Bayesian Analysis (2009) **4**, Number 1, pp. 63–84

Parameter Estimation in Continuous Time Markov Switching Models: A Semi-Continuous Markov Chain Monte Carlo Approach

Markus Hahn* and Jörn Sass†

Abstract. We want to estimate the parameters governing a continuous time Markov switching model given observations at discrete times only. For parameter estimation in a setting with continuous time and a latent state process, using MCMC methods two approaches are common: Using time-discretization and augmenting the unknowns with the (then discrete) state process, or working in continuous time and augmenting with the full state process. In this paper, we combine useful aspects of both approaches. On the one hand, we are inspired by the discretization, where filtering for the state process is possible, on the other hand, we catch attractive features of the continuous time method, like exact estimation (i.e. no discretization error) and direct estimation of the generator matrix rather than the transition matrix. This is achieved by taking not the whole state process for data augmentation but only the states at observation times. Using results on the distribution of occupation times in Markov processes, it is possible to compute the complete data likelihood exactly. We obtain a sampler that works more robustly and more accurately especially for fast switching in the state process.

Keywords: Bayesian inference, data augmentation, hidden Markov model, switching diffusion

1 Introduction

We consider a continuous time Markov switching model (MSM). The observation process can be seen as a diffusion where the drift and the volatility coefficients are modeled as continuous time, finite state Markov processes with a common state process. This model includes the hidden Markov model (HMM) where the volatility is constant. Continuous time MSMs (e.g. Buffington and Elliott 2002; Guo 2001) and HMMs (e.g. Liechty and Roberts 2001; Sass and Haussmann 2004) are widely used in finance, but are also applied in many other areas, for example in biophysics for the problem of restoration of ion channels (e.g. Ball et al. 1999).

It is our aim to estimate the states for drift and for volatility as well as the generator matrix of the underlying Markov process based on discretely observed data. This is a common situation that occurs e.g. in many inference problems encountered in finance, where it is more convenient to use continuous time models since these allow the derivation of closed form solutions while stock prices are observed only in discrete time.

*RICAM, Austrian Academy of Sciences, Linz, Austria, mailto:markus.hahn@oeaw.ac.at
†Department of Mathematics, University of Kaiserslautern, Kaiserslautern, Germany, mailto:sass@mathematik.uni-kl.de

 DOI:10.1214/09-BA402

We assume that the number of states is given, and refer to (Otranto and Gallo 2002) and (Frühwirth-Schnatter 2001) for the matter of model selection. For the problem of label switching, which arises if the different states are not well separated, methods like identifiability constraints, relabelling, or random permutation (cf. Jasra et al. 2005) can be incorporated into the sampler constructed.

In discrete time, the expectation maximization (EM) algorithm (Dempster et al. 1977) can be used to find maximum likelihood estimates for the parameters, the drift and covariance coefficients as well as the generator matrix governing the switching of the states, see e.g. (Elliott et al. 1997) or (Engel and Hamilton 1990). For low rates (i.e. rare jumps) and low volatility, the EM algorithm provides very accurate results. However, for high rates the law of the discretized model (using a discrete time Markov chain) varies considerably from the law of the continuous time model and the EM algorithm yields poor results. In continuous time, no finite dimensional filter for estimating the underlying state of the MSM is known, hence it is not possible to compute the maximum likelihood parameter estimate via an EM algorithm, see (Elliott et al. 1995) for details. Alternatively, one might use a moment based method, see e.g. (Elliott et al. 2008). This can yield good estimates, but it requires a lot of observations.

Taking a Bayesian approach, for parameter estimation in a setting with continuous time and a latent state process, two methods based on Markov chain Monte Carlo (MCMC) seem to be reasonable: First, using time discretization (replacing the generator matrix with a transition matrix) and augmenting the unknowns with the (then discrete) state process; or, second, working in continuous time and augmenting with the full state process (compare Ball et al. 1999; Liechty and Roberts 2001). Both methods are based on a multiple-block Metropolis-Hastings sampler. Results for the continuous time MSM in (Hahn et al. 2007) indicate that for data with high rates, considerable noise, and based on not too many observations, MCMC estimation in continuous time is not sufficiently stable and fast. Then again, using time discretization and discrete time MCMC methods turns out to work more robustly and faster, but introduces a discretization error.

The idea of the algorithm presented here is to combine the useful aspects of these two MCMC approaches. On the one hand, the algorithm is inspired from the discretization, where filtering for the state process is possible, on the other hand, we catch attractive features of the continuous time method, like exact estimation (i.e. no discretization error) and direct estimation of the generator matrix rather than the transition matrix (compare Israel et al. 2001). This is achieved by taking not the whole state process for data augmentation, but considering only the states at observation times. Working in continuous time but augmenting with the state process at discrete times only, we refer to this method as "semi-continuous".

For this setting, the conditional distribution of the observed data given the boundary states is an infinite mixture of normal distributions. As the weights of these mixtures depend rather on the occupation times of the state process than on the path of the state process itself, the calculations simplify. Using results on the distribution of occupation times in Markov processes in (Sericola 2000), it is possible to compute the complete data likelihood exactly.

We construct an exact scheme for a multiple-block Metropolis-Hastings sampler. More precisely, it is possible to perform a Gibbs step for the update of the state process. For the update of the generator matrix, we construct a proposal approximating a Gibbs step. The drift coefficients and the volatility coefficients are updated using a random walk Metropolis step.

Compared to augmenting with the full state process, we dramatically lower the dimension of the unknowns. Also the dependence between the parameters is lowered (especially between the state process and the generator matrix). Therefore, we expect faster convergence.

The remainder of the paper is organized as follows. In Section 2 we introduce the model and the problem of estimation given discretely observed data. In Section 3 we specify the prior distributions and give the complete data likelihood. In Section 4 we describe the proposal distributions used for the Metropolis-Hastings sampler. In Section 5 we derive the conditional distribution of the observations using results from (Sericola 2000) needed to implement the sampler (in order to improve readability we describe the case of 2 states; in the Appendix, details for the general case are given). Numerical results in comparison to other methods are given in Section 6 both for simulated data and financial market data.

2 Continuous time Markov switching model

2.1 Setting up the model

In this section we present the model which is a multidimensional continuous time MSM. On a filtered probability space $(\Omega, \mathcal{F} = (\mathcal{F}_t)_{t\in[0,T]}, \mathrm{P})$ we observe in $[0,T]$ the n-dimensional process $R = (R_t)_{t\in[0,T]}$,

$$R_t = \int_0^t \mu_s \,\mathrm{d}s + \int_0^t \sigma_s \,\mathrm{d}W_s, \tag{1}$$

where $W = (W_t)_{t\in[0,T]}$ is an n-dimensional Brownian motion with respect to $\mathcal{F}$, with $\mu = (\mu_t)_{t\in[0,T]}$ being the drift process, and $\sigma = (\sigma_t)_{t\in[0,T]}$ being the volatility process. The drift process μ and the volatility process σ, taking values in $\mathbb{R}^n$ and $\mathbb{R}^{n\times n}$, respectively, are continuous time, time homogeneous, irreducible Markov processes with d states, adapted to $\mathcal{F}$, independent of W, driven by the same state process $Y = (Y_t)_{t\in[0,T]}$. We denote the possible values of μ and σ with $B = (\mu^{(1)}, \ldots, \mu^{(d)})$ and $\Sigma = (\sigma^{(1)}, \ldots, \sigma^{(d)})$, respectively. We assume the volatility matrices $\sigma^{(k)}$ to be nonsingular and use the notation $C^{(k)} = \sigma^{(k)} (\sigma^{(k)})^\top$.

The state process Y, which is a continuous time, time homogeneous, irreducible Markov process adapted to $\mathcal{F}$, independent of W, with state space $\{1, \ldots, d\}$, allows for

the representations

$$\mu_t = \sum_{k=1}^{d} \mu^{(k)} \mathbb{1}_{\{Y_t=k\}}, \qquad \sigma_t = \sum_{k=1}^{d} \sigma^{(k)} \mathbb{1}_{\{Y_t=k\}}. \tag{2}$$

The state process Y is characterized by the generator matrix $Q \in \mathbb{R}^{d\times d}$ as follows: For rates $\lambda_k = -Q_{kk} = \sum_{l=1,l\neq k}^{d} Q_{kl}$, $k = 1,\dots,d$, in state k the waiting time for the next jump is λ_k-exponentially distributed, and the probability of jumping to state $l \neq k$ when leaving k is given by Q_{kl}/λ_k, see e.g. (Brémaud 1981).

Remark 1. One cannot distinguish between the pairs (σ, W) and $(\bar\sigma, \bar W)$, where $\bar\sigma$ is a square-root of $\sigma\sigma^\top$ and $\bar W = \bar\sigma^{-1}\sigma W$. However, without loss of generality we can assume $\sigma^{(k)}$ to be a lower triangular matrix, i.e. $\sigma^{(k)}(\sigma^{(k)})^\top$ equals the Cholesky factorization of the covariance matrix.

2.2 Estimation problem

Starting from a prior distribution of the unknown parameters B, Σ, Q, we wish to determine $\mathrm{P}(B, \Sigma, Q \mid \dot R)$, the posterior distribution of these parameters given discretely observed data $\dot R = (\dot R_m)_{m=0,\dots,N}$, $\dot R_m = R_{\Delta t\, m}$, for a uniform observation time interval $\Delta t > 0$. Equivalently, one can consider the corresponding increments $\Delta\dot R = (\Delta\dot R_m)_{m=1,\dots,N}$, where

$$\Delta\dot R_m = \int_{(m-1)\Delta t}^{m\,\Delta t} \mu_s \,\mathrm{d}s + \int_{(m-1)\Delta t}^{m\,\Delta t} \sigma_s \,\mathrm{d}W_s, \quad m = 1,\dots,N. \tag{3}$$

To this end, we augment the unknowns with $\dot Y = (\dot Y_m)_{m=0,\dots,N}$, $\dot Y_m = Y_{m\,\Delta t}$, the latent state process at the observation times. We construct a multiple-block Metropolis-Hastings sampler to draw from the joint posterior distribution of $B, \Sigma, Q, \dot Y$ given $\dot R$. We partition the unknowns into three blocks, namely (B, Σ), Q, $\dot Y$.

In Sections 3 and 4, we specify appropriate prior distributions and describe the construction of the updates.

2.3 Motivation

As pointed out in Section 1, existing algorithms for estimating MSMs work well for slow state switching, but show weaknesses in dealing with fast switching. While Rydén et al. (1998) and Timmermann (2000) show that MSMs are suitable approximations for modeling stock returns (catching a number of stylized facts), in such applications one often faces fast state switching, compare (Sass and Haussmann 2004) and (Hahn et al. 2008a) for daily and intra-day data, respectively. To avoid this problem, one might make the observation interval Δt smaller, but this may be problematic. More frequently quoted data might not be available (think of assets quoted on a daily basis

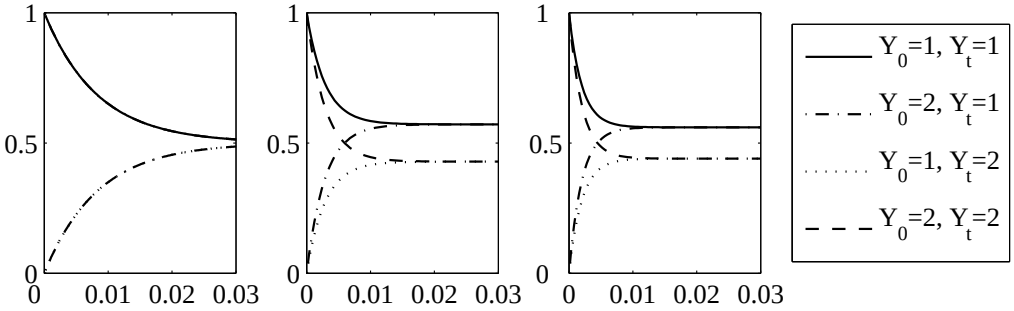

Figure 1: Transition probabilities depending on time for different rates (left: $\lambda_1 = \lambda_2 = 60$; middle: $\lambda_1 = 150, \lambda_2 = 200$; right: $\lambda_1 = 220, \lambda_2 = 280$)

only). Still, if high-frequent data is available, one should reconsider that appropriate statistical properties for stock returns are obtained only over time intervals that are not too short; e.g. for second-by-second quotations it is inevitable to include the possibility of jumps in the price process.

This is some motivation to look for methods giving reliable results even for fast state switching. And in fact it is possible to obtain stable results for such data (of course depending on the noise level), where "fast" switching means up to about one switch per observation interval on average (in accordance with observations in Sass and Haussmann 2004 and Hahn et al. 2008b).

To give some illustration, Figure 1 plots $\mathrm{P}(Y_t = l \,|\, Y_0 = k, Q)$, the transition probabilities depending on time, for different rates. Assuming $\Delta t = 1/250$ (as in Section 6), even in the third panel with high rates, transition probabilities for a time-step of length Δt are far from ergodic probabilities. This justifies our choice of augmenting with $\dot{Y}$. More illustrations and results for some reasonably chosen drift and volatility are found in Section 6.

3 Prior distributions and complete data likelihood

3.1 Prior distributions

Assumption 1. A priori, B, Σ, Q, $\dot{Y}_0$, are independent with distributions

$$\begin{aligned} \mu_i^{(k)} &\sim \mathrm{N}(m_{ik}, s_{ik}^2), \\ C^{(k)} &\sim \mathrm{IW}(\Xi^{(k)}, \nu_k), \\ Q_{kl} &\sim \mathrm{Ga}(f_{kl}, g_{kl}), \\ \dot{Y}_0 &\sim \mathrm{U}(\{1, \dots, d\}), \end{aligned} \tag{4}$$

where $i = 1, \dots, n$ and $k, l = 1, \dots, d$, $l \neq k$. With N, IW, Ga, and U, we denote the normal, inverse Wishart, Gamma, and uniform distribution, respectively. The off-

diagonal elements of Q are assumed to be independent. Without loss of generality, the volatility matrices $\sigma^{(k)}$ are assumed to be lower triangular matrices with positive diagonal entries.

If we think of time zero as the beginning of our observations after the process has already run for some time, it may be reasonable to alter the assumptions on the initial state as follows:

Assumption 1a. Assumption 1 holds, but B, Σ, $(Q, \dot{Y}_0)$ are independent and the state process starts from its ergodic probability ω, i.e.

$$\mathrm{P}(\dot{Y}_0 \,|\, Q) = \omega. \tag{5}$$

3.2 Complete data likelihood

For given $B, \Sigma, Q, \dot{Y}$, the process $\dot{R}$ is Markov and its increments $\Delta\dot{R}$ are independent. Hence, the complete data likelihood is given by

$$\mathrm{P}(\dot{R} \,|\, B, \Sigma, Q, \dot{Y}) = \prod_{m=1}^{N} \mathrm{P}(\Delta\dot{R}_m \,|\, B, \Sigma, Q, \dot{Y}_{m-1}, \dot{Y}_m), \tag{6}$$

where $\mathrm{P}(\Delta\dot{R}_m \,|\, B, \Sigma, Q, \dot{Y}_{m-1}, \dot{Y}_m)$ is an infinite mixture of normal distributions (for details see Section 5).

Note that while $\dot{R}$ given Y is independent of Q, this is not true for $\dot{R}$ given $\dot{Y}$.

4 Proposal distributions

4.1 Drift and volatility

For the update of B and Σ, a joint normal random walk proposal, reflected at zero for the diagonal entries of $\sigma^{(k)}$, is used:

$$\begin{aligned} \mu^{(k)\prime} &= \mu^{(k)} + r^{\mu}\, \phi, \\ \sigma_{ij}^{(k)\prime} &= \sigma_{ij}^{(k)} + r^{\sigma}\, \psi_{ij}^{(k)}, \\ \sigma_{ii}^{(k)\prime} &= \bigl|\sigma_{ii}^{(k)} + r^{\sigma}\, \psi_{ii}^{(k)}\bigr|, \end{aligned} \tag{7}$$

where ϕ and ψ are n-dimensional vectors and $n \times n$ matrices, respectively, of independent standard normal variates, $i = 1, \ldots, n$, $j = 1, \ldots, i-1$, $k = 1, \ldots, d$, and r^{μ} and r^{σ} are parameters scaling the step widths.

As the transition kernel is symmetric we have a Metropolis step with acceptance probability $\alpha_{B,\Sigma} = \min\{1, \bar{\alpha}_{B,\Sigma}\}$, where

$$\bar{\alpha}_{B,\Sigma} = \frac{\mathrm{P}(\dot{R} \,|\, B', \Sigma', Q, \dot{Y})\, \mathrm{P}(B')\, \mathrm{P}(\Sigma')}{\mathrm{P}(\dot{R} \,|\, B, \Sigma, Q, \dot{Y})\, \mathrm{P}(B)\, \mathrm{P}(\Sigma)}. \tag{8}$$

4.2 Generator matrix

For the off-diagonal elements of the generator matrix we sample from a Gamma distribution with parameters $f_{kl} + \hat{f}_{kl}$ and $g_{kl} + \hat{g}_{kl}$, where the expected number of jumps $\hat{f}$ and the expected occupation times $\hat{g}$ can be computed as follows. Denoting the transition matrix of $\dot{Y}$ with X,

$$X_{kl} = \mathrm{P}(\dot{Y}_m = l \,|\, \dot{Y}_{m-1} = k) = \big(\exp(Q\,\Delta t)\big)_{kl}, \tag{9}$$

and the number of jumps in $\dot{Y}$ with $\hat{n}_{kl}$,

$$\hat{n}_{kl} = \big|\{m \in \{1, \dots, N\} \,|\, \dot{Y}_{m-1} = k, \dot{Y}_m = l\}\big|, \tag{10}$$

the maximum likelihood estimate of the transition matrix given $\dot{Y}$ is given by $\hat{X}_{kl} = \hat{n}_{kl} / \sum_{j=1}^{d} \hat{n}_{kj}$, from which we can compute the corresponding generator matrix $\hat{Q}$, stationary distribution $\hat{\omega}$, and expected number of jumps $\hat{f}$ and expected occupation times $\hat{g}$,

$$\begin{aligned} \hat{f}_{kl} &= \hat{\omega}_k \, \hat{Q}_{kl} \, T, \\ \hat{g}_{kl} &= \hat{\omega}_k \, T. \end{aligned} \tag{11}$$

Then, the proposal distribution is independent of Q with density

$$q(Q') = \prod_{k \neq l} \mathrm{Ga}(Q'_{kl}; f_{kl} + \hat{f}_{kl}, g_{kl} + \hat{g}_{kl}). \tag{12}$$

For the conditional distribution of the generator matrix we have

$$\mathrm{P}(Q \,|\, \dot{R}, B, \Sigma, \dot{Y}) \propto \mathrm{P}(\dot{R} \,|\, B, \Sigma, Q, \dot{Y}) \, \mathrm{P}(\dot{Y} \,|\, Q) \, \mathrm{P}(Q), \tag{13}$$

where $\mathrm{P}(\dot{Y} \,|\, Q)$ is obtained as

$$\mathrm{P}(\dot{Y} \,|\, Q) = \mathrm{P}(\dot{Y}_0 \,|\, Q) \prod_{k,l=1}^{d} X_{kl}^{\hat{n}_{kl}}. \tag{14}$$

Hence, the acceptance probability is given by $\alpha_Q = \min\{1, \bar{\alpha}_Q\}$ with

$$\bar{\alpha}_Q = \frac{\mathrm{P}(\dot{R} \,|\, B, \Sigma, Q', \dot{Y}) \, \mathrm{P}(\dot{Y} \,|\, Q') \, \mathrm{P}(Q')}{\mathrm{P}(\dot{R} \,|\, B, \Sigma, Q, \dot{Y}) \, \mathrm{P}(\dot{Y} \,|\, Q) \, \mathrm{P}(Q)} \, \frac{q(Q)}{q(Q')}. \tag{15}$$

Remark 2. If we augment with Y instead of $\dot{Y}$, we know the number of jumps and the occupation times for the current sample path, and the above proposal constitutes a Gibbs step. Hence, for our setting using only $\dot{Y}$ it seems reasonable to keep the proposal approximating the parameters.

4.3 State process

For updating $\dot{Y}$, we describe an extension of the forward-filtering-backward-sampling algorithm (see e.g. Frühwirth-Schnatter 2006) allowing to draw from the full conditional distribution. We give details using the notation $\Delta\dot{R}^m = (\Delta\dot{R}_1, \ldots, \Delta\dot{R}_m)$ and $\eta = (B, \Sigma, Q)$.

Starting from $\mathrm{P}(\dot{Y}_0 \,|\, \Delta\dot{R}^0, \eta) = \mathrm{P}(\dot{Y}_0 \,|\, Q)$, for $m = 1, \ldots, N$ we compute the filtered probabilities

$$\mathrm{P}(\dot{Y}_m \,|\, \Delta\dot{R}^m, \eta) = \sum_{k=1}^{d} \mathrm{P}(\dot{Y}_m \,|\, \dot{Y}_{m-1} = k, \Delta\dot{R}^m, \eta)\, \mathrm{P}(\dot{Y}_{m-1} = k \,|\, \Delta\dot{R}^m, \eta), \tag{16}$$

where

$$\begin{aligned} \mathrm{P}(\dot{Y}_m \,|\, \dot{Y}_{m-1}, \Delta\dot{R}^m, \eta) &= \mathrm{P}(\dot{Y}_{m-1}, \Delta\dot{R}^{m-1}, \eta)\, \mathrm{P}(\dot{Y}_m \,|\, \dot{Y}_{m-1}, \Delta\dot{R}^{m-1}, \eta) \\ &\quad \times \mathrm{P}(\Delta\dot{R}_m \,|\, \dot{Y}_{m-1}, \dot{Y}_m, \Delta\dot{R}^{m-1}, \eta) / \mathrm{P}(\dot{Y}_{m-1}, \Delta\dot{R}^m, \eta) \\ &\propto X_{\dot{Y}_{m-1}, \dot{Y}_m}\, \mathrm{P}(\Delta\dot{R}_m \,|\, \dot{Y}_{m-1}, \dot{Y}_m, \eta) \end{aligned} \tag{17}$$

(using that given $\dot{Y}_{m-1}$, $\dot{Y}_m$ is independent of $\Delta\dot{R}^{m-1}$), and

$$\mathrm{P}(\dot{Y}_{m-1} \,|\, \Delta\dot{R}^m, \eta) \propto \mathrm{P}(\Delta\dot{R}_m \,|\, \dot{Y}_{m-1}, \eta)\, \mathrm{P}(\dot{Y}_{m-1} \,|\, \Delta\dot{R}^{m-1}, \eta) \tag{18}$$

in a forward-run.

Then we perform backward-sampling: we draw $\dot{Y}'_N$ from $\mathrm{P}(\dot{Y}_N \,|\, \Delta\dot{R}^N, \eta)$, and for $m = N-1, \ldots, 0$ we draw $\dot{Y}'_m$ from

$$\begin{aligned} \mathrm{P}(\dot{Y}_m \,|\, \dot{Y}_{m+1}, \Delta\dot{R}^N, \eta) &= \mathrm{P}(\dot{Y}_m \,|\, \dot{Y}_{m+1}, \Delta\dot{R}^{m+1}, \eta) \\ &\propto \mathrm{P}(\dot{Y}_{m+1} \,|\, \dot{Y}_m, \Delta\dot{R}^{m+1}, \eta)\, \mathrm{P}(\dot{Y}_m \,|\, \Delta\dot{R}^{m+1}, \eta), \end{aligned} \tag{19}$$

where we used that given $\dot{Y}_{m+1}$, $\dot{Y}_m$ is independent of $\Delta\dot{R}_{m+2}, \ldots, \Delta\dot{R}_N$.

5 Conditional distributions of the observations

The conditional distributions of the observed increments $\Delta\dot{R}_m$ given the initial state $\dot{Y}_{m-1}$ and the boundary states $\dot{Y}_{m-1}$, $\dot{Y}_m$, respectively, are infinite mixtures of normal distributions given by

$$\mathrm{P}(\Delta\dot{R}_m \,|\, \dot{Y}_{m-1} = k, \eta) = \int_{\mathcal{T}} \psi(\Delta\dot{R}_m;\, B, C, \tau)\, \mathrm{d}F_k^{\mathrm{in}}(\tau), \tag{20}$$

$$\mathrm{P}(\Delta\dot{R}_m \,|\, \dot{Y}_{m-1} = k, \dot{Y}_m = l, \eta) = \int_{\mathcal{T}} \psi(\Delta\dot{R}_m;\, B, C, \tau)\, \mathrm{d}F_{k,l}^{\mathrm{bd}}(\tau), \tag{21}$$

where

$$\psi(\Delta\dot{R}_m;\, B, C, \tau) = \varphi\Big(\Delta\dot{R}_m;\, \sum_{j=1}^{d} \mu^{(j)}\tau_j, \sum_{j=1}^{d} C^{(j)}\tau_j\Big), \tag{22}$$

$\varphi(\,\cdot\,; x, c)$ denotes the density of a multivariate normal distribution with mean vector x and covariance matrix c, $\mathcal{T} = [0, \Delta t]^d$, and F^{in} and F^{bd} denote the conditional joint distribution functions of the occupation times given the initial and boundary states, respectively,

$$F_k^{\text{in}}(\tau) = \mathrm{P}(O_{\Delta t}^1 \le \tau_1, \dots, O_{\Delta t}^d \le \tau_d \,|\, Y_0 = k), \tag{23}$$

$$F_{k,l}^{\text{bd}}(\tau) = \mathrm{P}(O_{\Delta t}^1 \le \tau_1, \dots, O_{\Delta t}^d \le \tau_d \,|\, Y_0 = k, Y_{\Delta t} = l). \tag{24}$$

Here O_t^k denotes the occupation times of Y in state k up to time t,

$$O_t^k = \int_0^t \mathbb{1}_{\{Y_s = k\}} \,\mathrm{d}s. \tag{25}$$

Since

$$\mathrm{P}(\Delta \dot{R}_m \,|\, \dot{Y}_{m-1} = k, \eta) = \sum_{l=1}^{d} \mathrm{P}(\Delta \dot{R}_m \,|\, \dot{Y}_{m-1} = k, \dot{Y}_m = l, \eta) X_{kl}, \tag{26}$$

it is sufficient to describe F^{bd}. In the following we use results of Sericola (2000) to obtain expressions for the corresponding density f^{bd}.

First we introduce some notation. We denote the uniformized Markov chain (see e.g. Ross 1996) associated to the Markov process Y with $\tilde{Z} = (\tilde{Z}_m)_{m \in \mathbb{N}_0}$ and the uniform jump rate of $\tilde{Z}$ with $\tilde{\lambda}$. We define the number of visits of $\tilde{Z}$ in state k during the first m transitions

$$\tilde{V}_m^k = \sum_{h=0}^{m} \mathbb{1}_{\{\tilde{Z}_h = k\}}. \tag{27}$$

We set

$$\mathcal{D} = \{\tau \in [0, \Delta t]^d \,|\, \tau_1 + \cdots + \tau_d = \Delta t\} \tag{28}$$

and use the notation $\operatorname{ri} \mathcal{D}$ for its relative interior. Clearly, we just have $\operatorname{ri} \mathcal{D} = \{\tau \in (0, \Delta t)^d \,|\, \tau_1 + \cdots + \tau_d = \Delta t\}$. We also define its relative boundary $\operatorname{rb} \mathcal{D} = \mathcal{D} \setminus \operatorname{ri} \mathcal{D}$.

It is useful to introduce the * operator that projects a vector on its first $d-1$ components, i.e. for a d-dimensional vector the last component is omitted, while a $d-1$-dimensional vector is unchanged.

To ease notation, in the remainder of this section we assume Q to be given.

Sericola (2000, Theorem 4.3) describes the joint distribution of the occupation times of $d-1$ states and the final state conditional on the initial state in terms of the distribution of the uniformized discrete chain $\tilde{Z}$. We adapt the conditioning and take partial derivatives to obtain an expression for the density f^{bd}.

Theorem 1. *On* $\operatorname{ri}\mathcal{D}$, F^{bd} *has a continuous density* f^{bd}, *and for* $\tau \in \operatorname{ri}\mathcal{D}$,

$$f^{\mathrm{bd}}_{k,l}(\tau) = \sum_{m=d-1}^{\infty} e^{-\tilde{\lambda}\,\Delta t} \frac{(\tilde{\lambda}\,\Delta t)^m}{m!} \sum_{v^* \in \mathcal{V}_m} \chi^{\Delta t;\tau^*}_{m;v^*}\, \mathrm{P}(\tilde{V}^*_m \le v^*, \tilde{Z}_m = l \,|\, \tilde{Z}_0 = k)/X_{kl}, \quad (29)$$

where

$$\mathcal{V}_m = \{(v_1, \dots, v_{d-1}) \,|\, v_i \ge 0, v_1 + \cdots + v_{d-1} \le m\} \quad (30)$$

and, setting $\bar{\tau}_d = \Delta t - \tau_1 - \cdots - \tau_{d-1}$, $\bar{v}_d = m - v_1 - \cdots - v_{d-1}$,

$$\begin{aligned}\chi^{\Delta t;\tau^*}_{m;v^*} &= \frac{\partial^{d-1}}{\partial\tau_1 \cdots \partial\tau_{d-1}} \left(\frac{m!\, \tau_1^{v_1} \cdots \tau_{d-1}^{v_{d-1}}\, \bar{\tau}_d^{\bar{v}_d}}{t^m\, v_1! \cdots v_{d-1}!\, \bar{v}_d!} \right) \\ &= \frac{m!}{\Delta t^m\, v_1! \cdots v_{d-1}!\, \bar{v}_d!} \\ &\quad \times \sum_{i=0}^{d-1} (-1)^i \sum_{\substack{J \subseteq \{1,\dots,d-1\} \\ |J| = i}} \left(\prod_{j \in J} v_j\, \tau_j^{v_j - 1} \right) \left(\prod_{j \notin J} \tau_j^{v_j} \right) \frac{\bar{v}_d!}{(\bar{v}_d - i)!}\, \bar{\tau}_d^{\bar{v}_d - i}. \end{aligned} \quad (31)$$

Remark 3. The conditional joint distribution of $\tilde{V}^*_m$ and $\tilde{Z}_m$ can be computed via its backward equation (see Sericola 2000, Section 4.1).

For a better understanding, consider the case of $d = 2$, where Theorem 1 takes a simpler form.

Corollary 2. *Let* $f^{\mathrm{bd}}_{k,l}(\tau_1, \tau_2)$ *denote the joint distribution density of the occupation times* (τ_1, τ_2) *given the initial state equals* k *and the final state equals* l. *For* $\tau_1 + \tau_2 = \Delta t$, $\tau_1 \in (0, \Delta t)$, *the density is continuous and*

$$\begin{aligned} f^{\mathrm{bd}}_{k,l}(\tau_1, \tau_2) = \sum_{m=1}^{\infty} \Bigg(& e^{-\tilde{\lambda}\,\Delta t} \frac{(\tilde{\lambda}\,\Delta t)^m}{m!} \sum_{v=0}^{m} \binom{m}{v} \frac{\tau_1^{v-1}\, \tau_2^{m-v-1} (v\,\tau_2 - (m-v)\tau_1)}{\Delta t^m} \\ & \times \mathrm{P}(\tilde{V}^1_m \le v, \tilde{Z}_m = l \,|\, \tilde{Z}_0 = k) \Bigg) / X_{kl}. \end{aligned} \quad (32)$$

At the boundary we have

$$\mathrm{P}(O_{\Delta t} = (0, \Delta t) \,|\, Y_0 = k, Y_{\Delta t} = l) = \mathbb{1}_{\{k=2, l=2\}}\, e^{-\lambda_2\,\Delta t}/X_{22}, \quad (33)$$

$$\mathrm{P}(O_{\Delta t} = (\Delta t, 0) \,|\, Y_0 = k, Y_{\Delta t} = l) = \mathbb{1}_{\{k=1, l=1\}}\, e^{-\lambda_1\,\Delta t}/X_{11}. \quad (34)$$

Hence, Equation (21) *can be rewritten as*

$$\begin{aligned}\mathrm{P}(\Delta\dot{R}_m \,|\, \dot{Y}_{m-1} = k, \dot{Y}_m = l, \eta) = \\ \psi\big(\Delta\dot{R}_m;\, B, C, (0, \Delta t)\big)\, \mathrm{P}(O_{\Delta t} = (0, \Delta t) \,|\, Y_0 = k, Y_{\Delta t} = l) \\ + \int_0^{\Delta t} \psi\big(\Delta\dot{R}_m;\, B, C, (\tau_1, \Delta t - \tau_1)\big) f^{\mathrm{bd}}_{k,l}(\tau_1, \Delta t - \tau_1)\, \mathrm{d}\tau_1 \\ + \psi\big(\Delta\dot{R}_m;\, B, C, (\Delta t, 0)\big)\, \mathrm{P}(O_{\Delta t} = (\Delta t, 0) \,|\, Y_0 = k, Y_{\Delta t} = l). \quad (35)\end{aligned}$$

In the appendix, we describe how to proceed for $d > 2$.

Remark 4. The error in the computation of the conditional distributions of the observations $\Delta\dot{R}_m$ results firstly from the truncation of the infinite sum in (29), and secondly from the numerical approximation of the integrals. However, opposite to the situation of the discrete approximation, using this approach we can control the error and reduce it arbitrarily. In fact, we have fast convergence for the error-affected terms and hence, it is possible to keep the numerical error negligible: Due to the exponential term in (29), only the very first summands are of importance, and as ψ as well as f^{bd} are very smooth functions, e.g. the composite Simpson's Rule provides fast and highly accurate results even for few subintervals (compare Figure 4).

We illustrate the results of this section with Figures 2, 3, and 4, where we consider a one-dimensional $d = 2$ states model with observation time interval length $\Delta t = 0.004$; for state one and two, the corresponding rates are set to $\lambda_1 = 150$ and $\lambda_2 = 200$, the drifts to $\mu^{(1)} = 2$ and $\mu^{(2)} = -2$, and the volatilities to $\sigma^{(1)} = 0.10$ and $\sigma^{(2)} = 0.12$.

Figure 2 shows the conditional distributions of the occupation time of the first state given the boundary and initial states. Note that the distributions for $\dot{Y}_0 = 1, \dot{Y}_1 = 2$ and $\dot{Y}_0 = 2, \dot{Y}_1 = 1$ are identical (also in Figure 3 and 4) and that the discrete contributions in the density are not shown.

Figure 3 shows the distributions of the observations $\Delta\dot{R}_m$ (rescaled to be comparable with μ) given the boundary (left) and initial (right) states; for comparison with the discrete time approximation, we included the normal densities $\varphi(\,\cdot\,;\, \mu^{(1)}, C^{(1)})$ and $\varphi(\,\cdot\,;\, \mu^{(2)}, C^{(2)})$ (dotted). Note that the conditional distributions are skew and obviously non-normal.

In Figure 4 the distributions of the occupation time of the first state given the boundary (left) and initial states (right) are plotted for varying precision. The solid plots (highest values at the modes) show the true functions. For the dotted plots, three summands in (32) and three sampling points for Simpson's Rule were used. For the dashed plots (which nearly totally coincide with the true functions), four summands and five integration points were used. This makes it evident that the approximation error is negligible even for fast state switching. Using five terms in the summation and seven integration points takes little more computation time and gives results numerically equal to the exact values.

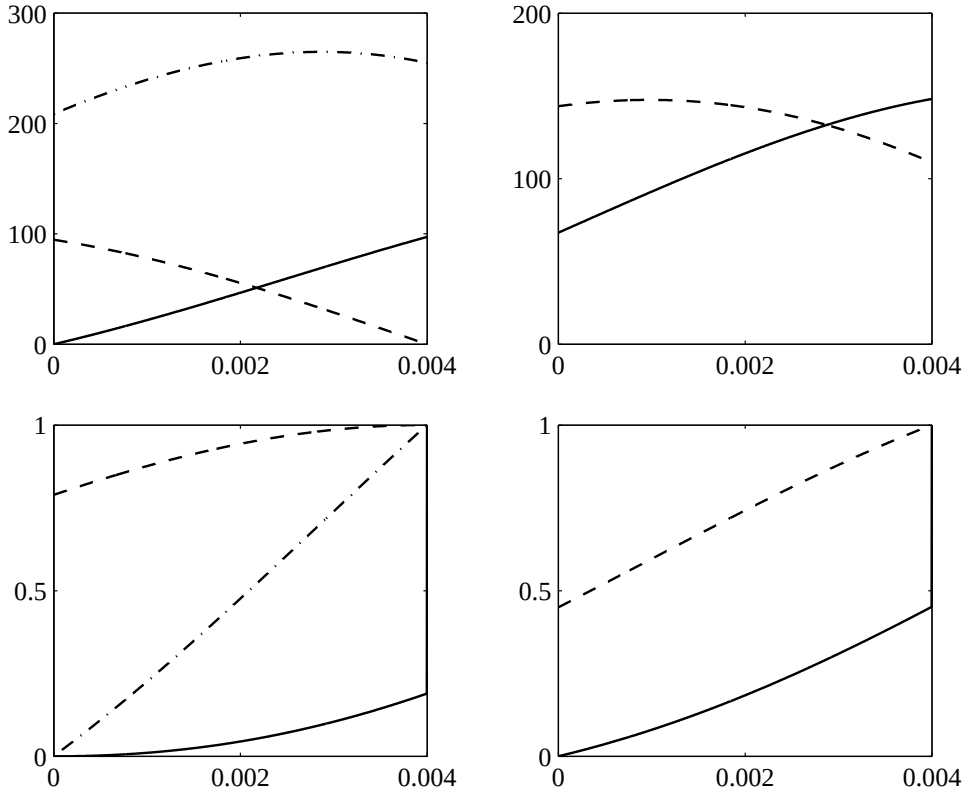

Figure 2: Distributions of occupation time of state one (densities – top, cumulative – bottom) given boundary (left; $\dot{Y}_0 = 1, \dot{Y}_1 = 1$ – solid; $\dot{Y}_0 \neq \dot{Y}_1$ – dash-dotted; $\dot{Y}_0 = 2, \dot{Y}_1 = 2$ – dashed) and initial (right; $\dot{Y}_0 = 1$ – solid; $\dot{Y}_0 = 2$ – dashed) states

6 Numerical results

6.1 Simulated data

First we consider 100 samples of simulated data ($n = 1$, $d = 2$, $N = 5000$, $\Delta t = 1/250$) for moderate speed of state switching. To avoid a bias, the parameters for the prior distributions (according to (4)) are chosen such that the mean values are the true ones, while the variances are large ($\mathrm{Var}(\mu^{(k)}) = 16$, $\mathrm{Var}(\sigma^{(k)}) = 0.0025$, $\mathrm{Var}(Q_{kl}) = 120$).

We compare three methods: a sampler based on time discretization and data augmentation with the state-vector (referred to as DT), a continuous time sampler based on data augmentation with the full state process (referred to as CT), and the semi-continuous method (referred to as SC) to estimate parameters; both DT and CT are described in (Hahn et al. 2007). The samplers were slightly adapted to make them fully comparable with respect to number of samples needed and run-time (see Remark 5).

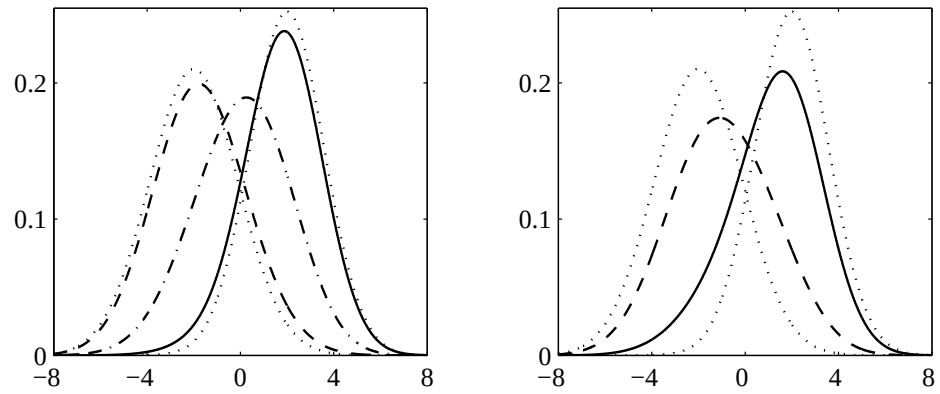

Figure 3: Distributions of observations given boundary (left; $\dot{Y}_0 = 1, \dot{Y}_1 = 1$ – solid; $\dot{Y}_0 \neq \dot{Y}_1$ – dash-dotted; $\dot{Y}_0 = 2, \dot{Y}_1 = 2$ – dashed) and initial (right; $\dot{Y}_0 = 1$ – solid; $\dot{Y}_0 = 2$ – dashed) states, and assuming that no jumps occur (dotted)

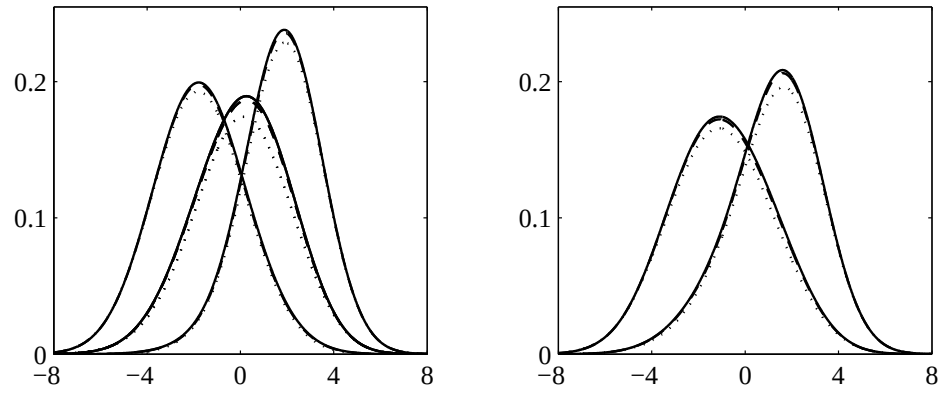

Figure 4: Approximation error for distributions of observations given boundary (left) and initial states (right): 3 summands, 3 integration points (dotted), 4 summands, 5 integration points (dashed), true (solid)

	$\mu^{(1)}$	$\mu^{(2)}$	$\sigma^{(1)}$	$\sigma^{(2)}$	λ_1	λ_2
True	3.00	-3.00	0.100	0.100	60.0	60.0
Discrete time approximation (DT)						
Mean	2.74	-2.70	0.107	0.105	51.1	50.7
Std. dev.	0.04	0.05	0.002	0.002	2.2	2.2
RMSE	0.26	0.30	0.007	0.005	9.2	9.5
Continuous time method (CT)						
Mean	3.03	-2.99	0.101	0.100	62.6	62.4
Std. dev.	0.05	0.05	0.001	0.001	3.1	3.5
RMSE	0.06	0.06	0.002	0.001	4.1	4.2
Semi-continuous approach (SC)						
Mean	3.03	-3.01	0.100	0.098	62.6	62.9
Std. dev.	0.04	0.03	0.002	0.002	3.7	2.9
RMSE	0.05	0.03	0.002	0.003	4.4	4.0

Table 1: Results for moderate state-switching

In Table 1, for each method we give means, standard deviations, and root mean square errors of the mean estimates.

For DT as well as SC, 100 000 samples (with a burn-in of 20 000) were sufficient to give stable results; for CT, 5 000 000 samples (1 000 000 burn-in) turned out to be necessary.

Remark 5. Note that these numbers are not "full" updates of the parameters for the following reason. For CT, updates for the state process are rejected more often the longer the state process is. Hence, to obtain a reasonable acceptance ratio, only a randomly chosen block is updated each time (cf. Hahn et al. 2007). In this example, on average about 1 % of the state process is updated. To facilitate comparison, we proceed similarly for DT and SC. As the current state process is crucial for the updates of the remaining parameters, the crude numbers given above should be divided by 100 to obtain a rough number of "full" updates.

The quality of the estimates is very accurate for CT and SC, while DT introduces some discretization error. Regarding the run-time per sample, CT is fastest; DT takes twice the time, and SC takes ten times longer. Altogether, DT is fastest but not very exact; SC takes five times longer but gives precise results; also giving accurate results, CT mixes much slower and takes 25 times longer.

Next, we turn to a more challenging example with less observations, a lower signal-to-noise-ratio, and very frequent state switching. Again we consider 100 samples of simulated data ($n = 1$, $d = 2$, $N = 2500$, $\Delta t = 1/250$); prior distributions are set

	$\mu^{(1)}$	$\mu^{(2)}$	$\sigma^{(1)}$	$\sigma^{(2)}$	λ_1	λ_2
True	2.00	-2.00	0.110	0.090	220.0	280.0
Discrete time approximation (DT)						
Mean	1.49	-1.35	0.113	0.097	147.6	191.7
Std. dev.	0.08	0.08	0.003	0.003	11.9	16.7
RMSE	0.52	0.65	0.005	0.008	73.3	89.7
Continuous time method (CT)						
Mean	1.88	-1.93	0.114	0.087	199.5	267.6
Std. dev.	0.36	0.17	0.007	0.007	50.4	41.8
RMSE	0.36	0.17	0.008	0.007	52.0	41.5
Semi-continuous approach (SC)						
Mean	1.69	-2.13	0.116	0.087	179.4	287.0
Std. dev.	0.13	0.15	0.004	0.005	14.4	11.8
RMSE	0.34	0.20	0.007	0.006	43.1	13.7

Table 2: Results for fast state-switching

such that the mean values are the true ones but with large variances ($\mathrm{Var}(\mu^{(k)}) = 16$, $\mathrm{Var}(\sigma^{(k)}) = 0.0025$, $\mathrm{Var}(Q_{kl}) = 1000$); the different samplers are run. Results for 200 000 samples per run for DT and SC and for 5 000 000 samples for CT (where again Remark 5 holds) are given in Table 2.

DT introduces a considerable discretization error; still, the results are acceptable in the sense that they give an idea about the magnitudes of and the relations between the parameters. For SC, with a total run-time that is about five times higher, results are much better and the RMSEs are substantially reduced compared to DT. As the volatility is higher for the first state, the corresponding estimates are less accurate than those for the second state. Nevertheless, the estimates are of good quality for all parameters. Also CT yields accurate results, however, the total runtime required is 60 times longer than DT and 12 times higher than SC.

6.2 Market data

As an example for financial market data we consider daily data for 20 stocks of the Dow Jones Industrial Average Index (DJIA) from 1972 to 2000. The data was organized in 25 overlapping blocks of 5 consecutive years' quotes, resulting in a total of 500 data sets each comprising about $N = 1260$ data points, where $\Delta t = 1/252$ and $T = 5$. This data set was considered in (Hahn et al. 2008b) for portfolio-optimization; there, DT was used for parameter estimation.

Here, for each of these 500 data sets we employed DT as well as SC to fit a sim-

ple 2 states HMM, that is, we assume to observe up and down periods for the drift but constant volatility. Obviously, this model is not sufficient to give good pathwise fits, however, it is already a serious improvement over the Black-Scholes model (where constant drift is assumed) and can lead to substantially better results e.g. in portfolio-optimization (cf. Hahn et al. 2008b; Sass and Haussmann 2004).

Prior distributions are chosen according to (4); for each data set, means for the drifts m_1 and m_2 are set to the 60 % and 40 % quantiles of the daily returns under consideration, and standard deviations are set to $s_1 = s_2 = (m_1 - m_2)/2$; for the volatility, the mean equals 0.9 times the standard deviation of the returns and the standard deviation equals half of the mean. For the rates, for all data sets we set means to 80 and standard deviations to 40, meaning that we expect average drift regimes to last between 2 and 6 trading days.

For DT as well as SC, 500 000 samples (with a burn-in of 100 000) were sufficient for all data sets to give stable results (where in each update step about 2 % of the state process is updated, compare Remark 5; so, roughly speaking, we needed about 10 000 "full" samples).

We also tried to employ CT, however, for almost all sets we obtained no plausible results; no well separated drifts were identified and often enormous rates were estimated. This is not surprising, since (according to the results of DT and SC) state switching occurs frequently; drifts for up and down states are close and volatility is high (i.e. the signal-to-noise ratio is much lower); moreover, only relatively few data points are used.

Figure 5 gives an overview of the results for all the 500 data sets, showing the distributions of the sample means both for SC and DT. For SC (DT), drifts are centered around 1.3 (0.93) and −0.78 (−0.44), volatility around 0.25 (0.26), and rates around 102 (92) and 89 (83), the latter meaning that on average we expect a change in the drift every 2–3 trading days. Average sample standard deviations for SC (DT) are around 0.50 (0.44) and 0.40 (0.34) for the drifts, 0.006 (0.006) for the volatility, and 39 (40) and 35 (37) for the rates. These results look quite plausible and reliable.

Clearly, this is only a very rough analysis of the results, however, we are less interested in details on the concrete results for each of the 500 data sets, but more in the comparison of the results of DT and SC. To this end, for each run we compute the relative deviation of the results of DT from SC; i.e. for some parameter θ we take mean estimates $\bar{\theta}^{\mathrm{DT}}$ and $\bar{\theta}^{\mathrm{SC}}$ from the output of the samplers DT and SC; then we plot the distribution of the relative deviation $(\bar{\theta}^{\mathrm{SC}} - \bar{\theta}^{\mathrm{DT}})/|\bar{\theta}^{\mathrm{SC}}|$. In Figure 6 these distributions for the different parameters over all data sets (as well as corresponding average values) are shown. Comparing DT to SC, on average the drift states are about 40 % less extreme, while the volatility is only 1.5 % higher; the rates are about 8 % lower. This fits perfectly with the results for simulated data, where SC yielded much more accurate results than DT; in many applications, this enhanced accuracy will justify the higher computational costs. Especially when challenging parameters are expected, SC should be the preferred estimation method.

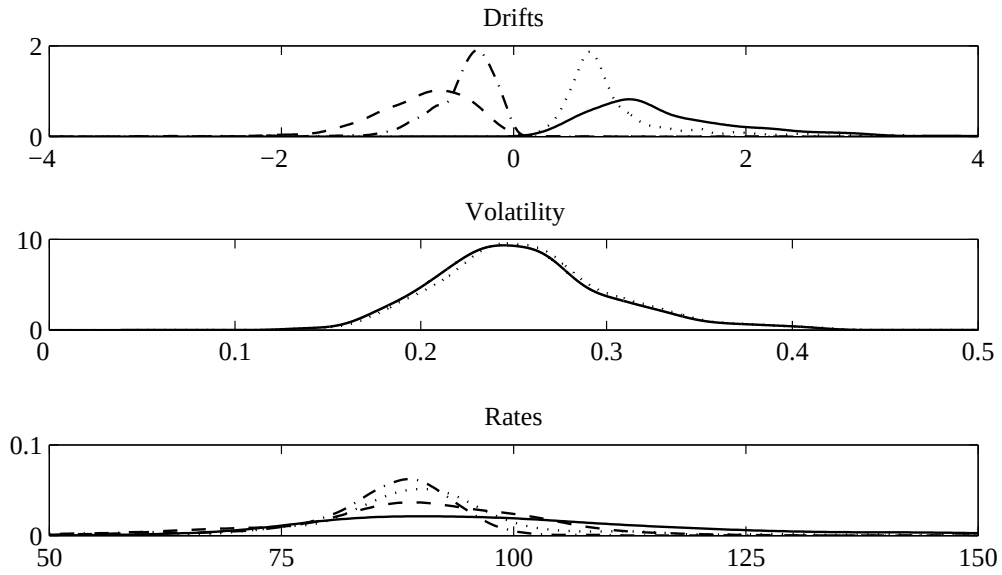

Figure 5: Distribution of sample means of SC and DT (solid: SC state 1, dotted: DT state 1, dashed: SC state 2, dash-dotted: DT state 2)

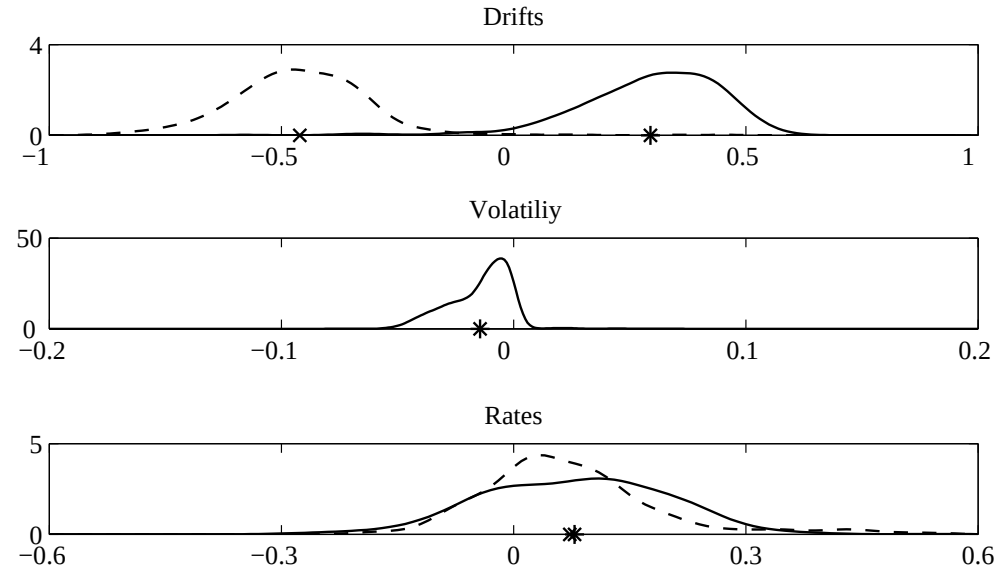

Figure 6: Distribution of deviations of results of DT and SC relative to SC (solid: state 1, dashed: state 2; crosses and asterisks indicate average values)

7 Conclusion

We present a so-called semi-continuous time MCMC algorithm for parameter estimation in multi-dimensional continuous time MSMs. This problem is especially difficult if few data points are available and the switching of the states happens fast. While methods of moments yield precise results for big data sets (say, more than 30 000 data points) very fast, we concentrate on the case where not more than 5 000 data points are available. Moreover, in many applications one observes fast regime switching (which makes the expectation maximization algorithm fail) and a low signal-to-noise ratio.

The resulting semi-continuous algorithm is a stable method that mixes quickly, as the dimension of the augmented variables as well as the dependence between the parameters is reduced significantly compared to augmenting with the full state process. In comparison to discrete time approximations, we still can employ exact filtering for the states, but no approximation error is introduced and the generator matrix is estimated directly rather than estimating the transition matrix. The numerical computations can be performed to arbitrary precision. The quality of the results is quite satisfying, and especially for fast state switching the higher computational costs pay off definitely.

While the theory presented here applies for an arbitrary number of states, due to computational complexity, it is applicable for a moderate number of states only; it remains for future work to devise a fast and efficient implementation to deal with high numbers of states.

The transfer of the semi-continuous method to other Markov switching models is straightforward whenever the distribution of the observed data depends rather on the occupation times than on the full path of the underlying Markov process. Otherwise, the semi-continuous approach still may be beneficial, but calculations will get more involved.

8 Appendix

In the following we describe how to compute $\mathrm{P}(\Delta\dot{R}_m \,|\, \dot{Y}_{m-1} = k, \dot{Y}_m = l, \eta)$, the conditional distributions of the observed increments $\Delta\dot{R}_m$ given the boundary states $\dot{Y}_{m-1}$, $\dot{Y}_m$, for $d > 2$.

For an index set $I \in \mathcal{I}$, where $\mathcal{I} = \{I : I \subsetneq \{1, \dots, d\}\}$, we define

$$\mathcal{D}_I^0 = \{\tau \in \mathcal{D} \,|\, \tau_k = 0 \text{ for } k \in I,\ \tau_k > 0 \text{ for } k \notin I\}, \tag{36}$$

the set of possible occupation times τ when states in I are not visited, but all others are. Note that $\mathcal{D}_\emptyset^0 = \operatorname{ri}\mathcal{D}$ and $\mathrm{P}(O_{\Delta t} \in \mathcal{D}_I^0 \,|\, Y_0 = k, Y_{\Delta t} = l) > 0$ if $k, l \notin I$, i.e. in general there is positive probability for occupation time 0 of single states, although these sets are of lower effective dimension. Hence, while Theorem 1 only tells us how to deal with $\operatorname{ri}\mathcal{D}$, we also need to consider the boundary $\operatorname{rb}\mathcal{D}$ explicitly. Note that as in Section 5, we assume Q to be given.

The idea is to split $\mathcal{D}$ into its interior and parts of its boundary. For the interior, we

need to solve a $(d-1)$-dimensional integral. For the parts of the boundary, the dimension is lower, but suitably adapting Q, we can proceed similarly as for the interior. In more detail, we use the representation $\mathcal{D} = \bigcup_{I\in\mathcal{I}} \mathcal{D}_I^0$ to evaluate the integral (21) conditioning on which states have occupation time 0, i.e. $O_{\Delta t} \in \mathcal{D}_I^0$. We obtain

$$\int_{\mathcal{T}} \psi(\Delta\dot{R}_m;\, B, C, \tau)\, \mathrm{d}F^{\mathrm{bd}}_{k,l}(\tau) = \sum_{I\in\mathcal{I}} \int_{\mathcal{D}_I^0} \psi(\Delta\dot{R}_m;\, B, C, \tau)\, \mathrm{d}F^{\mathrm{bd}}_{I;k,l}(\tau)\ \mathrm{P}(O_{\Delta t} \in \mathcal{D}_I^0 \,|\, Y_0 = k, Y_{\Delta t} = l), \quad (37)$$

where

$$F^{\mathrm{bd}}_{I;k,l}(\tau) = \mathrm{P}(O_{\Delta t} \le \tau \,|\, Y_0 = k, Y_{\Delta t} = l, O_{\Delta t} \in \mathcal{D}_I^0). \quad (38)$$

This removes difficulties arising from discontinuities of F^{bd} on $\mathrm{rb}\,\mathcal{D}$. In fact, we get a weighted sum of integrals over relatively open domains $\mathcal{D}_I^0$ of dimension $d - |I| - 1$ (except for $|I| = d-1$). The weights for each integral, $\mathrm{P}(O_{\Delta t} \in \mathcal{D}_I^0 \,|\, Y_0 = k, Y_{\Delta t} = l)$, are obtained from (Sericola 2000, Theorem 3.3). The corresponding conditional densities f_I^{bd} can be derived from Theorem 1.

Corollary 3. *For $I \in \mathcal{I}$, $|I| < d-1$, F_I^{bd} has a continuous density f_I^{bd} on $\mathcal{D}_I^0$.*

For sets I, $0 < |I| < d-1$, replace d with $d_I = d - |I|$ and adapt Q as follows. Reset the diagonal entries Q_{kk} to $-\sum_{l\notin I, l\neq k} Q_{kl}$ and drop lines and columns i for $i \in I$, which gives the corresponding $d_I \times d_I$ rate matrix. Accordingly adapt parameters of the uniformized Markov chain. In τ, components τ_k, where $k \in I$, are dropped.

Then, w.l.o.g. we can assume $I = \emptyset$. For $\tau \in \mathcal{D}_I^0$, we have

$$f^{\mathrm{bd}}_{I;k,l}(\tau) = \sum_{m=d-1}^{\infty} e^{-\tilde{\lambda}\,\Delta t} \frac{(\tilde{\lambda}\,\Delta t)^m}{m!} \sum_{v^*\in\mathcal{V}_m^+} \chi^{\Delta t;\tau^*}_{m;v^*}\ \mathrm{P}(\mathbf{1}_{d-1} \le \tilde{V}_m^* \le v^*, \tilde{Z}_m = l \,|\, \tilde{Z}_0 = k)/X_{kl}, \quad (39)$$

where $\mathbf{1}_{d-1}$ denotes the $(d-1)$-dimensional vector of ones and

$$\mathcal{V}_m^+ = \{(v_1, \ldots, v_{d-1}) \,|\, v_i \ge 1, v_1 + \cdots + v_{d-1} \le m\}. \quad (40)$$

For $|I| = d-1$, i.e. no jumps occur, we have

$$\mathrm{P}(O_{\Delta t} = u^j \Delta t \,|\, Y_0 = k, Y_{\Delta t} = l, O_{\Delta t} \in \mathcal{D}_I^0) = \mathbb{1}_{\{j=k, j=l, j\notin I\}}, \quad (41)$$

where u^j denotes the j-th d-dimensional unit vector.

Finally, we rewrite Equation (37) to obtain a numerically computable form for Equation (21).

Corollary 4. *Combining Equation* (37) *and Corollary 3, we obtain*

$$\begin{aligned}\mathrm{P}(\Delta\dot{R}_m \,|\, \dot{Y}_{m-1} = k, \dot{Y}_m = l, \eta) = \\ \sum_{\substack{I\in\mathcal{I}\\ |I|<d-1}} \int_{\mathcal{D}_I^0} \psi(\Delta\dot{R}_m;\, B, C, \tau)\, f_{I;k,l}^{\mathrm{bd}}(\tau)\,\mathrm{d}\tau\ \mathrm{P}(O_{\Delta t} \in \mathcal{D}_I^0 \,|\, Y_0 = k, Y_{\Delta t} = l) \\ + \mathbb{1}_{\{k=l\}}\, \psi(\Delta\dot{R}_m;\, B, C, u^k\Delta t)\, e^{-\lambda_k \Delta t}/X_{kk}. \quad (42)\end{aligned}$$

Remark 6. Finally, let us consider the numerical complexity of evaluating (42). The set of summands $\mathcal{V}_m^+$ in (40) contains $|\mathcal{V}_m^+| = \binom{m}{m+1-d}$ elements. Cutting the infinite summation at some M, we have to sum up $\binom{M+1}{M+1-d}$ terms to obtain $f_\emptyset^{\mathrm{bd}}$. To obtain all f_I^{bd} for $I \in \mathcal{I}$, we need to consider a total of $\sum_{d_I=0}^{d-1} \binom{M+1}{M+1-(d-d_I)}\binom{d}{d_I} = \binom{M+1+d}{M+1} - 1$ terms.

For the numerical integration, consider the following. Using some fixed grids, the terms $\chi_{m;v^*}^{\Delta t;\tau^*}$ are set up only once. As $\chi_{m;v^*}^{\Delta t;\tau^*}$ are rather smooth polynomials in τ, numerical integration works very well even for simple schemes with a low number of integration points.

Summing up, as integrals of dimensions up to $(d-1)$ need to be evaluated numerically, whose densities get more complex with increasing d, for an increasing number of states, using a straightforward implementation the computations get very time-consuming. However, for a moderate number of states computations can be carried out within reasonable time.

References

Ball, F. G., Cai, Y., Kadane, J. B., and O'Hagan, A. (1999). "Bayesian inference for ion-channel gating mechanisms directly from single-channel recordings, using Markov chain Monte Carlo." *Proceedings of the Royal Society Series A*, 455.

Brémaud, P. (1981). *Point Processes and Queues*. New York: Springer.

Buffington, J. and Elliott, R. J. (2002). "American options with regime switching." *International Journal of Theoretical and Applied Finance*, 5: 497–514.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society Series B*, 39(1): 1–38.

Elliott, R. J., Aggoun, L., and Moore, J. B. (1995). *Hidden Markov Models – Estimation and Control*, volume 29 of *Applications of mathematics*. New York: Springer.

Elliott, R. J., Hunter, W. C., and Jamieson, B. M. (1997). "Drift and volatility estimation in discrete time." *Journal of Economic Dynamics and Control*, 22: 209–218.

Elliott, R. J., Krishnamurthy, V., and Sass, J. (2008). "Moment based regression algorithm for drift and volatility estimation in continuous time Markov switching models." *Econometrics Journal*, 11: 244–270.

Engel, C. and Hamilton, J. D. (1990). "Long swings in the dollar: Are they in the data and do markets know it?" *American Economic Review*, 80(4): 689–713.

Frühwirth-Schnatter, S. (2001). "Fully Bayesian analysis of switching Gaussian state space models." *Annals of the Institute of Statistical Mathematics*, 53(1): 31–49.

— (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

Guo, X. (2001). "An explicit solution to an optimal stopping problem with regime switching." *Stochastic Processes and their Applications*, 38: 464–481.

Hahn, M., Frühwirth-Schnatter, S., and Sass, J. (2007). "Markov chain Monte Carlo methods for parameter estimation in multidimensional continuous time Markov switching models." Technical Report 2007-09, Johann Radon Institute for Computational and Applied Mathematics.

— (2008a). "Estimating models based on Markov jump processes given fragmented observation series." Submitted.

Hahn, M., Putschögl, W., and Sass, J. (2008b). "Optimizing consumption and investment: The case of partial information." In Kalcsics, J. and Nickel, S. (eds.), *Operations Research Proceedings 2007*, 57–62. German Operations Research Society, Berlin: Springer.

Israel, R. B., Rosenthal, J. S., and Wei, J. Z. (2001). "Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings." *Mathematical Finance*, 11: 245–265.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling." *Statistical Science*, 20(1): 50–67.

Liechty, J. C. and Roberts, G. O. (2001). "Markov chain Monte Carlo methods for switching diffusion models." *Biometrika*, 88(2): 299–315.

Otranto, E. and Gallo, G. M. (2002). "A nonparametric Bayesian approach to detect the number of regimes in Markov switching models." *Econometric Reviews*, 21(4): 477–496.

Ross, S. M. (1996). *Stochastic Processes*. New York: John Wiley & Sons, 2nd edition.

Rydén, T., Teräsvirta, T., and Åsbrink, S. (1998). "Stylized facts of daily return series and the hidden Markov models." *Journal of Applied Econometrics*, 13: 217–244.

Sass, J. and Haussmann, U. G. (2004). "Optimizing the terminal wealth under partial information: The drift process as a continuous time Markov chains." *Finance and Stochastics*, 8(4): 553–577.

Sericola, B. (2000). "Occupation times in Markov processes." *Communications in Statistics – Stochastic Models*, 16(5): 479–510.

Timmermann, A. (2000). "Moments of Markov switching models." *Journal of Econometrics*, 96: 75–111.

Acknowledgments

The authors acknowledge support by the Austrian Science Fund FWF, project P17947-N12, and by the Heisenberg Programme of the German Research Foundation DFG. We also thank the referee, the Associate Editor, and the Editor for their comments and suggestions which led to a considerable improvement of the paper.