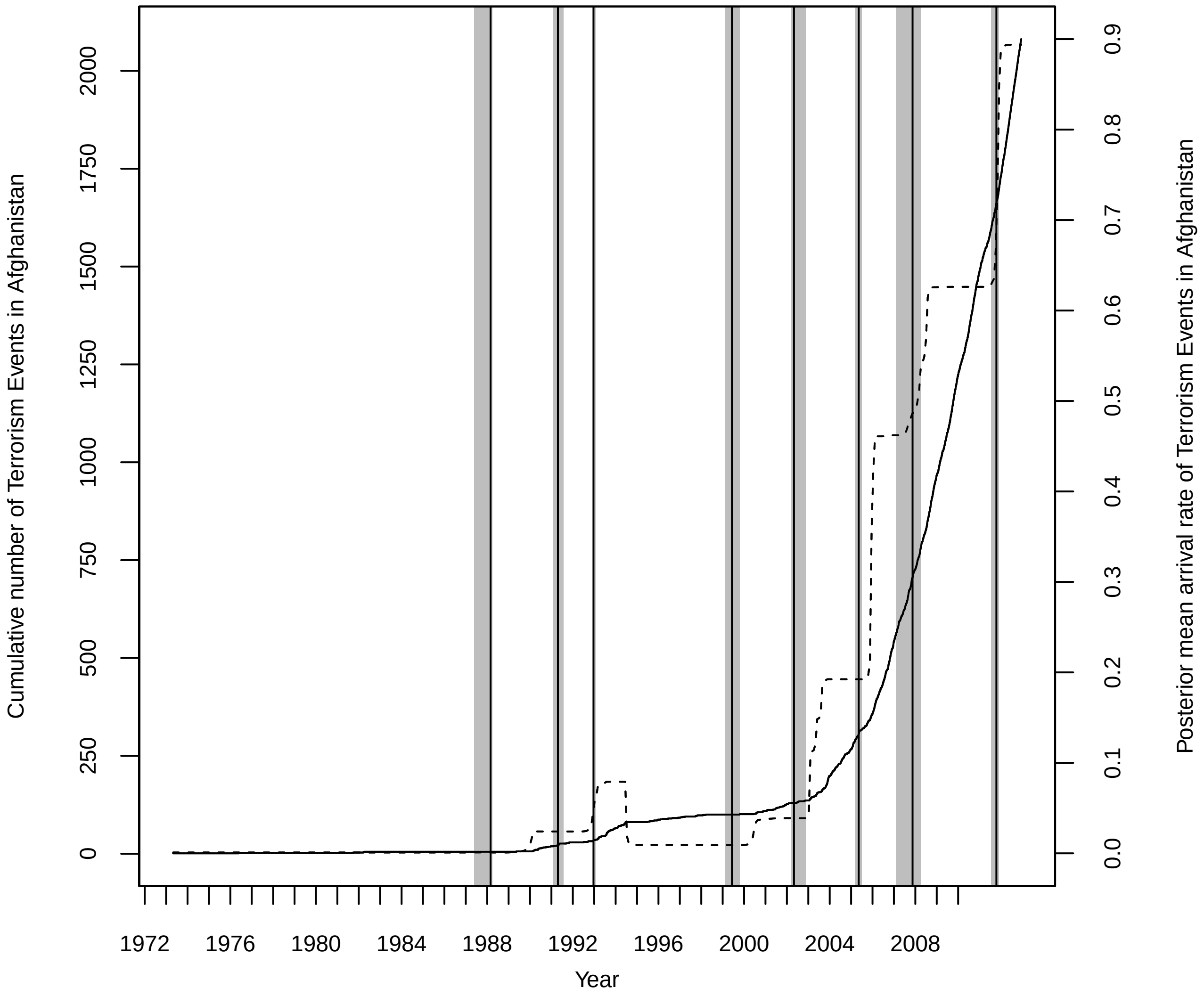

FIGURE 5.7: Location of change points and arrival rate of Terrorism Event in Afghanistan in years 1973-2012. The grey areas represent 95 % confidence intervals. Both here and in figure 5.6 the three first change point appear sharper probably due to sparser data. Unfortunately the sharpest result in 1993 cannot really be trusted, since the GTD reportedly has lost all data from that year.

In year 2002 emergency loya jirga (grand council) was held in Kabul 11.-19. June. This was called for by Bonn Agreement (December 2001)[3] where solution was sought for the government of Afghanistan after the US had ousted the Taliban regime. Since no nationally-agreed-upon government had existed in Afghanistan since 1979, it was felt necessary to have a transition period before establishing a permanent government. That would require at least one loya jirga[4] to be convened and immediate steps felt required.

Interestingly, the fifth change point given by the program matches the time of 2002 loya jirga very well.

Hamid Karzai was elected President of the Islamic Republic of Afghanistan in 9th October, 2004. In 2005 Taliban insurgency began after Pakistan decided to station around 80000 soldiers next to the porous Durand Line border with Afghanistan. There is also a change point and a clear increase in the rate of terrorism incidents in first half of 2005.

Since 1949 there have occurred cross-border shellings along the poorly marked Durand Line border between the unified Pakistan Armed forces and the Afghan National Security Forces called the Afghanistan–Pakistan skirmishes.[5] The latest hostility started in mid-2003 in the Khost province in Afghanistan and continued until 2013 when a dozen of missiles, reportedly were fired from Pakistan, killed an Afghan woman and wounded several others in Kunan Province of Afghanistan.

Particularly intensive attacks are reported in 2011 and 2012 when many sources report Pakistani missiles having hit civilian areas in Afghan provinces of Kunan, Nangarhar and Nuristan. Most of these are related to the US driven drone (unmanned aerial vehicle) attacks[6] in Northwest Pakistan along the Afghan border since 2004, the Taliban insurgency and the fact that border has never been clearly marked. The drone strikes began during Georg W. Bush administration and have increased substantially under president Barack Obama.

The last change point given by the RJMCMC program also dates around 2012 and elevates the hazard rate of terrorism to yet higher level.

The government of Pakistan has publicly condemned the drone attacks. However it also allegedly allowed the drones to operate from Shamsi Airfield in Pakistan until 21th

[3] http://en.wikipedia.org/wiki/Bonn_Agreement_(Afghanistan)

[4] http://en.wikipedia.org/wiki/2002_loya_jirga

[5] http://en.wikipedia.org/wiki/Afghanistan-Pakistan_Skirmishes

[6] http://en.wikipedia.org/wiki/Drone_attacks_in_Pakistan

April 2011. According to Wikileaks[7], Pakistan's army chief Ashfaq Parvez Kayani not only tacitly agreed to continue with the drone flights but in 2008 requested an increase of them. This matches very well with the time of the second last change point which is just before the change of year 2008.

[7] Allbritton, Chris (20 May 2011). "Pakistan army chief sought more drone coverage in '08: Wikileaks". Reuters. Retrieved 16 December 2011.

Chapter 6

Discussion

6.1 Application Areas

The applications of the Bayesian approach and hence the (RJ)MCMC-samplers are many. Since the acceptance probability

$$\alpha(x, y) = \min\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\} \tag{6.1}$$

depends on the proportion of the target $\pi(\cdot)$ times the proposal distribution $q(\cdot, \cdot)$ in two different points - the present sample point and the proposed one, one only needs to be able to evaluate these distributions up to a multiplicative constant.

Most of the problems in particle physics have been simulated with the method, and N-particle MCMC models are still being developed and run in Los Alamos[1]. There simply exists few other ways of doing this. In Los Alamos there has also been developed a software system YADAS[2] written in java for performing MCMC-simulations.

Many other examples of statistical problems outside the field of statistical physics where the number of elements could be unknown a priori can be listed. For example model selection, variable selection in regression, cluster analysis, partitioning problems, identifying mixture distributions [S.Richardson and P.J.Green, 1997], image segmentation analysis (two dimensional analogue of change-point problems), time-series models, classification problems object recognition, signal processing and Bayesian non-parametrics are such problems.

[1] http://la-science.lanl.gov/cat_math.shtml#monte

[2] http://www.ccs.lanl.gov/ccs6/yadas/yadas.html

The RJMCMC has also successfully been applied to population ecological models [Ruth King and P.Brooks, 2010], procedural modeling generating complex geometric structures such as trees, cities, buildings and Mondrian paintings [O.Talton et al., 2011] with formal grammars such as Lindenmayer L-systems defining the structural building blocks of such systems.

Given the grammar for enforcing location of houses in a suburbian residential area in a shape of say a whale or a shoe, a RJMCMC-program could perform the task of calculating the placemants, shapes and heights of the houses. Of course, the house façades could also be designed with the similar simulational approach. However, this raises the philosophically interesting question whether this can be considered as architecture or not? An architectural student [Palmer, 2013] from Bergen School of Architecture gives the clearcut answer: "No, in my opinion, this is not architecture!"

6.2 Pros and Cons of (RJ)MCMC

The power of the MCMC-simulation is based on in its ability to approximate any integral representing an expected value of a function relatively effectively by a simple sum consisting of points sampled from the Markov chain.

$$E_\pi[h(x)] = \int_M h(x)\pi(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i). \tag{6.2}$$

No matter how high-dimensional the state space M, the theory says that the approximating sum will converge at rate $\frac{1}{n}$ and precision can be made arbitrarily high by using arbitrarily high computing time. Thus the computing time for approximating a posterior distribution may easily become quite long. Of course, the development of computing speed also makes construction of more complicated models possible.

There is always the issue of monitoring the convergence as discussed in section (4.2.3). Especially in higher-dimensional models it may be hard to decide whether the convergence has yet occurred at all and one may need to rely on the results looking good.

The joint posterior

$$\pi(k, \theta_k|Y) = \frac{p(k, \theta_k)L(Y|k, \theta_k)}{\sum_{k' \in K} \int p(k', \theta_{k'})L(Y|k', \theta_{k'})d\theta_{k'}} \tag{6.3}$$

can always be factorized as the product of posterior model probabilities and model-specific parameter posteriors

$$\pi(k, \theta_k|Y) = \pi(k|Y)\pi(\theta_k|k, Y). \tag{6.4}$$

The generality of this formulation embraces both genuine model-choice situations and a single model with a variable-dimension parameter [P.J.Green and D.Hastie, 2009] .

The RJ-setup requires specifying all the regular diffusional within-model moves and the interdimensional move proposals for changing the submodel and calculating the Jacobian determinant (5.19) for the acceptance procedure (5.19). Hence any RJMCMC-sampler is necessarily rather problem-specific and the task of setting one up is hard to automatize. An attempt could be made for performing the differentiation with a symbol algebra system such as Maple.

The rather complicated setup may have stopped people from adopting the RJ-method. The possible complexity of the state space $M = \cup M_k$ itself may also present challenges in constructing proper across-model proposals as natural ideas of proximity and neigbourhood that help the design in within-model proposals may no longer be intuitive.

Inefficient proposal mechanisms lead to slow exploration of the state space, demonstrate slow convergence to the stationary distribution π and have high autocorrelation increasing the asymptotic variance of Monte Carlo estimators [P.J.Green and D.Hastie, 2009].

It can be argued that there is an inherently predetermined optimal dimension of the problem that has been a priori fixed, and once the dimension is estimated, the reversible jumping between dimensions does not bring anything new to the analyze. Yet, how to estimate that dimension? The RJ gives almost directly a posteriori estimates for the submodel probabilities.

The question whether it is good to jump seems to be little controversial [Green, 2001]. The most favourable situation for jumping is when the full posterior inference about $(k, \theta^{(k)})$ is required. In other cases, it might be interesting to compare the RJMCMC to a set of fixed-dimensional runs, and judge if the reversible jumping really gives a good contribution to the analysis in relation to the extended effort it takes.

The whole process of MCMC can be seen as quite inefficient approach (and with RJ even more so) as a lot of proposals need to be created and large part of them may need to be discarded. With the increasing computational power of today this is much less of a problem, but the length of simulation runs could easily be measured in days, even weeks. More efficient use of computing usually requires more effort on sophisticating the numerical or analytical methods. Usually this means calculating gradients.

6.3 Future Trends

The random walk Metropolis-Hastings is quite robust a method for doing simulations in a fixed dimension.

The search for improved proposal distributions, *tuning* the algorithm is often done manually and adaptive MCMC is attractive to let the computer to “learn” better parameter values while running the algorithm [Steve Brooks and Meng, 2011].

Creating a fully automated sampler in the RJ-setting has been an ideal of Green’s and his students David Hastie’s [Hastie, 2005]. It would be a tremendous practical advantage if the user could just specify the target in algebraic form and let the computer both construct an algorithm and then run it to create a reliable sample [Green, 2001].

The closest one could come at the time was a random walk Metropolis sampler for sampling from a fixed-dimensional density in the simplest form where all variables were simultaneously updated.

However, as Green puts it, "random walk Metropolis is not a panacea“. It has the drawbacks of not having geometric ergodicity[3] guaranteed and requiring conditions on the relative size of the tails of the target and proposal densities.

Reversible jump analogy of the random walk MH was proposed by Green in 2003 [Green, 2001]. The idea is to use estimates of the first- and second- order moments of θ_k denoted by μ_k and $B_k B_k^T$ where μ_k is a $k-$vector and B_k is a $k \times k-$ matrix. The proposed move

[3] A positive Harris recurrent and aperiodic Markov chain is called *ergodic*. An ergodic Markov chain with invariant distribution is *geometrically ergodic* if $\exists$ a nonnegative real-valued function M with $\pi|M| < \infty$ and a positive constant $r < 1$ such that $|P^n(x, \cdot) - \pi| \leq M(x)r^n$ for all x.

from model (k, θ_k) to model $M_{k'}$ is

$$\theta'_{k'} = \begin{cases} \mu_{k'} + B_{k'}[R_{k,k'}(B_k)^{-1}(\theta_k - \mu_k)]^{n_{k'}} & , n_{k'} < n_k \\ \mu_{k'} + B_{k'}R_{k,k'}(B_k)^{-1}(\theta_k - \mu_k) & , n_{k'} = n_k \\ \mu_{k'} + B_{k'}R_{k,k'} \left(\begin{array}{c} (B_k)^{-1}(\theta_k - \mu_k) \\ u \end{array} \right)' & , n_{k'} > n_k \end{cases} \tag{6.5}$$

where $[\cdot]^m$ picks the m first components of a vector, $R_{k,k'}$ is an orthogonal matrix and $u \sim q_{n_{k'}-n_k}(u)$ is a $(n_{k'} - n_k)-$dimensional stochastic vector, only needed when the dimension is going up. If $n_{k'} \leq n_k$ the proposal for $\theta'_{k'}$ is deterministic and calculating the Jacobian trivial. If $n_{k'} > n_k$, then (the orthogonal matrix $R_{k,k'}$ gives no contribution)

$$\left| \frac{\partial(\theta_{k'})}{\partial(\theta_k, u)} \right| = \frac{|B_{k'}|}{|B_k|}.$$

The acceptance probability is

$$\alpha[(k, \theta_k), (k', \theta_{k'})] = \frac{\pi(k', \theta'_{k'}|x)}{\pi(k, \theta_k|x)} \frac{q(k' \to k)}{q(k \to k')} \frac{|B_{k'}|}{|B_k|} \times \begin{cases} q_{n_{k'}-n_k}(u), n_{k'} < n_k \\ 1, \qquad n_{k'} = n_k \\ \frac{1}{q_{n_{k'}-n_k}(u)}, n_{k'} > n_k \end{cases} \tag{6.6}$$

where $q(k \to k')$ is the probability for attempting the change of a submodel.

If the model specific densities $\pi(k, \theta_k|x)$ are unimodal and the first and second moments μ_k and $B_k B_k^T$ available, then high between-model acceptance probabilities may be achieved.

The whole idea for proposing (6.5) was motivated by the fact that if the model-specific targets $\pi(k, \theta_k|x)$ were normal distributions with means μ_K and variance $B_k B_k^T$, if the innovation variables were standard normal and if $\frac{q(k \to k')}{q(k' \to k)} = \frac{p(k'|x)}{p(k|x)}$ could be chosen, then these proposals would already be in detailed balance and hence there would be no need to compute the MH accept/reject decision.

Bibliography

Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Herbert L. Anderson. Metropolis, Monte Carlo, and the MANIAC. *Los Alamos Science Special Issue*, 72(14), 1986.

Christian P.Robert and George Casella. A Short History of Markov Chain Monte Carlo Subjective Recollections from Incomplete Data. *Statistical Science*, 26(1):102–115, 2011. doi: 10.1214/10-STS351.

N.Metropolis, A.Rosenbluth, M.Rosenbluth, A.Teller, and E.Teller. Equations of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21:1087–1092, 1953. doi: 10.1063/1.1699114.

W. Hastings. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57:97–109, 1970.

P.H.Peskun. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60(3): 607–612, 1973.

Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741, 1984.

Alan E.Gelfand and Adrian F.M.Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

David Draper. Assesment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 57(1):45–97, 1995.

Robert E.Kass and Adrian E.Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Christian P.Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2007.

Hirotugu Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC-19(6), 1974.

Gideon Schwarz. Estimating the Dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

K. Takeuchi. Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18, 1976. In Japanese.

N.Murata, S.Yoshizawa, and S.Amari. Network information criterion-determining the number of hidden units for artificial neural network models. *IEEE Trans. Neur. Netwrks.*, 5:865–872, 1978.

David J.Spiegelhalter, Nicola G.Best, Bradley P.Carlin, and Angelika van der Linde. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(Part 4):583–639, 2002.

Roger Eckhardt. Stan Ulam, John von Neumann and the Monte Carlo method. *Los Alamos Science Special Issue*, 15, 1987.

Anna Mikusheva. course materials for 14.384 Time Series Analysis, 2007. URL http://ocw.mit.edu. MIT OpenCourseWare.

William Feller. *An Introduction to Probability Theory and its Applications*, volume I. John Wiley and Sons, 3rd edition, 1968.

Luke Tierney. Markov Chains for exploring the posterior distributions. *The Annals of Statistics*, 64(4):1701–1762, 1994.

S.P.Meyn and R.L.Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, 2005. URL probability.ca/MT.

O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, New York, 2005.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, 1995.

Jun S.Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

Esa Nummelin. MC's for MCMC'ists. *International Statistical Review*, 70(2):215–240, 2002.

Dani Gamerman and Hedibert F. Lopes. *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, 2nd edition, 2006.

Christian P.Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004.

Christian P.Robert and George Casella. *Introducing Monte Carlo Methods with R*. Springer, 2010.

Julian Besag. Markov Chain Monte Carlo for Statistical Inference. *Center for Statistics and the Social Sciences. Working Paper*, 9, 2001.

A. A. Barker. Monte Carlo Calculations of the Radial Distribution Functions for a Proton-Electron Plasma. *Aust. J. Phys.*, 18:119–133, 1965.

Alan E.Gelfand, Susan E.Hills, Amy Racine-Poon, and Adrian F.M.Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85(412):972–985, December 1990.

George Casella and Edward I.George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, 1992.

John M. Castelloe and Dale L. Zimmerman. Convergence Assessment for Reversible Jump MCMC Samplers. Technical report, Department of Statistics and Actuarial Science, University of Iowa, 2002.

Mary Kathryn Cowles and Bradley P. Carlin. Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

Stephen P. Brooks and Gareth O. Roberts. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8:319–335, 1998.

Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–511, 1992.

John Geweke. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In A.P.Dawid J.M.Bernardo, J.Berger and A.F.M.Smith, editors, *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, pages 169–188, Oxford, U.K., April 1991. Oxford University Press.

Jerry O.Talton, Yu Lou, Steve Lesser, Jared Duke, Radomir Měch, and Vladlen Koltun. Metropolis procedural modeling. *ACM Transactions on Graphics*, 30 Issue 2(11), April 2011. doi: 10.1145/1944846.1944851. URL http://doi.acm.org/10.1145/1944846.1944851.

P. Giudici S.P. Brooks and G. O. Roberts. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(Part 1):3–55, 2003.

E.Marinari and G.Parisi. Simulated Tempering: a New Monte Carlo Scheme. *Europhys. Lett.*, 19(6):451–458, 1992.

C.D.Gelatt Jr. S.Kirkpatrick and M.P.Vecchi. Optimization by Simulated Annealing. *Science, New Series*, 220(4598):671–680, May 1983. doi: 10.1126/science.220.4598.671.

N.Friel S.P. Brooks and R.King. Classical model selection via simulated annealing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(Part 2):503–520, 2003.

Peter J. Green. *Trans-Dimensional Markov Chain Monte Carlo*. Oxford University Press, Oxford, U.K., 2001.

S.A.Sisson and Y.Fan. A Distance-Based Diagnostic for Trans-Dimensional Markov Chains. *Statistics and Computing*, 17:357–367, 2007. doi: 10.1007/s11222-007-9025-z.

E.S.Pearson B.A.Maguire and A.H.A.Wynn. The time intervals between industrial accidents. *Biometrika*, 39(1):168–180, 1952.

R.G.Jarrett. A note on the time intervals between coal-mining disasters. *Biometrika*, 66(1):191–193, 1979.

A.E.Raftery and V.E.Akman. Bayesian analysis of a Poisson process with a change point. *Biometrika*, 73(1):85–89, 1986.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

S.Richardson and P.J.Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 59:731–792, 1997.

Todd Sandler Charlinda Santifort and Patrick T Brandt. Terrorist attack and target diversity: Changepoints and their drivers. *Journal of Peace Research*, 50(1):75–90, 2013. doi: 10.1177/0022343312445651.

Arthur Spirling. "Turning Points" in the Iraq Conflict: Reversible Jump Markov Chain Monte Carlo in Political Science. *The American Statistician*, 61(4), 2007. doi: 10.1198/000313007X247076.

Global Terrorism Database. National Consortium for the Study of Terrorism and Responses to Terrorism, GTD, Codebook: Inclusion Criteria and Variables. *Global Terrorism Database*, December 2013. URL http://www.start.umd.edu/gtd/.

Olivier Gimenez Ruth King, Byron J.T.Morgan and Stephen P.Brooks. *Bayesian Analysis for Population Ecology*. CRC Press, 2010.

Christian Victor Palmer. Personal communication, 2013.

P.J.Green and D.Hastie. Reversible jump MCMC. *Manuscript*, 2009.

Andrew Gelmanm Galin L.Jones Steve Brooks and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC, 2011.

David Hastie. *Towards Automatic Reversible Jump Markov Chain Monte Carlo*. PhD thesis, Univ. of Bristol, 2005.