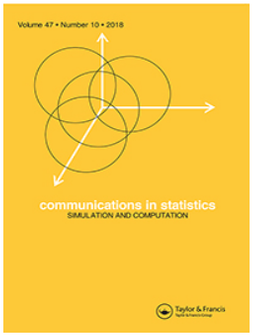

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

MCMC for Markov-switching models—Gibbs sampling vs. marginalized likelihood

Kjartan Kloster Osmundsen, Tore Selland Kleppe & Atle Oglend

To cite this article: Kjartan Kloster Osmundsen, Tore Selland Kleppe & Atle Oglend (2019): MCMC for Markov-switching models—Gibbs sampling vs. marginalized likelihood, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1565580

To link to this article: https://doi.org/10.1080/03610918.2019.1565580

Published online: 31 Jan 2019.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1565580

MCMC for Markov-switching models—Gibbs sampling vs. marginalized likelihood

Kjartan Kloster Osmundsen[a], Tore Selland Kleppe[a], and Atle Oglend[b]

[a]Department of Mathematics and Physics, University of Stavanger, Stavanger, Norway; [b]Department of Safety, Economics and Planning, University of Stavanger, Stavanger, Norway

ABSTRACT

This article proposes a method to estimate Markov-switching vector autoregressive models that combines (integrated over latent states) marginal likelihood and Hamiltonian Monte Carlo. The method is compared to commonly used implementations of Gibbs sampling. The proposed method is found to be numerically robust, flexible with respect to model specification and easy to implement using the Stan software package. The methodology is illustrated on a real data application exploring time-varying cointegration relationships in a data set consisting of crude oil and natural gas prices.

ARTICLE HISTORY
Received 31 August 2017
Accepted 29 November 2018

KEYWORDS
Hamiltonian Monte Carlo;
Markov-switching;
Marginalized likelihood;

1. Introduction

A Markov-switching (MS) model (also called hidden Markov model) is a mixture model governed by a (hidden) finite state Markov chain. It has a wide range of applications, and has successfully been used in fields such as speech recognition (Baker 1975) and image analysis (Yamato, Ohya, and Ishii 1992). In economics, MS models have also been widely applied as parsimonious specifications of non-linear time series dynamics since the seminal article by Hamilton (1989). Hamilton realized that the changing nature of contractions and expansions in economic activity can be modeled as an MS model where the growth rate of the economy is determined by a time-varying latent state. Other business cycle applications can be found in Ang and Bekaert (2002) and Bansal, Tauchen, and Zhou (2004). In addition to business cycles, MS models have also been applied to model interest rate dynamics (Garcia and Perron 1996; Gray 1996; Bansal and Zhou 2002), electricity pricing (Mount, Ning, and Cai 2006; Kanamura and Ōhashi 2008) and oil and natural gas pricing (Brigida 2014; Asche, Oglend, and Osmundsen 2017). The reason for the popularity of MS models in economics is the parsimonious yet flexible nature of the models, and the relative ease of providing an economic interpretation of the underlying states or regimes implied by the model.

Sims and Zha (2006) applies a Markov-switching vector autoregressive (MS-VAR) model to investigate structural breaks in monetary policies, and in a highly cited article, Bloom (2009) uses an MS specification to model shocks of uncertainty in a model of economic activity. The success of the MS framework in economics has spurred interest

CONTACT Kjartan Kloster Osmundsen kjartan.k.osmundsen@uis.no Department of Mathematics and Physics, University of Stavanger, Stavanger 4036, Norway

in the estimation of the models. Related to this article, Sims, Waggoner, and Zha (2008) discuss methods for inference in large multiple-equation MS models. Also, Lanne, Lütkepohl, and Maciejowska (2010) investigate structural MS-VAR models, while Bianchi (2016) develops methods to analyze multivariate MS models, specifically formulas for the evolution of first and second moments.

This article contributes to the growing literature on estimating MS-VAR models by comparing a proposed marginalized likelihood Markov chain Monte Carlo (MCMC) estimator to commonly applied Gibbs sampling. The two methods are distinguished by whether the hidden regime variables (discrete) are sampled or marginalized out from the likelihood function, with the former currently being the preferred method. However, the marginalized method lowers the dimension of the target distribution, enabling most general purpose MCMC procedures, such as Metropolis-Hastings (Robert and Casella 2013). The marginalized method also generates a continuous target distribution, enabling Hamiltonian Monte Carlo (HMC) (Neal 2011). By applying HMC with the no-U-turn sampler of Hoffman and Gelman (2014), fast exploration of the posterior distribution is achieved, resulting in close to iid samples. This can be seen as an MS analog to the currently popular pseudo-marginal methods (see e.g. Andrieu, Doucet, and Holenstein 2010) for state-space models and other models where latent states take continuous values.

The article compares the efficiency of the two methods numerically, using both simulated data and real data sets. The MCMC efficiency is measured as the effective sample size (Geyer 1992; Girolami and Calderhead 2011) per second (ESS/s). The experiments are conducted using appropriate statistical software packages for the two methods, and also using tailor-made Gibbs samplers. Using statistical software packages, the marginalization approach produces more efficient samples than Gibbs sampling. The tailor-made Gibbs sampler implementations are very fast, but require greater coding efforts and are less flexible to changes of the model such as parameter restrictions. The marginalization approach also gives stable performance across parameters, unlike the Gibbs implementations.

The article is organized as follows. Section 2 describes the MS-VAR model and introduces the notation used throughout the article. Moreover, the two different sampling methods and their implementations are discussed in detail. Section 3 compares the proposed methodology to relevant alternatives. Section 4 applies the sampling methods to estimate an MS-VAR model on the joint dynamics of crude oil and natural gas prices, extending the modeling approach in Asche, Oglend, and Osmundsen (2017) to a fully flexible bivariate MS-VAR. Final conclusions are given in Sec. 5.

2. Methodology

The methods presented in this section will be used for estimating MS-VAR models. The methods are not limited to the following model, but due to the challenges faced by estimating these types of models and the recent popularity of the models in econometric time series modeling, the generic MS-VAR model is used as a basis for comparing the different estimation methods.

2.1. Markov-switching vector autoregressive models

A Markov-switching vector autoregressive model with one lag may be expressed as

$$\begin{aligned} \boldsymbol{Y}_t &= \phi_{(S_t)}\boldsymbol{Y}_{t-1} + \boldsymbol{\mu}_{(S_t)} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_{(S_t)}\right), \quad \boldsymbol{Y}_t \in \mathbb{R}^d, \\ & S \in (1, 2, ..., m), \quad t \in (2, 3, ..., n), \end{aligned} \tag{1}$$

where $\boldsymbol{Y}_t, t \in (1, 2, 3, ..., n)$ are d-dimensional observations, n is the number of such observations, $\phi_{(S)}, S \in (1, 2, ..., m)$ are the autoregressive coefficient matrices, m is the number of states and $\boldsymbol{\mu}_{(S)}, S \in (1, 2, ..., m)$ are the mean vectors. $S_t \in (1, 2, ..., m), t \in (2, 3, ..., n)$ are the latent state variables, which follow a first-order time homogeneous Markov Chain characterized by a transition probability matrix:

$$\boldsymbol{Q} = \begin{bmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & \vdots & \vdots \\ p_{m1} & \cdots & p_{mm} \end{bmatrix}, \quad p_{ij} = p\left(S_t = j | S_{t-1} = i\right).$$

Here it is assumed that $\{S_t\}_t$ is irreducible and aperiodic, and thus admit a stationary distribution $\boldsymbol{\delta}$. The first observation is treated as known and hence is not modeled. For notational convenience, only one lag is considered in the discussion in this section, but extensions to several lags are straightforward.

For the collection of parameters $\boldsymbol{\Theta} = (\boldsymbol{Q}, \phi_{(1)}, \phi_{(2)}, ..., \phi_{(m)}, \boldsymbol{\mu}_{(1)}, \boldsymbol{\mu}_{(2)}, ..., \boldsymbol{\mu}_{(m)}, \boldsymbol{\Sigma}_{(1)}, \boldsymbol{\Sigma}_{(2)}, ..., \boldsymbol{\Sigma}_{(m)})$, the likelihood function, conditional on $\boldsymbol{Y}_1$, is given by

$$\begin{aligned} l(\boldsymbol{\Theta}) &= p(\boldsymbol{Y}_{2:n} | \boldsymbol{\Theta}, \boldsymbol{Y}_1) = \boldsymbol{\delta}\boldsymbol{P}(\boldsymbol{Y}_2)\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{Y}_3)\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{Y}_4)\cdots\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{Y}_n)1^\top, \\ \boldsymbol{P}(\boldsymbol{Y}_t) &= \text{diag}\left(\left\{p\left(\boldsymbol{Y}_t | \boldsymbol{Y}_{t-1}, \boldsymbol{\Theta}_{-\boldsymbol{Q}}, S_t = j\right)\right\}_{j=1}^m\right), \end{aligned} \tag{2}$$

where $\boldsymbol{Y}_{1:n} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_n)$, $1^\top$ is a row vector of length m where all elements equal 1 and $\boldsymbol{\Theta}_{-\boldsymbol{Q}}$ denotes the parameter collection excluding the latent Markov chain transition probability matrix.

2.2. Sampling methods for Θ

Given a prior distribution for the collection of parameters $\boldsymbol{\Theta}$, say $p(\boldsymbol{\Theta})$, the posterior distribution for $\boldsymbol{\Theta}$ has the following form:

$$p(\boldsymbol{\Theta} | \boldsymbol{Y}_{1:n}) \propto l(\boldsymbol{\Theta})p(\boldsymbol{\Theta}) = \left[\int p(\boldsymbol{Y}_{2:n} | \boldsymbol{S}_{2:n}, \boldsymbol{\Theta}, \boldsymbol{Y}_1)p(\boldsymbol{S}_{2:n} | \boldsymbol{\Theta})d\boldsymbol{S}_{2:n}\right]p(\boldsymbol{\Theta}). \tag{3}$$

The primary objective of this article is MCMC sampling of the posterior distribution of $\boldsymbol{\Theta}$. Two approaches are considered to this end. First, a marginalized (over the latent states $\boldsymbol{S}_{2:n}$) approach which targets (3) directly. This approach is not new per se (see e.g. Scott 2002, Sec. 2 for a discussion), but here the marginalized approach is combined with the no-U-turn sampler of Hoffman and Gelman (2014), which can produce close to iid samples while making the user responsible for very modest coding efforts. Second, variants of Gibbs sampling, which is the currently preferred method for Bayesian analysis of Markov-switching models (Scott 2002), are used as references for the marginalized approach. The Gibbs sampling variants directly target the latter

representation of the posterior in (3) by sampling both $\boldsymbol{\Theta}$ and $\boldsymbol{S}_{2:n}$, and implement the marginalization by simply disregarding the samples produced for $\boldsymbol{S}_{2:n}$.

It has been argued (see e.g. Andrieu, Doucet, and Holenstein 2010; Scharth and Kohn 2016) that MCMC methods for state space models that target the, typically relatively low-dimensional, collection of parameters directly can lead to better mixing properties of the parameter chains than for Gibbs sampling with parameters and latent states in different blocks. This is a consequence of the typically strong and non-linear dependence between latent states and parameters. Even if the latent vector is not partitioned into smaller blocks, a Gibbs sampler may require a considerable amount of iterations to traverse the joint parameters and latent space. Whether this also holds true for Markov-switching models, where both marginal likelihoods and sampling of $p(\boldsymbol{S}_{2:n}|\boldsymbol{\theta}, \boldsymbol{Y}_{1:n})$ are relatively cheap, is the main question this article sets out to answer. The remainder of this section describes the specific MCMC methodology used to this end.

2.2.1. MCMC based on marginal likelihood

Unlike pseudo-marginal methods for latent variable models, which must rely on computationally expensive unbiased Monte Carlo estimates of the marginal likelihood, the (integrated over latent states) marginal likelihood of a Markov-switching model can be computed analytically. The forward algorithm, detailed in Algorithm 1, is used to calculate pointwise in $\boldsymbol{\Theta}$ the marginal log-likelihood $l(\boldsymbol{\Theta}) = \log p(\boldsymbol{Y}_{2:n}|\boldsymbol{\Theta}, \boldsymbol{Y}_1)$, with complexity $\mathcal{O}(m^2 n)$.

Algorithm 1. Marginal log-likelihood

1: for $1 \leq i \leq m$ **do**
2: $\quad \boldsymbol{\alpha}_{2,i} = \log \boldsymbol{\delta}_{\boldsymbol{i}} + \log p(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\Theta}_{-\boldsymbol{Q}}, S_t = i)$
3: **end for**
4: **for** $3 \leq t \leq n$ **do**
5: $\quad$ **for** $1 \leq i \leq m$ **do**
6: $\quad\quad \boldsymbol{\alpha}_{t,i} = \log\left(\sum_{j=1}^{m} \exp\left(\boldsymbol{\alpha}_{t-1,i} + \log \boldsymbol{Q}_{\boldsymbol{j},\boldsymbol{i}} + \log p(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\Theta}_{-\boldsymbol{Q}}, S_t = i)\right)\right)$
7: $\quad$ **end for**
8: **end for**
9: **return** $l(\boldsymbol{\Theta}) = \log p(\boldsymbol{Y}_{2:n}|\boldsymbol{\Theta}, \boldsymbol{Y}_1) = \log \sum_{i=1}^{m} \exp\left(\boldsymbol{\alpha}_{n,i}\right)$

Based on the ability to compute the posterior log kernel $\log p(\boldsymbol{Y}_{2:n}|\boldsymbol{\Theta}, \boldsymbol{Y}_1) + \log p(\boldsymbol{\Theta})$, most general purpose MCMC methods, such as random walk Metropolis-Hastings and Metropolis-adjusted Langevin algorithms (Robert and Casella 2013), can in principle be applied. As the dimension of $\boldsymbol{\Theta}$ is often rather large (e.g. 20 for an unrestricted model with $m = d = 2$), it suits Hamiltonian Monte Carlo (HMC) (Neal 2011), which exploits gradient information from the posterior log kernel to generate proposals that admit fast exploration of the posterior. Specifically, HMC is used together with the no-U-turn sampler of Hoffman and Gelman (2014); an extension to HMC that eliminates the need for user-specified integration parameters. Marginalizing over the latent states result in no inference for the states themselves, but this can be obtained subsequently using the

Viterbi algorithm (Viterbi 1967), which calculates the most likely state sequence given the observations and the parameter estimates.

2.2.2. Gibbs sampling

The tailor-made Gibbs samplers considered here relies on Gibbs blocks:

- Block 1: $\phi_{(1)}, \phi_{(2)}, ..., \phi_{(m)}, \boldsymbol{\mu}_{(1)}, \boldsymbol{\mu}_{(2)}, ..., \boldsymbol{\mu}_{(m)}, \boldsymbol{\Sigma}_{(1)}, \boldsymbol{\Sigma}_{(2)}, ..., \boldsymbol{\Sigma}_{(m)} | \boldsymbol{Y}_{1:n}, \boldsymbol{S}_{2:n}$. For an unrestricted model (1) with conjugate priors (Gaussian for $\phi_{(S)}$ and $\boldsymbol{\mu}_{(S)}$, and Inverse-Wishart for $\boldsymbol{\Sigma}_{(S)}$), this step can be carried out using Bayesian regression software by treating observations and parameters corresponding to the different regimes separately.
- Block 2: $\boldsymbol{Q} | \boldsymbol{S}_{2:n}$. Based on Dirichlet conjugate priors, this step requires minimal effort.
- Block 3: $\boldsymbol{S}_{2:n} | \boldsymbol{\Theta}, \boldsymbol{Y}_{1:n}$.

Scott (2002) describes two different ways of implementing Block 3; (1) direct Gibbs (DG) sampler, which samples each individual state S_t given the most recent draws of the preceding state S_{t-1} and proceeding state S_{t+1}, and (2) the forward-backward (FB) Gibbs sampler, which uses recursive algorithms to sample the whole state vector $\boldsymbol{S}_{2:n}$ from its joint conditional posterior.

The DG sampler is the simpler of the two methods, drawing each state from its marginal conditional distribution (Albert and Chib 1993):

$$p(S_t | S_{-t}, \boldsymbol{Y}_{1:n}) \propto \begin{cases} \delta_{S_2} p(\boldsymbol{Y}_2 | \boldsymbol{Y}_1, S_2) p(S_3 | S_2) & t = 2, \\ p(S_t | S_{t-1}) p(\boldsymbol{Y}_t | \boldsymbol{Y}_{t-1}, S_t) p(S_{t+1} | S_t) & t = 3, 4, 5, ..., n-1, \\ p(S_n | S_{n-1}) p(\boldsymbol{Y}_n | \boldsymbol{Y}_{n-1}, S_n) & t = n, \end{cases}$$

where $S_{-t} = \{S_i : i \neq t\}$. The more sophisticated FB sampler reduces the number of highly correlated variables in the Gibbs Markov chain by sampling the complete state vector $\boldsymbol{S}_{2:n}$ in one operation, resulting in faster convergence. This is accomplished by adopting the following stochastic backward recursion (Chib 1996; Krolzig 1997):

$$p(\boldsymbol{S}_{2:n} | \boldsymbol{Y}_{1:n}) = p(S_n | \boldsymbol{Y}_{1:n}) \prod_{t=1}^{n-2} p(S_{n-t} | S_{n-t+1}, \boldsymbol{Y}_{1:n}).$$

The factor $p(S_n | \boldsymbol{Y}_{1:n})$ is efficiently calculated using the forward algorithm. As the distribution of $p(S_{n-t} | S_{n-t+1}, \boldsymbol{Y}_{1:n})$ is equal to $p(S_{n-t} | S_{n-t+1}, \boldsymbol{Y}_{1:n-t})$ (Kim 1994), it follows that

$$p(\boldsymbol{S}_{2:n} | \boldsymbol{Y}_{1:n}) \propto p(S_n | \boldsymbol{Y}_{1:n}) \prod_{t=1}^{n-2} p(S_{n-t+1} | S_{n-t}) p(S_{n-t} | \boldsymbol{Y}_{1:n-t}),$$

where $p(S_{n-t} | \boldsymbol{Y}_{1:n-t}), t \in (2, 3, ..., n-1)$ are by-products from the calculation of $p(S_n | \boldsymbol{Y}_{1:n})$.

2.3. Implementation

Simulations are conducted using both methods described in Sec. 2.2. The marginalized approach is implemented using Stan (Stan Development Team 2016); a programing language in which the user can code their models in familiar notation, that is transformed to efficient C++ code and compiled into an executable program. In particular, Stan has

automatic routines for tuning the HMC sampler and uses automatic differentiation (Griewank and Walther 2008) to compute the gradient of the log-target. Thus, the fine details of implementing HMC is hidden for the user, who is only responsible for providing prior specifications and specifying the relevant model via the marginal log-likelihood. In this case, the marginal log-likelihood is calculated using Algorithm 1, and is added to Stan using the log probability increment statement "target +=". The Viterbi algorithm may be efficiently implemented in Stan's *generated quantities* block Stan Development Team (2016, Section 9.6).

Gibbs sampling is implemented using a statistical software package called JAGS; *Just Another Gibbs Sampler* (Plummer 2013). JAGS uses Gibbs sampling in the form of univariate slice sampling updates to produce MCMC output, given a model specified in the BUGS language. Notice in particular that JAGS does not exploit conjugacy, and can therefore handle a wide range of models. Moreover, as JAGS uses univariate updates, it is a DG sampler. Both Stan and JAGS are coded in C++, and can both be used through *R* (R Core Team 2016) interfaces.

The efficiency of the Stan and JAGS implementations are also compared to tailor-made Gibbs samplers according to Blocks 1-3 above, where Block 3 is implemented both as DG and FB. To get comparable run times, the tailor-made Gibbs samplers are implemented using the *R* package **Rcpp** by Eddelbuettel and François (2011), which makes it possible to run compiled C++ code in *R* (The *R* interface for Stan is also made using this package). Aside from the mentioned difference in how the latent states are sampled, the DG and FB samplers are identical. The multivariate regressions in Block 1 are carried out using the *R* package **bayesm** by Rossi (2015).

At this point, it is worth mentioning that in many applications, restrictions on the parameters may severely complicate Block 1 of the Gibbs sampler. Such restrictions are routinely imposed, e.g in order to reduce the number of parameters or to carry out hypothesis tests. In the model considered in Sec. 3.3, the mean structure of (1) is shared by all regimes. Subsequently, the (weighted) regressions required to sample $\phi, \boldsymbol{\mu}$ become non-standard and must be coded from scratch. In addition, relaxing the Gaussian assumption on $\boldsymbol{\epsilon}_t$ may also complicate Step 1, as parameter conjugacy is typically lost. It is worth noticing that these complications only applies to Gibbs samplers of the type indicated in Blocks 1-3 above, whereas for Stan and JAGS, changing specification either by parameter restrictions or distributional assumptions does not lead to substantial changes to the code.

2.4. Label switching

Unsupervised Markov-switching models typically result in multimodal posteriors due to label switching; see, e.g., Stephens (2000); Frühwirth-Schnatter (2001); Jasra, Holmes, and Stephens (2005). If identical priors are chosen for the parameters belonging to each state, the posterior likelihood is invariant to relabeling of the states. This means that each state needs to be identifiable in some way to get meaningful results. A common approach is to apply one or more identifiability constraints, usually by ordering on one of the parameters, for example $\mu_{(1),1} < \mu_{(2),1} \ldots < \mu_{(m),1}$. Such orderings are often imposed using auxiliary variables, e.g.: $\mu_{(2),1} = \lambda_1 \mu_{(1),1}, ; \mu_{(3),1} = \lambda_2 \mu_{(2),1}, \ldots; \mu_{(m),1} = \lambda_{m-1} \mu_{(m-1),}\, 1, \quad \lambda_i > 1; \forall i.$

In the frequentist approach, artificial identifiability constraints can be used to break the symmetry in the likelihood (Jasra, Holmes, and Stephens 2005). In the Bayesian

context, constraints that ignore the geometry of the posterior does generally not induce a unique labeling (Frühwirth-Schnatter 2001), so the constraints should be chosen carefully. There has been some skepticism in the literature regarding the effects that identifiability constraints may have on MCMC samplers; see, e.g., Stephens (1997); Celeux, Hurn, and Robert (2000). It is argued that the identifiability constraints should be imposed after the MCMC run has finished, removing any risk of undesired effects on the sampler. In fact, applying constraints post-simulation is equivalent to changing the prior distribution (Stephens 1997). Another benefit of this method is the possibility to check the effects of different identifiability constraints without re-running the MCMC sampler.

In the following numerical comparisons, identifiability constraints are imposed post-simulation, using the *R* package **label.switching** (Papastamoulis 2016). This approach is in particular chosen as it does not interfere with exploiting conjugacy in the tailor-made Gibbs samplers. For the approach based on Stan, it is likely that implementations based on well-chosen identifiability constraints would lead to more efficient sampling by removing multimodality in the target distribution, while requiring minimal coding efforts.

3. Numerical comparison

In this section, the proposed methodology is compared to the relevant alternatives using two models:

- An unrestricted two-dimensional, two-state case of (1).
- A model used in an economic study with quarterly observations by Lanne, Lütkepohl, and Maciejowska (2010), namely a restricted three-dimensional, two-state case of (1), with an additional three autoregressive lags:

$$\boldsymbol{Y}_t = \phi_1 \boldsymbol{Y}_{t-1} + \phi_2 \boldsymbol{Y}_{t-2} + \phi_3 \boldsymbol{Y}_{t-3} + \phi_4 \boldsymbol{Y}_{t-4} + \boldsymbol{\mu} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_{(S_t)}\right), \quad \boldsymbol{Y}_t \in \mathbb{R}^3, \tag{4}$$

where the Markov-switching is confined to the covariance structure. MS-VAR models with this restriction has proven successful in several applications (Sims and Zha 2006; Sims, Waggoner, and Zha 2008; Bloom 2009).

For both models, the numerical comparisons will focus on data sets with transitions between states of low and high volatility, thereby making ordering of the variances a natural identifiability constraint. For simplicity, this ordering is only applied to the first component of the variance, making $\Sigma_{(1),11} < \Sigma_{(2),11}$ the identifiability constraint.

To ensure robustness, different sets of initial values are used for each MCMC chain. Each set of initial values is a random draw from the respective prior distributions, and the same sets are used for each method. However, the main purpose is to compare the efficiency for the stationary part of the MCMC simulations, so a sufficient amount of burn-in is applied to ensure stationarity. 1500 iterations are used for the simulations, of which the first 500 are considered as burn-in iterations. Eight MCMC chains are used, i.e. the same simulation is repeated eight times for each simulation method, resulting in a total of 8000 samples after warm-up. All the simulation methods are implemented with multi-core support and run on a computer with a quad-core processor (Intel Core i5-6500), meaning that eight chains are simulated in about double the computational time of a single chain.

3.1. Prior distributions

Conjugate priors are chosen to simplify the implementation of the tailor-made Gibbs samplers. A Dirichlet(1,1) prior is used for the rows of the transition probability matrix, a $\mathcal{N}(0, 0.2^2)$ prior for each component of the mean vector and a $\mathcal{N}(0, 1)$ prior for each of the elements of the autoregressive coefficient matrix. An Inverse-Wishart $(\boldsymbol{I}_d, d+1)$ prior ($\boldsymbol{I}_d$ is the $d \times d$ identity matrix) is used for the covariance matrices. JAGS only operates with precision matrix, so Wishart priors are used for those, before they finally are inverted to get the covariance matrix. It appears that the Wishart sampler in JAGS is quite limited, so instead of sampling directly from the Wishart distributions, Wishart samples are constructed using the Bartlett decomposition (Kshirsagar 1959).

Stan supports the Inverse-Wishart prior, but the Bartlett decomposition proved to be more efficient. Stan also offers the possibility to use a Cholesky LKJ prior Stan Development Team (2016, Section 59.2) for the correlation matrices, combined with separate priors for the standard deviations. This prior results in roughly 20% less computational time for Stan. To ensure that the results of Stan and JAGS are comparable, the Bartlett decomposition is used for both implementations. However, the reader should keep in mind that the Stan code could have run a bit faster just by changing the covariance prior.

3.2. Unrestricted model

In order to obtain a robust comparison of the methods presented in Sec. 2, both real and simulated data with diverse characteristics are considered. The simulated data set consists of 2500 observations generated using chosen values for all the parameters in (1). Both regimes are chosen to be highly persistent, with $\boldsymbol{Q}_{11} = 0.97$ and $\boldsymbol{Q}_{22} = 0.9$, with the variance of regime 1 chosen to be lower than the variance of regime 2. The complete set of chosen parameter values is included in Table 1. The real data sets contain exchange rates, interest rates and crude oil prices, and are shown in Figure 1 together with the simulated data set.

For the real data sets, log returns of the raw data are used, scaled by 100. The exchange rate data set ranges from January 2010 to November 2016, and includes 1725 observations of the exchange rate for US dollars in Norwegian kroner (NOK) and Swedish kronor (SEK). The interest rate data set includes 1469 observations of the 3-Month Treasury Bill and the 3-Month London Interbank Offered Rate (LIBOR), and ranges from January 2001 to December 2006. The oil price data set consists of 1999 observations of 1 and 5 months NYMEX WTI Crude Oil futures, from January 1984 to March 1992.

The simulated data set is treated as three different data sets with different observation size, namely the first 750, 1500 and 2500 observations of the simulated data set. The estimates of all parameters in $\boldsymbol{\Theta}$ are approximately the same for all four methods mentioned in Sec. 2.3, and can be found in Table A1 in Appendix for the three different observation sizes. Increasing the observation sample size results in less variation for the parameter estimates, as should be expected. Table 1 shows the ESS/s for these estimates. HMC applied to marginal likelihood (ML-HMC) clearly gives better ESS/s than JAGS, and also stands out regarding stable ESS/s values across the different parameters. The other methods display considerable variation in ESS/s levels for the different parameter types, with the autoregressive coefficients proving hardest to sample. Table 1 also shows that the tailor-made Gibbs samplers are extremely fast compared to the general software packages. This big difference

Table 1. The parameter estimates' effective sample size per second, for simulated data sets with three different observation sizes.

	n = 750				n = 1500				n = 2500			
	ML-HMC	JAGS	FB	DG	ML-HMC	JAGS	FB	DG	ML-HMC	JAGS	FB	DG
t(s)	513	243	3	2.3	598	568	4.4	2.8	863	975	6.3	3.9
Q_{11}	16	2	176	119	12	1.4	123	91	8.1	0.8	70	46
Q_{22}	16	0.2	235	130	13	1	205	117	8.2	1.1	139	53
$\mu_{(1),1}$	16	6.1	1317	613	13	3.3	1364	1038	9.3	2.6	870	660
$\mu_{(1),2}$	16	9.2	1666	596	13	4.4	1259	846	9.3	2.7	893	605
$\mu_{(2),1}$	16	6.6	1400	958	13	5.4	1271	887	9.3	2.1	840	604
$\mu_{(2),2}$	16	2.9	1124	395	13	4	1059	635	9.3	1.6	732	363
$\Sigma_{(1),11}$	16	0.3	906	346	12	1.5	569	144	9.3	1	399	180
$\Sigma_{(1),12}$	16	4.4	1626	399	13	1.9	917	167	9.3	1.5	863	550
$\Sigma_{(1),22}$	16	0.6	1337	424	13	1.7	883	175	9.3	1.2	593	169
$\Sigma_{(2),11}$	16	1	639	182	13	1.6	531	148	9.3	1.6	531	139
$\Sigma_{(2),12}$	16	19	1408	524	13	3.4	1456	663	8.4	2.2	995	854
$\Sigma_{(2),22}$	16	5	672	215	13	1.9	726	162	8.8	1.7	441	212
$\phi_{(1),11}$	12	0.4	873	279	8.3	0.4	692	236	6.1	0.2	617	200
$\phi_{(1),12}$	12	0.4	859	284	8.4	0.4	700	239	5.6	0.3	613	227
$\phi_{(1),21}$	10	0.4	1133	348	8.2	0.5	911	679	5.5	0.4	813	574
$\phi_{(1),22}$	10	0.4	1090	319	9	0.4	866	591	6	0.4	776	548
$\phi_{(2),11}$	14	0.5	1622	672	9	2.3	1262	359	6.6	1.5	945	513
$\phi_{(2),12}$	14	7.4	1441	648	9.2	1.8	932	322	6.3	1.5	884	497
$\phi_{(2),21}$	14	2.5	1786	717	9.6	2.6	1510	1511	7.2	1.9	1064	902
$\phi_{(2),22}$	14	0.3	1262	425	8.8	1.5	946	570	6.4	1.4	1010	734

The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as a single chain.

in computational time is also reflected in the ESS/s results. However, the greater performance of the tailor-made Gibbs samplers has to be weighed against the greater coding efforts required, in addition to less flexibility with respect to modeling changes.

Figure 2 shows how the computational time of ML-HMC and JAGS depends on the observation size. It appears to increase linearly for both ML-HMC and JAGS, with a steeper slope for JAGS. It is also quite clear that ML-HMC has a higher variance in computational time for the different MCMC chains, which is due to independent tuning of the integrator step size for each chain. Table 1 shows that ML-HMC produces the best ESS/s, even for the smallest data sets where JAGS runs much faster. This is because the difference in computational time is outweighed by a massive ESS advantage of ML-HMC. Figure 3 illustrates how the observation size affects the median effective sample sizes for two of the parameters in $\boldsymbol{\Theta}$. The ML-HMC samples are close to being perfect samples (ESS equal to the full chain sample size of 1000), and clearly superior to the JAGS samples. There are some variations in the ESS for both parameters with both methods, but it appears to be nearly independent of observation size. This means that larger datasets will further increase the efficiency gap between ML-HMC and JAGS.

For the three real data sets shown in Figure 1, the estimates of all parameters in $\boldsymbol{\Theta}$ can be found in Table A2 in Appendix. The parameter estimates are still approximately the same for the different simulation methods, as for the simulated data set. Table 2 shows the ESS/s values for these estimates, showing that the sampling efficiency is not just dependent on the simulation method used and the observation size; the data set itself also has a considerable impact. Computational time varies substantially for ML-HMC for the three different data sets, even though they have similar observation sizes.

Anyhow, the overall performance of the four different methods matches the findings for the simulated data set: Using statistical software packages, the marginalization

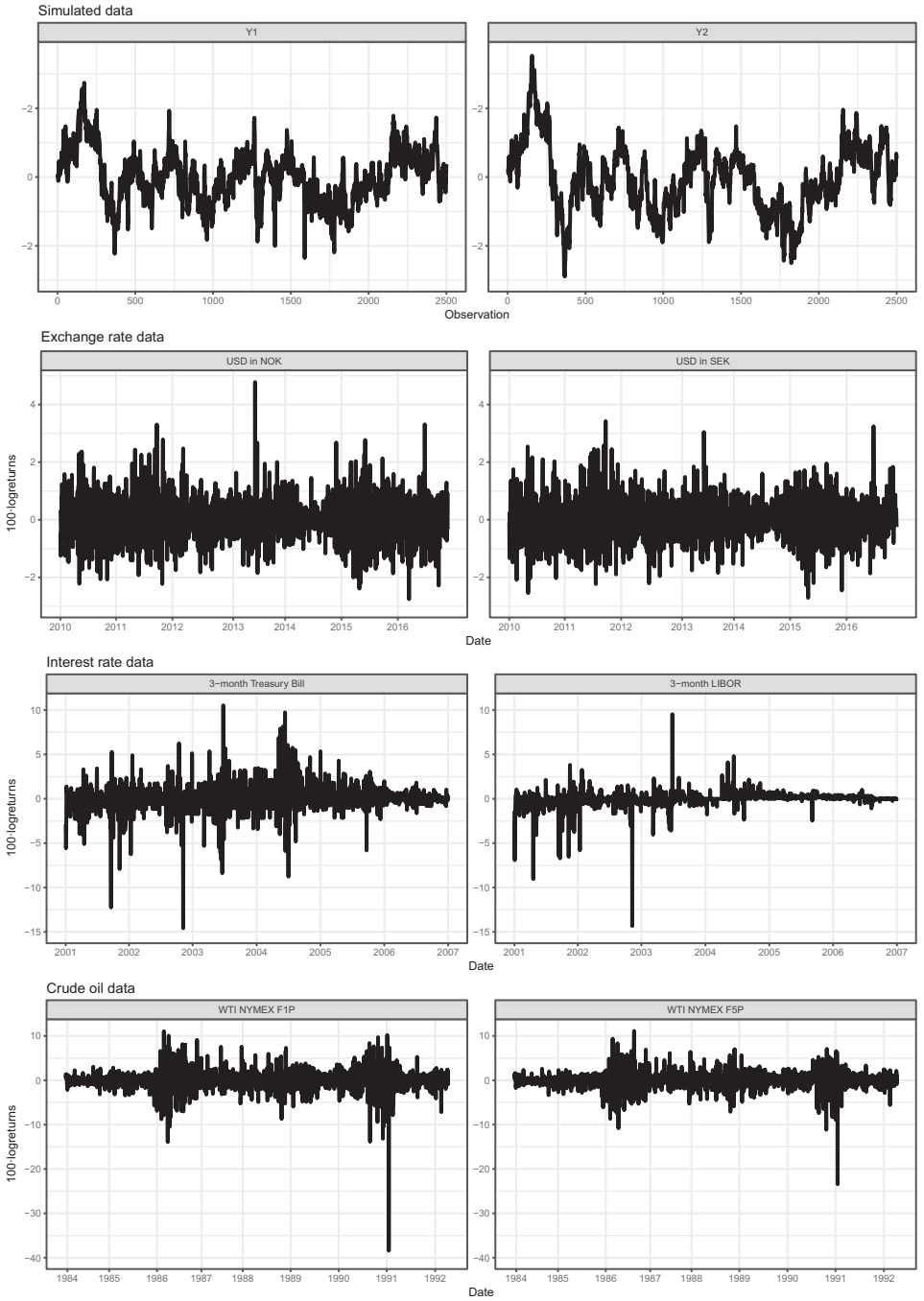

Figure 1. Plots of a simulated data set and three real data sets, all having two components. For the real data sets, log returns of the raw data are used, scaled by 100.

approach produces more efficient samples than Gibbs sampling, with stable performance across parameters. The tailor-made Gibbs sampler implementations are very fast, but have the same issue with unstable performance across parameters, in addition to greater coding efforts and less flexibility.

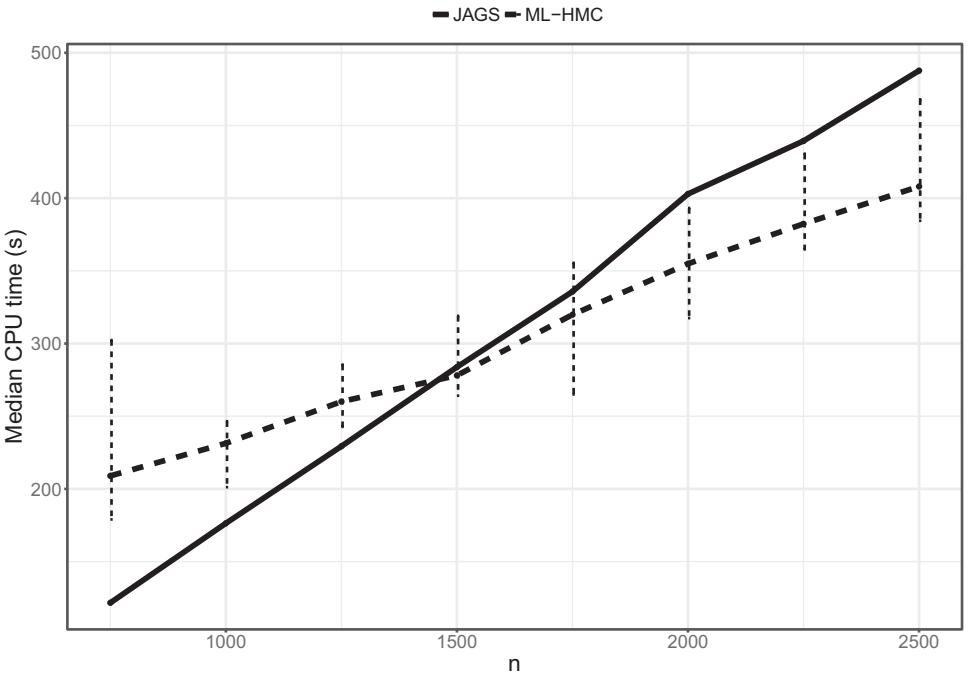

Figure 2. ML-HMC and JAGS median run time for 8 chains of 1000 (500 warm-up) iterations each, obtained from simulated data sets with different observation sizes, *n*. Vertical lines span from minimum to maximum run time.

3.3. Restricted model

Following Lanne, Lütkepohl, and Maciejowska (2010), (4) is fit to a quarterly US macro data set, consisting of inflation, unemployment and an interest rate (3-Month Treasury Bill). A completely unrestricted model would result in too many parameters to be determined by the relatively few observations available, and therefore the mean-structure of the model is restricted to be invariant of the state. Still, the model has 53 parameters. As mentioned in Sec. 2.3, the restriction of (4) complicates the implementation of the tailor-made Gibbs samplers relative to the previous model, while only minor code changes are needed for the software packages. Table 3 shows the resulting parameter estimates and their ESS, using quarterly data from Q1 1960 to Q4 2017 ($n = 232$). JAGS was converging slowly for this model fit, requiring the burn-in to be increased from 500 to 20000 iterations, but the resulting ESS was still very low. Stan produces almost perfect samples, except the autoregressive parameters. The computational time of the tailor-made Gibbs samplers is still superior, and the FB version is performing quite stable across the parameters. Based on these observations, we see that most of the observations made earlier also carries over to this situation with fewer observations and substantially more parameters.

4. The joint dynamics of natural gas and oil prices

This section applies the sampling methods from Sec. 2.2 to estimate parameters for a model of the joint dynamics of UK natural gas and Brent oil prices, extending the modeling approach in Asche, Oglend, and Osmundsen (2017). The model used by Asche, Oglend,

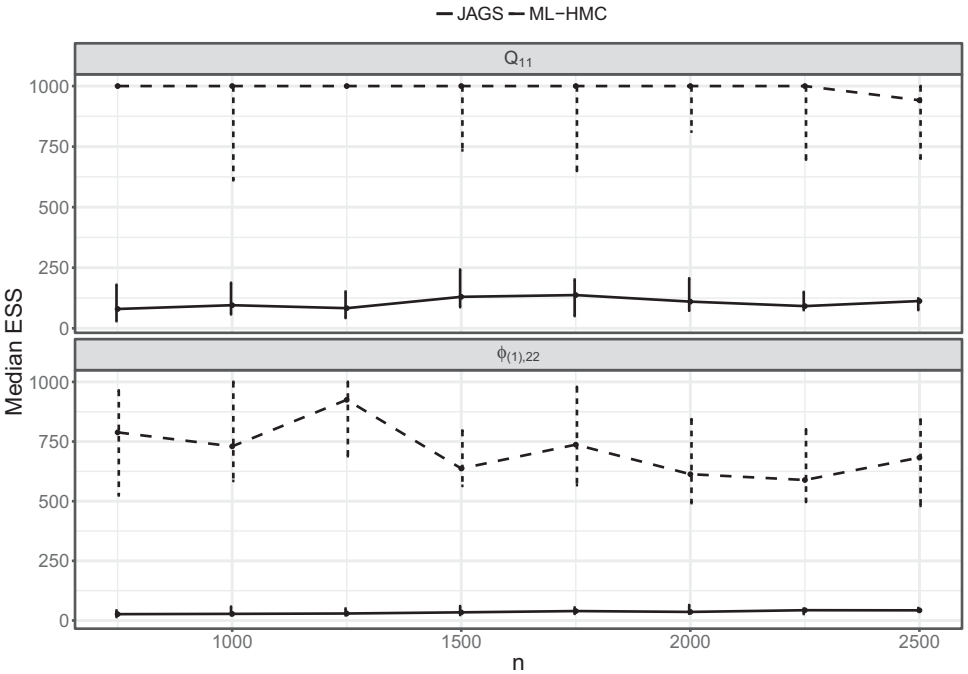

Figure 3. Median effective sample size for the estimates of parameters $\boldsymbol{Q}_{11}$ and $\boldsymbol{\phi}_{(1),22}$ for 8 chains of 1000 (500 warm-up) iterations each, obtained from simulated data sets of different observation sizes, *n*, using both ML-HMC and JAGS. Vertical lines span from minimum to maximum ESS.

Table 2. The parameter estimates' effective sample size per second, for three real data sets.

	Exchange rate, n = 1725				Interest rate, n = 1469				Oil, n = 1999			
	ML-HMC	JAGS	FB	DG	ML-HMC	JAGS	FB	DG	ML-HMC	JAGS	FB	DG
t(s)	605	654	4.7	2.9	187	537	4.2	2.7	1563	772	5	3.1
Q_{11}	8.2	0.2	29	27	34	2.2	221	426	5.1	1.3	218	204
Q_{22}	9.8	0.2	29	27	39	2.7	349	415	5.1	0.9	281	112
$\mu_{(1),1}$	12	1.5	858	319	35	8.1	1178	1726	5.1	0.7	1023	1278
$\mu_{(1),2}$	11	1.4	811	229	43	6.8	1183	1488	5.1	0.7	1165	1318
$\mu_{(2),1}$	10	1.4	137	51	39	12	1442	2384	5.1	1.6	1414	1866
$\mu_{(2),2}$	11	1.6	161	81	43	12	1441	2216	5.1	1.6	1410	1929
$\Sigma_{(1),11}$	10	0.3	36	16	34	1.9	270	273	5.1	0.7	196	73
$\Sigma_{(1),12}$	12	0.3	38	22	43	4.3	909	1352	5.1	0.7	215	79
$\Sigma_{(1),22}$	11	0.3	35	20	29	0.9	130	130	5.1	0.5	219	84
$\Sigma_{(2),11}$	10	0.3	82	15	33	2.1	341	282	5.1	0.6	230	83
$\Sigma_{(2),12}$	13	0.6	123	27	43	6.5	730	1199	5.1	0.7	268	90
$\Sigma_{(2),22}$	12	0.5	110	24	37	1.9	284	307	5	0.7	213	87
$\phi_{(1),11}$	7.8	0.3	645	252	43	7.6	1218	1578	2.9	0.05	854	695
$\phi_{(1),12}$	7.8	0.4	648	468	43	6.1	940	1166	2.8	0.04	861	681
$\phi_{(1),21}$	8	0.4	709	266	43	4.7	684	664	2.9	0.04	857	745
$\phi_{(1),22}$	8.2	0.4	513	287	37	3	533	687	2.8	0.04	860	692
$\phi_{(2),11}$	8.1	0.7	1163	1123	36	11	1802	3008	2.8	0.1	1584	2423
$\phi_{(2),12}$	9.9	0.7	1186	1182	40	10	1882	2898	2.9	0.1	1584	2572
$\phi_{(2),21}$	8.8	0.7	1151	1003	37	5.4	1015	949	2.8	0.1	1488	2572
$\phi_{(2),22}$	9	0.7	1034	987	36	11	1715	2880	2.8	0.1	1429	2572

The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as a single chain.

and Osmundsen (2017) is a one-dimensional, two-regime Markov-Switching Vector Error Correction (MS-VECM) model with one autoregressive lag. The model is closely related to the MS-VAR model, with the difference being an added error correction term. The latent

Table 3. The parameter estimates and their effective sample size per second, for the US macro data set.

	ML-HMC			JAGS			FB			DG		
	Mean	SD	ESS/s	Mean	SD	ESS/s	Mean	SD	ESS/s	Mean	SD	ESS/s
t[s]	3910			2021			5			5		
Q_{11}	0.9677	0.0146	2	0.9680	0.0145	1.4	0.9678	0.0144	1034	0.9680	0.0143	865
Q_{22}	0.8955	0.0470	2	0.8908	0.0486	0.5	0.8943	0.0456	1056	0.8943	0.0469	504
μ_1	0.1691	0.1319	2	0.1558	0.1316	0.1	0.1634	0.1342	1502	0.1647	0.1325	1046
μ_2	0.0897	0.0679	2	0.0926	0.0688	0.1	0.0912	0.0691	1398	0.0930	0.0688	903
μ_3	0.1126	0.0764	2	0.1229	0.0732	0.1	0.1119	0.0764	1594	0.1107	0.0757	1436
$\phi_{1,11}$	1.1521	0.0844	1.1	1.1537	0.0858	0.01	1.1533	0.0848	796	1.1540	0.0845	534
$\phi_{1,21}$	−0.0516	0.0268	1.1	−0.0524	0.0247	0.03	−0.0513	0.0270	1145	−0.0515	0.0269	947
$\phi_{1,31}$	0.0288	0.0315	1.3	0.0207	0.0283	0.03	0.0292	0.0312	1247	0.0293	0.0318	1030
$\phi_{1,12}$	−0.2614	0.1778	1.1	−0.2838	0.1615	0	−0.2599	0.1803	1223	−0.2648	0.1785	995
$\phi_{1,22}$	1.4312	0.0766	1	1.4390	0.0732	0	1.4302	0.0775	1214	1.4313	0.0776	871
$\phi_{1,32}$	−0.1249	0.0884	1.1	−0.1508	0.0587	0.01	−0.1251	0.0870	1212	−0.1242	0.0878	1025
$\phi_{1,13}$	0.0133	0.1610	1.1	0.0425	0.1910	0.01	0.0145	0.1613	998	0.0096	0.1609	812
$\phi_{1,23}$	0.1138	0.0630	1.2	0.0939	0.0739	0.01	0.1136	0.0639	1540	0.1134	0.0624	1389
$\phi_{1,33}$	1.3679	0.0739	1.4	1.3852	0.0594	0.01	1.3686	0.0732	1505	1.3701	0.0744	1076
$\phi_{2,11}$	−0.2031	0.1121	1.1	−0.1996	0.1232	0.01	−0.2018	0.1130	1075	−0.2023	0.1125	723
$\phi_{2,21}$	0.0458	0.0338	1.1	0.0452	0.0327	0.02	0.0451	0.0337	1399	0.0452	0.0337	1118
$\phi_{2,31}$	−0.0086	0.0406	1.2	−0.0044	0.0383	0.02	−0.0091	0.0396	1338	−0.0083	0.0402	1171
$\phi_{2,12}$	0.0701	0.2955	1	0.1569	0.2686	0	0.0632	0.2957	1439	0.0704	0.2951	1211
$\phi_{2,22}$	−0.3686	0.1313	1	−0.3834	0.1253	0	−0.3669	0.1317	1461	−0.3679	0.1314	1108
$\phi_{2,32}$	0.1801	0.1542	0.9	0.2181	0.1186	0	0.1808	0.1521	1216	0.1782	0.1539	982
$\phi_{2,13}$	−0.0352	0.2516	1.1	−0.0628	0.2607	0	−0.0379	0.2534	1248	−0.0321	0.2518	1124
$\phi_{2,23}$	−0.1145	0.1078	1	−0.0896	0.1058	0.01	−0.1139	0.1075	1338	−0.1143	0.1058	1369
$\phi_{2,33}$	−0.2255	0.1243	1	−0.2362	0.0948	0.01	−0.2249	0.1228	1470	−0.2287	0.1254	1010
$\phi_{3,11}$	0.1266	0.1108	1.1	0.1000	0.1097	0.01	0.1222	0.1113	846	0.1218	0.1132	611
$\phi_{3,21}$	−0.0256	0.0329	1.1	−0.0226	0.0369	0.01	−0.0249	0.0336	1249	−0.0252	0.0336	942
$\phi_{3,31}$	−0.0063	0.0436	1	−0.0009	0.0366	0.02	−0.0068	0.0423	806	−0.0079	0.0425	453
$\phi_{3,12}$	0.1810	0.2962	1	0.0308	0.2093	0	0.1829	0.2907	1318	0.1810	0.2957	1046
$\phi_{3,22}$	−0.1265	0.1277	1	−0.1254	0.1152	0	−0.1253	0.1270	1537	−0.1272	0.1271	1192
$\phi_{3,32}$	−0.0687	0.1503	0.8	−0.0753	0.0944	0	−0.0700	0.1459	1396	−0.0684	0.1486	1155
$\phi_{3,13}$	0.1814	0.2499	1	0.1540	0.1903	0.01	0.1809	0.2462	1224	0.1830	0.2445	927
$\phi_{3,23}$	0.0170	0.1042	1	0.0181	0.0770	0.01	0.0164	0.1035	1516	0.0162	0.1038	1427
$\phi_{3,33}$	−0.3157	0.1227	1	−0.3418	0.1033	0.01	−0.3173	0.1211	1288	−0.3149	0.1217	943
$\phi_{4,11}$	−0.1274	0.0748	1.2	−0.1034	0.0701	0.02	−0.1244	0.0744	957	−0.1245	0.0754	621
$\phi_{4,21}$	0.0398	0.0245	1.3	0.0379	0.0233	0.03	0.0396	0.0244	1305	0.0395	0.0242	1068
$\phi_{4,31}$	−0.0121	0.0316	1.1	−0.0135	0.0244	0.03	−0.0111	0.0306	794	−0.0109	0.0310	404
$\phi_{4,12}$	−0.0185	0.1683	1.1	0.0695	0.1355	0.01	−0.0143	0.1671	1024	−0.0151	0.1681	791
$\phi_{4,22}$	0.0327	0.0693	1.1	0.0381	0.0589	0	0.0304	0.0697	1356	0.0319	0.0692	941
$\phi_{4,32}$	−0.0014	0.0831	1.1	−0.0087	0.0400	0.01	−0.0005	0.0804	1143	−0.0002	0.0812	798
$\phi_{4,13}$	−0.0826	0.1449	1.1	−0.0613	0.1350	0.01	−0.0823	0.1413	1267	−0.0844	0.1400	716
$\phi_{4,23}$	−0.0049	0.0593	1.3	−0.0107	0.0519	0.01	−0.0045	0.0590	1576	−0.0032	0.0591	1432
$\phi_{4,33}$	0.1611	0.0686	1.3	0.1801	0.0591	0.01	0.1610	0.0689	1338	0.1609	0.0681	1157
$\Sigma_{(1),11}$	0.1536	0.0202	2	0.1567	0.0198	0.4	0.1565	0.0209	806	0.1564	0.0211	450
$\Sigma_{(1),12}$	−0.0219	0.0066	2	−0.0220	0.0067	0.3	−0.0225	0.0067	757	−0.0224	0.0066	513
$\Sigma_{(1),22}$	0.0327	0.0038	2	0.0328	0.0038	0.4	0.0328	0.0039	1133	0.0328	0.0039	629
$\Sigma_{(1),13}$	0.0183	0.0068	2	0.0188	0.0069	0.9	0.0187	0.0070	1042	0.0185	0.0069	955
$\Sigma_{(1),23}$	−0.0033	0.0030	2	−0.0033	0.0030	1	−0.0033	0.0030	1247	−0.0033	0.0030	1018
$\Sigma_{(1),33}$	0.0415	0.0051	2	0.0411	0.0050	0.4	0.0411	0.0051	881	0.0411	0.0050	536
$\Sigma_{(2),11}$	3.7647	0.8162	1.7	4.0257	0.9612	0.1	3.9467	0.9032	1024	3.9492	0.9357	544
$\Sigma_{(2),12}$	−0.3421	0.1380	1.9	−0.3705	0.1595	0.1	−0.3576	0.1455	1025	−0.3572	0.1475	759
$\Sigma_{(2),22}$	0.1806	0.0388	1.9	0.1865	0.0425	0.2	0.1827	0.0409	1080	0.1830	0.0413	790
$\Sigma_{(2),13}$	0.2853	0.1582	2	0.2973	0.1693	0.2	0.3020	0.1681	1135	0.3042	0.1691	954
$\Sigma_{(2),23}$	0.0049	0.0326	2	0.0061	0.0345	0.2	0.0038	0.0339	1337	0.0038	0.0335	1247
$\Sigma_{(2),33}$	0.2729	0.0606	2	0.2644	0.0587	0.1	0.2659	0.0585	1204	0.2655	0.0593	821

The results are calculated from a collection of 8 chains with 1000 (excluding warm − up) iterations each, treated as a single chain.

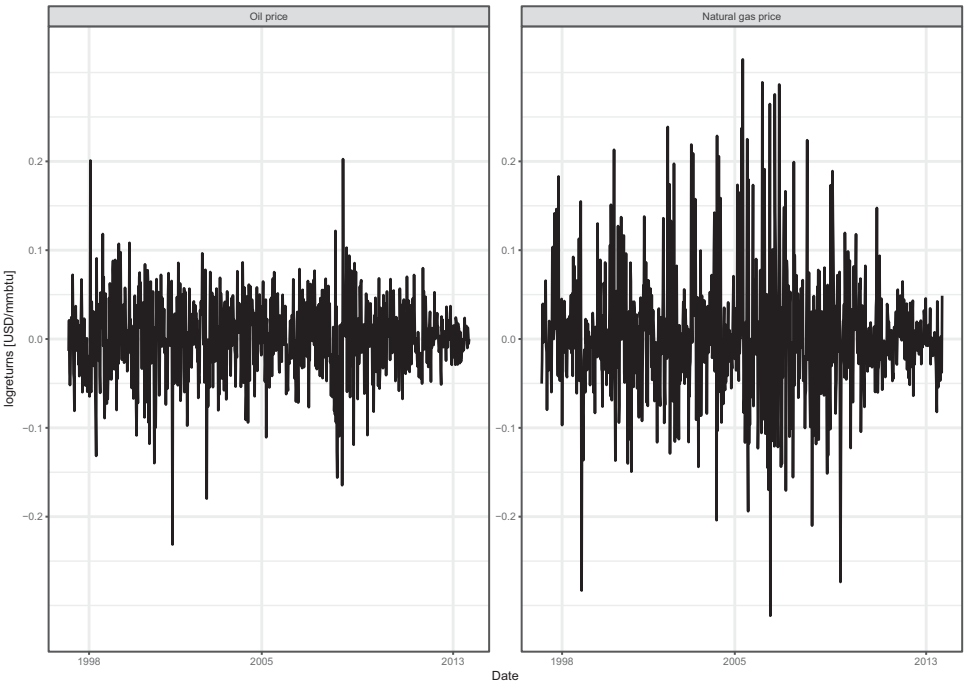

Figure 4. Plot of the log returns of oil and natural gas prices. The data set consists of 904 weekly observations, from March 1997 to April 2014.

state variables indicate the connection between the natural gas and oil prices at each given time, with state 1 indicating decoupled prices and state 2 indicating integrated prices. The use of only one dimension means that the natural gas prices' influence on oil prices has to be modeled separately from the oil prices' influence on natural gas prices. Asche, Oglend, and Osmundsen (2017) study mainly the oil prices' influence on natural gas prices, as the oil price is considered to be largely exogenous to the natural gas market.

The parameters for the joint dynamics of natural gas and oil prices are estimated using a two-dimensional, two-regime Markov-Switching Vector Error Correction (MS-VECM) model with one autoregressive lag:

$$\begin{bmatrix} \Delta \boldsymbol{Y}_{\mathrm{t},1} \\ \Delta \boldsymbol{Y}_{t,2} \end{bmatrix} = \begin{bmatrix} \phi_{(S_t),11} & \phi_{(S_t),12} \\ \phi_{(S_t),21} & \phi_{(S_t),22} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{Y}_{\mathrm{t}-1,1} \\ \Delta \boldsymbol{Y}_{t-1,2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\mu}_{(S_t),1} \\ \boldsymbol{\mu}_{(S_t),2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\alpha}_{(S_t),1} \\ \boldsymbol{\alpha}_{(S_t),2} \end{bmatrix} z_{t-1} + \begin{bmatrix} \boldsymbol{\epsilon}_{t,1} \\ \boldsymbol{\epsilon}_{t,2} \end{bmatrix},$$
$$\begin{bmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{(S_t)}\right), \quad \boldsymbol{Y}_t \in \mathbb{R}^2, \quad S_t \in (1, 2, ..., m), \quad t \in (2, 3, 4, ..., n),$$

where $\boldsymbol{\Delta Y}_t = \boldsymbol{Y}_t - \boldsymbol{Y}_{t-1}$ and $z_{t-1} = \boldsymbol{Y}_{t,1} - \boldsymbol{Y}_{t,2}$. Here $\boldsymbol{Y}$ is a $n \times 2$ matrix, where the first column contains natural gas prices and the second column contains oil prices. As the error correction term is the only deviation from (1), (5) can be treated as an MS-VAR with a varying mean term $\boldsymbol{\mu}^*_{(S_t)} = \boldsymbol{\mu}_{(S_t)} + \boldsymbol{\alpha}_{(S_t)} z_{t-1}$, only requiring a slight adjustment of the implementations used in Sec. 3.2.

The parameters are estimated using the log prices for natural gas and oil, thus modeling the log returns shown in Figure 4. The resulting values are not too far from unit scale, so no further scaling of the data is needed. However, smaller variance estimates are expected for this data set, so an Inverse-Wishart $(\frac{1}{30}\boldsymbol{I}_2, 3)$ prior is used on the covariance matrices. Asche,

Table 4. Parameter estimates, with their respective standard deviation in parenthesis, for the data set with oil and natural gas prices.

Method	Q_{11}	Q_{22}	$\mu_{(1),1}$	$\mu_{(1),2}$	$\mu_{(2),1}$	$\mu_{(2),2}$	$\Sigma_{(1),11}$	$\Sigma_{(1),12}$	$\Sigma_{(1),22}$	$\Sigma_{(2),11}$	$\Sigma_{(2),12}$	$\Sigma_{(2),22}$
ML-HMC	0.9533	0.9357	−0.0303	0.0206	−0.0088	−0.0048	0.0013	0.0001	0.0011	0.0086	0.0002	0.0028
	(0.0156)	(0.0234)	(0.0067)	(0.0058)	(0.0073)	(0.0042)	(0.0001)	(0.0001)	(0.0001)	(0.0008)	(0.0003)	(0.0002)
JAGS	0.9384	0.9351	−0.0224	0.0244	−0.0123	−0.0037	0.0010	0.0001	0.0009	0.0078	0.0001	0.0026
	(0.0197)	(0.0245)	(0.0098)	(0.0071)	(0.0070)	(0.0039)	(0.0001)	(0.0001)	(0.0001)	(0.0007)	(0.0002)	(0.0002)
FB	0.9388	0.9432	−0.0169	0.0045	−0.0222	0.0111	0.0058	0.0002	0.0022	0.0040	0.0002	0.0017
	(0.0335)	(0.0284)	(0.0122)	(0.0131)	(0.0119)	(0.0132)	(0.0036)	(0.0002)	(0.0009)	(0.0035)	(0.0002)	(0.0009)
DG	0.9428	0.9443	−0.0188	0.0086	−0.0203	0.0081	0.0049	0.0002	0.0019	0.0048	0.0002	0.0019
	(0.0265)	(0.0310)	(0.0120)	(0.0142)	(0.0122)	(0.0136)	(0.0037)	(0.0002)	(0.0009)	(0.0036)	(0.0002)	(0.0009)

$\phi_{(1),11}$	$\phi_{(1),12}$	$\phi_{(1),21}$	$\phi_{(1),22}$	$\phi_{(2),11}$	$\phi_{(2),12}$	$\phi_{(2),21}$	$\phi_{(2),22}$	$\alpha_{(1),1}$	$\alpha_{(1),2}$	$\alpha_{(2),1}$	$\alpha_{(2),2}$
0.2221	−0.0458	0.0566	0.2001	0.0991	−0.0144	0.0099	0.2022	−0.0473	0.0260	−0.0398	−0.0042
(0.0459)	(0.0516)	(0.0407)	(0.0488)	(0.0533)	(0.0940)	(0.0310)	(0.0545)	(0.0105)	(0.0089)	(0.0129)	(0.0073)
0.2116	−0.0381	0.0692	0.1954	0.1151	−0.0268	0.0071	0.2031	−0.0350	0.0319	−0.0410	−0.0036
(0.0536)	(0.0521)	(0.0484)	(0.0504)	(0.0505)	(0.0858)	(0.0292)	(0.0501)	(0.0154)	(0.0111)	(0.0118)	(0.0066)
0.1452	−0.0245	0.0282	0.2001	0.1766	−0.0358	0.0386	0.2012	−0.0430	0.0068	−0.0445	0.0147
(0.0747)	(0.0672)	(0.0384)	(0.0437)	(0.0740)	(0.0580)	(0.0392)	(0.0406)	(0.0112)	(0.0162)	(0.0102)	(0.0164)
0.1615	−0.0289	0.0340	0.2006	0.1631	−0.0333	0.0342	0.2012	−0.0426	0.0120	−0.0438	0.0109
(0.0757)	(0.0652)	(0.0399)	(0.0428)	(0.0753)	(0.0629)	(0.0386)	(0.0426)	(0.0112)	(0.0180)	(0.0112)	(0.0171)

Normal distributed error terms. The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as a single chain.

Table 5. The parameter's effective sample size per second, for the data set with oil and natural gas prices.

Method	t(s)	Q_{11}	Q_{22}	$\mu_{(1),1}$	$\mu_{(1),2}$	$\mu_{(2),1}$	$\mu_{(2),2}$	$\Sigma_{(1),11}$	$\Sigma_{(1),12}$	$\Sigma_{(1),22}$	$\Sigma_{(2),11}$	$\Sigma_{(2),12}$
ML-HMC	627	11	11	12	12	13	13	13	13	13	13	13
JAGS	335	0.7	0.4	0.1	0.5	1.1	4.6	0.3	4.2	1.1	0.4	9.3
FB	3	107	79	4.8	3.7	5.1	3.3	2.7	1831	2.3	2.5	673
DG	2	54	52	7.7	4.1	7.3	4.9	2.9	560	2.7	3	878

$\Sigma_{(2),22}$	$\phi_{(1),11}$	$\phi_{(1),12}$	$\phi_{(1),21}$	$\phi_{(1),22}$	$\phi_{(2),11}$	$\phi_{(2),12}$	$\phi_{(2),21}$	$\phi_{(2),22}$	$\alpha_{(1),1}$	$\alpha_{(1),2}$	$\alpha_{(2),1}$	$\alpha_{(2),2}$
13	13	13	13	13	13	13	13	13	12	12	13	13
0.4	1.6	3.3	1.7	4	4.9	12	7.3	15	0.1	0.4	2.1	5.2
2.3	6.2	221	25	605	6.2	207	27	449	147	4.3	144	4.8
3.1	8.1	362	25	491	9.5	310	22	483	64	5.5	72	4.6

Normal distributed error terms. The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as one single chain.

Oglend, and Osmundsen (2017) find that state 2 has eight times higher variance than state 1, meaning that it is reasonable using the identifiability constraint from Sec. 3. The resulting parameter estimates for (5) are given in Table 4. The ESS/s values are shown in Table 5.

To illustrate one of the main benefits of the software packages compared to tailor-made Gibbs samplers, (5) is also estimated using student-t distributed errors, with $\nu_{(s)}$ degrees of freedom. For JAGS and ML-HMC, this model adjustment simply involves re-specifying the distribution of the error terms, which is done in under a minute. For the Gibbs samplers, such a model adjustment would require almost a complete re-writing of the code. The parameter estimates for (5) with student-t distributed errors are given in Table 6. The estimated degrees of freedom are quite low (especially for the integrated regime), implying that student-t distributed errors give a better model fit. The ESS/s values are shown in Table 7. Increased computational time results in lower values than for the normal distributed errors.

Table 8 shows how the parameter estimates for (5) compare to the original estimates of Asche, Oglend, and Osmundsen (2017). The inclusion of the natural gas prices' influence on oil prices gives similar results for the regime of decoupled prices (state 1), while the parameter estimates for the integrated prices (state 2) are quite different, with the assumption of an exogenous oil price only holding in state 1 (as measured by the significance of the adjustment coefficient $\boldsymbol{\alpha}_{(S),2}$ on oil). The difference in the estimate of Q_{22} shows that (5) identifies longer-lasting integrated regimes, affecting the estimation of the other parameters.

5. Discussion

Two different ways of implementing MCMC estimation of the parameters in Markov switching/hidden Markov models with emphasis on MS-VAR models have been presented. Efficiency varies for different data sets and different observation sizes for both methods, and suitable priors and identifiability constraints are crucial to get good results. Overall, the marginalization approach gives robust results with reasonable efficiency, and was quickly implemented using the statistical software package Stan. The Gibbs sampling method was quickly implemented using the statistical software package JAGS, but is quite inefficient. The conjugacy-exploiting tailor-made Gibbs implementations are very efficient, but require substantially greater coding efforts, and are not easily adaptable to modeling changes, parameter restrictions or explicit ordering of regimes according to some form of interpretation. The marginalization approach has approximately equal efficiency for all parameters, unlike the Gibbs implementations.

Table 6. Parameter estimates, with their respective standard deviation in parenthesis, for the DATA set with oil and natural gas prices. Student-*t* distributed error terms.

	Q_{11}	Q_{22}	$\mu_{(1),1}$	$\mu_{(1),2}$	$\mu_{(2),1}$	$\mu_{(2),2}$	$\Sigma_{(1),11}$	$\Sigma_{(1),12}$	$\Sigma_{(1),22}$	$\Sigma_{(2),11}$	$\Sigma_{(2),12}$	$\Sigma_{(2),22}$	$\phi_{(1),11}$
ML-HMC	0.9591	0.9731	−0.0174	0.0341	−0.0164	−0.0012	0.0008	0.0001	0.0008	0.0045	0.0001	0.0016	0.1897
	(0.0143)	(0.0113)	(0.0090)	(0.0086)	(0.0055)	(0.0035)	(0.0001)	(0.0001)	(0.0001)	(0.0006)	(0.0001)	(0.0002)	(0.0610)
JAGS	0.9572	0.9711	−0.0167	0.0343	−0.0163	−0.0015	0.0008	0.0001	0.0008	0.0046	0.0001	0.0017	0.1926
	(0.0146)	(0.0122)	(0.0083)	(0.0085)	(0.0055)	(0.0035)	(0.0001)	(0.0001)	(0.0001)	(0.0006)	(0.0001)	(0.0002)	(0.0616)

$\phi_{(1),12}$	$\phi_{(1),21}$	$\phi_{(1),22}$	$\phi_{(2),11}$	$\phi_{(2),12}$	$\phi_{(2),21}$	$\phi_{(2),22}$	$\alpha_{(1),1}$	$\alpha_{(1),2}$	$\alpha_{(2),1}$	$\alpha_{(2),2}$	$\nu_{(1)}$	$\nu_{(2)}$
−0.0245	0.0945	0.2022	0.1435	−0.0296	0.0027	0.2197	−0.0275	0.0479	−0.0353	−0.0036	19.7326	6.7546
(0.0535)	(0.0648)	(0.0555)	(0.0403)	(0.0704)	(0.0240)	(0.0424)	(0.0142)	(0.0137)	(0.0090)	(0.0056)	(11.3163)	(1.5143)
−0.0238	0.0877	0.1983	0.1414	−0.0279	0.0037	0.2188	−0.0262	0.0481	−0.0356	−0.0041	19.7323	6.7905
(0.0516)	(0.0656)	(0.0555)	(0.0400)	(0.0708)	(0.0244)	(0.0434)	(0.0130)	(0.0133)	(0.0090)	(0.0055)	(11.9735)	(1.5004)

The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as one single chain.

Table 7. The parameter's effective sample size per second, for the data set with oil and natural gas prices.

Method	t(s)	Q_{11}	Q_{22}	$\mu_{(1),1}$	$\mu_{(1),2}$	$\mu_{(2),1}$	$\mu_{(2),2}$	$\Sigma_{(1),11}$	$\Sigma_{(1),12}$	$\Sigma_{(1),22}$	$\Sigma_{(2),11}$	$\Sigma_{(2),12}$	$\Sigma_{(2),22}$
ML-HMC	1654	4.8	3.4	2.9	3.2	4.2	4.8	3.1	4.8	4.1	2.1	4.8	3
JAGS	875	0.4	0.1	0.1	0.1	0.9	0.7	0.2	0.5	0.6	0.1	5.4	0.3

$\phi_{(1),11}$	$\phi_{(1),12}$	$\phi_{(1),21}$	$\phi_{(1),22}$	$\phi_{(2),11}$	$\phi_{(2),12}$	$\phi_{(2),21}$	$\phi_{(2),22}$	$\alpha_{(1),1}$	$\alpha_{(1),2}$	$\alpha_{(2),1}$	$\alpha_{(2),2}$	$\nu_{(1)}$	$\nu_{(2)}$
4.8	4.5	4.8	4.8	4.8	4.8	4.8	4.8	2.9	3.3	4.7	4.8	4.2	2.7
0.5	0.9	0.4	0.9	1.6	3.6	2.3	4.6	0.1	0.1	1.4	0.8	0.3	0.7

Student-t distributed error terms. The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as one single chain.

Table 8. The parameter estimates for the one-dimensional MS-VECM of Asche, Oglend, and Osmundsen (2017) compared to the estimates for the two-dimensional model, using both normal and student-t distributed errors.

	Asche, Oglend, and Osmundsen (2017)	Normal	Student t
Q_{11}	0.94	0.953	0.959
Q_{22}	0.732	0.936	0.973
$\mu_{(1),1}$	−0.03	−0.030	−0.017
$\mu_{(1),2}$	–	0.021	0.034
$\mu_{(2),1}$	0.042	−0.009	−0.016
$\mu_{(2),2}$	–	−0.005	−0.001
$\Sigma_{(1),11}$	0.001	0.001	0.001
$\Sigma_{(1),12}$	–	0.000	0.000
$\Sigma_{(1),22}$	–	0.001	0.001
$\Sigma_{(2),11}$	0.008	0.009	0.005
$\Sigma_{(2),12}$	–	0.000	0.000
$\Sigma_{(2),22}$	–	0.003	0.002
$\phi_{(1),11}$	0.237	0.222	0.190
$\phi_{(1),12}$	−0.053	−0.046	−0.024
$\phi_{(1),21}$	–	0.057	0.095
$\phi_{(1),22}$	–	0.200	0.202
$\phi_{(2),11}$	0.093	0.099	0.144
$\phi_{(2),12}$	0.026	−0.014	−0.030
$\phi_{(2),21}$	–	0.010	0.003
$\phi_{(2),22}$	–	0.202	0.220
$\alpha_{(1),1}$	−0.047	−0.047	−0.028
$\alpha_{(1),2}$	–	0.026	0.048
$\alpha_{(2),1}$	−0.022	−0.040	−0.035
$\alpha_{(2),2}$	–	−0.004	−0.004
$\nu_{(1)}$	–	–	19.733
$\nu_{(2)}$	–	–	6.755

The empirical part of this article focuses on a two-dimensional, two-state model, but the methodology extends easily to more complex models. Larger models will favor the general software packages, as increased model dimension complicates the implementation of tailor-made solutions. The marginalized approach combined with the NUTS sampler is likely to produce effective samples also for larger models, while increased dimensionality may cause slower mixing for Gibbs samplers.

The latter observation is particularly relevant for larger models such as Markov switching variants of structural form VARs, where the posterior distribution of the parameters is highly non-Gaussian (Waggoner, Wu, and Zha 2016). Application of marginalized approach and Stan to such situations holds scope for future research.

Acknowledgments

The authors would like to express gratitude to an anonymous referee for many useful comments that considerably improved the article.

Table A1. Parameter estimates, with their respective standard deviation in parenthesis, for simulated data sets with three different observation sizes.

	n = 750				n = 1500				n = 2500			
	ML-HMC	JAGS	FB	DG	ML-HMC	JAGS	FB	DG	ML-HMC	JAGS	FB	DG
Q_{11}	0.9690	0.9696	0.9660	0.9662	0.9704	0.9706	0.9699	0.9695	0.9754	0.9751	0.9754	0.9749
	(0.0102)	(0.0109)	(0.0144)	(0.0147)	(0.0066)	(0.0067)	(0.0076)	(0.0085)	(0.0047)	(0.0048)	(0.0052)	(0.0060)
Q_{22}	0.8773	0.8681	0.8723	0.8787	0.8998	0.8995	0.8994	0.9002	0.9116	0.9111	0.9117	0.9103
	(0.0388)	(0.0797)	(0.0489)	(0.0441)	(0.0217)	(0.0219)	(0.0233)	(0.0241)	(0.0157)	(0.0159)	(0.0171)	(0.0186)
$\mu_{(1),1}$	0.0060	0.0061	0.0057	0.0052	0.0041	0.0042	0.0041	0.0042	0.0048	0.0047	0.0048	0.0047
	(0.0073)	(0.0073)	(0.0072)	(0.0073)	(0.0048)	(0.0048)	(0.0048)	(0.0048)	(0.0037)	(0.0037)	(0.0037)	(0.0036)
$\mu_{(1),2}$	0.0010	0.0012	0.0007	0.0005	0.0012	0.0012	0.0011	0.0012	0.0007	0.0005	0.0006	0.0006
	(0.0073)	(0.0072)	(0.0072)	(0.0074)	(0.0046)	(0.0047)	(0.0046)	(0.0046)	(0.0036)	(0.0035)	(0.0035)	(0.0036)
$\mu_{(2),1}$	0.0306	0.0298	0.0308	0.0306	0.0090	0.0091	0.0089	0.0086	0.0072	0.0080	0.0071	0.0078
	(0.0246)	(0.0360)	(0.0228)	(0.0230)	(0.0160)	(0.0163)	(0.0154)	(0.0158)	(0.0138)	(0.0136)	(0.0132)	(0.0133)
$\mu_{(2),2}$	0.0172	0.0155	0.0172	0.0171	−0.0050	−0.0052	−0.0045	−0.0050	−0.0106	−0.0104	−0.0105	−0.0107
	(0.0257)	(0.0364)	(0.0242)	(0.0246)	(0.0150)	(0.0153)	(0.0148)	(0.0146)	(0.0123)	(0.0124)	(0.0120)	(0.0121)
$\Sigma_{(1),11}$	0.0248	0.0251	0.0246	0.0245	0.0230	0.0231	0.0229	0.0229	0.0228	0.0228	0.0228	0.0228
	(0.0017)	(0.0023)	(0.0017)	(0.0017)	(0.0011)	(0.0011)	(0.0011)	(0.0012)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
$\Sigma_{(1),12}$	0.0061	0.0062	0.0061	0.0060	0.0049	0.0049	0.0049	0.0049	0.0044	0.0044	0.0044	0.0044
	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
$\Sigma_{(1),22}$	0.0250	0.0251	0.0248	0.0246	0.0217	0.0217	0.0216	0.0216	0.0214	0.0213	0.0214	0.0213
	(0.0016)	(0.0019)	(0.0017)	(0.0017)	(0.0011)	(0.0011)	(0.0010)	(0.0011)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
$\Sigma_{(2),11}$	0.0720	0.0905	0.0701	0.0693	0.0764	0.0766	0.0756	0.0756	0.0797	0.0797	0.0793	0.0794
	(0.0098)	(0.3278)	(0.0097)	(0.0100)	(0.0065)	(0.0066)	(0.0066)	(0.0068)	(0.0053)	(0.0053)	(0.0054)	(0.0055)
$\Sigma_{(2),12}$	0.0016	0.0046	0.0017	0.0021	0.0075	0.0076	0.0075	0.0077	0.0065	0.0064	0.0064	0.0065
	(0.0065)	(0.1989)	(0.0063)	(0.0061)	(0.0040)	(0.0041)	(0.0040)	(0.0040)	(0.0032)	(0.0032)	(0.0032)	(0.0032)
$\Sigma_{(2),22}$	0.0695	0.0942	0.0666	0.0659	0.0629	0.0630	0.0620	0.0619	0.0628	0.0629	0.0623	0.0625
	(0.0095)	(0.8798)	(0.0093)	(0.0096)	(0.0055)	(0.0054)	(0.0055)	(0.0056)	(0.0042)	(0.0043)	(0.0042)	(0.0042)
$\phi_{(1),11}$	0.5870	0.5892	0.5847	0.5819	0.5780	0.5772	0.5766	0.5759	0.5773	0.5773	0.5775	0.5767
	(0.0309)	(0.0426)	(0.0317)	(0.0328)	(0.0226)	(0.0234)	(0.0230)	(0.0236)	(0.0171)	(0.0174)	(0.0170)	(0.0175)
$\phi_{(1),12}$	0.3057	0.3040	0.3076	0.3095	0.3118	0.3124	0.3129	0.3134	0.3151	0.3151	0.3151	0.3156
	(0.0231)	(0.0318)	(0.0237)	(0.0245)	(0.0171)	(0.0177)	(0.0172)	(0.0177)	(0.0131)	(0.0132)	(0.0131)	(0.0134)
$\phi_{(1),21}$	−0.0675	−0.0684	−0.0700	−0.0718	−0.0267	−0.0270	−0.0272	−0.0266	−0.0073	−0.0079	−0.0075	−0.0074
	(0.0302)	(0.0329)	(0.0310)	(0.0323)	(0.0210)	(0.0204)	(0.0213)	(0.0211)	(0.0153)	(0.0157)	(0.0154)	(0.0153)
$\phi_{(1),22}$	1.0418	1.0422	1.0436	1.0450	1.0077	1.0081	1.0082	1.0076	0.9885	0.9891	0.9887	0.9884
	(0.0230)	(0.0251)	(0.0236)	(0.0247)	(0.0163)	(0.0158)	(0.0164)	(0.0163)	(0.0120)	(0.0123)	(0.0120)	(0.0119)
$\phi_{(2),11}$	0.9269	0.9180	0.9275	0.9262	0.9563	0.9580	0.9565	0.9555	0.9486	0.9476	0.9490	0.9489
	(0.0453)	(0.1456)	(0.0460)	(0.0460)	(0.0275)	(0.0277)	(0.0272)	(0.0277)	(0.0208)	(0.0205)	(0.0209)	(0.0210)
$\phi_{(2),12}$	0.0396	0.0397	0.0397	0.0408	0.0130	0.0120	0.0134	0.0143	0.0247	0.0253	0.0243	0.0245
	(0.0404)	(0.1073)	(0.0407)	(0.0404)	(0.0257)	(0.0257)	(0.0253)	(0.0260)	(0.0171)	(0.0169)	(0.0171)	(0.0171)
$\phi_{(2),21}$	0.0843	0.0894	0.0854	0.0837	0.0667	0.0672	0.0661	0.0652	0.0558	0.0552	0.0553	0.0551
	(0.0460)	(0.1171)	(0.0445)	(0.0448)	(0.0236)	(0.0236)	(0.0235)	(0.0239)	(0.0176)	(0.0179)	(0.0180)	(0.0180)
$\phi_{(2),22}$	0.9027	0.8894	0.9033	0.9038	0.9261	0.9253	0.9269	0.9280	0.9524	0.9523	0.9526	0.9533
	(0.0428)	(0.1532)	(0.0417)	(0.0421)	(0.0233)	(0.0233)	(0.0237)	(0.0233)	(0.0146)	(0.0150)	(0.0147)	(0.0149)

The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as a single chain.

Table A2. Parameter estimates, with their respective standard deviation in parenthesis, for three real data sets.

	Exchange rate, n = 1725				Interest rate, n = 1469				Oil, n = 1999			
	ML-HMC	JAGS	FB	DG	ML-HMC	JAGS	FB	DG	ML-HMC	JAGS	FB	DG
Q_{11}	0.9540	0.9516	0.9392	0.9378	0.9099	0.9101	0.9080	0.9088	0.9481	0.9481	0.9476	0.9480
	(0.0156)	(0.0174)	(0.0450)	(0.0385)	(0.0121)	(0.0125)	(0.0130)	(0.0130)	(0.0079)	(0.0080)	(0.0085)	(0.0086)
Q_{22}	0.8641	0.8597	0.8465	0.8617	0.7821	0.7838	0.7818	0.7819	0.8736	0.8743	0.8741	0.8751
	(0.0418)	(0.0431)	(0.0744)	(0.0594)	(0.0275)	(0.0281)	(0.0295)	(0.0305)	(0.0201)	(0.0207)	(0.0208)	(0.0221)
$\mu_{(1),1}$	−0.0167	−0.0169	−0.0159	−0.0153	0.0335	0.0339	0.0342	0.0344	0.0236	0.0240	0.0260	0.0251
	(0.0187)	(0.0194)	(0.0199)	(0.0207)	(0.0301)	(0.0307)	(0.0309)	(0.0311)	(0.0329)	(0.0327)	(0.0337)	(0.0333)
$\mu_{(1),2}$	−0.0143	−0.0149	−0.0136	−0.0119	0.0495	0.0496	0.0497	0.0498	0.0088	0.0095	0.0108	0.0101
	(0.0187)	(0.0195)	(0.0199)	(0.0211)	(0.0082)	(0.0084)	(0.0083)	(0.0082)	(0.0283)	(0.0280)	(0.0290)	(0.0287)
$\mu_{(2),1}$	0.1298	0.1273	0.1383	0.1252	−0.0459	−0.0478	−0.0711	−0.0736	−0.0552	−0.0536	−0.1092	−0.1071
	(0.0574)	(0.0574)	(0.0658)	(0.0693)	(0.1111)	(0.1141)	(0.1352)	(0.1352)	(0.1156)	(0.1195)	(0.1665)	(0.1684)
$\mu_{(2),2}$	0.0895	0.0881	0.0979	0.0859	−0.0841	−0.0839	−0.0938	−0.0949	−0.0301	−0.0279	−0.0710	−0.0709
	(0.0524)	(0.0513)	(0.0579)	(0.0604)	(0.0626)	(0.0618)	(0.0657)	(0.0656)	(0.0943)	(0.0973)	(0.1322)	(0.1340)
$\Sigma_{(1),11}$	0.3834	0.3826	0.3736	0.3617	0.8041	0.8030	0.7956	0.8001	1.4043	1.4011	1.3915	1.3919
	(0.0226)	(0.0236)	(0.0359)	(0.0445)	(0.0480)	(0.0492)	(0.0504)	(0.0516)	(0.0761)	(0.0769)	(0.0809)	(0.0913)
$\Sigma_{(1),12}$	0.3150	0.3148	0.3065	0.2968	0.0022	0.0022	0.0016	0.0022	1.1275	1.1256	1.1177	1.1175
	(0.0209)	(0.0216)	(0.0315)	(0.0384)	(0.0076)	(0.0077)	(0.0077)	(0.0077)	(0.0625)	(0.0631)	(0.0660)	(0.0740)
$\Sigma_{(1),22}$	0.3826	0.3815	0.3719	0.3612	0.0557	0.0554	0.0544	0.0548	1.0517	1.0500	1.0426	1.0427
	(0.0241)	(0.0251)	(0.0375)	(0.0445)	(0.0046)	(0.0049)	(0.0051)	(0.0054)	(0.0565)	(0.0569)	(0.0594)	(0.0663)
$\Sigma_{(2),11}$	1.1914	1.1898	1.1690	1.1238	7.8381	7.8178	7.7435	7.7594	16.5159	16.4563	16.3451	16.3829
	(0.1118)	(0.1092)	(0.1343)	(0.1661)	(0.5888)	(0.6101)	(0.6185)	(0.6402)	(1.1421)	(1.1298)	(1.1851)	(1.3036)
$\Sigma_{(2),12}$	0.8108	0.8104	0.7993	0.7749	1.0651	1.0594	1.0503	1.0519	12.1196	12.0776	11.9979	12.0242
	(0.0828)	(0.0810)	(0.0904)	(0.1042)	(0.1946)	(0.1967)	(0.1960)	(0.1962)	(0.8628)	(0.8452)	(0.8918)	(0.9711)
$\Sigma_{(2),22}$	0.9430	0.9410	0.9292	0.9015	1.8325	1.8271	1.8039	1.8084	10.4835	10.4461	10.3761	10.3945
	(0.0869)	(0.0835)	(0.0946)	(0.1111)	(0.1425)	(0.1460)	(0.1462)	(0.1523)	(0.7217)	(0.7069)	(0.7456)	(0.8124)
$\phi_{(1),11}$	−0.0159	−0.0140	−0.0184	−0.0145	−0.0757	−0.0759	−0.0755	−0.0750	−0.0564	−0.0648	−0.0585	−0.0564
	(0.0524)	(0.0544)	(0.0573)	(0.0610)	(0.0274)	(0.0276)	(0.0277)	(0.0276)	(0.0722)	(0.0743)	(0.0711)	(0.0703)
$\phi_{(1),12}$	−0.0027	−0.0063	−0.0053	−0.0076	0.2077	0.2072	0.2084	0.2091	0.0282	0.0369	0.0300	0.0268
	(0.0525)	(0.0532)	(0.0565)	(0.0590)	(0.0875)	(0.0887)	(0.0894)	(0.0901)	(0.0843)	(0.0892)	(0.0836)	(0.0828)
$\phi_{(1),21}$	−0.0049	−0.0035	−0.0067	−0.0045	0.0546	0.0547	0.0542	0.0543	−0.1550	−0.1626	−0.1565	−0.1553
	(0.0509)	(0.0533)	(0.0551)	(0.0584)	(0.0079)	(0.0079)	(0.0081)	(0.0080)	(0.0620)	(0.0632)	(0.0611)	(0.0606)
$\phi_{(1),22}$	−0.0245	−0.0273	−0.0276	−0.0261	0.3205	0.3202	0.3213	0.3208	0.1414	0.1493	0.1426	0.1407
	(0.0515)	(0.0527)	(0.0565)	(0.0582)	(0.0265)	(0.0270)	(0.0271)	(0.0272)	(0.0727)	(0.0755)	(0.0721)	(0.0716)
$\phi_{(2),11}$	0.0066	0.0020	0.0071	0.0037	0.0445	0.0436	0.0427	0.0422	−0.1153	−0.1118	−0.1242	−0.1213
	(0.0825)	(0.0874)	(0.0868)	(0.0822)	(0.0536)	(0.0554)	(0.0546)	(0.0551)	(0.1091)	(0.1030)	(0.1106)	(0.1105)
$\phi_{(2),12}$	−0.0066	−0.0005	−0.0021	−0.0016	0.1864	0.1884	0.1872	0.1884	0.1968	0.1898	0.2059	0.2032
	(0.0935)	(0.0976)	(0.0954)	(0.0900)	(0.0849)	(0.0883)	(0.0857)	(0.0882)	(0.1364)	(0.1287)	(0.1389)	(0.1392)
$\phi_{(2),21}$	0.0437	0.0413	0.0430	0.0395	0.3608	0.3600	0.3596	0.3597	−0.0468	−0.0428	−0.0547	−0.0519
	(0.0735)	(0.0772)	(0.0769)	(0.0739)	(0.0275)	(0.0281)	(0.0280)	(0.0278)	(0.0874)	(0.0816)	(0.0886)	(0.0879)
$\phi_{(2),22}$	−0.0596	−0.0559	−0.0543	−0.0542	0.0910	0.0903	0.0907	0.0904	0.1040	0.0972	0.1123	0.1096
	(0.0839)	(0.0875)	(0.0866)	(0.0819)	(0.0422)	(0.0422)	(0.0417)	(0.0428)	(0.1094)	(0.1022)	(0.1108)	(0.1109)

The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as a single chain.

References

Albert, J. H., and S. Chib. 1993. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business & Economic Statistics* 11 (1): 1–15. doi:10.2307/1391303.

Andrieu, C., A. Doucet, and R. Holenstein. 2010. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (3):269–342. doi: 10.1111/j.1467-9868.2009.00736.x.

Ang, A., and G. Bekaert. 2002. International asset allocation with regime shifts. *Review of Financial Studies* 15 (4):1137–87. doi:10.1093/rfs/15.4.1137.

Asche, F., A. Oglend, and P. Osmundsen. 2017. Modeling UK natural gas prices when gas prices periodically decouple from the oil price. *The Energy Journal* 38 (2): 22.

Baker, J. 1975. The dragon system: An overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23 (1):24–9. doi:10.1109/TASSP.1975.1162650.

Bansal, R., G. Tauchen, and H. Zhou. 2004. Regime shifts, risk premiums in the term structure, and the business cycle. *Journal of Business & Economic Statistics* 22 (4):396–409. doi:10.1198/073500104000000398.

Bansal, R., and H. Zhou. 2002. Term structure of interest rates with regime shifts. *The Journal of Finance* 57 (5):1997–2043. doi:10.1111/0022-1082.00487.

Bianchi, F. 2016. Methods for measuring expectations and uncertainty in Markov-switching models. *Journal of Econometrics* 190 (1):79–99. doi:10.1016/j.jeconom.2015.08.004.

Bloom, N. 2009. The impact of uncertainty shocks. *Econometrica* 77 (3):623–85.

Brigida, M. 2014. The switching relationship between natural gas and crude oil prices. *Energy Economics* 43:48–55. doi:10.1016/j.eneco.2014.01.014.

Celeux, G., M. Hurn, and C. P. Robert. 2000. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* 95 (451):957–70. doi:10.1080/01621459.2000.10474285.

Chib, S. 1996. Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics* 75 (1):79–97. doi:10.1016/0304-4076(95)01770-4.

Eddelbuettel, D., and R. François. 2011. RCPP: Seamless R and C++ integration. *Journal of Statistical Software* 40 (8):1–18.

Frühwirth-Schnatter, S. 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* 96 (453): 194–209. doi:10.1198/016214501750333063.

Garcia, R., and P. Perron. 1996. An analysis of the real interest rate under regime shifts. *Review of Economics and Statistics* 78 (1):111–25. doi:10.2307/2109851.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo. *Statistical Science* 7 (4):473–83. doi: 10.1214/ss/1177011137.

Girolami, M., and B. Calderhead. 2011. Riemann Manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (2): 123–214. doi:10.1111/j.1467-9868.2010.00765.x.

Gray, S. F. 1996. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* 42 (1):27–62. doi:10.1016/0304-405X(96)00875-6.

Griewank, A., and A. Walther. 2008. Evaluating derivatives: principles and techniques of algorithmic differentiation. Siam.

Hamilton, J. D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society* 357–384.

Hoffman, M. D., A. Gelman. 2014. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15 (1):1593–623.

Jasra, A., C. Holmes, and D. Stephens. 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* 20 (1):50–67. doi:10.1214/088342305000000016.

Kanamura, T., and K. Ōhashi. 2008. On transition probabilities of regime switching in electricity prices. *Energy Economics* 30 (3):1158–72. doi:10.1016/j.eneco.2007.07.011.

Kim, C.-J. 1994. Dynamic linear models with Markov-switching. *Journal of Econometrics* 60 (1–2):1–22. doi:10.1016/0304-4076(94)90036-1.

Krolzig, H.-M. 1997. *Markov-switching vector autoregressive model: Modelling, Statistical inference, and application to business cycle analysis*. New York: Springer.

Kshirsagar, A. M. 1959. Bartlett decomposition and Wishart distribution. *The Annals of Mathematical Statistics* 30 (1):239–41. doi:10.1214/aoms/1177706379.

Lanne, M., H. Lütkepohl, and K. Maciejowska. 2010. Structural vector autoregressions with Markov switching. *Journal of Economic Dynamics and Control* 34 (2):121–31. doi:10.1016/j.jedc.2009.08.002.

Mount, T. D., Y. Ning, and X. Cai. 2006. Predicting price spikes in electricity markets using a regime-switching model with time-varying parameters. *Energy Economics* 28 (1):62–80. doi:10.1016/j.eneco.2005.09.008.

Neal, R. M. 2011. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* 2:113–62.

Papastamoulis, P. 2016. Label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software, Code Snippets* 69 (1):1–24.

Plummer, M. 2013. *JAGS Version 3.4*. 0 user manual.

R Core Team. 2016. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robert, C., and G. Casella. 2013. *Monte Carlo statistical methods*. New York: Springer Science & Business Media.

Rossi, P. 2015. Bayesm: Bayesian inference for marketing/Micro-Econometrics. R Package Version 3.0-2.

Scharth, M., and R. Kohn. 2016. Particle efficient importance sampling. *Journal of Econometrics* 190 (1):133–47. doi:10.1016/j.jeconom.2015.03.047.

Scott, S. L. 2002. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association* 97 (457):337–51.

Sims, C. A., D. F. Waggoner, and T. Zha. 2008. Methods for inference in large multiple-equation Markov-switching models. *Journal of Econometrics* 146 (2):255–74. doi:10.1016/j.jeconom.2008.08.023.

Sims, C. A., and T. Zha. 2006. Were there regime switches in us monetary policy? *The American Economic Review* 96 (1):54–81. doi:10.1257/000282806776157678.

Stan Development Team. 2016. Stan Modeling Language Users Guide and Reference Manual (2.14.0 ed.).

Stephens, M. 1997. Bayesian methods for mixtures of normal distributions.

Stephens, M. 2000. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62 (4):795–809. doi:10.1111/1467-9868.00265.

Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* 13 (2):260–9. doi:10.1109/TIT.1967.1054010.

Waggoner, D. F., H. Wu, and T. Zha. 2016. Striated metropolis–hastings sampler for high-dimensional models. *Journal of Econometrics* 192 (2):406–20. doi:10.1016/j.jeconom.2016.02.007.

Yamato, J., J. Ohya, and K. Ishii. 1992. Recognizing human action in time-sequential images using hidden markov model. In *1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92*. pp. 379–385. IEEE.