Conditional Prior Proposals in Dynamic Models

LEONHARD KNORR-HELD

Ludwig-Maximilians-Universität München

ABSTRACT. Dynamic models extend state space models to non-normal observations. This paper suggests a specific hybrid Metropolis–Hastings algorithm as a simple device for Bayesian inference via Markov chain Monte Carlo in dynamic models. Hastings proposals from the (conditional) prior distribution of the unknown, time-varying parameters are used to update the corresponding full conditional distributions. It is shown through simulated examples that the methodology has optimal performance in situations where the prior is relatively strong compared to the likelihood. Typical examples include smoothing priors for categorical data. A specific blocking strategy is proposed to ensure good mixing and convergence properties of the simulated Markov chain. It is also shown that the methodology is easily extended to robust transition models using mixtures of normals. The applicability is illustrated with an analysis of a binomial and a binary time series, known in the literature.

Key words: Bayesian computing, blocking, conditional prior proposal, discrete data, dynamic model, innovative outliers, Markov chain Monte Carlo

1. Introduction

Markov chain Monte Carlo (MCMC) simulation in dynamic models with non-normal observations is an on-going problem. Such dynamic models relate observations y_t, $t = 1, \ldots, T$, to unobserved state parameters α_t with a so-called observation model, typically a generalized linear model. Temporal dependence is modelled within a transition model, an autoregressive Gaussian prior for the latent parameters $\alpha = (\alpha_1', \ldots, \alpha_T')'$. Hyperparameters are included in a third level of hierarchy and some conditional independence assumptions complete the model specification.

Such models are known as state space models for Gaussian observations y_t. MCMC simulation in state space models is discussed in several papers. Carlin *et al.* (1992) discuss Gibbs sampling and update α_t with a sample from the corresponding full conditional. However, Carter & Kohn (1994) and Frühwirth-Schnatter (1994) observe bad mixing and convergence behaviour in such a "single move" blocking strategy. They propose to update α all at once instead, again using a Gibbs step, i.e. a sample from the (now high dimensional) full conditional. Special properties of this Gaussian distribution ensure an efficient algorithm.

Corresponding work for the more general class of dynamic (generalized linear) models is rather rudimentary; the full conditionals are now fundamentally non-Gaussian due to the non-Gaussian observation model. Fahrmeir *et al.* (1992) generalize the single move Gibbs sampler of Carlin *et al.* (1992) to non-Gaussian observations. As for Gaussian observations, the method may have poor performance when parameters α_t are highly correlated in the posterior.

Gamerman (1998) tries to counter this problem through a reparameterization of the model to *a priori* independent system disturbances and reports considerably improved mixing and convergence behaviour. The algorithm uses ideas from posterior mode estimation by iterative Kalman smoothing (Fahrmeir, 1992; Fahrmeir & Wagenpfeil, 1997) to construct a Hastings proposal that takes observations into account. The proposal is Gaussian and is built in the spirit of weighted least squares algorithms for generalized linear models (McCullagh & Nelder,

1989). However, the reparameterization destroys the simple structure of the full conditional, leading to an algorithm of quadratic computational complexity in T.

Shephard & Pitt (1997) propose, in contrast to Gamerman, to divide α into several blocks ("block move") as an intermediate strategy between updating α one at a time and all at once. They use several Fisher scoring type steps for every updating step to calculate the moments of a Gaussian Hastings proposal that tries to approximate the full conditional of the block through an analytic Taylor expansion.

Both algorithms have proposals in common which try to approximate full conditionals, imitating a Gibbs step with acceptance probability close to 1. In contrast, our approach does not seek to approximate the corresponding full conditional; in fact it seems to have optimal performance for acceptance rates significantly below 1. Performance is poor for acceptance rates close to 1, a feature known from other Hastings proposals such as the widely used Metropolis random walk proposal, which is known to have optimal performance for acceptance rates below 50% (Gelman *et al.*, 1995).

Our methodology uses specific Hastings proposals which reflect the autoregressive prior specification but are independent of the observation model. The resulting algorithm is conceptually simple, since all proposals are Gaussian with known moments. Updating is done within a certain blocking strategy to ensure good mixing and convergence of the simulated Markov chain. Tuning of the algorithm is done by choosing a block size, rather than the spread of the proposal as in the Metropolis random walk case. It will be shown through simulated examples that the procedure works well in situations where the prior is relatively strong compared to the likelihood.

The next section reviews dynamic models as a useful framework for the analysis of non-normal time series or panel data. MCMC simulation by conditional prior proposals is discussed in section 3. Some simulation results are given for a data set, known to be problematic for the single move algorithm. Furthermore a comparison with a Gibbs block move algorithm is given for the special case of Gaussian observations. The goal is here to assess how much statistical efficiency is lost for our proposal, which is built independently from observations y. Finally, extensions of the transition model to errors within the class of t-distributions are discussed in section 4. Such models allow abrupt jumps in the transition model, so-called innovative outliers. As a final example, we analyse a binary time series with an additional hyperprior on the degrees of freedom of the t-distribution.

2. Dynamic models

Let $y = (y_1, \ldots, y_T)$ denote the sequence of observations and $\alpha = (\alpha_1', \ldots, \alpha_T')'$ the sequence of state parameters. We assume that $\alpha_t|\alpha_{-t}$, Q_t $(t = z+1, \ldots, T)$ has a Gaussian distribution with mean $-F_1\alpha_{t-1} - F_2\alpha_{t-2} - \cdots - F_z\alpha_{t-z}$ and dispersion Q_t. Here α_{-t} denotes the sequence $(\alpha_{t-z}', \ldots, \alpha_{t-1}')'$, the matrices $F_1, \ldots, F_z$ are assumed to be known. In some models, for example in the state space representation of spline priors (Kohn & Ansley, 1987), a more general specification is needed with matrices $F_1, \ldots, F_z$ also depending on time t. In other applications the matrices might be (partially) unknown and could be estimated within an extended MCMC algorithm. We keep the simpler form here for reasons of presentation.

Let Q denote the sequence of dispersions $Q_{z+1}, \ldots, Q_T$. We place flat priors on the initial values $\alpha_1, \ldots, \alpha_z$, which gives

$$\mathrm{p}(\alpha|Q) \propto \prod_{t=z+1}^{T} \mathrm{p}(\alpha_t|\alpha_{-t}, Q_t)$$

as the Gaussian (vector) autoregressive prior of lag z for α. It is often called the transition model. Conditional independence of $y_t|\alpha_t,\ t = 1, \ldots, T$, gives the posterior distribution as

$$\mathrm{p}(\alpha, Q|y) \propto \prod_{t=1}^{T} \mathrm{p}(y_t|\alpha_t) \times \mathrm{p}(\alpha|Q) \times \mathrm{p}(Q)$$

with some hyperprior $\mathrm{p}(Q)$, independent of α and y.

Typical examples of such transition models (with time-constant variance Q) are first ($z = 1$) and second ($z = 2$) order random walks

$$\alpha_t|\alpha_{-t}, Q \sim N(\alpha_{t-1}, Q),$$

$$\alpha_t|\alpha_{-t}, Q \sim N(2\alpha_{t-1} - \alpha_{t-2}, Q),$$

or seasonal models $\alpha_t|\alpha_{-t}, Q \sim N(-\alpha_{t-1} - \alpha_{t-2} - \cdots - \alpha_{t-z}, Q)$ with period $z + 1$.

Gaussian autoregressive priors can be written as

$$\mathrm{p}(\alpha|Q) \propto \exp\left(-\frac{1}{2}\alpha' K \alpha\right)$$

with a so-called penalty matrix K. Note, that $\mathrm{p}(\alpha|Q)$ is improper due to the flat priors for the initial parameters $\alpha_1, \ldots, \alpha_z$; therefore K^{-1} does not exist. For the random walks given above the corresponding penalty matrices are

$$K = \frac{1}{Q}\begin{pmatrix} 1 & -1 & & & & & & \\ -1 & 2 & -1 & & & & & \\ & -1 & 2 & -1 & & & & \\ & & \vdots & \vdots & \vdots & & & \\ & & & -1 & 2 & -1 & & \\ & & & & -1 & 2 & -1 & \\ & & & & & -1 & 1 \end{pmatrix}$$

and

$$K = \frac{1}{Q}\begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \vdots & \vdots & \vdots & \vdots & \vdots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix},$$

respectively.

The penalty matrix K plays a key role in the derivation of the conditional distribution of a subvector of α. Let F_0 denote the identity matrix. Defining

$$F = \begin{pmatrix} F_z & F_{z-1} & \ldots & F_1 & F_0 & & \\ & F_z & F_{z-1} & \ldots & F_1 & F_0 & \\ & & \vdots & & & & \vdots \\ & & & F_z & F_{z-1} & \ldots & F_1 & F_0 \end{pmatrix}$$

and the block-diagonal matrix

$$Q = \begin{pmatrix} Q_{z+1} & & & \\ & Q_{z+2} & & \\ & & \ddots & \\ & & & Q_T \end{pmatrix},$$

it follows that $K = F'Q^{-1}F$. Since Q is symmetric, so is K. Furthermore, it can be shown that the elements of

$$K = \begin{pmatrix} k_{11} & k_{12} & \dots & k_{1T} \\ k_{21} & k_{22} & \dots & k_{2T} \\ \vdots & & & \vdots \\ k_{T1} & k_{T2} & & k_{TT} \end{pmatrix}$$

are given by

$$k_{t,t+s} = \sum_{j=\max(0,s,1+z-t)}^{\min(z,z+s,T-t)} F_j' Q_{t+j}^{-1} F_{j-s}, \quad |s| \leqslant z, \tag{1}$$

with zero elements for $|s| > z$.

We think of dynamic models as a module for flexible Bayesian analysis, which can be conveniently combined with other priors such as priors for the level of the sequence α, random effect priors for modelling heterogeneity among several units $y_1, \dots, y_n$ with $y_i = (y_{i1}, \dots, y_{iT})$, or priors for spatial dependence. Another useful extension is to allow for a non-zero prior trend in the state sequence, as a referee has noted. For example, the first-order random walk can be extended to

$$\alpha_t | \alpha_{-t}, Q, \tau \sim N(\alpha_{t-1} + \tau, Q),$$

in which τ is an unknown trend parameter. A recent review of dynamic models is given in Fahrmeir & Knorr-Held (1998), which also points out connections to non- and semiparametric smoothing methods.

Applications of dynamic models are widespread. Fahrmeir & Tutz (1994a) discuss smoothing of categorical time series, panel and survival data. Fahrmeir & Tutz (1994b) introduce dynamic models for ordered paired comparison data. Duration data is covered in Fahrmeir & Knorr-Held (1997). Breslow & Clayton (1993) and Clayton (1996) discuss biostatistical applications with second order random walk priors in mixed models, which is related. Berzuini & Clayton (1994) propose second order random walk priors in survival models with multiple time scales. Berzuini & Larizza (1996) use dynamic models for joint modelling of time series and failure time data. Besag *et al.* (1995) use second order random walk priors in age–period–cohort models. Finally Knorr-Held & Besag (1998) use dynamic models for time–space mapping of disease risk data. Most of these references use binomial or multinomial logistic or log-linear Poisson models as the observation model. For panel and survival data, several units $i = 1, \dots, n_t$ are observed at each time t, and conditional independence is usually assumed for $y_{ti} | \alpha_t$, $i = 1, \dots, n_t$.

3. MCMC simulation with conditional prior proposals

Our MCMC implementation is based on updating using full conditionals with the Hastings algorithm as described in full detail in Besag *et al.* (1995); we also use their terminology. We denote full conditionals by $p(\alpha_t | \quad)$, for example. We start this section with a technical note about the conditional distribution of $\alpha_a, \dots, \alpha_b$, given $\alpha_1, \dots, \alpha_{a-1}$ and $\alpha_{b+1}, \dots, \alpha_T$. Then the single and the block move with conditional prior proposals is introduced. We close with several simulation results.

3.1. Conditional properties of autoregressive priors

The conditional distribution of a subvector of α, given the rest of α plays a key role in our algorithm. Let α_{ab} denote the subvector $(\alpha_a', \alpha_{a+1}', \ldots, \alpha_b')$ and K_{ab} denote the submatrix out of K, given by the rows and columns a to b. Finally, let $K_{1,a-1}$ and $K_{b+1,T}$ denote the matrix to the left and right of K_{ab}, respectively:

$$K = \begin{pmatrix} & K_{1,a-1}' & \\ K_{1,a-1} & K_{ab} & K_{b+1,T} \\ & K_{b+1,T}' & \end{pmatrix}.$$

Then the following result can be proved by simple matrix manipulations: the conditional distribution of α_{ab}, given $\alpha_{1,a-1}$ and $\alpha_{b+1,T}$ is normal $N(\mu_{ab}, \Sigma_{ab})$ with moments

$$\mu_{ab} = \begin{cases} -K_{ab}^{-1} K_{b+1,T} \alpha_{b+1,T} & a = 1 \\ -K_{ab}^{-1} K_{1,a-1} \alpha_{1,a-1} & b = T \\ -K_{ab}^{-1}(K_{1,a-1}\alpha_{1,a-1} + K_{b+1,T}\alpha_{b+1,T}) & \text{otherwise} \end{cases} \tag{2}$$

and

$$\Sigma_{ab} = K_{ab}^{-1}. \tag{3}$$

It can be seen from (2) in connection with (1) that only $\alpha_{a-z}, \ldots, \alpha_{a-1}$ and $\alpha_{b+1}, \ldots, \alpha_{b+z}$ enter in μ_{ab}, since all elements in K outside the z off-diagonals are zero. We always make use of this property to reduce the computation involved in the multiplications $K_{1,a-1}\alpha_{1,a-1}$ and $K_{b+1,T}\alpha_{b+1,T}$.

3.2. Single move

The most natural blocking strategy for α is to update α_t one at a time. The main advantage is that the full conditional has a simple form, achieved by the hierarchical structure of the model:

$$\mathrm{p}(\alpha_t | \quad) \propto \mathrm{p}(y_t|\alpha_t) \times \mathrm{p}(\alpha_t|\alpha_{s\neq t}, Q).$$

One way to update α_t is to use a proposal α_t^*, distributed as $\mathrm{p}(\alpha_t|\alpha_{s\neq t}, Q)$. Such a "conditional prior proposal" is independent of the current state of α_t but, in general, depends on the current states of other parameters (here $\alpha_{s\neq t}$ and Q). Note, that "Gibbs proposals", i.e. samples from the full conditional, have exactly the same "conditional independence" property.

It is illustrative to discuss differences between conditional and unconditional independence proposals (Tierney, 1994). It is often very difficult, at least for higher dimensions and non-normal models, to construct an unconditional independence proposal with acceptance rates not too small. In contrast, a conditional proposal is far more constrained than the unconditional version because it depends on the current state of neighbouring parameters. However, the conditional proposal is still very flexible because its distribution changes at each iteration whenever neighbouring parameters are updated and accepted. (Unconditional independence proposals are generated from exactly the same distribution in every iteration step). If the states α_t are *a priori* independent, however, conditional prior proposals do not depend on neighbouring parameters, so they are no longer conditional but now unconditional independence proposals. The proposed method will not work in this case, as a referee has noted.

The Hastings acceptance probability simplifies for the conditional prior proposal to

$$\min\left\{1, \frac{\mathrm{p}(y_t|\alpha_t^*)}{\mathrm{p}(y_t|\alpha_t)}\right\},$$

the likelihood ratio for observation y_t. Conditional prior proposals have a natural interpretation: α_t^* is drawn independently of the observation model and just reflects the specific autoregressive prior specification. If it produces improvement in the likelihood at time t, it will always be accepted, if not, then the acceptance probability is equal to the likelihood ratio.

Of course, a simple random walk proposal can be used instead, but it has to be tuned. Other single move updating schemes are more demanding in their proposals and require more effort to calculate the acceptance probability. Shephard & Pitt (1997) use a predefined number of iterations (two to five) to calculate a reasonably good approximation to the mode of $\mathrm{p}(\alpha_t|\ \ )$ for every updating step. The approximative mode and the curvature are used in an analytic Taylor expansion to build a specific Gaussian (conditional) independence proposal and to perform a pseudo rejection sampling step (Tierney, 1994). The advantage is that the proposal takes the observation y_t into account for the cost of considerably more computational effort. The pseudo rejection sampling step avoids additional iterations, which are necessary for a real Gibbs step in a rejection sampling procedure.

However, the single move blocking scheme might be very slowly converging, especially if neighbouring parameters are highly correlated. This is typically the case when the likelihood at time t is very flat in α_t and does not give much information relative to the autoregressive prior specification. Smoothing of binary time series is a typical example. A simple modification of the single move conditional prior algorithm addresses this problem without losing its simplicity both in programming and computing time.

3.3. Block move

Instead of updating one parameter α_t at a time, the block move is based on updating one block $\alpha_{rs} = (\alpha_r', \ldots, \alpha_s')$ at a time, following suggestions of Shephard & Pitt (1997). The number of blocks may range from 2 up to T, which corresponds to the single move. Consider the breakpoints that divide α into blocks as fixed for the moment. The idea of this blocking strategy is to use blocks that are large enough to ensure a good mixing and convergence behaviour. So what kind of proposals are useful for the block move?

It is not clear how to choose the spread of a multivariate Metropolis random walk proposal because correlations between parameters are unknown. But, in contrast, the generalization of the conditional prior proposal is straightforward: the simple structure of the full conditional is retained, since $\mathrm{p}(\alpha_{rs}|\alpha_{1,r-1}, \alpha_{s+1,T}, Q)$ is still normal with known moments (see section 3.1). Therefore a conditional prior proposal can be implemented similarly as in the previous section: generate α_{rs}^* distributed as $\mathrm{p}(\alpha_{rs}|\alpha_{1,r-1}, \alpha_{s+1,T}, Q)$ to update the full conditional

$$\mathrm{p}(\alpha_{rs}|\ \ ) \propto \prod_{t=r}^{s} \mathrm{p}(y_t|\alpha_t) \times \mathrm{p}(\alpha_{rs}|\alpha_{1,r-1}, \alpha_{s+1,T}, Q).$$

The acceptance probability simplifies again to a likelihood ratio

$$\min\left\{1, \frac{\prod_{t=r}^{s} \mathrm{p}(y_t|\alpha_t^*)}{\prod_{t=r}^{s} \mathrm{p}(y_t|\alpha_t)}\right\}.$$

The block move provides a considerable improvement in situations where the single move has bad mixing behaviour. However, fixed blocks still cause convergence and mixing problems for parameters close to a breakpoint. Changing the block configuration in every iteration cycle is a simple remedy. This can be done either by a deterministic or a random scheme. In all following examples we use random blocking with a fixed standard block size. The first block has uniform random block size between 1 and the standard block length. All following blocks have the same standard size except for the last block. So, most of the updating involves blocks of a fixed block length, which has computational advantages, since the dispersion matrix K_{ab}^{-1} and the corresponding Cholesky decomposition of the standard block size full conditional can be computed in advance, at least for Gaussian transition models with time-constant dispersion Q. Nevertheless, calculation of μ_{ab} and Σ_{ab} via (2) and (3) may become computationally demanding for large blocks α_{ab}. In this case it will be useful to exploit the specific structure of K fully to implement a more numerically efficient version. Finally we note the block sizes proportional to the number of observations n_t per block may be considered in situations where n_t is changing over time as in survival models (Fahrmeir & Knorr-Held, 1997).

Shephard & Pitt (1997) propose a different proposal in the block move, which is similar to their version of the single move proposal. They use several additional Fisher scoring iterations within each updating step to get a reasonably good approximation to the mode of $p(\alpha_{rs}|\ \ )$ and add a pseudo rejection sampling step. These iterations can be rather time-consuming, especially for multivariate observation models such as models for multicategorical responses (e.g. Fahrmeir & Tutz, 1994a, ch. 3). In contrast, the conditional prior algorithm benefits of block updating without spending too much effort in the construction of the proposal.

3.4. An example: Tokyo rainfall data

To illustrate the gain of the block move, we analyse the Tokyo rainfall data (e.g. Fahrmeir & Tutz, 1994a), a single binomial time series of length $T = 366$. We assume a binomial logit model

$$y_t|\alpha_t \sim \begin{cases} B(2, \pi_t) & t \neq 60 \\ B(1, \pi_t) & t = 60 \end{cases}, \quad \pi_t = 1/(1 + \exp(-\alpha_t)),$$

with a second order random walk prior for $\{\alpha_t\}$. A highly dispersed, but proper inverse gamma prior was chosen for the random walk variance Q and updating of Q was implemented with a Gibbs step. The prior reflects sufficient ignorance about Q but avoids problems arising with improper posteriors. Figure 1 displays the data and some characteristics of the posterior distribution of $\{\pi_t\}$.

We separate our empirical analysis into two parts, speed of convergence and efficiency of estimation. First we focus on the empirical convergence behaviour. For block size 1, 5, 20 and 40 we computed the average trajectories of 100 parallel chains after 10, 50, 100 and 500 iterations, which are shown in Fig. 2. For every chain, the state parameters were initialized to zero and the variance Q to 0.1.

Figure 2 shows clear empirical evidence that the block move converges much faster for bigger block sizes, at least for this data set and model. The single move algorithm does not converge at all, at least for the first 500 iterations. The algorithm with blocksize 40 seems to have reached equilibrium after only 50 iterations. We also computed the average acceptance rate of the Hastings steps, averaged over all α_ts. The rates were 99.4% (block size 1), 94.4% (5), 65.5% (20) and 35.3% (40), indicating decreasing acceptance rates with increasing block size.

We repeated the same analysis, assuming a random walk of first order instead. Convergence was a bit faster and, again, the block move algorithm exhibited superior convergence performance.

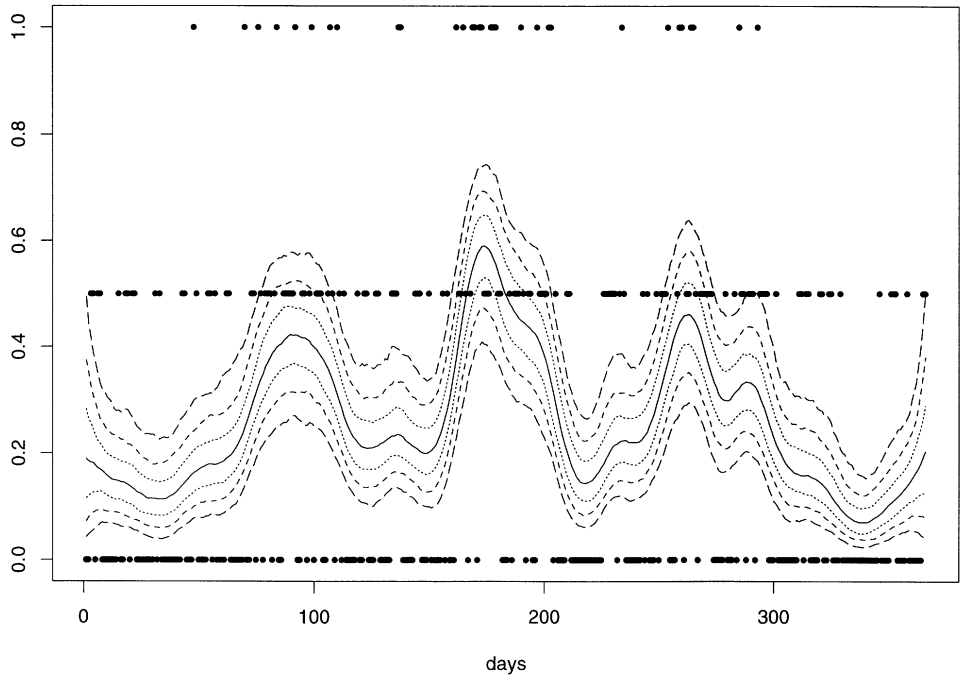

Fig. 1. Tokyo rainfall data. Data and fitted probabilities (posterior median within 50, 80 and 95% pointwise credible regions). The data is reproduced as relative frequencies with values 0, 0.5 and 1.

A measure of efficiency of estimation are the autocorrelations of parameters of the simulated Markov chain after reaching equilibrium. The larger these correlations are, the larger the variances of the estimate of the posterior mean. We started the chain in equilibrium, ran it for 1000 iterations and stored every 10th sample until we had 10,000 samples. We calculated autocorrelations for 12 parameters, namely for $t = 1, 33, 67, 100, 133, 167, 200, 233, 267, 300, 333, 366$ and for the hyperparameter Q. We did this analysis twice, for block size 1 and block size 20, both assuming a second order random walk prior. The results can be summarized as follows: For block size 1, all autocorrelations up to lag 40 were larger than 0.5. In contrast, for block size 20, the autocorrelations of all parameters considered were close to zero for lag 5 and higher. Autocorrelations for the hyperparameter Q were somewhat larger (around zero for lag 20 and higher) but still much smaller than for block size 1.

Figure 3 shows trajectories of the last 2000 iterations for three representative parameters α_1, α_{100}, α_{333} and the variance Q. Whereas the mixing behaviour of the block size 1 algorithm is catastrophic, the block size 20 algorithm shows well-behaved mixing. The plots for the other parameters look very similar.

3.5. A comparison with a Gibbs block move for Gaussian observations

To gain more insight into the behaviour of the proposed methodology, we add an extended study for the simple state space model with Gaussian observation model

$$\alpha_t \sim N(\alpha_{t-1}, Q) \quad t = 2, \ldots, T,$$
$$y_t \sim N(\alpha_t, R) \quad t = 1, \ldots, T,$$

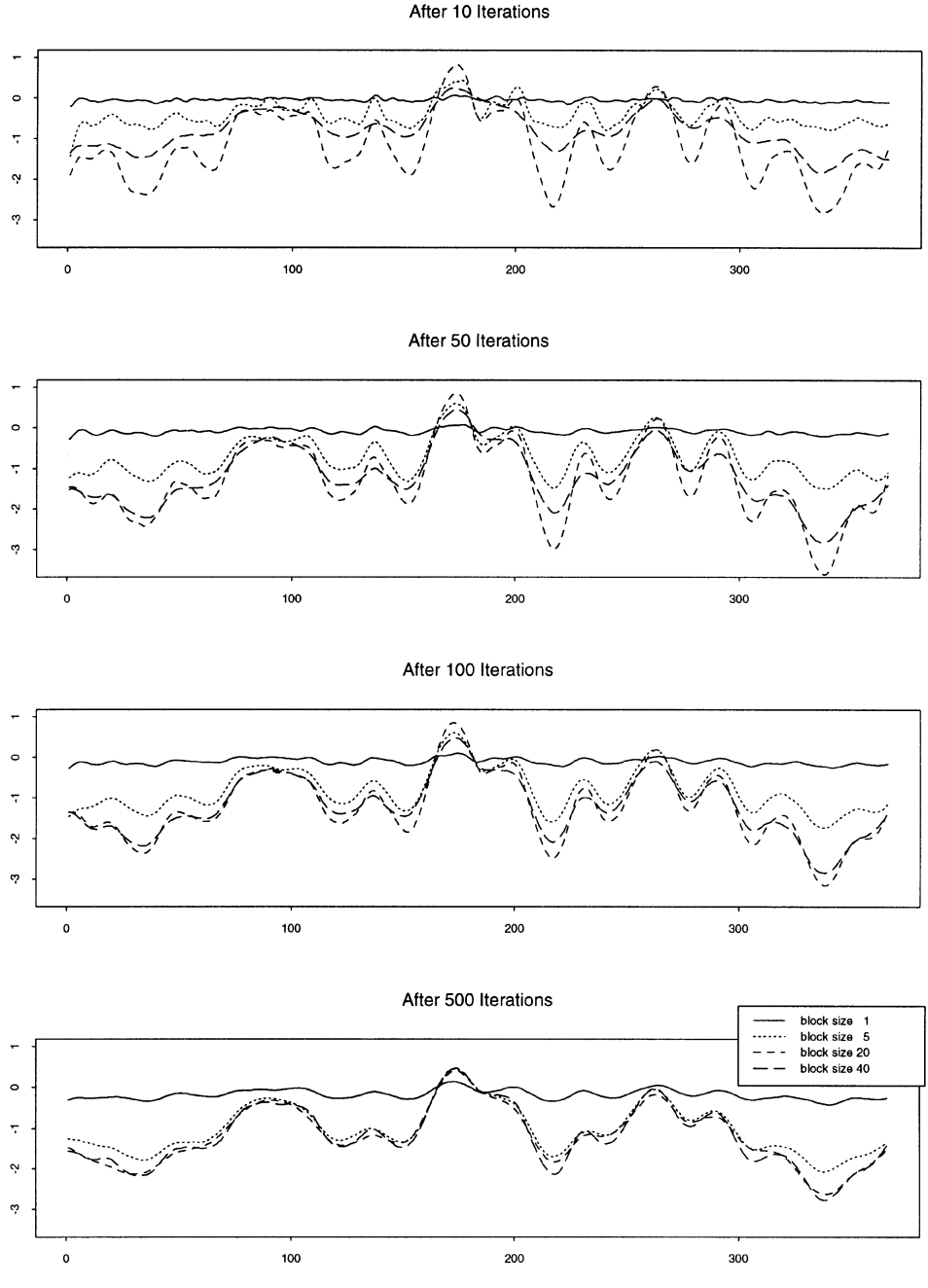

Fig. 2. Speed of convergence of the block move algorithm for different block sizes.

and values $R = 0.01$ and $T = 1000$. The value of Q and the block size is chosen in various combinations.

This model allows us to implement a block move algorithm which samples from the full conditionals due to the Gaussian observation model. Thus we can compare the conditional prior proposal methodology with a more standard Gibbs type block move algorithm. Note that the

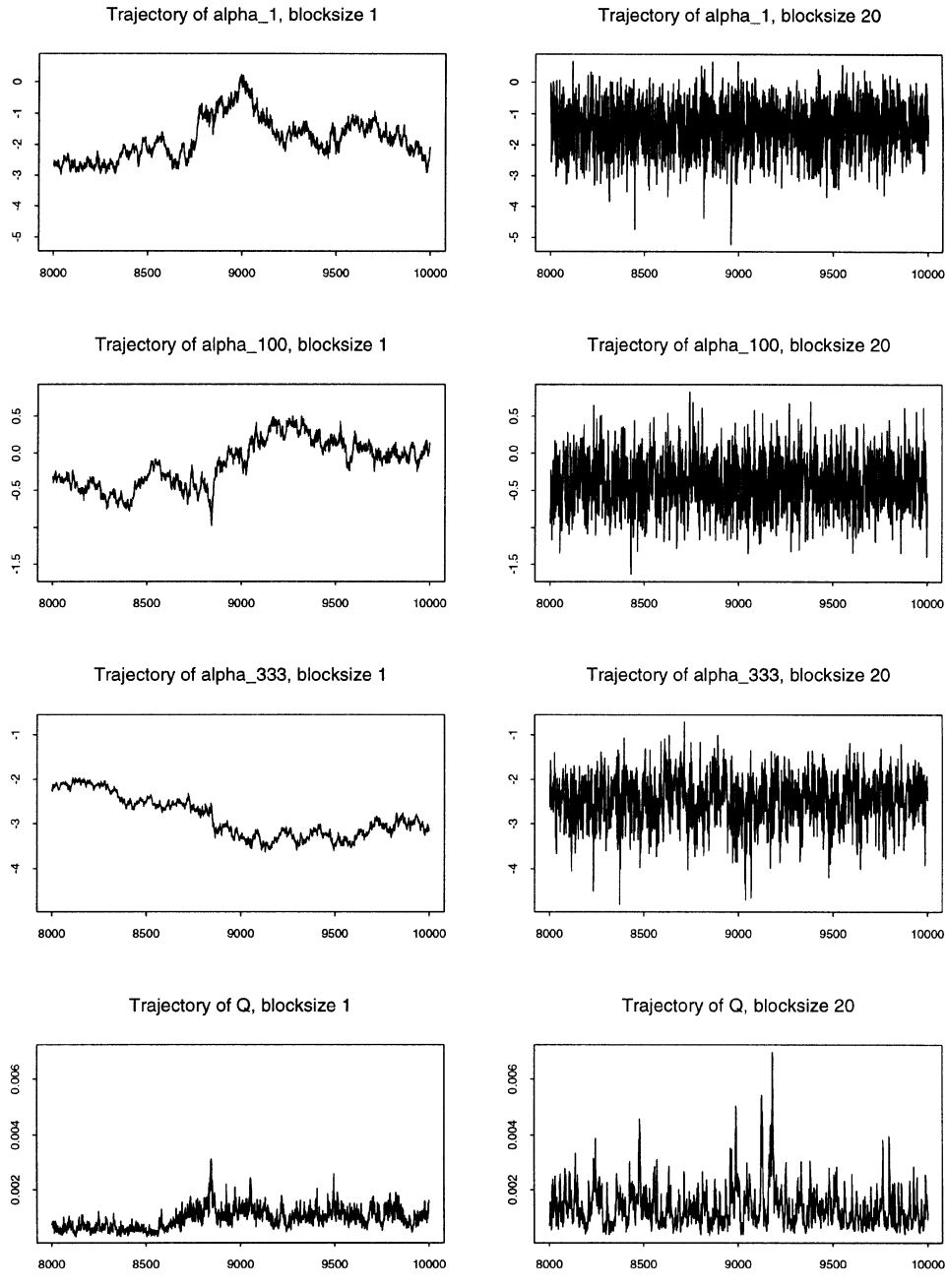

Fig. 3. Trajectories of α_1, α_{100}, α_{333} and Q for block size 1 and block size 20.

Gibbs sampler uses the observed values $y_a, \ldots, y_b$ in the construction of the proposal, whereas the observed values enter in the conditional prior proposal algorithm only in the calculation of the acceptance probabilities but not in the construction of the proposal. We have calculated acceptance rates and estimated autocorrelations for lag 1, 10 and 25 for every parameter $\alpha_1, \ldots, \alpha_{1000}$. Table 1 reports those quantities averaged over all 1000 parameters $\alpha_1, \ldots, \alpha_{1000}$

Table 1. *Results for the Gaussian state space model*

		Conditional prior proposal		Gibbs sampler
			ACF1 ACF10 ACF25	ACF1 ACF10 ACF25
Q	Block sise	Acceptance rate (in %) Mean (max, min)	Mean (max, min)	Mean (max, min)
1	1	12.72 (21.02, 0.00)	0.83 (1.00, 0.0) 0.21 (0.96, −0.13) 0.05 (0.91, −0.22)	0.00 (0.11, −0.11) 0.00 (0.10, −0.09) 0.00 (0.10, −0.10)
0.01	1	70.51 (85.81, 13.65)	0.49 (0.95, 0.25) 0.03 (0.56, −0.12) 0.00 (0.25, −0.14)	0.25 (0.34, 0.15) 0.00 (0.13, −0.10) 0.00 (0.09, −0.11)
0.01	3	36.53 (57.32, 3.91)	0.64 (0.97, 0.37) 0.06 (0.79, −0.16) 0.01 (0.61, −0.19)	0.11 (0.20, −0.01) 0.00 (0.12, −0.11) 0.00 (0.09, −0.12)
0.01	10	3.38 (23.66, 0.00)	0.95 (1.00, 0.0) 0.66 (0.97, 0.0) 0.41 (0.92, −0.14)	0.03 (0.13, −0.08) 0.00 (0.11, −0.10) 0.00 (0.09, −0.12)
0.0001	1	96.77 (99.82, 88.54)	0.89 (0.96, 0.76) 0.61 (0.83, 0.22) 0.42 (0.77, 0.09)	0.88 (0.95, 0.74) 0.58 (0.83, 0.28) 0.38 (0.75, 0.00)
0.0001	3	91.85 (98.09, 82.44)	0.84 (0.92, 0.72) 0.48 (0.70, 0.21) 0.26 (0.55, −0.09)	0.82 (0.89, 0.72) 0.43 (0.66, 0.18) 0.21 (0.52, −0.08)
0.0001	10	76.41 (87.99, 59.33)	0.72 (0.84, 0.58) 0.20 (0.46, −0.02) 0.03 (0.22, −0.17)	0.62 (0.73, 0.46) 0.10 (0.25, −0.08) 0.00 (0.17, −0.13)
0.0001	30	41.35 (56.51, 22.48)	0.69 (0.86, 0.54) 0.08 (0.37, −0.10) 0.00 (0.19, −0.21)	0.30 (0.40, 0.19) 0.00 (0.11, −0.11) 0.00 (0.09, −0.12)
0.000001	1	99.67 (100.00, 98.18)	0.93 (0.98, 0.84) 0.76 (0.94, 0.46) 0.62 (0.90, 0.21)	0.93 (0.98, 0.83) 0.75 (0.93, 0.37) 0.61 (0.88, 0.08)
0.000001	10	97.53 (99.55, 94.36)	0.93 (0.97, 0.82) 0.77 (0.88, 0.42) 0.63 (0.84, 0.23)	0.91 (0.97, 0.73) 0.69 (0.89, 0.26) 0.53 (0.81, 0.04)
0.000001	100	77.97 (85.99, 70.06)	0.75 (0.87, 0.61) 0.24 (0.50, 0.05) 0.05 (0.31, −0.14)	0.70 (0.79, 0.56) 0.22 (0.39, 0.07) 0.07 (0.16, −0.02)

as well as the maximum and the minimum value. The last column gives estimated autocorrelations for the corresponding Gibbs block move sampler with the same block size.

The results can be summarized as follows. For situations, where the prior is rather weak compared to the information in the likelihood ($Q = 1$), the conditional prior proposal in combination with the single move has rather poor performance with very low acceptance rates. In fact, for some parameters, proposals have been rejected for the whole run. For the corresponding outlying observations, the likelihood is not supported by the conditional prior, hence the posterior is substantially different from the conditional prior, even for possibly changing neighbouring parameters. The Gibbs block move sampler, in contrast, has very good performance

with virtually independent samples. In fact, this is no surprise, since the posterior will not differ much from the likelihood, so the states α_t are close to independence even in the posterior.

For $Q = 0.01$, we observe the same phenomenon, but less distinct. For block size equal to 1, the conditional prior proposal has better performance but is still outperformed by the Gibbs sampler. Keeping every 25th sample of the conditional prior proposal algorithm seems to be roughly equivalent to keeping every 10th by Gibbs sampling. Increasing the block size does not improve the performance of the proposed methodology, it seems that acceptance rates are too close to zero.

However, for situations where the prior is strong relative to the likelihood ($Q = 0.0001$ and $Q = 0.000001$) both methods perform similarly. The gain of the block move can be seen both from the Gibbs sampler as well from the conditional prior algorithm. For $Q = 0.000001$ a block size around 100 seems to be necessary for good performance. Note that the small differences in the estimated autocorrelations between both methods are slightly increasing for increasing block size. This feature is probably caused by the fact that increasing block size goes along with decreasing acceptance rates for the conditional prior proposal, which automatically increases autocorrelations to some extent.

The results suggest that, whenever the prior is relatively strong compared to the likelihood, resulting in strong dependence among neighbouring parameters, ignoring information from observations in constructing the proposal does not do serious harm in terms of statistical efficiency on a cycle per cycle basis. The main advantage of the proposed algorithm is that it is simpler and faster per cycle. It will therefore be more efficient in terms of CPU time, which is a more appropriate basis for comparisons, as noted by Besag (1994) and Tierney (1994) in the rejoinder.

For situations where there is low dependence between neighbouring parameters, an algorithm, which incorporates information from observations will outperform the proposed methodology. However, low correlation systems are less frequent (Shephard, 1994). In particular, for most categorical data, dependence is usually strong among parameters. It may, however, be sometimes worth exploring a hybrid scheme, where conditional prior proposals are combined with more elaborate proposals that take observations into account.

For practical implementation of the conditional prior proposal it will be useful to monitor acceptance rates for every parameter. Acceptance rates too close to one suggest a bigger block size, whereas acceptance rates too small indicate a block size too large. Theoretical considerations similar to the results of Gelman *et al.* (1995) would be very helpful to determine an optimal acceptance rate for tuning the algorithm.

4. Hierarchical *t*-transition models

The temporal variation of underlying parameters may have jumps, so-called innovative outliers. The Gaussian distributional assumption in the autoregressive prior, however, does not allow such abrupt movement. Distributions with heavier tails such as *t*-distributions are more adequate. In this section we will sketch how autoregressive priors can be extended via an hierarchical *t*-formulation with unknown degrees of freedom (Besag *et al.*, 1995).

4.1. Autoregressive t-distributed priors

Introducing hyperparameters $\gamma = (\gamma_{z+1}, \ldots, \gamma_T)'$, the autoregressive prior formulation can be extended to

$$\alpha_t | \alpha_{-t}, Q, \gamma_t \sim N\left(-\sum_{l=1}^{z} F_l \alpha_{t-l}, Q/\gamma_t \right), \quad t = z+1, \ldots, T.$$

Assuming $\gamma_t|\nu$ to be independently gamma distributed $\gamma_t \sim G(\nu/2, \nu/2)$, $\alpha_t|\alpha_{-t}, Q$ has a t-distribution with ν degrees of freedom.

The distribution $p(\alpha|\gamma, Q)$ can be expressed again in a penalty formulation with a penalty matrix K, now depending on γ, too. The elements in K have the same form as in (1) with $Q_t = Q/\gamma_t$. For example, the matrix

$$K = \frac{1}{Q}\begin{pmatrix} \gamma_2 & -\gamma_2 & & & & & & \\ -\gamma_2 & \gamma_2+\gamma_3 & -\gamma_3 & & & & & \\ & -\gamma_3 & \gamma_3+\gamma_4 & -\gamma_4 & & & & \\ & & \vdots & \vdots & \vdots & & & \\ & & & -\gamma_{T-2} & \gamma_{T-2}+\gamma_{T-1} & -\gamma_{T-1} & & \\ & & & & -\gamma_{T-1} & \gamma_{T-1}+\gamma_T & -\gamma_T & \\ & & & & & -\gamma_T & \gamma_T \end{pmatrix}$$

corresponds to a first order random walk t-transition model.

4.2. A second example: sleep data

Carlin & Polson (1992) present an analysis of a binary time series of length $T = 120$ min. The outcome variable y_t corresponds to the sleep status (REM ($y_t = 1$) or non-REM) of a specific child. We reanalyse this data to illustrate the hierarchical t-formulation. The response variable is assumed to depend on a latent "sleep status" α_t via a dynamic logistic model. We assume α_t to follow a first order hierarchical t-random walk and place an equally weighted hyperprior $p(\nu)$ on the values $\{2^k, k = -1, -0.9, -0.8, \ldots, 6.9, 7.0\}$. For updating ν, we use a discrete Metropolis random walk proposal which gives equal weight to the two neighbours of the current value. Note that for the limit cases $\nu = 0.5$ and $\nu = 128$, the proposal becomes deterministic, proposing the only neighbour. The acceptance probability has to be modified adequately for proposed jumps to or away from these limit values. All other hyperparameters are updated with Gibbs steps.

The following analysis is based on a run of length 505,000, discarding the first 5000 values and storing every 100th thereafter. The standard block length was chosen as 10 which resulted in an average acceptance rate of 68.6%. Starting values were zero for all α_ts. Since the posterior might be multimodal the chain might stay in one part of the posterior for a long time. To account for that we started several chains with different values for ν over the whole range of the prior: 0.5 to 128. However, all of these chains moved after not more than 1000 iterations into the region around $\nu = 1$.

Figure 4 shows the data and estimates. Note that our model formulation gives a significantly better fit to the data than the analysis by Carlin & Polson (1992, fig. 1, p. 583). The resulting posterior for the hyperparameter ν has its mode at $\nu = 2^{-0.3} \approx 0.81$. The 90 and 95% credible regions for ν are [0.66, 3.3] and [0.54, 13.0], respectively, showing strong evidence for highly non-normal system disturbances. The estimates of the sequence $\{\alpha_t\}$, the latent sleep status, exhibit some huge abrupt jumps, e.g. around $t = 53$ and $t = 62$. Note that the posterior of α_t is highly skewed for some values of t.

5. Discussion

Conditional prior proposals reflect the dependence of underlying parameters and therefore provide a useful tool for highly dependent parameters in dynamic models. The resulting algorithm is appealing since all proposals are easy to generate and all acceptance probabilities are easy to calculate. The choice of a blocking strategy serves as a tuning device.

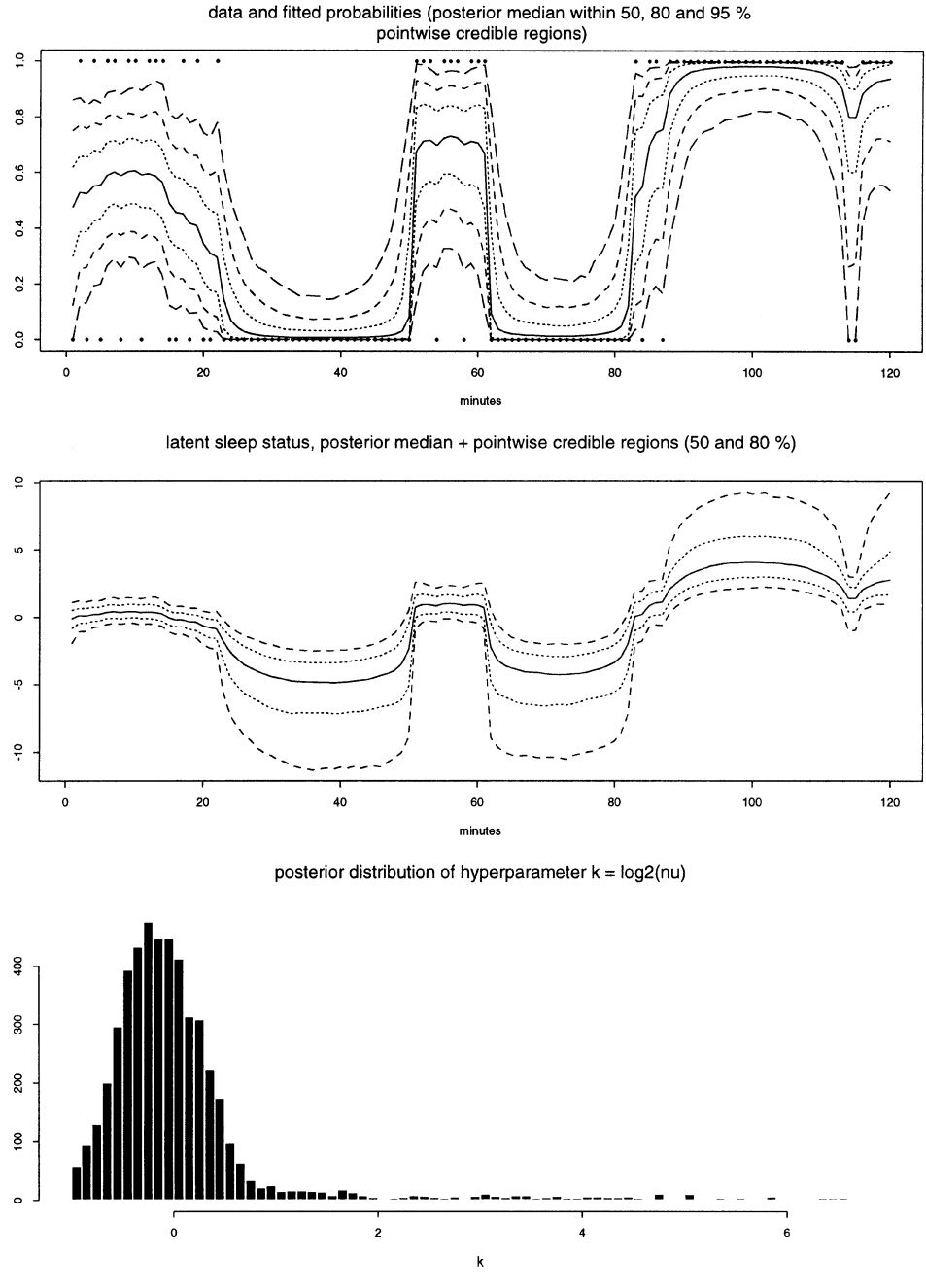

Fig. 4. Sleep data. Data and estimates.

We have also experimented with conditional prior proposals in dynamic models, where $p(\alpha)$ is a product of several autoregressive prior specifications. For example, each component of α_t may correspond to a certain covariate effect (plus intercept) and independent random walk priors are assigned to all components. Here two generalizations are possible: either updating

each component within its own blocking strategy or updating all components within one blocking strategy. The former approach provides more flexibility in tuning the algorithm and has been successfully implemented for duration time data. However, the latter is faster, especially for large dimension of α_t and is usually sufficiently accurate.

There might also be a wide field of applications in models for non-normal spatial data, e.g. Besag *et al.* (1991). Here intrinsic (or undirected) autoregressions replace directed autoregressions. Conditional prior proposals can be implemented in similar lines, since intrinsic autoregressions can be written in a penalty formulation as well, see Besag & Kooperberg (1995).

Acknowledgements

Part of this research was done during a visit to the Department of Statistics, University of Washington, Seattle, USA, whose hospitality is gratefully acknowledged. The visit was supported by a grant of the German Academic Exchange Service (DAAD). The author would like to thank Julian Besag for frequent discussions and helpful comments on a first version of this paper. The revision of the paper has substantially benefited from comments from the editor and two referees as well as from discussions with Ludwig Fahrmeir, Dani Gamerman and Neil Shephard.

References

Besag, J. E. (1994). Contribution to the discussion of the paper by Tierney (1994). *Ann. Statist.* **22**, 1734–1741.

Besag, J. E., Green, P. J., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10**, 3–66.

Besag, J. E. & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**, 733–746.

Besag, J. E., York, J. C. & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43**, 1–59.

Berzuini, C. & Clayton, D. (1994). Bayesian analysis of survival on multiple time scales. *Statist. Med.* **13**, 823–838.

Berzuini, C. & Larizza, C. (1996). A unified approach for modelling longitudinal and failure time data, with application in medical monitoring. *IEEE Trans. Pattern Anal. Machine Intell.* **18**, 109–123.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9–25.

Carlin, B. P. & Polson, N. G. (1992). Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian statistics 4* (eds J. Bernardo, J. Berger, A. P. Dawid & A. F. M. Smith), 577–586. Oxford University Press, Oxford.

Carlin, B. P., Polson, N. G. & Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space-modeling. *J. Amer. Statist. Assoc.* **87**, 493–500.

Carter, C. K. & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553.

Clayton, D. G. (1996). Generalized linear mixed models. In *Markov chain Monte Carlo in practice* (eds W. R. Gilks, S. Richardson & D. J. Spiegelhalter), 275–301. Chapman & Hall, London.

Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *J. Amer. Statist. Assoc.* **87**, 501–509.

Fahrmeir, L. & Knorr-Held, L. (1997). Dynamic discrete time duration models. *Sociological Methodology*, Vol. 27 (ed. A. E. Raftery), 417–452. Blackwell Publishers, Boston.

Fahrmeir, L. & Knorr-Held, L. (1998). Dynamic and semiparametric models. In *Smoothing and regression: approaches, computation and application* (ed. M. G. Schimek). Wiley, New York (forthcoming).

Fahrmeir, L. & Tutz, G. (1994a). *Multivariate statistical modelling based on generalized linear models.* Springer, New York.

Fahrmeir, L. & Tutz, G. (1994b). Dynamic stochastic models for time-dependent ordered paired comparison systems. *J. Amer. Statist. Assoc.* **89**, 1438–1449.

Fahrmeir, L. & Wagenpfeil, S. (1997). Penalized likelihood estimation and iterative Kalman smoothing for non-Gaussian dynamic regression models. *Comput. Statist. Data Anal.* **24**, 295–320.

Fahrmeir, L., Hennevogl, W. & Klemme, K. (1992). Smoothing in dynamic generalized linear models by Gibbs sampling. In *Advances in GLIM and statistical modelling* (eds L. Fahrmeir, B. Francis, R. Gilchrist & G. Tutz), 85–90. Springer, New York.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Ser. Anal.* **15**, 183–202.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika* **85**, 215–227.

Gelman, A., Roberts, G. O. & Gilks, W. R. (1995). Efficient Metropolis jumping rules. In *Bayesian statistics* 5 (eds J. Bernardo, J. Berger, A. P. Dawid & A. F. M. Smith), 599–607. Oxford University Press, Oxford.

Knorr-Held, L. & Besag, J. (1998). Modelling risk from a disease in time and space. *Statist. Med.* **17**, (forthcoming).

Kohn, R. & Ansley, C. (1987). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM J. Sci. Comput.* **8**, 33–48.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*, 2 edn. Chapman & Hall, London.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika* **81**, 115–131.

Shephard, N. & Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653–667.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701–1762.

Received June 1996, in final form February 1998

Leonhard Knorr-Held, Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstr. 33, D-80539 München, Germany.