Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2007-24

Marginal Likelihoods for Non-Gaussian Models Using Auxiliary Mixture Sampling

Sylvia Frühwirth-Schnatter and Helga Wagner

June 2007

Abstract

In this paper we consider several new estimators of the marginal likelihood for complex non-Gaussian models which make use of the output of auxiliary mixture sampling as developed in Frühwirth-Schnatter and Wagner (2006) for count data and in Frühwirth-Schnatter and Frühwirth (2007) for binary and multinomial data. One of these estimators is based on combining Chib's estimator (Chib, 1995) with data augmentation as in auxiliary mixture sampling, while the other estimators are importance sampling and bridge sampling based on constructing an unsupervised importance density from the output of auxiliary mixture sampling. These estimators are applied to a logit regression model, to a Poisson regression model, to a binomial model with random intercept, as well as to state space modeling of count data.

Keywords: auxiliary mixture sampling, Bayesian model selection, bridge sampling, importance sampling, Markov chain Monte Carlo

1 Introduction

For an applied statistician, choosing an appropriate model from a class of candidate models is a fundamental data analytical task. In a classical framework model selection problems are addressed either via hypothesis testing for nested models, or by using information criteria such as the Akaike information criterion (Akaike, 1974) or Schwarz's criterion or BIC (Schwarz, 1978).

In a Bayesian setting model selection relies on the posterior probabilities of a model given the data, see Bernardo and Smith (1994) as well as the recent rewiews by Godsill (2001), Green (2003) and Kadane and Lazar (2004). More formally, suppose there are K different models $\mathcal{M}_1, \ldots, \mathcal{M}_K$, which are candidates for having generated the data $\mathbf{y}$. Each of these models is assigned a prior probability $p(\mathcal{M}_k)$ and the goal is to derive the posterior model probabilities $p(\mathcal{M}_k|\mathbf{y})$ for each model $\mathcal{M}_k, k = 1, \ldots, K$.

There are basically two strategies to implement Bayesian model selection. Model space MCMC methods directly sample from the discrete models space $(\mathcal{M}_1, \ldots, \mathcal{M}_K)$ by drawing 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). A more classical strategy which dates back to Jeffreys (1948) and Zellner (1971) determines the posterior model probabilities $p(\mathcal{M}_k|\mathbf{y})$ 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 $\mathcal{M}_k$. The evaluation of the marginal likelihood typically requires computation of high-dimensional integrals. Let $\boldsymbol{\vartheta}_k$ denote the parameter of model $\mathcal{M}_k$ then the marginal likelihood is given as

$$p(\mathbf{y}|\mathcal{M}_k) = \int_{\Theta_k} p(\mathbf{y}|\boldsymbol{\vartheta}_k)p(\boldsymbol{\vartheta}_k)d\boldsymbol{\vartheta}_k, \tag{1}$$

with $p(\boldsymbol{\vartheta}_k)$ being the prior distribution of model parameter $\boldsymbol{\vartheta}_k$.

An analytical solution to (1) exists only for conjugate problems like linear regression models with normally distributed errors. For more complex, in particular for non-Gaussian models, practical Bayesian model choice requires the use of a numerical techniques to evaluate the marginal likelihood $p(\mathbf{y}|\mathcal{M}_k)$.

In general, the computation of the marginal likelihood for complex statistical models is a nontrivial integration problem. Marginal likelihoods have been estimated using methods such as standard importance sampling (Zellner and Rossi, 1984), importance sampling-based on mixture approximations (Frühwirth-Schnatter, 1995, 2004), Chib's estimator (Chib, 1995; Chib and Jeliazkov, 2001), combining MCMC simulations and asymptotic approximation (DiCiccio, Kass, Raftery, and Wasserman, 1997), and bridge sampling (Meng and Wong, 1996; Frühwirth-Schnatter, 2004). Han and Carlin (2001) provide a comparative review of MCMC methods for computing Bayes factors and marginal likelihoods for model selection.

In this paper we consider several new estimators of the marginal likelihood for complex non-Gaussian models which make use of the output of auxiliary mixture sampling. Auxiliary mixture sampling is a simple MCMC method for estimating a broad class of non-Gaussian models, developed in Frühwirth-Schnatter and Wagner (2006) for count data and in Frühwirth-Schnatter and Frühwirth (2007) for binary and multinomial data. One of these estimators is based on combining Chib's estimator (Chib, 1995) with data augmentation as in auxiliary mixture sampling, while the other estimators are importance sampling and bridge sampling based on constructing an unsupervised importance density as in Frühwirth-Schnatter (1995, 2004). These estimators are applied to logit regression models, Poisson regression models, binomial models with random intercept, as well as to state space modeling of count data.

2 Non-Gaussian Fixed Parameter Models

An important non-Gaussian fixed parameter model is the generalized linear model, see e.g. Dey, Ghosh, and Mallick (2000). An analytical expression for the marginal likelihood (1) exists only for very restricted non-Gaussian models, like observing iid observations from a Poisson, binomial or multinomial distribution. For more interesting models the posterior distribution $p(\boldsymbol{\vartheta}_k|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\vartheta}_k)p(\boldsymbol{\vartheta}_k)$ does not belong to a well-known distribution family, making analytical integration in (1) unfeasible.

2.1 Estimating the Marginal Likelihood

2.1.1 Chib's Estimator

In Chib (1995) the marginal likelihood is written as

$$p(\mathbf{y}|\mathcal{M}_k) = \frac{p(\mathbf{y}|\boldsymbol{\vartheta}_k)p(\boldsymbol{\vartheta}_k)}{p(\boldsymbol{\vartheta}_k|\mathbf{y})}, \tag{2}$$

which is known as the basic marginal likelihood equation. It is evaluated at a high density point $\boldsymbol{\vartheta}_k^*$ to obtain an estimator of $p(\mathbf{y}|\mathcal{M}_k)$:

$$\hat{p}_{CH}(\mathbf{y}|\mathcal{M}_k) = \frac{p(\mathbf{y}|\boldsymbol{\vartheta}_k^*)p(\boldsymbol{\vartheta}_k^*)}{\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})}. \tag{3}$$

As for many non-Gaussian fixed parameter models both the prior ordinate $p(\boldsymbol{\vartheta}_k)$ and the likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k)$ are easy to calculate, the application of Chib's estimator requires only an estimate $\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})$ of the posterior ordinate at the point $\boldsymbol{\vartheta}_k^*$. Chib (1995) demonstrates how the posterior ordinate $\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})$ may be estimated easily, if the model is estimated by a Markov chain Monte Carlo (MCMC) method which combines data augmentation and Gibbs sampling. For non-Gaussian models such a simple MCMC scheme has been available only for probit regression models whereas for other non-Gaussian models the Metropolis-Hastings algorithm is used and the posterior ordinate is estimated as in Chib and Jeliazkov (2001).

Only recently, auxiliary mixture sampling became available which provides a simple MCMC scheme based on data augmentation and Gibbs sampling for a much broader class of complex non-Gaussian models, like models for count data (Frühwirth-Schnatter and Wagner, 2006) and models for binary and multinomial data (Frühwirth-Schnatter and Frühwirth, 2007). The idea of auxiliary mixture sampling is to introduce a set of auxiliary variables $\mathbf{z}_k$, which allow the implementation of a two step Gibbs sampler for fixed parameter models:

(a) Sample $\boldsymbol{\vartheta}_k$ from $p(\boldsymbol{\vartheta}_k|\mathbf{z}_k, \mathbf{y})$;

(b) sample $\mathbf{z}_k$ from $p(\mathbf{z}_k|\boldsymbol{\vartheta}_k, \mathbf{y})$.

In this scheme, the augmented density $p(\boldsymbol{\vartheta}_k|\mathbf{z}_k, \mathbf{y})$ is of the same closed form as it would be for a model based on the normal distribution. On the other hand, the density $p(\mathbf{z}_k|\boldsymbol{\vartheta}_k, \mathbf{y})$ takes a simple form which is easy to sample from.

The MCMC output of auxiliary mixture sampling is used to estimate the posterior ordinate $\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})$ in Chib's estimator (3). If $\mathbf{z}_k^{(m)}, m = 1, \ldots, M$ denotes the draws from auxiliary mixture sampling, an appropriate estimator is

$$\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{\vartheta}_k^*|\mathbf{z}_k^{(m)}, \mathbf{y}),$$

where $p(\boldsymbol{\vartheta}_k^*|\mathbf{z}_k^{(m)}, \mathbf{y})$ is the closed form conditional density appearing in step (a) of auxiliary mixture sampling and therefore easily evaluated.

The resulting combination of Chib's estimator with auxiliary mixture sampling provides a new estimator of the marginal likelihood $p(\mathbf{y}|\mathcal{M}_k)$ for a non-Gaussian fixed parameter model which is much easier to implement than the Metropolis-Hastings based estimator of Chib and Jeliazkov (2001).

2.1.2 Importance Sampling

An alternative approximation of the marginal likelihood is obtained by applying importance sampling to the integral (1):

$$\hat{p}_{IS}(\mathbf{y}|\mathcal{M}_k) = \frac{1}{L}\sum_{l=1}^{L} \frac{p(\mathbf{y}|\boldsymbol{\vartheta}_k^{(l)})p(\boldsymbol{\vartheta}_k^{(l)})}{q(\boldsymbol{\vartheta}_k^{(l)})}, \tag{4}$$

where $\boldsymbol{\vartheta}_k^{(l)}, l = 1, \ldots, L$ denote iid draws from an importance density $q(\boldsymbol{\vartheta}_k)$.

Importance sampling may be combined with auxiliary mixture sampling in the following way. The MCMC scheme underlying auxiliary mixture sampling provides a useful way to construct the importance density in an unsupervised manner as in Frühwirth-Schnatter (1995, 2004), by considering following mixture density constructed from a subsequence of the posterior draws:

$$q(\boldsymbol{\vartheta}_k) = \frac{1}{S} \sum_{m'=1}^{S} p(\boldsymbol{\vartheta}_k | \mathbf{z}_k^{(m')}, \mathbf{y}). \tag{5}$$

Again, $p(\boldsymbol{\vartheta}_k | \mathbf{z}_k^{(m')}, \mathbf{y})$ is the closed form conditional density appearing in step (a) of auxiliary mixture sampling and therefore easily evaluated. This mixture importance density automatically has high mass in regions of high posterior probability, because $q(\boldsymbol{\vartheta}_k)$ converges to the posterior density $p(\boldsymbol{\vartheta}_k|\mathbf{y})$ as $S \rightarrow \infty$. On the other hand, the evaluation of $q(\boldsymbol{\vartheta}_k)$ requires S density evaluations for each draw $\boldsymbol{\vartheta}_k^{(l)}$ and is more expensive than standard importance sampling where the importance density $q(\boldsymbol{\vartheta}_k)$ is obtained from fitting an appropriate density like a Gaussian or a t-density to the posteriors draws $\boldsymbol{\vartheta}_k^{(1)}, \ldots, \boldsymbol{\vartheta}_k^{(M)}$ obtained from auxiliary mixture sampling.

2.2 Application to Binary Logit Regression

2.2.1 Data and Modelling

For illustration we reanalyze the prostatic nodal involvement data considered in Chib (1995), see also Collett (1991). The data were collected on $N = 53$ patients with cancer of the prostate. The binary response variable $y_i, i = 1, \ldots, N$ is coded 1, if the cancer has spread to the surrounding lymph nodes and 0 otherwise. Explanatory variables are the age of the patient at time of diagnosis (x_1), level of serum acid phosphate (x_2), the outcome of an X-ray examination (x_3), taking the value 1 if positive and 0 otherwise, the size of the tumor (x_4), coded 0 if small and 1 if large; and the pathological grade of the tumor (x_5), coded 0 if less serious and 1 if more serious.

To demonstrate the computation of the marginal likelihood, we consider the same models as in Chib (1995), however, substitute the probit by the logit link used originally by Collett (1991):

$$\Pr(y_i = 1 | \boldsymbol{\alpha}_k, \mathcal{M}_k) = \frac{\exp(\mathbf{x}_{ik} \boldsymbol{\alpha}_k)}{1 + \exp(\mathbf{x}_{ik} \boldsymbol{\alpha}_k)}, \tag{6}$$

where $\boldsymbol{\alpha}_k$ is an unknown regression parameter and $\mathbf{x}_{ik}$ is a row vector containing the regressors relevant for model $\mathcal{M}_k$, including 1 for the intercept.

To pursue a Bayesian approach, we assume that apriori $\boldsymbol{\alpha}_k$ follows a normal distribution, $\mathcal{N}(\mathbf{a}_0, \mathbf{A}_0)$. Since the parameters in a probit model should be multiplied by 1.6 to compare with them with the parameters in a logit model, we transform the prior used in Chib (1995) to the logit scale, i.e. the mean of each parameter is equal to 1.2 and the standard deviation is equal to 8.

2.2.2 Auxiliary Mixture Sampling

Following Frühwirth-Schnatter and Frühwirth (2007) estimation of model (6) using auxiliary mixture sampling is obtained in the following way[1]. Define for each $i = 1, \dots, N$, a latent utility u_i such that

$$\begin{aligned} u_i &= \mathbf{x}_i \boldsymbol{\alpha} + \varepsilon_i, \\ y_i &= 1, \text{ iff } u_i > \xi_{i0}, \end{aligned} \tag{7}$$

where $\exp(-\varepsilon_i)$ and $\exp(-\xi_{i0})$ are standard exponential random variables. According to McFadden (1974), model (7) is equivalent to the logit model (6), see also Scott (2005). The distribution of ε_i is then approximated by a mixture of normal distributions,

$$p_\varepsilon(\varepsilon_i) = \exp\{-\varepsilon_i - e^{-\varepsilon_i}\} \approx \sum_{r_i=1}^{10} w_{r_i} f_N(\varepsilon_i; m_{r_i}, s_{r_i}^2). \tag{8}$$

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

Auxiliary mixture sampling is based on data augmentation by introducing the auxiliary variables $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$, where $\mathbf{z}_i = (u_i, r_i)$. It is defined through following complete conditional densities

(a) Sample $\boldsymbol{\alpha}$ from $p(\boldsymbol{\alpha}|\mathbf{z}, \mathbf{y}) \sim \mathcal{N}(\mathbf{a}_N, \mathbf{A}_N)$-distribution, where

$$\mathbf{a}_N = \mathbf{A}_N \left(\sum_{i=1}^N \mathbf{x}_i'(u_i - m_{r_i})/s_{r_i}^2 + \mathbf{A}_0^{-1}\mathbf{a}_0 \right), \qquad \mathbf{A}_N^{-1} = \mathbf{A}_0^{-1} + \sum_{i=1}^N \mathbf{x}_i'\mathbf{x}_i/s_{r_i}^2. \tag{9}$$

(b) Sample the latent utility u_i conditional on $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\alpha})$ and y_i as

$$u_i = -\log \left(-\frac{\log(U_i)}{1+\lambda_i} - \frac{\log(V_i)}{\lambda_i} I_{\{y_i=0\}} \right), \tag{10}$$

where U_i and V_i are two independent uniform random numbers. Sample the component indicator r_i conditional on u_i and λ_i from the following discrete density:

$$\Pr(r_i = j|u_i, \boldsymbol{\alpha}) \propto \frac{w_j}{s_j} \exp\left\{ -\frac{1}{2} \left(\frac{u_i - \log \lambda_i - m_j}{s_j} \right)^2 \right\}. \tag{11}$$

2.2.3 Marginal Likelihoods

We select the same combination of covariates as Chib (1995) yielding nine different models, see also Table 1. For each model $\mathcal{M}_k$, auxiliary mixture sampling was run for $M = 20000$ draws after a burn-in of 5000 draws, yielding a sequence of

[1] The model index k is suppressed to simplify notation

Table 1: Marginal likelihoods for nodal involvement data based on the logit model in comparison to the probit model (Chib, 1995)

k	Model $\mathcal{M}_k$	$\log \hat{p}_{CH}(\mathbf{y}\|\mathcal{M}_k)$	$\log \hat{p}_{IS,1}(\mathbf{y}\|\mathcal{M}_k)$	$\log \hat{p}_{IS,2}(\mathbf{y}\|\mathcal{M}_k)$	probit
1	c	-37.60(.032)	-37.61(.0003)	-37.61(.0003)	-38.503(.005)
		-37.62(.034)	-37.61(.0008)	-37.61(.0008)	
		-37.57(.033)	-37.61(.0005)	-37.61(.0005)	
2	c, x_1	-41.412(.048)	-41.443(.0005)	-41.445(.0005)	-43.175(.007)
		-41.499(.048)	-41.443(.0006)	-41.443(.0006)	
		-41.444(.047)	-41.445(.0009)	-41.444(.0008)	
3	$c, \log(x_2)$	-35.806(.064)	-35.829(.0013)	-35.831(.0013)	-37.916(.007)
		-35.809(.062)	-35.830(.0008)	-35.829(.0008)	
		-35.807(.067)	-35.827(.0013)	-35.826(.0015)	
4	c, x_3	-33.459(.068)	-33.512(.0013)	-33.509(.0014)	-35.323(.009)
		-33.664(.062)	-33.508(.0037)	-33.510(.0013)	
		-33.703(.060)	-33.510(.0021)	-33.512(.0013)	
5	c, x_4	-35.435(.061)	-35.476(.0012)	-35.473(.0012)	-37.234(.009)
		-35.615(.063)	-35.474(.0010)	-35.473(.0011)	
		-35.487(.064)	-35.473(.0010)	-35.472(.0010)	
6	c, x_5	-37.212(.060)	-37.229(.0012)	-37.227(.0012)	-39.075(.007)
		-37.236(.059)	-37.230(.0008)	-37.229(.0009)	
		-37.164(.061)	-37.228(.0011)	-37.229(.0011)	
7	$c, \log(x_2), x_4$	-33.210(.105)	-33.224(.0026)	-33.222(.0030)	-36.140(.013)
		-33.225(.117)	-33.227(.0015)	-33.227(.0013)	
		-33.211(.121)	-33.228(.0022)	-33.228(.0019)	
8	$c, \log(x_2), x_3, x_4$	**-30.946**(.190)	**-30.720**(.0045)	**-30.719**(.0059)	**-34.553**(.020)
		-30.622(.213)	**-30.730**(.0023)	**-30.730**(.0023)	
		-30.687(.207)	**-30.726**(.0022)	**-30.729**(.0019)	
9	$c, \log(x_2), x_3, x_4, x_5$	-31.316(.361)	-31.440(.0022)	-31.438(.0022)	-36.233(.024)
		-31.623(.372)	-31.438(.0026)	-31.432(.0036)	
		-32.073(.497)	-31.437(.0022)	-31.435(.0028)	

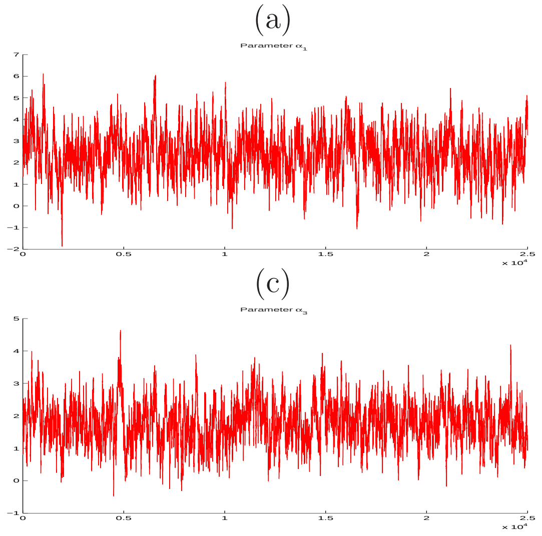

Figure 1: Posterior draws (including burn-in) obtained for the various components of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4)$ for model $\mathcal{M}_8$

draws $\boldsymbol{\alpha}_k^{(1)}, \ldots, \boldsymbol{\alpha}_k^{(M)}$ for the unknown regression parameters. For illustration, Figure 1 shows for model $\mathcal{M}_8$ that the sampler is converging quickly to the stationary distribution and mixing is pretty good.

To implement Chib's estimator $\hat{p}_{CH}(\mathbf{y}|\mathcal{M}_k)$, $\boldsymbol{\alpha}_k^*$ is chosen as the average of all draws $\boldsymbol{\alpha}_k^{(1)}, \ldots, \boldsymbol{\alpha}_k^{(M)}$. The estimate of $p(\boldsymbol{\alpha}_k^*|\mathbf{y})$ is

$$\hat{p}(\boldsymbol{\alpha}_k^*|\mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} f_N(\boldsymbol{\alpha}_k^*; \mathbf{a}_N^{(m)}, \mathbf{A}_N^{(m)}),$$

where $\mathbf{a}_N^{(m)}$ and $\mathbf{A}_N^{(m)}$ are the moments of the conditional normal distribution given in (9).

To implement importance sampling, two different importance functions $q(\boldsymbol{\vartheta}_k)$ were considered. First, a multivariate normal distribution was fitted to the MCMC draws of $\boldsymbol{\alpha}_k^{(m)}$. Second, an importance density based on the mixture approximation (5) was chosen with $S = 100$ components yielding the estimators $\hat{p}_{IS,1}(\mathbf{y}|\mathcal{M}_k)$ and $\hat{p}_{IS,2}(\mathbf{y}|\mathcal{M}_k)$. Importance sampling is based on 20000 draws from the importance density.

Table 1 compares the different estimators of the marginal likelihood for the various logit models. For each estimator, numerical standard errors where computed as in Chib (1995). Additionally, estimation was carried out for three independent runs to check stability. For both importance densities, importance sampling is very accurate and more precise than Chib's estimator, which tends to have quite large standard errors for the larger models. Among all logit models considered, the same combination of explanatory variables is chosen as in Chib (1995), namely $\log(x_2)$, x_3 and x_4. A comparison with the marginal likelihood obtained for the probit model in Chib (1995) yields that the Bayes factor favors the logit link over the probit link quite clearly for each model.

3 Non-Gaussian Latent Variable Models

Latent variable models provide a very flexible way of modelling complex data and encompass important examples as random effects models (Verbeke and Molenberghs, 2000) and state space models (Durbin and Koopman, 2001). Computing marginal likelihoods for latent variable models is a challenging task as it involves integration over the latent variable $\boldsymbol{\beta}_k$ and the model parameters $\boldsymbol{\vartheta}_k$, i.e. a high-dimensional integration:

$$p(\mathbf{y}|\mathcal{M}_k) = \int p(\mathbf{y}|\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_k, \mathcal{M}_k) p(\boldsymbol{\beta}_k|\boldsymbol{\vartheta}_k, \mathcal{M}_k) p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) d(\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_k). \tag{12}$$

3.1 Auxiliary Mixture Sampling

Recently, auxiliary mixture sampling has been developed for MCMC estimation of non-Gaussian models involving latent variables in Frühwirth-Schnatter and Wagner (2006) for count data and in Frühwirth-Schnatter and Frühwirth (2007) for binary and multinomial data. A set of auxiliary variables $\mathbf{z}_k$ is introduced which allows the implementation of a three step Gibbs sampler for many latent variable models:

(a) sample $\boldsymbol{\beta}_k$ from $p(\boldsymbol{\beta}_k|\boldsymbol{\vartheta}_k, \mathbf{z}_k, \mathbf{y})$;

(b) sample $\boldsymbol{\vartheta}_k$ from $p(\boldsymbol{\vartheta}_k|\boldsymbol{\beta}_k, \mathbf{z}_k, \mathbf{y})$;

(c) sample $\mathbf{z}_k$ from $p(\mathbf{z}_k|\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_k, \mathbf{y})$;

where the augmented densities $p(\boldsymbol{\beta}_k|\boldsymbol{\vartheta}_k, \mathbf{z}_k, \mathbf{y})$ and $p(\boldsymbol{\vartheta}_k|\boldsymbol{\beta}_k, \mathbf{z}_k, \mathbf{y})$ are of the same closed form as they would be for a model based on a normal rather than on a non-Gaussian distribution. Again, the density $p(\mathbf{z}_k|\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_k, \mathbf{y})$ takes a very simple form. As in Section 2 the output from auxiliary mixture sampling is used to estimate the marginal likelihood.

3.2 Estimating the Marginal Likelihood

3.2.1 Chib's Estimator

As for fixed parameter models, Chib's estimator is based on identity (2):

$$\hat{p}_{CH}(\mathbf{y}|\mathcal{M}_k) = \frac{p(\mathbf{y}|\boldsymbol{\vartheta}_k^*) p(\boldsymbol{\vartheta}_k^*)}{\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})}. \tag{13}$$

In the context of latent variable models, however, (13) is called the integrated marginal likelihood equation because $p(\mathbf{y}|\boldsymbol{\vartheta}_k)$ is the likelihood function where the latent variables $\boldsymbol{\beta}_k$ are integrated out:

$$p(\mathbf{y}|\boldsymbol{\vartheta}_k) = \int p(\mathbf{y}|\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_k, \mathcal{M}_k) p(\boldsymbol{\beta}_k|\boldsymbol{\vartheta}_k, \mathcal{M}_k) d\boldsymbol{\beta}_k. \tag{14}$$

If a numerical value for $p(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$ is available, then (13) may be implemented as in Subsection 2.1, by estimating $\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})$ by means of the draws obtained from auxiliary mixture sampling:

$$\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{\vartheta}_k^*|\boldsymbol{\beta}_k^{(m)}, \mathbf{z}_k^{(m)}, \mathbf{y}), \tag{15}$$

where $p(\boldsymbol{\vartheta}_k^*|\boldsymbol{\beta}_k^{(m)}, \mathbf{z}_k^{(m)}, \mathbf{y})$ is the conditional density appearing in step (b) of auxiliary mixture sampling.

In general, the latent variables $\boldsymbol{\beta}_k$ can analytically be integrated out in (14) only for models which are based on the normal distribution. For non Gaussian latent variable models, however, analytical integration is usually not feasible and some numerical technique is required to determine the likelihood value $p(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$. For models where the latent variables are conditionally independent given $\boldsymbol{\vartheta}_k$, like random effect models, techniques like GH-integration (Frühwirth-Schnatter, 1997) or importance sampling may be applied, see Subsection 3.3. For state space models an approximation to $p(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$ is obtained by particle filtering as in Chib, Nardari, and Shephard (2002) who face the same problem in the context of stochastic volatility models, see Subsection 3.4.

3.2.2 The Complete-data Likelihood Estimator

To avoid the use of particle filtering or other sampling techniques to approximate the integrated likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$, one might use a representation of the marginal likelihood which is based on the complete data likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k, \boldsymbol{\beta}_k)$:

$$\hat{p}_{CDL}(\mathbf{y}|\mathcal{M}_k) = \frac{p(\mathbf{y}|\boldsymbol{\vartheta}_k^*, \boldsymbol{\beta}_k^*)\, p(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*)p(\boldsymbol{\vartheta}_k^*)}{\hat{p}(\boldsymbol{\vartheta}_k^*, \boldsymbol{\beta}_k^*|\mathbf{y})}. \tag{16}$$

In contrast to the integrated likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$ appearing in (13), the complete-data likelihood value $p(\mathbf{y}|\boldsymbol{\vartheta}_k^*, \boldsymbol{\beta}_k^*)$ can be easily computed from the observation equation of the latent variable model. In the context of latent variable models, (16) is called the complete-data marginal likelihood equation.

The complete-data likelihood estimator $\hat{p}_{CDL}(\mathbf{y}|\mathcal{M}_k)$ is easily implemented using auxiliary mixture sampling. By writing the joint posterior as $p(\boldsymbol{\vartheta}_k^*, \boldsymbol{\beta}_k^*|\mathbf{y}) = p(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{y})p(\boldsymbol{\vartheta}_k^*|\mathbf{y})$, (16) may be written as:

$$\hat{p}_{CDL}(\mathbf{y}|\mathcal{M}_k) = \left(\frac{p(\mathbf{y}|\boldsymbol{\vartheta}_k^*, \boldsymbol{\beta}_k^*)\, p(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*)}{\hat{p}(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{y})}\right) \frac{p(\boldsymbol{\vartheta}_k^*)}{\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})}. \tag{17}$$

The marginal posterior ordinate $\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})$ is estimated from the output of auxiliary mixture sampling as in Subsection 3.2.1. To approximate the conditional posterior ordinate $\hat{p}(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{y})$ of the latent variables $\boldsymbol{\beta}_k^*$ given $\boldsymbol{\vartheta}_k^*$, reduced auxiliary mixture sampling is performed with holding $\boldsymbol{\vartheta}_k$ fixed at $\boldsymbol{\vartheta}_k^*$:

(a') Sample $\boldsymbol{\beta}_k$ from $p(\boldsymbol{\beta}_k|\boldsymbol{\vartheta}_k^*, \mathbf{z}_k, \mathbf{y})$;

(c') sample $\mathbf{z}_k$ from $p(\mathbf{z}_k|\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_k^*, \mathbf{y})$.

Reduced auxiliary mixture sampling leads to the following approximation of $\hat{p}(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{y})$:

$$\hat{p}(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{z}_k^{(m)}, \mathbf{y}). \tag{18}$$

While implementation of the complete-data likelihood estimator is straightforward, the case studies in Subsection 3.3 and 3.4 demonstrate that this estimator is extremely inaccurate. By comparing (17) with (13), we find that the complete-data likelihood estimator as that special case of Chib's estimator (13) where the integrated likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$ is represented by another marginal likelihood equation, namely

$$\hat{p}(\mathbf{y}|\boldsymbol{\vartheta}_k^*) = \frac{p(\mathbf{y}|\boldsymbol{\vartheta}_k^*, \boldsymbol{\beta}_k)\, p(\boldsymbol{\beta}_k|\boldsymbol{\vartheta}_k^*)}{\hat{p}(\boldsymbol{\beta}_k|\boldsymbol{\vartheta}_k^*, \mathbf{y})},$$

which is evaluated at the point $\boldsymbol{\beta}_k^*$. It turns out that this estimator of the integrated likelihood $\hat{p}(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$ is extremely unstable, causing high standard errors of the complete-data likelihood estimator of the marginal likelihood.

3.2.3 Blocked Estimators

Both estimators can be extended to the case, where in step (b) of auxiliary mixture sampling more than one block is needed to sample the model parameter. Assume that in step (b) $\boldsymbol{\vartheta}$ is divided in G blocks, $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_G)$, and in each block Gibbs sampling of $\boldsymbol{\vartheta}_g$ conditional on the remaining parameters is possible.[2] In this case the posterior ordinate $p(\boldsymbol{\vartheta}^*|\mathbf{y})$ in (13) or (17) can be decomposed as in Chib (1995):

$$p(\boldsymbol{\vartheta}^*|\mathbf{y}) = p(\boldsymbol{\vartheta}_1^*|\mathbf{y}) p(\boldsymbol{\vartheta}_2^*|\mathbf{y}, \boldsymbol{\vartheta}_1^*) \ldots p(\boldsymbol{\vartheta}_G^*|\mathbf{y}, \boldsymbol{\vartheta}_1^*, \ldots, \boldsymbol{\vartheta}_{G-1}^*). \qquad (19)$$

Each term $p(\boldsymbol{\vartheta}_g^*|\mathbf{y}, \boldsymbol{\vartheta}_1^*, \ldots, \boldsymbol{\vartheta}_{g-1}^*)$ can estimated from the output of a separate reduced auxiliary mixture sampler, where the value of $\boldsymbol{\vartheta}_i, i < g$ is set to $\boldsymbol{\vartheta}_i^*$. If $(\boldsymbol{\vartheta}_{g+1}^{(m)}, \ldots, \boldsymbol{\vartheta}_G^{(m)}, \mathbf{z}^{(m)})$ denotes the draws from reduced auxiliary mixture sampling, the estimate is

$$\hat{p}(\boldsymbol{\vartheta}_g^*|\mathbf{y}, \boldsymbol{\vartheta}_1^*, \ldots, \boldsymbol{\vartheta}_{g-1}^*) = \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{\vartheta}_g^*|\mathbf{y}, \boldsymbol{\vartheta}_1^*, \ldots, \boldsymbol{\vartheta}_{g-1}^*, \boldsymbol{\vartheta}_{g+1}^{(m)}, \ldots, \boldsymbol{\vartheta}_G^{(m)}, \mathbf{z}^{(m)}).$$

3.2.4 Importance Sampling and Bridge Sampling

A sampling based approximation to the marginal likelihood is obtained by applying importance sampling to the integral (12). This method requires the choice of an importance density $q(\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_k)$ which turns out to be quite a challenge. One way to fit such a density is to separate the latent variables $\boldsymbol{\beta}_k$ from the model parameters $\boldsymbol{\vartheta}_k$, by selecting $q(\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_k) = q(\boldsymbol{\beta}_k) q(\boldsymbol{\vartheta}_k)$. The importance density $q(\boldsymbol{\beta}_k)$ for the latent variables $\boldsymbol{\beta}_k$ is obtained by fitting an appropriate density to the MCMC draws $\boldsymbol{\beta}_k^{(1)}, \ldots, \boldsymbol{\beta}_k^{(M)}$ obtained by auxiliary mixture sampling. Similarly, an appropriate density could be fitted to the MCMC draws $\boldsymbol{\vartheta}_k^{(1)}, \ldots, \boldsymbol{\vartheta}_k^{(M)}$ of the model parameter. Alternatively, the importance density $q(\boldsymbol{\vartheta}_k)$ may be constructed in an unsupervised manner from the conditional densities appearing in step (b).

Importance sampling may be unstable if the ratio of the nonnormalized posterior density over the importance density is unbounded (Geweke, 1989). To reduce sensitivity to the choice of the importance density, bridge sampling may be implemented

[2] Again, the model index k is suppressed to simplify notation

(Meng and Wong, 1996; Frühwirth-Schnatter, 2004). Like importance sampling, bridge sampling is based on an i.i.d. sample from an importance density, however, this sample is combined with the MCMC draws from the posterior density in an appropriate way. An important advantage of bridge sampling is that the variance of the resulting estimator depends on a ratio that is bounded regardless of the tail behavior of the underlying importance density (Frühwirth-Schnatter, 2004). This allows far more flexibility in the construction of the importance density. For more details on the practical implementation of bridge sampling we refer to Frühwirth-Schnatter (2006, Section 5.4).

3.3 Application to Logit Regression with Random Effects

3.3.1 Data and Modelling

In this example, we reconsider the data given by Crowder (1978, Table 3) reporting the number Y_i of seeds that germinated among T_i seeds in $N = 21$ plates covered with a certain root extract. Covariates are the type of root extract, x_1 (bean or cucumber) and the type of seed, x_2 and the interaction term $x_1 x_2$.

The data are modelled as in Breslow and Clayton (1993) and Gamerman (1997), assuming that Y_i is generated by a binomial distribution, where the dependence of the success probability on the covariates $\mathbf{x}_i$ is modelled through a logit transform:

$$\begin{aligned} Y_i &\sim \text{BiNom}\left(T_i, \pi_i\right), \qquad (20) \\ \log \frac{\pi_i}{1-\pi_i} &= \mathbf{x}_{ik}\boldsymbol{\alpha}_k + \gamma_i, \qquad \gamma_i \sim \mathcal{N}\left(0, Q\right). \end{aligned}$$

$\boldsymbol{\alpha}_k$ is an unknown regression parameter and $\mathbf{x}_{ik}$ is a row vector containing the regressors relevant for model $\mathcal{M}_k$, including 1 for the intercept. The random intercept γ_i has been added in Breslow and Clayton (1993) to capture overdispersion. In this model, $\boldsymbol{\vartheta}_k = (\boldsymbol{\alpha}_k, Q)$ are unknown model parameters, whereas $\boldsymbol{\beta}_k = (\gamma_1, \ldots, \gamma_N)$ are the unknown latent variables.

Gamerman (1997) used a Metropolis-Hastings algorithm to estimate model (20) and used the estimated posterior densities to explore significance of the various covariates and the presence of unobserved heterogeneity. Here, we use marginal likelihoods to perform covariate selection and testing for unobserved heterogeneity. To make model comparison through marginal likelihoods feasible, the improper prior $p(\boldsymbol{\alpha}_k, Q) \propto 1/\sqrt{Q}$ used by Gamerman (1997) is substituted by the proper priors $\boldsymbol{\alpha}_k \sim \mathcal{N}\left(\mathbf{a}_0, \mathbf{A}_0\right)$ and $Q \sim \mathcal{G}^{-1}\left(c_0, C_0\right)$ where $\mathbf{a}_0 = \mathbf{0}$, $\mathbf{A}_0 = \mathbf{I}$, $c_0 = 0.5$ and $C_0 = 0.2275$.

3.3.2 Auxiliary Mixture Sampling

The binomial model (20) is estimated using auxiliary mixture sampling as in Frühwirth-Schnatter and Frühwirth (2007)[3]. Any observation Y_i from model (20) is equivalent with observing T_i repeated measurements y_{it} from a binary model with random effects,

$$\Pr(y_{it} = 1|\boldsymbol{\alpha}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\alpha} + \gamma_i)}{1 + \exp(\mathbf{x}_i\boldsymbol{\alpha} + \gamma_i)}, \qquad (21)$$

[3] Again, the model index k is suppressed to simplify notation

where

$$y_{it} = \begin{cases} 1, & 1 \leq t \leq Y_i, \\ 0, & Y_i < t \leq T_i. \end{cases}$$

Introducing for each $i = 1, \ldots, N$ and each $t = 1, \ldots, T_i$, a latent utility u_{it} such that

$$\begin{aligned} u_{it} &= \mathbf{x}_i \boldsymbol{\alpha} + \gamma_i + \varepsilon_{it}, \\ y_{it} &= 1, \text{ iff } u_{it} > \xi_{it0}, \end{aligned} \tag{22}$$

where $\exp(-\varepsilon_{it})$ and $\exp(-\xi_{it0})$ are standard exponential random variables, is equivalent to model (21) (McFadden, 1974). The distribution of ε_{it} is then approximated by a mixture of normal distributions as in (8) and the auxiliary variables $\mathbf{z} = (\mathbf{z}_{11}, \ldots, \mathbf{z}_{N,T_N})$, where $\mathbf{z}_{it} = (u_{it}, r_{it})$, are introduced.

Auxiliary mixture sampling is defined through the following conditional densities (Frühwirth-Schnatter and Frühwirth, 2007, Section 4.2):

(a) Sample γ_i for each $i = 1, \ldots, N$ from $\mathcal{N}(b_i, B_i)$ where

$$b_i = B_i \sum_{t=1}^{T_i} \frac{u_{it} - \mathbf{x}_i \boldsymbol{\alpha} - m_{r_{it}}}{s^2_{r_{it}}}, \qquad B_i = Q \left(1 + Q \sum_{t=1}^{T_i} 1/s^2_{r_{it}} \right)^{-1}. \tag{23}$$

(b1) Sample Q from $\mathcal{G}^{-1}\left(c_0 + N/2, C_0 + 1/2 \sum_{i=1}^N \gamma_i^2\right)$;

(b2) Sample $\boldsymbol{\alpha}$ from $\mathcal{N}(\mathbf{a}_N, \mathbf{A}_N)$-distribution, where

$$\begin{aligned} \mathbf{A}_N^{-1} &= \mathbf{A}_0^{-1} + \sum_{i=1}^N \mathbf{X}_i^{'} \mathbf{V}_i^{-1} \mathbf{X}_i, \quad \mathbf{a}_N = \mathbf{A}_N \left(\mathbf{A}_0^{-1} \mathbf{a}_0 + \sum_{i=1}^N \mathbf{X}_i^{'} \mathbf{V}_i^{-1} (\mathbf{u}_i - \boldsymbol{m}_i) \right), \\ \mathbf{V}_i^{-1} &= \text{Diag}(\boldsymbol{f}_i) - B_i \boldsymbol{f}_i \boldsymbol{f}_i^{'}, \end{aligned}$$

and the T_i rows of $\mathbf{X}_i$ are equal to $\mathbf{x}_i$, $\mathbf{u}_i = (u_{i1}, \ldots, u_{i,T_i})^{'}$, $\boldsymbol{m}_i = (m_{r_{i1}}, \ldots, m_{r_{i,T_i}})^{'}$, and $\boldsymbol{f}_i = (1/s^2_{r_{i1}}, \ldots, 1/s^2_{r_{i,T_i}})^{'}$.

(c) Sample the latent utility u_{it} conditional on $\lambda_i = \exp(\mathbf{x}_i \boldsymbol{\alpha} + \gamma_i)$ and y_{it} as

$$u_{it} = -\log \left(-\frac{\log(U_{it})}{1 + \lambda_i} - \frac{\log(V_{it})}{\lambda_i} I_{\{y_{it}=0\}} \right), \tag{24}$$

where U_{it} and V_{it} are two independent uniform random numbers. Sample the component indicator r_{it} conditional on u_{it} and λ_i from the following discrete density:

$$\Pr(r_{it} = j | u_{it}, \boldsymbol{\alpha}) \propto \frac{w_j}{s_j} \exp \left\{ -\frac{1}{2} \left(\frac{u_{it} - \log \lambda_i - m_j}{s_j} \right)^2 \right\}. \tag{25}$$

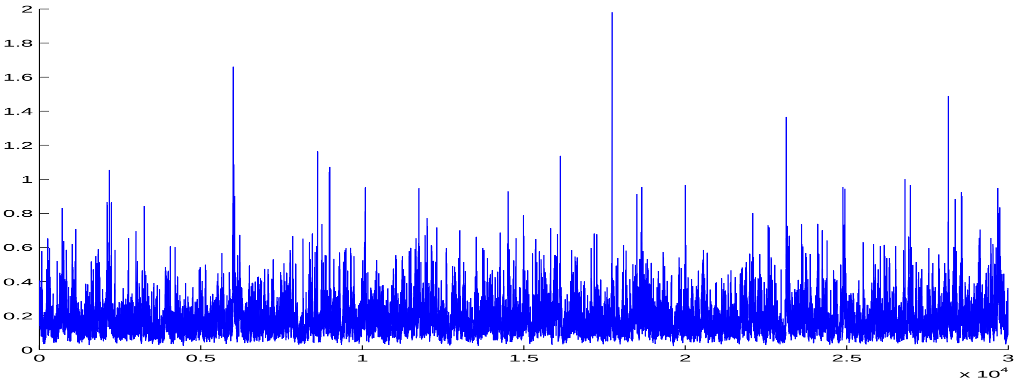

Figure 2: Posterior draws (including burn-in) obtained for Q for model $\mathcal{M}_6$

Table 2: Marginal likelihoods for the seed data

			logit including unobserved heterogeneity			
k	$\mathcal{M}_k$	logit	$\log \hat{p}_{CH}(\mathbf{y}\|\mathcal{M}_k)$	$\log \hat{p}_{CDL}(\mathbf{y}\|\mathcal{M}_k)$	$\log \hat{p}_{IS}(\mathbf{y}\|\mathcal{M}_k)$	$\log \hat{p}_{BS}(\mathbf{y}\|\mathcal{M}_k)$
1	c	-578.50	-555.77(0.016)	-530.21(0.853)	-555.85(0.041)	-555.78(0.020)
			-555.77(0.017)	-526.44(0.536)	-555.82(0.071)	-555.80(0.020)
			-555.76(0.017)	-522.37(0.705)	-555.81(0.076)	-555.74(0.021)
2	c, x_1	-553.11	-551.31(0.024)	-538.47(0.948)	-551.26(0.213)	-551.35(0.024)
			-551.31(0.023)	-538.72(0.997)	-551.47(0.060)	-551.37(0.023)
			-551.37(0.023)	-542.20(0.996)	-551.45(0.042)	-551.37(0.025)
3	c, x_2	-579.18	-556.06(0.017)	-530.58(0.889)	-556.17(0.104)	-556.11(0.030)
			-556.10(0.017)	-543.53(0.998)	-556.30(0.077)	-556.15(0.031)
			-556.09(0.015)	-531.53(0.656)	-555.89(0.337)	-556.11(0.029)
4	c, x_1x_2	-580.05	-556.72(0.018)	-539.36(0.998)	-556.79(0.104)	-556.77(0.032)
			-556.80(0.017)	-548.10(0.998)	-556.96(0.053)	-556.82(0.031)
			-556.70(0.017)	-531.37(0.985)	-556.94(0.065)	-556.76(0.031)
5	c, x_1, x_2	-553.46	-551.64(0.023)	-537.41(0.799)	-551.58(0.098)	-551.58(0.029)
			-551.57(0.023)	-557.34(0.997)	-551.66(0.067)	-551.60(0.028)
			-551.57(0.023)	-535.94(0.619)	-551.39(0.242)	-551.59(0.028)
6	c, x_1, x_1x_2	**-550.58**	**-550.40**(0.023)	-532.42(0.860)	**-550.42**(0.127)	**-550.32**(0.025)
			-550.38(0.024)	-536.22(0.574)	**-550.47**(0.069)	**-550.41**(0.026)
			-550.37(0.024)	-538.20(0.991)	**-550.50**(0.056)	**-550.37**(0.026)
7	$c, x_2,$	-578.47	-556.54(0.018)	-529.22(0.999)	-556.83(0.076)	-556.59(0.038)
	x_1x_2		-556.56(0.018)	-539.26(0.977)	-556.78(0.080)	-556.57(0.036)
			-556.56(0.018)	-526.47(0.966)	-556.75(0.156)	-556.56(0.035)
8	$c, x_1, x_2,$	-552.06	-551.48(0.025)	-536.60(0.988)	-551.54(0.097)	-551.49(0.031)
	x_1x_2		-551.47(0.025)	-536.39(0.905)	-551.51(0.082)	-551.47(0.033)
			-551.50(0.025)	-538.28(0.950)	-551.57(0.065)	-551.51(0.029)

3.3.3 Marginal Likelihoods

For each model $\mathcal{M}_k$, auxiliary mixture sampling was run for $M = 20000$ draws after a burn-in of 5000 draws, yielding a sequence of draws $(\boldsymbol{\alpha}_k^{(m)}, Q^{(m)})$ for the unknown parameter. For illustration, Figure 2 shows for model $\mathcal{M}_6$ that the sampler is converging quickly to the stationary distribution and mixing is pretty good also for $Q^{(m)}$ which is a parameter that is often slowly mixing.

For Chib's estimator (13) as well as for the complete-data likelihood estimator (16), the posterior ordinate $\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y}) = \hat{p}(\boldsymbol{\alpha}_k^*|Q^*, \mathbf{y})\hat{p}(Q^*|\mathbf{y})$ is estimated in two steps. First, Q^* is chosen as the average of $Q^{(1)}, \ldots, Q^{(M)}$ and $\hat{p}(Q^*|\mathbf{y})$ is estimated by

$$\hat{p}(Q^*|\mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} p(Q^*|\gamma_1^{(m)}, \ldots, \gamma_N^{(m)}, \mathcal{M}_k),$$

where the conditional densities are equal to the inverted Gamma distribution given in step (b1) of auxiliary mixture sampling. Then a reduced auxiliary mixture sampler is run where Q is fixed at Q^* and step (b1) is omitted. The average of the draws $\boldsymbol{\alpha}_k^{(m)}$ of this reduced sampler define $\boldsymbol{\alpha}_k^*$ while the draws $\mathbf{z}_k^{(m)}$ are used to estimate the posterior ordinate $p(\boldsymbol{\alpha}_k^*|Q^*, \mathbf{y})$:

$$p(\boldsymbol{\alpha}_k^*|Q^*, \mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{\alpha}_k^*|Q^*, \mathbf{z}_k^{(m)}, \mathbf{y}),$$

where the conditional densities are equal to the multivariate normal distribution given in step (b2) of auxiliary mixture sampling.

To implement Chib's estimator (13), the integrated likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$ is approximated by importance sampling:

$$\hat{p}(\mathbf{y}|\boldsymbol{\vartheta}_k^*) = \frac{1}{L} \sum_{l=1}^{L} \frac{p(\mathbf{y}|\boldsymbol{\alpha}_k^*, \gamma_1^{(l)}, \ldots, \gamma_N^{(l)}) p(\gamma_1^{(l)}, \ldots, \gamma_N^{(l)})|Q^*)}{q(\gamma_1^{(l)}, \ldots, \gamma_N^{(l)})}, \tag{26}$$

where $q(\gamma_1, \ldots, \gamma_N) = \prod_{i=1}^N q(\gamma_i)$ and a univariate normal distribution is fitted to the MCMC draws $\gamma_i^{(m)}$ to obtain $q(\gamma_i)$ for each $i = 1, \ldots, N$.

To implement the complete-data likelihood estimator (16), a further run of reduced auxiliary mixture sampling is added, where $(\boldsymbol{\alpha}_k, Q)$ is fixed at $(\boldsymbol{\alpha}_k^*, Q^*)$. This sampler consists only of the steps (a) and (c). The average of the MCMC draws $(\gamma_1^{(m)}, \ldots, \gamma_N^{(m)})$ defines $\boldsymbol{\beta}_k^* = (\gamma_1^*, \ldots, \gamma_N^*)$ while the draws $\mathbf{z}_k^{(m)}$ are used to estimate the posterior ordinate $\hat{p}(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{y})$ as in (18). The conditional density $p(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{z}_k^{(m)}, \mathbf{y}) = \prod_{i=1}^N p(\gamma_i^*|\boldsymbol{\vartheta}_k^*, \mathbf{z}_k^{(m)}, \mathbf{y})$ is simply the product of the N univariate normal densities appearing in step (a).

To implement importance sampling and bridge sampling, the importance density $q(\boldsymbol{\beta}_k, \boldsymbol{\vartheta}_k) = q(\boldsymbol{\alpha}_k)q(Q) \prod_{i=1}^N q(\gamma_i)$ is considered. A univariate normal distribution is fitted to the MCMC draws $\gamma_i^{(m)}$ to obtain $q(\gamma_i)$ for each $i = 1, \ldots, N$. The importance densities $q(\boldsymbol{\alpha}_k)$ and $q(Q)$ are constructed in an unsupervised manner as mixture densities with $S = 200$ components, based on step (b1) and (b2) of auxiliary mixture sampling:

$$q(\boldsymbol{\alpha}_k) = \frac{1}{S}\sum_{m=1}^{S} p(\boldsymbol{\alpha}_k|\mathbf{y}, Q^{(m)}, \mathbf{z}_k^{(m)}), \tag{27}$$
$$q(Q) = \frac{1}{S}\sum_{m=1}^{S} p(Q|\gamma_1^{(m)}, \ldots, \gamma_N^{(m)}).$$

Table 2 compares the different estimators of the (log) marginal likelihood of the various logit models including a random intercept. Each estimator is computed three times to evaluate accuracy and stability. Additionally, numerical standard errors were computed as in Chib (1995).

Chib's estimator $\hat{p}_{CH}(\mathbf{y}|\mathcal{M}_k)$ which is based on the importance sampling approximation (26) of the integrated likelihood $\hat{p}(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$ is very precise. On the other hand, the complete-data likelihood estimator $\hat{p}_{CDL}(\mathbf{y}|\mathcal{M}_k)$ which avoids the direct computation of $\hat{p}(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$ is extremely imprecise. The inaccuracy of this estimator is even larger than the standard errors would indicate and in comparison to the other estimators an upwards bias seems to be present. Furthermore, the estimator is very unstable, see for instance the repeated estimation of the marginal likelihoods of model $\mathcal{M}_1$ or $\mathcal{M}_4$.

The importance sampling estimator $\hat{p}_{IS}(\mathbf{y}|\mathcal{M}_k)$ tends to be much more inaccurate than Chib's estimator, see in particular the standard errors of the third estimator for model $\mathcal{M}_3$ or $\mathcal{M}_5$. This indicates considerable sensitivity to the unsupervised importance density (27). In contrast to that, the bridge sampling estimator $\hat{p}_{BS}(\mathbf{y}|\mathcal{M}_k)$ is stable over repeated estimation and the standard errors are comparable to that of Chib's estimator. This indicates robustness to the unsupervised importance density (27) as observed earlier in Frühwirth-Schnatter (2004).

The marginal likelihoods of a logit model without heterogeneity, based on importance sampling as in Subsection 2.2, were added to Table 2. Among all models considered, the model including x_1 (the root extract), the interaction term x_1x_2 between the seed and the root extract, and a random intercept has the largest marginal likelihood. Evidence in favor of this model compared to a model with the same predictors, but no random intercept, however, is pretty weak.

3.4 Application to State Space Modelling of Count Data

3.4.1 Data and Modelling

For further illustration, we consider two time series of counts. First, we reanalyze a time series of reported cases of purse snatching y_t in the Hyde park neighborhood in Chicago, taken from Harvey (1989). The data cover the period from January 1968 to September 1973 and are 28 days apart, see also Figure 3. We assume that $y_t \sim \mathcal{P}(\lambda_t)$ where the intensity λ_t is time-varying:

$$\log(\lambda_t) = \mu_t, \tag{28}$$

and μ_t is a stochastic trend, following a random walk with drift a_{t-1}:

$$\mu_t = \mu_{t-1} + a_{t-1} + w_{1t}, \qquad w_{1t} \sim \mathcal{N}(0, \theta_1), \tag{29}$$

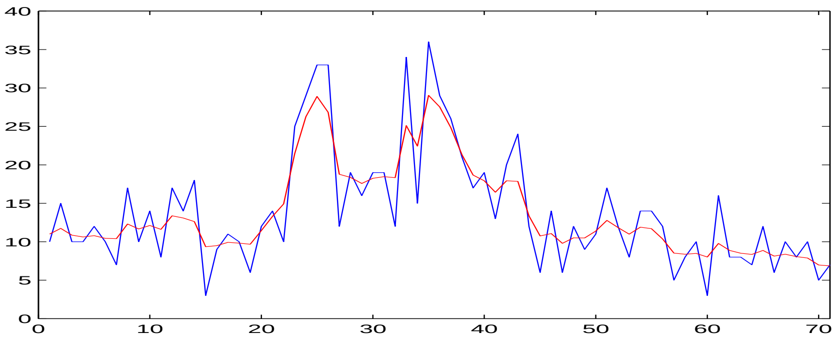

Figure 3: Purse snatching in Hyde Park, Chicago

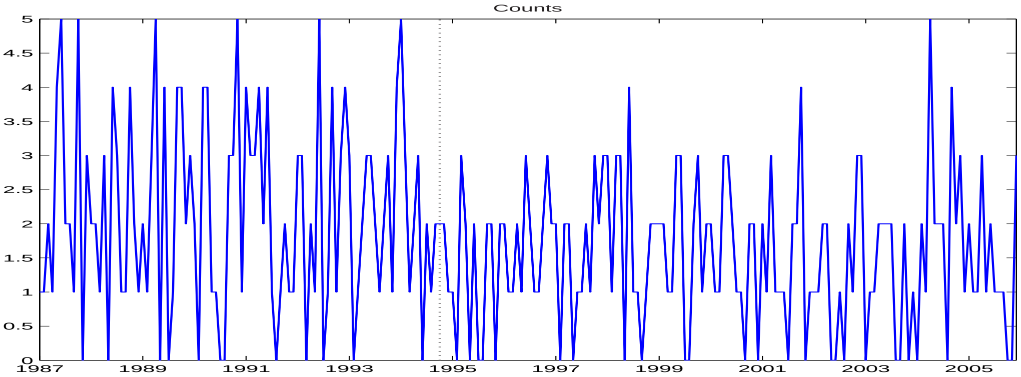

Figure 4: Children data

and $\mu_0 \sim \mathcal{N}(0,1)$. In its most general form, the model assumes that the drift changes over time and itself follows a random walk:

$$a_t = a_{t-1} + w_{2t}, \qquad w_{2t} \sim \mathcal{N}(0, \theta_2), \tag{30}$$

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

The second time series consists of monthly counts of killed or injured pedestrians, aged 6-10, from 1987-2005 in Linz, which is the third largest town in Austria.[4] The observations are a series of small counts not exceeding 5, see also Figure 4. 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. For these data, a basic structural model with intervention effect for Poisson counts as in Durbin and Koopman (2000) and Frühwirth-Schnatter and Wagner (2006) is fitted to the number y_t of persons killed or seriously injured in time period t, $y_t \sim \mathcal{P}(e_t\lambda_t)$, where e_t is the number of children living in Linz. λ_t is a very small intensity, assumed to

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

have a multiplicative trend as well as a multiplicative seasonal component:

$$\log(\lambda_t) = \mu_t + s_t. \tag{31}$$

μ_t is a stochastic trend as in (29), however $\mu_0 \sim \mathcal{N}(\log(y_1/e_1), 1)$. To capture the intervention effect, equation (29) is slightly modified by including a level shift δ at the time point $t = t_{\text{int}}$, when the legal amendments became effective:

$$\mu_t = \mu_{t-1} + a_{t-1} + \delta + w_{1t}. \tag{32}$$

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

$$s_t = -s_{t-1} - \cdots - s_{t-11} + w_{3t}, \qquad w_{3t} \sim \mathcal{N}(0, \theta_3), \tag{33}$$

where $\sum_{j=0}^{11} s_{-j} = 0$, and $\mathbf{s}_0 = (s_{-1}, \ldots, s_{-11})'$ is an unknown initial pattern following the prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. As MCMC did not converge for this formulation of the seasonal component, we use a reparametrization where the seasonal component is non-centered:

$$\log(\lambda_t) = \mu_t + s_t = \mu_t + \mathbf{Z}_t\mathbf{s}_0 + \theta_4 \frac{s_t - \mathbf{Z}_t\mathbf{s}_0}{\theta_4} = \mu_t + \mathbf{Z}_t\mathbf{s}_0 + \theta_4\tilde{s}_t, \tag{34}$$

where $\theta_4 = \pm\sqrt{\theta_3}$. Here the initial seasonal pattern $\mathbf{s}_0$ is introduced as fixed effect and $\mathbf{Z}_t$ is a row vector selecting the appropriate initial value according to the season of time point t. For t being a multiple of 12, $\mathbf{Z}_t$ is a row vector of -1, otherwise all elements of $\mathbf{Z}_t$ are 0, apart from the element in the column corresponding to the actual season, which takes the value 1.

Model selection within the basic structural model amounts to determining which of the components, level μ_t, drift a_t and seasonal s_t should be included, and whether these are stochastic or not. Additionally, for the road safety data, we have to test for the presence of an intervention effect.

For an unrestricted model, the parameters θ_1, θ_2 and θ_4 appearing in (29), (30) and (34) are unknown and estimated under the conditionally conjugate priors $\theta_i \sim \mathcal{G}^{-1}(c_0, C_0), i = 1, 2$ and $\theta_4 \sim \mathcal{N}(0, 1)$. The model simplifies considerably, if some of these variances are equal to zero. The stochastic trend, for instance, reduces to a log-linear deterministic trend with intercept μ_0 and slope a_0, if both variances θ_1 and θ_2 are zero. Choosing $\theta_4 = 0$ leads to a fixed seasonal pattern over the whole observation period, whereas $\theta_4 \neq 0$ allows a smooth change in this pattern. If all variances are equal to zero, then the basic structural model reduces to a Poisson regression model with log-linear trend, fixed seasonal pattern, and intervention effect:

$$y_t \sim \mathcal{P}(e_t\lambda_t), \qquad \lambda_t = \exp(\mu_0 + ta_0 + \mathbf{Z}_t\mathbf{s}_0 + I_{\{t \geq t_{\text{int}}\}}\delta). \tag{35}$$

Otherwise, the basic structural model is regarded as a state space or dynamic generalized linear model for discrete observations, see e.g. West, Harrison, and Migon (1985) with state vector equals $\boldsymbol{\beta}_t = (\mu_t, a_t, \tilde{s}_t, \ldots, \tilde{s}_{t-10}, \delta)$. The state evolution in this state space model follows a first order Markov chain,

$$\boldsymbol{\beta}_t = \mathbf{F}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{36}$$

where the matrices $\mathbf{F}_t$ and $\mathbf{Q}$ are obtained from (29) to (34). The observation equation reads:

$$y_t \sim \mathcal{P}\left(e_t\lambda_t\right), \qquad \lambda_t = \exp(\mu_t + \mathbf{Z}_t\mathbf{s}_0 + \theta_4\tilde{s}_t).$$

The starting values $\boldsymbol{\beta}_0 = (\mu_0, a_0)$ are added to the latent variables, $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_T)$, while $\boldsymbol{\vartheta}_k = (\theta_1, \theta_2, \theta_4, \mathbf{s}_0)$. Note that the starting values for the non-centered seasonal component are equal to 0, $(\tilde{s}_{-1}, \dots, \tilde{s}_{-11})' = \mathbf{0}$. If any of the variances θ_1, θ_2 or θ_4^2 is equal to 0, or if no seasonal component is included, a reduced model parameter $\boldsymbol{\vartheta}_k$ results.

3.4.2 Auxiliary Mixture Sampling

Auxiliary mixture sampling is implemented as in 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, $t = 1, \dots, T$, and introduces the inter-arrival times τ_{tj}, $j = 1, \dots, (y_t + 1)$ of this Poisson process as missing data. Since each $\tau_{tj} \sim \mathcal{E}\left(e_t\lambda_t\right)$ we have

$$-\log \tau_{tj} = \log e_t + \log \lambda_t + \varepsilon_{tj}, \tag{37}$$

where $\varepsilon_{tj} = -\log \xi_{tj}$ with $\xi_{tj} \sim \mathcal{E}\left(1\right)$. The distribution of ε_{tj} is then approximated by a mixture of normal distributions as in (8) with component indicator r_{tj} and the auxiliary variables $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_T)$, where $\mathbf{z}_t = (\tau_{tj}, r_{tj}, j = 1, \dots, y_t + 1)$, are introduced.

Auxiliary mixture sampling is defined through the following conditional densities (Frühwirth-Schnatter and Wagner, 2006, Section 3.3):

(a) Sample $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_T)$ from $p(\boldsymbol{\beta}|\boldsymbol{\vartheta}, \mathbf{z}, \mathbf{y})$ by forward-filtering backward sampling as in Frühwirth-Schnatter (1994), Carter and Kohn (1994), or De Jong and Shephard (1995) for a conditionally Gaussian state space model being defined for each $t = 1, \dots, T$ by state equation (36) and $y_t + 1$ independent observation equations:

$$-\log \tau_{tj} - m_{r_{tj}} - \log e_t = \mu_t + Z_t\mathbf{s}_0 + \theta_4\tilde{s}_t + \varepsilon_{tj}, \quad \varepsilon_{tj} \sim \mathcal{N}\left(0, s^2_{r_{tj}}\right), \tag{38}$$

$j = 1, \dots, y_t + 1.$

(b1) Sample θ_1 (if unknown) and θ_2 (if unknown) independently from following inverted Gamma densities:

$$\theta_1 \sim \mathcal{G}^{-1}\left(c_0 + T/2, C_0 + 1/2\sum_{t=1}^{T}(\mu_t - \mu_{t-1} - a_{t-1} - \delta I_{\{t=t_{\text{int}}\}})^2\right), \tag{39}$$

$$\theta_2 \sim \mathcal{G}^{-1}\left(c_0 + T/2, C_0 + 1/2\sum_{t=1}^{T}(a_t - a_{t-1})^2\right). \tag{40}$$

(b2) Sample $\mathbf{s}_0$ (if a seasonal component is included) and θ_4 (if the seasonal component is stochastic) jointly from a normal distribution, obtained by combining all observation equations (38) which are regarded as a normal heteroscedastic linear regression model in the unknown parameters with a standard normal prior. Because the sign of θ_4 and $\tilde{s}_t$ is not defined uniquely from (38), a random sign switch is performed. With probability 0.5, θ_4 and $\tilde{s}_t$ are unchanged, whereas with probability 0.5 the signs of θ_4 and $\tilde{s}_t, t = 1, \dots, T$ are changed.

(c) For each $t = 1, \dots, T$, sample the inter-arrival times $\{\tau_{tj}, j = 1, \dots, y_t + 1\}$. If $y_t > 0$, sample the order statistics $u_{t,(1)}, \dots, u_{t,(n)}$ of $n = y_t$ $\mathcal{U}[0, 1]$ random variables, see Robert and Casella (1999, p.47) for details, and define the inter-arrival times τ_{tj} as their increments: $\tau_{tj} = u_{t,(j)} - u_{t,(j-1)}, j = 1, \dots, n$, where $u_{i,(0)} := 0$. Sample the final arrival time as $\tau_{t,n+1} = 1 - \sum_{j=1}^{n} \tau_{tj} + \xi_t$, where $\xi_t \sim \mathcal{E}(\lambda_t)$. Sample the component indicator r_{tj} conditional on τ_{tj} and λ_t from the following discrete density:

$$\Pr(r_{tj} = l|\tau_{tj}, \lambda_t) \propto \frac{w_l}{s_l} \exp\left\{-\frac{1}{2}\left(\frac{-\log \tau_{tj} - \log e_t - \log \lambda_t - m_l}{s_l}\right)^2\right\}. \quad (41)$$

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

3.4.3 Marginal Likelihoods

Both for Chib's estimator (13) as well as for the complete-data likelihood estimator (16), the posterior ordinate $\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})$ can be estimated in one step. θ_1^*, θ_2^*, and $\mathbf{s}_0^*$ are estimated as the average of the corresponding MCMC draws. θ_4^* is estimated as the positive square root of θ_3^*, i.e. $\theta_4^* = +\sqrt{\theta_3^*}$, where θ_3^* is the average of the MCMC draws $\theta_3^{(m)} = (\theta_4^{(m)})^2$. The posterior ordinate $\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y})$ is given by:

$$\hat{p}(\boldsymbol{\vartheta}_k^*|\mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} p(\theta_1^*|\boldsymbol{\beta}_k^{(m)}) p(\theta_2^*|\boldsymbol{\beta}_k^{(m)}) p(\mathbf{s}_0^*, \theta_4^*|\boldsymbol{\beta}_k^{(m)}, \mathbf{z}_k^{(m)}), \quad (42)$$

where $p(\theta_i^*|\boldsymbol{\beta}_k^{(m)})$ and $p(\mathbf{s}_0^*, \theta_4^*|\boldsymbol{\beta}_k^{(m)}, \mathbf{z}_k^{(m)})$ are equal to the inverted Gamma density in step (b1) and the multivariate normal density in step (b2), respectively.

To implement Chib's estimator, the integrated likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k^*)$ is approximated by particle filtering (Pitt and Shephard, 1999) as in Chib et al. (2002) and Omori, Chib, Shephard, and Nakajima (2007).

To implement the complete-data likelihood estimator, another run of reduced auxiliary mixture sampling is added, where $\boldsymbol{\vartheta}_k$ is fixed at $\boldsymbol{\vartheta}_k^*$. This sampler consists only of the steps (a) and (c). The average of the MCMC draws $(\boldsymbol{\beta}_0^{(m)}, \dots, \boldsymbol{\beta}_T^{(m)})$ defines $\boldsymbol{\beta}_k^*$ and $\hat{p}(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{y})$ is estimated as in (18). The evaluation of the conditional density $p(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{z}_k^{(m)}, \mathbf{y})$ where $\mathbf{z}_k^{(m)} = (\boldsymbol{\tau}_k^{(m)}, \mathbf{r}_k^{(m)})$ is feasible through:

$$p(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*, \mathbf{z}_k^{(m)}, \mathbf{y}) = \frac{p(-\log \boldsymbol{\tau}_k^{(m)}|\mathbf{r}_k^{(m)}, \boldsymbol{\beta}_k^*, \boldsymbol{\vartheta}_k^*) p(\boldsymbol{\beta}_k^*|\boldsymbol{\vartheta}_k^*)}{p(-\log \boldsymbol{\tau}_k^{(m)}|\mathbf{r}_k^{(m)}, \boldsymbol{\vartheta}_k^*)}, \quad (43)$$

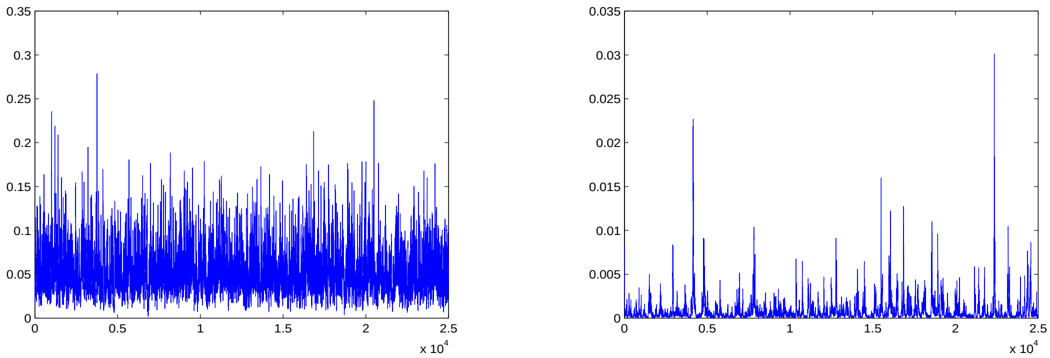

Figure 5: Purse snatching data; MCMC draws for θ_1 (left) and θ_2 (right) based on auxiliary mixture sampling

Table 3: Marginal likelihoods for the Purse snatching data

Model $\mathcal{M}_k$	$\log \hat{p}_{CH}(\mathbf{y}\vert\mathcal{M}_k)$	$\log \hat{p}_{CDL}(\mathbf{y}\vert\mathcal{M}_k)$
iid Poisson	-291.18 (0.0057)	—
Poisson regression with deterministic trend	-292.47(0.0093)	—
local level	-230.76(0.0095)	-230.30(0.4557)
local trend	-252.09(0.0193)	-251.31(0.4107)
local trend with $\theta_2 = 0$	-234.72(0.0105)	-238.60(0.4383)
local trend with $\theta_1 = 0$	-239.96(0.0295)	-230.49(0.5083)

where the numerator is the product of $T + \sum_{t=1}^{T} y_t$ univariate normal densities and T multivariate normal densities and the denominator is a by-product of running Kalman-filtering.

3.4.4 Application to Purse Snatching Data

For the purse snatching data we consider two fixed parameter models, namely a standard Poisson distribution with unknown intensity and a Poisson regression model with deterministic trend model, and four different state space models: a local level model, a local trend model, and two restricted local trend models with $\theta_1 = 0$ and $\theta_2 = 0$, respectively. Model selection is carried out under the priors $\theta_i \sim \mathcal{G}^{-1}(0.5, 0.2275), i = 1, 2.$[5]

Auxiliary mixture sampling was run for $M = 20000$ draws after a burn-in of 10000 draws. For illustration, Figure 5 shows for a local trend model with $\theta_1 > 0$ and $\theta_2 > 0$ that the sampler is converging quickly to the stationary distribution and mixing is pretty good.

Chib's estimator $\hat{p}_{CH}(\mathbf{y}|\mathcal{M}_k)$, the complete-data likelihood estimator $\hat{p}_{CDL}(\mathbf{y}|\mathcal{M}_k)$ and their standard errors are displayed for each model in Table 3. All models but the local level model have a posterior probability of practically zero. Again, the estimator based on the complete-data likelihood has very high standard errors for

[5] For the local trend model with $\theta_1 = 0$ we used the prior $\mathcal{G}^{-1}(2.5, 0.001)$, because the sampler did not converge for the priors $\mathcal{G}^{-1}(0.5, 0.2275)$, $\mathcal{G}^{-1}(2.5, 0.05)$ and $\mathcal{G}^{-1}(2.5, 0.005)$.

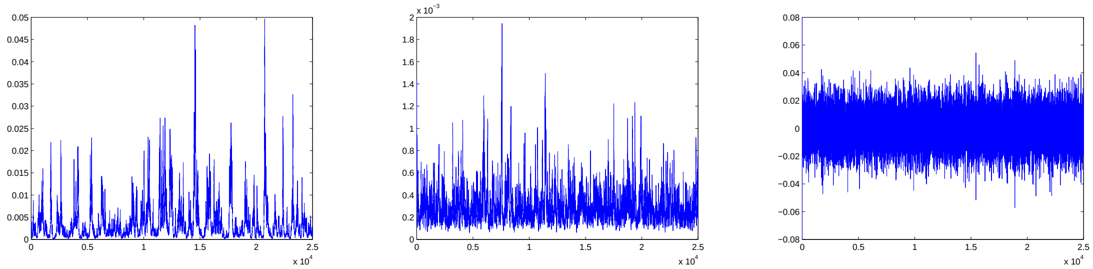

Figure 6: MCMC draws for θ_1 (left), θ_2 (middle) and $\theta_4 = \pm\sqrt{\theta_3}$ (right) for a basic structural model with intervention effect

all models. Chib's estimator, which is based on approximating the integrated likelihood by particle filtering, is very precise both for the fixed parameter models and the state space model. It should, however, be noted that uncertainty of estimating the integrated likelihood by particle filtering is not taken into account.

3.4.5 Application to the Road Safety Data

For the road safety data, we consider six different state space models: the basic structural model with and without intervention effect ($\delta = 0$), the local trend model with fixed seasonal pattern ($\theta_3 = 0$) and the local level model with fixed seasonal pattern ($\theta_2 = 0, \theta_3 = 0$) with and without intervention effect respectively. Model selection is carried out under the priors $\theta_i \sim \mathcal{G}^{-1}(2.5, 0.05), i = 1, 2$.

Auxiliary mixture sampling was run for $M = 20000$ draws after a burn-in of 10000 draws. Introducing the non-centered state vector $\boldsymbol{\beta}_t = (\mu_t, a_t, \tilde{s}_t, \ldots, \tilde{s}_{t-10}, \delta)$ led to a Gibbs sampler with rather quick convergence to the stationary distribution, see Figure 6 for illustration.

Chib's estimator $\hat{p}_{CH}(\mathbf{y}|\mathcal{M}_k)$, the complete-data likelihood estimator $\hat{p}_{CDL}(\mathbf{y}|\mathcal{M}_k)$ and their standard errors are displayed for each model in Table 4. The complete-data likelihood estimator has extreme standard errors for all models but the simple local level model. Also Chib's estimator based on the integrated likelihood is rather imprecise for this case study.

The local level model is clearly dominating the other state space models, however, the local level model with and without intervention are not well discriminated with regard to their marginal likelihoods. This becomes clear from Figure 7 indicating that the estimated level contains a smooth intervention effect, even if no explicit intervention effect is included, making model discrimination between these models difficult.

The estimated level of a local level model with intervention shown in Figure 7 suggest that before and after the intervention the level hardly changed and a regression model might fit the data equally well. For this reason, we added several Poisson regression models with a seasonal pattern based on dummy variables, with and without an intervention effect, and a model with a holiday effect (a dummy variable taking the value 1 in the months July and August). As the Poisson regression model results as the limiting form of a state space model, where the process variances θ_1, θ_2 and θ_3 are 0, we use the same prior for the regression coefficients as

Table 4: Marginal likelihoods for the road safety data

Model $\mathcal{M}_k$	$\log \hat{p}_{CH}(\mathbf{y}\vert\mathcal{M}_k)$	$\log \hat{p}_{CDL}(\mathbf{y}\vert\mathcal{M}_k)$	$\log \hat{p}_{BS}(\mathbf{y}\vert\mathcal{M}_k)$
loc. level – fixed seas., no int.	-377.72(0.407)	-376.40(0.651)	
loc. level – fixed seas., int.	-378.68(0.636)	-379.18(0.905)	
loc. trend – fixed seas., no int.	-414.18(0.351)	-404.96($9.1 \cdot 10^{28}$)	
loc. trend – fixed seas., int.	-415.71(0.261)	-403.00($3.7 \cdot 10^{30}$)	
basic struct. model – no int.	-418.93(0.451)	-405.80($2.4 \cdot 10^{27}$)	
basic struct. model – int.	-420.32(0.565)	-407.40($8.9 \cdot 10^{30}$)	
Poisson reg. – seas., no int.	-373.90(0.503)		-374.89(0.091)
Poisson reg. – seas., int.	-368.72(0.349)		-371.37(0.100)
Poisson reg. – holiday, no int.	-364.20(0.029)		-364.25(0.006)
Poisson reg. – holiday, int.	-360.23(0.037)		-360.26(0.008)

we used for μ_0, $\mathbf{s}_0$ and δ in the state space model.

The marginal likelihoods of the various Poisson regression models reported in Table 4 are estimated using Chib's estimator $\hat{p}_{CH}(\mathbf{y}|\mathcal{M}_k)$. As in Subsection 2.2, Chib's estimator tends to be rather imprecise for a Poisson regression model with seasonal dummy variables, where the dimension of the regression parameter is equal to 12 and 13, respectively, while it is very precise for a Poisson regression model with holiday effect, where the dimension of the regression parameter is equal to 2 and 3, respectively. For comparison, we add a bridge sampling estimator of the marginal likelihood, based on unsupervised construction of a mixture importance density as in (5) with $S = 100$ components. The conditional densities are obtained from auxiliary mixture sampling and are multivariate normal densities as in Subsection 2.2. Again, bridge sampling yields rather precise estimators of the marginal likelihoods even for the larger regression models.

When comparing all marginal likelihoods, we find that the simple Poisson regression model is dominating all state space models in terms of marginal likelihoods. The intervention effect is significant and the monthly seasonal pattern reduces to a holiday effect.

4 Concluding remarks

We investigated different estimators of the marginal likelihood for complex non-Gaussian models which makes use of of a simple MCMC method for estimating a broad class of non-Gaussian models called auxiliary mixture sampling (Frühwirth-Schnatter and Wagner, 2006; Frühwirth-Schnatter and Frühwirth, 2007).

For a fixed parameter model, Chib's estimator (Chib, 1995) is directly available from the output of auxiliary mixture sampling. An application to modelling nodal involvement data (Chib, 1995) through a logistic regression model and road safety data (Frühwirth-Schnatter and Wagner, 2006) through a Poisson regression model indicate that the combination of Chib's estimator with auxiliary mixture sampling yields precise estimators of the marginal likelihood if the dimension of the regression

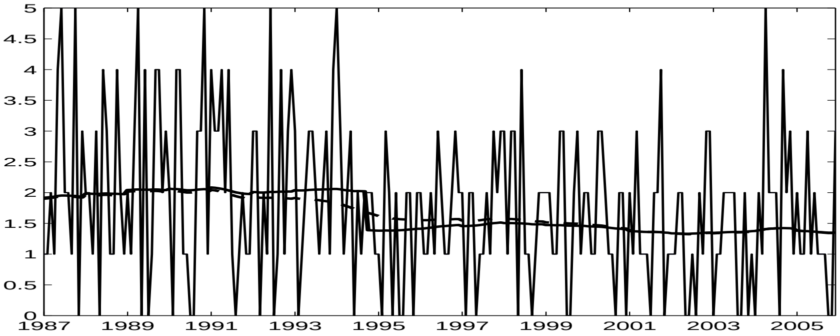

Figure 7: Counts of killed and injured children with estimated level for the local level model with intervention (full line) and without intervention (dashed)

parameter is not too high.

For latent variable models, the combination of Chib's estimator with auxiliary mixture sampling requires some additional method of determining the integrated likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k)$, like importance sampling for random effects models and particle filtering for state space models. This estimator turned out to be rather precise when modelling the seed data (Gamerman, 1997) and the purse snatching data (Harvey, 1989), while the marginal likelihood of the various state space models of the road safety data (Frühwirth-Schnatter and Wagner, 2006) were quite imprecise.

We also investigated an estimator based on the complete-data likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k, \boldsymbol{\beta}_k)$ which avoids the explicit computation of the integrated likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k)$. We showed that this estimator is based on estimating the integrated likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}_k)$ by the application of the marginal likelihood equation to a latent model with fixed model parameters $\boldsymbol{\vartheta}_k^*$. While this estimator is easily implemented by running reduced auxiliary mixture sampling, its performance turned out to be very disappointing for all of our case studies, yielding unstable and very imprecise results. Thus its application is not to be recommended.

Finally, importance and bridge sampling estimators have been combined with auxiliary mixture sampling. Auxiliary mixture sampling allows the unsupervised construction of an importance density as in Frühwirth-Schnatter (1995, 2004). While importance sampling turned out to be somewhat sensitive to this unsupervised importance density, bridge sampling yielded precise and stable estimators of the marginal likelihood.

Both Chib's estimator as well as the bridge sampling estimator can be accommodated to a wide range of further complex non-Gaussian models using auxiliary mixture sampling. It is, however, strongly recommended to evaluate the accuracy of these estimators carefully by computing standard errors as in Chib (1995) and by repeating marginal likelihood estimation several times to check the stability and reliability of the ensuing Bayesian model selection procedure.

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory.* Chichester: Wiley.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

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

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

Chib, S. and I. Jeliazkov (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association 96*, 270–281.

Chib, S., F. Nardari, and N. Shephard (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics 108*, 281–316.

Collett, D. (1991). *Modelling Binary Data.* London: Chapman & Hall.

Crowder, M. J. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics 27*, 34–37.

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

Dey, D., S. K. Ghosh, and B. K. Mallick (Eds.) (2000). *Generalized Linear Models: a Bayesian Perspective*, New York/Basel. Marcel Dekker.

DiCiccio, T. J., R. E. Kass, A. Raftery, and L. Wasserman (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association 92*, 903–915.

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.

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods.* Oxford: Oxford University Press.

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. (1997). Discussion of the paper by Diggle and AlWasel on "Spectral analysis of replicated biomedical time series ". *Applied Statistics 46*, 62–63.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal 7*, 143–167.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.

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

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.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing 7*, 57–68.

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.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–1339.

Godsill, S. J. (2001). On the relation between MCMC model uncertainty methods. *Journal of Computational and Graphical Statistics 10*, 230–248.

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

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, pp. 179–198. Oxford: Oxford University Press.

Han, C. and B. P. Carlin (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association 96*, 1122–1132.

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

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

Kadane, J. B. and N. Lazar (2004). Methods and criteria for model selection. *Journal of the American Statistical Association 99*, 279–290.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers of Econometrics*, pp. 105–142. New York: Academic.

Meng, X.-L. and W. H. Wong (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica 6*, 831–860.

Omori, Y., S. Chib, N. Shephard, and J. Nakajima (2007). Stochastic volatility with leverage: Fast likelihood inference. *Journal of Econometrics*, in press.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association 94*, 590–599.

Robert, C. P. and G. Casella (1999). *Monte Carlo Statistical Methods.* Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Scott, S. L. (2005). Data augmentation, frequentistic estimation, and the Bayesian analysis of multinomial logit models. Technical report, The Marshall School of Business, University of Southern California, Los Angelos, CA.

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data.* Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.

West, M., P. J. Harrison, and H. S. Migon (1985). Dynamic generalized linear models and Bayesian forecasting (C/R: p84-97). *Journal of the American Statistical Association 80*, 73–83.

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

Zellner, A. and P. E. Rossi (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics 25*, 365–393.