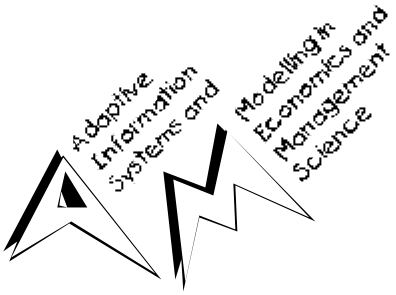
Working Paper Series

Selection of the Number of States by Birth-Death Processes

Leopold Sögner

Working Paper No. 69
May 2000

Working Paper Series



May 2000

SFB
'Adaptive Information Systems
and Modelling in Economics and Management Science'

Vienna University of Economics
and Business Administration
Augasse 2-6, 1090 Vienna, Austria

in cooperation with

University of Vienna
Vienna University of Technology

http://www.wu-wien.ac.at/am

This piece of research was supported by the Austrian Science
Foundation (FWF) under grant SFB # 010 ('Adaptive Information Systems
and Modelling in Economics and Management Science').

Selection of the Number of States by Birth-Death Processes

Leopold Sögner
E-mail: Leopold.Soegner@wu-wien.ac.at
Department of Economics
Vienna University of Economics
and Business Administration
Augasse 2 – 6, A-1090 Vienna, AUSTRIA

May 22, 2000

Abstract

In this article we use spatial birth-death processes to estimate the number of states k of a switching model. Following Preston (1976) and Stephens (1998) matching the *detailed balance condition* for the underlying birth-death process results in an unique invariant probability measure with the corresponding stationary distribution of the number of states. This concept could be easily integrated to Bayesian sampling to derive the marginal posterior distribution of number of states within the sampling procedure. We apply this technique to simulated $AR(1)$ data and to quarterly Austrian data on unemployment and real gross domestic product.

Keywords: Markov-Chain Monte Carlo, Birth-Death Processes, Switching Models.

1 Introduction

In this article we imbed a spatial birth-death processes into a *Markov-chain Monte Carlo* (MCMC) algorithm to derive an irreducible and a-periodic Markov chain, resulting in the full posterior distribution of the number of states k and the model parameters of a switching regression model.

In Bayesian modeling inference of the number of components is carried out by calculating the model likelihood from MCMC output for each model with a particular number of components separately, or by deriving the posterior distribution of the number of components within the sampler. For the former methods the reader is referred to Chib and Greenberg (1996) (*Candidates formula*), Frühwirth-Schnatter (1995) (Importance Sampling), DiCiccio *et al.* (1997) and Meng and Wong (1996) (*Bridge-sampling*). For problems with these methods especially when *label switching* occurs see Frühwirth-Schnatter (1999b). Methods considering a stochastic k, where k is sampled from the posterior distribution are *reversible jump* MCMC methods (RJ-MCMC) and *birth-death* MCMC methods (BD-MCMC). It is the posterior distribution of the number of components of a switching regression model we want derive and analyze in this paper. We prove of detailed balance and provide examples to describe the performance of BD-MCMC.

Let us start to relate our work to recent literature on models sampling from the posterior distribution of the number of states: Green (1995), Richardson and Green (1997) and Robert *et al.* (1999) used *reversible jump MCMC* algorithms (RJ-MCMC) to allow for jumps between a finite number of parameter sub-spaces in mixture models, where each sub-space is associated with a certain number of components. Within this algorithm the sampling of the number of components is based on the Metropolis-Hastings algorithm. Green and Mira (1999) provide a modified rejection algorithm. Richardson and Green (1997) give a detailed description and discussion on the properties and the implementability of the RJ-MCMC algorithm. The authors demonstrate and describe the dependence of the estimate of the number of components on the prior distribution, the ability of the sampler to separate between different components and on the occurrence of *label switching*. The authors conclude that with label switching or "bad priors" the performance of the RJ-MCMC algorithm is poor. In Brooks and Giudici (1998) the problem of convergence with RJ-MCMC is investigated, where the authors provide a convergence assessment technique that could also be applied to samples from BD-MCMC.

As an alternative to RJ-MCMC methods, Stephens (1998) suggested a sampler based on a *birth-death* process (BD) to estimate the number of components of a standard mixture model within the sampling algorithm. This method heavily relies on the properties of birth-death point process as already analyzed by Preston (1976). Ripley (1977) provides an idea how BD-processes can be simulated and applied to problems in biology and clustering. By applying the BD-MCMC methodology to mixture models new components are added to existing components (*birth*) or some of the existing components are erased (*death*).

Within this article the BD-MCMC algorithm is applied to a switching regression model and to a switching vector autoregressive model. By a proper specification of births and deaths the *detailed balance condition* for the underlying birth-death process is satisfied, resulting in a unique invariant probability measure with the corresponding stationary distribution of the number of states. In our applications to simulated data and Austrian data on unemployment and gross national product (GDP) we conclude that the performance of this method heavily

depends on the prior assumptions, which govern the ability of the sampler to detect structural breaks. We present examples where BD-MCMC jumps either to a number of states of one very rapidly and further births are followed by sudden deaths or to a high number of states. In the first case the sampler is not able to detect structural breaks while in the latter model over-fitting occurs. By our examples, we demonstrate that these effects depend primarily on the model priors.
This article is organized as follows: Section 2 describes the statistical model and specifies births and deaths. Detailed balance is checked in appendix A. An ergodic Markov-chain is constructed in section 3. The performance of BD-MCMC in switching models is presented in section 4, where simulated VAR and unemployment-GDP data are used to check the performance of this method.

2 BD Processes and Switching Models

Consider the random variable k, the number of components within a switching model. Given a prior $\pi(k)$ and data $\mathcal{Y}^N$, it is the posterior distribution of $\pi(k|\mathcal{Y}^N)$ we are going to investigate within a Bayesian model. For some particular k, $k = \{1, \ldots, k_{max}\}$, let us consider the following sub-model:

$$y_t \quad = \quad \beta_k^i z_t + \varepsilon_t^i \ . \tag{1}$$

where y_t is the vector of response variables, z_t is the vector of prediction variables and k is the number of states. We label each state by $i \in \{1, 2, ..., k\}$. The components of y_t are labeled by the index ι, $\iota = 1, \ldots, p$. ε_t^i is an independent identically distributed (iid) vector of normal random variables with zero mean and variance $(\sigma_\iota^2)^i$, i.e. $\varepsilon_{\iota,t}^i \ \sim \ iid\ \mathcal{N}(0, (\sigma_\iota^2)^i)$. For a sub-model with k components the parameters θ_k consist of the state specific model parameters $\beta_k := (\beta_k^i)_{i=1}^k = ((\beta_{k,\iota}^i)_{\iota=1}^p)_{i=1}^k$ and $\sigma_k^2 := ((\sigma_k^2)^i)_{i=1}^k = (((\sigma_{k,\iota}^2)^i)_{\iota=1}^p)_{i=1}^k$, and the matrix of the Markov transition probabilities $\eta_k := (\eta_{1,1}, \ldots, \eta_{1,k}, \eta_{2,1}, \ldots, \eta_{k,k})$; row l of η_k is expressed by $\eta_{\diamond l}$. The latent switching variable $I_{k,t}$ follows a homogenous Markov process in discrete time, taking values on a k-dimensional simplex $\mathcal{E}$ (see Karlin and Taylor (1975)). Since we only observe data $\mathcal{Y}^N := (y_t, z_t)_{t=1}^N$ the corresponding sequence of switching variables is defined by $I_k^N := (I_{k,t})_{t=1}^N$. This results in the *augmented parameters* $\Psi_k = (\theta_k, I_k^N)$, where θ_k consists of common parameters, the state specific model parameters, and the matrix of the Markov transition probabilities η_k. Furthermore we abbreviate $(\beta_k)_{k=1}^{k_{max}}$, $(\sigma_k^2)_{k=1}^{k_{max}}$, $(\eta_k)_{k=1}^{k_{max}}$, $(\theta_k)_{k=1}^{k_{max}}$ and $(\Psi_k)_{k=1}^{k_{max}}$ by β, σ^2, η, θ and Ψ respectively.

Remark 1 *In the following analysis we do not include common parameters to keep the model and the notation simple. The following approach can easily extended to a model where some parameters of β_k and/or σ_k are common, i.e. they do not switch with $I_{k,t}$.*

Prior Assumptions: We assume that the priors of the switching regression model fulfill:

A1: $\pi(\beta_k, \sigma_k^2, \eta_k, i_k^N|k) = \pi(\beta_k, \sigma_k^2|k)\pi(\eta_k|k)\pi(i_k^N|\eta_k, k)$.
$\pi(\beta_k, \sigma_k^2|k) = \pi(\beta_k^1, (\sigma_k^2)^1|k) \cdot \ldots \cdot \pi(\beta_k^k, (\sigma_k^2)^k|k)$.

A2: *The switching process* $(I_{k,t})$ *is Markov*, i.e.

$$\pi(i_k^N|\theta_k,\eta_k,k)=\pi(i_k^N|\eta_k,k)=\prod_{\iota=1}^{k}\prod_{i=1}^{k}\eta_{ij}^{N_{ij}}\pi(i_0|\eta_k,k)\ ,$$

where $\pi(i_{k,0}|\eta_k,k)$ is the initial distribution of $I_{k,0}$, $N_{il}=\#(I_{k,t}=l|I_{k,t-1}=i)$ and η_{il} are the conditional probabilities to switch from state $I_{k,t-1}=i$ to $I_{k,t}=l$. i_k^N is a realization of the process I_k^N defined on the set $\mathcal{I}_k$

A3: *Priors are invariant to permutations:* $\pi(\beta_k,\sigma_k^2|k)$, $\pi(\eta_k|k)$, $\pi(i_k^N|k)$ are invariant to permutations $\rho(.)$ of the indices $i=1,\ldots,k$.

A4: *Cascasde structure of priors:* For every k, $k=1,\ldots,k_{max}$ we require that: $\pi(\beta_{k+1},\sigma_{k+1}^2|k+1)=r(\beta_k,\sigma_k^2)\pi(\beta_k,\sigma_k^2|k)\pi(\beta_{k+1}^{k+1},(\sigma_{k+1}^2)^{k+1}|k+1)$ and $\pi(\eta_{k+1}|k+1)=r(\eta_k)\pi(\eta_k|k)\Pi_{i=1}^{k}\pi((1-\eta_{i,k+1}),\eta_{i,k+1}|k+1)\pi(\eta_{\diamond k+1})$. $r(.)$ are Radon-Nykodim derivatives. This assumption on the structure of sub-model priors is needed to prove detailed balance.

In the following paragraph we briefly describe the estimation of Ψ_k for a standard switching model with k fixed (see e.g. Casella and George (1992), Albert and Chib (1993), Robert (1994) and Chib and Greenberg (1996)). The reader familiar with these concepts could skip this paragraph. For every fixed k the parameters $\beta_k,\sigma_k^2,\eta_k$ and the latent switching variable I_k^N are derived by MCMC. Due to the hierarchical structure of this Bayesian model the vector of parameters can be derived from successive sampling from the conditional distributions of the parameters. By the convergence properties of ergodic Markov-chains, (geometric) convergence to the invariant distribution of Ψ_k for k fixed is guaranteed (for regularity conditions see e.g. Robert (1994) and Chib and Greenberg (1996)). Table 1 presents the *conjugate priors* usually used with switching models. These distributions share the property that if the a-priori distribution is in the class $\mathcal{C}$, the a-posteriori distribution lies in $\mathcal{C}$. A detailed discussion on the advantages and disadvantages of conjugate prior modeling is provided in Robert (1994). The parameters are defined as follows for every k: $N_{il}:=\#(I_{k,t}=l|I_{k,t-1}=i)$ is the number of jumps from $[I_{k,t-1}=i]$ to $[I_{k,t}=l]$ in I_k^N. The assumption that the states are multinomial $(\mathcal{M}_k)$ of degree k results in the Dirichlet distribution $(\mathcal{D}_k)$ of the vectors of transition probabilities $\eta_{\diamond 1},\eta_{\diamond 2},...,\eta_{\diamond k}$. Concerning the variance terms $(\sigma_\iota^2)^i$, $i=1,...,k$ and $\iota=1,\ldots,p$, the parameters of the inverse gamma distributions $(\mathcal{IG})$ are given by $\nu_\iota^i:=\nu_{0,\iota}^i+0.5N_i$ and $D_\iota^i:=D_{0,\iota}^i+0.5\sum_{v=1}^{N}S_{k,v}^{i,N}(y_{\iota,v}-\beta_\iota^i z_v)^2$, where $N_i:=\#(I_t=i)$ is the frequency the chain has hit state i, $y_{\iota,v}$ is the vth element of component ι and z_v is the vth observation of the vector of prediction variables. To derive $S_{k,v}^{i,N}$, let us express $I_{k,v}^N$ by an $N\times k$ matrix S_k^N, where each row S_w is a vector where the l-th element of this row is equal to 1 if $[I_{k,w}=l]$, the other elements of the row $S_{k,w}$ are zero. $S_k^{i,N}$ is the ith column of the S_k^N and $S_{k,v}^{i,N}$ the vth element of this column. Considering the distribution of $\beta_{k,\iota}^i$ we define $Z_i=S_k^{i,N}z^N$, where the wth row of Z_i is equal to $z_w\in z^N$ if $[I_{k,w}=i]$, and it is equal to vector of zeros in all other states. I.e. by means of $S_k^{i,N}$ we project on the states $[I_{k,t}=i]$ in I_k^N. Last but not least, $\kappa_\iota=:(Z_i'Z_i+(B_{0,\iota}^i)^{-1})^{-1}$. Moreover, if we have no real prior information to discriminate between the different states a permutation of the labels results in the same data likelihood $\mathcal{L}(\mathcal{Y}^N|\Psi_k)$ for at least two different vectors of parameters. Therefore the unrestricted model is not identifiable. Imposing a restriction R on

Ψ_k, i.e. $\Psi_1 < ... < \Psi_i < ... < \Psi_k$ makes the problem identifiable. In this article we keep the model identifiable by applying *permutation sampling* (see Frühwirth-Schnatter (1999a)).

Parameter	a-priori	a-posteriori
$\eta_{i.}$		Dirichlet
	$\mathcal{D}(e_{0,i1}, ..., e_{0,ik})$	$\mathcal{D}(e_{0,i1} + N_{i1}, ..., e_{0,ik} + N_{ik})$
$(\sigma_\iota^2)^i$		Inverse Gamma
	$\mathcal{IG}(\nu_{0,\iota}^i, D_{0,\iota}^i)$	$\mathcal{IG}(\nu_\iota^i, D_\iota^i)$
β_ι^i		Normal
	$\mathcal{N}(b_{0,\iota}^i, B_{0,\iota}^i(\sigma_\iota^2)^i)$	$\mathcal{N}(\kappa_i(Z_i' y_\iota^N + (B_{0,\iota}^i)^{-1} b_{0,\iota}^i), \kappa_i(\sigma_\iota^2)^i)$

Table 1: Conjugate Priors for the Switching Model

As already stated I_k^N and k are considered to be random variables on the corresponding probability space generated by the data set $\mathcal{Y}^N = (y_t, z_t)_{t=1}^N$ and the parameters of the model. Since we consider a Bayesian model we want to derive the posterior $\pi(\Psi, k|\mathcal{Y}^N)$, $\Psi = (\Psi_k)_{k=1}^{k_{max}}$, given the prior $\pi(\Psi, k)$. In the following $f(\mathcal{Y}^N|\Psi_k, k)$ is the density function of the data $\mathcal{Y}^N$ under the augmented parameters Ψ_k and a certain number of states k. In the further analysis this density function will also be called marginal likelihood of the data $\mathcal{L}(\mathcal{Y}^N|\Psi_k) := f(\mathcal{Y}^N|\Psi_k, k)$. After this brief description of the switching model – which is standard by the way – let us focus on the estimation of the unknown number of states k by means of a birth-death process. Spatial birth-death processes have been introduced by Preston (1976), while first applications to statistical modeling are provided in Ripley (1977). In the following we closely follow the seminal work of Stephens (1998), who applied the birth-death framework to mixture models. The usual birth-death process is a Markov chain in continuous time having non-negative integers as state space. In this setup the parameter space for the parameters x_k changing with k is denoted by Ω, where $\Omega = \bigcup_{k\geq 1} \Omega_k$. Ω_k is a set of parameters for a sub-model (1) with k components, where the labelling of the parameters is ignored. Since we do not consider common parameters the parameters $x_k \in \Omega_k$ are equal to Ψ_k.

Births: Let k jump to $k+1$ at time s, where the arrival times are exponentially distributed resulting in an independent Poisson process. If a birth occurs the parameters β_k, σ_k^2, η_k, and I_k^N have to be adapted to fit into a $k+1$ state sub-model:

B1 Parameters β_k and σ_k^2: The components β_{k+1}^{k+1} and $(\sigma_{k+1}^2)^{k+1}$ are added to a sub-model with k states.

B2 Transition probabilities η_k: The transition probabilities are altered in the following way: Suppose that for each row $\eta_{\diamond i}$ we assume a Dirichlet prior $\mathcal{D}_k(\gamma_1, \ldots, \gamma_k)$, where $\gamma_l = v_2$ if $i \neq l$ and $\gamma_i = v_1$ of $i = l$. For each row $i = 1, ..., k$ of the matrix η_k we add an $\eta_{i,k+1}$ to each row i and multiply $\eta_{\diamond i}$ by $(1 - \eta_{i,k+1})$. The $k+1$-th row of η_{k+1} is given by $\eta_{\diamond k+1} = (\eta_{k+1,1}, ..., \eta_{k+1,k+1})$. We assume that $((1 - \eta_{i,k+1}), \eta_{i,k+1})$ are independent with a Dirichlet $\mathcal{D}_2(v_2, v_2)$ density and $\eta_{\diamond k+1} \sim \mathcal{D}_{k+1}(v2, \ldots, v2, v1)$. It is easily checked that these assumptions result in $\eta_{\diamond i} \sim \mathcal{D}_{k+1}(\gamma_1, \ldots, \gamma_{k+1})$ for each row of η_{k+1}, where $\gamma_l = v_2$

for $l \neq i$ and $\gamma_i = v_1$ for $l = i$ (e.g. see the poof of (16) in appendix A), which satisfies the assumption $A4$ with $r(\eta_k) = 1$.

B3 Switching variable: Consider the switching variable $I_k^N(t)$, where each component $I_w(t)$ takes values $i = 1, .., k$ and the prior is given by assumption A2. If k changes to $k+1$ at time s, the switching variable changes to $I_{k+1}^N(s)$, i.e. $I_k^N(s)$ dies with probability one and is replaced by $I_{k+1}^N(s)$, where the prior is given by $\pi(i^N|\eta_{k+1}, k+1)$.

Summing up, we derive parameters $x_{k+1} \in \Omega_{k+1}$, where the density of the parameters to be borne $\beta_{k+1}^{k+1}, \sigma_{k+1}^{k+1}, \eta_{1,k+1}, ..., \eta_{k,k+1}, \eta_{k+1,1}, ..., \eta_{k+1,k+1}, i_{k+1}^N$ is given by:

$$\begin{aligned}
\pi^b \quad (\quad & \beta_{k+1}^{k+1}, (\sigma_{k+1}^2)^{k+1}, \eta_{1,k+1}, ..., \eta_{k,k+1}, \eta_{k+1,1}, ..., \eta_{k+1,k+1}, i_{k+1}^N) = \\
= \quad & 1/\mathrm{C}_b \cdot b(x_k)\pi(\beta_{k+1}^{k+1}, (\sigma_{k+1}^2)^{k+1}|k+1) \left(\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)\Gamma(v1)}\right)^k \\
\cdot \quad & \eta_{1,k+1}^{v2-1} \cdots \eta_{k,k+1}^{v2-1} (1-\eta_{1,k+1})^{kv2-1} \cdots (1-\eta_{k,k+1})^{kv2-1} \\
\cdot \quad & \frac{\Gamma(kv_2+v_1)}{\Gamma(v2)^k\Gamma(v1)} \eta_{k+1,1}^{v1-1} \eta_{k+1,2}^{v2-1} \cdots \eta_{k+1,k+1}^{v2-1} \pi(i^N|\eta_{k+1}, k+1) \ . \qquad (2)
\end{aligned}$$

$b(x_k)$ is a function in the parameters $x_k \in \Omega_k$ into the positive real number and C_b is the normalizing constant to make $\pi^b(.)$ a density.

Deaths: Suppose that the process is at $x_k \in \Omega_k$:

D1 Parameters θ_k:
Point $\tilde{x}_i := \beta_k^i, (\sigma_k^2)^i, \eta_{1,i}, ..., \eta_{k,i}, \eta_{i,1}, ..., \eta_{i,k+1}, i_k^N$ dies independently as a Poisson process with a death rate $d_i(x_k)$, where $d_i(x_k)$ is – like $b(x_k)$ – a function of the parameters x_k which is independent of the index i.

D2 Switching variable: I_k^N is derived in the same way we have analyzed births, i.e. we get I_{k-1}^N from the density $\pi(i^N|\eta_{k-1}, k-1)$.

Overall birth and death rate: In this setup we assume that births evolve as a Poisson process with overall birth rate $\mathcal{B}(x)$ and the transition of $x_k \in \Omega_k$ to $x_{k+1} \in \Omega_{k+1}$ is described by (2), i.e. it defines a transition kernel $K_{\mathcal{B}}^k(x_k, F)$ from x_k to a set $F \in \Omega_{k+1}$. Define the parameters added or deleted from x_k by x^b and x^d respectively; then $x_{k+1} = x_k \setminus i_k^N \cup x^b$ and $x_{k-1} = x_k \setminus x^d \cup i_{k-1}^N$ for births and deaths respectively. Let $\mathcal{Z}_k^b$ be the support of the parameters x_k. For the overall birth rate $\mathcal{B}(x_k)$ we assume $\mathcal{B}(x_k) = \mathrm{C}_b$, i.e.

$$\begin{aligned}
\mathcal{B}(x_k) \quad = \quad & \int_{\mathcal{Z}_k^b} b(x_k)\pi^b(\beta_{k+1}^{k+1}, \ldots, i_{k+1}^N) \\
& \cdot d\beta d\sigma^2 d\eta_{1,k+1} \cdots d\eta_{k,k+1} d\eta_{k+1,1} \cdots d\eta_{k+1,k+1} di_{k+1}^N \ . \qquad (3)
\end{aligned}$$

Equivalently we derive from the transition kernel for deaths $K_\delta^k(x_k, G)$, with $G \in \Omega_{k-1}$, the overall death rate $\delta(x_k)$:

$$\delta(x_k) = \int_{\mathcal{I}_{k-1}^N} \sum_{i=1}^{k} d_i(x_k)\pi(i^N|\eta_{k-1}, k-1)di_{k-1}^N \ . \tag{4}$$

Thus, we have constructed an independent Poisson process, where the expected arrival time of a birth or a death is equal to $1/(\mathcal{B}(x_k) + \delta(x_k))$. The probability of a birth is $\frac{\mathcal{B}(.)}{\mathcal{B}(.)+\delta(.)}$ while the probability of a death is $\frac{\delta(.)}{\mathcal{B}(.)+\delta(.)}$. The question arises whether this birth-death setup has a stationary distribution $\pi(\Psi, k|\mathcal{Y}^N)$. As shown in appendix A a stationary distribution exists if:

Lemma 1 *Consider model (1) and assume that the prior assumptions A1, A2, A3 and A4 are satisfied. The birth-death process specified above has a stationary distribution $\pi(\Psi, k|\mathcal{Y}^N)$ if the following condition is fulfilled:*

$$(k+1)\frac{\pi(k+1)}{\pi(k)} r_{k+1} d_i(y) f(\mathcal{Y}^N|y, k+1) = b(z) f(\mathcal{Y}^N|z, k) \ , \tag{5}$$

for all $y \in \Omega_{k+1}$ and $z \in \Omega_k$, where y and z are realizations of the parameter vectors Ψ_{k+1} and Ψ_k respectively. $\pi(k)$ is the prior of the number of states k, and $f(\mathcal{Y}^N|\Psi_k, k) = \mathcal{L}(\mathcal{Y}^N|\Psi_k)$ is the marginal likelihood of model (1) with k parameters.

3 Construction of a Markov Chain

For every fixed k the parameters $\beta_k, \sigma_k^2, \eta_k$ and the latent switching variable I_k^N are derived by standard MCMC. The sampling algorithm for sampling periods $j = 1, ..., T$ works as follows:

Step G0: Define a Restriction $\mathcal{R}$: $\Psi_1 < ... < \Psi_i < ... < \Psi_k$.

Step G1: Sampling on the unrestricted set

$$\begin{array}{lll}
 & \vdots & \\
I_k^{N\ (j)} & \text{from} & \pi(i_k^N|\mathcal{Y}^N, \theta^{(j-1)}, k) \\
\eta_k^{(j)} & \text{from} & \pi(\eta_k|\mathcal{Y}^N, I_k^{N\ (j)}, \beta_k^{(j-1)}, (\sigma_k^2)^{(j-1)}, k) \\
(\sigma_k^2)^{(j)} & \text{from} & \pi(\sigma_k^2|\mathcal{Y}^N, I_k^{N\ (j)}, \beta_k^{(j-1)}, \eta_k^{(j)}, k) \\
\beta_k^{(j)} & \text{from} & \pi(\beta_k|\mathcal{Y}^N, I_k^{N\ (j)}, \eta_k^{(j)}, (\sigma_k^2)^{(j)}, k) \\
 & \vdots &
\end{array}$$

Step G2: Sort $\Psi_k^{(j)}$ according to the restriction $\mathcal{R}$.

The birth-death process starts at $s_0 = j - 1$ with a number of components $k^{(j-1)}$, where the continuous time scale is adapted to the sampling steps j of MCMC such that $j = s$ for all non-negative integers. The parameters $x_k \in \Omega_{k,\mathcal{R}} \in \Omega_k$ are sampled in the following way:

Step S0: Assume a constant birth rate $b(x_k) = b$. Due to (2) and (3) the overall birth rate $\mathcal{B}(x_k)$ is equal to b.

Step S1: Derive $d_i(x_k)$ from (5) and calculate $\delta(x_k)$.

Step S2: Simulate an exponentially distributed random number with mean $\frac{1}{b+\delta(x_k)}$. Simulate whether this jump is a birth or a death, where the corresponding probabilities are given by $\frac{b}{b+\delta(x_k)}$ and $\frac{\delta(x_k)}{b+\delta(x_k)}$.

Step S3_: Derive $\Psi_{k+1} = x_{k+1} \in \Omega_{k+1}$ or $\Psi_{k-1} = x_{k-1} \in \Omega_{k-1}$ as described in section 2.

Step S3: To reduce the number of sampling steps we keep the last samples of $\Psi^{(l)} = (\Psi_k^{(l)})_{k=1}^{k_{max}}$ in memory. When running the birth-death process we replace Step $S3\_$, where x or y are sampled from the corresponding priors, by the corresponding $\Psi_{(.)}^{(l)}$. This augmentation of the BD-MCMC algorithm reduces the number of sampling steps enormously. This can be motivated as follows: E.g. if a birth takes place and the last sample of $\Psi_{k+1}^{(l)}$ is used, then this corresponds to sampling the new parameters from (2) and running MCMC for another $\tau \geq 1$ steps. Despite the extra memory required to store $\Psi^{(l)}$, the performance of the sampler has been improved enormously.

To imbed the birth-death process we use the following algorithm:

Step M1: After $\Psi_k^{(j-1)}$ is samped by MCMC, run the birth-death process starting at $(j-1) = s$, with starting parameters $(k(s), \Psi_k(s)) = (k^{(j-1)}, \Psi^{(j-1)})$, for a time span of one period.

Step M2: The j-th sample of the Ψ_k is derived as usual by MCMC (Steps $G1$ and $G2$, after $\mathcal{R}$ is defined on every Ω_k in step $G0$), where $(k(s+1), \Psi_k(s+1))$ are used instead of $(k^{(j-1)}, \Psi^{(j-1)})$ in MCMC.

If the full conditional posterior distributions for each parameter give support to all parts of the parameter space, the algorithm M1-M2 results in an reducible Markov chain with stationary distribution $\pi(\Psi, k|\mathcal{Y}^N)$ (for a proof see Stephens (1997)[pp.84]).

4 Applications

This section presents two applications of the algorithm described in section 2 and section 3. The first example uses simulated data from a two-dimensional VAR(1) process, while the second example uses Austrian unemployment and real GDP data to estimate Okun's Law. In the first application we want to check the performance of BD-MCMC, while in the second example we compare the BD-MCMC estimates of the number states to an estimate of k by means of calculation the model likelihood (See Chib and Greenberg (1996), for the corresponding data see Sögner (2000)). Especially in the VAR(1) examples we demonstrate how the performance of BD-MCMC depends on the ability of MCMC (Steps G1-G2) to identify structural breaks within the sub-model with the true number of components. This property depends crucially on the model priors. If the priors are "wrong" or "too diffuse" either the sub-models with $k \geq 2$ die out quickly or the number of components sampled by BD-MCMC climbs up rapidly resulting in model over-fitting.
In both sub-sections the number of states has a truncated Poisson prior $\pi(k) \propto \frac{\lambda^k}{k!}$ with $1 \leq k \leq k_{max} = 5$. In all examples of this section a constant birth rate $b = 1$ is used. Experiments on

b show that the results do not change drastically if a $b \in [0.1, 10]$ is used. To keep the sampler as fast as possible we do not use a birth-rate depending on model parameters; we choose the priors such that the Randon-Nikodim derivative r is constant and equal to one. Estimates from BD-MCMC output are denoted by the superscript $\hat{}$. Provided with the estimate of the distribution of number of components $\hat{\pi}(k|\mathcal{Y}^N)$, derived from BD-MCMC output, inference on k can be carried out by a given loss function, taking the k with the highest $\hat{\pi}(k|\mathcal{Y}^N)$ or by the Bayes-factor. Given the prior distribution of the number of states $\pi(k)$, the Bayes-factor $BF_{l,k}$ is derived from:

$$BF_{l,m} := \frac{\pi(\mathcal{Y}^N|k=l)}{\pi(\mathcal{Y}^N|k=m)} = \frac{\pi(k=l|\mathcal{Y}^N)}{\pi(k=m|\mathcal{Y}^N)} \cdot \frac{\pi(k=m)}{\pi(k=l)} \ . \tag{6}$$

For the truncated Poisson prior used in this article, i.e. $\pi(k) \propto \lambda^k/k!$, $BF_{l,m} = \lambda^{m-l}l!/m!$. Convergence of the sampler is checked by comparing the distributions of sub-samples of the samples supposed to be derived from the posterior distribution. This is done by applying non-parametric techniques developed in Fan and Ullah (1999).

4.1 Simulated VAR(1) data

Within this subsection we investigate a two-dimensional first order autoregressive process:

$$\begin{aligned} y_{1,t} &= \tilde{\beta}^i_{1,0} + \tilde{\beta}^i_{1,1} y_{1,t-1} + \tilde{\beta}^i_{1,2} y_{2,t-1} + \varepsilon^i_{1,t} \\ &= z_t' \tilde{\beta}^i_1 + \varepsilon^i_{1,t} \\ y_{2,t} &= \tilde{\beta}^i_{2,0} + \tilde{\beta}^i_{2,1} y_{1,t-1} + \tilde{\beta}^i_{2,2} y_{2,t-1} + \varepsilon^i_{2,t} \\ &= z_t' \tilde{\beta}^i_2 + \varepsilon^i_{2,t} \ , \end{aligned} \tag{7}$$

where the error terms are independent and *iid* normal with mean zero and variances $(\tilde{\sigma}^2_1)^i$ and $(\tilde{\sigma}^2_2)^i$ respectively. We simulated $N = 40$ time steps of (7) where $k = 2$, $\tilde{\beta}^1_{1,0} = 0$, $\tilde{\beta}^1_{1,1} = 0.1$, $\tilde{\beta}^1_{1,2} = 0.05$, $\tilde{\beta}^1_{2,0} = 0.1$, $\tilde{\beta}^1_{2,1} = 0.2$, $\tilde{\beta}^1_{2,2} = 0$, $\tilde{\beta}^2_{1,0} = 0.5$, $\tilde{\beta}^1_{1,1} = 0.1$, $\tilde{\beta}^1_{1,2} = 0.3$, $\tilde{\beta}^1_{2,0} = 0.5$, $\tilde{\beta}^1_{2,1} = 0.3$, $\tilde{\beta}^1_{2,2} = 0.1$, $(\tilde{\sigma}^2_1)^1 = 0.02$, $(\tilde{\sigma}^2_2)^2 = 0.02$, $(\tilde{\sigma}^2_1)^2 = 0.03$ and $(\tilde{\sigma}^2_1)^2 = 0.01$. The state process I_t jumps at $t = 10$ from state 1 to state 2 and at period $t = 30$ from state 2 to state 1. The dimension of y_t is $p = 2$.[1] In all samples within this subsection we use permutation sampling where the restriction $\mathcal{R}$ is put on the intercept of the first component, i.e. $\beta^1_{0,1} < \ldots < \beta^k_{0,1}$. With this specification no label switching has been observed. In the following we analyze the dependence of $\hat{\pi}(k|\mathcal{Y}^N)$ on the underlying model priors. We demonstrate how "wrong" model priors could result in sampling either $k = 1$ or k_{max} too often, given the true number of states $\tilde{k}$. These examples should confront the reader with the problems arising with BD-MCMC and demonstrate that BD-MCMC needs a lot of tuning.

Prior Assumptions and the Performance of BD-MCMC: In the first step we use very precise prior information and look at the relative frequencies of the number of states k after a burn-in

[1] In these simulation runs only 40 time steps of the VAR process are simulated. This relatively low number of data points is due to the relatively low computing power and memory of my PC, which has a Pentium 200 processor with a working memory of 48 KB. The software used is MATLAB 5.1.

period of 1000 samples. The last sample for each k is kept in memory as described in section 3. Before the birth-death process is started and sampling step $[S3]$ is applied, we use standard MCMC (sampling steps $[G0 - G2]$) to get 500 runs for every fixed k, $k = 1, \ldots, 5$. After these simulation runs the birth-death process (sampling steps $[M1]$ and $[M2]$) is started. These prior runs have the advantage that if MCMC is able the discriminate between different states the samples drawn from this algorithm are already close to samples form the posterior given the underlying data. If we would start the BD-MCMC algorithm immediately, the likelihoods for sub-models with $k \geq 1$ usually had a low likelihood compared to the sub-model with one state. This resulted in immediate deaths of the components as well as of births taken from the priors. Therefore the argument in Stephens (1998) that 'bad components' die and 'good components' are kept alive by BD-MCMC cannot be observed within the models we consider in this article. Our experience was that if we start with diffuse priors the components will die out very fast and new components sampled from the priors will die out as well. Therefore, we started with sampling parameters for k fixed before the birth-death process starts. Nevertheless if MCMC cannot almost approximately identify structural breaks within this prior runs within the sub-model with the true number of states, this step-wise procedure will indeed fail as described in some of the following examples.
Let us start with the following priors:

PD: For $k = 2$: $b^i_{0,\iota} = \tilde{\beta}^i_\iota$, while with $k = 1$ we used $b_{0,\iota} = \tilde{\beta}^1_\iota$. In the other sub-models with $k = 3, 4, 5$, the first two components are given by the priors of the model with $k = 2$. For $i = 3, \ldots, k$, we use $b^i_{0,\iota} = (0, 0, 0)'$. For $i = 1, \ldots, k$, $k = 1, \ldots, k_{max}$ let $B^i_{0,\iota} = 0.01\mathrm{I}_3$, where I_d is the identity matrix of dimension d. $\nu^i_{0,\iota} = 0.02$ and $D^i_{0,\iota} = 1$ for $\iota = 1, 2$. For the distribution of the switching parameters η we assume $e_{0,i\iota} = 1$ for $i \neq \iota$ and $e_{0,ii} = 4$.

PM: Alternatively, let $b^i_{0,\iota} = \left(\mu_{y,\iota} - 1^i \; i/k \; \sqrt{\mathrm{VAR}_{y,\iota}}, 0, 0\right)'$, where $i = 1, \ldots, k$, $\iota = 1, 2$, $\mu_{y,\iota}$ is the sample mean of the ι-th component of y_t and $\mathrm{VAR}_{y,\iota}$ is the sample variance of the ι-th component of y_t. $B^i_{0,\iota}$ is equal to $0.01\ \mathrm{I}_3$. The parameters of the inverse gamma distributions of the variance terms are given by: $\nu^i_{0,\iota} = 0.02$ and $D^i_{0,\iota} = 1$ for all $i = 1, \ldots, k$ and $\iota = 1, 2$. For the distribution of the switching parameters η we assume $e_{0,ij} = 1$ for $i \neq \iota$ and $e_{0,ii} = 4$.

Table 2 presents the estimates of the posterior distribution of the number of states $\pi(k|\mathcal{Y}^N)$. By calculating the relative frequencies h_i of states $i = 1, \ldots, k_{max}$ we derive $\hat{\pi}(k = i|\mathcal{Y}^N) =: h_i$ given the parameter of the prior λ. $MEAN_k$ is the mean value over the number of states from BD-MCMC output, with standard deviation SD_k. These are taken over 1000 samples after a burn-in phases of 1000 steps. In some runs of BD-MCMC a burn-in of even more than 1000 steps was necessary. However, in these runs it was easy to see (even by simple plots of the model parameters) that the sampler had not converged and longer burn-in periods were necessary. The samples of k at sampling step j for the BD-MCMC model with prior PD with $\lambda = 2$ are presented in Figure 1 (this prior assumption will be called $PD(\lambda = 2)$ in the further analysis). The estimates from BD-MCMC output of the parameters β_2, σ^2_2 and η_2 are very close to their true values with the precise priors ($PD(\lambda)$, $\lambda = 1, \ldots, 5.$). The cases with priors $PM(\lambda)$ result in parameters that are within $\pm$ one time the standard deviation of the corresponding parameter estimated from MCMC output. However, the standard deviations with $PM(\lambda)$ are

large. I.e. the estimates of the parameters remain within $\pm$ the standard deviation but the distance between the estimate and the true parameter value may become very high. From Table 2 we conclude that with priors $PD(\lambda)$, we are able to infer the true number of states correctly using the number of states with the highest posterior probability or the Bayes-factor. Considering the results with the priors $PM(\lambda)$ figure a totally different result. With these priors the sampler is not able to detect the structural breaks within the data. This results in low marginal likelihoods for sub-models with $k \geq 2$ compared to the samples with $PD(\lambda)$. Therefore, sub-models with more than one component die out quickly, resulting in an inference of $\hat{k} = 1$. Thus, if "wrong" priors are used, BD-MCMC will result in an incorrect number of states. Nevertheless, we have to note that with empirical data, we do not know what are the "right" priors – or do sub-models with $k \geq 2$ die since the true number of states is simply one? Nevertheless, the examples with priors $PM(\lambda)$ should demonstrate, the sensitivity of the posterior distribution of states with respect to the prior assumptions. We conclude that if MCMC (Steps $G1 - G3$) is able to detect structural breaks within the sub-model with the true number of components BD-MCMC performs well. If this were not the case, sub-models with $k \geq 2$ die quickly due to their low likelihoods. This has been checked by starting BD-MCMC immediately after we have generated one sample for each sub-model. The results are presented in Table 3 and motivate why $[Step3\_]$ has been replaced by $[Step3]$ in the BD-MCMC algorithm.

	$PD(\lambda)$					$PM(\lambda)$				
λ	1	2	3	4	5	1	2	3	4	5
h_1	0.3250	0.1263	0.0088	0.2112	0.0437	0.9750	0.9650	0.9463	0.9337	0.8962
h_2	0.6700	0.8512	0.9675	0.7662	0.9225	0.0163	0.0325	0.0537	0.0650	0.0825
h_3	0.0050	0.0150	0.0238	0.0213	0.0325	0.0050	0.0025	0.0000	0.0013	0.0113
h_4	0.0000	0.0000	0.0000	0.0013	0.0013	0.0025	0.0000	0.0000	0.0000	0.0013
h_5	0.0000	0.0075	0.0000	0.0000	0.0000	0.0013	0.0000	0.0000	0.0000	0.0088
$MEAN_k$	1.7428	1.8634	2.0180	1.8648	1.9980	1.0333	1.0473	1.0660	1.0773	1.1386
SD_k	0.4653	0.4389	0.1761	0.5161	0.3214	0.2399	0.2499	0.2588	0.2769	0.4627

Table 2: Estimates of the posterior distribution of the number of states $\hat{\pi}(k|\mathcal{Y}^N)$ with $PD(\lambda)$ and $PM(\lambda)$, $\tilde{k} = 2$.

In the next step we change the prior assumptions on β and σ^2. In $PDIFF$ we use diffuse priors, i.e. $b^i_{0,\iota} = (0,0,0)'$, $B_0 = \mathrm{I}_3$, $\nu^i_{0,\iota} = 2$ and $D^i_{0,\iota} = \max\{\mathrm{VAR}\,(y_\iota)\}$, where $\mathrm{VAR}\,(y_\iota)$ is the sample variance of the ι-th component of the data. In this case the MCMC algorithm is not able to detect the structural breaks. A careful look at the latent variable I^N_k figures this problem occurring with MCMC applications. Subplot (a) of Figure 2 presents the samples of k with $PDIFF$. One particular sample of the latent variable I^N_2 is presented in subplot (b), figuring that the sampler is not able to detect the breaks at $t = 10$ and $t = 30$. This results in low likelihoods $\mathcal{L}(\mathcal{Y}^N|\Psi_k)$ of sub-models with $k > 1$, resulting in deaths of sub-models with more than one component. The posterior probabilities $\hat{\pi}(k|\mathcal{Y}^N)$ of the number of states k is presented in the second column of Table 4.
In $PD(\zeta)$ and $PM(\zeta)$ the prior assumption on the parameter ζ are altered, while $\lambda = 2$. We define $B^i_{0,\iota} = \zeta\ \mathrm{I}_3$, where the parameter ζ is set to $\zeta = 0.05$, $\zeta = 0.1$ and $\zeta = 1$. Table 4 presents

λ	1	2	3	4	5
h_1	0.9663	0.9463	0.9075	0.9038	0.8750
h_2	0.0325	0.0512	0.0887	0.0900	0.1088
h_3	0.0013	0.0025	0.0037	0.0063	0.0138
h_4	0.0000	0.0000	0.0000	0.0000	0.0025
h_5	0.0000	0.0000	0.0000	0.0000	0.0000
$MEAN_k$	1.0227	1.0533	1.0966	1.1113	1.1392
SD_k	0.1533	0.2363	0.3276	0.3468	0.3898

Table 3: Estimates of the posterior distribution of the number of states $\hat{\pi}(k|\mathcal{Y}^N)$ with $PD(\lambda)$ and $PM(\lambda)$, $\tilde{k} = 2$, BD-MCMC starts at immediately.

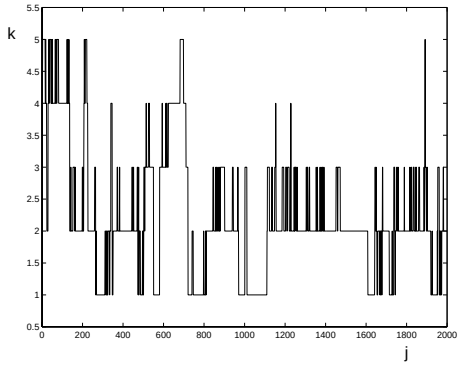


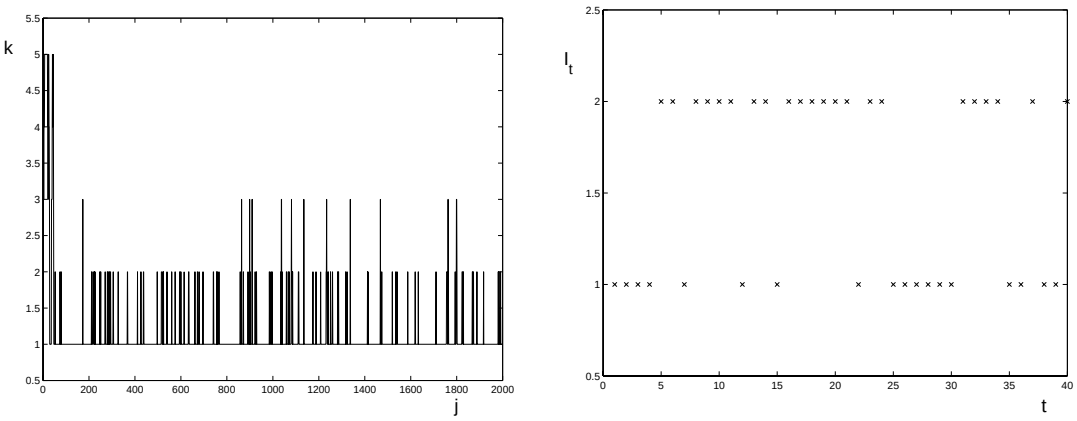
Figure 1: BD-MCMC output on k with $PD(\lambda = 2)$.

the results, where we conclude that the higher ζ the worse the performance of BD-MCMC. If ζ rises, the prior variance of the parameters β increases. If this is the case MCMC becomes more and more unable to identify the structural breaks within the time series. This effect starts approximately with $\zeta \geq 0.1$ with priors $PD(\zeta)$. Additionally we recommend to check whether label switching occurs with k fixed. If this is the case different restrictions $\mathcal{R}$ should be used to check whether label switching disappears. In the ongoing analysis label switching was not observed in the sub-model with $k = \tilde{k} = 2$. Therefore, we leave $\mathcal{R}$ unchanged. With $PM(\zeta)$ an increase in ζ does not improve the performance of the sampler as already observed in Table 2, where a $\zeta = 0.01$ is used. Therefore, we conclude from our experiments with simulated data, that BD-MCMC works well if the priors are "well" specified and MCMC is able to detect structural breaks.

VAR Model without Switching: In contrast to the above analysis we want to check the performance of BD-MCMC in simulated data with no structural breaks, i.e. $\tilde{k} = 1$. The data used are 40 observations from the VAR process (7) with the parameters for $i = 1$. Additionally we want to highlight the problem of "wrong priors" and model over-fitting. E.g. if the expectation of σ_i^2 with respect to the prior distribution becomes very small this results in model over-fitting. From the above analysis we might conclude that e.g. "keeping the expected value and the variance with respect to the prior distribution of $(\sigma_i^2)^i$ small results in a good performance of

	$PDIFF$	$PD(\zeta)$			$PM(\zeta)$		
ζ		0.05	0.1	1	0.05	0.1	1
h_1	0.8791	0.1525	0.9337	0.9287	0.9275	0.9087	0.9287
h_2	0.1109	0.8213	0.0475	0.0625	0.0688	0.0762	0.0550
h_3	0.0090	0.0262	0.0050	0.0088	0.0013	0.0150	0.0088
h_4	0.0010	0.0000	0.0113	0.0000	0.0025	0.0000	0.0075
h_5	0.0000	0.0000	0.0025	0.0000	0.0000	0.0000	0.0000
$MEAN_k$	1.1339	1.8694	1.0933	1.1093	1.0906	1.1852	1.0786
SD_k	0.3830	0.4183	0.3907	0.3882	0.3242	0.4820	0.3252

Table 4: Estimates of the posterior distribution of the number of states $\hat{\pi}(k|\mathcal{Y}^N)$ with $PD(\zeta)$ and $PM(\zeta)$, $\tilde{k} = 2$.



(a) $PDIFF$: k (b) $PDIFF$: I_2^N

Figure 2: Samples of I^N.

BD-MCMC". In the following we show that this claim cannot be supported and demonstrate that this kind of "wrong priors" results in model over-fitting. In a first stage of the analysis we use the diffuse prior assumption $PDIFF$ and the priors $PD(\zeta)$ and $PM(\zeta)$. The estimates of the posterior distribution are presented in Table 5, where the sampler works well with $PD(\zeta)$ and $PM(\zeta)$. Even with the diffuse prior assumptions BD-MCMC performs well.
To demonstrate how "wrong priors" can result in too many components we change the prior assumptions on the variance, while the other parameters correspond to those of $PDIFF$. The inverse gamma prior with parameters $\nu_{0,\iota}^i \geq 2$ and $D_{0,\iota}^i$ results in $\mathbb{E}((\sigma_\iota^2)^i) = D_{0,\iota}^i/(\nu_{0,\iota}^i - 1)$ and $\mathrm{VAR}\left((\sigma_\iota^2)^i\right) = (D_{0,\iota}^i)^2/((\nu_{0,\iota}^i - 1)^2(\nu_{0,\iota}^i - 2))$. Altering the expectations of $(\sigma_\iota^2)^i$ (and its variance) result in high likelihoods for the sub-models with a high number of components, since due to the small bandwidths in the normal kernels of the parameters β_k – given by $\kappa_i(\sigma_\iota^2)^i$ – we put much more weight to a good fit with the data compared to normal densities with a higher bandwidth. With 40 data points from the VAR(1) process the sub-models with four and five components are sufficient to fit the data almost very well. This results – in contrast

	$PDIFF$	$PD(\zeta)$			$PM(\zeta)$		
ζ		0.05	0.1	1	0.05	0.1	1
h_1	0.9033	0.9487	0.9150	0.0000	0.9612	0.9663	0.3212
h_2	0.0600	0.0475	0.0750	0.0000	0.0388	0.0338	0.0125
h_3	0.0208	0.0038	0.0100	0.0000	0.0000	0.0000	0.0325
h_4	0.0017	0.0000	0.0000	0.0375	0.0000	0.0000	0.1787
h_5	0.0142	0.0000	0.0000	0.9625	0.0000	0.0000	0.4550

Table 5: Estimates of the posterior distribution of the number of states $\hat{\pi}(k|\mathcal{Y}^N)$, with $PDIFF$, $PD(\zeta)$ and $PM(\zeta)$, $\tilde{k} = 1$.

to the problem of the detection of structural breaks in the problem of over-fitting. The effect of over-fitting is presented in Table 6. In this analysis we use $PR(\rho, \varrho)$, where $b^i_{0,\iota} = (0,0,0)'$, $B_0 = 0.5\ \mathrm{I}_3$, $\nu^i_{0,\iota} = 2$ and $D^i_{0,\iota} = \mathrm{VAR}\,(y_\iota)$, where $\mathrm{VAR}\,(y_\iota)$ is the sample variance of the ι-th component, $\zeta = 0.5$ and $\lambda = 1$. The hyperparameters ρ and ϱ control the expectation and the variance of $(\sigma^2_\iota)^i$, $\iota = 1, 2$, $i = 1, \ldots, k$. More precisely we use $\nu^i_{0,\iota} = 2\varrho$ and $D^i_{0,\iota} = \mathrm{VAR}\,(y_\iota)\rho\varrho$ resulting in $\mathbb{E}((\sigma^2_\iota)^i)) = \rho\mathrm{VAR}\,(y_\iota)$ and $\mathrm{VAR}\,((\sigma^2_\iota)^i)) = (\rho\varrho\mathrm{VAR}\,(y_\iota))^2/((2\varrho-1)^1)^2(2\varrho-2))$. In Table 6 we use $\varrho = 3$ while ρ is altered, resulting in too small variance terms causing wrong and too high estimates of the number of components. Sampling a high number of components could also be caused by empty states in I^N_k. E.g. the sampler might generate I^N_5 with four states empty. Within this analysis this effect has not been observed. To overcome the problem of over-fitting we recommend to check whether the k sampled by BD-MCMC climbs up rapidly to a high k. If this high number of components is very unreasonable the priors on the variance should be changed. Using a smaller k_{max} need not solve this problem since if k_{max} is still larger than $\tilde{k}$ the BD-MCMC too often samples k_{max} due to the "wrong priors" while the inference on the number of components would still be wrong.

ρ	0.010	0.025	0.050	0.100
h_1	0.0000	0.1225	0.2925	0.2488
h_2	0.0000	0.0025	0.0150	0.0100
h_3	0.0000	0.0050	0.0113	0.0262
h_4	0.0688	0.0075	0.0437	0.0788
h_5	0.9313	0.8625	0.6375	0.6262
$MEAN_k$	4.8947	4.7182	3.6449	4.3691
SD_k	0.3113	0.9999	1.8253	1.3730

Table 6: Estimates of the posterior distribution of the number of states $\hat{\pi}(k|\mathcal{Y}^N)$ with $PR(\rho, \varrho))$, $\tilde{k} = 1$.

Thus we conclude that BD-MCMC can be efficient especially if no structural breaks are in the underlying data set. Nevertheless this sub-section should highlight the dependence of the performance of BD-MCMC on the prior assumptions. First BD-MCMC fails if the priors

are definitely wrong such as in the $PM(.)$ examples. Secondly, the prior assumptions on the variance terms have countervailing effects. If the expectation is too high the sampler hardly detects structural breaks while with a too small expectation of the variance over-fitting occurs. Since we usually do not know what is a "good" and what is a "bad" prior in an empirical investigation, we recommend to perform a sensitivity analysis by taking short runs of BD-MCMC with different priors to derive a first impression of the properties of the sequence of k sampled by BD-MCMC. Thus implementing BD-MCMC needs a lot of experience and the problems discussed in this section should not be neglected.

4.2 Okun's Law

Okun's law states that the annual growth in unemployment Δu_t is a function of the growth in real gross domestic product ΔGDP_t. For quarterly data this results in $\Delta u_t := u_t - u_{t-4}$, $\Delta GDP_t := GDP_t - GDP_{t-4}$, and $\mathcal{Y}^N = (\Delta u_t \ , \ \Delta GDP_t)$. In this article we consider the following model:

$$\begin{aligned} \Delta u_t &= \beta_0^i + \beta_1^i \Delta GDP_t + \beta_2^i \Delta u_{t-1} + \varepsilon_t^i \ , \\ &= \beta^i z_t + \varepsilon_t^i \end{aligned} \tag{8}$$

where the index $i \in \{1, 2, ..., k\}$ is the label of the state I_t. $\beta_{0,i}$ is the intercept if state $[I_t = i]$ is realized. $\beta_{1,i}$ shows the dependence of Δu_t on GDP growth ΔGDP_t. $\varepsilon_{t,i}$ is an independent identically distributed (iid) normal variable with zero mean and variance σ_i^2, i.e. $\varepsilon_{t,i} \sim iid\, \mathcal{N}(0, \sigma_i^2)$. Due to autocorrelation in the residuals ε_t^i – in a model without β_2^i – the lagged variable Δu_{t-1} has been added to eliminate autocorrelation in the residuals. β^i is the vector of regression parameters, i.e. $\beta^i := (\beta_0^i, \ \beta_1^i, \ \beta_2^i)$. The prediction variables are $z_t := (1 \ , \ \Delta GDP_t \ , \ \Delta u_{t-1})$, the response variable is $y_t = \Delta u_t$. The data used are 80 observations of quarterly Austrian real GDP and unemployment from 1976 to 1995 resulting in $N = 75$. First differences Δu_t and ΔGDP_t are required to derive stationary time series. Δu_t is measured in percents, while ΔGDP_t is measured in billions of Austrian Schillings.

Prior Distributions: In the following analysis the estimates of the parameters $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\eta}$ and $\hat{k}$ are performed for different prior assumptions to investigate the dependence of $\hat{k}$ on the corresponding priors. In all the following models we use $\lambda = 2$ in the truncated Poisson prior of the number of states.

Strong Priors: The distribution of the parameters is given by: $e_{0,il} = 4$ for $i = l$ and $e_{0,il} = 3$ for all $i \neq l$. Nevertheless, this prior assumption is relatively vague compared to some other investigations, for example McCulloch and Tsay (1994) use a $\mathcal{D}(45, 2)$ prior in a model with two states. For the parameters of the state specific priors we use the ordinary least squares estimator and the corresponding variance of the residuals s_y^2 in the following way: for all i, $i = 1, \ldots, k$ and all k, $k = 1, \ldots, k_{max}$: $b_0^i = (0.2682, \ -0.0246, \ 0.7110)'$, $B_0^i = 0.1 \cdot I_3$, $\nu_0 = 2$, $D_0^i = s_y^2$, where $s_y^2 = 0.0611$. In the model with *semi strong priors* we used $B_0 = I_3$ and $e_{0,il} = 1$, $\forall\ i, l$, while in the model with *weak priors* D_0 was additionally altered to $D_0 = 1$. In the model with *diffuse priors* we additionally set $b_0 = (0, 0, 0)'$.

Results: The estimates of the state dependent parameters are derived from 500 samples from the posterior after a burn-in of 1525 sweeps. The estimates $\hat{\beta}$ and $\hat{\sigma}^2$ are derived from taking the

	BDMCMC				MoL	
	Strong	*Semi Strong*	*Weak*	*Diffuse*	*Strong*, $\lambda = 1$	*Strong*,$\lambda = 2$
h_1	0.9750	0.9595	0.9287	0.9672	0.9929	0.9856
h_2	0.0231	0.0347	0.0597	0.0289	0.0070	0.0139
h_3	0.0019	0.0058	0.0058	0.0039	0.0001	0.0005
h_4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
h_5	0.0000	0.0000	0.0058	0.0000	0.0000	0.0000

Table 7: BD-MCMC and Model Likelihood (MoL) estimates of the posterior distribution of the number of states $\hat{\pi}(k|\mathcal{Y}^N)$.

mean value over the samples from the posterior. $\hat{k}$ is derived from the Bayes-factor. Considering the posterior probabilities presented in Table 7 and using $\lambda = 2$ in (6) results in $\hat{k} = 1$ for all prior assumptions. In Table 7 these posterior probabilities derived by BD-MCMC are compared to estimates of $\pi(k|\mathcal{Y}^N)$ by means of the model likelihood (MoL). Applying candidate's formula, as described in Chib (1995) and Frühwirth-Schnatter (1999b), results in estimates of the model likelihood $\hat{\mathcal{L}}_k(\mathcal{Y}^N)$, with standard deviations in parentheses, of -114.145 (0.017), -118.408 (0.066), -121.373 (0.251), -121.910 (1.068) and -122.446 (3.092) for models with one to five states respectively. Applying Bayes rule results in $\pi(k|\mathcal{Y}^N) \propto \mathcal{L}_k(\mathcal{Y}^N)\pi(k)$. Using the truncated Poisson prior with parameter λ and a straightforward normalization yields the estimates of Table 7. The BD-MCMC estimates of the regression parameters and the variance are presented in Table 8. The terms in parentheses are the standard deviations of the corresponding estimate. Next we investigate the unemployment–GDP relationship as stated in economic textbooks, where a 2%–3% increase in GDP above the *normal growth* is supposed to cause unemployment to decline by 1% point for US data (See Romer (1996)). We express the normal growth by the mean growth rate ($G := 1/N \ \sum_{t=1}^{N}(\Delta GDP_t/GDP_{t-4}) \ \cdot \ [100\%]$), which is equal to 2.3276% for the underlying data. Since the autoregressive variable Δu_{t-1} has been included in the regression model (8) we derive the annual excess growth rate G_1 which is necessary to decrease the unemployment rate by 1% point per year. The estimates are presented in Table 9, where the mean GDP ($1/(N+4) \ \sum_{t=1-4}^{N} GDP_t$) was inserted for GDP_{t-4} to calculate the extra growth rate. Additionally, we want to investigate the necessary excess growth rate G_2 to decrease unemployment by 1% point in one particular period. Considering the regression model, this question depends on the unemployment rate of the last period. Nevertheless, ΔGDP_t is easily derived if we insert the mean unemployment growth ($1/N \ \sum_{t=1}^{N} \Delta u_t = 0.2737$) into (8). The results on G_2 are presented in Table 9.

5 Conclusions

In this article we applied the method of BD-MCMC to switching models. By this method the full posterior distribution of parameters including the number of states is derived within the sampler. BD-MCMC has been applied to simulated VAR data and to unemployment-GDP data. From our experience with BD-MCMC, we conclude that the problems occurring with RJ-

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}^2$
'Strong Priors'			
0.2697	-0.0246	0.7100	0.0643
(0.0468)	(0.0053)	(0.0472)	(0.0116)
'Semi Strong Priors'			
0.2723	-0.0249	0.7108	0.0650
(0.0465)	(0.0052)	(0.0481)	(0.0120)
'Weak Priors'			
0.2790	-0.0238	0.4964	0.1017
(0.0582)	(0.0064)	(0.0604)	(0.0157)
'Diffuse Priors'			
0.2858	-0.0244	0.4939	0.1024
(0.0579)	(0.0067)	(0.0612)	(0.0168)

Table 8: Estimated parameters

	Strong Priors	*Semi Strong Priors*	*Weak Priors*	*Diffuse Priors*
G_1	4.4962	4.4339	7.6229	7.4910
G_2	4.1985	4.0826	5.0621	4.7882

Table 9: Excess growth rates

MCMC cannot be solved by BD-MCMC. The performance of BD-MCMC preliminarily depends on the performance of MCMC within BD-MCMC. If the sampler is able to detect structural breaks and no label-switching is generated by the sampler then BD-MCMC performs well. Otherwise the inference is definitely wrong. Thus, BD-MCMC needs a lot of tuning to obtain correct estimates. From these reasons a standard implementation cannot be recommended without checking in properties of MCMC within BD-MCMC.

A Appendix: Proof of Lemma 1

In this section we show that the birth-death process specified in section 2 has a stationary distribution. In this proof we follow Stephens (1998) who proved the detailed balance condition for the standard Bayesian mixture model. The *detailed balance condition* providing a sufficient condition for the existence of an invariant measure is due to Preston (1976). Consider a birth-death process on state space $\Omega = \bigcup_{k\geq 1} \Omega_k$, Ω_k are disjoint, where $\mathcal{F}_k$ is the corresponding sigma field. Let $\mathcal{F}$ be the sigma field generated by $\mathcal{F}_k$. We consider a jump process with state space $(\Omega, \mathcal{F})$, where $z \in \Omega_k$ can only jump to Ω_{k+1} (*birth*) or to Ω_{k-1} (*death*). Next, we define the set F_k by $F_k = F \cap \Omega_k$, where $F \in \mathcal{F}$. If μ is a measure on $(\Omega, \mathcal{F})$ the restriction on $\mathcal{F}_k$ is μ_k. The functions $\mathcal{B}(z)$, $\delta(z) : \Omega \to \mathrm{R}^+$ are assumed $\mathcal{F}$ measurable and $\delta(z) = 0$ for all $z \in \Omega_0$. In section 2 $\mathcal{B}(z)$ and $\delta(z)$ are the overall birth and the death rate. $\alpha(z) = \alpha(z)^b + \alpha(z)^d$ is

the normalizing constant. To describe the transition of the model parameters with births and deaths we have defined transition kernels $K_{\mathcal{B}}^k(z, F_{k+1})$ and $K_{\delta}^k(z, F_{k-1})$ in section 2 for births and deaths respectively. Let $z \in \Omega_k$ and $k \geq 1$, then the birth-death process is described by the probability kernel $K : \Omega \times \mathcal{F} \to \mathrm{R}^+$ where

$$K(z, F) = \frac{\mathcal{B}(z)}{\alpha(z)} K_{\mathcal{B}}^k(z, F_{k+1}) + \frac{\delta(z)}{\alpha(z)} K_{\delta}^k(z, F_{k-1}) \ . \tag{9}$$

Following Preston (1976) and Stephens (1998), a birth-death process exhibits an invariant measure $\tilde{\mu}$ on $(\Omega, \mathcal{F})$, if $\int \alpha(z) d\mu < \infty$ and the *detailed balance conditions* (10) and (11) are satisfied:

$$\int_F \mathcal{B}(z) d\tilde{\mu}_k(z) = \int_{\Omega_{k+1}} \delta(y) K_{\delta}^{k+1}(y, F) d\tilde{\mu}_{k+1}(y) \tag{10}$$

for $k \geq 1$ and $F \in \mathcal{F}_k$, i.e. $F \subset \Omega_k$.

$$\int_G \delta(y) d\tilde{\mu}_{k+1}(y) = \int_{\Omega_k} \mathcal{B}(z) K_{\mathcal{B}}^k(z, G) d\tilde{\mu}_k(z) \tag{11}$$

for any $k \geq 0$ and $G \subset \Omega_{k+1}$. Let $x \in \Omega_{k-1}$, $z \in \Omega_k$ and $y \in \Omega_{k+1}$ and consider the specification of the birth death process in section 2. Since (3) and (4) are assumed to exist $\int \alpha d\mu < \infty$ is fulfilled. In the case of a birth z^b are the new parameters added to $z \in \Psi_k$, while in the case of a death z^d represents the eliminated parameters; the labeling of the parameters is not taken into account on Ω_k. Since, $I_k^N \in \mathcal{I}_k$ is replaced by $I_{k+1}^N \in \mathcal{I}_{k+1}$ or $I_{k-1}^N \in \mathcal{I}_{k-1}$, the new parameters $y \in \Omega_{k+1}$ and $x \in \Omega_{k-1}$ are given by $y = z \setminus i_k^N \cup z^b$ and $x = z \cup i_{k-1}^N \setminus z^d$, in the case of a birth and the case of a death, respectively. This yields:

$$\begin{aligned}
\mathcal{B}(z) K_{\mathcal{B}}^k(z, F_{k+1}) \quad = \quad & \int_{z^b : z \setminus i_k^N \cup z^b \in F_{k+1}} b(z) \pi(\beta_{k+1}^{k+1}, (\sigma_{k+1}^2)^{k+1} | k+1) \left(\frac{\Gamma(kv_2 + v_1)}{\Gamma(v2)\Gamma(v1)} \right)^k \\
& \cdot \eta_{1,k+1}^{v2-1} \cdots \eta_{k,k+1}^{v2-1} (1 - \eta_{1,k+1})^{kv2-1} \cdots (1 - \eta_{k,k+1})^{kv2-1} \\
& \cdot \frac{\Gamma(kv_2 + v_1)}{\Gamma(v2)^k \Gamma(v1)} \eta_{k+1,1}^{v1-1} \eta_{k+1,2}^{v2-1} \cdots \eta_{k+1,k+1}^{v2-1} \pi(i^N | \eta_{k+1}, k+1) \\
& \cdot d\beta d\sigma^2 d\eta_{1,k+1} \cdots d\eta_{k,k+1} d\eta_{k+1,1} \cdots d\eta_{k+1,k+1} di_{k+1}^N \ ,
\end{aligned} \tag{12}$$

and

$$\delta(z) K_{\delta}^k(z, F_{k-1}) \quad = \quad \int_{i_{k-1}^N : z \setminus z^d \cup i_{k-1}^N \in F_{k-1}} \sum_{i=1}^{k} d_i(z) \pi(i^N | \eta_{k-1}, k-1) di_{k-1}^N \ . \tag{13}$$

Let us pause on the problem that each sub-model demands for a restriction $\mathcal{R}$ on the parameters to get the model identified. Therefore the parameters Ψ_k are sampled from a restricted space $\Omega_{k,\mathcal{R}}$ where the labeling of the states is taken into account. Parameters from $\Omega_{k,\mathcal{R}}$ will be denoted with a subscript $\mathcal{R}$. Let $\Omega_{\mathcal{R}} = \bigcup_{k \geq 1} \Omega_{k,\mathcal{R}}$ with sigma field $\mathcal{F}_{\mathcal{R}}$. The prior and the posterior probability measures on $(\Omega_{k,\mathcal{R}}, \mathcal{F}_{k,\mathcal{R}})$ are $\mu_{k,\mathcal{R}}$ and $\tilde{\mu}_{k,\mathcal{R}}$ respectively. Let μ and $\tilde{\mu}$ be probability measures on Ω induced by $\mu_{\mathcal{R}}$ and $\tilde{\mu}_{\mathcal{R}}$ respectively, the restrictions to Ω_k are μ_k and

$\tilde{\mu}_k$ respectively. Then, since $\Omega_{k,\mathcal{R}}$ and Ω_k only differ in the labeling of the parameters, we have for every $\mathcal{F}$ measurable function $g(.)$:

$$\int_{\Omega_k} g(z)d\mu_k(z) = \int_{\Omega_{k,\mathcal{R}}} g(z_{\mathcal{R}})d\mu_{k,\mathcal{R}}(z_{\mathcal{R}}) \tag{14}$$

By applying the Bayes theorem we additionally get:

$$\begin{aligned}\int_{\Omega_k} g(z)d\tilde{\mu}_k(z) &= \int_{\Omega_{k,\mathcal{R}}} g(z_{\mathcal{R}})d\tilde{\mu}_{k,\mathcal{R}}(z_{\mathcal{R}}) \\ &= \int_{\Omega_{k,\mathcal{R}}} g(z_{\mathcal{R}})f(\mathcal{Y}^N|z_{\mathcal{R}},k)d\mu_{k,\mathcal{R}}(z_{\mathcal{R}}) \\ &= \int_{\Omega_k} g(z)f(\mathcal{Y}^N|z,k)d\mu_k(z)\ . \end{aligned} \tag{15}$$

In the next step we derive the following interrelationship between the prior for parameters $y \in \Omega_{k+1,\mathcal{R}}$ and $z \in \Omega_{k,\mathcal{R}}$:

$$\begin{aligned}\pi(i^N|\eta_k,k)di_k^N d\mu_{k+1,\mathcal{R}}(y) &= \frac{\pi(k+1)}{\pi(k)}r_{k+1}\pi(\beta_{k+1}^{k+1},(\sigma_{k+1}^2)^{k+1}|k+1)\left(\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)\Gamma(v1)}\right)^k \\ &\cdot\eta_{1,k+1}^{v2-1}\cdots\eta_{k,k+1}^{v2-1}(1-\eta_{1,k+1})^{kv2-1}\cdots(1-\eta_{k,k+1})^{kv2-1} \\ &\cdot\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)^k\Gamma(v1)}\eta_{k+1,1}^{v1-1}\eta_{k+1,2}^{v2-1}\cdots\eta_{k+1,k+1}^{v2-1} \\ &\cdot\pi(i^N|\eta_{k+1},k+1)d\beta d\sigma^2 d\eta_{1,k+1}\cdots d\eta_{k,k+1} \\ &\cdot d\eta_{k+1,1}\cdots d\eta_{k+1,k+1}di_{k+1}^N d\mu_{k,\mathcal{R}}(z)\ . \end{aligned} \tag{16}$$

where $\pi(k)$ is the prior on the number of states; (16) is derived from

$$\begin{aligned}d\mu_{k+1,\mathcal{R}}(y) &= \pi(k+1)\pi(\beta_{k+1},\sigma_{k+1}^2|k+1)\pi(\eta_{k+1}) \\ &\cdot\pi(i^N|\eta_{k+1},k+1)d\beta_{k+1}d\sigma_{k+1}^2 d\eta_{k+1}di_{k+1}^N \qquad (17) \\ &= \pi(k+1)r_{k+1}\pi(\beta_k,(\sigma^2)_k|k)\pi(\eta_k) \\ &\cdot\pi(\beta_{k+1}^{k+1},(\sigma_{k+1}^2)^{k+1}|k+1)\left(\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)\Gamma(v1)}\right)^k \\ &\cdot\eta_{1,k+1}^{v2-1}\cdots\eta_{k,k+1}^{v2-1}(1-\eta_{1,k+1})^{kv2-1}\cdots(1-\eta_{k,k+1})^{kv2-1} \\ &\cdot\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)^k\Gamma(v1)}\eta_{k+1,1}^{v1-1}\eta_{k+1,2}^{v2-1}\cdots\eta_{k+1,k+1}^{v2-1}\pi(i^N|\eta_{k+1},k+1) \\ &\cdot d\beta d\sigma^2 d\eta_{1,k+1}\cdots d\eta_{k,k+1}d\eta_{k+1,1}\cdots d\eta_{k+1,k+1}di_{k+1}^N \qquad (18) \\ &= \frac{\pi(k+1)}{\pi(k)}r_{k+1}\pi(\beta_{k+1}^{k+1},(\sigma_{k+1}^2)^{k+1}|k+1)\left(\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)\Gamma(v1)}\right)^k \\ &\cdot\eta_{1,k+1}^{v2-1}\cdots\eta_{k,k+1}^{v2-1}(1-\eta_{1,k+1})^{kv2-1}\cdots(1-\eta_{k,k+1})^{kv2-1} \\ &\cdot\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)^k\Gamma(v1)}\eta_{k+1,1}^{v1-1}\eta_{k+1,2}^{v2-1}\cdots\eta_{k+1,k+1}^{v2-1}\frac{\pi(i^N|\eta_{k+1},k+1)}{\pi(i^N|\eta_k,k)di_k^N} \\ &\cdot d\beta d\sigma^2 d\eta_{1,k+1}\cdots d\eta_{k,k+1}d\eta_{k+1,1}\cdots d\eta_{k+1,k+1} \\ &\cdot di_{k+1}^N d\mu_{k,\mathcal{R}}(z)\ , \qquad (19) \end{aligned}$$

where (17) is derived from the prior assumptions $A1 - A3$. (18) is obtained from the prior assumption $A4$ and the specification of the birth-death process, resulting in (19), which proves (16).
In the last step we check the detailed balance condition (10). $\chi(.)$ is the indicator function on the corresponding set. Let $\mathcal{Z}_k^b$ be the support of the parameters born z^b if the system jumps from k to $k+1$, and $\mathcal{Z}^{b-} := \mathcal{Z}^b \setminus \mathcal{I}_{k+1}$. By the prior assumptions and the assumptions on the birth death process this implies $\Omega_k \cup \mathcal{Z}_k^{b-} \cup \mathcal{I}_{k+1} \setminus \mathcal{I}_k = \Omega_{k+1}$. For a set $F \in \Omega_k$ we derive:

$$
\begin{aligned}
LHS &= \int_F \mathcal{B}(z) d\tilde{\mu}_k(z) && (20)\\
&= \int_{\Omega_k} \chi(z \in F) \mathcal{B}(z) f(\mathcal{Y}^N | z, k) d\mu_k(z) && (21)\\
&= \int_{\Omega_k} \chi(z \in F) \int_{\mathcal{Z}_k^b} b(z) \pi(\beta_{k+1}^{k+1}, (\sigma_{k+1}^2)^{k+1} | k+1) \left(\frac{\Gamma(kv_2 + v_1)}{\Gamma(v2)\Gamma(v1)} \right)^k \\
&\quad \cdot \eta_{1,k+1}^{v2-1} \cdots \eta_{k,k+1}^{v2-1} (1 - \eta_{1,k+1})^{kv2-1} \cdots (1 - \eta_{k,k+1})^{kv2-1} \\
&\quad \cdot \frac{\Gamma(kv_2 + v_1)}{\Gamma(v2)^k \Gamma(v1)} \eta_{k+1,1}^{v1-1} \eta_{k+1,2}^{v2-1} \cdots \eta_{k+1,k+1}^{v2-1} \pi(i^N | \eta_{k+1}, k+1) \\
&\quad \cdot d\beta d\sigma^2 d\eta_{1,k+1} \cdots d\eta_{k,k+1} d\eta_{k+1,1} \cdots d\eta_{k+1,k+1} di_{k+1}^N \\
&\quad \cdot f(\mathcal{Y}^N | y, k) d\mu_{k,\mathcal{R}}(z) && (22)\\
&= \int_{\Omega_k} \int_{\mathcal{Z}_k^{b-}} \int_{\mathcal{I}_{k+1}} \chi(z \in F) b(z) \pi(\beta_{k+1}^{k+1}, (\sigma_{k+1}^2)^{k+1} | k+1) \left(\frac{\Gamma(kv_2 + v_1)}{\Gamma(v2)\Gamma(v1)} \right)^k \\
&\quad \cdot \eta_{1,k+1}^{v2-1} \cdots \eta_{k,k+1}^{v2-1} (1 - \eta_{1,k+1})^{kv2-1} \cdots (1 - \eta_{k,k+1})^{kv2-1} \\
&\quad \cdot \frac{\Gamma(kv_2 + v_1)}{\Gamma(v2)^k \Gamma(v1)} \eta_{k+1,1}^{v1-1} \eta_{k+1,2}^{v2-1} \cdots \eta_{k+1,k+1}^{v2-1} \pi(i^N | \eta_{k+1}, k+1) \\
&\quad \cdot d\beta d\sigma^2 d\eta_{1,k+1} \cdots d\eta_{k,k+1} d\eta_{k+1,1} \cdots d\eta_{k+1,k+1} \\
&\quad \cdot di_{k+1}^N f(\mathcal{Y}^N | z, k) d\mu_k(z) \ , && (23)
\end{aligned}
$$

Where (i) (21) is derived from (15). (ii) Using (3) yields (22). (iii) Collecting terms results in (23). The right hand side (RHS) is derived by: (i) Use (15) to get (25). (ii) Insert (13) to derive (26). Use the symmetry in d_i to derive (27). (iii) Apply (16) to obtain (28). The fact that the integrals agree over the unrestricted set Ω_k and the restricted set $\Omega_{k,\mathcal{R}}$ – as already stated in (14) and (15) – ends up in (29).

$$
\begin{aligned}
RHS &= \int_{\Omega_{k+1}} \delta(y) K_\delta^{k+1}(y, F) d\tilde{\mu}_{k+1}(y) && (24)\\
&= \int_{\Omega_{k+1}} \delta(y) K_\delta^{k+1}(y, F) f(\mathcal{Y}^N | y, k+1) d\mu_{k+1}(y) && (25)\\
&= \int_{\Omega_{k+1,\mathcal{R}}} \int_{\mathcal{I}_k} \chi(z_\mathcal{R} = y_\mathcal{R} \setminus y_d \cup i_k^N \in F) \\
&\quad \cdot \sum_{i=1}^{k+1} d_i(y_\mathcal{R}) \pi(i^N | \eta_k, k) di_k^N f(\mathcal{Y}^N | y_\mathcal{R}, k+1) d\mu_{k+1,\mathcal{R}}(y) && (26)
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\Omega_{k+1,\mathcal{R}}} \int_{\mathcal{I}_k} \chi(z_{\mathcal{R}} = y_{\mathcal{R}} \setminus y_d \cup i_k^N \in F) && (27)\\
&\quad \cdot(k+1) d_i(y_{\mathcal{R}}) \pi(i^N|\eta_k, k) di_k^N f(\mathcal{Y}^N|y_{\mathcal{R}}, k+1) d\mu_{k+1,\mathcal{R}}(y)\\
&= \int_{\Omega_{k,\mathcal{R}}} \int_{\mathcal{Z}_k^{b-}} \int_{\mathcal{I}_{k+1}} \chi(z_{\mathcal{R}} \in F)(k+1) d_i(y_{\mathcal{R}}) f(\mathcal{Y}^N|y_{\mathcal{R}}, k+1)\\
&\quad \frac{\pi(k+1)}{\pi(k)} r_{k+1} \pi(\beta_{k+1}^{k+1}, (\sigma_{k+1}^2)^{k+1}|k+1) \left(\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)\Gamma(v1)}\right)^k\\
&\quad \cdot \eta_{1,k+1}^{v2-1} \cdots \eta_{k,k+1}^{v2-1} (1-\eta_{1,k+1})^{kv2-1} \cdots (1-\eta_{k,k+1})^{kv2-1}\\
&\quad \cdot \frac{\Gamma(kv_2+v_1)}{\Gamma(v2)^k \Gamma(v1)} \eta_{k+1,1}^{v1-1} \eta_{k+1,2}^{v2-1} \cdots \eta_{k+1,k+1}^{v2-1} \pi(i^N|\eta_{k+1}, k+1)\\
&\quad \cdot \cdot d\beta d\sigma^2 d\eta_{1,k+1} \cdots d\eta_{k,k+1} d\eta_{k+1,1} \cdots d\eta_{k+1,k+1} di_{k+1}^N\\
&\quad \cdot d\mu_{k,\mathcal{R}}(z) && (28)\\
&= \int_{\Omega_k} \int_{\mathcal{Z}_k^{b-}} \int_{\mathcal{I}_{k+1}} \chi(z \in F)(k+1) d_i(y) f(\mathcal{Y}^N|y, k+1)\\
&\quad \frac{\pi(k+1)}{\pi(k)} r_{k+1} \pi(\beta_{k+1}^{k+1}, (\sigma_{k+1}^2)^{k+1}|k+1) \left(\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)\Gamma(v1)}\right)^k\\
&\quad \cdot \eta_{1,k+1}^{v2-1} \cdots \eta_{k,k+1}^{v2-1} (1-\eta_{1,k+1})^{kv2-1} \cdots (1-\eta_{k,k+1})^{kv2-1}\\
&\quad \cdot \frac{\Gamma(kv_2+v_1)}{\Gamma(v2)^k \Gamma(v1)} \eta_{k+1,1}^{v1-1} \eta_{k+1,2}^{v2-1} \cdots \eta_{k+1,k+1}^{v2-1} \pi(i^N|\eta_{k+1}, k+1)\\
&\quad \cdot d\beta d\sigma^2 d\eta_{1,k+1} \cdots d\eta_{k,k+1} d\eta_{k+1,1} \cdots d\eta_{k+1,k+1} di_{k+1}^N\\
&\quad \cdot d\mu_k(z)\ . && (29)
\end{aligned}
$$

Now we check for (11):

$$
\begin{aligned}
LHS &= \int_G \delta(y) d\tilde{\mu}_{k+1}(y) && (30)\\
&= \int_{\Omega_{k+1}} \chi(y \in G) \delta(y) f(\mathcal{Y}^N|y, k+1) d\mu_{k+1}(y) && (31)\\
&= \int_{\Omega_{k+1}} \chi(y \in G) \int_{\mathcal{I}_k^N} \sum_{i=1}^{k+1} d_i(y) \pi(i^N|\eta_k, k) di_k^N && (32)\\
&= \int_{\Omega_{k+1}} \int_{\mathcal{I}_k^N} \chi(y \in G)(k+1) d_i(y) \pi(i^N|\eta_k, k) di_k^N\ . && (33)
\end{aligned}
$$

$$
\begin{aligned}
RHS &= \int_{\Omega_k} \mathcal{B}(z) K_{\mathcal{B}}^k(z, G) d\tilde{\mu}_k(z) && (34)\\
&= \int_{\Omega_k} \mathcal{B}(z) K_{\mathcal{B}}^k(z, G) f(\mathcal{Y}^N|z, k) d\mu_k(z) && (35)\\
&= \int_{\Omega_k} \int_{z^b : z \setminus i_k^N \cup z^b \in G} b(z) \pi(\beta_{k+1}^{k+1}, (\sigma_{k+1}^2)^{k+1}|k+1) \left(\frac{\Gamma(kv_2+v_1)}{\Gamma(v2)\Gamma(v1)}\right)^k\\
&\quad \cdot \eta_{1,k+1}^{v2-1} \cdots \eta_{k,k+1}^{v2-1} (1-\eta_{1,k+1})^{kv2-1} \cdots (1-\eta_{k,k+1})^{kv2-1}\\
&\quad \cdot \frac{\Gamma(kv_2+v_1)}{\Gamma(v2)^k \Gamma(v1)} \eta_{k+1,1}^{v1-1} \eta_{k+1,2}^{v2-1} \cdots \eta_{k+1,k+1}^{v2-1} \pi(i^N|\eta_{k+1}, k+1)
\end{aligned}
$$

$$
\begin{aligned}
& \cdot d\beta d\sigma^2 d\eta_{1,k+1} \cdots d\eta_{k,k+1} d\eta_{k+1,1} \cdots d\eta_{k+1,k+1} di_{k+1}^N \\
& f(\mathcal{Y}^N|z,k) d\mu_k(z) \qquad (36) \\
= & \int_{\Omega_{k,\mathcal{R}}} \int_{\mathcal{Z}_k^b} \chi(y_{\mathcal{R}} : z_{\mathcal{R}} \setminus i_k^N \cup z^b \in G) \\
& b(z_{\mathcal{R}}) \pi(\beta_{k+1}^{k+1}, (\sigma_{k+1}^2)^{k+1} | k+1) \left(\frac{\Gamma(kv_2 + v_1)}{\Gamma(v2)\Gamma(v1)} \right)^k \\
& \cdot \eta_{1,k+1}^{v2-1} \cdots \eta_{k,k+1}^{v2-1} (1 - \eta_{1,k+1})^{kv2-1} \cdots (1 - \eta_{k,k+1})^{kv2-1} \\
& \cdot \frac{\Gamma(kv_2 + v_1)}{\Gamma(v2)^k \Gamma(v1)} \eta_{k+1,1}^{v1-1} \eta_{k+1,2}^{v2-1} \cdots \eta_{k+1,k+1}^{v2-1} \pi(i^N | \eta_{k+1}, k+1) \\
& \cdot d\beta d\sigma^2 d\eta_{1,k+1} \cdots d\eta_{k,k+1} d\eta_{k+1,1} \cdots d\eta_{k+1,k+1} di_{k+1}^N \\
& f(\mathcal{Y}^N|z,k) d\mu_{k,\mathcal{R}}(z) \qquad (37) \\
= & \int_{\Omega_{k+1,\mathcal{R}}} \int_{\mathcal{I}_k^N} \chi(y_{\mathcal{R}} = z_{\mathcal{R}} \cup z^b \setminus i_k^N \in G) \\
& b(z_{\mathcal{R}}) \frac{\pi(k)}{r_{k+1}\pi(k+1)} \pi(i^N | \eta_k, k) di_k^N \\
& f(\mathcal{Y}^N|z,k) d\mu_{k+1,\mathcal{R}}(y) \qquad (38) \\
= & \int_{\Omega_{k+1}} \int_{\mathcal{I}_k^N} \chi(y \in G) b(z) \frac{\pi(k)}{r_{k+1}\pi(k+1)} \pi(i^N | \eta_k, k) di_k^N \\
& f(\mathcal{Y}^N|z,k) d\mu_{k+1}(y) \ . \qquad (39)
\end{aligned}
$$

(31) is derived from (15). (ii) Using (4) yields (32). (iii) Collecting terms results in (33). The right hand side (RHS) is derived by: (i) Use (15) to get (35). (ii) Insert (12) to derive (36). (iii) Substitute $\mu_{k,\mathcal{R}}(z)$ as shown in (16) to obtain (38). The fact that the integrals agree over the unrestricted set Ω_k and the restricted set $\Omega_{k,\mathcal{R}}$ – as already stated in (14) and (15) – ends up in (39). Thus $LHS = RHS$ if:

$$
(k+1) r_{k+1} d_i(y) f(\mathcal{Y}^N|y, k+1) \frac{\pi(k+1)}{\pi(k)} = b(z) f(\mathcal{Y}^N|z,k) \ , \qquad (40)
$$

for $y \in \Omega_{k+1}$ and $z \in \Omega_k$.
Therefore, (10) and (11) are fulfilled if

$$
(k+1) r_{k+1} d_i(y) f(\mathcal{Y}^N|y, k+1) \frac{\pi(k+1)}{\pi(k)} = b(z) f(\mathcal{Y}^N|z,k) \ , \qquad (41)
$$

for $y \in \Omega_{k+1}$ and $z \in \Omega_k$. This proves Lemma 1. □

References

James H. Albert, Siddhartha Chib (1993). Bayesian Inference via Gibbs Sampling of Autoregressive Time-Series subject to Markov Mean and Variance Shifts. *Journal of Business and Economic Statistics*, 11(1).

Stephen P. Brooks, Paolo Giudici (1998). Convergence Assessment for reversible jump MCMC Simulations. *Bayesian Statistics*, 6.

George Casella, Edward I. George (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174.

Siddhartha Chib, Edward Greenberg (1996). Markov Chain Monte Carlo Simulation Methods in Econometrics. *Econometric Theory*, 12:409–431.

Siddhartha Chib (1995). Marginal Likelihoods from the Gibbs Output. *Journal of the American Statistical Association*, 90(432):1313–1321.

T.J. DiCiccio, R. Krass, A. Raftery, L. Wasserman (1997). Computing Bayes Factors by Combining simulations and Asymptotic Approximations. *Journal of the American Statistical Association*, 92.

Yanqin Fan, Aman Ullah (1999). On goodsness-of-fit Tests for weakly dependent Processes using Kernel Method. *Nonparametric Statistics*, 11:337–360.

Sylvia Frühwirth-Schnatter (1995). Bayesian Model Discrimination and Bayes Factors for Linear Gaussian State Space Models. *Journal of the Royal Statistical Society*, 57:237–246.

Sylvia Frühwirth-Schnatter (1999). MCMC Estimation of Classical and Dynamic Switching and Mixture Models. *mimeo, Department of Statistics, Vienna University of Economics and Business Administration.*

Sylvia Frühwirth-Schnatter (1999). Model Likelihoods for Switching and Mixture Models. *mimeo, Department of Statistics, Vienna University of Economics and Business Administration.*

Peter J. Green, Antonietta Mira (1999). Delayed rejection in reversible jump Metropolis Hastings. *mimeo, Department of Mathematics, University of Bristol.*

Peter Green (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4).

Samuel Karlin, Howard M. Taylor (1975). *A First Course in Stochastic Processes.* Harcourt Brace and Company, Boston, 2 edition.

R.E. McCulloch, R.S. Tsay (1994). Statistical Analysis of Economic time Series via Markov Switching Models. *Journal of Time Series Analysis*, 15:523–539.

X.L. Meng, W.H. Wong (1996). Simulating Ratios if Normalizing Constants via a simple Identity. *Statistica Sinica*, 6:831–860.

Chris Preston (1976). Spatial Birth-and-Death processes. *Bulletin of the Institute of International Statistics*, 46:371–391.

Sylvia Richardson, Peter J. Green (1997). On bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society*, 59(4):731–792.

B.D. Ripley (1977). Modelling Spatial Patterns. *Journal of the Royal Statistical Society*, 39(2):172–212.

Christian P. Robert, Tobias Ryden, D.M. Titterington (1999). Bayesian Inference in Hidden Markov Models through Reversible Jump Markov Chain Monte Carlo. *mimeo, Laboraoire de Statistique, CREST, Paris.*

Christian P. Robert (1994). *The Bayesian Choice.* Springer, New York.

David Romer (1996). *Advanced Macroeconomics.* McGraw-Hill, New York.

Leopld Sögner (2000). Okun's law: Does the Austrian Unemployment-GDP Relationship exhibit Structural Breaks. *SFB-Working Paper, No. 61, Vienna University of Economics and Business Administration, http://www.wu-wien.ac.at/am/.*

Matthew Stephens (1997). *Bayesian Methods for Mixtures of Normal Distributions.* PHD-Dissertation, Department of Statistics, Oxford University, Oxford U.K.

Matthew Stephens (1998). Bayesian Analysis of Mixture Models with an Unknown Number of Components – an alternative to reversible jump methods. *mimeo, Department of Statistics, Oxford, U.K.*