# BCCS 2008/09: Graphical models and complex stochastic systems: Lecture 9: Markov chain Monte Carlo

## 9.1 Computational Bayesian statistics

For Bayesian methods to be practical in real problems, not just 'toy' special cases with a few parameters and simple distributions, we need a general-purpose 'inference engine' to do the computation. Ideally we would want

- complete freedom in model specification
- complete freedom in inference
- capacity to handle moderate and large problems
- algorithms generated from problem specification automatically

Remarkably there is a class of methods, Markov chain Monte Carlo (MCMC), discovered at Los Alamos in the 1940's, and popular in physics since 1953, that gets quite close to satisfying these requirements. Since 1989 they have become popular in statistical science, and hence given a great impetus to the use of Bayesian statistical methods. There continues to be a lot of research effort both in designing new general methods, and analysing their properties, and in developing methods for particular kinds of model or data.

## 9.2 Cyclones example: point processes and change points

We are going to illustrate the ideas of MCMC with a running example: the observations are a point process of *events* at times $y_1, y_2, \ldots, y_N$ in an observation interval $[0, L)$. For simplicity, we suppose the events occur *at random* — that is, as a Poisson process — but at a possibly non-uniform rate: say rate $x(t)$ per unit time, at time $t$. The objective is to make inference about $x(t)$. We will work up through a series of models, ultimately allowing an unknown number of change points, unknown hyperparameters, and a parametric periodic component.

The models and the respective algorithms and inferences will be illustrated by an analysis of the point process of cyclones hitting the Bay of Bengal: there were 141 cyclones over a period of 100 years. The data are plotted, both as a jittered dot plot, and by means of its cumulative counting process, in Figure 1.

**Model 1: constant rate** First suppose that $x(t) \equiv x$ for all $t$.

Then the times of the events are immaterial: we observe $N$ events in a time interval of length $L$; the obvious estimate of $x$ is

$$\widehat{x} = \frac{N}{L}.$$

This is the *maximum likelihood estimator* of $x$ under the assumption (implied by the 'randomness' assumption above), that $N$ has a Poisson distribution, with mean $xL$:

$$p(N|x) = e^{-xL}\frac{(xL)^N}{N!}.$$

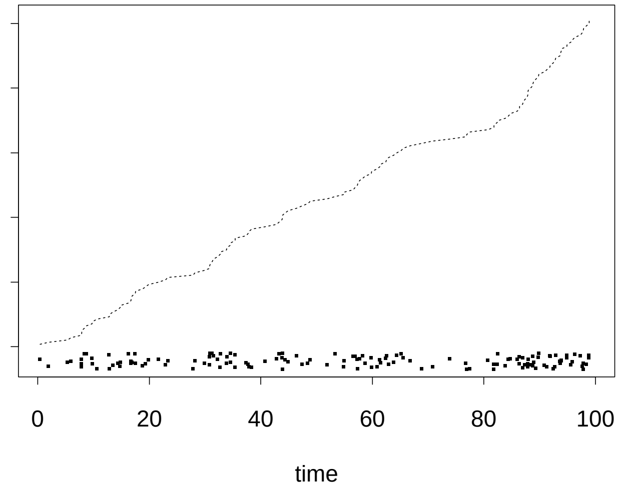

Figure 1: Cyclones data, as a jittered dot plot, and its cumulative counting process.

**Model 2: constant rate, the Bayesian way** To take a Bayesian approach to this example, suppose that we have prior information about $x$ (from previous studies, for example). Let us suppose we can model this by saying

$$x \sim \Gamma(\alpha, \beta),$$

a Gamma distribution (with mean $\alpha/\beta$ and variance $\alpha/\beta^2$).

Then since

$$p(x|N) \propto p(x)p(N|x),$$

we find that

$$\begin{aligned} p(x|N) &\propto \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} e^{-xL} \frac{(xL)^N}{N!} \\ &\propto x^{\alpha+N-1} \exp(-(\beta+L)x) \end{aligned}$$

or in other words

$$x|N \sim \Gamma(\alpha + N, \beta + L).$$

So $x$ has a Gamma distribution with mean $(\alpha + N)/(\beta + L)$, or approximately $N/L$ if $N$ and $L$ are large compared with $\alpha$ and $\beta$. Thus *with a lot of data*, the Bayesian posterior mean is close to the maximum likelihood estimator. The posterior distribution of $x$ for model 2 fitted to the cyclones data is shown in Figure 2; we used $\alpha = \beta = 1$ here.

There is no need for MCMC in this model: you can calculate the posterior exactly, and recognise it as a standard distribution. It would not have worked out like this for any other prior; this choice is called *conjugate*.

## 9.3 The Gibbs sampler for a Normal random sample

Before we elaborate the cyclones example to a point where exact calculation is no longer practicable, let us consider an even simpler, and completely familiar, example, but following an elementary Bayesian approach.

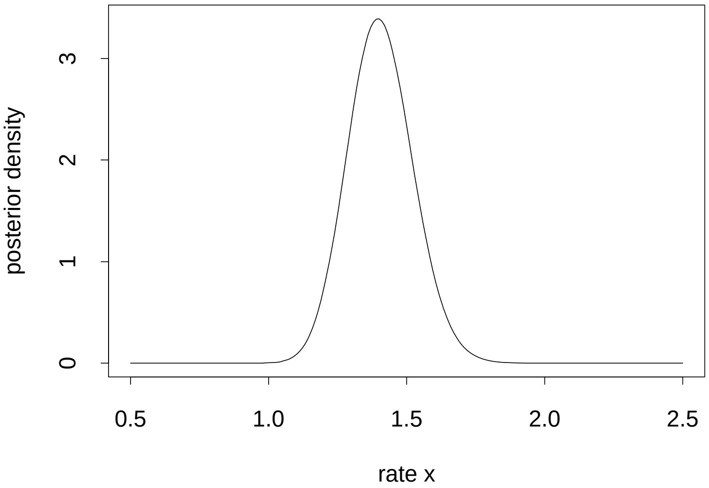

Figure 2: Cyclones data: posterior for $x$ in model 2.

Our data are a random sample of size $n$ from $N(\mu, \sigma^2)$. We place independent priors on $\mu$ and $\sigma$:

$$\begin{aligned} \mu &\sim N(\xi, \kappa^{-1}) \\ \sigma^{-2} &\sim \Gamma(\alpha, \beta), \end{aligned}$$

and it is easy to see that the resulting joint posterior is

$$\begin{aligned} \pi(\mu, \sigma^2|Y) &\propto (\sigma^2)^{-\alpha-n/2-1} \\ &\quad \times \exp\left\{-\frac{\beta}{\sigma^2} - \frac{\kappa(\mu-\xi)^2}{2} - \frac{\sum(Y_i-\mu)^2}{2\sigma^2}\right\}. \end{aligned} \tag{1}$$

This is somewhat awkward to handle; the parameters are dependent *a posteriori*, although they were independent *a priori*. However, the *full conditionals* — the conditional distributions of each parameter given the other parameter(s) and the data — are easily found:

$$\begin{aligned} \mu|\sigma, Y &\sim N\left(\frac{\sigma^{-2}\sum Y_i + \kappa\xi}{\sigma^{-2}n + \kappa}, \frac{1}{\sigma^{-2}n + \kappa}\right) \\ \sigma^{-2}|\mu, Y &\sim \Gamma(\alpha + n/2, \beta + \textstyle\sum(Y_i - \mu)^2/2). \end{aligned}$$

What happens if we generate a sample of $(\mu, \sigma)$ pairs by alternately drawing $\mu$ and $\sigma^{-2}$ from these distributions? The beginning of this process is illustrated in Figure 3.

This is a simple example of a *Gibbs sampler*, a particular kind of MCMC. The alternating updates of one variable conditioned on the other induces Markov dependence: the successively sampled pairs form a Markov chain, and it is readily shown that the joint posterior (1) is the (unique) invariant distribution of the chain. Standard theorems imply that the chain converges to this invariant distribution, so that we can treat the realised values as an approximate sample from the posterior. (The approximation error vanishes in the simulation limit.) A sample of 1000 pairs is shown in Figure 4, and the shape of the joint distribution can now be discerned. Examples of possible outputs of interest are the marginal distributions shown in Figure 5.

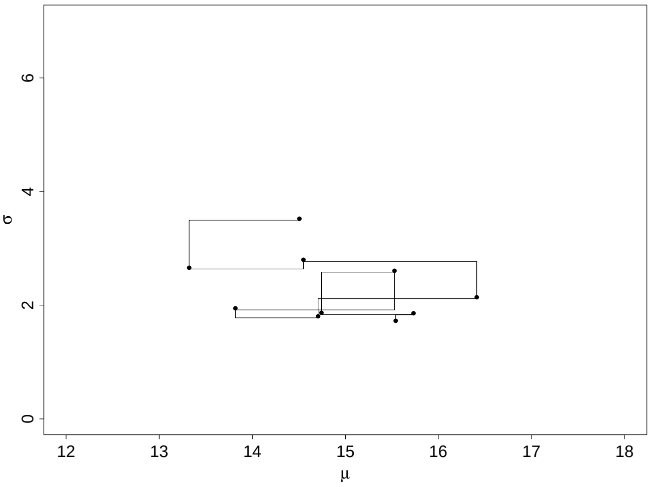

Figure 3: First 10 samples from a Gibbs sampler of $(\mu, \sigma)$ from Normal random sample with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.

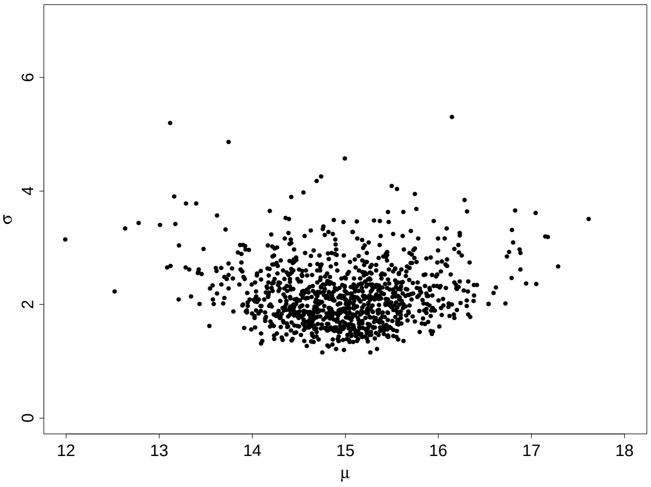

Figure 4: Posterior sample of $(\mu, \sigma)$ from Normal random sample with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.

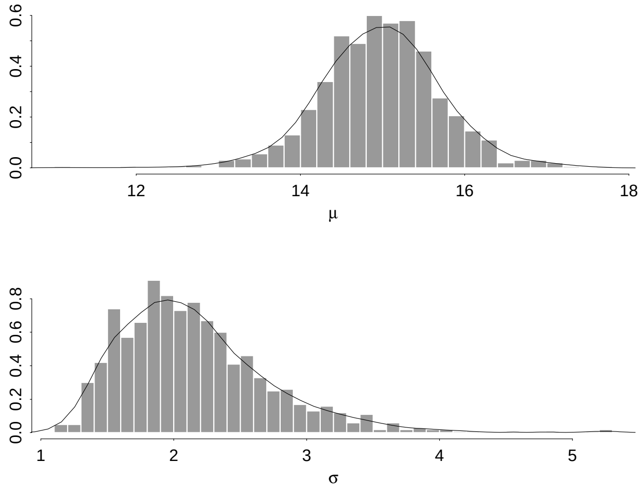

Figure 5: Posterior distributions of $\mu$ and $\sigma$ from Normal random sample with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.

However, we need not be confined to pictorial displays of marginal posteriors. One of the great liberating influences of MCMC in Bayesian inference has been the flexibility of inference allowed by sample-based computation. For example, consider prediction: we can calculate $P\{Y_{n+1} > 19\}$ by averaging $1 - \Phi(\{19 - \mu\}/\sigma)$:

$$\frac{1}{N}\sum_{t=1}^{N}\left[1 - \Phi(\{19 - \mu^{(t)}\}/\sigma^{(t)})\right] \approx 0.045$$

for the sample of Figure 4. Incidentally, it is interesting that this is more than twice the value (0.0175) that a frequentist would obtain by plugging the maximum likelihood estimates into $1 - \Phi(\{19 - \mu\}/\sigma)$.

### 9.4 Cyclones example, continued

For a more interesting and substantial application, let us return to the cyclones example, and consider some elaborations of the basic model 2.

**Model 3: constant rate, with hyperparameter** Suppose you are reluctant to specify your prior fully: you are happy to say

$$x \sim \Gamma(\alpha, \beta)$$

and can specify $\alpha$ but not $\beta$, and want to state only

$$\beta \sim \Gamma(e, f)$$

for fixed $e$ and $f$. (This formulation actually makes rather more sense in our next formulation, model 4).

Now $p(x|N, \alpha, e, f)$ is no longer available: it does not have an explicit form. But $p(x|N, \alpha, \beta, e, f)$ and $p(\beta|x, N, \alpha, e, f)$ are simple:

$$x|N, \alpha, \beta, e, f \sim \Gamma(\alpha + N, \beta + L)$$

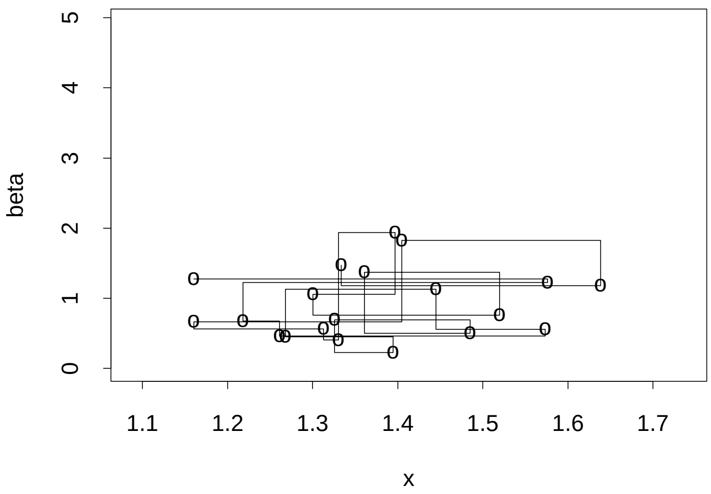

Figure 6: First few moves of the Gibbs sampler for the cyclones data, model 3.

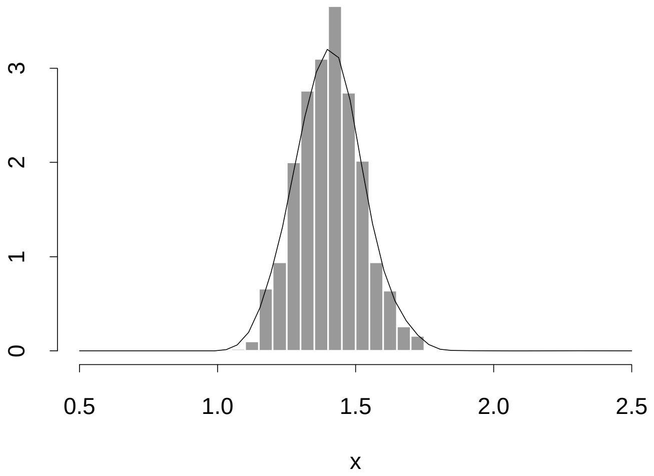

Figure 7: Marginal distribution for $x$ for the cyclones data, model 3.

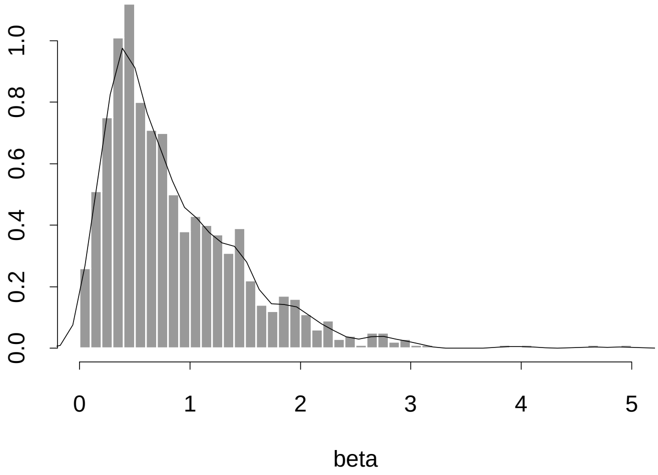

Figure 8: Marginal distribution for $\beta$ for the cyclones data, model 3.

as before, and

$$\beta | x, N, \alpha, e, f \sim \Gamma(e + \alpha, f + x).$$

So we can use the Gibbs sampler, and sample from these distributions in turn, updating $x$ and $\beta$ alternately. This creates a Markov chain with states $(x, \beta)$, the unknown parameters in this model.

Figure 6 shows the first few moves of a Gibbs sampler applied to model 3 on the cyclones data; we took $e = 1$ and $f = N/L = 1.41$. The marginal distributions for $x$ and $\beta$, as accumulated from the first 1000 sweeps of this Gibbs sampler are displayed in Figures 7 and 8.

**Model 4: constant rate, with change point** Now let us allow $x(t)$ to vary, but in a particular way.

Suppose $x(t)$ is piecewise constant, that is, a step function. This might be a suitable model if we postulate one or more *change points*; the process is completely random, but the rate switches between levels, perhaps as part of an underlying process, perhaps due to the recording mechanism.

Let us first take one change point, at *known* time $T \in (0, L)$, so that

$$x(t) = \begin{cases} x_0 & \text{if } 0 \leq t < T \\ x_1 & \text{if } T \leq t < L. \end{cases}$$

Suppose that $x_0$ and $x_1$ are *a priori* independently drawn from Gamma distributions, as before:

$$x_j \sim \Gamma(\alpha, \beta).$$

Then if $N_0$ and $N_1$ are the numbers of events before and after $T$, the above method extends to sampling in turn from

$$x_0 | \cdots \sim \Gamma(\alpha + N_0, \beta + T),$$

$$x_1 | \cdots \sim \Gamma(\alpha + N_1, \beta + (L - T)),$$

and

$$\beta|\cdots \sim \Gamma(e+2\alpha, f+x_0+x_1),$$

forming a Markov chain with a three-dimensional state space $\{(x_0, x_1, \beta)\}$. Note that to preserve our sanity we write '$|\cdots$' to mean 'given all other variables' — including the data.

The hierarchical model using random $\beta$ makes more sense now: the effect is to 'borrow strength' in estimation from both halves of the data together: $x_0$ and $x_1$ are conditionally independent given $\beta$, but are *un*conditionally *dependent*. In inference their values will be shrunk together.

**Model 5: multiple change points** If there are $k$ change points $T_1, T_2, \ldots, T_k$ with

$$x(t) = \begin{cases} x_0 & \text{if } 0 \le t < T_1 \\ x_1 & \text{if } T_1 \le t < T_2 \\ \cdots & \cdots \\ x_k & \text{if } T_k \le t < L. \end{cases}$$

then everything is extended in a very similar way, giving a Markov chain with states $(x_0, x_1, \ldots, x_k, \beta)$.

### 9.5 Gibbs sampling in general: deriving full conditionals

In a nutshell, the Gibbs sampler requires us iteratively to resample each unknown parameter from its full conditionals – its conditional distributions given the other parameter(s) and the data. For a model defined as a DAG, there is a simple recipe for finding the full conditional. Recall that

$$p(x) = \prod_v p(x_v|x_{\mathrm{pa}(v)}) \tag{2}$$

where $x$ are all the variables in a DAG – it doesn't matter whether these are parameters ('hidden variables') or data ('observed variables'). Consider a particular variable $x_a$: let $x_{-a}$ denote all other variables. Then the full conditional for $x_a$ is

$$p(x_a|x_{-a}) = \frac{p(x_a, x_{-a})}{p(x_{-a})} = \frac{p(x)}{p(x_{-a})} \propto p(x)$$

(proportional as a function of $x_a$), so we need just the terms from the DAG factorisation (2) that involve $x_a$. This variable appears once in (2) as a child and possibly several times as a parent, so

$$p(x_a|x_{-a}) \propto p(x_a|x_{\mathrm{pa}(a)}) \times \prod_{v: a \in \mathrm{pa}(v)} p(x_v|x_{\mathrm{pa}(v)})$$

Gibbs sampling is straightforward if this expression simplifies to a standard distribution.

### 9.6 Other approaches to Bayesian computation

Do we have to resort to Gibbs sampling for this application and examples like it? Under the posterior distribution in a Bayesian formulation, the parameters $\boldsymbol{\theta}$ are generally *dependent*, so we have to compute with a multivariate distribution, often in a high number of dimensions, with arbitrarily complex patterns of dependence. Here, "compute with" could mean almost anything; examples would be to calculate a marginal (posterior) density or make a probabilistic prediction.

There are various possible alternative approaches to Bayesian computation:

- Exact analytic integration: this is usually not available (and we do not want to be restricted in our model construction to use conjugate priors, etc., to make it possible).

- Asymptotic analytic approximations (e.g. Laplace; see, for example, Kass *et al.*, 1988): these are somewhat awkward to set up, and can be unreliable.

- Conventional numerical methods: these require expertise and careful design to set up, and are only efficient in a low number of dimensions.

- Ordinary ("static") simulation: this is always available in principle, since any posterior distribution can be factorised as

$$p(\boldsymbol{\theta}|Y) = p(\theta_1, \theta_2, \ldots, \theta_p|Y)$$

$$= p(\theta_1|Y)p(\theta_2|\theta_1, Y)\ldots p(\theta_p|\theta_1, \ldots, \theta_{p-1}, Y)$$

  but the univariate distributions on the right hand side are rarely all available for simulation purposes (even after re-ordering).

- Mean field methods and more advanced variational approximations are popular in the machine learning community for inference in Bayesian models. They typically run very much faster than MCMC methods, but on the downside, they are only available for models in very specific classes (for example, exponential family models with conjugate priors), and it is currently not possible to quantify the quality of the approximations made.

### 9.7 MCMC beyond the Gibbs sampler

Markov chain Monte Carlo (MCMC, also sometimes known as iterative or dynamic simulation) works even where static simulation does not, essentially because

- All simulation methods rely on the Law of Large Numbers, and this remains true when you have a Markov chain instead of an i.i.d. sequence.

- If you can tolerate Markov dependence, then you can update the parameters $\theta_1, \theta_2, \ldots, \theta_p$ one-by-one (or in small groups).

The result of combining these two simple points is very far-reaching!

Gibbs sampling is not always straightforward – in general modelling we cannot count on being able to draw randomly from full conditionals as they may not have a standard form. Fortunately there are other recipes for constructing MCMC methods – the key idea is to arrange that all updates to the current state of the unknowns ($\boldsymbol{\theta}$ say) preserves *detailed balance.*

$$p(\boldsymbol{\theta}|\boldsymbol{y})P(\boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{\theta}'|\boldsymbol{y})P(\boldsymbol{\theta}', \boldsymbol{\theta})$$

where $\boldsymbol{y}$ are the data, and $P(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is the probability (density) of changing $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$.

The Metropolis-Hastings method (Metropolis et al, 1953; Hastings, 1970) is the most flexible way of doing this - essentially you *propose* an update from any distribution you like, and then accept or reject it to ensure detailed balance – the formula for the acceptance probability involves the full conditional – but this only has to be evaluated, not sampled from. In 1995, Green extended these methods to deal with models whether 'the number of things you don't know is one of the things you don't know' – i.e. parameters whose dimension is not fixed.

### 9.8 WinBugs

`WinBugs` is a software system in which hierarchical Bayesian models can be set up, either via a program text or using a GUI. The system then automatically creates a MCMC method to simulate from the posterior. The resulting program is not as efficient as specially-written code, but much faster for the user. The online manual has dozens of useful examples. `WinBugs` is installed on the BCCS PC's.

### 9.9 Reading

The book of Gilks, Richardson and Spiegelhalter (1996), comprising articles contributed by 32 authors, provides an excellent introduction and overview to the theory, implementation and application of Bayesian MCMC (although some sections are rather dated now). There is a useful online directory of papers and preprints at `http://www.statslab.cam.ac.uk/∼mcmc/`. These notes are extracted/adapted from my chapter 'A Primer on Markov chain Monte Carlo' in the Complex Stochastic Systems book (ed. Barndorff-Nielsen, Cox and Klüppelberg).