Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2008-36

Stochastic Model Specification Search for Gaussian and Non-Gaussian State Space Models

Sylvia Frühwirth-Schnatter and Helga Wagner

May 2008

Abstract

Model specification for state space is a difficult task as one has to decide which components to include in the model and to specify whether these components are fixed or time-varying. To this aim a new model space MCMC method is developed in this paper. It is based on extending the Bayesian variable selection approach which is usually applied to model selection for regression models to state space models. For non-Gaussian state space models stochastic model search MCMC makes use of auxiliary mixture sampling. We focus on structural time series models including seasonal components, trend or intervention. The method is applied to various well-known time series.
Key words: auxiliary mixture sampling, Bayesian econometrics, noncentered parameterization, Markov chain Monte Carlo, variable selection

1 Introduction

State space models are widely used in time series analysis to deal with processes which gradually change over time. Model specification, however, is a challenge for these models as one has to specify which components to include and to decide whether they are fixed or time-varying. For state space models, like for many other complex models, this often leads to testing problems which are non-regular from the view-point of classical statistics. Thus, a classical approach toward model selection which is based on hypothesis testing such as a likelihood ratio test or information criteria such as AIC or BIC cannot be easily applied, because it relies on asymptotic arguments based on regularity conditions that are violated in this context.

Consider, for example, modeling a time series $\mathbf{y} = (y_1, \dots, y_T)$ through the dynamic linear trend model, defined for $t = 1, \dots, T$ as:

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{1}$$

where μ_t follows a random walk with a random drift starting from unknown initial values μ_0 and a_0:

$$\mu_t = \mu_{t-1} + a_{t-1} + \omega_{1t}, \qquad \omega_{1t} \sim \mathcal{N}\left(0, \theta_1\right), \tag{2}$$

$$a_t = a_{t-1} + \omega_{2t}, \qquad \omega_{2t} \sim \mathcal{N}\left(0, \theta_2\right). \tag{3}$$

A typical model specification problem for this model is to decide if the drift a_t is really time-varying or if it is fixed. This could be handled by testing $\theta_2 = 0$ versus $\theta_2 > 0$, however, this is a nonregular testing problem, because the null hypothesis lies on the boundary of the parameter space. Another model specification problem is selecting the components in this times series model. Is it necessary to include a dynamic drift term a_t? Testing the null hypothesis $a_0 = a_1 = \cdots = a_T = 0$ versus the alternative where a_t follows a random walk is, again, non-regular because the size of the hypothesis increases with the number of observations.

The Bayesian approach is, in principle, able to deal with such non-regular testing problems. Suppose that K different models $\mathcal{M}_1, \dots, \mathcal{M}_K$ are considered to be candidates for having generated the time series $\mathbf{y}$. In a Bayesian setting each of these models is assigned a prior probability $p(\mathcal{M}_k)$ and the goal is to derive the posterior model probability $p(\mathcal{M}_k|\mathbf{y})$ for each model $\mathcal{M}_k, k = 1, \dots, K$.

There are basically two strategies to cope with the challenge associated with computing the posterior model probabilities. The traditional approach dating back to Jeffreys (1948) and Zellner (1971) determines the posterior model probabilities of each model separately by using Bayes' rule, $p(\mathcal{M}_k|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_k)p(\mathcal{M}_k)$, where $p(\mathbf{y}|\mathcal{M}_k)$ is the marginal likelihood for model $\mathcal{M}_k$. An explicit expression for the marginal likelihood exists only for conjugate problems like linear regression models with normally distributed errors, whereas for more complex models numerical techniques are required. For Gaussian state space models, marginal likelihoods have been estimated using methods such as importance sampling (Frühwirth-Schnatter, 1995; Durbin and Koopman, 2000), Chib's estimator (Chib, 1995), numerical integration (Shively and Kohn, 1997) and bridge sampling (Frühwirth-Schnatter, 2001). Recently, Frühwirth-Schnatter and Wagner (2008) considered estimation of the marginal likelihood for non-Gaussian state space models and demonstrated that the resulting estimators can be pretty inaccurate.

The modern approach to Bayesian model selection is to apply model space MCMC methods by sampling jointly model indicators and parameters, using e.g. the reversible jump MCMC algorithm (Green, 1995) or the stochastic variable selection approach (George and McCulloch, 1993, 1997). The stochastic variable selection approach is commonly applied to model selection for regression models and aims at identifying non-zero regression effects, but it is useful far beyond this problem and allows parsimonious covariance modelling for longitudinal data (Smith and Kohn, 2002) and covariance selection in random effects models (Chen and Dunson, 2003; Frühwirth-Schnatter and Tüchler, 2008).

In the present paper we show that the variable selection approach is also useful for many model selection problems occurring in state space modelling. To perform stochastic model specification search for the dynamic linear trend model defined in (1) to (3), for instance, we introduce three binary stochastic indicators in such a way that the unconstrained model corresponds to setting all indicators equal to 1. Reduced model specifications result by setting certain indicators equal to 0. One of those models, for instance, is the local level model, where the drift component a_t completely disappears:

$$\mu_t = \mu_{t-1} + \omega_{1t}, \qquad \omega_{1t} \sim \mathcal{N}(0, \theta_1). \tag{4}$$

Another interesting special case is the linear trend model, where

$$y_t = \mu_0 + ta_0 + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right). \tag{5}$$

We derive an MCMC method for Gaussian as well as non-Gaussian state space models that performs stochastic model specification search in practice by sampling the indicators simultaneously with the state process and the models parameters. For non-Gaussian state space models applied to binary, multinomial or count data we make use of auxiliary mixture sampling (Frühwirth-Schnatter and Wagner, 2006; Frühwirth-Schnatter and Frühwirth, 2007) which is a simple MCMC method for estimating a broad class of non-Gaussian models.

It is well-known that variable selection is sensitive to the choice of the prior, see e.g. Fernández, Ley, and Steel (2001). Based on a noncentered parameterization of the state space model, we define a new prior for the process variances of the state space model and show that it is far less influential than the usually applied inverted Gamma prior.

Throughout the paper we focus on structural time series models including seasonal components, trend and an intervention effect and apply the method to various well-known time series.

2 The Dynamic Linear Trend Model

Our method is based on a noncentered parameterization of the dynamic linear trend model which is discussed in the next subsection.

2.1 A Noncentered Parameterization

Define two independent random walk processes $\tilde{\mu}_t$ and $\tilde{a}_t$ with standard normal independent increments as well as an integrated process $\tilde{A}_t$:

$$\tilde{\mu}_t = \tilde{\mu}_{t-1} + \tilde{\omega}_{1t}, \qquad \tilde{\omega}_{1t} \sim \mathcal{N}(0,1), \tag{6}$$

$$\tilde{a}_t = \tilde{a}_{t-1} + \tilde{\omega}_{2t}, \qquad \tilde{\omega}_{2t} \sim \mathcal{N}(0,1),$$

$$\tilde{A}_t = \tilde{A}_{t-1} + \tilde{a}_{t-1}, \tag{7}$$

which all are assumed to start at zero: $\tilde{\mu}_0 = \tilde{a}_0 = \tilde{A}_0 = 0$. Combine the state equations (6) to (7) with following observation equation:

$$y_t = \mu_0 + ta_0 + \sqrt{\theta_1}\tilde{\mu}_t + \sqrt{\theta_2}\tilde{A}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right), \tag{8}$$

where μ_0 and a_0 are equal to the initial values for the level and the drift component and θ_1 and θ_2 are equal to the variances in the dynamic linear trend model defined in (1) to (3). The resulting state space model is a noncentered parameterization of the dynamic linear trend model. To verify this define

$$a_t = a_0 + \sqrt{\theta_2}\tilde{a}_t,$$
$$\mu_t = \mu_0 + ta_0 + \sqrt{\theta_1}\tilde{\mu}_t + \sqrt{\theta_2}\tilde{A}_t.$$

Then

$$\begin{aligned}
a_t - a_{t-1} &= \sqrt{\theta_2}(\tilde{a}_t - \tilde{a}_{t-1}) = \sqrt{\theta_2}\tilde{\omega}_{2t} = \omega_{2t}, \qquad \omega_{2t} \sim \mathcal{N}(0,\theta_2), \\
\mu_t - \mu_{t-1} &= \sqrt{\theta_1}(\tilde{\mu}_t - \tilde{\mu}_{t-1}) + a_0 + \sqrt{\theta_2}(\tilde{A}_t - \tilde{A}_{t-1}) \\
&= \sqrt{\theta_1}\tilde{\omega}_{1t} + a_0 + \sqrt{\theta_2}\tilde{a}_{t-1} = \omega_{1t} + a_{t-1}, \qquad \omega_{1t} \sim \mathcal{N}(0,\theta_1),
\end{aligned}$$

which corresponds to the state equations (2) and (3).

The noncentered parameterization of the dynamic linear model has a representation as a state space model with a state vector of dimension 3:

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{9}$$

$$y_t = \mathbf{H}\mathbf{x}_t + \boldsymbol{z}_t^f\boldsymbol{\alpha} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right), \tag{10}$$

where $\mathbf{x}_0 = \mathbf{0}_{3\times 1}$ and

$$\mathbf{x}_t = \begin{pmatrix} \tilde{\mu}_t \\ \tilde{a}_t \\ \tilde{A}_t \end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad \mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$
$$\mathbf{H} = \begin{pmatrix} \sqrt{\theta_1} & 0 & \sqrt{\theta_2} \end{pmatrix}, \qquad \boldsymbol{z}_t^f = \begin{pmatrix} 1 & t \end{pmatrix}, \qquad \boldsymbol{\alpha} = \begin{pmatrix} \mu_0 & a_0 \end{pmatrix}'.$$

This state space form could be used to perform Kalman filtering and to compute the integrated likelihood $p(\mathbf{y}|\boldsymbol{\vartheta})$ for $\boldsymbol{\vartheta} = (\sqrt{\theta_1}, \sqrt{\theta_2}, \sigma_\varepsilon^2, \mu_0, a_0)$.

The noncentered parameterization of the dynamic linear model, however, is not identified, because in the observation equation (8), the sign of $\sqrt{\theta_1}$ and the sequence $\{\tilde{\mu}_t\}_1^T$ may be changed by multiplying all elements with –1 without changing the distribution of $y_1, \ldots, y_T$. If we define a state vector $\mathbf{x}_t^\star = (-\tilde{\mu}_t, \tilde{a}_t, \tilde{A}_t)'$ and a parameter $\boldsymbol{\vartheta}^\star = (-\sqrt{\theta_1}, \sqrt{\theta_2}, \sigma_\varepsilon^2, \mu_0, a_0)$, then $\boldsymbol{\vartheta}^\star$ and $\boldsymbol{\vartheta}$, although being different, define the same integrated likelihood:

$$p(\mathbf{y}|\boldsymbol{\vartheta}) = \int p(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_T, \sqrt{\theta_1}, \sqrt{\theta_2}, \sigma_\varepsilon^2, \mu_0, a_0) p(\mathbf{x}_1, \ldots, \mathbf{x}_T) d(\mathbf{x}_1, \ldots, \mathbf{x}_T)$$
$$= \int p(\mathbf{y}|\mathbf{x}_1^\star, \ldots, \mathbf{x}_T^\star, -\sqrt{\theta_1}, \sqrt{\theta_2}, \sigma_\varepsilon^2, \mu_0, a_0) p(\mathbf{x}_1^\star, \ldots, \mathbf{x}_T^\star) d(\mathbf{x}_1^\star, \ldots, \mathbf{x}_T^\star) = p(\mathbf{y}|\boldsymbol{\vartheta}^\star).$$

Similarly, the sign of $\sqrt{\theta_2}$ and the sequences $\{\tilde{a}_t\}_1^T$ and $\{\tilde{A}_t\}_1^T$ may be changed without changing the distribution of $y_1, \ldots, y_T$ and $\boldsymbol{\vartheta}^\star = (\sqrt{\theta_1}, -\sqrt{\theta_2}, \sigma_\varepsilon^2, \mu_0, a_0)$ and $\boldsymbol{\vartheta}$ define the same integrated likelihood, $p(\mathbf{y}|\boldsymbol{\vartheta}) = p(\mathbf{y}|\boldsymbol{\vartheta}^\star)$.

As a consequence, the likelihood function $p(\mathbf{y}|\boldsymbol{\vartheta})$ is symmetric around 0 in the direction of $\sqrt{\theta_1}$ and $\sqrt{\theta_2}$ and therefore multimodal. If the data are generated by a dynamic linear trend model with parameters $(\theta_1^{\text{tr}}, \theta_2^{\text{tr}}, \boldsymbol{\xi}^{\text{tr}})$, where $\boldsymbol{\xi}^{\text{tr}} = (\sigma_\varepsilon^{2,\text{tr}}, \mu_0^{\text{tr}}, a_0^{\text{tr}})$, then with increasing number of observations T, the modes of the likelihood function will be close to $(\sqrt{\theta_1^{\text{tr}}}, \sqrt{\theta_2^{\text{tr}}}, \boldsymbol{\xi}^{\text{tr}})$, $(-\sqrt{\theta_1^{\text{tr}}}, \sqrt{\theta_2^{\text{tr}}}, \boldsymbol{\xi}^{\text{tr}})$, $(\sqrt{\theta_1^{\text{tr}}}, -\sqrt{\theta_2^{\text{tr}}}, \boldsymbol{\xi}^{\text{tr}})$, and $(-\sqrt{\theta_1^{\text{tr}}}, -\sqrt{\theta_2^{\text{tr}}}, \boldsymbol{\xi}^{\text{tr}})$. If the true variances θ_1^{tr} and θ_2^{tr} are positive, then the likelihood function concentrates around four modes. If one the true variances is equal to 0 while the other is positive, two of those modes collapse and the likelihood is bimodal with increasing T. If both variances θ_1^{tr} and θ_2^{tr} are equal to zero, then the likelihood function becomes unimodal as T increases.

For illustration, Figure 1 shows contour and surface plots of the (scaled) likelihood $p(\mathbf{y}|\sqrt{\theta_1}, \sqrt{\theta_2}, \sigma_\varepsilon^{2,\text{tr}}, \mu_0^{\text{tr}}, a_0^{\text{tr}})$ for a time series of length $T = 1000$ simulated from a dynamic linear trend model with $\mu_0^{\text{tr}} = 0.3$, $a_0^{\text{tr}} = -0.1$ and $\sigma_\varepsilon^{2,\text{tr}} = 1$ and four different combinations of θ_1^{tr} and θ_2^{tr}. There are clearly four modes, if both process variances are positive, two modes, if one of the variances is restricted to zero and a single mode, if both variances are restricted to 0.

Thus by considering the non-centered parameterization and allowing for nonidentifiability we gain important information about the hypothesis whether the variances of the state space model are zero.

2.2 The Parsimonious Dynamic Linear Trend Model

The noncentered parameterization of the dynamic linear model is very useful for model selection both for the components and the dynamics. The observation equation (8) of the noncentered parameterization represents the level of the time series

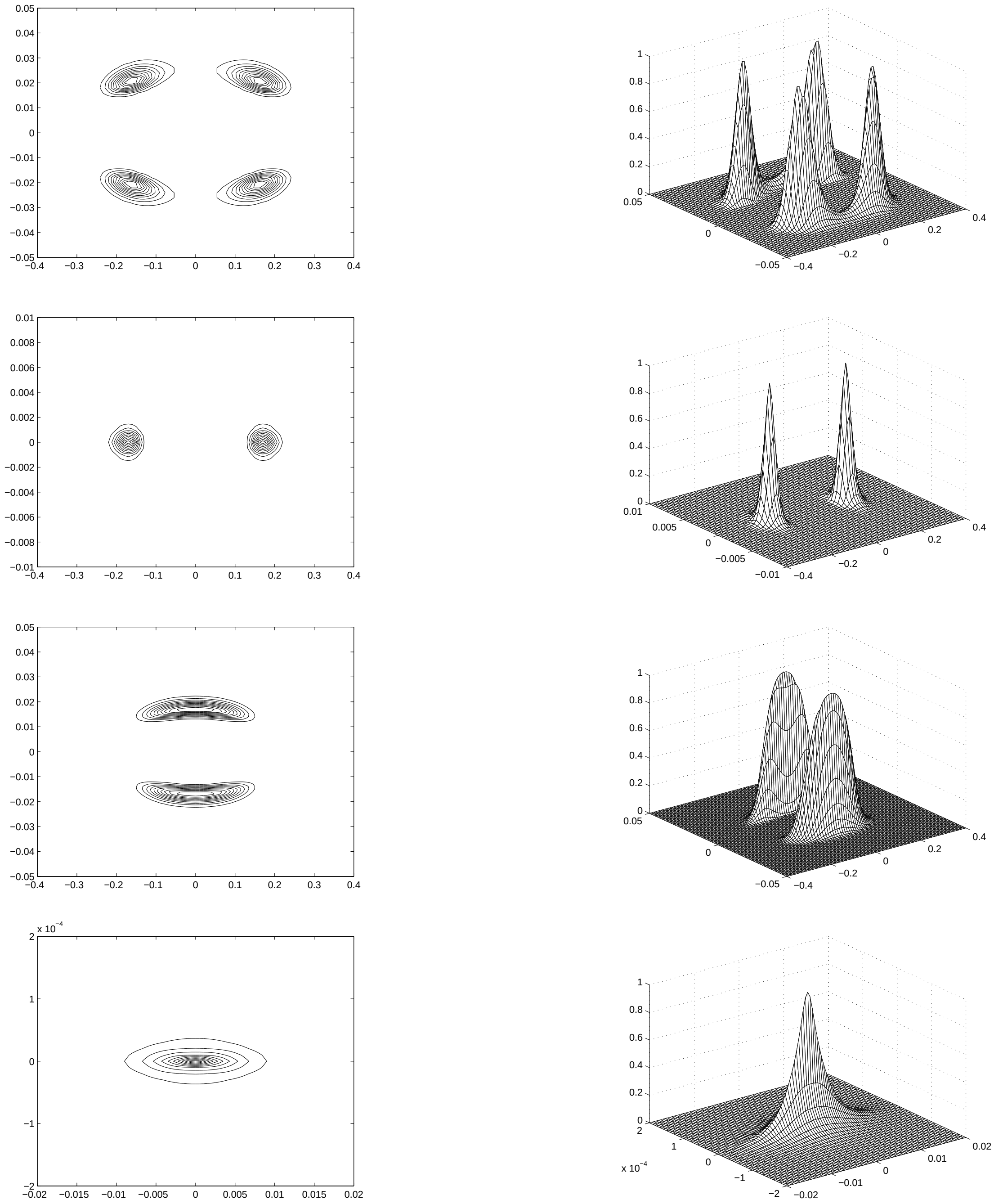

Figure 1: Contour and surface plots of the (scaled) profile likelihood $l(\sqrt{\theta_1}, \sqrt{\theta_2})/\max(l(\sqrt{\theta_1}, \sqrt{\theta_2}))$, where $l(\sqrt{\theta_1}, \sqrt{\theta_2}) = p(\mathbf{y}|\sqrt{\theta_1}, \sqrt{\theta_2}, \sigma_\varepsilon^{2,\text{tr}}, \mu_0^{\text{tr}}, a_0^{\text{tr}})$ for simulated data with $(\theta_1^{\text{tr}}, \theta_2^{\text{tr}}) = (0.15^2, 0.02^2)$ (first row), $(\theta_1^{\text{tr}}, \theta_2^{\text{tr}}) = (0.15^2, 0)$ (second row), $(\theta_1^{\text{tr}}, \theta_2^{\text{tr}}) = (0, 0.02^2)$ (third row), and $(\theta_1^{\text{tr}}, \theta_2^{\text{tr}}) = (0, 0)$ (last row)

y_t as a superposition of the components at time $t = 0$ and the random processes $\tilde{\mu}_t$ and $\tilde{A}_t$. Note that neither $\tilde{\mu}_t$ nor $\tilde{A}_t$ degenerate to a static component. A static component is obtained by setting the appropriate variance equal to 0. For instance, if the variance θ_1 is equal to 0, then $\sqrt{\theta_1} = 0$ and $\tilde{\mu}_t$ is not used to explain y_t. Similarly, if the variance θ_2 is equal to 0, then $\sqrt{\theta_2} = 0$ and $\tilde{A}_t$ is not used to explain y_t. This suggests to consider the choice of the variances θ_1 and θ_2 as a variable selection problem in regression model (8).

To this aim we introduce two binary indicators γ_1 and γ_2, where $\sqrt{\theta_i}$, and consequently θ_i, is equal to 0, if $\gamma_i = 0$. If $\gamma_i = 1$, then $\sqrt{\theta_i}$ is an unconstrained unknown parameter which is estimated from the data under a suitable prior. Evidently, the indicators γ_1 and γ_2 decide if a certain component of the state vector is fixed or changes over time. If both $\gamma_1 = 0$ and $\gamma_2 = 0$, then the model reduces to a regression model with a linear trend, given by (5).

To include or delete the trend, an additional indicator δ is introduced which decides, if the initial slope a_0 is equal to 0 or not. If $\delta = 0$, then a_0 is equal to 0; otherwise, if $\delta = 1$, then a_0 is an unknown parameter which is estimated from the data under a suitable prior. This leads to following parsimonious dynamic linear trend model:

$$\tilde{\mu}_t = \tilde{\mu}_{t-1} + \tilde{\omega}_{1t}, \qquad \tilde{\omega}_{1t} \sim \mathcal{N}(0,1), \tag{11}$$
$$\tilde{a}_t = \tilde{a}_{t-1} + \tilde{\omega}_{2t}, \qquad \tilde{\omega}_{2t} \sim \mathcal{N}(0,1), \tag{12}$$
$$\tilde{A}_t = \tilde{a}_{t-1} + \tilde{A}_{t-1}, \tag{13}$$
$$y_t = \mu_0 + \delta t a_0 + \gamma_1\sqrt{\theta_1}\tilde{\mu}_t + \gamma_2\sqrt{\theta_2}\tilde{A}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0,\sigma_\varepsilon^2\right). \tag{14}$$

For a direct comparison with the usual dynamic linear trend model it is useful to rewrite the parsimonious model in the centered parameterization. Define

$$a_t = \delta a_0 + \sqrt{\theta_2}\tilde{a}_t, \tag{15}$$
$$\mu_t = \mu_0 + \delta t a_0 + \gamma_1\sqrt{\theta_1}\tilde{\mu}_t + \gamma_2\sqrt{\theta_2}\tilde{A}_t. \tag{16}$$

Then (11) to (14) may be rewritten as:

$$\mu_t = \mu_{t-1} + \delta a_0 + \gamma_2(a_{t-1} - \delta a_0) + \gamma_1\omega_{1t}, \qquad \omega_{1t} \sim \mathcal{N}(0,\theta_1), \tag{17}$$
$$a_t = a_{t-1} + \omega_{2t}, \qquad \omega_{2t} \sim \mathcal{N}(0,\theta_2), \tag{18}$$
$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0,\sigma_\varepsilon^2\right). \tag{19}$$

Evidently, $(\delta,\gamma_1,\gamma_2) = (1,1,1)$ corresponds to the unrestricted dynamic linear trend model (2). The combination $(\delta,\gamma_1,\gamma_2) = (0,1,0)$ leads to the local level model (4) which is also known as exponential smoothing, $(\delta,\gamma_1,\gamma_2) = (1,1,0)$ leads to double exponential smoothing, and $(\delta,\gamma_1,\gamma_2) = (1,0,1)$ leads to a smooth trend as in Hodrick-Prescot filtering. The combination $(\delta,\gamma_1,\gamma_2) = (1,0,0)$ leads to a regression model with a deterministic linear trend, given by (5) and $(\delta,\gamma_1,\gamma_2) = (0,0,0)$ leads to i.i.d. normal data, $y_t \sim \mathcal{N}(\mu_0,\sigma_\varepsilon^2)$.

The indicators δ, γ_1 and γ_2 have to be introduced carefully into the centered parametrization. Consider the following alternative choice which appears more natural than (17) and (18), but leads to nonindentifiability:

$$\mu_t = \mu_{t-1} + \delta a_{t-1} + \gamma_1\omega_{1t}, \qquad \omega_{1t} \sim \mathcal{N}(0,\theta_1),$$
$$a_t = a_{t-1} + \gamma_2\omega_{2t}, \qquad \omega_{2t} \sim \mathcal{N}(0,\theta_2).$$

After recursive substitution we get following representation of the model as a normal linear mixed model:

$$y_t = \mu_0 + \delta t a_0 + \gamma_1 \sum_{j=1}^{t} \omega_{1j} + \delta\gamma_2 \sum_{j=1}^{t-1} (t-j)\omega_{2j} + \varepsilon_t,$$

with fixed effects (μ_0, a_0) and random effects $(\omega_{1j}, \omega_{2j}), j = 1 \dots, t$. Only 6 models among the 8 possible combinations of the indicators $(\delta, \gamma_1, \gamma_2)$ are identifiable, because γ_2 is not identified, if $\delta = 0$. In contrast to that, model (17) to (19) has the representation

$$y_t = \mu_0 + \delta t a_0 + \gamma_1 \sum_{j=1}^{t} \omega_{1j} + \gamma_2 \sum_{j=1}^{t-1} (t-j)\omega_{2j} + \varepsilon_t.$$

Evidently, all 8 combinations of the indicators $(\delta, \gamma_1, \gamma_2)$ are identifiable.

The noncentered parameterization of the parsimonious dynamic linear trend model given by (11) to (14) has the following representation as a state space model:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}\right), \\ y_t &= \mathbf{H}(\gamma_1, \gamma_2)\mathbf{x}_t + \boldsymbol{z}_t^f(\delta)\boldsymbol{\alpha} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \end{aligned} \tag{20}$$

where $\mathbf{x}_t$, $\mathbf{F}$, $\mathbf{Q}$ and $\boldsymbol{\alpha}$ are the same as in (9), while $\mathbf{H}$ and $\boldsymbol{z}_t^f$ depend on the model indicators:

$$\mathbf{H}(\gamma_1, \gamma_2) = \begin{pmatrix} \gamma_1\sqrt{\theta_1} & 0 & \gamma_2\sqrt{\theta_2} \end{pmatrix}, \qquad \boldsymbol{z}_t^f(\delta) = \begin{pmatrix} 1 & \delta t \end{pmatrix}.$$

2.3 Prior Distributions

To perform Bayesian estimation one has to choose a prior distribution $p(\delta, \gamma_1, \gamma_2)$ for all possible combinations of indicators. Subsequently, we assume a uniform distribution over all 8 combinations of the indicators. A more flexible distribution is discussed in Section 5.

As common for dynamic linear models, we assume that apriori μ_0 and a_0 are independently normally distributed, $\mu_0 \sim \mathcal{N}\left(y_1, P_{0,11}\sigma_\varepsilon^2\right)$ and $a_0 \sim \mathcal{N}\left(0, P_{0,22}\sigma_\varepsilon^2\right)$. Furthermore we assume an inverted Gamma prior $\mathcal{G}^{-1}\left(c_0, C_0\right)$ for the observation variance σ_ε^2.

In contrast to previous work, we do not use the usual inverted Gamma priors $\theta_1 \sim \mathcal{G}^{-1}\left(d_{0,1}, D_{0,1}\right)$ and $\theta_2 \sim \mathcal{G}^{-1}\left(d_{0,2}, D_{0,2}\right)$ which are the conditionally conjugate priors in the centered version of the dynamic linear model. For reasons that will become clear in Subsection 2.4 MCMC estimation is based on the noncentered version of the dynamic linear trend model. As in this parameterization the parameters $\sqrt{\theta_1}$ and $\sqrt{\theta_2}$ appear as regression coefficients in the regression model (14), the conditionally conjugate priors are given by the normal priors $\sqrt{\theta_1} \sim \mathcal{N}\left(0, B_{0,1}\sigma_\varepsilon^2\right)$ and $\sqrt{\theta_2} \sim \mathcal{N}\left(0, B_{0,2}\sigma_\varepsilon^2\right)$. It be should noted that the two priors are equivalent only under the limiting case of following improper priors: an inverted Gamma prior where $d_{0,i} = -0.5$ and $D_{0,i} = 0$, i.e. $p(\theta_i) \propto \sqrt{\theta_i}$, and a normal prior where $B_{0,i}^{-1} = 0$, i.e. $p(\sqrt{\theta_i}) \propto$ constant.

Apart from being conditionally conjugate for the noncentered parameterization, the normal prior turns out to be more suitable under model specification uncertainty than the inverted Gamma prior. It is well-known, that the hyperparameters in the inverted Gamma prior $\theta_i \sim \mathcal{G}^{-1}(d_{0,i}, D_{0,i})$ strongly influence the posterior density of θ_i, if the true value of θ_i is close to 0.

Consider, for example, a local level model,

$$\begin{aligned} \mu_t &= \mu_{t-1} + \omega_{1t}, \qquad \omega_{1t} \sim \mathcal{N}(0, \theta_1), \\ y_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \end{aligned} \tag{21}$$

where θ_1 is unknown and $\sigma_\varepsilon^2 = 1$ is assumed to be known. To compare the inverted Gamma prior to the normal prior we consider the posterior density of the parameter $\pm\sqrt{\theta_1}$ which is obtained from θ_1 by multiplying the square root of θ_1 with a random sign. We added the $\pm$ sign to make it clear that the sign of this parameter is not identified.

The posterior of $\pm\sqrt{\theta_1}$ allows to explore the hypothesis that $\theta_1 = 0$. Due to the symmetry of the likelihood discussed in Subsection 2.1, the posterior density of $\pm\sqrt{\theta_1}$ is symmetric around zero as long as the prior is also symmetric around 0. If the unknown variance θ_1^{tr} is significantly different from zero, then the posterior density of $\pm\sqrt{\theta_1}$ is likely to be bimodal with the modes being close to $\pm\sqrt{\theta_1^{\mathrm{tr}}}$. Otherwise, if θ_1^{tr} is close to or equal to zero, then the posterior density of $\pm\sqrt{\theta_1}$ is likely be centered around zero.

For illustration, we consider posterior inference for $T = 100$ observations simulated from the local level model (21). In Figure 2 the posterior of $\pm\sqrt{\theta_1}$ is plotted under the $\mathcal{G}^{-1}(0.5, D_0)$-prior for θ_1 and under the normal $\mathcal{N}(0, B_0)$-prior for $\pm\sqrt{\theta_1}$ for two values of θ_1^{tr} and various scale parameters D_0 and B_0. Whereas the posterior is fairly robust to the choice of the variance B_0 in the normal prior, it turns out to be rather sensitive to the scale parameter D_0 of the inverted Gamma prior.

Both posteriors are roughly the same for $\theta_1^{\mathrm{tr}} = 0.01$ and clearly indicate that $\theta_1^{\mathrm{tr}} > 0$. A remarkable difference, however, occurs if $\theta_1^{\mathrm{tr}} = 0$. Under the normal prior, the posterior of $\pm\sqrt{\theta_1}$ is centered at 0 strongly supporting the hypothesis that $\theta_1^{\mathrm{tr}} = 0$. The inverted Gamma density, however, shrinks the posterior of $\pm\sqrt{\theta_1}$ away from 0, falsely indicating that $\theta_1^{\mathrm{tr}} > 0$.

2.4 MCMC Estimation

An MCMC approach is implemented to sample jointly the indicators $(\boldsymbol{\delta}, \boldsymbol{\gamma}) = (\delta, \gamma_1, \gamma_2)$, the unrestricted elements of the parameter $\boldsymbol{\beta} = (\mu_0, a_0, \sqrt{\theta_1}, \sqrt{\theta_2})$, the observation variance σ_ε^2, and the latent state process $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t$ is the state vector defined in (9).

When sampling the indicators $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ it is essential to marginalize over the parameters for which variable selection is carried out, see George and McCulloch (1993, 1997) for a full account. To make this feasible, we use the non-centered parameterization of the dynamic linear trend model. Conditional on the state process $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, the observation equation (14) defines a standard regression model

$$y_t = \mathbf{z}_t^{\boldsymbol{\delta},\boldsymbol{\gamma}} \boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right). \tag{22}$$

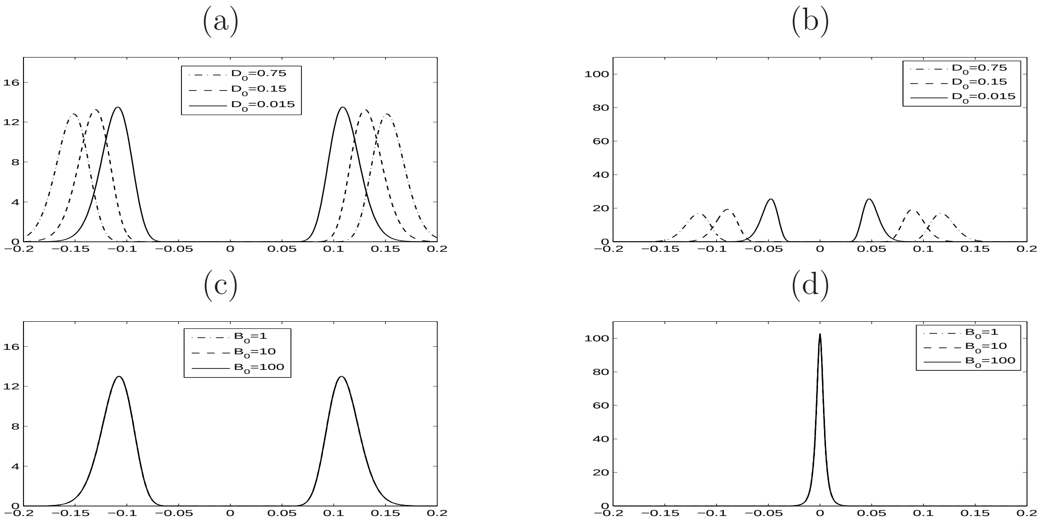

Figure 2: Posterior density for $\pm\sqrt{\theta_1}$ under different priors: top: $\mathcal{G}^{-1}(0.5, D_0)$-prior for θ_1; bottom: $\mathcal{N}(0, B_0\sigma_\varepsilon^2)$ prior for $\pm\sqrt{\theta_1}$; true values: $\sigma_\varepsilon^2 = 1$, $\theta_1 = 0.01$ (left); $\theta_1 = 0$ (right);

If all indicators take the value one, then $\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \boldsymbol{\beta}$ and $\mathbf{z}_t^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \mathbf{z}_t$, where $\mathbf{z}_t = (1, t, \tilde{\mu}_t, \tilde{A}_t)$. Otherwise the restricted parameter $\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ and the corresponding predictors $\mathbf{z}_t^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ contain only those elements of $\boldsymbol{\beta}$ and $\mathbf{z}_t$, respectively, for which the corresponding indicator is equal to 1. Under the conjugate prior

$$\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}} \sim \mathcal{N}\left(\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}, \mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}\sigma_\varepsilon^2\right), \qquad \sigma_\varepsilon^2 \sim \mathcal{G}^{-1}(c_0, C_0), \tag{23}$$

the posterior $p(\boldsymbol{\delta}, \boldsymbol{\gamma}|\mathbf{x}, \mathbf{y})$ is obtained from Bayes' theorem:

$$p(\boldsymbol{\delta}, \boldsymbol{\gamma}|\mathbf{x}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x})p(\boldsymbol{\delta}, \boldsymbol{\gamma}), \tag{24}$$

where $p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x})$ is equal to the marginal likelihood of the regression model (22):

$$p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x}) = \frac{1}{(2\pi)^{T/2}} \frac{|\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}|^{1/2}}{|\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}|^{1/2}} \frac{\Gamma(c_T)C_0^{c_0}}{\Gamma(c_0)(C_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{c_T}}. \tag{25}$$

Here $\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}$, c_T and $C_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ denote the posterior moments of $\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ and σ_ε^2 given below in (26) to (28). It should be noted that such a closed form expression for $p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x})$ is not available if any of the indicators γ_1 and γ_2 is equal to 1 and an inverted Gamma prior is chosen for θ_1 and θ_2. The MCMC scheme reads:

(a) Sample the indicators $(\boldsymbol{\delta}, \boldsymbol{\gamma}) = (\delta, \gamma_1, \gamma_2)$, the initial values μ_0 and a_0, all variance parameters $\sqrt{\theta_1}$ and $\sqrt{\theta_2}$ and the observation variance σ_ε^2 jointly in one block:

(a1) Sample the indicators from $p(\boldsymbol{\delta}, \boldsymbol{\gamma}|\mathbf{x}, \mathbf{y})$ given in (24);

(a2) sample σ_ε^2 from $\mathcal{G}^{-1}\left(c_T, C_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}\right)$, and, conditional on σ_ε^2, sample μ_0, a_0 (if unrestricted), and all unrestricted variance parameters $\sqrt{\theta_1}$ and $\sqrt{\theta_2}$

jointly from the normal posterior $\mathcal{N}\left(\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}, \mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}\sigma_\varepsilon^2\right)$ where

$$\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \left((\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}})'\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}} + (\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\right)^{-1}, \tag{26}$$

$$\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}\left((\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}})'\mathbf{y} + (\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}\right),$$

$$c_T = c_0 + T/2, \tag{27}$$

$$C_T^{\boldsymbol{\delta},\boldsymbol{\gamma}} = C_0 + \frac{1}{2}\left(\mathbf{y}'\mathbf{y} + (\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}})'(\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}} - (\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})'(\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}\right), \tag{28}$$

and $\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ is the regressor matrix with rows equal to $\mathbf{z}_t^{\boldsymbol{\delta},\boldsymbol{\gamma}}$;

(a3) set all restricted initial values and all restricted variances equal to 0.

(b) Sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ from the state space form (20).

(c) Perform a random sign switch for $\sqrt{\theta_1}$ and $\{\tilde{\mu}_t\}_1^T$. Thus with probability 0.5 the draws of these parameters remain unchanged, while they are substituted by $-\sqrt{\theta_1}$ and $\{-\tilde{\mu}_t\}_1^T$ with the same probability. Perform another random sign switch for $\sqrt{\theta_2}$, $\{\tilde{a}_t\}_1^T$ and $\{\tilde{A}_t\}_1^T$.

A few comments are in order. The dimension of the normal distribution appearing in step (a2) depends on the number of unrestricted components and is equal to $1 + \delta + \gamma_1 + \gamma_2$.

In step (b), forward-filtering-backward-sampling (FFBS, Frühwirth-Schnatter (1994); Carter and Kohn (1994); De Jong and Shephard (1995)) is used to sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. To speed up sampling, a reduced state space form is used if γ_1 or γ_2 is 0. If, for instance, $\gamma_1 = 0$, then the observation equation is independent of $\{\tilde{\mu}_t\}_1^T$. FFBS is applied to the reduced state vector $\mathbf{x}_t = (\tilde{a}_t, \tilde{A}_t)'$, while $\tilde{\mu}_1, \ldots, \tilde{\mu}_T$ is sampled from (11). A similar method applies, if $\gamma_2 = 0$, with reduced state vector $x_t = \tilde{\mu}_t$. If both indicators γ_1 and γ_2 are equal to 0, then no FFBS is needed, as sampling of $\mathbf{x}$ from the prior is straightforward.

Sampling of the state process in step (b) is based on the noncentered parameterization. The unknown components a_t and μ_t in the centered parameterization are easily reconstructed from the MCMC draws using (15) and (16).

We found it useful to start from an unrestricted model and to run the first say 1000 draws of burn-in without variable selection. This allows to generate sensible starting values for the state process and the parameters of the unrestricted model before variable selection actually sets in.

3 Extension to the Basic Structural Model

3.1 The Parsimonious Basic Structural Model

In the basic structural model, a seasonal component is added to the dynamic linear trend model discussed in Section 2, see e.g. Harvey (1989):

$$s_t = -s_{t-1} - \cdots - s_{t-S+1} + \omega_{3t}, \qquad \omega_{3t} \sim \mathcal{N}(0, \theta_3), \tag{29}$$

$$y_t = \mu_t + s_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{30}$$

where μ_t is the same as in (2) and (3) and S is the number of seasons. The initial seasonal pattern is given by $\mathbf{s}_0 = (s_{-S+1}, \ldots, s_0)$ with $s_{-S+1} + \ldots + s_0 = 0$. In addition to the model specification problems discussed in Section 2, a decision has to made if a seasonal pattern is present and if this pattern is fixed or dynamic. To this aim, two additional binary stochastic indicators δ_3 and γ_3 are introduced. δ_3 decides, if the initial seasonal pattern is equal to 0, whereas γ_3 controls if it changes over time. As before, the indicators are introduced into the noncentered version of the model.

Combine the following stochastic difference equation:

$$\tilde{s}_t = -\tilde{s}_{t-1} - \cdots - \tilde{s}_{t-S+1} + \tilde{\omega}_{3t}, \qquad \tilde{\omega}_{3t} \sim \mathcal{N}(0, 1), \tag{31}$$

where $\tilde{s}_{-S+1} = \ldots = \tilde{s}_0 = 0$ with the state equations (6) to (7) and following observation equation:

$$y_t = \mu_t + \delta_3 s_{0,q(t)} + \gamma_3 \sqrt{\theta_3} \tilde{s}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right), \tag{32}$$

where θ_3 is equal to the variance of the error term in (29), μ_t is the same as in (16) and $s_{0,q(t)}$ with $q(t) = 1 + (t-1) \mod S$ is the seasonal component corresponding to time t. The resulting state space model is a noncentered parameterization of the basic structural model.

If $\gamma_3 = 0$, we define $\theta_3 = 0$ and the resulting seasonal pattern is fixed. If $\delta_3 = 0$, we set the initial seasonal pattern to zero, $\mathbf{s}_0 = \mathbf{0}$. If both indicators are equal to 0, then no seasonal pattern is present in the time series and the model reduces to the dynamic linear trend model studied in Section 2.

The non-centered parameterization (32) could be written as

$$y_t = \mu_0 + \delta t a_0 + \delta_3 s_{0,q(t)} + \gamma_1 \sum_{j=1}^{t} \omega_{1j} + \gamma_2 \sum_{j=1}^{t-1} (t-j)\omega_{2j} + \gamma_3 \sum_{j=1}^{t} \omega_{3j} + \varepsilon_t,$$

with fixed effects μ_0, a_0 and $s_{0,q(t)}$ and random effects ω_{1j}, ω_{2j} and ω_{3j}. Evidently, all $2^5 = 32$ combinations of indicators are identifiable.

As before, the noncentered model is not identified, as the sign of $\sqrt{\theta_3}$ and the sequence $\{\tilde{s}_t\}_1^T$ may be changed without changing the likelihood function. As a consequence, the likelihood function $p(\mathbf{y}|\boldsymbol{\vartheta})$ where $\boldsymbol{\vartheta} = (\sqrt{\theta_1}, \sqrt{\theta_2}, \sqrt{\theta_3}, \mu_0, a_0, \mathbf{s}_0, \sigma^2_\varepsilon)$ is symmetric around 0 in the direction of $\sqrt{\theta_i}, i = 1, 2, 3$. With an increasing number of observations T, the modes of the likelihood function will be close to all combinations of $(\pm\sqrt{\theta_1^{\text{tr}}}, \pm\sqrt{\theta_2^{\text{tr}}}, \pm\sqrt{\theta_3^{\text{tr}}}, \boldsymbol{\xi}^{\text{tr}})$, where $\boldsymbol{\xi}^{\text{tr}} = (\mu_0^{\text{tr}}, a_0^{\text{tr}}, \mathbf{s}_0^{\text{tr}}, \sigma_\varepsilon^{2,\text{tr}})$. Thus with an increasing number of observations, the likelihood function has eight modes as long as in the data generating process the true variances θ_1^{tr}, θ_2^{tr} and θ_3^{tr} are positive. If one of the true variances is equal to 0 while the others are positive, half of those modes are identical leaving four modes. If two of the true variances are equal to 0 while the other is positive, only two modes are different leaving a bimodal likelihood with an increasing number of observations T. If all variances are equal to zero, then the likelihood function will be unimodal with an increasing number of observations T.

It is easy to verify that in the centered parameterization the parsimonious model is equivalent to combining (2) and (3) with state equation (29) and following observation equation:

$$y_t = \mu_t + \delta_3 s_{0,q(t)} + \gamma_3 (s_t - \delta_3 s_{0,q(t)}) + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right). \tag{33}$$

3.2 MCMC Sampling Scheme

An MCMC approach is implemented to sample the indicators $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$, the model parameters $\boldsymbol{\beta} = (\mu_0, a_0, \mathbf{s}_0, \sqrt{\theta_1}, \sqrt{\theta_2}, \sqrt{\theta_3})$, the observation variance σ_ε^2, and the latent state process $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t$ is following state vector:

$$\mathbf{x}_t = \begin{pmatrix} \tilde{\mu}_t & \tilde{a}_t & \tilde{A}_t & \tilde{s}_t & \ldots & \tilde{s}_{t-S+2} \end{pmatrix}'.$$

The MCMC sampling scheme introduced in Subsection 2.4 is easily modified to deal with a basic structural model. Conditional on the state process $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, the observation equation (32) of the non-centered parameterization of the basic structural model is a standard regression model as in (22) with appropriate regressors $\mathbf{z}_t^{\boldsymbol{\delta},\boldsymbol{\gamma}}$. Under the same conditionally conjugate prior for $\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ and σ_ε^2 as in (23), the marginal likelihood $p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x})$ and all posterior moments are then computed exactly as in Subsection 2.4. This leads to following MCMC scheme:

(a) Sample the indicators $(\boldsymbol{\delta}, \boldsymbol{\gamma})$, the observation variance σ_ε^2 and the initial values μ_0, a_0, and $\mathbf{s}_0$ and all variance parameters $\sqrt{\theta_1}$, $\sqrt{\theta_2}$ and $\sqrt{\theta_3}$ jointly in one block as in Subsection 2.4.

(b) Sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ from the state space form corresponding to (32).

(c) Perform two random sign switches as in step (c) in Subsection 2.4. Perform a third random sign switch for $\sqrt{\theta_3}$ and $\{\tilde{s}_t\}_1^T$.

As in Subsection 2.4, FFBS is applied to a reduced state vector, if any of the indicators $\gamma_i = 0$ is equal to 0, while the remaining components are sampled from the prior.

3.3 Prior Specification

To run the MCMC schemes, prior distributions have to be defined. As before, we assume a uniform prior distribution over all possible indicators $\boldsymbol{\delta} = (\delta, \delta_3)$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$.

For the observation variance σ_ε^2 we choose a hierarchical prior where $\sigma_\varepsilon^2 \sim \mathcal{G}^{-1}(c_0, C_0)$ and $C_0 \sim \mathcal{G}(g_0, G_0)$ with $c_0 = 2.5$, $g_0 = 5$ and $G_0 = g_0/(0.75\text{Var}(y)(c_0 - 1))$. For this hierarchical prior it is necessary to add an additional sampling step were C_0 is sampled conditional on σ_ε^2 from the conditional Gamma posterior $C_0|\sigma_\varepsilon^2 \sim \mathcal{G}(g_0 + c_0, G_0 + 1/\sigma_\varepsilon^2)$ at each sweep of the sampler.

Prior (23) assumes normality not only for the initial values μ_0, a_0, and $\mathbf{s}_0$, but also for all remaining parameters. For the same reasons as in Subsection 2.3, we do not use inverted Gamma priors for the variances $\theta_1, \ldots, \theta_3$ as usual in the basic structural model, but assume that the parameters $\pm\sqrt{\theta_1}$, $\pm\sqrt{\theta_2}$ and $\pm\sqrt{\theta_3}$ follow a normal prior.

In our case studies, we found the following prior choices useful for variable selection. First, we use a partially proper prior which combines the improper prior $p(\mu_0) \propto 1$ for μ_0 with a proper prior $\mathcal{N}\left(\mathbf{0}, \mathbf{B}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}\sigma_\varepsilon^2\right)$ on the remaining unrestricted elements of $\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}}$, where $\mathbf{B}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}} = B_0\mathbf{I}$. This prior corresponds to choosing $\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \mathbf{0}$ and

$$\left(\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}\right)^{-1} = \begin{pmatrix} 0 & \\ & (\mathbf{B}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1} \end{pmatrix}. \tag{34}$$

Under this prior, the sampling scheme described above has to be changed slightly, because the marginal likelihood $p(\mathbf{y}|\boldsymbol{\delta},\boldsymbol{\gamma},\mathbf{x})$ and the posterior parameter c_T read:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\delta},\boldsymbol{\gamma},\mathbf{x}) &= \frac{1}{(2\pi)^{T/2}} \frac{|\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}|^{1/2}}{|\mathbf{B}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}|^{1/2}} \frac{\Gamma(c_T)C_0^{c_0}}{\Gamma(c_0)(C_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{c_T}}, \\ c_T &= c_0 + (T-1)/2. \end{aligned}$$

Another prior commonly used in model selection is the fractional prior (O'Hagan, 1995). In the present context, this is a conditional fractional prior for regression model (22) which depends on the state vector $\mathbf{x}$ and is defined as

$$p(\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}}|\sigma_\varepsilon^2) \quad \propto p(\mathbf{y}|\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}},\sigma_\varepsilon^2)^b = \left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{Tb/2} \exp\left(-\frac{b}{2\sigma_\varepsilon^2}(\mathbf{y}-\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}})'(\mathbf{y}-\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}})\right).$$

The fractional prior can be interpreted as posterior of a non-informative prior and a fraction b of the data $\mathbf{y}$. It reads

$$\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}}|\sigma_\varepsilon^2 \sim \mathcal{N}\left(\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}, \mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}\sigma_\varepsilon^2/b\right),$$

where $\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ and $\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ are the posterior moments under a non-informative prior:

$$\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \left((\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}})'\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}}\right)^{-1}, \qquad \mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}(\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}})'\mathbf{y}. \tag{35}$$

In the MCMC sampling scheme all posterior moments as well as the marginal likelihood $p(\mathbf{y}|\boldsymbol{\delta},\boldsymbol{\gamma},\mathbf{x})$ have to be modified according to:

$$\begin{aligned} c_T &= c_0 + \frac{(1-b)}{2}T, \qquad C_T^{\boldsymbol{\delta},\boldsymbol{\gamma}} = C_0 + \frac{(1-b)}{2}(\mathbf{y}'\mathbf{y} - (\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})'(\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}), \\ p(\mathbf{y}|\boldsymbol{\delta},\boldsymbol{\gamma},\mathbf{x}) &= \frac{b^{q/2}\Gamma(c_T)C_0^{c_0}}{(2\pi)^{T(1-b)/2}\Gamma(c_0)(C_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{c_T}}, \end{aligned}$$

where q is the dimension of $\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}}$, while $\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ and $\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ are the same as in (35).

3.4 UK coal consumption data

We reconsider the series of UK coal consumption, analyzed in Harvey (1989), Frühwirth-Schnatter (1994) and Frühwirth-Schnatter (1995), among others. Data are quarterly from 1/1960 to 4/1982, see Figure 3, panel (a). We model the series on the log scale by a basic structural model.

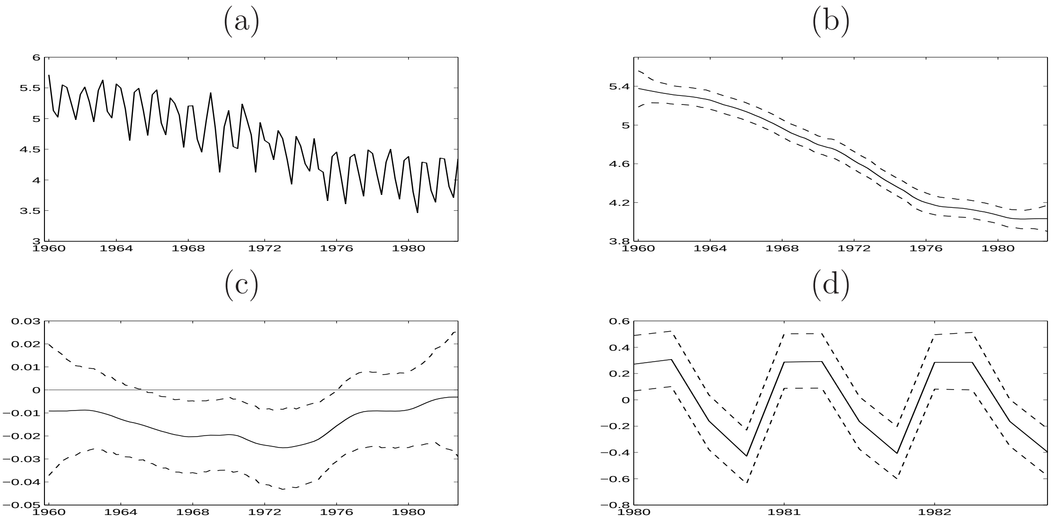

Figure 3: UK coal consumption; (a) observations 1/1960 to 4/1982 (log scale), posterior means and point-wise 95% credible regions of (b) the level μ_t, (c) the drift a_t and (d) the seasonal component s_t in the last three years under the centered parameterization

3.4.1 Comparing the Centered and the Non-centered Parameterization

For illustration, we compare the centered parameterization with priors $\theta_i \sim \mathcal{G}^{-1}(-0.5, 10^{-7})$ with the noncentered parameterization where $\pm\sqrt{\theta_i} \sim \mathcal{N}(0,1)$ for the unrestricted basic structural model without variable selection. The remaining priors are $\mu_0 \sim \mathcal{N}(0, 100\sigma_\varepsilon^2)$ and $\sigma_\varepsilon^2 \sim \mathcal{G}^{-1}(0,0)$. Gibbs sampling was run for 40000 iterations after a burn-in of 10000.

Estimated state components are plotted for the centered parameterization in Figure 3. The posterior densities of the transformed process variances $\pm\sqrt{\theta_i}, i = 1,\ldots,3$ are plotted in Figure 4. Under the centered parameterization, MCMC draws for $\pm\sqrt{\theta_i}$ are obtained by multiplying the square root of the MCMC draws $\theta_i^{(m)}$ with a random sign. Evidently, the posterior density of any parameter $\pm\sqrt{\theta_i}$ has to be symmetric around zero. If the unknown variance θ_i is systematically different from zero, then the posterior density of $\pm\sqrt{\theta_i}$ is likely to be bimodal; otherwise, if θ_i is close to zero, the posterior density of $\pm\sqrt{\theta_i}$ will be centered around zero. This should allow to explore the hypothesis that $\theta_i = 0$.

For the noncentered parameterization, the posterior densities of $\pm\sqrt{\theta_1}$ and $\pm\sqrt{\theta_3}$ are unimodal and centered at 0, while the posterior of $\pm\sqrt{\theta_2}$ is bimodal. This indicates that θ_1 and θ_3 are equal to 0, while $\theta_2 > 0$. This finding is confirmed by stochastic model selection search in Subsection 3.4.2. Under the inverted Gamma prior, all posterior densities are bimodal and $\pm\sqrt{\theta_i}$ is bounded away from 0, providing spurious evidence for an unrestricted model.

A further difference between the centered and the non-centered parameterization lies in the mixing properties of the corresponding MCMC draws. If some variances are equal to or close to 0, the corresponding MCMC draws mix badly under the centered parameterizaton, while mixing is perfect under the noncentered parameterizaton, see Figure 5.

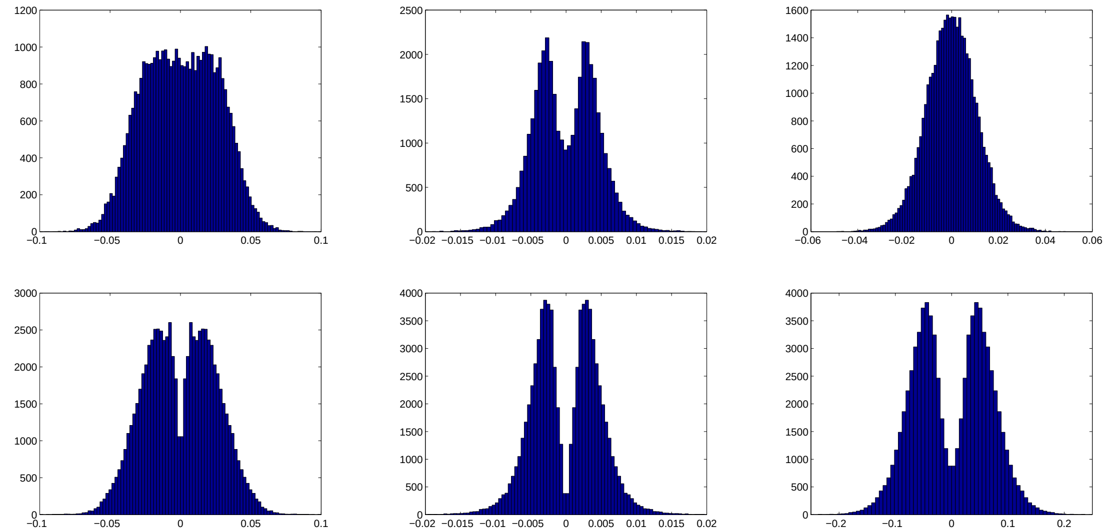

Figure 4: UK coal consumption; posterior densities of $\pm\sqrt{\theta_1}$ (left), $\pm\sqrt{\theta_2}$ (middle) and $\pm\sqrt{\theta_3}$ (right) estimated from the MCMC draws under different priors; top: $\mathcal{N}(0,1)$ prior for $\pm\sqrt{\theta_i}$; bottom: $\mathcal{G}^{-1}(-0.5, 10^{-7})$-prior for θ_i

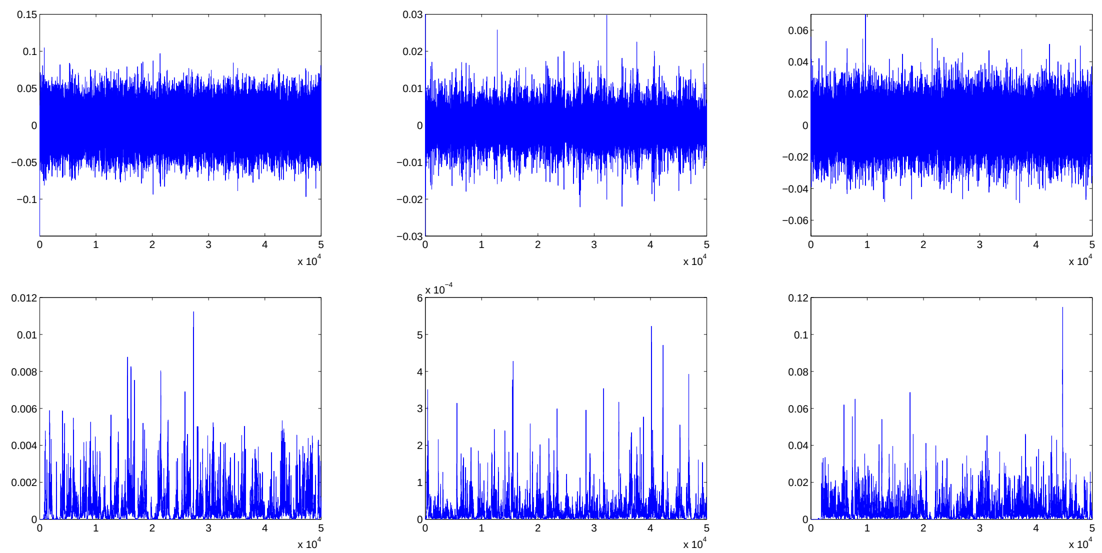

Figure 5: UK coal consumption; MCMC draws for $\pm\sqrt{\theta_1}$ (left), $\pm\sqrt{\theta_2}$ (middle) and $\pm\sqrt{\theta_3}$ (right) under different parameterizations; top: noncentered parameterization with a $\mathcal{N}(0,1)$-prior for $\pm\sqrt{\theta_i}$; bottom: centered parameterization with a $\mathcal{G}^{-1}(-0.5, 10^{-7})$ prior for θ_i

Table 1: Coal data; the three most frequently visited models (among 40000 MCMC iterations) for various prior distributions

prior	δ	δ_3	γ_1	γ_2	γ_3	frequency
$p(\mu_0) \propto 1, B_0 = 1$	**0**	**1**	**0**	**1**	**0**	20331
	1	1	1	0	0	7032
	0	1	1	0	0	6453
$p(\mu_0) \propto 1, B_0 = 100$	**0**	**1**	**0**	**1**	**0**	26454
	0	1	1	0	0	11870
	1	1	1	0	0	818
$b = 10^{-3}$	**0**	**1**	**0**	**1**	**0**	25647
	0	1	1	1	0	4173
	1	1	0	1	0	3994
$b = 10^{-4}$	**0**	**1**	**0**	**1**	**0**	34364
	1	1	0	1	0	1799
	0	1	1	1	0	1675
$b = 10^{-5}$	**0**	**1**	**0**	**1**	**0**	37012
	1	1	0	1	0	1150
	0	1	1	1	0	680

Table 2: Coal data; marginal posterior probability of selecting each indicator under various priors

Prior	δ	δ_3	γ_1	γ_2	γ_3
$p(\mu_0) \propto 1, B_0 = 1$	0.2375	**1.0000**	0.4131	**0.6192**	0.0597
$p(\mu_0) \propto 1, B_0 = 100$	0.0315	**1.0000**	0.3246	**0.6728**	0.0051
$b = 10^{-3}$	0.1845	**1.0000**	0.2048	**0.9295**	0.0698
$b = 10^{-4}$	0.0647	**1.0000**	0.0765	**0.9694**	0.0214
$b = 10^{-5}$	0.0347	**1.0000**	0.0386	**0.9792**	0.0078

3.4.2 Stochastic Model Specification Search

Stochastic model specification search was carried out using partially proper priors with different prior variances B_0 and using fractional priors with different fractions b, see Subsection 3.3. MCMC sampling was carried out for $M = 40000$ draws after a burn-in of 10000 draws. The first 1000 draws of the burn-in were drawn from the unrestricted model, model selection began after these first 1000 draws.

Results of the variable selection procedure are summarized in Table 1 and 2. The most frequently visited model in Table 1 is robust against the prior choice, only the frequency with which this model is selected varies. The same model results for all priors, if in Table 2 an indicator is estimated to be 1, if the corresponding marginal posterior probability is greater or equal to 0.5.

As expected from panel (a) and (d) in Figure 3, a seasonal pattern is present in the selected model ($\delta_3 = 1$), but it is fixed and does not change over time ($\gamma_3 = 0$). The drift a_t is stochastic ($\gamma_2 = 1$), but the initial value a_0 is selected to be 0 ($\delta = 0$). This is plausible from panel (c) in Figure 3, where the pointwise confidence band covers $a_t = 0$ at $t = 0$, but does not contain the restricted line where $a_t = 0$ for all t. Finally, no additional noise ω_{1t} is added in (2), since $\gamma_1 = 0$. This finding confirms model choice based on the marginal likelihoods as in Frühwirth-Schnatter (1995).

4 Model Selection for Non-Gaussian State Space Models

The variable selection approach developed for Gaussian state space model may be extended to nonnormal state space models using auxiliary mixture sampling (Frühwirth-Schnatter and Wagner, 2006; Frühwirth-Schnatter and Frühwirth, 2007). This allow variable selection for state space modelling of times series of small counts based on the Poisson distribution and of binary as well as categorical time series based on the logit transform. We provide an illustrative application to two time series of small counts.

4.1 A Basic structural model for Count Data including Intervention

For count data the basic structural model reads (Harvey and Durbin, 1986):

$$y_t \sim \mathcal{P}\left(e_t \lambda_t\right),$$
$$\log \lambda_t = \mu_t + s_t, \tag{36}$$
$$\mu_t = \mu_{t-1} + a_{t-1} + \omega_{1t}, \qquad \omega_{1t} \sim \mathcal{N}\left(0, \theta_1\right) \tag{37}$$
$$a_t = a_{t-1} + \omega_{2t}, \qquad \omega_{2t} \sim \mathcal{N}\left(0, \theta_2\right), \tag{38}$$
$$s_t = -s_{t-1} - \cdots - s_{t-S+1} + \omega_{3t}, \qquad \omega_{3t} \sim \mathcal{N}\left(0, \theta_3\right). \tag{39}$$

To account for the intervention at $t = t_{int}$, equation (38) is modified in the following way:

$$\mu_t = \mu_{t-1} + a_{t-1} + \Delta + \omega_{1t}.$$

4.1.1 Stochastic model specification search

Indicators $\delta, \delta_3, \gamma_1, \gamma_2$ and γ_3 are introduced as in Section 3 to select the structural components, and an additional indicator δ_4 is introduced for the intervention effect. In the centered parameterization, equations (36) and (37) are modified in the following way:

$$\log \lambda_t = \mu_t + \delta_3 s_{0,q(t)} + \gamma_3 (s_t - \delta_3 s_{0,q(t)}), \tag{40}$$
$$\mu_t = \mu_{t-1} + \delta a_0 + \gamma_2 (a_{t-1} - \delta a_0) + \delta_4 I_{\{t=t_{int}\}} \Delta + \gamma_1 \omega_{1t}, \tag{41}$$

while the (38) and (39) are unaffected. For MCMC estimation, the noncentered version of this model is required which reads:

$$\log \lambda_t = \mu_0 + \delta t a_0 + \delta_3 s_{0,q(t)} + \delta_4 I_{\{t \geq t_{int}\}} \Delta + \gamma_1 \sqrt{\theta_1} \tilde{\mu}_t + \gamma_2 \sqrt{\theta_2} \tilde{A}_t + \gamma_3 \sqrt{\theta_3} \tilde{s}_t,$$

where $\tilde{\mu}_t$ and $\tilde{A}_t$ are defined as in (11) to (13), while $\tilde{s}_t$ is defined as in (31).

4.1.2 MCMC Estimation

MCMC estimation is implemented using auxiliary mixture sampling for count data (Frühwirth-Schnatter and Wagner, 2006). For each t, the distribution of $y_t|\lambda_t$ is regarded as the distribution of the number of jumps of an unobserved Poisson process with intensity $e_t\lambda_t$, having occurred in the time interval [0,1]. The first step of data augmentation creates such a Poisson process for each y_t, $t = 1, \ldots, T$, and introduces the inter-arrival times τ_{tj}, $j = 1, \ldots, (y_t + 1)$ of this Poisson process as missing data. Since each $\tau_{tj} \sim \mathcal{E}(e_t\lambda_t)$ we have

$$-\log \tau_{tj} = \log e_t + \log \lambda_t + \varepsilon_{tj},$$

where $\varepsilon_{tj} = -\log \xi_{tj}$ with $\xi_{tj} \sim \mathcal{E}(1)$. The distribution of ε_{tj} is then approximated by a mixture of normal distributions with component indicator r_{tj}:

$$p_\varepsilon(\varepsilon_{tj}) = \exp\{-\varepsilon_{tj} - e^{-\varepsilon_{tj}}\} \approx \sum_{r_{tj}=1}^{10} w_{r_{tj}} f_N(\varepsilon_{tj}; m_{r_{tj}}, s^2_{r_{tj}}).$$

The quantities $(w_j, m_j, s^2_j), j = 1, \ldots, 10$ are the parameters of the finite mixture approximation tabulated in Frühwirth-Schnatter and Frühwirth (2007, Table 1).

Introducing the auxiliary variables $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_T)$, where $\mathbf{u}_t = (\tau_{tj}, r_{tj}, j = 1, \ldots, y_t + 1)$, leads to a conditionally Gaussian state space model:

$$\begin{aligned} -\log \tau_{tj} &= \mu_0 + \delta t a_0 + \delta_3 s_{0,q(t)} + \delta_4 I_{\{t \geq t_{int}\}} \Delta \\ &+ \gamma_1 \sqrt{\theta_1} \tilde{\mu}_t + \gamma_2 \sqrt{\theta_2} \tilde{A}_t + \gamma_3 \sqrt{\theta_3} \tilde{s}_t + m_{r_{tj}} + \varepsilon_{tj}, \qquad \varepsilon_{tj} \sim \mathcal{N}\left(0, s^2_{r_{tj}}\right) \end{aligned} \tag{42}$$

where $j = 1, \ldots, y_t + 1$. An improved version of auxiliary mixture sampling discussed in Frühwirth-Schnatter, Frühwirth, Held, and Rue (2007) could be applied where the maximum dimension of $\mathbf{u}_t$ is equal to 4 rather than $2(y_t + 1)$.

In (42) we are dealing with a state space model that is conditionally Gaussian with the state vector $\mathbf{x}_t$ being the same as in Subsection 3.2. The MCMC scheme introduced in Subsection 3.2 for Gaussian state space models needs only a few modifications. First, an additional step has to be added to draw the auxiliary variables $\mathbf{u}$. Second, conditional on the state vector, we are dealing with a regression model with heteroscedastic normal errors with known error variance:

$$\tilde{\mathbf{y}} = \mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}} \boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{43}$$

where $\tilde{\mathbf{y}}$ denotes the collection of the auxiliary variables $(-\log \tau_{tj} - m_{r_{tj}})$ and $\boldsymbol{\Sigma}$ is a diagonal matrix with elements $s^2_{r_{tj}}$. Under the normal prior $\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}} \sim \mathcal{N}\left(\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}, \mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}\right)$, the marginal likelihood in this regression model defines $p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x}, \mathbf{u})$:

$$\begin{aligned} &p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x}, \mathbf{u}) \\ &= \frac{|\boldsymbol{\Sigma}|^{-1/2} |\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}|^{1/2}}{(2\pi)^{T/2} |\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}|^{1/2}} \exp\left(-\frac{1}{2}\left(\tilde{\mathbf{y}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{y}} + (\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}})'(\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}} - (\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})'(\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}\right)\right), \end{aligned} \tag{44}$$

where

$$(\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1} = ((\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}})'\boldsymbol{\Sigma}^{-1}\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}} + (\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}), \tag{45}$$

$$\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}((\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}})'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{y}} + (\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}). \tag{46}$$

The MCMC scheme reads:

(a1) Sample $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ from $p(\boldsymbol{\delta}, \boldsymbol{\gamma}|\mathbf{x}, \mathbf{u}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x}, \mathbf{u})p(\boldsymbol{\delta}, \boldsymbol{\gamma})$ conditional on the state process $\mathbf{x}$ and the auxiliary variables $\mathbf{u}$ using the marginal likelihood (44) obtained from regression model (42).

(a2) Sample all *unrestricted* elements of the initial values of $\mathbf{x}_0$ and all *unrestricted* variance parameters $\sqrt{\theta_i}$ jointly from the multivariate normal distribution $\mathcal{N}\left(\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}, \mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}\right)$ conditional on $\mathbf{x}$ and $\mathbf{u}$ using the moments (45) and (46); set all remaining initial values of $\mathbf{x}_0$ and all remaining variances equal to 0.

(b) Sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ from an appropriate state space form;

(c) Perform random sign switches as in step (c) in Subsection 3.2.

(d) Sample the auxiliary variables $\mathbf{u}$ conditional on the current risk $\lambda_1, \ldots, \lambda_T$, see Frühwirth-Schnatter and Wagner (2006) for more details:

  (d1) For each $t = 1, \ldots, T$, sample the order statistics of y_t uniform random variables and define the inter-arrival times τ_{tj} for $j = 1, \ldots, y_t$ as their increments. Sample the final arrival time as $\tau_{t,n+1} = 1 - \sum_{j=1}^{n} \tau_{tj} + \xi_t$, where $\xi_t \sim \mathcal{E}(\lambda_t)$.

  (d2) Sample the component indicator r_{tj} conditional on τ_{tj} and λ_t from a discrete density.

Note that in step (a) marginalizing over the variables and components which are subject to model selection would not be possible for non-Gaussian state space models without the use of auxiliary mixture sampling or another augmentation scheme that leads to a conditionally Gaussian model. Such data augmentation schemes which enable variable selection in non-Gaussian models have been applied earlier by Holmes and Held (2006) for binary and multinomial regression model and by Tüchler (2008) for binary and multinomial regression models with random effects.

The partially proper normal prior and the fractional prior considered in Subsection 3.3 are easily adjusted for non-Gaussian state space model. A partially proper normal prior which combines $p(\mu_0) \propto 1$ with a proper prior $\mathcal{N}\left(\mathbf{0}, \mathbf{B}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}\right)$ on the remaining unrestricted elements of $\boldsymbol{\beta}^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ where $\mathbf{B}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}} = B_0\mathbf{I}$ corresponds to $\mathbf{a}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \mathbf{0}$ and $\mathbf{A}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ being the same as in (34). The marginal likelihood reads

$$p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x}, \mathbf{u}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}|\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}|^{1/2}}{(2\pi)^{T/2}|\mathbf{B}_0^{\boldsymbol{\delta},\boldsymbol{\gamma}}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\tilde{\mathbf{y}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{y}} - (\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})'(\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})\right).$$

For a fractional prior, derived as in Subsection 3.3, the marginal likelihood is given as

$$p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{x}, \mathbf{u}) = b^{q/2}\left(\frac{|\boldsymbol{\Sigma}|^{-1}}{(2\pi)^T}\right)^{(1-b)/2} \cdot \exp\left(-\frac{(1-b)}{2}(\tilde{\mathbf{y}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{y}} - (\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})'(\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1}\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})\right), \tag{47}$$

where $(\mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}})^{-1} = (\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}})'\boldsymbol{\Sigma}^{-1}\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}}$ and $\mathbf{a}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}} = \mathbf{A}_T^{\boldsymbol{\delta},\boldsymbol{\gamma}}(\mathbf{Z}^{\boldsymbol{\delta},\boldsymbol{\gamma}})'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{y}}$.

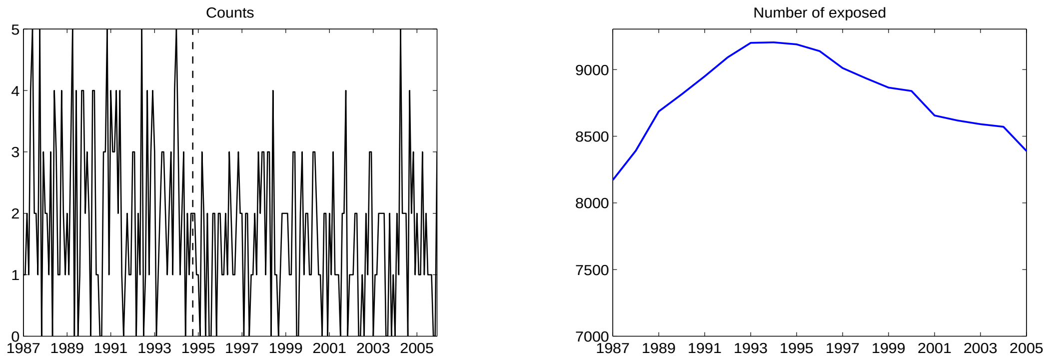

Figure 6: Road safety data; (a) counts of killed or injured children, (b) number of children exposed

4.2 Road Safety Data

We analyze a time series consisting of monthly counts of killed or injured pedestrians, aged 6-10, from 1987-2005 in Linz, which is the third largest town in Austria.[1] The observations are a series of small counts not exceeding 5, see Figure 6. A new law intended to increase road safety came into force in Austria on October 1, 1994, since when pedestrians who want to use a pedestrian crossing have to be allowed to cross. Of interest is the effect of this law on the (monthly) risk of being killed or seriously injured in a road accident as a child living in Linz.

The basic structural model with intervention effect for Poisson counts defined in Subsection 4.1 is fitted to the number y_t of children killed or seriously injured in time period t, $y_t \sim \mathcal{P}\left(e_t \lambda_t\right)$, where e_t is the number of children living in Linz. Model specification search is carried out to identify an appropriate model.

4.2.1 Comparing the Centered and the Noncentered Parameterization

For these data, we were not able to estimate the model under the completely centered parameterization as MCMC did not convergence. For this reason, we compare a parameterization where only the season is non-centered (Frühwirth-Schnatter and Wagner, 2006) under the priors $\theta_i \sim \mathcal{G}^{-1}\left(0.1, 0.001\right), i = 1, 2$, $\pm\sqrt{\theta_3} \sim \mathcal{N}\left(0, 1\right)$ with a fully noncentered model with priors $\pm\sqrt{\theta_i} \sim \mathcal{N}\left(0, 1\right), i = 1, 2, 3$. For both parameterizations we assume that $\mu_0 \sim \mathcal{N}\left(\log(y_1/e_1), 1\right) = \mathcal{N}\left(-9.0084, 1\right)$ and that the unknown initial values of the other components and the intervention effect follow a standard normal prior distribution. We used 20000 iterations after a burn-in of 5000 for each parameterization. As in Subsection 3.4.1, we observe much better mixing behavior of the MCMC sampler under the non-centered parametrization, see Figure 7.

Figure 8 shows histograms of the MCMC draws for $\pm\sqrt{\theta_i}, i = 1, \ldots, 3$ for both parameterizations. For the noncentered parameterization with the normal prior the posterior of all parameters $\pm\sqrt{\theta_i}, i = 1, \ldots, 3$ is clearly centered at 0, suggesting

[1] A shorter version of this time series ranging from 1987-2002 was analyzed in Frühwirth-Schnatter and Wagner (2006).

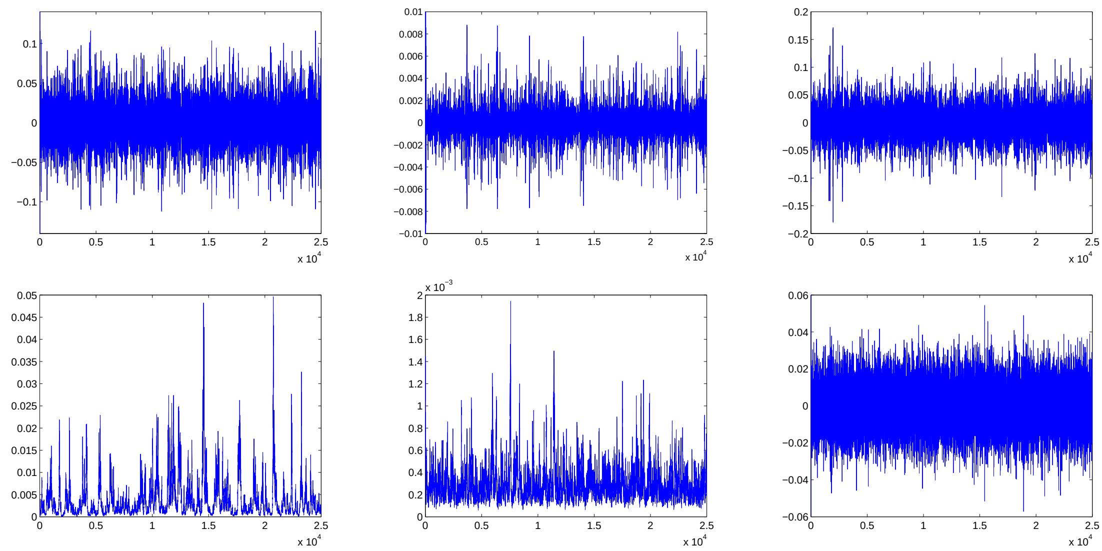

Figure 7: Road safety data; top: MCMC draws for $\pm\sqrt{\theta_1}$ (left), $\pm\sqrt{\theta_2}$ (middle) and $\pm\sqrt{\theta_3}$ (right) for the noncentered parameterization; bottom: MCMC draws for θ_1 (left), θ_2 (middle) and $\pm\sqrt{\theta_3}$ (right) under the partially noncentered parameterization

that the state space model is overfitting and the data may be explained by a simply Poisson regression model. This finding is confirmed by stochastic model specification search in Subsection 4.2.2. As in Subsection 3.4.1 the inverted Gamma density is very influential and shrinks the posterior densities of $\pm\sqrt{\theta_1}$ and $\pm\sqrt{\theta_2}$ away from 0, spuriously suggesting that $\theta_1 > 0$ and $\theta_2 > 0$.

Figure 9 shows the various components of the unconstrained model like the smoothed level μ_t with pointwise 95% credibility intervals under the centered parameterization. The estimated monthly risk λ_t for a child to be seriously injured or killed seems to decrease at the time of intervention. The drift a_t is not significantly different from 0 over the whole observation period. The seasonal component has significantly lower values than the annual average in the holiday months July and August and higher values in June and October.

4.2.2 Stochastic Model Specification Search

Stochastic model specification search was carried out using partially proper priors with different prior variances B_0 and using fractional priors with different fractions b. MCMC sampling was carried out for 40000 iterations after a burn-in of 10000. The first 1000 draws of the burnin were drawn from the unrestricted model, model selection began after these first 1000 draws.

Results of the variable selection procedure are summarized in Table 3 and 4.

The most frequently visited model is fairly robust against the choice of the prior, only the fractional prior with the smallest b leads to a more parsimonious model. No trend is present in the selected model, because $\delta = 0$ and $\gamma_2 = 0$ imply that $a_t = a_0 = 0$ for the whole observation period. The initial seasonal pattern is significant ($\delta_3 = 1$), but does not change over time ($\gamma_3 = 0$). The level of the model

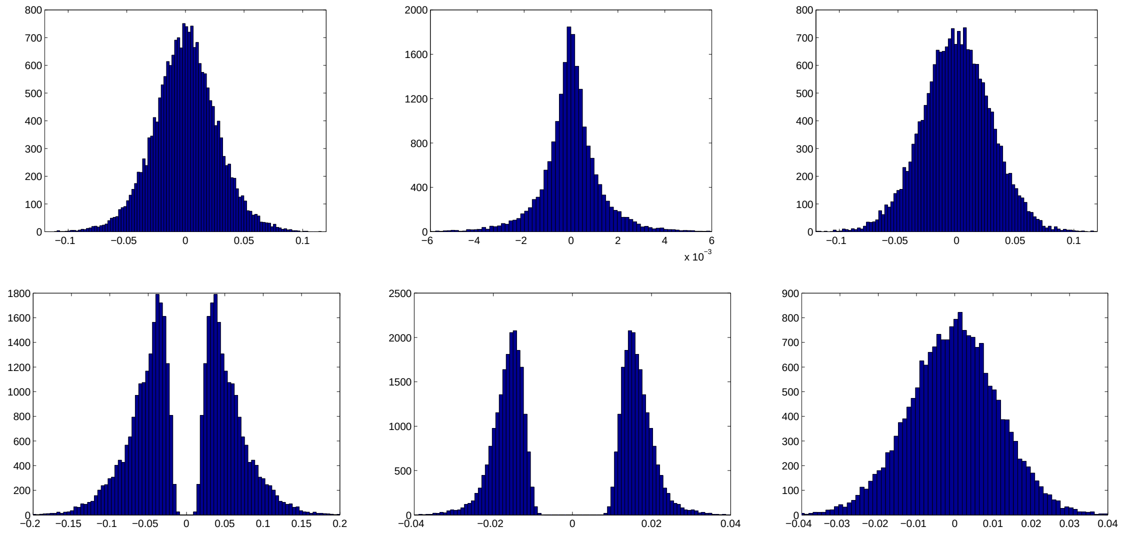

Figure 8: Road safety data; histograms of the MCMC draws for $\pm\sqrt{\theta_1}$ (left), $\pm\sqrt{\theta_2}$ (middle) and $\pm\sqrt{\theta_3}$ (right); top: $\mathcal{N}(0,1)$ prior for $\pm\sqrt{\theta_i}, i = 1, 2, 3$; bottom: $\mathcal{G}^{-1}(0.1, 0.001)$-prior for θ_1 and θ_2, $\mathcal{N}(0,1)$ prior for $\pm\sqrt{\theta_3}$

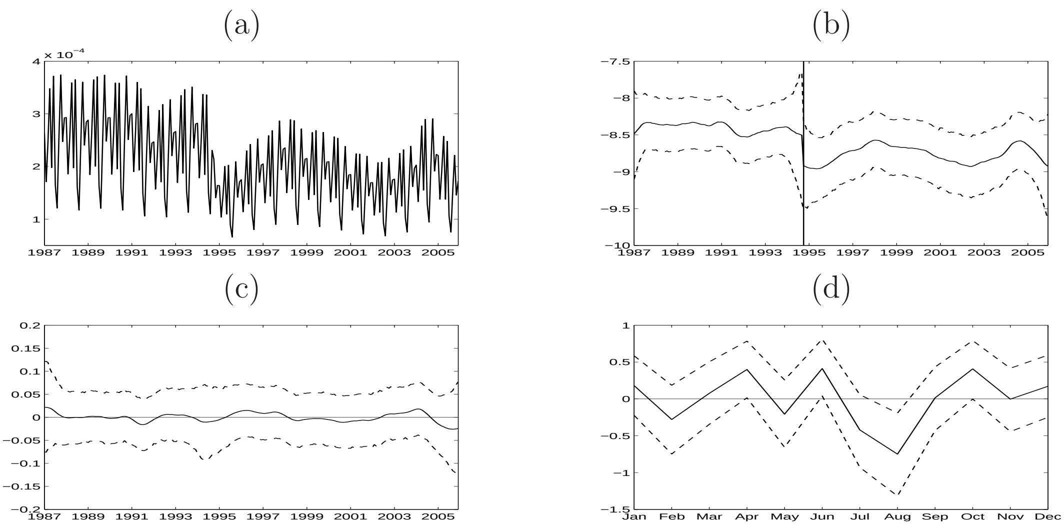

Figure 9: Road safety data; (a) posterior mean of the risk λ_t; posterior means and point-wise 95% credible regions of (b) the level μ_t, (c) the drift a_t and (d) the seasonal component s_t in year 2005 under the centered parameterization

Table 3: Road safety data; the three most frequently visited models (among 40000 MCMC iterations) for various prior distributions

prior	δ	δ_3	δ_4	γ_1	γ_2	γ_3	frequency
$p(\mu_0) \propto 1, B_0 = 1$	**0**	**1**	**1**	**0**	**0**	**0**	37544
	0	1	1	0	0	1	937
	0	1	1	1	0	0	611
$p(\mu_0) \propto 1, B_0 = 100$	**0**	**1**	**1**	**0**	**0**	**0**	34874
	0	1	0	0	0	0	4743
	0	1	1	0	0	1	125
$b = 10^{-2}$	**0**	**1**	**1**	**0**	**0**	**0**	9595
	0	1	1	0	0	1	4154
	0	1	1	1	0	0	3414
$b = 10^{-3}$	**0**	**1**	**1**	**0**	**0**	**0**	18528
	1	1	0	0	0	0	4048
	0	1	1	0	0	1	2686
$b = 10^{-4}$	**0**	**1**	**1**	**0**	**0**	**0**	24871
	1	1	0	0	0	0	4717
	0	1	0	0	1	0	2514
$b = 10^{-5}$	**0**	**0**	**1**	**0**	**0**	**0**	14298
	0	0	1	0	0	1	7823
	1	0	0	0	0	0	3991

is constant before and after intervention, because $\gamma_1 = 0$. Most importantly, the intervention effect is significant, because $\delta_4 = 1$ is selected. Interestingly, the selected model is no longer a state space model ($\gamma_1 = \gamma_2 = \gamma_3 = 0$), but a simple Poisson regression model with monthly seasonal dummies and an intervention effect. This finding is confirmed by the marginal likelihoods computed in Frühwirth-Schnatter and Wagner (2008).

In Table 5 and Figure 10, we compare posterior inference for the intervention effect for the unconstrained basic structural model and the model obtained by variable selection. We observe here an impressive gain of statistical efficiency for this parameter of interest. For the unconstrained basic structural model, making the level dynamic before and after the intervention causes quite a loss of information, leading to an intervention effect that is not significant.

Table 4: Road safety data; marginal posterior probability of selecting each indicator

	trend	season	intervention	process variances		
prior	δ	δ_3	δ_4	γ_1	γ_2	γ_3
$p(\mu_0) \propto 1, B_0 = 1$	0.0047	**1.0000**	**0.9798**	0.0209	0.0005	0.0244
$p(\mu_0) \propto 1, B_0 = 100$	0.0019	**1.0000**	**0.8767**	0.0042	0.0001	0.0035
$b = 10^{-2}$	0.3140	**1.0000**	**0.7769**	0.2872	0.2767	0.3015
$b = 10^{-3}$	0.2152	**1.0000**	**0.7094**	0.1567	0.1576	0.1289
$b = 10^{-4}$	0.1563	**1.0000**	**0.7196**	0.0772	0.0963	0.0501
$b = 10^{-5}$	0.1753	0	**0.5718**	0.0734	0.0971	0.3514

Table 5: Road safety data; posterior inference for the intervention effect Δ

Δ	Mean	Std.dev.	95%H.P.D. regions
Basic structural model	-0.4070	0.4954	[-1.3741; 0.5756]
Poisson regression model with seasonal dummies	-0.3579	0.0977	[-0.5519; -0.1726]

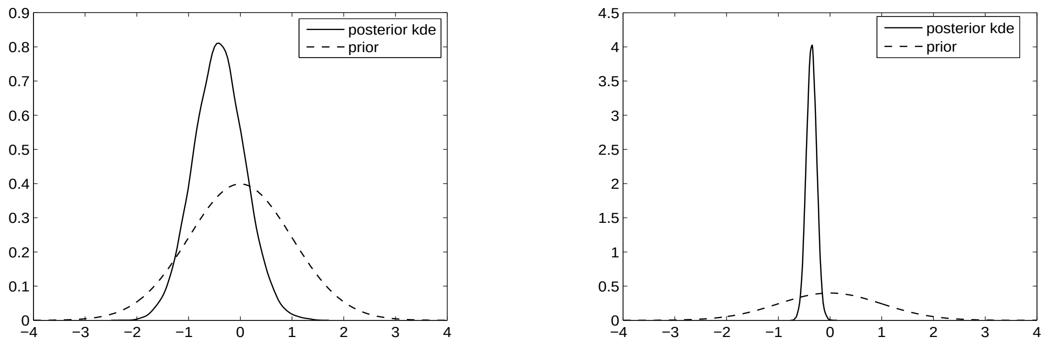

Figure 10: Road safety data; posterior density of the intervention effect Δ in comparison to the prior; left: unrestricted basic structural model, right: Poisson regression model with seasonal pattern (selected model)

The seasonal pattern disappears, if b is rather small in the fractional prior, see Table 3 and 4. Decreasing b forces more parsimonious models. Not surprisingly the seasonal pattern disappears first in a more parsimonious model, because in the selected model only a few seasonal dummies are different from 0, see also Figure 9.

4.3 Purse snatching in Hyde Park, Chicago

For further illustration, we reanalyze a time series of cases of purse snatching y_t in the Hyde park neighborhood in Chicago (Harvey, 1989) reported for the period from January 1968 to September 1973. We consider a simplified version of the model introduced in Subsection 4.1, where no seasonal and no intervention effect is present and the exposures e_t are equal to 1.

First, Gibbs sampling was run without variable selection for 15000 iterations after a burn-in of 10000 for both parameterizations. We selected the normal prior $\mathcal{N}(0, 10)$ both for μ_0 and a_0, while $\theta_i \sim \mathcal{G}^{-1}(-0.5, 0.0001)$ under the centered and $\pm\sqrt{\theta_i} \sim \mathcal{N}(0, 1)$ under the noncentered parameterization, for $i = 1, 2$. Figure 11 shows histograms of $\pm\sqrt{\theta_i}$ under both priors. The posterior for $\pm\sqrt{\theta_1}$ is roughly the same under both priors and clearly indicates that $\theta_1 > 0$. Again, the inverted Gamma density is too influential for θ_2 and shrinks the draws away from 0, while for the normal prior the posterior is clearly centered at 0, suggesting that $\theta_2 = 0$.

Second, stochastic model specification search was carried out using partially proper priors with various prior variances B_0 and using fractional priors with various fractions b. The sampling scheme was run for $M = 40000$ iterations after a burn-in of 10000. The first 1000 draws of the burn-in were drawn from the unrestricted

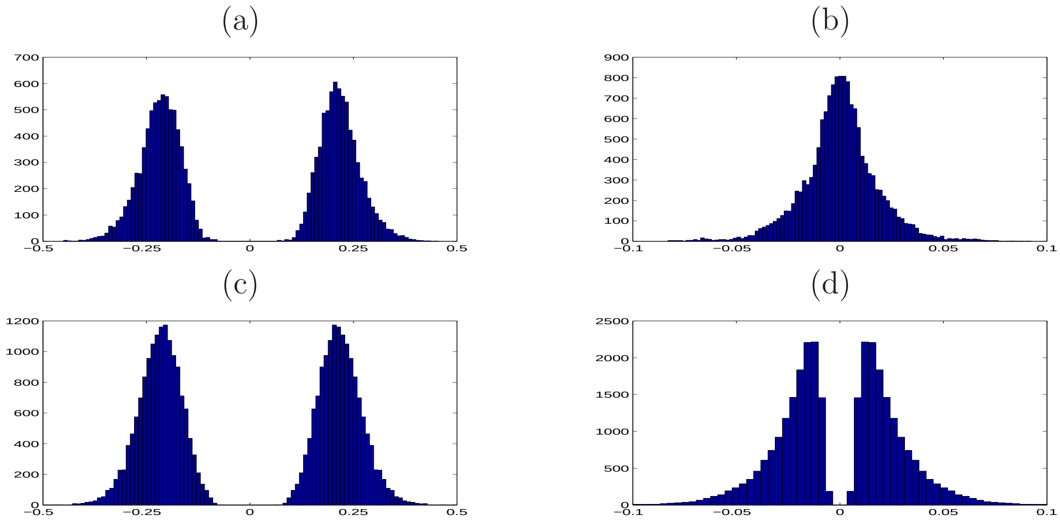

Figure 11: Purse snatching data; histograms for $\pm\sqrt{\theta_1}$ (left) and $\pm\sqrt{\theta_2}$ (right); top: $\mathcal{N}(0,1)$ prior for $\pm\sqrt{\theta_1}$ and $\pm\sqrt{\theta_2}$; bottom: $\mathcal{G}^{-1}(-0.5, 0.0001)$-prior for θ_1 and θ_2

Table 6: Purse snatching; the three most frequently visited models (among 40000 MCMC iterations) for various prior distributions

δ	γ_1	γ_2	$B_0 = 1$	$B_0 = 100$	$b = 10^{-3}$	$b = 10^{-4}$	$b = 10^{-5}$
0	**1**	**0**	39436	39822	20944	32971	38135
1	1	0	442	166	8263	3705	957
0	1	1	118	12	8110	3090	899

model, model selection began after these first 1000 draws.

Results of the variable selection procedure are presented in Table 6 and 7. Model selection is extremely robust to the prior choice and clearly picks a local level model. The drift disappears because $\delta = 0$ and $\gamma_2 = 0$ imply that $a_t \equiv a_0 = 0$ for all t. This finding is confirms model selection by the marginal likelihoods in Frühwirth-Schnatter and Wagner (2008).

Table 7: Purse snatching; marginal posterior probability of selecting each indicator

prior	δ	γ_1	γ_2
$p(\mu_0) \propto 1, B_0 = 1$	0.0112	**1.0000**	0.0031
$p(\mu_0) \propto 1, B_0 = 100$	0.0042	**1.0000**	0.0003
$b = 10^{-3}$	0.2698	**1.0000**	0.2737
$b = 10^{-4}$	0.0985	**1.0000**	0.0831
$b = 10^{-5}$	0.0227	**1.0000**	0.0242

5 Concluding remarks

The model space MCMC approach discussed in this paper could be easily adapted to other state space models. Auxiliary mixture sampling as discussed in Frühwirth-Schnatter and Frühwirth (2007), for instance, allows to consider state space modelling of binary and categorical time series. Another important extension is searching for fixed and time-varying coefficients in a regression model.

A couple of modifications of our approach are worth being considered. First, the uniform prior over all models may be substituted by a more flexible prior which is obtained by assuming that the prior occurrence of $\delta_i = 1$ and $\gamma_i = 1$ is different:

$$\Pr(\delta_i = 1|\alpha_\delta) = \alpha_\delta, \quad \Pr(\gamma_i = 1|\alpha_\gamma) = \alpha_\gamma.$$

In this prior, α_δ and α_γ may be chosen as fixed values, if prior information on the occurrence probabilities is available. If this is not the case, a hyperprior may be put on α_δ and α_γ as in Smith and Kohn (2002) and Frühwirth-Schnatter and Tüchler (2008). If both hyperparameters α_δ and α_γ are iid Uniform on [0,1], then

$$p(\boldsymbol{\delta}, \boldsymbol{\gamma}) = B(1 + \sum_i I_{\{\delta_i=1\}}, 1 + \sum_i I_{\{\delta_i=0\}})B(1 + \sum_i I_{\{\gamma_i=1\}}, 1 + \sum_i I_{\{\gamma_i=0\}}),$$

where $B(\cdot, \cdot)$ is the Beta function. This prior leads to a uniform distribution over model size and outperforms the uniform prior over all models in variable selection for large regression models, see Ley and Steel (2007). In our applications, where model size is small, posterior inference under both priors is virtually the same.

Second, sampling the indicators could be modified. In our MCMC schemes, the indicators $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ are sampled jointly from the discrete posterior $p(\boldsymbol{\delta}, \boldsymbol{\gamma}|\mathbf{x}, \mathbf{y})$ by evaluating the right hand side of (25) for all combinations of indicators at each sweep of the sampler. This multi-move sampling is rather time-consuming and may be substituted single-move sampling, i.e. sampling recursively from $p(\delta_j|\boldsymbol{\delta}_{-j}, \boldsymbol{\gamma}, \mathbf{x}, \mathbf{y})$ and $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\delta}, \mathbf{x}, \mathbf{y})$.

An open issue of our approach is the influence the prior on the initial values and the process variances exercises on final model selection. We demonstrated that the normal prior put on the signed square root of the process variances is far less influential than the usual inverted Gamma for the process variances themselves. The sensitivity analysis carried out for all of our case studies revealed a surprising robustness of the finally selected model against variation in the normal prior. A concise statement which prior scale leads to model consistency in the sense of Casella, Girón, Martínez, and Moreno (2006), however, is far beyond the scope of the present paper.

References

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

Casella, G., F. J. Girón, M. L. Martínez, and E. Moreno (2006). Consistency of Bayesian procedures for variable selection. Research Report.

Chen, Z. and D. Dunson (2003). Random effects selection in linear mixed models. *Biometrics 59*, 762–769.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

De Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika 82*, 339–350.

Durbin, J. and S. J. Koopman (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society, Ser. B 62*, 3–56.

Fernández, C., E. Ley, and M. F. J. Steel (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics 100*, 381–427.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis 15*, 183–202.

Frühwirth-Schnatter, S. (1995). Bayesian model discrimination and Bayes factors for linear Gaussian state space models. *Journal of the Royal Statistical Society, Ser. B 57*, 237–246.

Frühwirth-Schnatter, S. (2001). Fully Bayesian analysis of switching Gaussian state space models. *Annals of the Institute of Statistical Mathematics 53*, 31–49.

Frühwirth-Schnatter, S. and R. Frühwirth (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics and Data Analysis 51*, 3509–3528.

Frühwirth-Schnatter, S., R. Frühwirth, L. Held, and H. Rue (2007). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. IFAS Research Report 2007-25, `http://ifas.jku.at`, Johannes Kepler University Linz.

Frühwirth-Schnatter, S. and R. Tüchler (2008). Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing*, 1–13.

Frühwirth-Schnatter, S. and H. Wagner (2006). Auxiliary mixture sampling for parameter-driven models of time series of small counts with applications to state space modelling. *Biometrika 93*, 827–841.

Frühwirth-Schnatter, S. and H. Wagner (2008). Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling. *Computational Statistics and Data Analysis*, accepted for publication.

George, E. I. and R. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*, 881–889.

George, E. I. and R. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*, 339–373.

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

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A. C. and J. Durbin (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society, Ser. A 149*, 187–227.

Holmes, C. C. and L. Held (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis 1*, 145–168.

Jeffreys, S. H. (1948). *Theory of Probability* (2 ed.). Oxford: Clarendon.

Ley, E. and M. F. J. Steel (2007). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. Technical Report CRiSM Working Paper 07-08, University of Warwick.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (Disc: p118-138). *Journal of the Royal Statistical Society, Ser. B 57*, 99–118.

Shively, T. S. and R. Kohn (1997). A Bayesian approach to model selection in stochastic coefficient regression models and structural time series models. *Journal of Econometrics 76*, 39–52.

Smith, M. and R. Kohn (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association 97*, 1141–1153.

Tüchler, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling. *Journal of Computational and Graphical Statistics 17*, 76–94.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.