Biometrika (2006), **93,** 4, *pp.* 827–841

Printed in Great Britain

Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling

BY SYLVIA FRÜHWIRTH-SCHNATTER AND HELGA WAGNER

Department of Applied Statistics and Econometrics, Johannes Kepler Universität Linz, A-4040 Linz, Austria

sylvia.fruehwirth-schnatter@jku.at helga.wagner@jku.at

SUMMARY

We consider parameter-driven models of time series of counts, where the observations are assumed to arise from a Poisson distribution with a mean changing over time according to a latent process. Estimation of these models is carried out within a Bayesian framework using data augmentation and Markov chain Monte Carlo methods. We suggest a new auxiliary mixture sampler, which possesses a Gibbsian transition kernel, where we draw from full conditional distributions belonging to standard distribution families only. Emphasis lies on application to state space modelling of time series of counts, but we show that auxiliary mixture sampling may be applied to a wider range of parameter-driven models, including random-effects models and panel data models based on the Poisson distribution.

Some key words: Count data; Data augmentation; Finite mixture approximation; Gibbs sampling; Partially Gaussian state space model.

1. INTRODUCTION

In the analysis of time series of counts the observed process y_t is often assumed to follow a Poisson distribution. If there is a vector of covariates Z_t, a log-linear model could be applied, so that

$$y_t|\lambda_t \sim \mathrm{Po}(\lambda_t), \quad \lambda_t = \exp(Z_t'\beta),$$

in which λ_t is the mean of the time series observation y_t given β, and β is a vector of unknown coefficients. In the standard log-linear model it is assumed that the count observations are independent. To account for temporal dependence, various extensions of the log-linear model have been suggested which, following Cox (1981), may be broadly characterised as parameter-driven and observation-driven models. For observation-driven models, the mean structure of the conditional distribution of y_t given past observations $y_{t-1}, y_{t-2}, \ldots$ is directly specified as a function of these observations; see for instance Zeger & Qaqish (1988). In this paper we consider parameter-driven models, in which dependence among observations is introduced indirectly through a latent process, such as a hidden Markov chain, as in Leroux & Puterman (1992), or a latent stationary autoregressive process, as in Zeger (1988) and Chan & Ledolter (1995). More general state-space models based on a first-order hidden Markov process have been considered

by, among others, West et al. (1985) and Harvey & Fernandes (1989). A key property of parameter-driven models is that the distribution of y_t is allowed to depend on this latent process, and, although the observations are correlated marginally, they are independent, conditional on the latent process.

Estimation of parameter-driven Poisson time series models often turns out to be a challenging problem. In some cases, as in hidden Markov chain models, maximum likelihood estimation is straightforward, but for other models maximum likelihood is hampered by the fact that the marginal likelihood, with the latent process integrated out, is not available in closed form. Each evaluation of the likelihood function requires some numerical method for solving the necessary high-dimensional integration. For example, importance sampling was applied by Durbin & Koopman (2000) to state space modelling of count data; see also Durbin & Koopman (2001).

Alternatively, a Bayesian approach is feasible using data augmentation, as in Tanner & Wong (1987), and Markov chain Monte Carlo methods, as illustrated first by Zeger & Karim (1991) for generalised linear models with random effects and later by others for parameter-driven models for count data, in particular by Albert (1992) for Poisson random-effects models, by Wakefield et al. (1994) for more general random effect models, by Shephard & Pitt (1997) for non-Gaussian time series models based on distributions from the exponential family, by Gamerman (1998) for dynamic generalised linear models, by Chib et al. (1998) for panel count data models with multiple random effects, by Lenk & DeSarbo (2000) for mixtures of generalised linear models with random effects, and by Chib & Winkelmann (2001) for correlated multivariate count data. However, a major difficulty with any of the existing Markov chain Monte Carlo approaches is that practical implementation requires the use of a Metropolis–Hastings algorithm at least for part of the unknown parameter vector, and this makes it necessary to define suitable densities in rather high-dimensional parameter spaces. Single-move sampling for this type of model is likely to be very inefficient (Shephard & Pitt, 1997).

The main contribution of the present paper is to show how to design an approximate, yet very accurate, straightforward Gibbs sampling scheme for all unknown quantities, requiring only random draws from standard distributions such as multivariate normal, inverse Gamma and exponential distributions and discrete distributions with a few categories. Although we focus on state space models for Poisson counts, any other model with Poisson counts and some linear structure in the log intensity may be estimated in a similar way.

This is achieved by introducing two sequences of latent variables through data augmentation. The first of these sequences are the unobserved inter-arrival times of suitably chosen Poisson processes. Their introduction eliminates the nonlinearity of the observation equation, whereas the nonnormality of the error term, which is the negative of the logarithm of a random variable from the standard exponential distribution, remains. The distribution of the error term is then approximated by a mixture of normal distributions. If we introduce the component indicators of the normal mixture approximation as a second sequence of missing data, the resulting model may be thought of as a partially Gaussian model as in Shephard (1994), and Gibbs sampling becomes feasible. This will be shown to be particularly useful for state space models for Poisson time series, as multi-move sampling is feasible of the whole state process through forward-filtering backward sampling as in Frühwirth-Schnatter (1994b), Carter & Kohn (1994), de Jong & Shephard (1995) and Durbin & Koopman (2002).

2. Data augmentation for parameter-driven models based on the Poisson distribution

2·1. *Model specification*

Let $y_1, \ldots, y_T$ be a sequence of count data, observed at discrete, evenly spaced time points. We assume that $y_t|\lambda_t \sim \text{Po}(\lambda_t)$, where λ_t depends on covariates $Z_t = (Z_t^{(1)}, Z_t^{(2)})$ through fixed model parameters α and time-varying model parameters β_t:

$$y_t|\lambda_t \sim \text{Po}(\lambda_t), \tag{1}$$

$$\lambda_t \sim \exp\{(Z_t^{(1)})'\alpha + (Z_t^{(2)})'\beta_t\}. \tag{2}$$

The precise model for β_t will be left unspecified at this stage; we only assume that the joint distribution $p(\alpha, \beta_1, \ldots, \beta_T|\theta)$ follows a normal distribution, for some unknown parameter θ. Furthermore we assume that, conditional on knowing $\alpha, \beta_1, \ldots, \beta_T$, the observations $y_t|\lambda_t$ and $y_s|\lambda_s$ are mutually independent.

One may then derive the conditional posterior density $p(\alpha, \beta_1, \ldots, \beta_T|\theta, y)$ formally by Bayes' theorem, given the whole time series $y = (y_1, \ldots, y_T)$, but in general the resulting posterior density does not belong to a well-known distribution family. Although $\log \lambda_t$ in (2) is linear in $\alpha, \beta_1, \ldots, \beta_T$, the presence of the Poisson distribution in equation (1) causes nonnormality as well as nonlinearity of the mean λ_t in $\alpha, \beta_1, \ldots, \beta_T$. We now introduce our two latent processes.

2·2. *Step* 1: *Data augmentation through hidden interarrival times*

For each t, the distribution of $y_t|\lambda_t$ may be regarded as the distribution of the number of jumps of an unobserved Poisson process with intensity λ_t, having occurred in the time interval $[0, 1]$. The first step of data augmentation creates such a Poisson process for each y_t, for $t = 1, \ldots, T$, and introduces the inter-arrival times τ_{tj}, for $j = 1, \ldots, (y_t + 1)$, of this Poisson process as missing data. Since each $\tau_{tj} \sim \text{Ex}(\lambda_t)$ we have

$$\tau_{tj}|\alpha, \beta_t = \frac{\xi_{tj}}{\lambda_t}, \quad \xi_{tj} \sim \text{Ex}(1).$$

This may be reformulated as the linear model

$$-\log \tau_{tj}|\alpha, \beta_t = (Z_t^{(1)})'\alpha + (Z_t^{(2)})'\beta_t + \varepsilon_{tj}, \tag{3}$$

where $\varepsilon_{tj} = -\log \xi_{tj}$ with $\xi_{tj} \sim \text{Ex}(1)$.

Let $\tau = \{\tau_{tj}, j = 1, \ldots, (y_t + 1), t = 1, \ldots, T\}$ denote the collection of all inter-arrival times. Then the full-conditional distribution $p(\alpha, \beta_1, \ldots, \beta_T|\theta, \tau, y)$ depends on y only through τ: $p(\alpha, \beta_1, \ldots, \beta_T|\theta, \tau, y) = p(\alpha, \beta_1, \ldots, \beta_T|\theta, \tau)$. Secondly, conditional on τ, we are dealing with model (3), which is nonnormal, but in which the mean of the observation equation is linear in the parameters $\alpha, \beta_1, \ldots, \beta_T$:

$$E(-\log \tau_{tj}|\alpha, \beta_t) = (Z_t^{(1)})'\alpha + (Z_t^{(2)})'\beta_t + 0{\cdot}57722.$$

2·3. *Step* 2: *Data augmentation through a mixture approximation*

The error term in (3) may be regarded as the negative of the logarithm of an Ex(1) random variable, the density $p_\varepsilon(\varepsilon)$ of which can be written

$$p_\varepsilon(\varepsilon) = \exp(-\varepsilon - e^{-\varepsilon}).$$

To obtain a model that is conditionally Gaussian, we approximate this nonnormal density by a mixture of R normal components:

$$p_\varepsilon(\varepsilon) = \exp(-\varepsilon - e^{-\varepsilon}) \approx q_{R,\varepsilon}(\varepsilon) = \sum_{r=1}^{R} w_r N(\varepsilon; m_r, s_r^2), \tag{4}$$

where, for $r = 1, \ldots, R$, m_r and s_r^2 are the mean and the variance and $N(\varepsilon; m_r, s_r^2)$ denotes the Gaussian density. This idea is influenced by the related articles of Kim et al. (1998) and Chib et al. (2002), who used a normal mixture approximation to the density of a log χ^2-distribution in the context of stochastic volatility models. The appropriate parameters (w_r, m_r, s_r^2), for $r = 1, \ldots, R$, were determined numerically for $R = 2, \ldots, 10$ by minimising the Kullback–Leibler distance between the true density and the mixture approximation; see Frühwirth-Schnatter & Frühwirth (2007) for further details. The parameters (w_r, m_r, s_r^2) are tabulated in Table 1 for $R = 10$. The true density and the approximation are almost indistinguishable.

Table 1. *Ten-component normal mixture approximation for the density of* $-\log \xi$, *where* $\xi \sim \mathrm{Ex}(1)$

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	$r=8$	$r=9$	$r=10$
w_r	0·00397	0·00396	0·168	0·147	0·125	0·101	0·104	0·116	0·107	0·088
m_r	5·09	3·29	1·82	1·24	0·764	0·391	0·0431	−0·306	−0·673	−1·06
s_r^2	4·50	2·02	1·10	0·422	0·198	0·107	0·0778	0·0766	0·0947	0·146

Similarly to Kim et al. (1998) and Chib et al. (2002), the density $p_\varepsilon(\varepsilon_{tj})$ in (3) is approximated for each t and j by the mixture approximation $q_{R,\varepsilon}(\varepsilon_{tj})$. The second step of our data augmentation scheme introduces for each ε_{tj} the latent component indicator r_{tj} as missing data. Let $S = \{r_{tj}, j = 1, \ldots, (y_t + 1), t = 1, \ldots, T\}$. Then, conditional on τ and S, the nonnormal, nonlinear model (1) and (2) reduces to a linear, Gaussian model where the mean of the observation equation is linear in $\alpha, \beta_1, \ldots, \beta_T$ and the error term follows a normal distribution:

$$-\log \tau_{tj} | \alpha, \beta_t, r_{tj} = (Z_t^{(1)})'\alpha + (Z_t^{(2)})'\beta_t + m_{r_{tj}} + \varepsilon_{tj}, \quad \varepsilon_{tj} | r_{tj} \sim N(0, s_{r_{tj}}^2).$$

Consequently, the conditional posterior $p(\alpha, \beta_1, \ldots, \beta_T | \theta, \tau, S, y)$ is given by

$$p(\alpha, \beta_1, \ldots, \beta_T | \theta, \tau, S, y) \propto p(\alpha, \beta_1, \ldots, \beta_T | \theta) \prod_{t=1}^{T} \prod_{j=1}^{y_t+1} N\{-\log \tau_{tj}; (Z_t^{(1)})'\alpha + (Z_t^{(2)})'\beta_t + m_{r_{tj}}, s_{r_{tj}}^2\}, \tag{5}$$

which is proportional to a multivariate normal density.

3. Auxiliary mixture sampling for parameter-driven models for time series of counts

3·1. *The basic three-block auxiliary mixture sampler*

A three-block Gibbs sampler involves selecting starting values for τ, S and θ, and repeating the following steps.

Step 1. Carry out multi-move sampling of α and the whole sequence $\beta = \{\beta_1, \ldots, \beta_T\}$ from the multivariate normal distribution (5), conditional on τ, S, θ and y.

Step 2. Sample θ conditional on α, β, τ, S and y.

Step 3. Sample the inter-arrival times τ and the component indicators S conditional on y, θ, α and β by running the following steps, for $t = 1, \ldots, T$.

(a) Sample the inter-arrival times $\{\tau_{tj}, j = 1, \ldots, y_t + 1\}$. If $y_t > 0$, sample the order statistics $u_{t,(1)}, \ldots, u_{t,(n)}$ of $n = y_t$ Un$[0, 1]$ random variables, see Robert & Casella (1999, p. 47) for details, and define the inter-arrival times τ_{tj} as their increments: $\tau_{tj} = u_{t,(j)} - u_{t,(j-1)}$, for $j = 1, \ldots, n$, where $u_{t,(0)} := 0$. Sample the final arrival time as $\tau_{t,n+1} = 1 - \sum_{j=1}^{n} \tau_{tj} + \xi_t$, where $\xi_t \sim \mathrm{Ex}(\lambda_t)$.

(b) Sample the component indicators r_{tj} for $j = 1, \ldots, y_t + 1$.

Steps 1 and 2 are model-dependent, but for many models involve only standard draws, as we are dealing with a Gaussian model, once we condition on τ and S. Step 3, however, deserves detailed investigation. This step is based on decomposing the joint posterior $p(\tau, S|y, \alpha, \beta, \theta)$ as

$$p(\tau, S|y, \theta, \alpha, \beta) = p(S|\tau, y, \theta, \alpha, \beta)p(\tau|y, \theta, \alpha, \beta).$$

We first sample the inter-arrival times τ from $p(\tau|y, \theta, \alpha, \beta)$. The inter-arrival times $\{\tau_{tj}, j = 1, \ldots, y_t + 1\}$ are independent for different time points t, given β, θ, α and y:

$$p(\tau|y, \theta, \alpha, \beta) = \prod_{t=1}^{T} p(\tau_{t1}, \ldots, \tau_{t,y_t+1}|y_t, \theta, \alpha, \beta).$$

For fixed t, the inter-arrival times $\tau_{t1}, \ldots, \tau_{t,n+1}$, where $n = y_t$, are stochastically dependent, and the joint distribution factorises as

$$p(\tau_{t1}, \ldots, \tau_{tn}, \tau_{t,n+1}|y_t = n, \theta, \alpha, \beta) = p(\tau_{t,n+1}|y_t = n, \theta, \alpha, \beta, \tau_{t1}, \ldots, \tau_{tn})p(\tau_{t1}, \ldots, \tau_{tn}|y_t = n).$$

The first n inter-arrival times are independent of all model parameters, and are determined only by the observed number of counts y_t. By well-known properties of a Poisson process, the first n arrival times are distributed as the order statistics of n Un$[0, 1]$ random variables. Only the final inter-arrival time $\tau_{t,n+1}$ depends on the actual model parameters α, β and θ through the intensity λ_t. Conditionally on y_t, only $n = y_t$ arrivals occur in $[0, 1]$, so that arrival $(n + 1)$ is known to occur after 1. As a result of the zero-memory property of the exponential distribution, the waiting time after 1 is distributed as $\mathrm{Ex}(\lambda_t)$ and $\tau_{t,n+1} = 1 - \sum_{j=1}^{n} \tau_{tj} + \xi_t$, where $\xi_t \sim \mathrm{Ex}(\lambda_t)$. This justifies Step 3(a).

To sample the indicators S from $p(S|\tau, y, \theta, \alpha, \beta)$, we use the fact that all indicators are conditionally independent given y, θ, α, β and τ:

$$p(S|\tau, y, \theta, \alpha, \beta) = \prod_{t=1}^{T} \prod_{j=1}^{y_t+1} p(r_{tj}|\tau_{tj}, \theta, \beta_t, \alpha).$$

Thus, for each $t = 1, \ldots, T$ and each $j = 1, \ldots, y_t + 1$, the indicator r_{tj} is sampled independently from $p(r_{tj}|\tau_{tj}, \theta, \beta_t, \alpha)$. This density depends on the data only through τ_{tj} and depends on θ, α and β_t only through λ_t:

$$\mathrm{pr}(r_{tj} = k|\tau_{tj}, \theta, \beta_t, \alpha) \propto p(\tau_{tj}|r_{tj} = k, \beta_t, \alpha, \theta)w_k,$$

where

$$p(\tau_{tj}|r_{tj} = k, \theta, \beta_t, \alpha) \propto \frac{1}{s_k} \exp\left\{-\frac{1}{2}\left(\frac{-\log \tau_{tj} - \log \lambda_t - m_k}{s_k}\right)^2\right\}.$$

This explains Step 3(b).

Starting values for τ and S are obtained in the following way. Each component indicator r_{tj} is drawn uniformly from 1 to R. Step 3(a) is used to sample starting values for $\tau_{t1}, \ldots, \tau_{tn}$ for each t, given the observed counts y_t. To obtain a starting value for $\tau_{t,n+1}$, we sample ξ_t from $\mathrm{Ex}(\lambda_t)$ with $\lambda_t = y_t$, if $y_t > 0$. For all t, where $y_t = 0$, λ_t is set to a small value; in our examples we used $\lambda_t = 0{\cdot}1$.

3·2. *Adding a rejection step*

In principle a rejection step could be added, as suggested by referees. However, the rejection step would deprive auxiliary mixture sampling of its simplicity. Secondly, it is not necessary to introduce a mixture approximation for $p_\varepsilon(\varepsilon_{tj})$ if a rejection step is incorporated, because we could then use a single normal distribution as an approximation for $p_\varepsilon(\varepsilon_{tj})$. The acceptance rate of this independence Metropolis sampler is quite high, as will be shown in the example below. By increasing the number of components, we can render the mixture approximation arbitrarily accurate, and we claim that such a rejection step is not necessary, since the acceptance rate is then effectively 100 percent.

We evaluated this acceptance rate for a simple example, namely Bayesian inference for T independent observations $y_1, \ldots, y_T$ from the $\mathrm{Po}(\lambda)$ distribution under the prior $\lambda \sim \mathrm{Ga}(a_0, b_0)$, in which case $\lambda|y \sim \mathrm{Ga}(a_0 + T\bar{y}, b_0 + T)$, with $\bar{y}$ being the sample mean. The augmented model, after the first step of data augmentation, is

$$-\log \tau_{tj} = \beta + \varepsilon_{tj}, \tag{6}$$

with $\beta = \log \lambda$. To evaluate how the approximation error introduced in the second data augmentation step influences the acceptance rate, we consider a marginal two-step sampler without introducing the indicators, where we sample in a first step the inter-arrival times as in Step 3(a) and propose β^{new} from the proposal density $q_R(\beta|\tau) \propto q_R(\tau|\beta)p(\beta)$, where $q_R(\tau|\beta)$ is the likelihood for an approximation to model (6), obtained by substituting $p_\varepsilon(\varepsilon_{tj})$ by the mixture approximation $q_{\varepsilon,R}(\varepsilon_{tj})$:

$$q_R(\tau|\beta) = \prod_{t=1}^{T} \prod_{j=1}^{y_t+1} q_{\varepsilon,R}(-\log \tau_{tj} - \beta).$$

The acceptance rate depends on the ratio

$$r(\beta, \tau) = \frac{p(\tau|\beta)}{q_R(\tau|\beta)},$$

where $p(\tau|\beta)$ is the likelihood of the exact augmented model (6):

$$p(\tau|\beta) = \prod_{t=1}^{T} \prod_{j=1}^{y_t+1} p_\varepsilon(-\log \tau_{tj} - \beta).$$

The acceptance rate is random, depending both on the new draw β^{new} and on draws β^{old} and τ from the stationary distribution $p(\beta, \tau|y)$, which is known explicitly for this example. The expected acceptance rate is calculated for data $y = (y_1, \ldots, y_T)$ simulated from the $\mathrm{Po}(\lambda)$ distribution, where the expectation is taken with respect to the joint distribution $p(\beta^{\mathrm{new}}, \beta^{\mathrm{old}}, \tau|y) = q_R(\beta^{\mathrm{new}}|\tau)p(\tau|\beta^{\mathrm{old}}, y)p(\beta^{\mathrm{old}}|y)$:

$$\int \left\{ \int \min\left(1, \frac{r(\beta^{\mathrm{new}}, \tau)}{r(\beta^{\mathrm{old}}, \tau)}\right) q_R(\beta^{\mathrm{new}}|\tau) d\beta^{\mathrm{new}} \right\} p(\tau|\beta^{\mathrm{old}}, y)p(\beta^{\mathrm{old}}|y) d\tau \, d\beta^{\mathrm{old}}.$$

Table 2. *Expected acceptance rate* (%) *for a Metropolis–Hastings algorithm based on a mixture approximation with* R *components, for* $R = 1, \ldots, 10$

λ	T	$(\lambda+1)T$	$R=1$	$R=2$	$R=3$	$R=4$	$R=5$	$R=6$	$R=7$	$R=8$	$R=9$	$R=10$
1	1	2	77·3	91·4	95·7	97·4	98·9	99·4	99·6	99·7	99·8	99·9
3	1	4	71·7	87·9	93·5	96·1	98·4	99·1	99·4	99·6	99·7	99·8
10	1	11	70·7	85·9	92·9	96·2	98·2	99·0	99·4	99·5	99·7	99·7
1	10	20	67·9	84·8	92·0	95·1	97·6	98·5	99·0	99·4	99·5	99·6
3	10	40	68·4	85·8	92·4	95·5	97·6	98·6	99·0	99·3	99·4	99·6
0·1	100	110	21·3	83·1	84·7	92·8	96·9	98·1	98·5	99·2	99·3	99·5
10	10	110	66·6	83·9	91·9	94·7	97·5	98·6	99·1	99·2	99·3	99·4
1	100	200	42·0	75·2	91·4	94·0	97·1	97·9	98·7	99·1	99·3	99·3
3	100	400	62·8	80·7	89·5	94·1	97·2	98·2	98·8	99·1	99·2	99·4
1	1000	2000	5·2	35·7	88·5	90·1	95·4	97·2	97·8	98·8	99·1	99·1

Table 2 shows that running a Metropolis–Hastings algorithm with $R = 1$, in which case $p_\varepsilon(\varepsilon_{tj})$ is approximated by a single $N(0{\cdot}5772, 1{\cdot}6625)$ density and only the first augmentation step has to be implemented, is a reasonable alternative to auxiliary mixture sampling. As the number of components is increased, the acceptance rate evidently approaches 100%.

Note that the mixture approximation is applied to equation (6) not just once, but $\sum_{t=1}^{T} y_t + T$ times, so that on average $(\lambda+1)T$ approximations take place. Table 2 demonstrates how the approximation error accumulates when T and λ increase. For a smaller number of components the acceptance rate rapidly decreases, as the number of expected approximations increases. For the 10-component mixture approximation, however, it remains above 99 percent, even for $\lambda = 1$ and $T = 1000$, where the expected number of approximations is equal to 2000.

3·3. *Auxiliary mixture sampling for state space modelling of time series of counts*

We consider in detail state space modelling of time series of small counts as introduced by West et al. (1985) and Harvey & Fernandes (1989). In its most general form, the model is

$$y_t|\alpha, \beta_t \sim \mathrm{Po}[\exp\{(Z_t^{(1)})'\alpha + (Z_t^{(2)})'\beta_t\}],$$
$$\beta_t = F\beta_{t-1} + c + w_t, \quad w_t \sim N(0, Q),$$

where β_t is a latent Markov process. The expectation $E(\beta_t|\beta_{t-1}) = F\beta_{t-1} + c$ is linear in β_{t-1}, whereas the variance-covariance matrix is $\mathrm{var}(\beta_t|\beta_{t-1}) = Q$. The matrices F and Q and the vector c may be known, or may depend on unknown model parameters θ. A simple example is the local-level model, which is

$$y_t|\mu_t \sim \mathrm{Po}\{\exp(\mu_t)\},$$
$$\mu_t = \mu_{t-1} + w_t, \quad w_t \sim N(0, \theta),$$

with θ being the only unknown parameter.

Previous approximation methods for estimating state space models for time series of counts have included Monte Carlo EM (Chan & Ledolter, 1995), the assumption of natural conjugate priors for β_t, based on discounting information from the past (Harvey & Fernandes, 1989), and approaches which also allow for smoothing, based on the posterior

mode filter of Fahrmeir (1992) and the integration-based Kalman filter of Frühwirth-Schnatter (1994a). Each of these approximation methods is likely to introduce an approximation error of unknown magnitude, that is not reducible by increasing the computational effort of the investigator. A first attempt to compute the exact likelihood function for the Poisson local-level model, reported in Kashiwagi & Yanagimoto (1992), is an application of the numerical integration filter of Kitagawa (1987), and therefore limited to one- or two-dimensional state vectors. Advantages of Markov chain Monte Carlo methods in comparison to any of these methods are that increasing the computational effort leads to increased accuracy of the algorithm, and that rather high-dimensional state vectors can be handled.

The first data augmentation step described above introduces a total of $n_t = y_t + 1$ inter-arrival times τ_{tj} ($j = 1, \ldots, n_t$), for each of the T observations y_t ($t = 1, \ldots, T$). The second data augmentation step introduces a component indicator r_{tj} for each of the $T + \sum_{t=1}^{T} y_t$ inter-arrival times τ_{tj}. Conditioning on all inter-arrival times and the component indicators leads to the observation equation

$$-\log \tau_{tj} | \mu_t, r_{tj} = (Z_t^{(1)})'\alpha + (Z_t^{(2)})'\beta_t + m_{r_{tj}} + \varepsilon_{tj}, \quad \varepsilon_{tj} \sim N(0, s^2_{r_{tj}}).$$

If we define a multivariate observation vector $\tilde{y}_t$ of dimension $n_t = y_t + 1$ as

$$\tilde{y}_t = \begin{pmatrix} -\log \tau_{t1} - m_{r_{t1}} \\ \vdots \\ -\log \tau_{t,n_t} - m_{r_{t,n_t}} \end{pmatrix},$$

the augmented model may be written in the following linear Gaussian state space form:

$$\tilde{y}_t = \tilde{Z}_t^{(1)}\alpha + \tilde{Z}_t^{(2)}\beta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, R_t), \tag{7}$$

$$\beta_t = F\beta_{t-1} + u_t + w_t, \quad w_t \sim N(0, Q), \tag{8}$$

where $R_t = \mathrm{diag}(s^2_{r_{t1}}, \ldots, s^2_{r_{t,n_t}})$. Here $\tilde{Z}_t^{(1)}$ and $\tilde{Z}_t^{(2)}$ are matrices with n_t rows, containing n_t copies of the design vectors $Z_t^{(1)}$ and $Z_t^{(2)}$:

$$\tilde{Z}_t^{(1)} = \begin{pmatrix} (Z_t^{(1)})' \\ \vdots \\ (Z_t^{(1)})' \end{pmatrix}, \quad \tilde{Z}_t^{(2)} = \begin{pmatrix} (Z_t^{(2)})' \\ \vdots \\ (Z_t^{(2)})' \end{pmatrix}.$$

Thus, for a state space model for count data we have a partially Gaussian state space model for repeated measurements, where the transition equation is the same as for the original Poisson state space model. However, the Poisson observation equation for the single count observation y_t is substituted by a Gaussian observation equation with the multivariate observation vector $\tilde{y}_t$ appearing as repeated measurements.

The three-block auxiliary mixture sampler described in § 3·1 works as follows.

Step 1. Carry out multi-move sampling for the whole sequence $\alpha, \beta_0, \ldots, \beta_T$ by forward-filtering backward sampling as in Frühwirth-Schnatter (1994b), Carter & Kohn (1994), de Jong & Shephard (1995) or Durbin & Koopman (2002) for the conditionally Gaussian state space form (7) and (8).

Step 2. Sample θ conditional on α, β, τ and S from the conditionally Gaussian state space form (7) and (8).

Step 3. For each $t = 1, \ldots, T$, compute $\log \lambda_t = (Z_t^{(1)})'\alpha + (Z_t^{(2)})'\beta_t$, and sample the inter-arrival times $\{\tau_{tj}, j = 1, \ldots, y_t + 1\}$ and the component indicators r_{tj} $(j = 1, \ldots, y_t + 1\}$ as described earlier.

The precise details of Step 2 depend on the specific state space form. If Q is an unrestricted variance-covariance matrix, then Q is sampled from an inverse Wishart distribution. If only some diagonal elements of Q are unknown, as with the basic structural model to be considered in § 4, these parameters are sampled independently from inverse Gamma distributions.

3·4. *Auxiliary mixture sampling for other parameter-driven models of count data*

The auxiliary mixture sampler has wider application. The two latent sequences τ and S eliminate nonnormality and nonlinearity, whenever the log intensity is linear in the unknown parameters. To implement Steps 1 and 2 for a particular model, we may exploit any result that is available for Markov chain Monte Carlo estimation of this particular model class within the Gaussian family. To sample the inter-arrival times τ_{tj} in Step 3 we only need to know the observed counts and the conditional mean λ_t, whereas to sample the component indicator r_{tj} we need to know τ_{tj} and λ_t. Although λ_t depends on α, β and θ in a specific way described by the model, Step 3 is independent of the specific structure of the model, once we have determined λ_t.

Suppose for further illustration that we are fitting a random-effects model to panel count data y_{it}, for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, as in Chib et al. (1998), who considered a model with multiple random effects based on the Poisson distribution $y_{it} \sim \mathrm{Po}\{\exp(\lambda_{it})\}$. At each sweep of the auxiliary mixture sampler, each count observation y_{it} is augmented by inter-arrival times $\tau_{it,j}$ and indicators $r_{it,j}$, for $j = 1, \ldots, y_{it} + 1$. Through our data augmentation scheme, the random-effects model for count data reduces to the same random-effects model, but with $y_{it} + 1$ repeated Gaussian measurements $-\log \tau_{it,j} - m_{r_{it,j}}$, for $j = 1, \ldots, y_{it} + 1$, with observation variance $s^2_{r_{it,j}}$. Implementation of Steps 1 and 2 are now standard. To sample τ and S in Step 3, we only need to determine λ_{it}.

4. Application to road safety data

We illustrate the new sampler on time series provided by the Austrian Road Safety Board. These time series are monthly counts of killed or injured pedestrians from 1987–2002 in Linz, which is the third largest town in Austria. We use series for children aged 6–10 years and senior persons above 60 years. These are series of small counts not exceeding 5 and 15, respectively. A new law intended to increase road safety came into force in Austria on 1 October 1994, since when pedestrians who want to use a pedestrian crossing have to be allowed to cross. We are going to analyse, separately for each age group, the effect of this law on the intensity λ_t in month t of being killed or seriously injured as a pedestrian living in Linz.

If y_t is the number of persons killed or seriously injured in time period t, e_t is the number of children or senior people living in Linz then and λ_t is the typically very small intensity, a suitable model is

$$y_t \sim \mathrm{Po}(e_t \lambda_t).$$

Note that the correct unit of analysis would be rates not counts and, theoretically, the rate in the Poisson model is allowed to exceed 1. In practice, however, this is not really a problem because the rates are very small in this application.

Both time series are modelled using a basic structural model for Poisson counts as in Harvey & Durbin (1986), where the intensity λ_t is assumed to have a multiplicative trend as well as a multiplicative seasonal component:

$$\log(\lambda_t) = \mu_t + s_t. \tag{9}$$

In (9), it is assumed that μ_t is a stochastic trend, following a random walk with drift a_t:

$$\mu_t = \mu_{t-1} + a_{t-1} + w_{1t}, \quad w_{1t} \sim N(0, \theta_1), \tag{10}$$

where $\mu_0 \sim N\{\log(y_1/e_1), 1\}$. To capture the effect of the change in law, equation (10) is slightly modified, by including a level shift δ at the time point $t = t_{\text{int}}$, when the legal intervention occurred:

$$\mu_t = \mu_{t-1} + \alpha_{t-1} + \delta + w_{1t}. \tag{11}$$

In its most general form, the basic structural model assumes that the drift a_t changes over time and itself follows a random walk:

$$a_t = a_{t-1} + w_{2t}, \quad w_{2t} \sim N(0, \theta_2), \tag{12}$$

where $a_0 \sim N(0, 1)$. In the context of state space models, a_t is usually called the slope, as it determines the expected increase in the level of μ_{t+1} compared to μ_t.

Finally, $\exp(s_t)$ is a monthly multiplicative seasonal component generated by

$$s_t = -s_{t-1} - \ldots - s_{t-11} + w_{3t}, \quad w_{3t} \sim N(0, \theta_3), \tag{13}$$

where $\sum_{j=0}^{11} s_{-j} = 0$, and $(s_{-1}, \ldots, s_{-11})'$ is an unknown initial pattern.

In equations (10) to (13), the parameters θ_1, θ_2 and θ_3 are unknown variances, which are either estimated from the data or assumed to be zero. Note that the stochastic trend reduces to a linear deterministic trend function with intercept μ_0 and slope a_0, if the variances θ_1 and θ_2 are zero. Choosing $\theta_3 = 0$ leads to a fixed seasonal pattern over the whole observation period, whereas choosing $\theta_3 > 0$ allows a smooth change in this pattern.

We rewrite the model in state space model form. For a model with nonzero variances, for instance, the state vector β_t has 14 dimensions, namely $\beta_t = (\mu_t, a_t, s_t, \ldots, s_{t-11}, \delta)'$, where only the first three components are actually dynamic. We now turn to the application of the auxiliary mixture sampling scheme. Data augmentation through the mixture approximation leads to a partly dynamic model in the sense of Frühwirth-Schnatter (1994b) with $\theta = (\theta_1, \theta_2, \theta_3)'$. In this model the variances can be sampled independently from inverse Gamma distributions, assuming inverse Gamma prior distributions on each variance. We choose $\theta_i \sim \text{IG}(0{\cdot}1, 0{\cdot}001)$, for $i = 1, \ldots, 3$.

The Gibbs sampler described in § 3·3 was run 12 000 times with a burn-in of 2000 runs. As the chain did not converge for the original formulation of the model we used a reparameterisation in which the seasonal component was noncentred, as in Frühwirth-Schnatter (2004). The noncentred seasonal component $\tilde{s}_t$ is the standardised deviation of s_t from the initial seasonal pattern $\alpha = (s_{-1}, \ldots, s_{-11})'$:

$$\tilde{s}_t = \frac{s_t - (Z_t^{(1)})'\alpha}{\theta_4}, \quad \log(\lambda_t) = \mu_t + (Z_t^{(1)})'\alpha + \theta_4 \tilde{s}_t,$$

where $\theta_4^2 = \theta_3$ and $Z_t^{(1)}$ is a vector that selects the appropriate seasonal components for time point t. For t being a multiple of 12, $Z_t^{(1)}$ is a vector of -1's; otherwise all elements of $Z_t^{(1)}$ are 0, apart from the element in the row corresponding to the actual season, which

takes the value 1. Introducing the state vector $\beta_t = (\mu_t, a_t, \tilde{s}_t, \ldots, \tilde{s}_{t-11}, \delta)'$ and choosing $\theta = (\theta_1, \theta_2, \theta_4)$ led to a Gibbs sampler with quick convergence to the stationary distribution.

Figure 1 shows the observed counts, the number of exposed e_t, the smoothed level μ_t with pointwise 95% credibility intervals and the risk intensity λ_t for the children. Figure 2

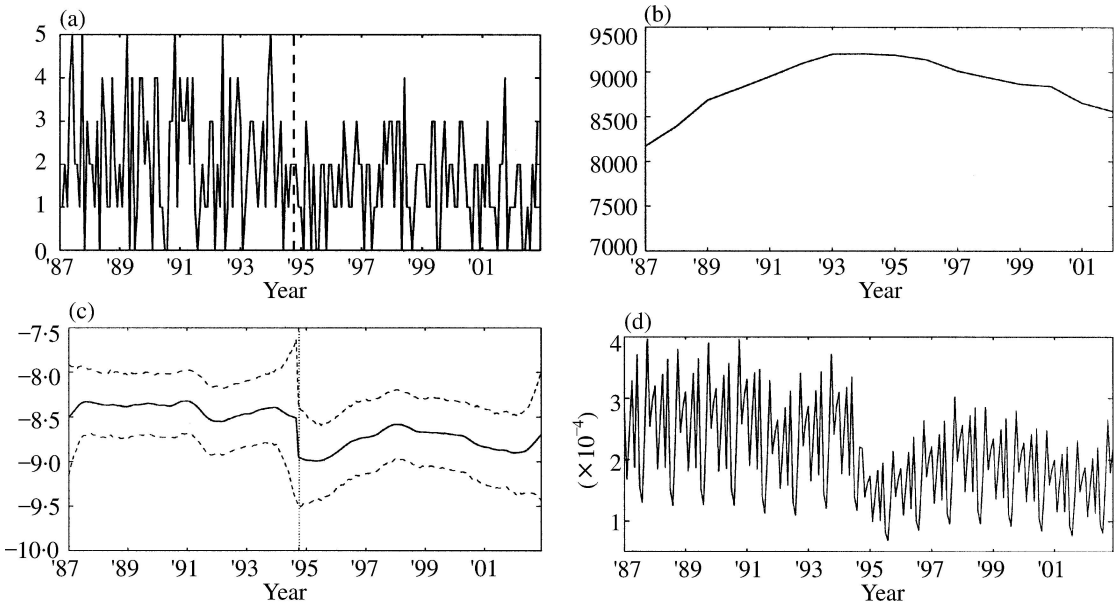

Fig. 1: Road safety data 1987–2002. (a) Counts of killed or injured children, (b) number of children exposed, (c) estimated mean $\mu_t|y$ characterised by the posterior mean and 95% credible regions, (d) estimated risk λ_t.

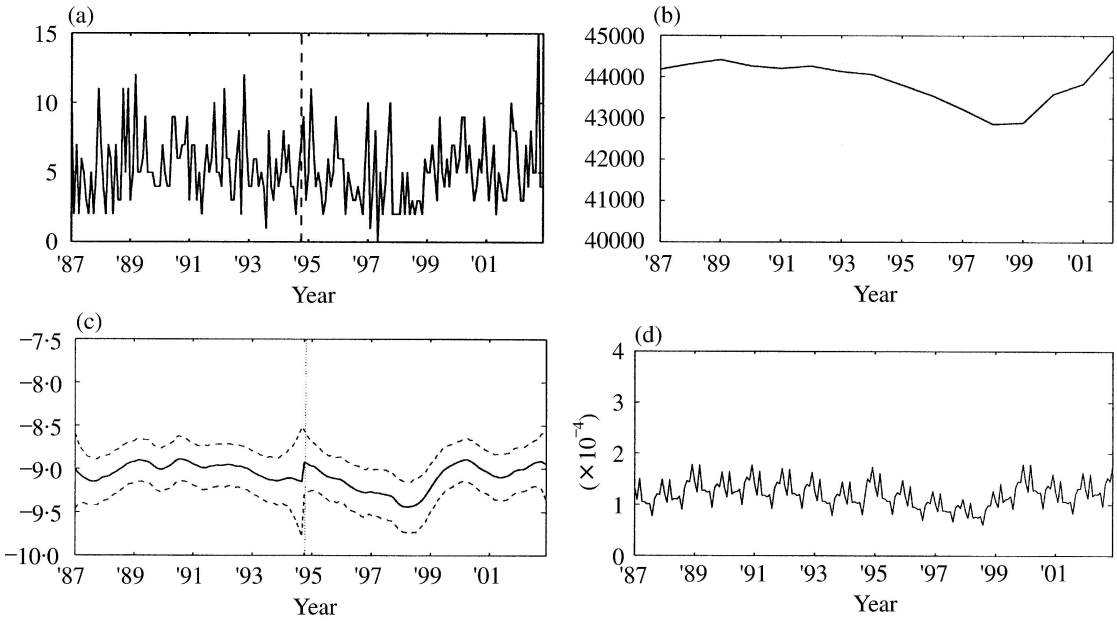

Fig. 2: Road safety data 1987–2002. (a) Counts of killed or injured senior persons, (b) number of people exposed, (c) estimated mean $\mu_t|y$ with 95% credible regions, (d) estimated risk λ_t.

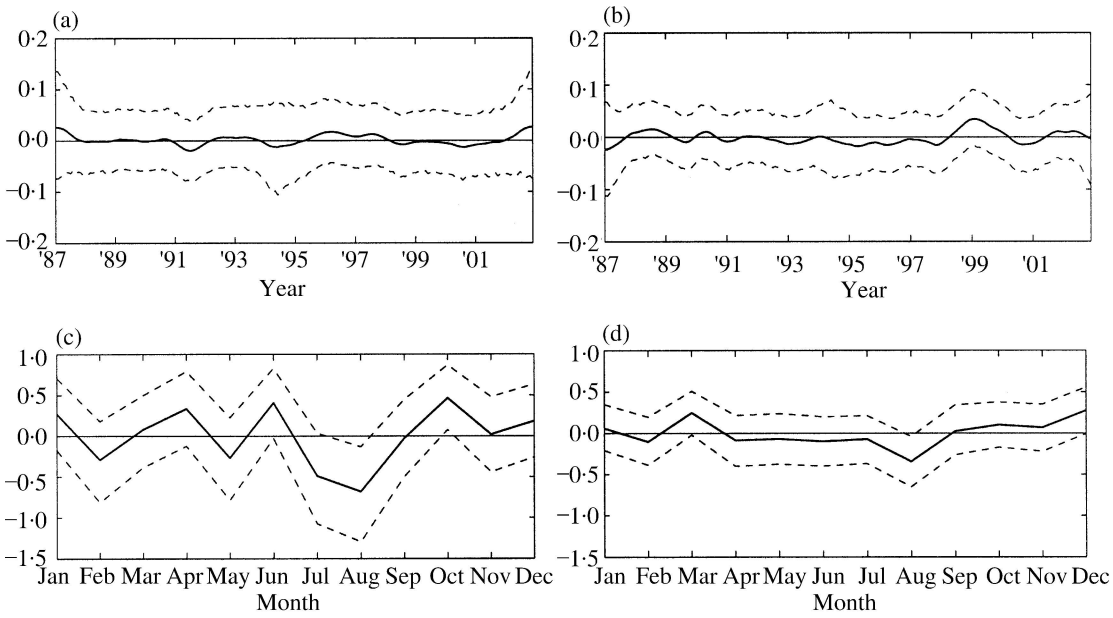

Fig. 3: Road safety data. Posterior means and 95% credible regions of the drift a_t 1987–2002 (a) for the number of killed and injured children, (b) for the number of killed and injured senior people, and for the seasonal components s_t in year 2002 (c) for the children, (d) for the senior people.

shows the same for the senior series. The trend component a_t and the seasonal pattern s_t in the last year are shown in Fig. 3 for both age groups.

The estimated risk intensity is much larger for the children than for the senior people, and there is a pronounced decrease for the children's time series after the change in law. The seasonal patterns in the two series are quite different: for the children, rates are significantly lower than the annual average in the holiday months of July and August and higher in June and October, whereas for senior people there is solely a significant decrease in August.

Figure 4 shows the posterior density of θ_4 for each time series. Recall that θ_4 is defined as $\pm(\theta_3)^{1/2}$, so that θ_4 has a negative or a positive sign with equal probability. Therefore, the posterior density of θ_4 is symmetric around zero. If the unknown variance θ_3 is systematically different from zero, then the posterior density of θ_4 is likely to be bimodal;

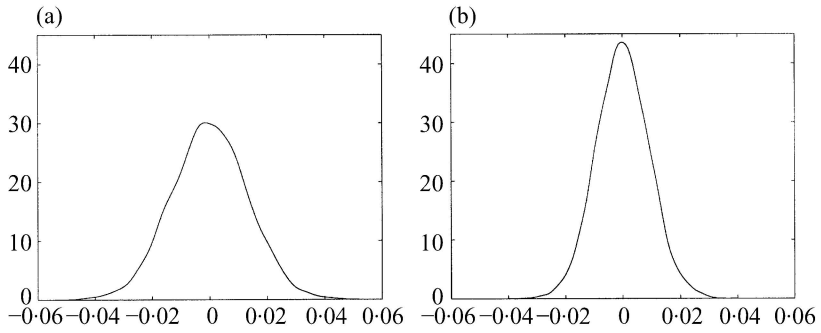

Fig. 4: Road safety data. Kernel density estimates of the posterior distribution of θ_4 (a) for children and (b) for senior people.

otherwise, if θ_3 is close to zero, the posterior density of θ_4 will be centred around zero. Figure 4 shows that for both time series θ_4 is centred around zero, and we may conclude that for these time series the seasonal pattern is stable over time.

The drift term a_t is not significantly different from zero in either of the series and we therefore consider the simpler local-level model with fixed seasonal pattern.

Table 3 reports point estimates as well as 95% highest posterior density regions for the variance θ_1 and the intervention effect δ in the local-level model with fixed seasonal pattern for both series. These regions are obtained as the shortest interval containing 95% of the simulations (Chen et al., 2000).

Table 3: *Road safety data. Parameter estimates for killed and injured children and for killed and injured senior people*

	Children			Seniors		
Parameter	Mean	Std dev.	95% HPD regions	Mean	Std dev.	95% HPD regions
θ_1	0·0022	0·0017	[0·0002, 0·0055]	0·0020	0·0016	[0·0002, 0·0052]
δ	−0·4029	0·2453	[−0·9018, 0·0772]	0·0417	0·1992	[−0·3401, 0·4512]

Std dev., posterior standard deviation; HPD, highest posterior density.

The process variances are nearly equal in both series. For children, the posterior of the intervention effect δ is clearly shifted away from 0, which is not the case for senior people. We take this as evidence for a positive effect of the legal change on reducing the risk of being seriously injured or killed for children and of no effect for senior people.

5. Discussion

Care must be exercised with respect to parameterisation issues, as straightforward Gibbs sampling often leads to convergence problems. Such problems are well known for Gaussian random-effects models (Gelfand et al., 1995; van Dyk & Meng, 2001) and Gaussian state space models (Roberts et al., 2004; Frühwirth-Schnatter, 2004). For Poisson count data parameterisation issues are also addressed in Chib et al. (1998). Our application demonstrates that the mixing properties of auxiliary mixture sampling can improve dramatically if we use a noncentred parameterisation similar to the one studied in Frühwirth-Schnatter (2004).

The auxiliary mixture sampler is mainly designed for, but not restricted to, small counts, as larger counts can be safely treated with a normal approximation. It is, however, useful for time series including both small and large counts. To give an example, we generated a series y_t of length $T = 101$ from a $\text{Po}(\lambda_t)$ distribution with exponential trend $\lambda_t = \exp(Z_t'\alpha)$, Z_t evenly spaced from 0 to 5 and $\alpha = 0{\cdot}9$.

The Gibbs sampler was run without difficulty for 12 000 iterations with a burn-in of 2000, using the improper normal prior $p(\alpha) \propto 1$. It gave a posterior mean for α of 0·8967 with a standard error of 0·0058; the 95% credible interval was [0·8852, 0·9073].

The auxiliary mixture sampler introduced in this paper is easily modified to deal with various extensions of the model structure. If the latent process follows a t distribution, as in Chib & Winkelmann (2001), rather than a normal distribution, our estimation approach needs to be adapted only slightly along the lines of Shephard (1994), by expressing the t distribution as a scale mixture of normals.

The results of the present paper are to some extent also useful outside the framework of data from a Poisson distribution. MacDonald & Zucchini (1997, p. 68) note that a

particularly useful model for capturing overdispersion is a negative binomial distribution based on a hidden Markov process S_t. By writing the negative binomial distribution as an infinite mixture of Poisson distributions, we can easily design an auxiliary mixture sampler for any model with a linear structure in the log of the mean.

Finally, as shown in Frühwirth-Schnatter & Frühwirth (2007), a similar auxiliary mixture sampler is feasible for parameter-driven models for other discrete-valued observations such as binary and multinomial data. The first data augmentation step for such observations is different and relies on the latent utility approach introduced by McFadden (1974), and the second augmentation step involves a normal mixture approximation to the type I extreme value distribution, which is equal to the density of the negative of the logarithm of an exponentially distributed variable, and is essentially the same as that used in the present paper.

ACKNOWLEDGEMENT

We wish to thank Stefan Höglinger from the Austrian Road Safety Board for providing the data analysed in § 4. This project was supported by a grant from the Austrian Science Foundation.

REFERENCES

ALBERT, J. (1992). A Bayesian analysis of a Poisson random effects model for homerun hitters. *Am. Statistician* **46**, 246–53.

CARTER, C. & KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–53.

CHAN, K. S. & LEDOLTER, J. (1995). Monte Carlo EM estimation for time series models involving counts. *J. Am. Statist. Assoc.* **90**, 242–52.

CHEN, M.-H., SHAO, Q.-M. & IBRAHIM, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

CHIB, S. & WINKELMANN, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *J. Bus. Econ. Statist.* **19**, 428–35.

CHIB, S., GREENBERG, E. & WINKELMANN, R. (1998). Posterior simulation and Bayes factors in panel count data models. *J. Economet.* **86**, 33–54.

CHIB, S., NARDARI, F. & SHEPHARD, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *J. Economet.* **108**, 281–316.

COX, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scand. J. Statist.* **8**, 93–115.

DE JONG, P. & SHEPHARD, N. (1995). The simulation smoother for time series models. *Biometrika* **82**, 339–50.

DURBIN, J. & KOOPMAN, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with Discussion). *J. R. Statist. Soc.* B **62**, 3–56.

DURBIN, J. & KOOPMAN, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

DURBIN, J. & KOOPMAN, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89**, 603–15.

FAHRMEIR, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *J. Am. Statist. Assoc.* **87**, 501–9.

FRÜHWIRTH-SCHNATTER, S. (1994a). Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering. *Statist. Comp.* **4**, 259–69.

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

FRÜHWIRTH-SCHNATTER, S. (2004). Efficient Bayesian parameter estimation. In *State Space and Unobserved Component Models: Theory and Applications. Proceedings of a Conference in Honour of James Durbin*, Ed. A. Harvey, S. J. Koopman and N. Shephard, pp. 123–51. Cambridge University Press.

FRÜHWIRTH-SCHNATTER, S. & FRÜHWIRTH, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Comp. Statist. Data Anal.* To appear.

GAMERMAN, D. (1998). Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika* **85**, 215–27.

GELFAND, A., SAHU, S. K. & CARLIN, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82**, 479–88.

HARVEY, A. C. & DURBIN, J. (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *J. R. Statist. Soc.* A **149**, 187–227.
HARVEY, A. C. & FERNANDES, C. (1989). Time series models for count or qualitative observations (with Discussion). *J. Bus. Econ. Statist.* **7**, 407–22.
KASHIWAGI, N. & YANAGIMOTO, T. (1992). Smoothing serial count data through a state-space model. *Biometrics* **48**, 1187–94.
KIM, S., SHEPHARD, N. & CHIB, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Rev. Econ. Studies* **65**, 361–93.
KITAGAWA, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with Discussion). *J. Am. Statist. Assoc.* **82**, 1032–63.
LENK, P. J. & DESARBO, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika* **65**, 93–119.
LEROUX, B. G. & PUTERMAN, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48**, 545–58.
MACDONALD, I. & ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series.* London: Chapman and Hall.
MCFADDEN, D. (1974). Conditional logit analysis of qualitative choice behaviour. In *Frontiers of Econometrics*, Ed. P. Zarembka, pp. 105–42. New York: Academic Press.
ROBERT, C. & CASELLA, G. (1999). *Monte Carlo Statistical Methods.* New York: Springer.
ROBERTS, G., PAPASPILIOPOULOS, O. & DELLAPORTAS, P. (2004). Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. *J. R. Statist. Soc.* B **66**, 369–93.
SHEPHARD, N. (1994). Partial non-Gaussian state space. *Biometrika* **81**, 115–31.
SHEPHARD, N. & PITT, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653–67.
TANNER, M. A. & WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Statist. Assoc.* **82**, 528–40.
VAN DYK, D. & MENG, X.-L. (2001). The art of data augmentation (with Discussion). *J. Comp. Graph. Statist.* **10**, 1–111.
WAKEFIELD, J. C., SMITH, A. F. M., RACINE-POON, A. & GELFAND, A. E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Appl. Statist.* **43**, 201–21.
WEST, M., HARRISON, P. J. & MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with Discussion). *J. Am. Statist. Assoc.* **80**, 73–97.
ZEGER, S. & KARIM, M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Am. Statist. Assoc.* **86**, 79–86.
ZEGER, S. L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621–9.
ZEGER, S. L. & QAQISH, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44**, 1019–31.

[*Received May* 2004. *Revised May* 2006]