Essays in statistics and econometrics

by

Kjartan Kloster Osmundsen

Thesis submitted in fulfilment of
the requirements for the degree of
PHILOSOPHIAE DOCTOR
(PhD)

Faculty of Science and Technology
Department of Mathematics and Physics
2020

University of Stavanger
NO-4036 Stavanger
NORWAY
www.uis.no

ISBN:978-82-7644-930-3
ISSN:1890-1387
PhD: Thesis UiS No. 523

Preface

This thesis is submitted in partial fulfilment of the requirements for the degree of Philosophiae Doctor (PhD) at the University of Stavanger, Faculty of Science and Technology, Norway. The research has been carried out at the University of Stavanger from August 2016 to January 2020.

I would like to thank my supervisor, Professor Tore Selland Kleppe, for his brilliant guidance during my work on this thesis. In our frequent meetings, you have always gladly shared your knowledge and experience. Through great discussions, detailed feedback and solid advice, you have inspired and enabled me to keep a steady progression throughout the project.

Thanks are also due to my co-supervisors, the professors Atle Øglend and Jan Terje Kvaløy. I am also grateful to Professor Roman Liesenfeld for co-authoring two of the papers in the thesis. I would also like to extend my thanks and appreciation to my fellow PhD students and the members of the mathematical statistics group at the University of Stavanger, for creating a pleasant work environment. Special thanks to Berent Ånund Strømnes Lunde and Birthe Aarekol, for interesting and educational discussions, and for being fun travel companions to various conferences and meetings throughout the world.

Kjartan Kloster Osmundsen
Stavanger, January 2020

Abstract

Helped by cheaper data computation, companies make more use of sophisticated statistical analysis in decision making and economic management. In the dissertation I evaluate and develop statistical methods and apply them for economic applications, e.g. credit risk evaluation and commodity pricing.

Recent developments in modern Monte Carlo methods have made statistical inference possible for complex non-linear and non-Gaussian latent variable models. It is typically computationally expensive to fit data to such dynamic models, due to a large number of unobserved parameters. However, the flexibility of the models has ensured a wide range of applications.

This thesis mainly considers non-linear cases of a latent variable model class called state-space models. The main objective is Bayesian inference for all model parameters, based on the information in the observed data. The presented work considers the existing methods for dealing with latent variables, and propose modifications to some of the most promising methods. The performance of the proposed methods is investigated through applications on economic time series data.

The thesis also includes research of a more applied nature, where an existing economic model for commodity prices is extended with a stochastic trend, to obtain a state-space model. It also contains applied economic research outside the latent variable domain, where different risk measures are compared in the context of credit risk regulation.

List of papers

Paper I

Osmundsen, Kjartan Kloster (2018). Using expected shortfall for credit risk regulation. *Journal of International Financial Markets, Institutions and Money* 57, 80-93.

Paper II

Osmundsen, Kjartan Kloster, Tore Selland Kleppe, and Atle Oglend (2019). MCMC for Markov-switching models - Gibbs sampling vs. marginalized likelihood. *Communications in Statistics - Simulation and Computation*, 1-22.

Paper III

Osmundsen, Kjartan Kloster, Tore Selland Kleppe, and Roman Liesenfeld (2019). Importance Sampling-based Transport Map Hamiltonian Monte Carlo for Bayesian Hierarchical Models. *Submitted for publication in Journal of Computational and Graphical Statistics.*

Paper IV

Osmundsen, Kjartan Kloster, Tore Selland Kleppe, Roman Liesenfeld, and Atle Oglend (2020). Estimating the Competitive Storage Model with Stochastic Trends in Commodity Prices. *Submitted for publication in Journal of Applied Econometrics.*

Table of Contents

Preface iii

Abstract iv

List of papers vi

1 Introduction 1

2 Bayesian inference 3

3 Hamiltonian Monte Carlo 5

4 State-space models 7

5 Particle filters 9

6 Credit risk 11

7 Summary of the papers 13

References 15

Appendix

Using expected shortfall for credit risk regulation 19

MCMC for Markov-switching models - Gibbs sampling vs. marginalized likelihood 35

Importance Sampling-based Transport Map Hamiltonian Monte Carlo for Bayesian Hierarchical Models 61

Estimating the Competitive Storage Model with Stochastic Trends in Commodity Prices 98

1 Introduction

To an increasing extent, more sophisticated statistical methods are applied in economic decision making and management. A clear indication of this is the observation that companies now recruit data analysts. A crucial cause of this change in business methods is technical developments that have made data computations less expensive. In the dissertation I evaluate and develop existing statistical methods and use them for economic applications, e.g. credit risk evaluation and commodity pricing.

The following sections give a basic and informal introduction of the statistical concepts and methods relevant for the papers of this thesis.

More specifically, Section 2 introduces Bayesian inference and Markov chain Monte Carlo. Hamiltonian Monte Carlo, which plays a key part in papers II and III, is introduced in Section 3. Then, Section 4 presents state-space models, which are employed in papers II-IV. Paper III estimates the parameters of the state-space model using particle filters and particle Markov chain Monte Carlo, methods which are presented Section 5. Section 6 introduces credit risk, which is the topic of Paper IV, while Section 7 summarizes the papers of the thesis.

2 Bayesian inference

In Bayesian inference, the unknown model parameters θ are given a *prior* distribution $p(\theta)$, which represents the believed distribution of the parameters before any data enters the analysis. The chosen prior distribution may be based on expert knowledge, or it can simply be chosen on the basis of appealing computational properties (conjugate prior). A *vague* prior is chosen if one wants the prior to play a minimal role in the resulting *posterior* distribution (Gelman et al., 2014).

Given n data observations $y = Y_1, \ldots, Y_n$ and a statistical model $p(y|\theta)$ that reflects the beliefs about the data given the model parameters, the posterior distribution $p(\theta|y)$ is obtained by Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}.$$

That is, the posterior distribution is the updated believed distribution for the model parameters, given the observed data.

One is typically interested in finding the posterior mean of the parameters. This can be achieved through integration:

$$E[\theta|y] = \int \theta p(\theta|y)\, d\theta = \int \theta \frac{p(y|\theta) \cdot p(\theta)}{\int p(y|\theta) \cdot p(\theta)\, d\theta}\, d\theta. \tag{2.1}$$

Note that the analytic form of the normalizing constant $p(y)$ is not necessarily known, thus expressed as an integral in Eq. (2.1). For high dimensions and/or non-Gaussian distributions, it quickly becomes infeasible to solve Eq. (2.1) analytically.

2.1 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is an algorithm constructed to simulate from complex and high-dimensional probability distributions. The idea of MCMC is to construct a *Markov chain* that has the desired distribution as its equilibrium distribution. A discrete-time Markov chain

is a sequence of random variables, $X_1, X_2, \ldots, X_n$, with the Markov property, meaning that the probability distribution of the next random variable depends only on the present variable:

$$p(X_n|X_{n-1}, X_{n-2}, \ldots, X_0) = p(X_n|X_{n-1}). \tag{2.2}$$

To estimate an integral on the form $\int h(x)f(x)dx$, we need to construct a Markov chain whose stationary distribution is $f(x)$. Starting at a chosen initial state X_0, the Markov chain X_i, $i \in (2, 3, \ldots, N)$ is generated according to Eq. (2.2). Following from the law of large numbers for Markov chains, we have that

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i) \xrightarrow{P} E[h(X)] = \int h(x)f(x)dx.$$

2.2 Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) is the most common MCMC algorithm. It has two very appealing properties: First, it is not needed to sample from a Markov chain with the exactly correct equilibrium distribution, as the algorithm makes the adjustments needed. Second, the normalizing constant is not needed, as only probability ratios are considered.

Given the arbitrarily starting state X_0, the remaining Markov chain is constructed as follows for each time step i:

1. Generate a value from the *proposal distribution*: $Y \sim Q(y|X_{i-1})$
2. Evaluate the *acceptance probability*:

$$\alpha(Y, X_{i-1}) = \min\left\{1, \frac{f(Y)}{f(X_{i-1})}\frac{Q(X_{i-1}|Y)}{Q(Y|X_{i-1})}\right\}.$$

3. Determine the next state of the Markov chain:

$$X_i = \begin{cases} Y, \text{ with probability } \alpha \\ X_{i-1}, \text{ with probability } \alpha - 1. \end{cases}$$

3 Hamiltonian Monte Carlo

Over the past decade, Hamiltonian Monte Carlo (HMC) introduced by Duane et al. (1987) has been extensively used as a general-purpose MCMC method, often applied for simulating from posterior distributions arising in Bayesian models (Neal, 2011). HMC offers the advantage of producing close to perfectly mixing MCMC chains by using the dynamics of a synthetic Hamiltonian system as proposal mechanism. The method has its origins from physics, and the total energy of the Hamiltonian dynamical system is described by a 'position coordinate' q and a 'momentum variable' p:

$$H(q, p) = -\log \pi(q) + \frac{1}{2} p^T M^{-1} p, \tag{3.1}$$

where M is a 'mass matrix' representing an HMC tuning parameter. When using HMC to sample from an analytically intractable target distribution $\pi(q)$, the variable of interest (q) is taken as the position coordinate, while the momentum is treated as an auxiliary variable, typically assumed to be independently Gaussian distributed.

Hamilton's equations describe how q and p change over time:

$$\begin{aligned} \frac{d}{dt} p(t) &= -\nabla_q H\left(q(t), p(t)\right) = \nabla_q \log \pi(q), \\ \frac{d}{dt} q(t) &= \nabla_p H\left(q(t), p(t)\right) = M^{-1} p. \end{aligned} \tag{3.2}$$

It can be shown that the dynamics associated with Hamilton's equations are time-reversible, and that it keeps the Hamiltonian (Eq. (3.1)) invariant. However, for all but very simple scenarios, the transition dynamics according to Eq. (3.2) does not admit a closed-form solution, making it necessary to approximate the dynamics using a numerical integrator. The approximation error can be exactly corrected by introducing an accept-reject step (see, e.g., Neal, 2011).

More specifically, each iteration of the HMC algorithm involves the following steps:

1. Refresh the momentum $p^{(k)} \sim N(0, M)$.

2. Propagate approximately the dynamics (3.2) from $(q(0), p(0)) = (q^{(k)}, p^{(k)})$ to obtain $(q^*, p^*) \approx (q(L\varepsilon), p(L\varepsilon))$ using L integrator steps with step size ε.

3. Set $q^{(k+1)} = q^*$ with probability

$$\min\left\{1, \exp\left(H(q^{(k)}, p^{(k)}) - H(q^*, p^*)\right)\right\}$$

and $q^{(k+1)} = q^{(k)}$ with remaining probability.

It is critical that the selection of the time-discretizing step size accounts for the inherent trade-off between the computing time required for generating accept-reject proposals and their quality, reflected by their corresponding acceptance rates. However, the energy preservation properties of the numeric integrator also rely on the nature of the target distribution (for any given step size). High-dimensional, highly non-Gaussian targets typically require small step sizes, whereas high-dimensional near-Gaussian targets can be sampled efficiently with rather large step sizes (Neal, 2011).

4 State-space models

Over the last decades, state-space models (SSMs) have gained interest in the field of time series analysis. It is an extremely flexible model class, with a hierarchical probabilistic structure. SSMs include the widely used and less flexible ARIMA models as special cases.

SSMs are based on the assumption of an unobservable *state* process, which in turn generates an observed time series. The observations are typically a noisy function of this underlying process, which may have a physical interpretation, but can also simply be an auxiliary random process to facilitate a more flexible model specification. The general SSM may be expressed as:

$$\begin{aligned} y_t &= g_t(\cdot|x_t, \theta, e_t), \\ x_t &= h_t(\cdot|x_{t-1}, \theta, \eta_t), \end{aligned} \tag{4.1}$$

where y_t and x_t are the observed value and unobservable state at time t, respectively, θ is a constant parameter vector, while e_t and η_t are two independent noise sequences. The term state-space model is used when the state variables are continuous. For discrete state variables, one usually use the term Markov-switching model, or hidden Markov model. The model specification in Eq. (4.1) results in the dependence structure shown in Figure 4.1. We see that y_t is conditionally independent from past observations, given the value of x_t. This means that the unobservable state always contains the full information of the past observations, making the process (x_t, y_t) Markovian. Thus, the joint

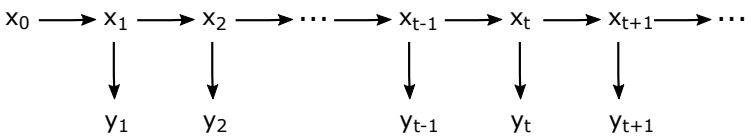

Figure 4.1: SSM dependence structure.

likelihood of $y_{1:T}$ and $x_{1:T}$ can easily be expressed recursively:

$$\begin{aligned} p(x_{1:T}, y_{1:T}|\theta) &= p(x_1) \prod_{t=1}^{T} f(y_t|x_t, \theta) p(x_t|x_{t-1}, \theta) \\ &= p(x_1) \prod_{t=1}^{T} g_t(\cdot|x_t, \theta, e_t) h_t(\cdot|x_{t-1}, \theta, \eta_t), \end{aligned} \tag{4.2}$$

where T is the number of available observations and $x_{1:T}$ denotes the vector $(x_1, \ldots, x_T)$. If the flexible functions g_t and h_t in Eq. (4.1) are both linear and Gaussian, the marginal likelihood of $y_{1:t}$ is available in closed form. This is not the general case, and recent developments in modern Monte Carlo methods have made statistical inference possible for more complex non-linear and non-Gaussian models. It is typically computationally expensive to fit data to an SSM (Shephard and Pitt, 1997; Durbin and Koopman, 1997; Andrieu et al., 2010), due to the naturally large number of latent parameters. The increased computation power available in recent years have made complex SSMs a field of interest, and the flexibility of the models have ensured a wide range of applications.

5 Particle filters

For state-space models and similar dynamic models, sequential Monte Carlo methods/particle filters can be used to produce unbiased estimates of the joint likelihood, while being relatively straightforward to implement. Particle filters are a set of flexible and powerful simulation-based methods, which approximates the marginal likelihood $p(y_{1:t})$ by generating samples that targets the state distribution $p(x_{1:t}|y_{1:t})$.

5.1 Sequential importance sampling

Importance sampling is a general estimation method particularly useful for cases where it is infeasible to sample from the distribution of interest. It consists of replacing the original sampler $p(x)$ by an auxiliary sampler $q(x)$:

$$E[x] = \int x \cdot w(x) \cdot q(x)\, dx, \quad w(x) = \frac{p(x)}{q(x)},$$

where $w(x)$ is the *weight function*.

Sequential importance sampling (SIS) is a computationally effective algorithm for distributions on the form of Eq. (4.2), which entails choosing an importance density with a recursive structure:

$$q_T(x_{1:T}) = q_1(x_1) \prod_{t=2}^{T} q_t(x_t|x_{1:t-1}).$$

This results in the overall importance weight $w_{1:T}$ being the product of all the incremental importance weights:

$$w_{1:T} = w_1 \prod_{t=2}^{T} w_t = \frac{p(x_1, y_1)}{q_1(x_1)} \prod_{t=2}^{T} \frac{p(x_t, y_t|x_{t-1}, y_{t-1})}{q_t(x_t|x_{1:t-1})}, \qquad x_t \sim q_t.$$

Thus, the density in Eq. (4.2) can be approximated by drawing $x_{1:T}^{(i)}, i = 1, \ldots, N$ from the importance density $q_T(x_{1:T})$, and calculate the corresponding importance weights. In particular, an estimate of the marginal likelihood is given by $\hat{p}(y_{1:T}) = \prod_{t=1}^{T} \left\{ \frac{1}{N} \sum_{i=1}^{N} w_{1:t}^{(i)} \right\}$.

SIS is known to fail in the long run, as the weights become highly degenerate (see, e.g., Cappé et al., 2007; Doucet and Johansen, 2009). The sampling importance resampling particle filter of Gordon et al. (1993) mitigates this effect by adding a resampling step, where the importance samples $x_t^{(i)}$ are sampled with probability $w_t^{(i)}$, with replacement.

5.2 Particle Markov chain Monte Carlo

When employing a particle filter to estimate the marginal likelihood $p(y_{1:T})$, the particle marginal Metropolis-Hastings (PMMH) approach developed by Andrieu et al. (2010) is well suited for Bayesian inference. The PMMH uses unbiased Monte Carlo (MC) estimates of the marginal likelihood inside a standard MH algorithm, targeting the posterior for the parameters $p(\theta|y_{1:T})$. The MC estimation error of the likelihood estimate does not affect the invariant distribution of the MH, so that the PMMH allows for exact inference.

The PMMH produces an MCMC sample $\{\theta_i\}_{i=1}^S$ from the target distribution by the following MH updating scheme: given the previously sampled θ_{i-1} and the corresponding likelihood estimate $\hat{p}_{\theta_{i-1}}(p_{1:T})$, a candidate value θ_* is drawn from a proposal density $Q(\theta|\theta_{i-1})$, and the estimate of the associated likelihood is $\hat{p}_{\theta_*}(p_{1:T})$ computed. Then the candidate θ_* is accepted as the next simulated θ_i with probability

$$\alpha(\theta_*, \theta_{t-1}) = \min\left\{1, \frac{\hat{p}_{\theta_*}(p_{1:T})p(\theta*)}{\hat{p}_{\theta_{i-1}}(p_{1:T})p(\theta_{i-1})}\frac{Q(\theta_{i-1}|\theta_*)}{Q(\theta_*|\theta_{i-1})}\right\},$$

otherwise θ_i is set equal to θ_{i-1}. Under weak regularity conditions, the resulting sequence $\{\theta_i\}_{i=1}^S$ converges to samples from the target density $p(\theta|y_{1:T})$ as $S \to \infty$ (Andrieu et al., 2010, Theorem 4).

6 Credit risk

Credit risk models encompass all of the policies, procedures and practices used by a bank in estimating a credit portfolio's probability density function of future credit losses. Such models enable banks to identify, measure and manage risk. As credit risk models have gained a large role in banks' internal risk management processes, they are now also utilized for supervisory and regulatory purposes (Bank for International Settlements, 1999).

The three main parameters of credit risk are *probability of default* (PD), *exposure at default* (EAD) and *loss given default* (LGD). The PD is the probability of a borrower not meeting the debt obligations, typically defined for a time horizon of one year. EAD is the total value a bank is exposed to when a loan defaults, while LGD is the proportion of the EAD the bank is likely to lose in case of default.

Multiplying these three risk parameters, one obtains the *expected loss* (EL):

$$EL = PD \cdot EAD \cdot LGD, \tag{6.1}$$

which is the bank's expected credit loss over the chosen time horizon, typically covered by provisioning and pricing policies (Bank for International Settlements, 2005).

Banks typically use the *unexpected loss* (UL) to express the risk of a portfolio, which is the amount by which the incurred credit loss exceeds the expected loss. The economic capital held to support a bank's credit risk exposure is usually determined by a target insolvency rate. The potential unexpected loss for which it is judged too expensive to allocate capital is called *stress loss*, and leads to insolvency. This is illustrated in Figure 6.1. The estimated probability density function of future credit losses is the basis for calculating the unexpected loss, and the target insolvency rate is normally chosen so that the resulting economic capital will cover all but the most extreme events.

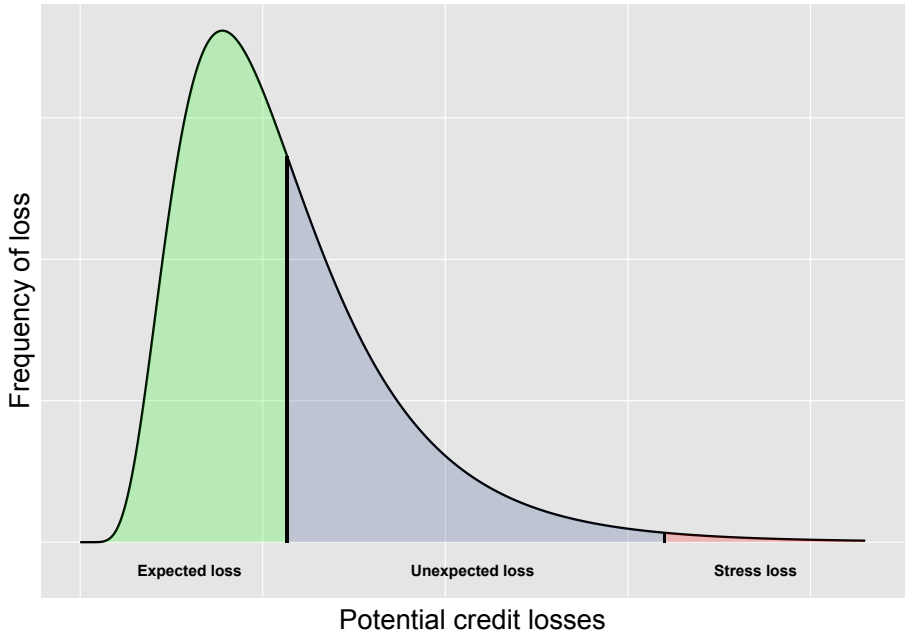

Figure 6.1: The three different types of loss in credit risk modelling.

7 Summary of the papers

The first paper of the thesis, "Using expected shortfall for credit risk regulation", considers the consequences of employing a different risk measure as the basis for the Basel Committee's minimum capital requirement function for banks' credit risk exposures. The currently used risk measure, *value at risk*, is compared to *expected shortfall*, which is already replacing value at risk for market risk regulation. For both risk measures, the paper examines in detail the sensitivity to the tail of the loss distribution. It also compares confidence levels, estimation uncertainty, model validation and parameter sensitivity. The empirical analysis is carried out by both theoretical simulations and real data from a Norwegian savings bank group's corporate credit portfolio. The findings indicate that a transition to a correctly calibrated expected shortfall results in similar capital requirement levels, with slightly increased levels for exposures with very low default probability. The estimation precision is not inferior to value at risk, even at very high confidence levels.

In the second paper, "MCMC for Markov-switching models - Gibbs sampling vs. marginalized likelihood", written in collaboration with the professors Tore Selland Kleppe and Atle Øglend, we propose a method for estimating 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.

The third paper, "Importance Sampling-based Transport Map Hamiltonian Monte Carlo for Bayesian Hierarchical Models", written in collaboration with the professors Tore Selland Kleppe and Roman Liesenfeld,

proposes an importance sampling (IS)-based transport map Hamiltonian Monte Carlo procedure for performing full Bayesian analysis in general non-linear high-dimensional hierarchical models. Using IS techniques to construct a transport map, the proposed method transforms the typically highly challenging target distribution of a hierarchical model into a target which is easily sampled using standard Hamiltonian Monte Carlo. Conventional applications of high-dimensional IS, where infinite variance of IS weights can be a serious problem, require computationally costly high-fidelity IS distributions. An appealing property of our method is that the IS distributions employed can be of rather low fidelity, making it computationally cheap. We illustrate our algorithm in applications to challenging dynamic state-space models, where it exhibits very high simulation efficiency compared to relevant benchmarks, even for variants of the proposed method implemented using a few dozen lines of code in the Stan software package.

In the fourth paper, "Estimating the Competitive Storage Model with Stochastic Trends in Commodity Prices", written in collaboration with the professors Tore Selland Kleppe, Atle Øglend and Roman Liesenfeld, we propose a state-space model (SSM) for commodity prices. The model decomposes the observed price into a component explained by the competitive storage model and a stochastic trend component, and use a particle filter to jointly estimate the structural parameters of the storage model and the trend parameters. Our storage SSM with stochastic trend fits into the economic rationality of storage decisions, and expands the range of commodity markets for which storage models can be empirically applied. The storage SMM is applied to cotton, aluminium, coffee and natural gas markets, and is compared to reduced form stochastic trend models without a structural storage price component, as well as the deterministic trend approach of Gouel and Legrand (2017). Results suggest that the storage component in the SSM adds empirically relevant non-linear price behaviour to reduced form stochastic trend representations and leads to estimates for storage costs and price elasticities of demand which are larger than those obtained under storage models with deterministic trends.

References

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.

Bank for International Settlements (1999). Credit risk modelling: Current practices and applications. `http://www.bis.org/publ/bcbs49.pdf`.

Bank for International Settlements (2005). An Explanatory Note on the Basel II IRB Risk Weight Functions. `http://www.bis.org/bcbs/irbriskweight.pdf`.

Cappé, O., S. J. Godsill, and E. Moulines (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE 95*(5), 899–924.

Doucet, A. and A. M. Johansen (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering 12*(656-704), 3.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Physics letters B 195*(2), 216–222.

Durbin, J. and S. J. Koopman (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika 84*(3), 669–684.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. Rubin (2014). *Bayesian Data Analysis* (3 ed.). CRC Press.

Gordon, N. J., D. J. Salmond, and A. F. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, Volume 140, pp. 107–113. IET.

Gouel, C. and N. Legrand (2017). Estimating the competitive storage model with trending commodity prices. *Journal of Applied Econometrics 32*(4), 744–763.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics 21*(6), 1087–1092.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2*, 113–162.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*(3), 653–667.

Appendix

Using expected shortfall for credit risk regulation 19

MCMC for Markov-switching models - Gibbs sampling vs. marginalized likelihood.....35

Importance Sampling-based Transport Map Hamiltonian Monte Carlo for Bayesian Hierarchical Models 61

Estimating the Competitive Storage Model with Stochastic Trends in Commodity Prices 98

Paper I

Using expected shortfall for credit risk regulation

J. Int. Financ. Markets Inst. Money 57 (2018) 80–93

Contents lists available at ScienceDirect

Journal of International Financial Markets, Institutions & Money

journal homepage: www.elsevier.com/locate/intfin

Using expected shortfall for credit risk regulation ☆

Kjartan Kloster Osmundsen

Department of Mathematics and Physics, University of Stavanger, Norway

ARTICLE INFO

Article history:
Received 12 December 2017
Accepted 4 July 2018
Available online 6 July 2018

Keywords:
Expected shortfall
Credit risk
Bank regulation
Basel III
Tail risk

ABSTRACT

The Basel Committee's minimum capital requirement function for banks' credit risk is based on value at risk. This paper performs a statistical analysis that examines the consequences of instead basing it on expected shortfall, a switch that has already been set in motion for market risk regulation. The ability to capture tail risk as well as diversification is examined in detail for the two risk measures. In addition, the article compares confidence levels, estimation uncertainty, model validation and parameter sensitivity. The empirical analysis is carried out by both theoretical simulations and real data from a Norwegian savings bank group's corporate portfolio. The findings indicate that the use of correctly calibrated expected shortfall results in similar capital requirement levels, with slightly increased levels for exposures with very low default probability. The estimation precision is not inferior to value at risk, even at very high confidence levels.

1. Introduction

Since Artzner et al. (1997) showed that value at risk (VaR) is not sub-additive in general, and thus not always reflecting the positive effect of diversification, several sub-additive risk measures have been proposed. Among these, Expected Shortfall (ES) (Acerbi and Tasche, 2002) has gained most interest. A comprehensive literature compares the properties and relative performance of these two risk measures, see e.g. Yamai and Yoshiba (2005) and Emmer et al. (2015). There are also several comparisons of VaR and ES in a regulatory context specific to market risk, see e.g. Basel Committee on Banking Supervision (2011), Kinateder (2016), Chang et al. (2016), while Frey and McNeil (2002) compare the two risk measures for optimization of credit risk portfolios. Guegan and Hassani (2018) point out the importance of taking into account the distribution and confidence levels when comparing VaR and ES. The present paper compares VaR and ES in a regulatory context specific to credit risk, which, to the best of the author's knowledge, is not yet present in the literature.

The Basel Committee on Banking Supervision aims to enhance financial stability worldwide, partly by setting minimum standards for the regulation and supervision of banks (Basel Committee on Banking Supervision, 2014). In 2004, the introduction of the Committee's second international regulatory accord, Basel II (Basel Committee on Banking Supervision, 2004), opened the possibility for banks to calculate their minimum capital requirements using risk parameters estimated by internal models, instead of using given standard rates (the standardised approach). The minimum capital requirement function is *portfolio-invariant*, so that a single loan's marginal contribution to the total credit risk of a portfolio can be calculated

☆ I would like to direct great thanks to Jacob Laading for stimulating discussions and constructive feedback. I would also like to express gratitude to one anonymous referee for many useful suggestions. Thanks are also due to a Norwegian savings bank group that provided the corporate portfolio data material used in the paper, and to Roy Endré Dahl, Tore Selland Kleppe, Sindre Lorentzen and Atle Øglend for contributing with valuable inputs and feedback.

E-mail address: kjartan.k.osmundsen@uis.no

https://doi.org/10.1016/j.intfin.2018.07.001
1042-4431/

independently from the rest of the portfolio. It is designed to limit each financial institution's isolated risk, while systematic risk is limited by requiring systematically important institutions to hold additional capital.

In January 2016, the Basel Committee published revised standards for calculation of minimum capital requirements for market risk (Basel Committee on Banking Supervision, 2016a), which include a shift from VaR to ES as the underlying risk measure. The Committee stated that the former reliance on VaR largely stems from historical precedent and common industry practice. This has been reinforced over time by the requirement to use VaR for regulatory capital purposes. The Committee recognizes that a number of weaknesses have been identified with VaR, including its inability to capture tail risk. They believe the new ES model will provide a broadly similar level of risk capture as the existing VaR model, while providing a number of benefits, including generally more stable model output and often less sensitivity to extreme outlier observations (Basel Committee on Banking Supervision, 2013b).

A transition from VaR to ES for measuring credit risk has so far not been considered. In fact, default risk in the trading book (market risk) is still to be calculated using VaR, as for the banking book (credit risk), to mitigate the risk of regulatory arbitrage. The Committee has also argued that ES might be too unstable at such high confidence levels (Basel Committee on Banking Supervision, 2013b). In 2013, ES was proposed for the securitisation framework (Basel Committee on Banking Supervision, 2013a), but VaR was retained to keep consistency with the credit risk framework (Basel Committee on Banking Supervision, 2013c). The Committee has also introduced more objective rules for determining whether instruments should be assigned to the trading book or the banking book, and imposed strict constraints on switching between books.[1] However, regulatory arbitrage is still present in the form of internal risk-weight model manipulation (Mariathasan and Merrouche, 2014; Ferri and Pesic, 2017).

The development of credit risk models lies a few years behind the market risk models,[2] due to challenges related to the infrequent nature of default events and the long time horizons involved, making it difficult to collect enough relevant data. This makes it plausible that ES might be considered for credit risk in a not so distant future, and this paper sets out to explore the potential effects.

The paper is structured as follows. Section 2 introduces VaR and ES and compares general theoretical properties, estimation uncertainty, confidence level calibration and model validation. Section 3 introduces the risk parameters used for credit risk modelling, and shows how the Basel Committee has derived a capital requirement function with VaR as the underlying risk measure, and how the same can be done for ES. Section 4 compares VaR and ES within the Basel Committee's minimum capital requirement framework, focusing on confidence level calibration and parameter sensitivity. Lastly, the VaR and ES version of the capital requirement function are compared for cases where the Basel Committee's assumption of normally distributed losses does not hold, using real risk parameter estimates from a Norwegian savings bank group's corporate portfolio. The final conclusions are given in Section 5.

2. Value at risk versus expected shortfall

Throughout this paper it is assumed that losses are expressed as positive numbers, thus focusing on the upper quantile of the profit-loss distribution. For a given confidence level α, VaR is simply defined as the α-quantile of the profit-loss distribution. This implies that the probability of losses exceeding VaR_α equals $(1-\alpha)$. This conceptual simplicity, together with its easy implementation, has made VaR a very popular risk measure.

The simple nature of VaR is also the reason for its shortcomings. By definition, VaR gives no information about the magnitude of losses beyond the VaR level. Consequently, VaR calculations are not affected by the shape of the loss distribution beyond the α-quantile. This is commonly referred to as tail risk, and can be particularly problematic if the loss distribution is heavy-tailed. Assets with higher potential for large losses may appear less risky than assets with lower potential for large losses. For assets with a probability of loss less than α, this leads to VaR disregarding the increase of potential loss due to portfolio concentration, see e.g. Example 2 in Yamai and Yoshiba (2002b). As a consequence, Pillar 2 of the Basel Committee's regulatory framework contain complementary measures to reduce credit concentration (Basel Committee on Banking Supervision, 2004).

Based on similar argumentation, Artzner et al. (1997) proved that VaR is not sub-additive[3] in general, i.e. not always reflecting the positive effect of diversification. More precisely, VaR only satisfies sub-additivity when the loss distribution belongs to the elliptical distribution family and has finite variance (Embrechts et al., 2002). For these distributions, VaR becomes a scalar multiple of the distribution's standard deviation, which satisfies sub-additivity. This includes the normal distribution, Student's t distribution (for $\nu > 2$) and Pareto distribution (for $\alpha > 2$).

Daníelsson et al. (2013) investigate the sub-additivity of VaR for fat-tailed distributions, and theoretically show that VaR is sub-additive except for the fattest tails. However, they still find that VaR estimated from historical simulations (HS) may lead to violations of sub-additivity, due to what they call the *tail coarseness problem*: "When only using a handful of observations in the estimation of HS, where the estimate is equal to one of the most extreme quantiles, the uncertainty about the location of a specific quantile is considerable, and one could easily get draws whereby a particular loss quantile of a relatively

[1] See Basel Committee on Banking Supervision (2013b), page 52, for more information about the imposed constraints.

[2] Banks have been allowed to use internal models as a basis for calculating their market risk capital requirements since 1997 (Basel Committee on Banking Supervision, 2014), i.e. seven years before the same applied for credit risk.

[3] A risk measure is sub-additive when the risk of a portfolio is less than or equal to the sum of the risk of the individual assets.

fat distribution is lower than the same quantile from a thinner distribution". Through an empirical Monte Carlo study, they show that the sub-additivity of VaR fails most frequently in practice for high confidence levels and fat distribution tails. This can be problematic for credit risk regulation, as it involves particularly high confidence levels.

VaR has become a common industry practice for internal risk calculations, e.g. portfolio optimization. This has been reinforced over time by the requirement to use it for regulatory capital purposes. If VaR is used as a constraint when optimizing the (expected) return on a portfolio, the resulting portfolio is likely to exploit the tail risk of VaR. If the risk capital is determined using VaR, portfolio managers have incentives to choose their portfolios as if operating directly under an unconstrained optimization, by investing in assets where the risk lies beyond the VaR level (Frey and McNeil, 2002). Similarly, the profit–loss distribution can be manipulated so that VaR becomes small while the tail becomes fat (Yamai and Yoshiba, 2005).

As an alternative to VaR, Artzner et al. (1997) proposed a risk measure called tailed conditional expectation (TCE). ES was proposed by Acerbi and Tasche (2002) as an extended version of TCE, that is sub-additive also for non-continuous probability distributions.

Given a confidence level $\alpha \in (0, 1)$, the ES of a position L is defined as

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha}\int_{u=\alpha}^{1} \mathrm{VaR}_u(L)du. \tag{1}$$

Another useful representation illustrates how ES differs from TCE:

$$\mathrm{ES}_\alpha(L) = E[L|L \geqslant \mathrm{VaR}_\alpha(L)] + (E[L|L \geqslant \mathrm{VaR}_\alpha(L)] - \mathrm{VaR}_\alpha(L))\left(\frac{P[L \geqslant \mathrm{VaR}_\alpha(L)]}{1-\alpha} - 1\right). \tag{2}$$

When $P[L \geqslant \mathrm{VaR}_\alpha(L)] = 1 - \alpha$, as is the case for continuous distributions, the last term in Eq. (2) vanishes, and ES equals TCE.

From Eqs. (1) and (2) it is clear that ES does not have the same degree of tail risk as VaR. Unlike VaR, ES can distinguish between two distributions of future net worth that have the same quantile but differ otherwise. ES is also more consistent with expected utility maximization (Yamai and Yoshiba, 2005), and is not easily manipulated like VaR (Danielsson and Zhou, 2016).

A critique of ES is the fact that tail behaviour is taken into account through an averaging procedure. Koch-Medina and Munari (2016) claim that averages are poor indicators of risk, and show that surplus outcomes in the tail can compensate for outcomes with large losses and high default probabilities. Comparing different capital positions, this can in some cases cause inconsistency between default behaviour and capital requirement. Thus making VaR a less "deceiving" risk measure, "because it does not purport to contain any information about the tail risk". However, they emphasize that their results do not invalidate ES as a risk measure, but highlights the need for cautious implementation. Note that the above is not an issue if the tail only contains loss outcomes, which is the case for credit risk regulation.[4]

2.1. Estimation methods and backtesting

The reason that VaR remains the most widely used risk measure, seems to be that its practical advantages are perceived to outweigh its theoretical shortcomings. VaR has been considered to have smaller data requirements, easier backtesting (model validation) and in some cases easier calculation than alternative risk measures (Yamai and Yoshiba, 2002a; Kerkhof and Melenberg, 2004; Daníelsson et al., 2013).

There exist multiple promising parametric estimation methods for both VaR and ES. For example, Daníelsson et al. (2013) show that VaR estimated with semi-parametric extreme value techniques tends to violate sub-additivity less frequently than VaR estimated using HS. However, HS is still the preferred method in practice. In addition to its easier implementation, there is also some scepticism towards parametric methods. Danielsson and Zhou (2016) claim that the good performance of a specific parametric model is usually driven by the fact that the model is close to the data generating process (DGP), and that it is not possible to find a parametric model that performs consistently well across all DGPs.

HS uses previous loss data $L_1, L_2, \ldots, L_n$ for estimation. Let $L_{(1)}, \leqslant L_{(2)}, \ldots, \leqslant L_{(n)}$ denote the corresponding order statistics. Then, VaR and ES can be estimated as

$$\widehat{\mathrm{VaR}}_\alpha(L) = L_{(\lfloor n\cdot\alpha \rfloor)}, \quad \widehat{\mathrm{ES}}_\alpha(L) = \left(\sum_{i=\lfloor n\cdot\alpha\rfloor}^{n} L_{(i)}\right) \Big/ (n - \lfloor n\cdot\alpha\rfloor + 1), \tag{3}$$

where $\lfloor x \rfloor$ denotes the largest integer not greater than x.

Gneiting (2011) proved in 2010 that ES is not *elicitable*, as opposed to VaR. This discovery led many to erroneously conclude that ES would not be backtestable, see for instance Carver (2013).

A statistic of a random variable is said to be elicitable if there exists a scoring function (error measure) that is *strictly consistent* for this statistic, meaning that the statistic strictly minimizes the expected value of the scoring function. The mean and the median represent popular examples, minimizing the mean square error and absolute error, respectively (Gneiting,

[4] This is due to the combination of bounded profits and very high confidence levels.

2011). The q^{th} quantile (VaR) is elicitable with the scoring function $S(x,y) = (\mathbf{1}_{\{x \geqslant y\}} - q)(x - y)$, where x is the forecast, y is the corresponding realization and $\mathbf{1}_{\{\}}$ is the indicator function (Acerbi and Szekely, 2014).

It turns out that even if ES is not elicitable, it is still conditionally elicitable, meaning it can be split up in two elicitable components, as both the quantiles and the mean are elicitable (Acerbi and Szekely, 2014). Yet, Danielsson (2013) claims that it is more difficult to backtest ES than VaR, because, when using ES, model predictions cannot be directly compared with observed outcomes. The model predictions are actually compared with model outcomes, a practice that is likely to increase the underlying model risk.

A popular backtesting method for VaR is based on the following *violation process*:

$$I_t(\alpha) = \mathbf{1}\{L(t) > \text{VaR}_\alpha(L(t))\},$$

where $\mathbf{1}\{\}$ denotes the indicator function and t denotes the time period.

Following the definition of VaR, the violations are iid Bernoulli random variables with success probability $1 - \alpha$. Thus, backtesting VaR involves checking if the observed violation process behaves as expected, by satisfying the unconditional coverage hypothesis, $E[I_t(\alpha)] = 1 - \alpha$, in addition to the independence condition (Christoffersen, 1998).

Backtesting ES does not have to be more complicated than backtesting VaR. Emmer et al. (2015) propose a backtesting method for ES that is an extension of the VaR violation method, based on the following approximation:

$$\text{ES}_\alpha(L) = \frac{1}{1-\alpha}\int_{u=\alpha}^{1} \text{VaR}_u(L)du \approx \frac{1}{4}[\text{VaR}_\alpha(L) + \text{VaR}_{0.75\alpha+0.25}(L) + \text{VaR}_{0.5\alpha+0.5}(L) + \text{VaR}_{0.25\alpha+0.75}(L)]. \quad (4)$$

If the four different VaR values in Eq. (4) are successfully backtested, then also the estimate of $\text{ES}_\alpha(L)$ can be considered reliable subject to careful manual inspection of the observations exceeding $\text{VaR}_{0.25\alpha+0.75}(L)$. These tail observations must at any rate be manually inspected in order to separate data outliers from genuine fair tail observations. For market risk, the Basel Committee uses a similar backtesting approach for 97.5% ES, which is based on testing VaR violations for the 97.5% and 99% confidence levels (Basel Committee on Banking Supervision, 2016a).

The literature on backtesting for ES is still increasing, see e.g. Kerkhof and Melenberg (2004), Acerbi and Szekely (2014) and Du and Escanciano (2016) for other promising methods.

2.2. Confidence level

Several comparisons of VaR and ES in the literature use the same confidence level for both risk measures. For example, Yamai and Yoshiba (2002a) conclude that the estimation error of ES is larger than that of VaR when the underlying loss distribution is fat-tailed, by comparing 95% and 99% ES to 95% and 99% VaR. Danielsson and Zhou (2016) also compare 99% VaR to 99% ES as part of a simulation experiment, claiming that "ES is always estimated more inaccurately than VaR". Kerkhof and Melenberg (2004) emphasize that, for capital reserve determination, it makes more sense to compare VaR and ES for confidence levels resulting in the same level of capital requirement. Given the definition of ES in Eq. (1), this means that the ES confidence level must be lower than the VaR confidence level.

For a normally distributed profit-loss function, 99.9% VaR results in the same level of capital requirement as 99.738% ES.[5] For 99% VaR, the corresponding confidence level for ES is 97.423%. For market risk regulation, the Basel Committee replaced 99% VaR with 97.5% ES, i.e. the exact confidence level for ES was rounded up.[6] Based on this, the rest of this paper will consider a rounded up confidence level of 99.75% for credit risk regulation using ES. In practice, one usually encounter distributions with heavier tails than the standard normal distribution (Mandelbrot, 1963; Fama, 1965; Jansen and De Vries, 1991), so this can be seen as an upper bound on the ES confidence level (Kerkhof and Melenberg, 2004). For example, the equivalent ES confidence levels for Student-t(5) and Student-t(2.5) distributions are 99.70% and 99.64%, respectively.

The simulation experiment of Danielsson and Zhou (2016) estimates VaR and ES using simulated values drawn from Student-t distributions with different degrees of freedom. This distribution is very sensible for this kind of experiment, as the degrees of freedom equals the tail index[7] of the distribution. Instead of using the same confidence intervals for VaR and ES, Table 1 shows the results of using comparable confidence levels for their simulation experiment. Compared to the original results, the difference in standard deviations between VaR and ES are significantly reduced. Aside from the case combining the least number of observations and the most heavy-tailed distribution, the accuracy of the VaR and ES estimates is approximately equal. Table 1 also illustrates that higher confidence levels need more observations to get precise estimations, regardless of the chosen risk measure.

Actually, Danielsson and Zhou (2016) also include a comparison of 99% Var to 97.5% ES, using a different empirical approach, where the main result is that the lower bounds of the VaR estimates are significantly higher than that of ES across all sample sizes. This finding matches to some degree the results of Table 1 for $\nu = 2.5$, but not for $\nu = 5$.

[5] $99.9\%\text{VaR} = \Phi^{-1}(0.999) = 3.09$, $99.738\%\text{ES} = \phi\left(\Phi^{-1}(0.99738)\right)/(1 - 0.99738) = 3.09$, where ϕ and Φ denote the density and distribution function of the standard normal distribution, respectively.

[6] $99\%\text{VaR} = 2.3263 \approx 97.5\%\text{ES} = 2.3378$.

[7] In Extreme Value Theory, the tail index β describes how heavy the tail of the distribution is (Haan, 1975). It is defined as regular variation in the tail of the distribution function F: $\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-\beta}$.

Table 1
Finite sample performance of VaR and ES, following the same Monte Carlo method as the results from Table 1 in Danielsson and Zhou (2016). N observations are sampled from a Student-t distribution with ν degrees of freedom. For this group of samples, VaR and ES are estimated using Eq. (3) (Danielsson and Zhou (2016) uses slightly different estimators) and divided by the theoretical value. The resulting ratio is regarded as the relative estimation error, which is simulated 2×10^7 times for each combination of N, ν and α. The table shows the standard deviations and 99% confidence intervals of these ratios.

		VaR			ES		
N	ν	α	sd	99% conf.int.	α	sd	99% conf.int.
300	2.5	99%	0.23	[0.56,1.86]	97.5%	0.31	[0.52,2.2]
300	2.5	99.9%	0.22	[0.26,1.53]	99.75%	0.45	[0.25,2.66]
300	5	99%	0.15	[0.68,1.45]	97.5%	0.14	[0.67,1.4]
300	5	99.9%	0.15	[0.44,1.26]	99.75%	0.2	[0.45,1.61]
1000	2.5	99%	0.13	[0.72,1.42]	97.5%	0.19	[0.68,1.73]
1000	2.5	99.9%	0.35	[0.44,2.49]	99.75%	0.39	[0.39,2.53]
1000	5	99%	0.08	[0.8,1.24]	97.5%	0.08	[0.81,1.24]
1000	5	99.9%	0.18	[0.62,1.63]	99.75%	0.16	[0.6,1.49]
12500	2.5	99%	0.04	[0.91,1.1]	97.5%	0.06	[0.88,1.19]
12500	2.5	99.9%	0.11	[0.75,1.32]	99.75%	0.17	[0.71,1.61]
12500	5	99%	0.02	[0.94,1.06]	97.5%	0.02	[0.94,1.06]
12500	5	99.9%	0.06	[0.85,1.17]	99.75%	0.06	[0.85,1.18]

3. Credit risk modelling

The introduction of Basel II in 2004 opened the possibility for banks to calculate the assets' risk weights using parameter estimates from internal models. To be able to use this *internal ratings based* (IRB) approach, the bank's risk models have to be approved by the national supervisory authorities (Basel Committee on Banking Supervision, 2004).

This section introduces the risk parameters involved, and describes the model choices made by the Basel Committee when deriving the mathematical function for calculating regulatory capital under the IRB approach.

3.1. Risk parameters

The *expected loss* (EL) is the credit loss a bank can expect on its credit portfolio over the chosen time horizon, typically one year. EL is calculated as the mean of the loss distribution, and is typically covered by provisioning and pricing policies (Basel Committee on Banking Supervision, 2005). The expected loss of a single loan can be calculated as follows:

$$EL = PD \cdot LGD \cdot EAD, \tag{5}$$

where *probability of default* (*PD*) is the probability that a borrower will be unable to meet the debt obligations within the given time horizon. The *exposure at default* (*EAD*) is the bank's outstanding exposure to the borrower in case of default, while the *loss given default* (*LGD*) is the bank's likely loss in case of default, usually stated as a percentage of *EAD*.

Banks typically express the risk of a portfolio with the *unexpected loss* (UL), which is the amount by which the actual credit loss exceeds the expected loss. The economic capital held to support a bank's credit risk exposure is usually determined so that the estimated probability of UL exceeding economic capital is less than a target insolvency rate. The potential UL which is judged too expensive to hold capital against is called *stress loss*, and leads to insolvency. The assumed probability density function of future credit losses is the basis for calculating the unexpected loss, and the target insolvency rate is chosen so that the resulting economic capital will cover all but the most extreme events.

3.2. The Basel Committee's capital requirement function

The Basel Committee's capital requirement function for credit risk is based on Gordy's *Asymptotic Single Risk Factor* (ASRF) model (Gordy, 2003), which models risk using a systematic risk factor, which may be interpreted as reflecting the state of the global economy. The model is constructed to be *portfolio-invariant*, so that the marginal capital requirement for a loan does not depend on the properties of the portfolio in which it is held.

The probability of default conditional on the systematic risk factor is calculated by Vasicek's adaptation of the Merton model (Vasicek, 2002), which assumes a normal distribution for the systematic risk factor X:

$$PD(X) = \Phi\left(\frac{\Phi^{-1}(PD) - X\sqrt{R}}{\sqrt{1-R}}\right), \tag{6}$$

where Φ is the distribution function of the standard normal distribution and R is the loan's correlation with the systematic risk factor, i.e. the degree of the bank's exposure to the systematic risk. The unconditional *PD* on the right-hand side reflects the expected default rate under normal business conditions, and is estimated by the bank.

The ASRF model uses VaR as the underlying risk measure, meaning that the required capital is calculated so that the loss probability does not exceed a set target α. This is achieved by holding capital that covers the α^{th} quantile of the assumed loss distribution, i.e. letting the systematic risk factor equal the α^{th} quantile, $q_\alpha(X) = \Phi^{-1}(1-\alpha)$:

$$PD(\Phi^{-1}(1-\alpha)) = PD(-\Phi^{-1}(\alpha)) = \Phi\left(\frac{\Phi^{-1}(PD) + \Phi^{-1}(\alpha)\sqrt{R}}{\sqrt{1-R}}\right). \tag{7}$$

The expected loss is calculated using Eq. (5) without the *EAD*-factor, thus being expressed as a percentage of the exposure at default. Inserting Eq. (7) for *PD* gives the α^{th} quantile of the expected loss conditional on the systematic risk factor *X*, i.e. the VaR (Gordy, 2003):

$$q_\alpha(E[L|X]) = E[L|q_\alpha(X)] = PD(q_\alpha(X)) \cdot LGD. \tag{8}$$

The *LGD* value in Eq. (8) must reflect economic downturn conditions in circumstances where loss severities are expected to be higher during cyclical downturns than during typical business conditions (Basel Committee on Banking Supervision, 2005). This so-called "*downturn*" *LGD* value is not computed with a mapping function similar to Eq. (7). Instead, the Basel Committee has decided to let the banks provide downturn *LGD* values based on their internal assessments. The reason for this is the evolving nature of bank practices in the area of *LGD* quantification.

The Basel Committee's capital requirement function only considers the unexpected loss. As the ASRF model delivers the entire VaR, the expected loss $PD \cdot LGD$ has to be subtracted from Eq. (8):

$$K = LGD \cdot \Phi\left(\frac{\Phi^{-1}(PD) + \Phi^{-1}(0.999) \cdot \sqrt{R}}{\sqrt{1-R}}\right) - PD \cdot LGD, \tag{9}$$

where *K* denotes the capital requirement, as a percentage of total exposure (*EAD*). The Committee has chosen the confidence level $\alpha = 0.999$, which means that unexpected losses on a loan should exceed the capital requirement only once in a thousand years. The reason why the confidence level is set so high is partly to protect against inevitable estimation error in the banks' internal models (Basel Committee on Banking Supervision, 2005).

Fig. 1 shows how Eqs. (8) and (9) depend on the *PD* parameter. The total loss is strictly increasing for larger *PD* values, while the unexpected loss is a concave function of *PD*.

As mentioned above, *R* is the loan's correlation with the systematic risk factor, and it is determined from information about the borrower. For loans to states, institutions and large enterprises (annual revenues above 50 million euros) (Basel Committee on Banking Supervision, 2005) the following formula applies:

$$R = 0.24 - 0.12\left(\frac{1 - e^{-50 \cdot PD}}{1 - e^{-50}}\right). \tag{10}$$

Because Eq. (9) is expressed as a percentage of total exposure, one must multiply by *EAD* to get the capital requirement stated as a money amount. The total money amount shall constitute at least 8% of the risk-weighted assets:

$$\sum_{i=1}^{n} K_i \cdot EAD_i \geqslant 0.08 \cdot \sum_{i=1}^{n} RW_i \cdot EAD_i.$$

Thus, the marginal risk-weight of a single asset is calculated as $RW_i = K_i / 0.08 = 12.5 K_i$.

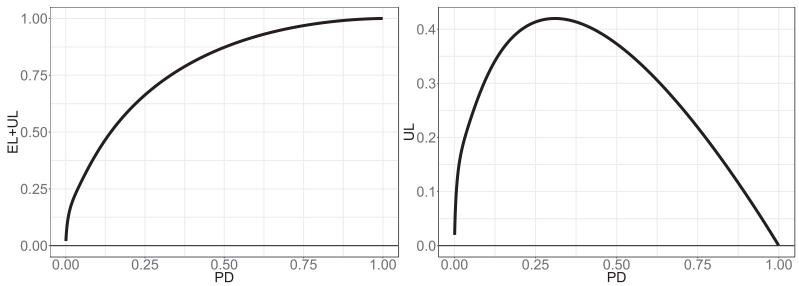

Fig. 1. Total loss (left) and unexpected loss (right) of the Basel Committee's capital requirement function, plotted as a function of the probability of default. Calculated for LGD = 1, with confidence level 99.9%.

The ASRF model is also applicable for ES, with the resulting capital charges being portfolio invariant under the same assumptions as VaR-based capital charges (Gordy, 2003), making it possible to derive a version of Eq. (9) that is based on ES (Hibbeln, 2010):

$$K_{\text{ES}} = \frac{LGD}{1-\alpha}\Phi_2\left(\Phi^{-1}(PD), -\Phi^{-1}(\alpha); \sqrt{R}\right) - PD \cdot LGD, \tag{11}$$

where $\Phi_2(\cdot)$ is the bivariate cumulative normal distribution function.

For market risk regulation, the Committee has decided that the ES calculations should be calibrated as if the relevant risk factors were experiencing a period of stress (Basel Committee on Banking Supervision, 2016a). This will also be the case for Eq. (11) as it depends on downturn *LGD* values and contains a mapping function for the *PD* values, similar to the existing VaR version.

4. VaR versus ES for credit risk

This section examines how the Basel Committee's capital requirement function is affected by the choice of its underlying risk measure, as a practical counterpart to the theoretical comparison of VaR and ES in Section 2.

4.1. Confidence level

Although the derived ES version of the capital requirement function in Eq. (11) is based on the same assumptions as the VaR version in Eq. (9), the difference between the two risk measures is significant enough that the two functions behave quite differently. As already mentioned, 99.75% is the ES confidence level that matches 99.9% VaR. However, as Eqs. (9) and (11) are parametric functions, the difference between the functions depends on the *PD* value.[8]

Conducting a least squares fit over the interval $PD \in (0, 1)$, a confidence level of 99.742% is found to make the ES capital requirement most similar to the 99.9% VaR version, making it reasonable to continue using the 99.75% confidence level for ES. There are however notable differences, as shown in Fig. 2. Compared to the VaR version, the ES version increases capital charges for loans with a low probability of default, and very slightly decreases capital charges for loans with a probability of default exceeding 45% (the maximum reduction is 0.13%, for $PD = 82\%$). As the Basel Committee has proposed to apply lower bounds on the PD estimates (Basel Committee on Banking Supervision, 2016b), the significant increase for the very lowest *PD* values may be desirable from a regulation perspective. However, the proposed floor is only $PD = 0.0005$, so the increased capital charges for $0.0005 < PD < 0.45$ may not be desirable. The increase exceeds 1% for $PD < 0.07$ and exceeds 2% for $PD < 0.02$. At the same time, this could slightly reduce the incentive to purposely report artificially small *PD* estimates.

The ES capital requirement shares the properties of VaR illustrated in Fig. 1, with the total loss strictly increasing for larger *PD* values, so banks are not directly incentivised to shift their exposure to higher risk. However, compared to the existing 99.9% VaR regulation, the relative capital charge will increase for less risky loans and decrease for more risky loans. One might argue that a switch to ES could promote slightly more risk-taking from the banking institutions, potentially threatening the financial stability. The validity of such an argument depends on whether the current credit risk regulation is optimal, from a socioeconomic point of view. There is also the possibility that the current regulation is slightly too much shifted towards the less risky loans, leading to some socially beneficial projects not being granted loans.

4.2. Parameter sensitivity

The parameter sensitivity of the capital requirement function is also of great interest, i.e. how the uncertainty of the *PD* and *LGD* estimates affects the function output. This is examined by applying a simulation experiment. In addition to determine how a change from VaR to ES will influence the parameter sensitivity, it is also explored which parameter the capital requirement function is most sensitive to.

Varying degrees of estimation uncertainty are represented by the variance of the distribution function the parameter values are simulated from. This is achieved by simply sampling random values from a normal distribution with expected value 1 and standard deviation σ, and then multiply these values with the true parameter value, so the distribution is centred around the correct value:

$$\begin{aligned} \widehat{PD} &= PD \cdot \delta, \quad \delta \sim \mathcal{N}(1, \sigma_{PD}) \\ \widehat{LGD} &= LGD \cdot \delta, \quad \delta \sim \mathcal{N}(1, \sigma_{LGD}). \end{aligned} \tag{12}$$

Using Eq. (12), N values of $\widehat{PD}$ and $\widehat{LGD}$ are generated for five different σ values $(0.05, 0.10, 0.15, 0.20, 0.25)$, indicating five different degrees of estimation uncertainty. Fig. 3 illustrates the resulting parameter distributions for three of the σ values.

[8] The *LGD* parameter is not of interest in this capacity, as it has a linear relationship with both versions of the capital requirement function.

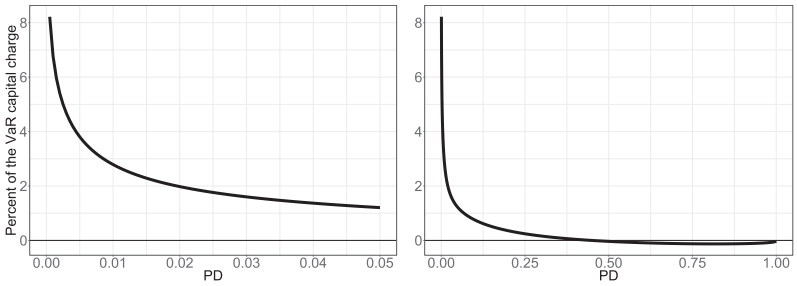

Fig. 2. The percentage difference between the calculated capital requirement from the ES version with confidence level 99.75% and the standard 99.9% VaR version. Positive values mean that the ES version results in a higher capital charge. The left graph gives a detailed view for small PD values, while the right graph shows the whole (0,1) interval.

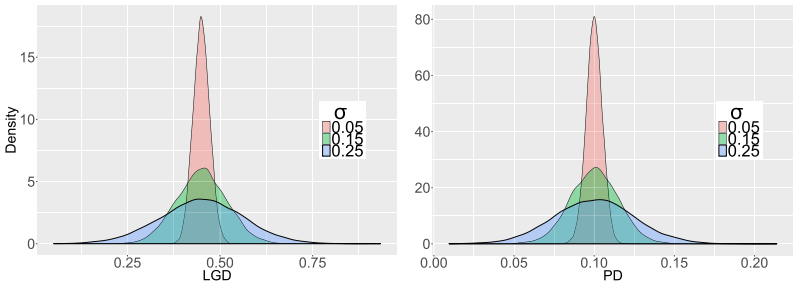

Fig. 3. The distributions of $N = 10,000$ simulated $\widehat{LGD}$ and $\widehat{PD}$ values, for three degrees of estimation uncertainty, σ. True parameter values are $LGD = 0.45$ and $PD = 0.1$.

The simulated $\widehat{LGD}$ and $\widehat{PD}$ values are used as input parameters for Eqs. (9) and (11) to calculate the capital requirements, where the correlation factor R is calculated using Eq. (10). The capital requirement levels $K_i,\ i \in (1, 2, \ldots, N)$ are calculated for all 25 possible combinations of σ_{LGD} and σ_{PD}. For each combination, the relative standard deviation of the capital requirement levels are calculated, for both risk measures:

$$\sigma_{K\text{rel}} = \sqrt{\frac{\sum_{i=1}^{N}(K_i - \overline{K})^2}{N-1}} \Big/ \overline{K}.$$

This simulation process is repeated for different values of PD, to see how this impacts the results. Different LGD values will only result in a linear scaling of the capital requirement, so a constant value of 0.45 is used. In fact, because the capital requirement's uncertainty is expressed as its relative standard deviation, the results are independent of the chosen value for LGD.

4.2.1. Results

Fig. 4 shows the results for the 99.9% VaR capital requirement. As one would expect, the capital requirement's relative uncertainty increases with increasing parameter uncertainty.

More interesting, the parameter uncertainties have different relative impact depending on the PD value. For small PD values, the PD estimation uncertainty is almost as influential as the LGD uncertainty. For medium PD values, precise estimation is less important, as the LGD uncertainty is the main contributor to variations in the calculated capital requirement. For $PD = 0.35$, the impact of PD estimation uncertainty is practically non-existent, as indicated by the solid-colour columns in Fig. 4. The reason for this is proximity to the vertex ($PD = 0.31$) of the concave capital requirement function, as shown in Fig. 1. For larger PD values, the estimation uncertainty steadily increases towards $PD = 1$. Fig. 5 shows the results for $PD = 0.8$, where the PD uncertainty is clearly dominant, as indicated by solid-colour rows. Note that the magnitude of the

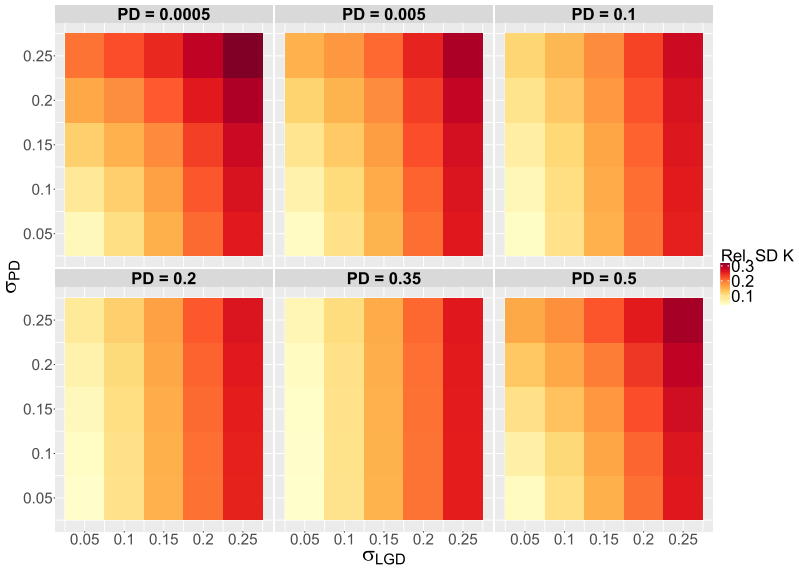

Fig. 4. Relative standard deviation of the 99.9% VaR capital requirement, given σ_{LGD} and σ_{PD}. Calculated for six different *PD* values, with $LGD = 0.45$. Using $N = 10,000$ simulations for each calculation.

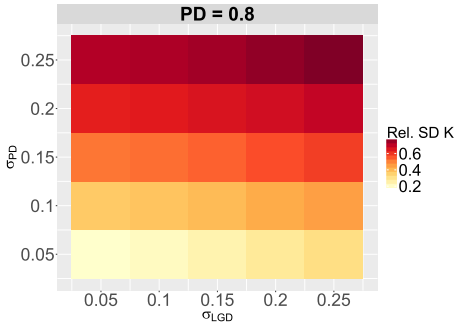

Fig. 5. Relative standard deviation of 99.9% VaR capital requirement, given σ_{LGD} and σ_{PD}. Calculated for $PD = 0.8$ and $LGD = 0.45$, using $N = 10,000$ simulations for each calculation.

capital requirement uncertainty is considerably larger than for the lower *PD* values, as the *LGD* contribution to uncertainty is constant.

Looking at Eqs. (9) and (10) it is clear that $e^{-50\,PD}$ is the part of the capital requirement function that explains the influence of the *PD* values' uncertainty for small *PD* values, as it is very sensitive for *PD* values close to zero. This sensitivity gradually becomes smaller for larger *PD* values, and for $PD > 0.1$ this part of the function is approximately constant. The function part $\Phi^{-1}(PD)$ is particularly sensitive for *PD* values in the far ends of the $(0, 1)$ interval. The major impact of large *PD* values on the capital requirement's uncertainty is due to the last term in Eq. (9). For large *PD* values, this term is no longer small compared to the first term.

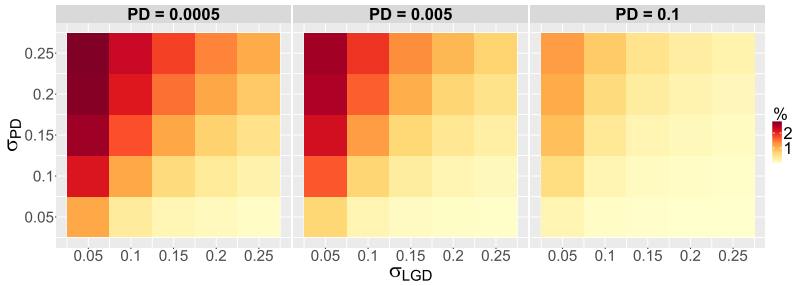

Fig. 6. Percentage reduction of the relative standard deviation of the capital requirement by switching from 99.9% VaR to 99.75% ES, given σ_{LGD} and σ_{PD}. Calculated for three different *PD* values, with $LGD = 0.45$. Using $N = 10{,}000$ simulations for each calculation.

The relative standard deviations of the simulated 99.75% ES capital requirements behave quite similar as for the 99.9% VaR, with the largest deviations for small *PD* values, consistent with the results in Fig. 2. As shown in Fig. 6, these deviations are all favourable of ES, resulting in relative uncertainty reductions of almost 3% in some cases. The relative reduction is largest when the *LGD* uncertainty is low and the *PD* uncertainty is high.

There could be both advantages and disadvantages with a capital requirement function that is less sensitive to the *PD* parameter at the lowest end of the scale. One could argue that this to some degree reduces the banks' incentive to estimate artificially low PD values. At the same time it might be viewed as counter-productive, since the fundamental idea behind the IRB approach is a more risk-sensitive capital requirement.

4.3. Loss distributions

The capital requirement function is derived on the assumption of a normal distributed systematic risk factor. As mentioned in Section 2, asset returns are usually distributed with heavier tails than the normal distribution. This section uses real risk parameter values from a Norwegian savings bank group's corporate portfolio to examine the behaviour of the capital requirement function for such loss distributions.

Loss realizations distributed with different tail weights are generated by simulating values for the systematic risk factor X from the Student-t distribution, a natural choice as its degrees of freedom parameter ν coincides with the tail index of the resulting distribution. 2.5, 5 and 10 degrees of freedom are used. The standard normal distribution (equivalent to $\nu \to \infty$) is used as a baseline.

The real risk parameter data set consists of PD, LGD, R and EAD values for J loans. Using the simulated X values, conditional *PD* values are calculated from Eq. (6) for all the loans in the data set[9] To simulate losses, loan j is considered defaulted if a uniformly distributed random number $U_j \in [0, 1]$ is smaller than or equal to the conditional *PD*. For the defaulted loans, the conditional *PD* is multiplied with the associated risk parameter values to obtain the money amount lost:

$$L(X) = \sum_{j=1}^{J} \mathbf{1}\left\{ \Phi\left(\frac{\Phi^{-1}(PD_j) - X\sqrt{R_j}}{\sqrt{1-R_j}} \right) \geqslant U_j \right\} \cdot LGD_j \cdot EAD_j, \tag{13}$$

where $\mathbf{1}\{\}$ is the indicator function.

For each value of ν, the loss simulation shown in Eq. (13) is repeated for N different X values, so that the resulting N loss values constitute a representation of the loss distribution for the given tail weight. The unexpected losses are then obtained by subtracting $\sum_{j=1}^{J} PD_j \cdot LGD_j \cdot EAD_j$ from Eq. (13), and lastly VaR and ES estimates for the unexpected losses are calculated using Eq. (3). Confidence levels of 99% and 99.9% are used for the VaR estimates, while 97.5% and 99.75% are used for the ES estimates. The entire simulation process is repeated M times to enable the calculation of the mean and relative standard deviation of the VaR and ES estimates.

The simulation code is written in R (R Core Team, 2014). The size of the data set makes this process quite time consuming, so multicore computer processing is enabled to speed up the process, using the packages `foreach`, `parallel` and `doParallel`.

[9] The results using the ES equivalent $\Phi_2(\Phi^{-1}(PD), X; \sqrt{R})/\Phi(X)$ are similar, and thus not reported.

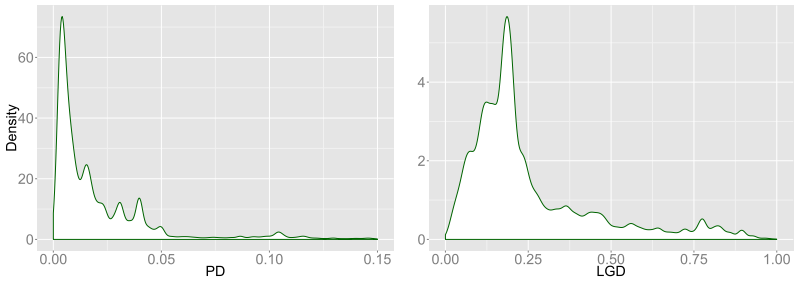

Fig. 7. The distribution of *PD* and *LGD* estimates from a Norwegian savings bank group's corporate portfolio, for loans issued between March 2015 and January 2016. The *PD* values exceeding 0.15 are not included in this figure (2.1% of the loans).

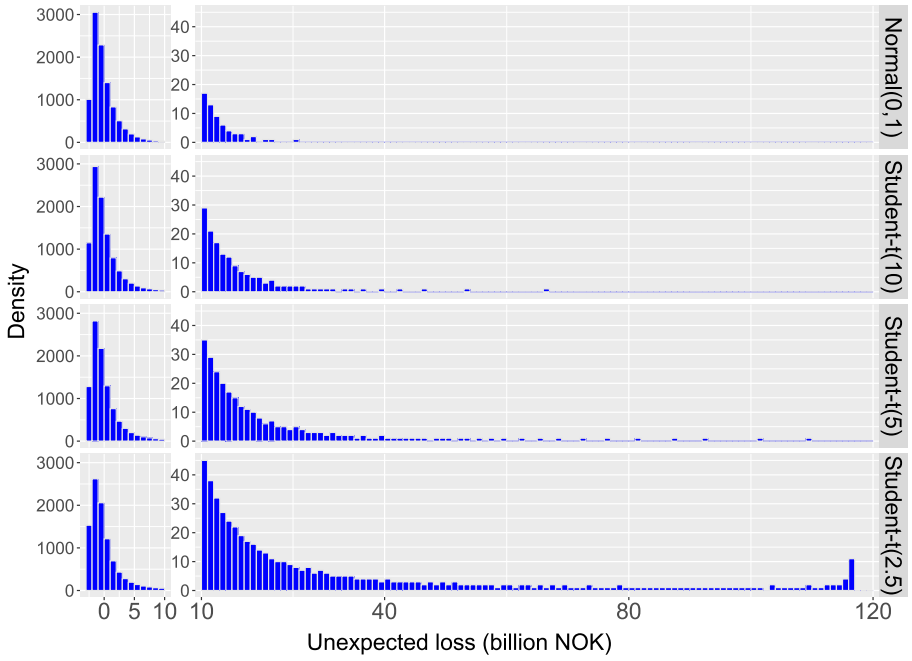

Fig. 8. Unexpected losses given different probability distributions for the systematic risk factor. From the top: Standard normal distribution, Student-t distribution with 10, 5 and 2.5 degrees of freedom. $M = 100$ sets of $N = 10,000$ loss values are simulated for each distribution. The figure shows the distribution of the set means. Two different y-axis are used to make the tail events more visible.

4.3.1. Results

The data set consists of real risk parameter values for corporate loans issued by a Norwegian savings bank group from March 2015 to January 2016. The data set contains about a fifth of the group's total corporate portfolio from this period, picked randomly. This amounts to a total of $J = 109,045$ loans.

Fig. 7 provides some insight about the data set, by displaying density plots of the *LGD* and *PD* values. Most of the loans have low risk, with 90.8% of the loans having been assigned a probability of default of 0.05 or less. Only 2.1% of the loans have

Table 2
The mean and relative standard deviation of $M = 100$ VaR and ES estimates, for the confidence levels corresponding to market risk and credit risk regulation. Each of the M estimates are calculated using $N = 10,000$ loss values, simulated using the probability distribution indicated in the leftmost column for the systematic risk factor.

	Mean (billion NOK)				Relative SD			
	99% VaR	97.5% ES	99.9% VaR	99.75% ES	99% VaR	97.5% ES	99.9% VaR	99.75% ES
Normal(0,1)	8.592	8.897	15.392	15.655	0.037	0.034	0.07	0.064
Student-t(10)	12.224	13.565	29.358	31.125	0.038	0.041	0.1	0.085
Student-t(5)	18.542	22.136	59.66	61.146	0.054	0.053	0.14	0.107
Student-t(2.5)	50.683	53.296	116.095	114.426	0.092	0.057	0.012	0.016

a *PD* value greater than 0.15. The majority of the *LGD* values are also at the low end of the scale, with 68% of the loans having a *LGD* value of 0.25 or less. There are however also a substantial number of loans that have high *LGD* values.

Fig. 8 shows the mean distribution for $M = 100$ sets of $N = 10,000$ (unexpected) loss values simulated using the method described above. As Eq. (6) is derived on the assumption that X follows a normal distribution, the simulations using the more heavy-tailed distributions naturally produce larger maximal estimates for $PD(X)$, thus producing larger maximum losses. The tail of the Student-t(2.5) distribution is so extreme compared to the normal distribution, that the tail simulations result in $PD(X) = 1$ for almost all the loans. For the given data set, this corresponds to $UL = 116.6$ billion NOK. For comparison, the expected loss is only 2.8 billion NOK.

Table 2 shows the means and relative standard deviations of the $M = 100$ VaR and ES estimates calculated from the simulated loss values. The VaR and ES estimates are most dependent on the confidence level for the most heavy-tailed loss distributions. The difference between the 99% VaR and the 97.5% ES increase for the more heavy-tailed distributions, while the high confidence level of the 99.9% VaR makes it behave closer to its ES equivalent, as it takes into account a greater part of the loss distribution function. The relative standard deviations are fairly similar for the VaR and ES estimates. The sum of the capital requirements for all loans in the data set, calculated using Eqs. (9) and (11) respectively, result in a total capital charge of 15.37 billion NOK using 99.9% VaR and 15.70 billion NOK using 99.75% ES, which correspond well with the simulation results for the normal distributed risk factor.

Table 3 in the appendix shows how the number of simulations affects the relative standard deviations of the VaR and ES estimates. A graphic representation of a selection of these results is shown in Fig. 9 in the appendix. As one would expect, the relative SD decreases when you increase the number of simulations. The size of this reduction appears to be approximately equal for the VaR and ES estimates.

To examine the practical implications of real losses being more heavy-tailed than the assumed normal distribution, a version of Eq. (6) assuming a Student-t distributed systematic risk factor is emulated by transforming the simulated X values[10] so that X values simulated from a Student-t(ν) distribution result in the same VaR and ES estimates as for the normal distributed X values. Using this transformation for $\nu = 10, 5$ and 2.5 on normal distributed X values, both VaR and ES estimates decrease by 34%, 50% and 65%, respectively. Compared to the 99.9% VaR, this implies VaR confidence levels of 99.4%, 98.7% and 96.8%, meaning that the *true* confidence level of the capital requirement function is lower than 99.9% if the loss distribution is more heavy-tailed than the assumed normal distribution.

5. Conclusion

The Basel Committee's minimum capital requirement function for banks' credit risk is based on VaR. The paper performs an analysis of the consequences of replacing VaR with ES, a switch that has already been set in motion for market risk.

ES has some well known conceptual advantages over VaR, primarily a better ability to accurately capture tail risk. ES is also sub-additive in general, always reflecting the positive effect of diversification. Additionally, it is more consistent with expected utility maximization, and cannot easily be manipulated like VaR. Nevertheless, VaR is still favoured for its conceptual simplicity, and is considered to have superior estimation precision and model validation methods.

However, the theoretical shortcomings of VaR have no practical impact on the present capital requirement function, as it is a closed-form expression that assumes a normally distributed systematic risk factor. This paper finds that by correctly calibrating the confidence level, an ES of the present capital requirement function will produce approximately the same capital requirement, with the largest differences occurring for low default probabilities, where the capital requirement is slightly higher and the parameter sensitivity is slightly lower. The paper also finds that the use of ES results in a capital requirement function with satisfactory estimation precision, and points out that VaR may be wrongly perceived to have superior estimation precision if non-comparable confidence levels are used.

The Basel Committee's revised standards for calculation of minimum capital requirements for market risk include a shift from VaR to ES as the underlying risk measure for most risk classes. Default risk is the only risk class of the trading book (market risk) that is still to be calculated using VaR, as for the banking book (credit risk), to mitigate the risk of regulatory

[10] $X_t = \Phi^{-1}(F(X, \nu))$, where Φ and F are the cumulative distribution function of the standard normal distribution and the Student-t(ν) distribution, respectively.

arbitrage. This decision is based on the argument that ES might be too unstable at such high confidence levels. As the findings of this paper indicate that this might not be the case, it could be worth to evaluate adapting the existing credit risk regulation to the risk measure used for most of the new market risk regulation.

If this were to be implemented, all possible implications of the increased relative capital charge for low default probabilities must be accounted for. In addition, one must also address how a transition to ES can be adapted to the supervisory review and market discipline pillars of credit risk regulation. As the switch to ES has already been set in motion for market risk, banks are going to have practical experience with ES before this switch potentially would happen for credit risk. In addition, ES is closely related to VaR, so the new implementations will in some cases require only minor adjustments, thus limiting the transition cost.

Appendix A

See Table 3 and Fig. 9.

Table 3
The relative standard deviation of $M = 100$ VaR and ES estimates, for the confidence levels corresponding to market risk and credit risk regulation, for different number of simulated loss values, N. The loss values are simulated using the probability distribution indicated in the table headers for the systematic risk factor.

	Normal(0,1)				Student-t(10)			
N	99% VaR	97.5% ES	99.9% VaR	99.75% ES	99% VaR	97.5% ES	99.9% VaR	99.75% ES
1000	0.093	0.084	0.192	0.164	0.144	0.141	0.293	0.277
2500	0.066	0.057	0.109	0.099	0.079	0.093	0.223	0.2
5000	0.042	0.035	0.088	0.075	0.061	0.068	0.164	0.149
10,000	0.037	0.034	0.07	0.064	0.038	0.041	0.1	0.085
	Student-t(5)				Student-t(2.5)			
N	99% VaR	97.5% ES	99.9% VaR	99.75% ES	99% VaR	97.5% ES	99.9% VaR	99.75% ES
1000	0.197	0.187	0.372	0.306	0.285	0.183	0.116	0.111
2500	0.13	0.129	0.259	0.203	0.166	0.103	0.041	0.04
5000	0.078	0.072	0.165	0.124	0.118	0.072	0.015	0.023
10,000	0.054	0.053	0.14	0.107	0.092	0.057	0.012	0.016

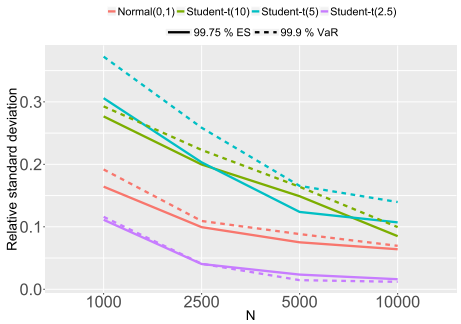

Fig. 9. Relative standard deviation of $M = 100$ VaR (dotted lines) and ES (solid lines) estimates, each estimated using N simulated loss values. 99.9% confidence level is used for VaR, while 99.75% is used for ES.

References

Acerbi, C., Szekely, B., 2014. Back-testing expected shortfall. Risk, 76.
Acerbi, C., Tasche, D., 2002. On the coherence of expected shortfall. J. Bank. Finan. 26 (7), 1487–1503.
Artzner, P., Delbaen, F., Eber, J., Heath, D., 1997. Thinking coherently. Risk 10 (11), 68–71.
Basel Committee on Banking Supervision, 2004. International Convergence of Capital Measurement and Capital Standards.
Basel Committee on Banking Supervision, 2005. An Explanatory Note on the Basel II IRB Risk Weight Functions.
Basel Committee on Banking Supervision, 2011. Messages from the Academic Literature on Risk Measurement for the Trading Book.
Basel Committee on Banking Supervision, 2013a. Foundations of the Proposed Modified Supervisory Formula Approach.
Basel Committee on Banking Supervision, 2013b. Fundamental Review of the Trading Book: A Revised Market Risk Framework.

Basel Committee on Banking Supervision, 2013c. Revisions to the Securitisation Framework.
Basel Committee on Banking Supervision, 2014. A Brief History of the Basel Committee.
Basel Committee on Banking Supervision, 2016a. Minimum Capital Requirements for Market Risk.
Basel Committee on Banking Supervision, 2016b. Reducing Variation in Credit Risk-weighted Assets - Constraints on the use of Internal Model Approaches.
Carver, L., 2013. Mooted VaR substitute cannot be back-tested, says top quant. Risk (March 8).
Chang, C.-L., Jiménez-Martín, J.-Á., Maasoumi, E., McAleer, M., Perez Amaral, T., 2016. Choosing Expected Shortfall over VaR in Basel III Using Stochastic Dominance. USC-INET Research Paper 16(05). Available at SSRN: <https://ssrn.com/abstract=2746710>.
Christoffersen, P.F., 1998. Evaluating interval forecasts. Int. Econ. Rev., 841–862
Danielsson, J., 2013. The New Market-risk Regulations. Article, VOX. <www.voxeu.org/article/new-market-risk-regulations>.
Daníelsson, J., Jorgensen, B.N., Samorodnitsky, G., Sarma, M., de Vries, C.G., 2013. Fat tails, VaR and subadditivity. J. Econometr. 172 (2), 283–291.
Danielsson, J., Zhou, C., 2016. Why Risk is so Hard to Measure. De Nederlandsche Bank Working Paper No. 494. Available at SSRN: <https://ssrn.com/abstract=2597563>.
Du, Z., Escanciano, J.C., 2016. Backtesting expected shortfall: accounting for tail risk. Manage. Sci. 63 (4), 940–958.
Embrechts, P., McNeil, A., Straumann, D., 2002. Correlation and dependence in risk management: properties and pitfalls. Risk Manage.: Value Risk Beyond, 176223.
Emmer, S., Kratz, M., Tasche, D., 2015. What is the best risk measure in practice? A comparison of standard measures. J. Risk 18 (2), 31–60.
Fama, E.F., 1965. The behavior of stock-market prices. J. Bus. 38 (1), 34–105.
Ferri, G., Pesic, V., 2017. Bank regulatory arbitrage via risk weighted assets dispersion. J. Financ. Stab. 33, 331–345.
Frey, R., McNeil, A.J., 2002. Var and expected shortfall in portfolios of dependent credit risks: conceptual and practical insights. J. Bank. Finan. 26 (7), 1317–1334.
Gneiting, T., 2011. Making and evaluating point forecasts. J. Am. Stat. Assoc. 106 (494), 746–762.
Gordy, M.B., 2003. A risk-factor model foundation for ratings-based bank capital rules. J. Financ. Intermed. 12 (3), 199–232.
Guegan, D., Hassani, B.K., 2018. More accurate measurement for enhanced controls: VaR vs ES?. J. Int. Financ. Markets Inst. Money 54, 152–165.
Haan, L.F.M., 1975. On regular variation and its application to the weak convergence of sample extremes. MC Tracts 32, 1–124.
Hibbeln, M., 2010. Risk Management in Credit Portfolios: Concentration Risk and Basel II. Springer Science & Business Media.
Jansen, D.W., De Vries, C.G., 1991. On the frequency of large stock returns: putting booms and busts into perspective. Rev. Econ. Stat., 18–24
Kerkhof, J., Melenberg, B., 2004. Backtesting for risk-based regulatory capital. J. Bank. Finan. 28 (8), 1845–1865.
Kinateder, H., 2016. Basel II versus III – a comparative assessment of minimum capital requirements for internal model approaches. J. Risk 18 (25–45).
Koch-Medina, P., Munari, C., 2016. Unexpected shortfalls of Expected Shortfall: Extreme default profiles and regulary arbitrage. J. Bank. Finan. 62, 141–151.
Mandelbrot, B., 1963. The variation of certain speculative prices. J. Bus. 36 (4), 394–419.
Mariathasan, M., Merrouche, O., 2014. The manipulation of basel risk-weights. J. Financ. Intermed. 23 (3), 300–321.
R Core Team, 2014. R: A Language and Environment for Statistical Computing. <http://www.R-project.org/>.
Vasicek, O., 2002. The distribution of loan portfolio value. Risk 15 (12), 160–162.
Yamai, Y., Yoshiba, T., 2002a. Comparative analyses of expected shortfall and value-at-risk: their estimation error, decomposition, and optimization. Monet. Econ. Stud. 20 (1), 87–121.
Yamai, Y., Yoshiba, T., 2002b. On the validity of value-at-risk: comparative analyses with expected shortfall. Monet. Econ. Stud. 20 (1), 57–85.
Yamai, Y., Yoshiba, T., 2005. Value-at-risk versus expected shortfall: a practical perspective. J. Bank. Finan. 29 (4), 997–1015.

Paper II

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

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

Kjartan Kloster Osmundsen[†1], Tore Selland Kleppe[1], and Atle Oglend[2]

[1]Department of Mathematics and Physics, University of Stavanger, Norway
[2]Department of Safety, Economics and Planning, University of Stavanger, Norway

24th August 2018

Abstract

This paper 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.

Keywords: Markov-switching, Marginalized likelihood, Hamiltonian Monte Carlo

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 like speech recognition (Baker, 1975), image analysis (Yamato et al., 1992) and network security (Scott, 2001). In economics, MS models have also been widely applied as parsimonious specifications of non-linear time series dynamics since the seminal paper 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 et al. (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 et al., 2006; Kanamura and Ōhashi, 2008) and oil and natural gas

*We would like to express gratitude to an anonymous referee for many useful comments that considerably improved the paper.
†Corresponding author. Email: kjartan.osmundsen@gmail.com

pricing (Brigida, 2014; Asche et al., 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 paper, 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 in the estimation of the models. Related to this paper, Sims et al. (2008) discuss methods for inference in large multiple-equation MS models. Also, Lanne et al. (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 paper 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 analogue to the currently popular pseudo-marginal methods (see e.g. Andrieu et al., 2010) for state-space models and other models where latent states take continuous values.

The paper 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 paper is organized as follows. Section 2 describes the MS-VAR model and introduces the notation used throughout the paper. 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 et al. (2017) to a fully flexible bivariate MS-VAR. Final conclusions are given in Section 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

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

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

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

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 $\boldsymbol{\Theta} = \Big(\boldsymbol{Q}, \boldsymbol{\phi}_{(1)}, \boldsymbol{\phi}_{(2)}, \ldots, \boldsymbol{\phi}_{(m)}, \boldsymbol{\mu}_{(1)}, \boldsymbol{\mu}_{(2)}, \ldots, \boldsymbol{\mu}_{(m)}, \boldsymbol{\Sigma}_{(1)}, \boldsymbol{\Sigma}_{(2)}, \ldots, \boldsymbol{\Sigma}_{(m)}\Big)$, the likelihood function, conditional on $\boldsymbol{Y}_1$, is given by

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

where $\boldsymbol{Y}_{1:n} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$, $\mathbf{1}^\top$ 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 $\boldsymbol{\Theta}$

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 paper 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, Section 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 et al., 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 paper 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

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

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: $\boldsymbol{\phi}_{(1)}, \boldsymbol{\phi}_{(2)}, \ldots, \boldsymbol{\phi}_{(m)}, \boldsymbol{\mu}_{(1)}, \boldsymbol{\mu}_{(2)}, \ldots, \boldsymbol{\mu}_{(m)}, \boldsymbol{\Sigma}_{(1)}, \boldsymbol{\Sigma}_{(2)}, \ldots, \boldsymbol{\Sigma}_{(m)} | \boldsymbol{Y}_{1:n}, \boldsymbol{S}_{2:n}$. For an unrestricted model (1) with conjugate priors (Gaussian for $\boldsymbol{\phi}_{(S)}$ 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):

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

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 $p(S_{n-t}|\boldsymbol{Y}_{1:n-t}), t \in (2, 3, \ldots, n-1)$ 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 Section 2.2. The marginalized approach is implemented using Stan (Stan Development Team, 2016); a programming 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 Section 3.3, the mean structure of (1) is shared by all regimes. Subsequently, the (weighted) regressions required to sample $\boldsymbol{\phi}, \boldsymbol{\mu}$ 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 et al. (2005). If identical priors are chosen for the parameters belonging to each state, the posterior likelihood is invariant to relabelling 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 $\mu_{(1),1} < \mu_{(2),1} \cdots < \mu_{(m),1}$. 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 et al., 2005). In the Bayesian context, constraints that ignore the geometry of the posterior does generally not induce a unique labelling (Frühwirth-Schnatter, 2001), so the constraints should

be chosen carefully. There has been some scepticism in the literature regarding the effects that identifiability constraints may have on MCMC samplers; see, e.g., Stephens (1997); Celeux et al. (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 et al. (2010), namely a restricted three-dimensional, two-state case of (1), with an additional three autoregressive lags:

$$\boldsymbol{Y}_t = \boldsymbol{\phi}_1 \boldsymbol{Y}_{t-1} + \boldsymbol{\phi}_2 \boldsymbol{Y}_{t-2} + \boldsymbol{\phi}_3 \boldsymbol{Y}_{t-3} + \boldsymbol{\phi}_4 \boldsymbol{Y}_{t-4} + \boldsymbol{\mu} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{(S_t)}\right), \qquad \boldsymbol{Y}_t \in \mathbb{R}^3, \quad (4)$$

  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 et al., 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 Section 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

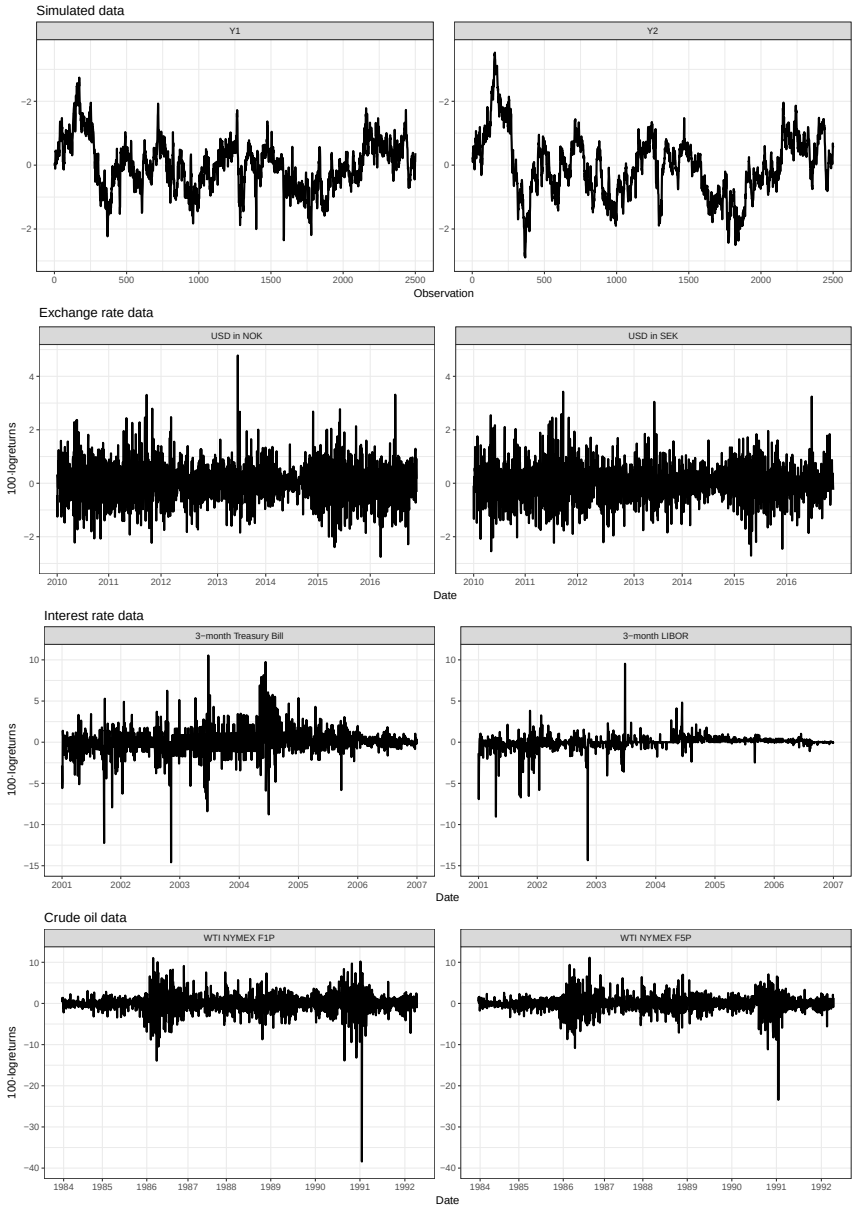

Figure 1: Plots of a simulated data set and three real data sets, all having two components. For the real data sets, log returns of the raw data are used, scaled by 100.

	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

Table 1: The parameter estimates' effective sample size per second, for simulated data sets with three different observation sizes. The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as a single chain.

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 Section 2.3, and can be found in Table 9 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 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 modelling 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

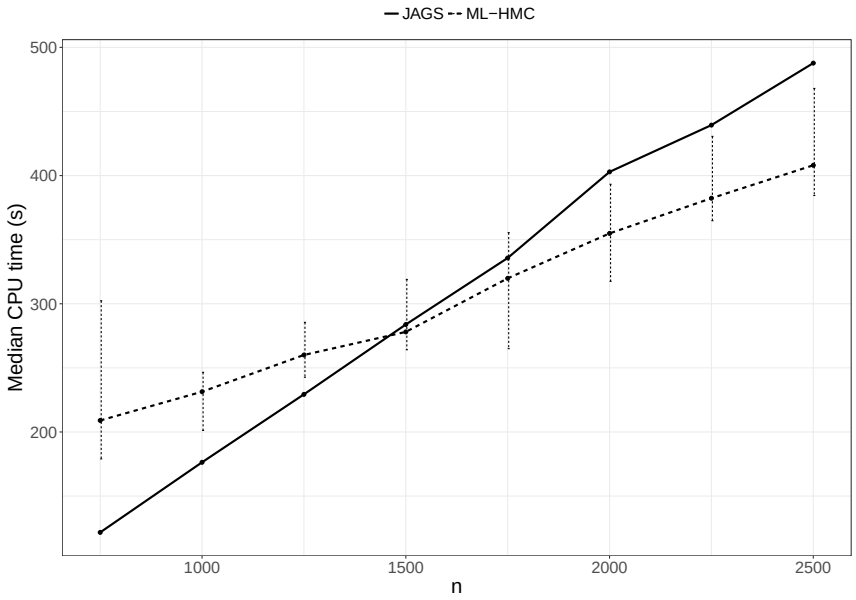

Figure 2: ML-HMC and JAGS median run time for 8 chains of 1000 (500 warm-up) iterations each, obtained from simulated data sets with different observation sizes, n. Vertical lines span from minimum to maximum run time.

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 10. 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

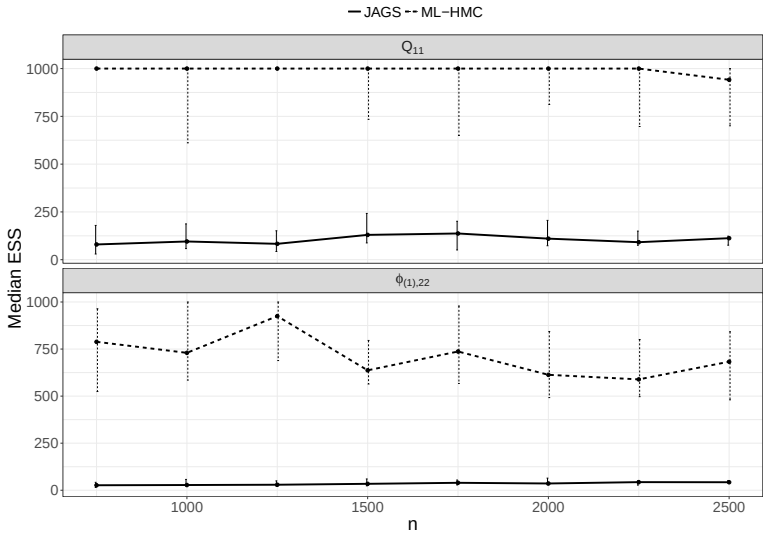

Figure 3: Median effective sample size for the estimates of parameters $\boldsymbol{Q}_{11}$ and $\boldsymbol{\phi}_{(1),22}$ for 8 chains of 1000 (500 warm-up) iterations each, obtained from simulated data sets of different observation sizes, n, using both ML-HMC and JAGS. Vertical lines span from minimum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

	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

Table 2: The parameter estimates' effective sample size per second, for three real data sets. The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as a single chain.

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 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.

3.3 Restricted model

Following Lanne et al. (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 Section 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 Section 2.2 to estimate parameters for a model of the joint dynamics of UK natural gas and Brent oil prices, extending the modelling approach in Asche et al. (2017). The model used by Asche et al. (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 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 modelled separately from the oil prices' influence on natural gas

	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

Table 3: The parameter estimates and their effective sample size per second, for the US macro data set. The results are calculated from a collection of 8 chains with 1000 (excluding warm-up) iterations each, treated as a single chain.

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)

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

prices. Asche et al. (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:

$$\begin{bmatrix} \Delta \boldsymbol{Y}_{t,1} \\ \Delta \boldsymbol{Y}_{t,2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\phi}_{(S_t),11} & \boldsymbol{\phi}_{(S_t),12} \\ \boldsymbol{\phi}_{(S_t),21} & \boldsymbol{\phi}_{(S_t),22} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{Y}_{t-1,1} \\ \Delta \boldsymbol{Y}_{t-1,2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\mu}_{(S_t),1} \\ \boldsymbol{\mu}_{(S_t),2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\alpha}_{(S_t),1} \\ \boldsymbol{\alpha}_{(S_t),2} \end{bmatrix} z_{t-1} + \begin{bmatrix} \boldsymbol{\epsilon}_{t,1} \\ \boldsymbol{\epsilon}_{t,2} \end{bmatrix}, \tag{5}$$
$$\begin{bmatrix} \boldsymbol{\epsilon}_{t,1} \\ \boldsymbol{\epsilon}_{t,2} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{(S_t)}\right), \qquad \boldsymbol{Y}_t \in \mathbb{R}^2, \qquad S_t \in (1, 2, \ldots, m), \qquad t \in (2, 3, 4, \ldots, n),$$

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 Section 3.2.

The parameters are estimated using the log prices for natural gas and oil, thus modelling 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 et al. (2017) find that state 2 has eight times higher variance than state 1, meaning that it is reasonable using the identifiability constraint from Section 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,

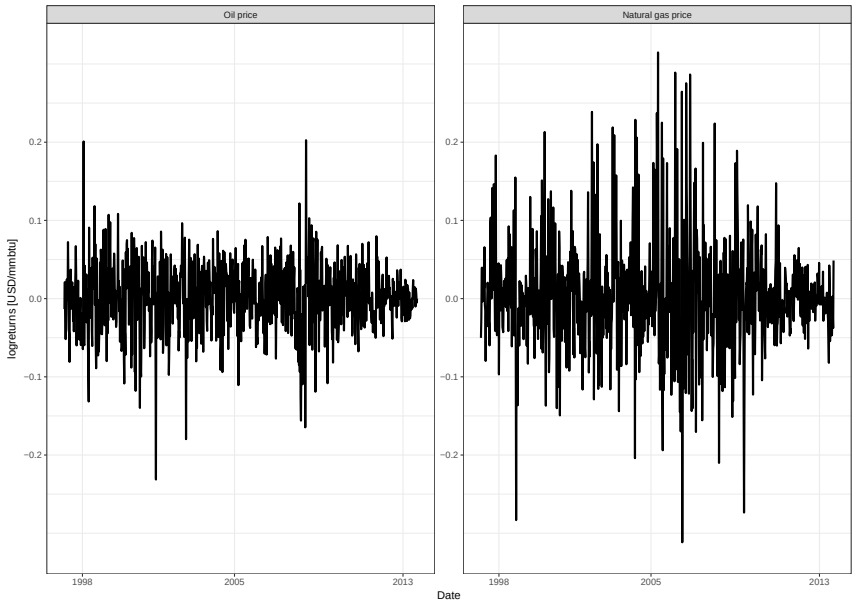

Figure 4: Plot of the log returns of oil and natural gas prices. The data set consists of 904 weekly observations, from March 1997 to April 2014.

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

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

	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)

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

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	4.8	4.5	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

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

(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 et al. (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

	Asche et al. (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

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

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 modelling 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.

The empirical part of this paper focuses on a two-dimensional, two-state model, but the methodology extends easily to more complex models. Larger models will favour 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 et al., 2016). Application of marginalized approach and Stan to such situations holds scope for future research.

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.

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.

Ang, A. and G. Bekaert (2002). International asset allocation with regime shifts. *The Review of Financial Studies 15*(4), 1137–1187.

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).

Baker, J. (1975). The dragon system–an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing 23*(1), 24–29.

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.

Bansal, R. and H. Zhou (2002). Term structure of interest rates with regime shifts. *The Journal of Finance 57*(5), 1997–2043.

Bianchi, F. (2016). Methods for measuring expectations and uncertainty in markov-switching models. *Journal of Econometrics 190*(1), 79–99.

Bloom, N. (2009). The impact of uncertainty shocks. *econometrica 77*(3), 623–685.

Brigida, M. (2014). The switching relationship between natural gas and crude oil prices. *Energy Economics 43*, 48–55.

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), 957–970.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics 75*(1), 79–97.

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.

Garcia, R. and P. Perron (1996). An analysis of the real interest rate under regime shifts. *Review of Economics and Statistics 78*, 111–125.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science*, 473–483.

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.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics 42*(1), 27–62.

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*.

Hoffman, M. D. and A. Gelman (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

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

Kanamura, T. and K. Ōhashi (2008). On transition probabilities of regime switching in electricity prices. *Energy Economics 30*(3), 1158–1172.

Kim, C.-J. (1994). Dynamic linear models with markov-switching. *Journal of Econometrics 60*(1-2), 1–22.

Krolzig, H.-M. (1997). *Markov-Switching Vector Autoregressive Model - Modelling, Statistical Inference, and Application to Business Cycle Analysis*. Springer.

Kshirsagar, A. M. (1959). Bartlett decomposition and Wishart distribution. *The Annals of Mathematical Statistics 30*(1), 239–241.

Lanne, M., H. Lütkepohl, and K. Maciejowska (2010). Structural vector autoregressions with markov switching. *Journal of Economic Dynamics and Control 34*(2), 121–131.

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.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2*, 113–162.

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*. 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), 133–147.

Scott, S. L. (2001). Detecting network intrusion using a markov modulated nonhomogeneous poisson process. *Submitted to the Journal of the American Statistical Association*.

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive Computing in the 21st Century. *Journal of the American Statistical Association*.

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), 255–274.

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

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.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory 13*(2), 260–269.

Waggoner, D. F., H. Wu, and T. Zha (2016). Striated metropolis–hastings sampler for high-dimensional models. *Journal of Econometrics 192*(2), 406–420.

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

A Appendix

	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)

Table 9: Parameter estimates, with their respective standard deviation in parenthesis, for simulated data sets with three different observation sizes. The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as a single chain.

	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)

Table 10: Parameter estimates, with their respective standard deviation in parenthesis, for three real data sets. The results are calculated from a collection of 8 chains with 1000 (500 warm-up) iterations each, treated as a single chain.

Paper III

Importance Sampling-based Transport Map Hamiltonian Monte Carlo for Bayesian Hierarchical Models

Importance Sampling-based Transport Map Hamiltonian Monte Carlo for Bayesian Hierarchical Models

Kjartan Kloster Osmundsen*[1], Tore Selland Kleppe[1], and Roman Liesenfeld[2]

[1]Department of Mathematics and Physics, University of Stavanger, Norway
[2]Institute of Econometrics and Statistics, University of Cologne,Germany

December 11, 2019

Abstract

We propose an importance sampling (IS)-based transport map Hamiltonian Monte Carlo procedure for performing full Bayesian analysis in general nonlinear high-dimensional hierarchical models. Using IS techniques to construct a transport map, the proposed method transforms the typically highly challenging target distribution of a hierarchical model into a target which is easily sampled using standard Hamiltonian Monte Carlo. Conventional applications of high-dimensional IS, where infinite variance of IS weights can be a serious problem, require computationally costly high-fidelity IS distributions. An appealing property of our method is that the IS distributions employed can be of rather low fidelity, making it computationally cheap. We illustrate our algorithm in applications to challenging dynamic state-space models, where it exhibits very high simulation efficiency compared to relevant benchmarks, even for variants of the proposed method implemented using a few dozen lines of code in the Stan statistical software.

Keywords: Hamiltonian Monte Carlo; Importance Sampling; Transport Map; Bayesian hierarchical models; State-space models; Stan

1 Introduction

Computational methods for Bayesian nonlinear/non-Gaussian hierarchical models is an active field of research, and advances in such computational methods allow researchers to build and fit progressively more complex models. Existing Markov chain Monte Carlo (MCMC) methods for such models fall broadly into four categories. Firstly, Gibbs sampling is widely used, in part due to its simple implementation (see e.g. Robert and Casella, 2004). However, a naive implementation updating latent variables in one block and

*Corresponding author. Email: kjartan.osmundsen@gmail.com

model parameters in another block can suffer from a very slow exploration (see e.g. Jacquier et al., 1994) of the target distribution if this joint distribution implies a strong, typically nonlinear dependence structure of the variables in the two blocks. Secondly, methods that update latent variables and parameters jointly avoid the nonlinear dependence problem of Gibbs sampling. One such approach for joint updates is to use Riemann manifold Hamiltonian Monte Carlo (RMHMC) methods (see e.g. Girolami and Calderhead, 2011; Zhang and Sutton, 2014; Kleppe, 2018). However, they critically require update proposals which are properly aligned with the (typically rather variable) local geometry of the target, the generation of which can be computationally demanding for complex high-dimensional joint posteriors of the parameters and latent variables.

The third category is pseudo-marginal methods (see e.g. Andrieu et al., 2010; Pitt et al., 2012, and references therein), which bypasses the problematic parameters and latent variables dependency by targeting directly the marginal posterior of the parameters. Pseudo-marginal methods require, however, a low variance, unbiased Monte Carlo (MC) estimate of said posterior, which can often be extremely computationally demanding for high-dimensional models (see e.g. Flury and Shephard, 2011). Moreover, for models with many parameters, it can be difficult to select an efficient proposal distribution for updating the parameters if the MC estimates for the marginal posterior are noisy and/or contain many discontinuities, which is typically the case if the MC estimator is implemented using particle filtering techniques.

Finally, the fourth category is transport map/dynamic rescaling methods (see e.g. Parno and Marzouk, 2018; Hoffman et al., 2019), which rely on introducing a modified parameterization related to the original parameterization via the nonlinear transport map. The transport map is chosen so that the target distribution in the modified parameterization is more well behaved and allows MCMC sampling using standard techniques. The Dynamically rescaled Hamiltonian Monte Carlo (DRHMC) approach of Kleppe (2019) involves a recipe for constructing transport maps suitable for a large class of Bayesian hierarchical models, and where the models are fitted using the (fixed scale) No-U-Turn Sampler (NUTS) Hamiltonian Monte Carlo (HMC) algorithm (Hoffman and Gelman, 2014) implemented in Stan (Stan Development Team, 2019b).

The present paper also considers a transport map approach for Bayesian hierarchical models, and sample from the modified target using HMC methods. However, the strategy for constructing the transport map considered here is different from that of DRHMC. Specifically, DRHMC involves deriving the transport maps from the model specification itself, and in particular it requires the availability of closed-form expressions for certain precision- and Fisher information matrices associated with the model. Moreover, the DRHMC approach is in practice limited to models containing only a certain class of nonlinearities which lead to so-called constant information parameterizations.

Here, on the other hand, we consider transport maps derived from well-known importance sampling

(IS) methods for the latent variables only. This approach relies only on the ability to evaluate the log-target density (and potentially it's derivatives) pointwise, and therefore bypasses the substantial analytic tractability requirement of DRHMC. The proposed approach is consequently more automatic in nature, and in particular applicable to a wider range of nonlinear models than DRHMC. Still, some analytical insight into the model is beneficial in terms of computational speed when choosing the initial iterates of the involved iterative processes.

A fortunate property of the proposed methodology, relative to conventional applications of high-dimensional importance sampling (see e.g. Koopman et al., 2009), is that the importance densities applied within the present framework may be of relatively low fidelity as long as they reflect the location and scale of the distribution of the latent state conditioned both on data and parameters. Since parameters and latent variables are updated simultaneously, the slow exploration of the target associated with Gibbs sampling is avoided. Moreover, being transport map-based, rather than say RMHMC-based, the proposed methodology allows for the application of standard HMC and in particular can be implemented with minimal effort in Stan.

The application of IS methods to construct transport maps also allows the proposed methodology to be interpreted as a pseudo-marginal method, namely a special case (with simulation sample size $n = 1$) of the pseudo-marginal HMC method of Lindsten and Doucet (2016). However, our focus on models with high-dimensional latent variables generally precludes the application of 'brute force' IS estimators that do not reflect information from the data (see, e.g., Danielsson, 1994). This is the case even for increased simulation sample size of the IS estimate, as is possible in the general setup of Lindsten and Doucet (2016).

The rest of the paper is laid out as follows: Section 2 provides some background and Section 3 introduces IS-based transport maps. Section 4 discusses specific choices of IS-based transport maps and Section 5 provides a simulation experiment where the fidelity vs computational cost tradeoff of the different transport maps is explored numerically. Finally, Section 6 presents a realistic application and Section 7 provides some discussion. The paper is accompanied by supplementary material giving further details in several regards, and the code used for the computations is available at `https://github.com/kjartako/TMHMC`.

2 Background

This section outlines some background on HMC and why the application of HMC in default formulations of hierarchical models is problematic. In what follows, we use $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the probability density function of a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random vector evaluated at $\mathbf{x}$, while $\nabla_{\mathbf{z}}$ and $\nabla^2_{\mathbf{z}}$ are used, respectively, for the gradient/Jacobian and Hessian operator with respect to the vector $\mathbf{z}$.

2.1 HMC

Over the past decade, HMC introduced by Duane et al. (1987) has been extensively used as a general-purpose MCMC method, often applied for simulating from posterior distributions arising in Bayesian models (Neal, 2011). HMC offers the advantage of producing close to perfectly mixing MCMC chains by using the dynamics of a synthetic Hamiltonian system as proposal mechanism. The popular Bayesian modelling software Stan (Stan Development Team, 2019b) is an easy to use HMC implementation based on the NUTS HMC algorithm of Hoffman and Gelman (2014).

Suppose one seeks to sample from an analytically intractable target distribution with density kernel $\tilde{\pi}(\mathbf{q}),\ \mathbf{q} \in \Omega \subseteq \mathbb{R}^s$. To this end, HMC takes the variable of interest $\mathbf{q}$ as the 'position coordinate' of a Hamiltonian system, which is complemented by an (artificial) 'momentum variable' $\mathbf{p} \in \mathbb{R}^s$. The corresponding Hamiltonian function specifying the total energy of the dynamical system is given by

$$H(\mathbf{q}, \mathbf{p}) = -\log \tilde{\pi}(\mathbf{q}) + \frac{1}{2}\mathbf{p}'\mathbf{M}^{-1}\mathbf{p}, \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{s \times s}$ is a symmetric, positive definite 'mass matrix' representing an HMC tuning parameter. For near-Gaussian target distributions, for instance, setting $\mathbf{M}$ close to the precision matrix of the target ensures the best performance. The law of motions under the dynamic system specified by the Hamiltonian H is determined by Hamilton's equations given by

$$\frac{d}{dt}\mathbf{p}(t) = -\nabla_{\mathbf{q}} H\left(\mathbf{q}(t), \mathbf{p}(t)\right) = \nabla_{\mathbf{q}} \log \tilde{\pi}(\mathbf{q}), \qquad \frac{d}{dt}\mathbf{q}(t) = \nabla_{\mathbf{p}} H\left(\mathbf{q}(t), \mathbf{p}(t)\right) = \mathbf{M}^{-1}\mathbf{p}. \tag{2}$$

It can be shown that the dynamics associated with Hamilton's equations preserves both the Hamiltonian (i.e. $dH\left(\mathbf{q}(t), \mathbf{p}(t)\right)/dt = 0$) and the Boltzmann distribution $\pi(\mathbf{q}, \mathbf{p}) \propto \exp\{-H(\mathbf{q}, \mathbf{p})\} \propto \tilde{\pi}(\mathbf{q})\, \mathcal{N}(\mathbf{p}|\mathbf{0}_s, \mathbf{M})$, in the sense that if $[\mathbf{q}(t), \mathbf{p}(t)] \sim \pi(\mathbf{q}, \mathbf{p})$, then $[\mathbf{q}(t+\tau), \mathbf{p}(t+\tau)] \sim \pi(\mathbf{q}, \mathbf{p})$ for any (scalar) time increment τ. Based on the latter property, a valid MCMC scheme for generating $\{\mathbf{q}^{(k)}\}_k \sim \tilde{\pi}(\mathbf{q})$ would be to alternate between the following two steps: (i) Sample a new momentum $\mathbf{p}^{(k)} \sim N(\mathbf{0}_s, \mathbf{M})$ from the $\mathbf{p}$-marginal of the Boltzmann distribution; and (ii) use the Hamiltonian's equations (2) to propagate $[\mathbf{q}(0), \mathbf{p}(0)] = [\mathbf{q}^{(k)}, \mathbf{p}^{(k)}]$ for some increment τ to obtain $[\mathbf{q}(\tau), \mathbf{p}(\tau)] = [\mathbf{q}^{(k+1)}, \mathbf{p}^*]$ and discard $\mathbf{p}^*$. However, for all but very simple scenarios (like those with a Gaussian target $\tilde{\pi}(\mathbf{q})$) the transition dynamics according to (2) does not admit closed-form solution, in which case it is necessary to rely on numerical integrators for an approximative solution. Provided that the numerical integrator used for that purpose is symplectic, the numerical approximation error can be exactly corrected by introducing an accept-reject (AR) step, which uses the Hamiltonian to compare the total energy of the new proposal for the pair $(\mathbf{q}, \mathbf{p})$ with that of the old pair inherited from the

previous MCMC step (see, e.g., Neal, 2011). More specifically each iteration of the HMC algorithm involves the following steps

- Refresh the momentum $\mathbf{p}^{(k)} \sim N(\mathbf{0}_s, \mathbf{M})$.
- Propagate approximately the dynamics (2) from $(\mathbf{q}(0), \mathbf{p}(0)) = (\mathbf{q}^{(k)}, \mathbf{p}^{(k)})$ to obtain $(\mathbf{q}^*, \mathbf{p}^*) \approx (\mathbf{q}(L\varepsilon), \mathbf{p}(L\varepsilon))$ using L symplectic integrator steps with time-step size ε.
- Set $\mathbf{q}^{(k+1)} = \mathbf{q}^*$ with probability $\min(1, \exp(H(\mathbf{q}^{(k)}, \mathbf{p}^{(k)}) - H(\mathbf{q}^*, \mathbf{p}^*))$ and $\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)}$ with remaining probability.

The most commonly used symplectic integrator is the Störmer-Verlet or leapfrog integrator (see, e.g., Leimkuhler and Reich, 2004; Neal, 2011). When implementing numerical integrators with AR-corrections it is critical that the selection of the step size accounts for the inherent trade-off between the computing time required for generating AR proposals and their quality reflected by their corresponding acceptance rates. $(\mathbf{q}, \mathbf{p})$-proposals generated by using small (big) step sizes tend to be computationally expensive (cheap) but imply a high (low) level of energy preservation and thus high (low) acceptance rates. Finally, the energy preservation properties of the symplectic integrator for any given step size critically relies on the nature of the target distribution. It is taken as a rule of thumb for the remainder of the text that high-dimensional, highly non-Gaussian targets typically require small step sizes and many steps, whereas high-dimensional near-Gaussian targets can be sampled efficiently with rather large step sizes and few steps.

2.2 Hierarchical models and HMC

Consider a stochastic model for a collection of observed data $\mathbf{y}$ involving a collection of latent variables $\mathbf{x}$ and a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ with prior density $p(\boldsymbol{\theta})$. The conditional likelihood for observations $\mathbf{y}$ given a value of the latent variable $\mathbf{x} \in \mathbb{R}^D$ is denoted by $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ and the prior for $\mathbf{x}$ by $p(\mathbf{x}|\boldsymbol{\theta})$. This latent variable model is assumed to be nonlinear and/or non-Gaussian so that both the joint posterior for $(\mathbf{x}, \boldsymbol{\theta})$ as well as the marginal posterior for $\boldsymbol{\theta}$ are analytically intractable.

The joint posterior for $(\mathbf{x}, \boldsymbol{\theta})$ under such a latent variable model, given by $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, can have a complex dependence structure. In particular, when the scale of $\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}$ varies substantially as a function of $\boldsymbol{\theta}$ in the typical range of $p(\boldsymbol{\theta}|\mathbf{y})$, the joint posterior will be "funnel-shaped" (see Kleppe, 2019, Figure 1 for an illustration). In this case, the HMC algorithm, as described in Section 2.1, for $\mathbf{q} = (\mathbf{x}^T, \boldsymbol{\theta}^T)^T$ must be tuned for the most extremely scaled parts of the target distribution to ensure exploration of the complete target distribution. This, in turn lead to a computationally wasteful exploration of the more moderately scaled parts of the target, as the tuning parameters cannot themselves depend on $\mathbf{q}$ (under

regular HMC). In addition, automated tuning of integrator step sizes (and mass matrices) crucially relies on the most extremely scaled parts being visited during the initial tuning phase. If not, they may not be explored at all.

3 Transport maps based on IS densities

To counteract such undesired extreme tuning, while avoiding computationally costly $\mathbf{q}$-dependent tuning such as RMHMC, the approach taken here involves "preconditioning" the original target so that the resulting modified target is close to Gaussian and thus suitable for statically tuned HMC. Such preconditioning with the aim of producing more tractable target distributions for MCMC methods have a long tradition, and prominent examples are the affine re-parameterizations common for Gibbs sampling applied to regression models (see, e.g., Gelman et al., 2014, Chapter 12). More recent approaches with such ends involve semi-parametric transport map approach of Parno and Marzouk (2018), and, neural transport as described by Hoffman et al. (2019). The approach taken here share many similarities with the dynamically rescaled HMC approach of Kleppe (2019), but the strategy for constructing the transport map considered here is very different and is applicable to more general models.

In a nutshell, a transport map, say T, is a smooth bijective mapping relating the original parameterization $\mathbf{q} \sim \pi_{\mathbf{q}}(\mathbf{q})$ and some modified parameterization $\mathbf{q}'$ via $\mathbf{q} = T(\mathbf{q}')$. If $\mathbf{q}'$ is some random draw $\sim \pi_{\mathbf{q}'}(\mathbf{q}') = \pi_{\mathbf{q}}(T(\mathbf{q}'))|\nabla_{\mathbf{q}'}T(\mathbf{q}')|$, then a draw distributed according to $\pi_{\mathbf{q}}$ is achieved by simply applying the transport map to $\mathbf{q}'$. The aim of introducing this construction, is that T can be chosen so that $\pi_{\mathbf{q}'}$ is loosely speaking "more suitable for MCMC sampling". In practice, this rather vague aim is replaced by making $\pi_{\mathbf{q}'}$ close to a Gaussian distribution with independent components, which can be sampled very efficiently using HMC.

3.1 Transport maps for Bayesian hierarchical models

In the current situation involving a Bayesian hierarchical model, a transport map T that is non-trivial for the latent variables only,

$$\mathbf{q} = \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{x} \end{bmatrix} = T(\mathbf{q}') = \begin{bmatrix} \boldsymbol{\theta} \\ \gamma_{\boldsymbol{\theta}}(\mathbf{u}) \end{bmatrix}, \ \mathbf{q}' = \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{u} \end{bmatrix},$$

is considered. The transport map specific to the latent variables, $\gamma_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}^D$ is assumed to be a smooth bijective mapping for each $\boldsymbol{\theta}$. As we have $\nabla_{\mathbf{u}}\boldsymbol{\theta} = \mathbf{0}$ in the above transport map, it follows that $|\nabla_{\mathbf{q}'}T(\mathbf{q}')| = |\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})|$, and thus the modified target distribution has the form:

$$\tilde{\pi}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) \propto |\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})| p(\boldsymbol{\theta}) \left[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\right]_{\mathbf{x}=\gamma_{\boldsymbol{\theta}}(\mathbf{u})}. \tag{3}$$

Notice in particular that the original parameterization of the latent variables is computed in each evaluation of (3), and thus obtaining MCMC samples in the $(\boldsymbol{\theta}, \mathbf{x}) = (\boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}}(\mathbf{u}))$ parameterization comes at no additional cost when MCMC samples targeting (3) are available.

Further, let $m(\mathbf{x}|\boldsymbol{\theta})$ denote the density of $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$ *when* $\mathbf{u} \sim N(\mathbf{0}_D, \mathbf{I}_D)$. In particular, $m(\mathbf{x}|\boldsymbol{\theta})$ is implicitly related to the underlying standard Gaussian distribution via the change of variable formula: $\mathcal{N}(\mathbf{u}|\mathbf{0}_D, \mathbf{I}_D) = |\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})| \left[m(\mathbf{x}|\boldsymbol{\theta})\right]_{\mathbf{x}=\gamma_{\boldsymbol{\theta}}(\mathbf{u})}$. Consequently, eliminating the Jacobian determinant in (3) results in

$$\tilde{\pi}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) \propto \mathcal{N}(\mathbf{u}|\mathbf{0}_D, \mathbf{I}_D) p(\boldsymbol{\theta}) \omega_{\boldsymbol{\theta}}(\mathbf{u}), \; \omega_{\boldsymbol{\theta}}(\mathbf{u}) = \left[\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})}{m(\mathbf{x}|\boldsymbol{\theta})}\right]_{\mathbf{x}=\gamma_{\boldsymbol{\theta}}(\mathbf{u})}. \tag{4}$$

Representation (4) reveal that if $m(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ (i.e. $\gamma_{\boldsymbol{\theta}}(\mathbf{u}) \sim \mathbf{x}|\mathbf{y}, \boldsymbol{\theta}$), the parameters and latent variables exactly "decouples" and (3) and (4) reduces to $\mathcal{N}(\mathbf{u}|\mathbf{0}_D, \mathbf{I}_D) p(\boldsymbol{\theta}|\mathbf{y})$ (see also Lindsten and Doucet, 2016, for a similar discussion). Such a situation will be well suited for HMC sampling (provided of course that the marginal likelihood $p(\boldsymbol{\theta}|\mathbf{y})$ is reasonably well-behaved). Of course, such an ideal situation is in practice unattainable when the model in question is nonlinear/non-Gaussian as neither $p(\boldsymbol{\theta}|\mathbf{y})$ nor $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ will have analytical forms. The strategy pursued here is therefore to take $m(\mathbf{x}|\boldsymbol{\theta})$ as an approximation to $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ in order to obtain an approximate decoupling effect, i.e. so that $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ is fairly flat across the region where $\mathcal{N}(\mathbf{u}|\mathbf{0}_D, \mathbf{I}_D)$ has significant probability mass.

3.2 Relation to importance sampling and pseudo-marginal methods

The $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ of (4) is recognized to be an importance weight targeting the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ (i.e. $E_{\mathbf{u}}(\omega_{\boldsymbol{\theta}}(\mathbf{u})) = p(\mathbf{y}|\boldsymbol{\theta})$) when $\mathbf{u} \sim N(\mathbf{0}_D, \mathbf{I}_D)$. This observation is important for at least three reasons. Firstly, it is clear that the large literature on importance sampling- and similar methods for hierarchical models (among many others, Shephard and Pitt, 1997; Richard and Zhang, 2007; Rue et al., 2009; Durbin and Koopman, 2012) may be leveraged to suggest suitable choices for importance density $m(\mathbf{x}|\boldsymbol{\theta})$ or $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$. Specific choices considered here are discussed in more detail in Section 4.

Secondly, as discussed, e.g., in Koopman et al. (2009), importance sampling-based likelihood estimates such as $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ may have infinite variance and thus become unreliable, in particular in high-dimensional applications. This occurs when the tails of $m(\mathbf{x}|\boldsymbol{\theta})$ are thinner than those of the target distribution $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})$, making $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ unbounded as a function of $\mathbf{u}$. However, under the modified target (4) the likelihood estimate is combined with the thin-tailed standard normal distribution in $\mathbf{u}$, which counteracts the potential unboundedness of the IS weight in the $\mathbf{u}$-direction. This robustness with respect to the infinite-variance problem is also evident in the representation (3) of the target, which does not explicitly involve the importance sampling weight. Affine transport maps $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$, and consequently thin-tailed Gaussian

importance densities $m(\mathbf{x}|\boldsymbol{\theta})$, lead to the Jacobian determinant $|\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})|$ being constant with respect to $\mathbf{u}$. Consequently, in this case the tail behavior of (3) with respect to $\mathbf{u}$ will be the same as the tail behavior of $p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})$ in $\mathbf{x}$. Thus, the proposed methodology may be seen as a resolution of the infinite variance problems complicating the application of high-dimensional importance sampling.

Finally, the proposed methodology may be seen as a special case of the pseduo-marginal HMC (PM-HMC) method of Lindsten and Doucet (2016). PM-HMC relies on joint HMC sampling of a Monte Carlo estimate of the marginal likelihood and the random variables used to generate said estimate. Lindsten and Doucet (2016) find a similar decoupling effect by admitting their Monte Carlo estimate be based on $n \geq 1$ importance weights (at the cost of increasing the dimensionality of $\mathbf{u}$ in their counterpart to (4)), and are to a lesser degree reliant on choosing high-quality importance densities. In particular, Lindsten and Doucet (2016) use $m(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ in their illustrations, which for moderately dimensional and low-signal-to noise situations will produce a good decoupling effect for moderate n. However, in the present work we focus on high-dimensional applications where it is well known that such "brute force" importance sampling estimators can suffer from prohibitively large variances for any practical n (see, e.g., Danielsson, 1994), and thus focus rather on higher fidelity importance densities and $n = 1$.

Lindsten and Doucet (2016) also propose a symplectic integrator suitable for HMC applications with target distributions on the form (4) under the "close to decoupling" assumption. In the decoupling case $\mathbf{u} \mapsto \omega_{\boldsymbol{\theta}}(\mathbf{u}) \propto 1$, the integrator reduces to a standard leapfrog integrator in the dynamics of $\boldsymbol{\theta}$, whereas the dynamics of $\mathbf{u}$ (typically high-dimensional) are simulated exactly. This integrator will be referred to as the LD-integrator in the example applications and is detailed in the supplementary material, Section A.

4 Specific choices of $m(\mathbf{x}|\boldsymbol{\theta})$ and $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$

As alluded to above, taking $m(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ may in cases where data $\mathbf{y}$ are rather un-informative with respect to the latent variable $\mathbf{x}$ lead to satisfactory results (see e.g. Stan Development Team, 2019b, Section 2.5). However, as illustrated by e.g. Kleppe (2019), such procedures can lead to misleading MCMC results if data are more informative with respect to the latent variables. An even more challenging situation with $m(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ is when one or more elements of $\boldsymbol{\theta}$ determine how informative the data are with respect to the latent variables (e.g. σ when $y_i \sim N(x_i, \sigma^2)$), as this may still lead to a funnel-shaped target distribution. On the other hand, as illustrated by Kleppe (2019), rather crude transport maps reflecting only roughly the location and scale of $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ may lead to dramatic speedups, and the resolution of funnel-related problems. In the rest of this section, two families of strategies for locating transport maps are discussed. Both are well known in the context of importance sampling, and are typically applicable when $p(\mathbf{x}|\boldsymbol{\theta})$ is non-Gaussian.

4.1 $m(\mathbf{x}|\boldsymbol{\theta})$ and $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$ derived from approximate Laplace approximations

As explained e.g. in Rue et al. (2009), the Laplace approximation (also often referred to as the second order approximation) for integrating out latent variables relies on approximating $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ with a $N(\mathbf{h}_{\boldsymbol{\theta}},\mathbf{G}_{\boldsymbol{\theta}}^{-1})$ density, where

$$\begin{aligned}\mathbf{h}_{\boldsymbol{\theta}} &= \arg\max_{\mathbf{x}} \log\left[p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right],\\ \mathbf{G}_{\boldsymbol{\theta}} &= -\nabla_{\mathbf{x}}^2 \log\left[p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}}.\end{aligned}$$

Namely, the first and second order derivatives of $\log p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ at the mode are matched with the same derivatives of the approximating Gaussian log-density. Due to conditional independence assumptions often involved in modelling, the negative Hessian of $-\log p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ is typically sparse which, when exploited, can substantially speed up the associated Cholesky factorizations.

In the present situation, obtaining the exact mode $\mathbf{h}_{\boldsymbol{\theta}}$ is typically not desirable from a computational perspective. Rather, given an initial guesses for $\mathbf{h}_{\boldsymbol{\theta}}$ and $\mathbf{G}_{\boldsymbol{\theta}}$, say $\mathbf{h}_{\boldsymbol{\theta}}^{(0)}$ and $\mathbf{G}_{\boldsymbol{\theta}}^{(0)}$, a sequence of gradually more refined approximate solutions $\mathbf{h}_{\boldsymbol{\theta}}^{(k)}$ and $\mathbf{G}_{\boldsymbol{\theta}}^{(k)}$ are calculated via iterations of Newton's method for optimization or an approximation thereof (see supplementary material, Sections C and D for details specific to the models considered shortly).

Finally, for some fixed number of iterations, $K = 0, 1, 2, \ldots$, the transport map is taken to be

$$\gamma_{\boldsymbol{\theta}}(\mathbf{u}) = \mathbf{h}_{\boldsymbol{\theta}}^{(K)} + \left(\mathbf{L}_{\boldsymbol{\theta}}^{(K)}\right)^{-T}\mathbf{u}, \tag{5}$$

where $\mathbf{L}^{(K)}$ is the lower triangular Cholesky factor of $\mathbf{G}_{\boldsymbol{\theta}}^{(K)}$, so that $m(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{x}|\mathbf{h}_{\boldsymbol{\theta}}^{(K)}, \left[\mathbf{G}_{\boldsymbol{\theta}}^{(K)}\right]^{-1}\right)$. Notice in particular that the Jacobian determinant of $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$, required in representation (3) (or in the normalization constant of $m(\mathbf{x}|\boldsymbol{\theta})$ in (4)), takes a particularly simple form, namely $|\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})| = |\mathbf{L}_{\boldsymbol{\theta}}^{(K)}|^{-1}$, when applying the affine transport map (5). It should be noted that the applicability of the Laplace approximation relies critically on that $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ is unimodal and log-concave in a region around the mode that also contains $\mathbf{h}_{\boldsymbol{\theta}}^{(0)}$.

Choices of $\mathbf{h}_{\boldsymbol{\theta}}^{(0)}$, $\mathbf{G}_{\boldsymbol{\theta}}^{(0)}$ and the iteration over k are inherently model specific. However, for a rather general class of models, the initial guesses may be taken to be

$$\mathbf{G}_{\boldsymbol{\theta}}^{(0)} = \mathbf{G}_{\boldsymbol{\theta},\mathbf{x}} + \mathbf{G}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}} \tag{6}$$

$$\mathbf{h}_{\boldsymbol{\theta}}^{(0)} = \left(\mathbf{G}_{\boldsymbol{\theta}}^{(0)}\right)^{-1}(\mathbf{G}_{\boldsymbol{\theta},\mathbf{x}}\mathbf{h}_{\boldsymbol{\theta},\mathbf{x}} + \mathbf{G}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}}\mathbf{h}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}}), \tag{7}$$

where $\mathbf{h}_{\boldsymbol{\theta},\mathbf{x}}$ and $\mathbf{G}_{\boldsymbol{\theta},\mathbf{x}}$ are the mean and precision matrix associated with $\mathbf{x}|\boldsymbol{\theta}$. Further, $\mathbf{h}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}}$ and $\mathbf{G}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}}$ are

the mode, and the negative Hessian at the mode of $\mathbf{x} \mapsto \log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. Note that Equations 6 and 7 correspond to the precision and mean of the crude approximation $\propto \mathcal{N}\left(\mathbf{x}|\mathbf{h}_{\boldsymbol{\theta},\mathbf{x}}, \mathbf{G}_{\boldsymbol{\theta},\mathbf{x}}{}^{-1}\right)\mathcal{N}\left(\mathbf{x}|\mathbf{h}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}}, \mathbf{G}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}}{}^{-1}\right)$ to $p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. Moreover, it is also in some cases possible to find approximations to the involved negative Hessian that do not depend on $\mathbf{x}$ (see e.g. Kleppe, 2019), reducing the number of Cholesky factorization per evaluation of (3) to one.

Interestingly, the approximate pseudo-marginal MCMC method of Gómez-Rubio and Rue (2018) is closely connected to the proposed methodology with Laplace approximation-based transport maps. Specifically, $\omega_{\boldsymbol{\theta}}(\mathbf{0}_D)$ is the conventional Laplace approximation (see e.g. Tierney and Kadane, 1986) of $p(\mathbf{y}|\boldsymbol{\theta})$ (modulus the usage of an approximate mode and Hessian). By substituting $\omega_{\boldsymbol{\theta}}(\mathbf{0}_D)$ for $\omega_{\boldsymbol{\theta}}(\mathbf{u})$ in (3) (and integrating analytically over $\mathbf{u}$), the target distribution of Gómez-Rubio and Rue (2018) is obtained. Thus, the proposed methodology with Laplace approximation-based transport maps may be regarded as variant of the Gómez-Rubio and Rue (2018) method that corrects for the approximation error of the underlying Laplace approximation.

4.2 $m(\mathbf{x}|\boldsymbol{\theta})$ and $\gamma_{\boldsymbol{\theta}}(\mathbf{u})$ derived from the Efficient Importance Sampler

The efficient importance sampler (EIS) algorithm of Richard and Zhang (2007) is a widely used technique for constructing close to optimal importance densities, typically in the context of integrating out latent variables. At its core, the EIS relies initially on eliciting a family of sampling mechanisms, say $\mathbf{x} = \Gamma_{\mathbf{a}}(\mathbf{u})$, $\Gamma_{\mathbf{a}} : \mathbb{R}^D \mapsto \mathbb{R}^D$, indexed by some, typically high-dimensional parameter $\mathbf{a} \in \mathcal{A}$. Moreover, for all $\mathbf{a} \in \mathcal{A}$, and for $\mathbf{u} \sim N(\mathbf{0}_D, \mathbf{I}_D)$, the density of $\Gamma_{\mathbf{a}}(\mathbf{u})$ is denoted by $m_{\mathbf{a}}(\mathbf{x})$. The EIS algorithm proceeds by first sampling a collection of "common random numbers" $\mathbf{Z} = \left\{\mathbf{z}^{(i)}\right\}_{i=1}^{r}$, $\mathbf{z}^{(i)} \sim$ iid $N(\mathbf{0}_D, \mathbf{I}_D)$, $i = 1, \ldots, r$, then selecting an initial parameter $\mathbf{a}^{[0]}$, and finally iterate over the below steps for $j = 1, \ldots, J$:

- Sample latent states $\mathbf{x}^{(i)} = \Gamma_{\mathbf{a}^{[j-1]}}(\mathbf{z}^{(i)})$, $i = 1, \ldots, r$.
- Locate a new $\mathbf{a}^{[j]}$ as a (generally approximate) minimizer (over $\mathbf{a}$) of the sample variance of the importance weights $w_{\mathbf{a}}^{(i)} = p(\mathbf{y}|\mathbf{x}^{(i)}, \boldsymbol{\theta})p(\mathbf{x}^{(i)}|\boldsymbol{\theta})/m_{\mathbf{a}}(\mathbf{x}^{(i)})$, $i = 1, \ldots, r$.

An unbiased estimate of $p(\mathbf{y}|\boldsymbol{\theta})$ is given by the means of conventional importance sampling (Robert and Casella, 2004, Section 3.3) based on importance density $m_{\mathbf{a}^{[J]}}(\mathbf{x})$, with random draws (from $m_{\mathbf{a}^{[J]}}(\mathbf{x})$) generated based on random numbers independent from $\mathbf{z}^{(i)}$, $i = 1, \ldots, n$.

Notice that the near optimal EIS parameter $\mathbf{a}^{[J]} = \mathbf{a}^{[J]}(\boldsymbol{\theta}, \mathbf{Z})$ generally depends both on $\boldsymbol{\theta}$ and $\mathbf{Z}$. In the present context, for some fixed set of common random numbers $\mathbf{Z}$ and number of EIS iterations J, the importance density of (4) is simply set equal to the EIS importance density, i.e. $m(\mathbf{x}|\boldsymbol{\theta}) = m_{\mathbf{a}^{[J]}(\boldsymbol{\theta},\mathbf{Z})}(\mathbf{x})$.

Notice in particular that the EIS iterations above must be repeated for each evaluation of (4), and that the common random numbers must be kept fixed during each HMC iteration (which typically involve several evaluations of (4) and its gradient), or throughout the whole MCMC simulation.

The EIS importance density is often regarded as more reliable than the Laplace approximation counterpart, as it explicitly seeks to minimize the importance weight variation across typical outcomes of importance density. In addition, the family of importance densities $m_{\mathbf{a}}(\mathbf{x})$ may be constructed to highly non-Gaussian densities, whereas the Laplace approximation importance density is multivariate Gaussian. On the other hand, the EIS algorithm typically is substantially more costly in a computational perspective, whether this additional computational effort pays of in terms of a better decoupling effect in (3,4) is sought to be answered here.

The sketch of the EIS algorithm above is intentionally kept somewhat vague, as the actual details, both in terms of selecting $m_{\mathbf{a}}(\mathbf{x})$ and how the optimization step is implemented, depends very much on the model specification at hand. A more detailed description of the EIS suitable for the models considered in the simulation study discussed shortly is given in Section B of the supplementary material.

4.3 Implementation and Tuning Parameters

The proposed methodology has been implemented in two ways. Firstly, the Laplace approximation-based methods are implemented in Stan using the modified target representation (3). This is also the case for the reference method corresponding to $m_{\boldsymbol{\theta}}(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})$.

Secondly, we also consider a bespoke HMC implementation as outlined in Section 2.1, for $\mathbf{q} = (\boldsymbol{\theta}^T, \mathbf{u}^T)^T$, targeting either (3, for Laplace approximation-based methods) or (4, for EIS-based methods). This HMC method is based on the LD-integrator (see supplementary material, Section A) in order to better exploit the approximate decoupling effects in the target, and was in particular included to explore the advantage of using the LD-integrator over the leapfrog integrator in the present situation.

The mass matrix in the bespoke implementation was taken to be

$$\mathbf{M} = \begin{bmatrix} \hat{\mathbf{M}}_{\boldsymbol{\theta}} & \mathbf{0}_{d\times D} \\ \mathbf{0}_{D\times d} & \mathbf{I}_D \end{bmatrix},$$

where $\hat{\mathbf{M}}_{\boldsymbol{\theta}} = -\nabla^2_{\boldsymbol{\theta}} \log\left[\hat{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ and the simulated MAP $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log\left[\hat{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right]$ is obtained from an EIS importance sampling estimate $\hat{p}(\mathbf{y}|\boldsymbol{\theta})$ of $p(\mathbf{y}|\boldsymbol{\theta})$. Finding the approximate parameter marginal posterior precision $\hat{\mathbf{M}}_{\boldsymbol{\theta}}$ is very fast and requires minimal additional effort as gradients of the importance weight with respect to $\boldsymbol{\theta}$ are already available via automatic differentiation (AD, to be discussed shortly).

Notice that the mass matrix specific to $\mathbf{u}$ is take to be the identity to match the precision of the $N(\mathbf{0}_D, \mathbf{I}_D)$ "prior" of $\mathbf{u}$ in (3,4). As for the integrator step size ε and the number of integrator steps L, we retain L as a tuning parameter while keeping the total integration time εL per HMC proposal fixed at $\approx \pi/2$. This choice of total integration time is informed by the expectation that $\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}$ under (3,4) will be close to a Gaussian with precision matrix $\mathbf{M}$. Moreover, whenever $\tilde{\pi}(\mathbf{q})$ in (1) is Gaussian with precision $\mathbf{M}$, the dynamics (2) are periodic with period $t = 2\pi$, and choosing a quarter of such a cycle leads to HMC proposals $\mathbf{q}^*$ independent of the current configuration $\mathbf{q}^{(k)}$ (see e.g. Neal, 2011; Mannseth et al., 2018). Finally, L is tuned by hand to obtain acceptance rates around 0.9.

Both implementations rely on the ability to compute gradients of log-targets (3,4) with respect to both $\boldsymbol{\theta}$ and $\mathbf{u}$. To this end, we rely on Automatic Differentiation (AD). In Stan, this is done automatically, whereas in the bespoke implementation, the Adept C++ automatic differentiation software library (Hogan, 2014) is applied. Notice that for the Laplace approximation-based method, AD is applied to calculations of band-Cholesky factorizations, and thus there may be room for improvement in CPU times if the AD libraries supported such operations natively. The bespoke algorithm is implemented using the R (R Core Team, 2019) package `Rcpp` by Eddelbuettel and François (2011), which makes it possible to run compiled C++ code in R. Stan is used through its R interface `rstan` (Stan Development Team, 2019a), version 2.19.2. The same C++ compiler was used for both the bespoke and Stan methods. All computations are performed using R version 3.6.1 on a PC with an Intel Core i5-6500 processor running at 3.20 GHz.

5 Simulation study

This section presents applications of the proposed methodology to three non-Gaussian/nonlinear state-space latent variable models for the purpose of benchmarking against alternative methods. State-space models with univariate state were chosen as the Laplace approximation-based methods only require tri-diagonal Cholesky factorizations, which are easily implemented in the Stan language. The specific models are selected to illustrate the performance under different, empirically relevant, scenarios. In particular, the three models exhibit significantly different, and variable signal-to-noise ratios, which as discussed above may modulate the need for (non-trivial) transport map methods.

In the proceeding, different combinations of implementation ($\in$ {Stan, LD}) and transport map method ($\in$ {Prior, Laplace, EIS, Fisher}) are considered, where "LD" refers to the bespoke HMC implementation with LD integrator. Transport map "Prior" correspond to $m_{\boldsymbol{\theta}}(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})$ and is equivalent to carrying out the simulations in an $(\boldsymbol{\theta}, \boldsymbol{\eta})$-parameterization where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_D)'$ are a-priori standard normal disturbances of the models to be discussed.

Transport map method "Fisher" corresponds to Fisher information-based DRHMC approach of Kleppe (2019) applied to the latent variables only (i.e. general DRHMC involves non-trivial transport maps for the parameters also). Fisher also leads to an affine transport map $\gamma_{\boldsymbol{\theta}}(\mathbf{u}) = \mathbf{h}_F + \mathbf{L}_F^{-T}\mathbf{u}$, $\mathbf{L}_F\mathbf{L}_F^T = \mathbf{G}_F$. Here, $\mathbf{G}_F$ is the sum of the a-priori precision matrix of $\mathbf{x}$ and the Fisher information of the observations with respect to $\mathbf{x}$. Notice that this method requires both that said Fisher information is constant with respect to the latent state, and that the $p(\mathbf{x}|\boldsymbol{\theta})$ precision matrix has closed form, where the latter requirement limits its applicability to the first two models considered below.

Methods LD-Prior and LD-Fisher were not carried out as the default tuning discussed in Section 4.3 work poorly in these cases. Moreover, Stan-EIS was also not considered as it was impractical to implement the EIS algorithm in the Stan language. For each of the three models, the LD algorithm is simulated for 1,500 iterations, where the draws from the first 500 burn-in iterations are discarded. Stan uses (the default) 2,000 iterations with 1,000 burn-in steps also used for automatic tuning of the integrator step size and the mass matrix. The reported computing times are for the 1,000 sampling iterations for both methods. Further details for the different example models, including prior assumptions and details related to the Newton iterations for the Laplace maps, are found in the supplementary material, Section C.

5.1 Stochastic Volatility Model

The first example model is the discrete-time stochastic volatility (SV) model for financial returns given by (Taylor, 1986)

$$y_t = \exp(x_t/2)e_t, \quad e_t \sim \text{iid } N(0,1),\ t = 1,\dots,D, \tag{8}$$

$$x_t = \gamma + \delta x_{t-1} + \nu\eta_t, \quad \eta_t \sim \text{iid } N(0,1),\ t = 2,\dots,D, \tag{9}$$

where y_t is the return observed on day t, x_t is the latent log-volatility with initial condition $x_1 \sim N(\gamma/[1-\delta], \nu^2/[1-\delta^2])$, while e_t and η_t are mutually independent innovations. The data consists of daily log-returns on the U.S. dollar against the U.K. Pound Sterling from October 1, 1981 to June 28, 1985 with $D = 945$.

Under this SV model the data density $p(y_t|x_t) = \mathcal{N}(y_t|0, \exp\{x_t\})$ is fairly uninformative about the states x_t, with a Fisher information (w.r.t. x_t) which is independent of $\boldsymbol{\theta}$ and given by $-E[\nabla^2_{x_t} \log p(y_t|x_t)] = 1/2$, whereas the states are fairly volatile under typical estimates for $\boldsymbol{\theta}$. This low signal-to-noise ratio together with a shape of the data density which is independent of the parameters implies that the conditional posterior of the innovations $\boldsymbol{\eta}$ given $\boldsymbol{\theta}$ are close to a normal distribution regardless of $\boldsymbol{\theta}$, leading to a correspondingly well-behaved joint posterior of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Hence, this represents a scenario where the Stan-Prior sampling on the joint space of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ used as a benchmark can be expected to exhibit a comparably good performance.

		LD-EIS		Stan-Prior		LD-Laplace		Stan-Laplace		Stan-Fisher	
		Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean
	CPU time (s)	276.5	278	12.4	15	10.6	10.6	9.7	16.7	6.1	7.6
γ											
	Post. mean		-0.021		-0.021		-0.021		-0.021		-0.021
	Post. std.		0.012		0.01		0.011		0.011		0.011
	ESS	201	337	237	348	275	354	268	494	218	321
	ESS/s	0.7	1.2	18.1	23.5	25.7	33.3	16.7	37.2	5.6	27
δ											
	Post. mean		0.98		0.98		0.98		0.98		0.98
	Post. std.		0.01		0.01		0.01		0.01		0.01
	ESS	269	380	192	309	320	363	290	423	239	319
	ESS/s	1	1.4	15.3	20.6	30.1	34.1	13.9	32	5	27.2
v											
	Post. mean		0.15		0.15		0.15		0.15		0.15
	Post. std.		0.03		0.03		0.03		0.03		0.03
	ESS	363	503	243	332	360	512	274	431	226	293
	ESS/s	1.3	1.8	16.7	23	33.9	48.1	14.1	32.8	3.8	25.6

Table 1: Simulation study results for the SV model (8,9). ESS corresponds to the effective sample size (out of 1,000 iterations) and ESS/s is the number of effective samples produced per second of computing time. The columns "Min", "Mean" correspond to the minimum, mean across 8 independent replicas of the experiment. Burn-in iterations are not included in the reported CPU times. The tuning parameters are: LD-EIS: $J = 2$, $r = 6$, $\varepsilon = 0.4$ and $L = 4$. LD-Laplace: $K = 2$, $\varepsilon = 0.4$ and $L = 4$. Stan-Laplace: $K = 0$.

For the Fisher transport map method, $\mathbf{G}_F = \mathbf{G}_{\boldsymbol{\theta},\mathbf{x}} + \mathbf{G}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}}$, and as suggested by Table 4 of Kleppe (2019), we set $\mathbf{h}_F = \mathbf{0}_d$.

Table 1 shows the HMC posterior mean and standard deviation for the parameters, which are sample averages computed from 8 independent replications. It also reports the effective sample size (ESS) (Geyer, 1992) and the ESS per second of CPU time (ESS/s), where the latter will be the main performance measure (provided of course that the MCMC method properly explores the target distribution) considered here. Several settings of the tuning parameters (i.e. some subset of r, J, K, and L) where considered, and the presented results are the best considered, in terms of ESS/s. Table 1 indicates firstly that all five methods produce a good exploration of the target distribution with posterior moments being essentially the same. For the Stan-based methods, there is substantial variation in the CPU times due to variation in the automatic tuning of the integrator step size ε over the replica. Judging from the ESS values, on average there is not much to be gained from introducing the Laplace approximation- and EIS-based transport map for this model. This finding mirrors to some extent what was found by Kleppe (2019, Section 5.2), and is also as expected since the observations carry very little information regarding the states. In terms of ESS/s, there is no uniform winner, but the computational overhead of locating the EIS importance density is clearly not worthwhile for this model, relative to the computationally cheaper Laplace- and Fisher transport maps.

5.2 Gamma Model for Realized Volatilities

The second example model is a dynamic state-space model for the realized variance of asset returns (see, e.g., Golosnoy et al., 2012, and references therein). It has the form

$$y_t = \beta \exp(x_t) e_t, \quad e_t \sim \text{iid } G(1/\tau, \tau),\ t = 1, \ldots, D, \tag{10}$$

$$x_t = \delta x_{t-1} + \nu \eta_t, \quad \eta_t \sim \text{iid } N(0,1),\ t = 2, \ldots, D, \tag{11}$$

where y_t is the daily realized variance measuring the latent integrated variance $\beta \exp(x_t)$, and $G(1/\tau, \tau)$ denotes a Gamma-distribution for e_t normalized such that $E(e_t) = 1$ and $\text{Var}(e_t) = \tau$. The innovations e_t and η_t are independent and the initial condition for the log-variance is $x_1 \sim N(0, \nu^2/[1-\delta^2])$. This Gamma volatility model is applied to a data set consisting of $D = 2,514$ observations of the daily realized variance for the American Express stock (more information concerning the data is given in Section 6; y_t here is identical to the 1,1-element of realized covariance matrices $\mathbf{Y}_t$).

In contrast to the SV model, this Gamma model applied to the realized variance data has both a considerably higher signal-to-noise ratio and a shape of the data density $x_t \mapsto p(y_t|x_t, \boldsymbol{\theta})$ which depends on the parameters. In particular, the Fisher information of its data density with respect to x_t is $1/\tau$ with an estimate of $\tau \simeq 0.13$ (see Table 2), while the estimated volatility of the states is roughly as large as under the SV model. Hence, it can be expected that the conditional posterior of the innovations $\boldsymbol{\eta}$ given $\boldsymbol{\theta}$ deviates distinctly from a Gaussian form and exhibits nonlinear dependence on $\boldsymbol{\theta}$, which makes the Gamma model a more challenging scenario for the Stan-Prior benchmark than the SV model.

The same initial guess $\mathbf{h}^{(0)}$ in the Laplace scaling as for the SV model above was applied, and also here $\mathbf{G}_F$ coincides with $\mathbf{G}_{\boldsymbol{\theta},\mathbf{x}} + \mathbf{G}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}}$. Choosing $\mathbf{h}_F = \mathbf{0}$ leads to poor results, and we therefore set $\mathbf{h}_F$ equal to (7) (see also Kleppe, 2019, Equation 20). Consequently, Stan-Fisher coincides with Stan-Laplace, $K = 0$ (which was also found to be the optimal Stan-Laplace method in this situation). The remaining experiment setup is also identical to that for the SV model, and the results are given in Table 2. Stan-Prior produces substantially lower ESSes than the EIS- and Laplace methods, which we attribute to the failure to take the higher information content from the observations into account in the transport map. LD-Laplace and Stan-Laplace are the winners in terms of ESS/s and again it is not beneficial to opt for the presumably more accurate and expensive EIS-transport map over the cruder and computationally faster Laplace-approximation.

		LD-EIS		Stan-Prior		LD-Laplace		Stan-Laplace	
		Min	Mean	Min	Mean	Min	Mean	Min	Mean
	CPU time (s)	935.4	938.1	150.5	171.1	50.9	51.1	40.8	62
τ									
	Post. mean		0.13		0.13		0.13		0.13
	Post. std.		0.006		0.006		0.006		0.006
	ESS	1000	1000	194	238	1000	1000	623	873
	ESS/s	1.1	1.1	1.1	1.4	19.5	19.6	10.1	15.2
β									
	Post. mean		2.7		2.8		2.5		2.8
	Post. std.		0.8		1		0.8		0.9
	ESS	460	542	65	281	216	568	103	505
	ESS/s	0.5	0.6	0.4	1.7	4.2	11.1	2.5	8.3
δ									
	Post. mean		0.98		0.98		0.98		0.98
	Post. std.		0.004		0.004		0.004		0.004
	ESS	497	641	207	282	384	685	382	719
	ESS/s	0.5	0.7	1.3	1.7	7.5	13.4	8.7	11.9
ν									
	Post. mean		0.22		0.22		0.22		0.22
	Post. std.		0.01		0.01		0.01		0.01
	ESS	827	976	139	178	1000	1000	416	785
	ESS/s	0.9	1	0.6	1.1	19.5	19.6	8.3	13.4

Table 2: Simulation study results for the Gamma model (10,11). ESS corresponds to the effective sample size (out of 1,000 iterations) and ESS/s is the number of effective samples produced per second of computing time. The columns "Min", "Mean" correspond to the minimum, mean across 8 independent replicas of the experiment. Burn-in iterations are not included in the reported CPU times. The tuning parameters are: LD-EIS: $J = 2$, $r = 5$, $\varepsilon = 0.64$ and $L = 3$, LD-Laplace: $K = 1$, $\varepsilon = 0.64$ and $L = 3$. Stan-Laplace: $K = 0$. Notice that Stan-Fisher and Stan-Laplace coincide in this case.

5.3 Constant Elasticity of Variance Diffusion Model

The last example model is a time-discretized version of the constant elasticity of variance (CEV) diffusion model for short-term interest rates (Chan et al., 1992), extended by a measurement error to account for microstructure noise (Aït-Sahalia, 1999; Kleppe and Skaug, 2016). The resulting model for the interest rate y_t observed at day t with a corresponding latent state $x_t > 0$, is described as

$$y_t = x_t + \sigma_y e_t, \quad e_t \sim \text{iid } N(0,1),\ t = 1, \ldots, D, \tag{12}$$

$$x_t = x_{t-1} + \Delta(\alpha - \beta x_{t-1}) + \sigma_x x_{t-1}^{\gamma}\sqrt{\Delta}\eta_t, \quad \eta_t \sim \text{iid } N(0,1),\ t = 2, \ldots, D, \tag{13}$$

where e_t and η_t are mutually independent and $\Delta = 1/252$. The parameters are $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \sigma_x, \sigma_y)$ and the initial condition $x_1 \sim N(y_1, 0.01^2)$. The data consist of $D = 3,082$ daily 7-day Eurodollar deposit spot rates from January 2, 1983 to February 25, 1995 (see Aït-Sahalia, 1996 for a description of this data set).

The estimated standard deviation of the noise component σ_y is very small with an estimate of 0.0005 (see

		LD-EIS		LD-Laplace		Stan-Laplace	
		Min	Mean	Min	Mean	Min	Mean
	CPU time (s)	615.6	618.8	60.3	60.6	482.2	515.7
α							
	Post. mean		0.01		0.01		0.01
	Post. std.		0.01		0.01		0.01
	ESS	869	984	876	972	1000	1000
	ESS/s	1.4	1.6	14.5	16	1.9	1.9
β							
	Post. mean		0.17		0.17		0.17
	Post. std.		0.17		0.17		0.17
	ESS	707	963	745	957	1000	1000
	ESS/s	1.1	1.6	12.4	15.8	1.9	1.9
γ							
	Post. mean		1.18		1.18		1.18
	Post. std.		0.06		0.06		0.06
	ESS	759	957	1000	1000	631	852
	ESS/s	1.2	1.5	16.4	16.5	1.3	1.6
σ_x							
	Post. mean		0.41		0.41		0.41
	Post. std.		0.06		0.06		0.06
	ESS	769	946	1000	1000	650	890
	ESS/s	1.2	1.5	16.4	16.5	1.3	1.7
σ_y							
	Post. mean		0.0005		0.0005		0.0005
	Post. std.		0.00002		0.00002		0.00002
	ESS	769	963	1000	1000	1000	1000
	ESS/s	1.2	1.6	16.4	16.5	1.9	1.9

Table 3: Simulation study results for the CEV model (12,13). ESS corresponds to the effective sample size (out of 1,000 iterations) and ESS/s is the number of effective samples produced per second of computing time. The columns "Min", "Mean" correspond to the minimum, mean across 8 independent replicas of the experiment. Burn-in iterations are not included in the reported CPU times. The tuning parameters are: LD-EIS: $J = 1$, $r = 7$, $\epsilon = 0.57$ and $L = 3$. LD-Laplace: $K = 2$, $\epsilon = 0.57$ and $L = 3$, Stan-Laplace: $K = 1$.

Table 3) so that the data density $x_t \mapsto p(y_t|x_t, \boldsymbol{\theta})$ is strongly peaked at $x_t = y_t$ and by far more informative about x_t than in the SV- and Gamma model with a Fisher information given by $1/\sigma_y^2$. Also, the volatility of the states is not constant and depends, unlike in the previous models, nonlinearly on the level of the states. As a result, the posterior of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ strongly deviates from being Gaussian. Consequently, Stan-Prior fails to produce meaningful results and is therefore not reported on. Moreover, since the prior on $\mathbf{x}$ is nonlinear and its precision matrix does not seem to have closed-form, Fisher-scaling is not feasible.

Table 3 reports results for LD-EIS, LD-Laplace and Stan-Laplace, and it is seen that all three methods produce reliable results. In terms of ESS per computing time, the LD-Laplace is a factor 5-10 faster than the other methods, where the difference between LD-Laplace and Stan-Laplace is due to the substantially higher number of integrator steps required for Stan-Laplace.

The same model and data set was also considered by Kleppe (2018, Section 5), who compare the modified

Cholesky Riemann manifold HMC algorithm and a Gibbs sampling procedure. Both methods were implemented in C++ and thus the orders of magnitude of produced ESS per computing time are comparable to the present situation. It is seen that for the "most difficult" parameters γ, σ_x, the proposed methodology is roughly two order of magnitude faster than the Riemann manifold HMC method and roughly three orders of magnitude faster than the Gibbs sampler.

5.4 Summary from simulation experiment

For models with higher signal-to-noise ratios than the SV model, the proposed methodology produces large speedups (or makes challenging models feasible as for the CEV model) relative to the benchmarks, even if the per evaluation cost of the modified target is higher than in the default parameterization. For the considered models, the EIS transport map is not competitive relative to the Laplace approximation counterpart due to the relatively higher computational cost. For the Laplace-based methods, it is seen that relatively few Newton iterations is optimal in an ESS per computing time perspective. Overall, and very much in line with Kleppe (2019), this is indicative that rather crude representations of the location and scale of $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ are sufficient. Moreover, this latter observation ties in with the second point discussed in Section 3.2: Due to the thin-tailed Gaussian distribution entering explicitly in representation (4) of the modified target, the importance sampling rule of thumb that you should seek high-fidelity approximations to $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ as the importance density is less relevant in the present situation.

With respect to the choice of integrator, it is seen that the LD-integrator and the leapfrog-integrator-based Stan produces similar raw ESSes, but that that the LD-integrator in general requires non-trivially fewer integration steps to accomplish this. E.g., the reported (automatically tuned) Stan-Laplace results for the CEV model required on average 63 leapfrog steps whereas the corresponding (manually tuned) number for LD-Laplace was 3. For the two other models, the performance of the LD integrator is roughly on par with Stan when Laplace scaling was employed. Further, the LD integrator generally needs more refined Laplace maps (higher K) to work satisfactory, whereas under Stan, more crude Laplace transport maps are permissible.

6 High-dimensional application

6.1 Model

To illustrate the proposed methodology in a high-dimensional situation, we consider the dynamic inverted Wishart model for realized covariance matrices proposed in Grothe et al. (2019, Section 6). More specifically,

for a time series of $r \times r$ symmetric positive definite observed realized covariance matrices $\mathbf{Y}_t,\ t = 1, \ldots, D$, the observations are modeled conditionally inverse-Wishart distributed,

$$p(\mathbf{Y}_t|\mathbf{\Sigma}_t, \nu) \propto |\mathbf{Y}_t|^{-\frac{\nu+r+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}_t \mathbf{Y}_t^{-1}\right)\right), \tag{14}$$

so that $E(\mathbf{Y}_t) = (\nu + r + 1)^{-1}\mathbf{\Sigma}_t$. Here, the degrees of freedom $\nu > r + 1$ is a parameter, and $\mathbf{\Sigma}_t$ is a (latent) time-varying scale matrix, given by

$$\mathbf{\Sigma}_t = \mathbf{H}\mathbf{D}_t\mathbf{H}^T,\ \mathbf{D}_t = \mathrm{diag}(\exp(x_{1,t}), \ldots, \exp(x_{r,t})),$$

where $\boldsymbol{H}$ is a lower triangular matrix with ones along the main diagonal and unrestricted parameters $h_{i,j}, i > j, 1 \leq j < r$ below the main diagonal. Moreover, $\mathbf{x}_s = \{x_{s,t}\}_{t=1}^D,\ s = 1, \ldots, r$ are latent Gaussian AR(1) processes

$$x_{s,t} = \mu_s + \delta_s(x_{s,t-1} - \mu_s) + \sigma_s\eta_{s,t},\ t = 2, \ldots, D,\ s = 1, \ldots, r, \tag{15}$$

$$x_{s,1} = \mu_s + \frac{\sigma_s}{\sqrt{1-\delta_s^2}}\eta_{s,1},\ s = 1, \ldots, r \tag{16}$$

where $\eta_{s,t} \sim$ iid $N(0, 1),\ t = 1, \ldots, D,\ s = 1, \ldots, r$. In total, the model contains $1+3r+r(r-1)/2$ parameters $\boldsymbol{\theta} = (\nu, \mu_{1:r}, \delta_{1:r}, \sigma_{1:r}, h_{2:r,1}, h_{3:r,2}, \ldots, h_{r,r-1})$. Further details concerning the model specification and priors can be found in the supplementary material (Section D).

A fortunate property of this model is that the conditional posterior of the latent states are independent over s, i.e. $p(\mathbf{x}_{1:r}|\boldsymbol{\theta}, Y_{1:D}) = \prod_{s=1}^r p(\mathbf{x}_s|\boldsymbol{\theta}, Y_{1:D})$. This implies that the transport map for $\mathbf{x}$ also may be split into r individual transport maps, say $\mathbf{x}_s = \gamma_{\boldsymbol{\theta},s}(\mathbf{u}_s),\ \mathbf{u}_s = \{u_{s,t}\}_{t=1}^D,\ s = 1, \ldots, r$, without losing fidelity. The (combined) transport map becomes $\gamma_{\boldsymbol{\theta}}(\mathbf{u}) = [(\gamma_{\boldsymbol{\theta},1}(\mathbf{u}_1))^T, \ldots, (\gamma_{\boldsymbol{\theta},r}(\mathbf{u}_r))^T]^T$, where $\mathbf{u} = [\mathbf{u}_1^T, \ldots, \mathbf{u}_r^T]^T$, and in particular $|\nabla_{\mathbf{u}}\gamma_{\boldsymbol{\theta}}(\mathbf{u})| = \prod_{s=1}^r |\nabla_{\mathbf{u}_s}\gamma_{\boldsymbol{\theta},s}(\mathbf{u}_s)|$ due to the block-diagonal nature of the Jacobian of $\gamma_{\boldsymbol{\theta}}$.

Further, each of the factors of the conditional posterior have a shape corresponding that of a state-space model with univariate state-process $\mathbf{x}_s$:

$$p(\mathbf{x}_s|\boldsymbol{\theta}, Y_{1:D}) \propto p(\mathbf{x}_s|\boldsymbol{\theta}) \prod_{t=1}^D \exp\left(\frac{\nu}{2}x_{s,t} - \frac{\tilde{y}_{s,t}}{2}\exp(x_{s,t})\right),\ \tilde{y}_{s,t} = (\mathbf{H}_{1:s,s})^T \mathbf{Y}_t^{-1}\mathbf{H}_{1:s,s},\ s = 1, \ldots, r. \tag{17}$$

Thus, individual transport maps $\gamma_{\boldsymbol{\theta},s}$ may be constructed to target (17) as described in the previous Sections. In particular, individual Laplace approximation-based maps, $\gamma_{\boldsymbol{\theta},s}$, involve only tri-diagonal Cholesky factorizations. It is, however, worth noticing that the proposed methodology does not rely on such a conditional independence structure in order to be applicable per se.

	Stan-Prior	Stan-Laplace $K=0$	Stan-Laplace $K=1$	Stan-Laplace $K=2$
CPU time (s)	6437	910	1196	1452
$\mu_{1:5}$ ESS (min , max)	(832 , 918)	(967 , 1000)	(987 , 1000)	(985 , 1000)
$\sigma_{1:5}$ ESS (min , max)	(301 , 349)	(1000 , 1000)	(1000 , 1000)	(1000 , 1000)
$\delta_{1:5}$ ESS (min , max)	(357 , 501)	(980 , 1000)	(986 , 1000)	(975 , 1000)
$h_{i,j}$ ESS (min , max)	(972 , 1000)	(984 , 1000)	(1000 , 1000)	(1000 , 1000)
ν ESS	562	1000	1000	1000
$x_{1:5,1}$ ESS (min , max)	(986 , 1000)	(1000 , 1000)	(1000 , 1000)	(1000 , 1000)
$u_{1:5,1}$ ESS (min , max)	(871 , 959)	(1000 , 1000)	(1000 , 1000)	(1000 , 1000)

Table 4: Effective sample sizes and CPU times for the inverse Wishart model (14-16). The parameters are grouped, and the reported ESS figures are (min, max) across each group. All of the results are averages across 8 independent replica of each experiment. Here, $u_{s,1}$ is the first element in $\mathbf{u}_s$. Under Prior transport map, $u_{1:5,1}$ is identical to $\eta_{1:5,1}$ in (16).

The observed Fisher information (w.r.t. $x_{s,t}$) of the marginal "measurement densities" $\propto \exp(\frac{\nu}{2}x_{s,t} - \frac{\tilde{y}_{s,t}}{2}\exp(x_{s,t}))$ equals $\nu/2$, with an estimate of $\nu \simeq 33.6$ for the data set considered here (see Table 5 in supplementary material). Thus, the signal to noise ratio here is similar to that of the Gamma model considered in section 5.2. As the LD- and Stan- results are similar for the Gamma model, we consider only Stan for this model, as it entails only a few dozen lines of Stan code and tuning is fully automated. EIS was found not to be competitive and is not considered here. The initial guess $\mathbf{h}_{\boldsymbol{\theta}}^{(0)}$ under Laplace scaling is given by (7), whereas $\mathbf{G}_{\boldsymbol{\theta}}^{(K)} = \mathbf{G}_{\boldsymbol{\theta}}^{(0)}$ given in (6). This (fixed) matrix was also used as the scaling matrix in the approximate Newton iterations for $K = 0, 1, 2$ (see supplementary material, Section D for more details).

6.2 Data and results

The data set of $D = 2,514$ observations of daily realized covariance matrices of $r = 5$ stocks (American Express, Citigroup, General Electric, Home Depot, and IBM) spanning Jan. 1st, 2000 to Dec. 31, 2009 is described in detail in Golosnoy et al. (2012). The same model and data set was considered in Grothe et al. (2019), where Gibbs sampling procedures were considered. From Grothe et al. (2019), it is seen that even with close to iid sampling from $p(\mathbf{x}_{1:r}|\boldsymbol{\theta}, Y_{1:D})$, the chains for ν and $\sigma_s,\ s = 1, \ldots, r$ mix rather poorly under Gibbs sampling.

The ESSes for the parameters and the first elements of $\mathbf{x}_s$ and $\mathbf{u}_s$, and CPU times for Stan-Prior and Stan-Laplace are given in Table 4. Corresponding posterior means and standard deviations for Stan-Laplace ($K = 0$) are given in Table 5 in the supplementary material and these are very much in line with Grothe et al. (2019, Table 5).

From Table 4 it is seen that the proposed methodology Stan-Laplace outperforms the benchmark Stan-Prior, both in terms CPU time (the modified target is highly non-Gaussian and thus requires many integration

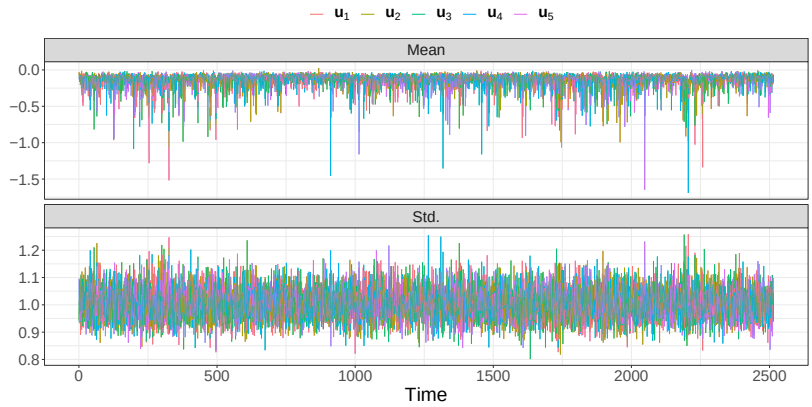

Figure 1: Posterior mean and standard deviation of $\mathbf{u}_s,\ s = 1, \ldots, 5$, for the inverse Wishart model (14-16) under Laplace transport map with $K = 0$. The results are for a single representative simulation replica with 1000 sampling iterations.

steps) and ESS. Indeed, Stan-Laplace with $K = 0$ is at least a full order of magnitude faster in terms of ESS per CPU time than Stan-Prior for the "difficult" parameters ν and $\sigma_s,\ s = 1, \ldots, r$. The added per evaluation computational cost of the more accurate Laplace approximations ($K = 1$ and $K = 2$) is not worthwhile, and this again corroborates the finds above that only crude location- and scale information with respect to $p(\mathbf{x}_{1:r}|\boldsymbol{\theta}, Y_{1:D})$ is needed. Figure 1 depicts the posterior mean and (marginal) standard deviation of each $\mathbf{u}_s$, for Stan-Laplace with $K = 0$. It is seen that the posterior standard deviations are close to 1, which one would expect in the case of close to perfect decoupling, i.e. is indicative that any funnel effects have been removed. The posterior means, on the other hand, are somewhat off 0, which is related both to the usage of the initial guess (7) and the fact that (17) is non-Gaussian and thus cannot be exactly decoupled using a Gaussian importance density. Figures 2,3 in the supplementary material shows corresponding plots for $K = 2$ and $K = 10$, and it is seen that the posterior means of $\mathbf{u}_s$ are closer to zero, but some deviation still exact due to the non-Gaussian target.

Comparing the computational performance to the Gibbs sampler in Grothe et al. (2019), it is seen that Stan-Laplace is also roughly an order of magnitude faster than a Gibbs sampler. This comparison is somewhat complicated by that Grothe et al. (2019) employ parallel processing (over s) when sampling the latent states $\mathbf{x}_s$, and that the computations in Grothe et al. (2019) are done in MATLAB, whereas Stan is based on compiled C++ code. In this consideration, also the fact that a model with 20 parameters and 12,570 latent variables can be fitted using a few minutes of CPU time and minimal coding efforts in Stan must be weighed against the typically time consuming and error-prone development efforts to develop Gibbs

samplers tailored for any given model.

7 Discussion

The paper proposes and evaluates importance sampler-based transport map HMC for Bayesian hierarchical models. The methodology relies on using off-the-shelf importance sampling strategies for high-dimensional latent variables to construct a modified target distribution that is easily sampled using (fixed metric) HMC. Indeed, as illustrated, the proposed methodology can lead to large speedups relative to relevant benchmarks for models with high-dimensional latent variables, while still being easily implemented using e.g. Stan.

Two strategies for selecting the involved importance samplers were considered in order to assess the optimal accuracy versus computational cost-tradeoff. The main insight in this regard is that only rather crude importance densities/transport maps (e.g. Laplace or DRHMC-type) are required when these are applied in the present framework. This observation is very much to the contrary to the importance sampling literature at large, where typically very accurate importance densities are required to produce reliable approximations to marginal likelihood functions when integrating over high-dimensional latent variables.

The proposed methodology, with Laplace transport maps and few or no Newton iterations lead to similar transport maps as those used in DRHMC in the cases where DRHMC is applicable. Thus the Laplace transport map approach may, in a rather broad sense, be seen as a generalization of DRHMC to models with nonlinear structures where DRHMC is not applicable.

Finally, there is scope for future research in developing software that can encompass a large class of models, and which implements the proposed methodology in a user-friendly manner. In particular, such software should include a sparse Cholesky algorithm for more general sparsity structures so that Laplace-based transport maps for e.g. multivariate latent state dynamic models and spatial models can be considered.

References

Aït-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate. *The Review of Financial Studies 9*(2), 385–426.

Aït-Sahalia, Y. (1999). Transition densities for interest rate and other nonlinear diffusions. *The Journal of Finance 54*(4), 1361–1395.

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.

Chan, K. C., G. A. Karolyi, F. A. Longstaff, and A. B. Sanders (1992). An empirical comparison of alternative models of the short-term interest rate. *The Journal of Finance 47*(3), 1209–1227.

Danielsson, J. (1994). Stochastic volatility in asset prices: Estimation with simulated maximum likelihood. *Journal of Econometrics 64*, 375–400.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Physics letters B 195*(2), 216–222.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods* (2 ed.). Number 38 in Oxford Statistical Science. Oxford University Press.

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

Flury, T. and N. Shephard (2011). Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. *Econometric Theory 27*(Special Issue 05), 933–956.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. Rubin (2014). *Bayesian Data Analysis* (3 ed.). CRC Press.

Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science 7*(4), 473–483.

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.

Golosnoy, V., B. Gribisch, and R. Liesenfeld (2012). The conditional autoregressive Wishart model for multivariate stock market volatility. *Journal of Econometrics 167*(1), 211–223.

Gómez-Rubio, V. and H. Rue (2018). "markov chain monte carlo with the integrated nested laplace approximation". *Statistics and Computing 28*(5), 1033–1051.

Grothe, O., T. S. Kleppe, and R. Liesenfeld (2019). The Gibbs sampler with particle efficient importance sampling for state-space models. *Econometric Reviews 38*(10), 1152–1175.

Hoffman, M., P. Sountsov, J. V. Dillon, I. Langmore, D. Tran, and S. Vasudevan (2019). NeuTra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport. arXiv:1903.03704.

Hoffman, M. D. and A. Gelman (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

Hogan, R. J. (2014). Fast reverse-mode automatic differentiation using expression templates in c++. *ACM Transactions on Mathematical Software (TOMS) 40*(4), 26.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994). Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics 12*(4), 371–89.

Kleppe, T. S. (2018). Modified Cholesky Riemann manifold Hamiltonian Monte Carlo: exploiting sparsity for fast sampling of high-dimensional targets. *Statistics and Computing 28*(4), 795–817.

Kleppe, T. S. (2019). Dynamically rescaled Hamiltonian Monte Carlo for Bayesian hierarchical models. *Journal of Computational and Graphical Statistics 28*(3), 493–507.

Kleppe, T. S. and H. J. Skaug (2016). Bandwidth selection in pre-smoothed particle filters. *Statistics and Computing 26*(5), 1009–1024.

Koopman, S. J., N. Shephard, and D. Creal (2009). Testing the assumptions behind importance sampling. *Journal of Econometrics 149*(1), 2 – 11.

Leimkuhler, B. and S. Reich (2004). *Simulating Hamiltonian dynamics*. Cambridge University Press.

Lindsten, F. and A. Doucet (2016). Pseudo-Marginal Hamiltonian Monte Carlo. arXiv preprint arXiv:1607.02516.

Mannseth, J., T. S. Kleppe, and H. J. Skaug (2018). On the application of improved symplectic integrators in Hamiltonian Monte Carlo. *Communications in Statistics-Simulation and Computation 47*(2), 500–509.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2*, 113–162.

Parno, M. and Y. Marzouk (2018). Transport map accelerated Markov chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification 6*(2), 645–682.

Pitt, M. K., R. dos Santos Silva, P. Giordani, and R. Kohn (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics 171*(2), 134 – 151.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richard, J.-F. and W. Zhang (2007). Efficient high-dimensional importance sampling. *Journal of Econometrics 141*(2), 1385–1411.

Robert, C. and G. Casella (2004). *Monte Carlo methods* (2 ed.). Springer, Berlin.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*(3), 653–667.

Stan Development Team (2019a). RStan: the R interface to Stan, Version 2.19.2. http://mc-stan.org.

Stan Development Team (2019b). *Stan user's guide, version 2.21.* http://mc-stan.org.

Taylor, S. J. (1986). *Modelling Financial Time Series.* Wiley, Chichester.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*(393), 82–86.

Zhang, Y. and C. Sutton (2014). Semi-separable Hamiltonian Monte Carlo for inference in Bayesian hierarchical models. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27*, pp. 10–18. Curran Associates, Inc.

Supplementary Material for "Importance Sampling-based Transport map Hamiltonian Monte Carlo for Bayesian Hierarchical Models"

Equation numbers < 18 refer to the equations in the main text.

A The Lindsten and Doucet (2016)-integrator

The the pseduo-marginal HMC (PM-HMC) algorithm of Lindsten and Doucet (2016) can be viewed as a standard HMC algorithm for simulating the random vector $\mathbf{q} = (\boldsymbol{\theta}', \mathbf{u}')'$ from the modified target densities (3) or (4). Proceeding with representation (4), the Hamiltonian is taken to be

$$H(\boldsymbol{\theta}, \mathbf{u}, \mathbf{p}_{\boldsymbol{\theta}}, \mathbf{p}_{\mathbf{u}}) = -\log \omega_{\boldsymbol{\theta}}(\mathbf{u}) - \log p(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{u}'\mathbf{u} + \frac{1}{2}\mathbf{p}_{\boldsymbol{\theta}}'\mathbf{M}_{\boldsymbol{\theta}}^{-1}\mathbf{p}_{\boldsymbol{\theta}} + \frac{1}{2}\mathbf{p}_{\mathbf{u}}'\mathbf{p}_{\mathbf{u}}, \tag{18}$$

where $\mathbf{p}_{\boldsymbol{\theta}} \in \mathbb{R}^d$ and $\mathbf{p}_{\mathbf{u}} \in \mathbb{R}^D$ are the artificial momentum variables specific to $\boldsymbol{\theta}$ and $\mathbf{u}$, respectively. Note that for this form of the extended Hamiltonian the mass matrix ($\mathbf{M}$) of the compound vector $(\boldsymbol{\theta}', \mathbf{u}')'$ is selected to be block diagonal, where the mass matrix specific to $\boldsymbol{\theta}$ is denoted by $\mathbf{M}_{\boldsymbol{\theta}} \in \mathbb{R}^{d\times d}$, while the mass for $\mathbf{u}$ is set equal to the identity in order to match the a-priori precision matrix of $\mathbf{u}$. Straight forward modifications of (18) and the proceeding theory applies if representation (3) is computationally more convenient.

Applying Hamilton's equations (2) to the extended Hamiltonian (18), for $\mathbf{q} = (\boldsymbol{\theta}', \mathbf{u}')'$ and $\mathbf{p} = (\mathbf{p}_{\boldsymbol{\theta}}', \mathbf{p}_{\mathbf{u}}')'$, we get the following equations of motion

$$\frac{d}{dt}\begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{p}_{\boldsymbol{\theta}} \\ \mathbf{u} \\ \mathbf{p}_{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{M}_{\boldsymbol{\theta}}^{-1}\mathbf{p}_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log \omega_{\boldsymbol{\theta}}(\mathbf{u}) \\ \mathbf{p}_{\mathbf{u}} \\ -\boldsymbol{u} + \nabla_{\boldsymbol{u}} \log \omega_{\boldsymbol{\theta}}(\mathbf{u}) \end{pmatrix}. \tag{19}$$

Equation (19) shows that the Hamiltonian transition dynamics of $(\boldsymbol{\theta}, \mathbf{p}_{\boldsymbol{\theta}})$ and $(\mathbf{u}, \mathbf{p}_{\mathbf{u}})$ are linked together via their joint dependence on the importance weight $\omega_{\boldsymbol{\theta}}(\mathbf{u})$. However, this link vanishes as the MC variance of the MC estimator $\mathrm{Var}_{\mathbf{u}}[\omega_{\boldsymbol{\theta}}(\mathbf{u})]$ tend to zero. In fact, an 'exact' MC estimate with zero MC variance implies that $\nabla_{\mathbf{u}} \log \omega_{\boldsymbol{\theta}}(\mathbf{u}) = \mathbf{0}_D$, in which case the transition dynamics of $(\boldsymbol{\theta}, \mathbf{p}_{\boldsymbol{\theta}})$ would be completely decoupled from that of $(\mathbf{u}, \mathbf{p}_{\mathbf{u}})$ and would be (marginally) the dynamics of the 'ideal' HMC algorithm for $p(\boldsymbol{\theta}|\mathbf{y})$. Moreover, the resulting marginal $(\mathbf{u}, \mathbf{p}_{\mathbf{u}})$-dynamics would reduce to that of a harmonic oscillator

with analytical solutions given by $\mathbf{u}(t) = \cos(t)\mathbf{u}(0) + \sin(t)\mathbf{p}_{\mathbf{u}}(0)$ and $\mathbf{p}_{\mathbf{u}}(t) = \cos(t)\mathbf{p}_{\mathbf{u}}(0) - \sin(t)\mathbf{u}(0)$.

In order to approximate the Hamiltonian transition dynamics (19), Lindsten and Doucet (2016) develop a symplectic integrator which for exact likelihood estimates produces exact simulations for the dynamics of $(\mathbf{u}, \mathbf{p}_{\mathbf{u}})$ and reduces for $(\boldsymbol{\theta}, \mathbf{p}_{\boldsymbol{\theta}})$ to the conventional leapfrog integrator. They derive this integrator for the special case where the mass matrix $\mathbf{M}_{\boldsymbol{\theta}}$, in (18) and (19) is restricted to be the identity. For the more general case with an unrestricted $\mathbf{M}_{\boldsymbol{\theta}}$ this integrator for approximately advancing the dynamics from time $t = 0$ to time $t = \varepsilon$ is given by

$$\boldsymbol{\theta}(\varepsilon/2) = \boldsymbol{\theta}(0) + (\varepsilon/2)\mathbf{M}_{\boldsymbol{\theta}}^{-1}\mathbf{p}_{\boldsymbol{\theta}}(0), \tag{20}$$

$$\mathbf{u}(\varepsilon/2) = \cos(\varepsilon/2)\mathbf{u}(0) + \sin(\varepsilon/2)\mathbf{p}_{\mathbf{u}}(0), \tag{21}$$

$$\mathbf{p}_{\mathbf{u}}^{*} = \cos(\varepsilon/2)\mathbf{p}_{\mathbf{u}}(0) - \sin(\varepsilon/2)\mathbf{u}(0), \tag{22}$$

$$\mathbf{p}_{\mathbf{u}}^{**} = \mathbf{p}_{\mathbf{u}}^{*} + \varepsilon\,\nabla_{\mathbf{u}}\big\{\log\omega_{\boldsymbol{\theta}(\varepsilon/2)}(\mathbf{u}(\varepsilon/2))\big\}, \tag{23}$$

$$\mathbf{p}_{\boldsymbol{\theta}}(\varepsilon) = \mathbf{p}_{\theta}(0) + \varepsilon\,\nabla_{\boldsymbol{\theta}}\big\{\log p\big[\boldsymbol{\theta}(\varepsilon/2)\big] + \log\omega_{\boldsymbol{\theta}(\varepsilon/2)}(\mathbf{u}(\varepsilon/2))\big\}, \tag{24}$$

$$\boldsymbol{\theta}(\varepsilon) = \boldsymbol{\theta}(\varepsilon/2) + (\varepsilon/2)\mathbf{M}_{\boldsymbol{\theta}}^{-1}\mathbf{p}_{\boldsymbol{\theta}}(\varepsilon), \tag{25}$$

$$\mathbf{u}(\varepsilon) = \cos(\varepsilon/2)\mathbf{u}(\varepsilon/2) + \sin(\varepsilon/2)\mathbf{p}_{\mathbf{u}}^{**}, \tag{26}$$

$$\mathbf{p}_{\mathbf{u}}(\varepsilon) = \cos(\varepsilon/2)\mathbf{p}_{\mathbf{u}}^{**} - \sin(\varepsilon/2)\mathbf{u}(\varepsilon/2). \tag{27}$$

B The EIS principle

In order to minimize the variance of IS estimates for the likelihood $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}$ of non-Gaussian and/or nonlinear latent variable models, EIS aims at sequentially constructing an IS density which approximates, as closely as possible, the (infeasible) optimal IS density $m^{*}(\mathbf{x}|\boldsymbol{\theta}) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})$, which would reduce the variance of likelihood estimates to zero.

With reference to the likelihood it is assumed that the conditional data density $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ and the prior for the latent variables $p(\mathbf{x}|\boldsymbol{\theta})$ under the latent variable model can be factorized as functions in $\mathbf{x} = (x_1, \ldots, x_D)$ into

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{t=1}^{D} g_t(x_t, \boldsymbol{\delta}), \qquad p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{t=1}^{D} f_t(x_t|\mathbf{x}_{(t-1)}, \boldsymbol{\delta}), \tag{28}$$

where $\mathbf{x}_{(t)} = (x_1, \ldots, x_t)$ with $\mathbf{x}_{(D)} = \mathbf{x}$ and $\boldsymbol{\delta} = (\boldsymbol{\theta}, \mathbf{y})$. Such factorizations can be found for a broad class of models, including dynamic non-Gaussian/nonlinear state-space models for time series, non-Gaussian/nonlinear models with a latent correlation structure for cross-sectional data as well as static hierarchical models without

latent correlation for which $f_t(x_t|\mathbf{x}_{(t-1)}, \boldsymbol{\delta}) = f_t(x_t, \boldsymbol{\delta})$. E.g., variants of EIS for univariate and multivariate linear Gaussian states subject to nonlinear measurements are given in Liesenfeld and Richard (2003, 2006) and for more general nonlinear models in Kleppe et al. (2014); Moura and Turatti (2014). EIS implementations with more flexible IS densities such as mixture of normal distributions are found in Kleppe and Liesenfeld (2014), Scharth and Kohn (2016), Grothe et al. (2019), and Liesenfeld and Richard (2010) use truncated normal distributions. Applications of EIS to models with non-Markovian latent variables for spatial data are provided in Liesenfeld et al. (2016, 2017). In our applications we consider univariate time series models, which is why we use t to index the elements in $\mathbf{x}$ and restrict x_t in (28) to be one-dimensional.

EIS-MC estimation of likelihood functions $p(\mathbf{y}|\boldsymbol{\theta})$ associated with (28) is based upon an IS density m for $\mathbf{x}$ which is decomposed conformably with the factorization in (28) into

$$m(\mathbf{x}|\mathbf{a}) = \prod_{t=1}^{D} m_t(x_t|\mathbf{x}_{(t-1)}, \mathbf{a}_t), \tag{29}$$

with conditional densities m_t such that

$$m_t(x_t|\mathbf{x}_{(t-1)}, \mathbf{a}_t) = \frac{k_t(\mathbf{x}_{(t)}, \mathbf{a}_t)}{\chi_t(\mathbf{x}_{(t-1)}, \mathbf{a}_t)}, \quad \chi_t(\mathbf{x}_{(t-1)}, \mathbf{a}_t) = \int k_t(\mathbf{x}_{(t)}, \mathbf{a}_t) dx_t, \tag{30}$$

where $\mathcal{K} = \{k_t(\cdot, \mathbf{a}_t), \mathbf{a}_t \in \mathcal{A}_t\}$ is a preselected parametric class of density kernels indexed by auxiliary parameters $\mathbf{a}_t$ and with a point-wise computable integrating factor χ_t. As required for the proposed methodology, it is assumed that the IS density (29) can be simulated by sequentially generating draws from the conditional densities (30) using smooth deterministic functions γ_t such that $x_t = \gamma_t(\mathbf{a}_t, v_t)$ for $t = 1, \ldots, D$, where $v_t \sim N(0, 1)$.

From (28)-(30) results the following factorized IS representation of the likelihood:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int \left[\chi_1(\mathbf{a}_1, \boldsymbol{\delta}) \prod_{t=1}^{D} \omega_t(\mathbf{x}_{(t)}, \mathbf{a}_{(t+1)}, \boldsymbol{\delta}) \right] m(\mathbf{x}|\mathbf{a}) d\mathbf{x}, \tag{31}$$

where the period-t IS weight is given by

$$\omega_t(\mathbf{x}_{(t)}, \mathbf{a}_{(t+1)}, \boldsymbol{\delta}) = \frac{g_t(x_t, \boldsymbol{\delta}) f_t(x_t|\mathbf{x}_{(t-1)}, \boldsymbol{\delta}) \chi_{t+1}(\mathbf{x}_{(t)}, \mathbf{a}_{t+1}, \boldsymbol{\delta})}{k_t(\mathbf{x}_{(t)}, \mathbf{a}_t)}, \tag{32}$$

with $\chi_{D+1}(\cdot) \equiv 1$. For any given $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_D) \in \mathcal{A} = \times_{t=1}^{D} \mathcal{A}_t$, the corresponding MC likelihood estimate

is given by

$$\hat{p}(\mathbf{y}|\boldsymbol{\theta},\mathbf{u}) = \omega(\mathbf{x},\mathbf{a}), \qquad \omega(\mathbf{x},\mathbf{a}) = \prod_{t=1}^{D} \omega_t(\mathbf{x}_{(t)},\mathbf{a}_{(t+1)}), \tag{33}$$

where $\mathbf{x}$ is a draw simulated from the sequential IS density $m(\mathbf{x}|\mathbf{a})$ in (29) (which is obtained by transforming $\mathbf{u}$ using the sequence of smooth deterministic functions γ_t).

In order to minimize the MC variance of the likelihood estimate (33), EIS aims at selecting values for the auxiliary parameters $\mathbf{a}$ that minimize period-by-period the MC variance of the IS weights ω_t in (32) with respect to $m(\mathbf{x}|\mathbf{a})$. This requires that the kernels $k_t(\mathbf{x}_{(t)},\mathbf{a}_t)$ as functions in $\mathbf{x}_{(t)}$ provide the best possible fit to the products $g_t(x_t,\boldsymbol{\delta})f_t(x_t|\mathbf{x}_{(t-1)},\boldsymbol{\delta})\chi_{t+1}(\mathbf{x}_{(t)},\mathbf{a}_{t+1})$. For an approximate solution to this minimization problem under the preselected class of kernels $\mathcal{K}$, EIS solves the following back-recursive sequence of least squares (LS) approximation problems:

$$\begin{aligned}(\hat{c}_t,\hat{\mathbf{a}}_t) = \arg\min_{c_t\in\mathbb{R},\mathbf{a}_t\in\mathcal{A}_t} \sum_{i=1}^{r} \Bigg\{ &\log\Big[g_t\big(x_t^{(i)},\boldsymbol{\delta}\big)\; f_t\big(x_t^{(i)}|\mathbf{x}_{(t-1)}^{(i)},\boldsymbol{\delta}\big)\; \chi_{t+1}\big(\mathbf{x}_{(t)}^{(i)},\hat{\mathbf{a}}_{t+1}\big)\Big] \\ &- c_t - \log k_t\big(\mathbf{x}_{(t)}^{(i)},\mathbf{a}_t\big)\Bigg\}^2, \qquad t = D, D-1,\ldots,1,\end{aligned} \tag{34}$$

where c_t represents an intercept, and $\{\mathbf{x}^{(i)}\}_{i=1}^r$ denote r iid draws simulated from $m(\mathbf{x}|\mathbf{a})$ itself. Thus, the EIS-optimal values for the auxiliary parameters $\hat{\mathbf{a}}$ result as a fixed-point solution to the sequence $\{\hat{\mathbf{a}}^{[0]},\hat{\mathbf{a}}^{[1]},\ldots\}$ in which $\hat{\mathbf{a}}^{[j]}$ is given by (34) under draws from $m(\mathbf{x}|\hat{\mathbf{a}}^{[j-1]})$. In order to ensure convergence to a fixed-point solution it is critical that all the $\mathbf{x}$ draws simulated for the sequence $\{\hat{\mathbf{a}}^{[j]}\}$ be generated by using the smooth deterministic functions γ_t to transform a *single set* of rD Common Random Numbers (CRNs), say $\mathbf{z} \sim N(\mathbf{0}_{rD},\mathbf{I}_{rD})$. To initialize the fixed-point iterations $j = 0,\ldots,J$, the starting value $\hat{\mathbf{a}}^{[0]}$ can be found, e.g., from an analytical local approximation (such as Laplace) of the EIS targets $\ln(g_t f_t \chi_{t+1})$ in (34). Convergence of the iterations to a fixed-point solution is typically fast to the effect that a value for the number of iterations J between 2 and 4 often suffices to produce a (close to) optimal solution (Richard and Zhang, 2007). The MC-EIS likelihood estimate, for a given $\boldsymbol{\theta}$, is then calculated by substituting in (33) the EIS-optimal value $\hat{\mathbf{a}}$ for $\mathbf{a}$. In order to highlight its dependence on $\boldsymbol{\theta}$ and $\mathbf{z}$ we shall use $\hat{\mathbf{a}} = \mathbf{a}(\boldsymbol{\theta},\mathbf{z})$ to denote the EIS-optimal value.

The selection of the parametric class $\mathcal{K}$ of EIS density kernels k_t is inherently specific to the latent variable model under consideration as those kernels are meant to provide a functional approximation in $\mathbf{x}_{(t)}$ to the product $g_t f_t \chi_{t+1}$. In the applications below, we consider models with data densities g_t which are log-concave in x_t and Gaussian conditional densities for x_t with a Markovian structure so that $f_t(x_t|\mathbf{x}_{(t-1)},\boldsymbol{\delta}) =$

$f_t(x_t|x_{t-1}, \boldsymbol{\delta})$. This suggests selection of the k_t's as Gaussian kernels and to exploit that such kernels are closed under multiplication in order to construct the k_t's as the following parametric extensions of the prior densities f_t:

$$k_t(x_t, x_{t-1}, \mathbf{a}_t) = f_t(x_t|x_{t-1}, \boldsymbol{\delta})\xi_t(x_t, \mathbf{a}_t), \tag{35}$$

where ξ_t is a Gaussian kernel in x_t of the form $\xi_t(x_t, \mathbf{a}_t) = \exp\{a_{1t}x_t + a_{2t}x_t^2\}$ with $\mathbf{a}_t = (a_{1t}, a_{2t})$. In this case the EIS approximation problems (34) take the form of simple *linear* LS-problems where $\log[g_t(x_t^{(i)}, \boldsymbol{\delta})$ $\chi_{t+1}(x_t^{(i)}, \hat{\mathbf{a}}_{t+1})]$ are regressed on a constant, $x_t^{(i)}$ and $[x_t^{(i)}]^2$. In fact, (34) reduces to linear LS regressions for all kernels k_t chosen within the exponential family (Richard and Zhang, 2007), which simplifies implementation. However, it is important to note that EIS is by no means restricted to the use of IS densities from the exponential family nor to models with low-order Markovian specifications for the latent variables.

The EIS approach as outlined above differs from standard IS in that it uses IS densities whose parameters $\hat{\mathbf{a}} = \mathbf{a}(\boldsymbol{\theta}, \mathbf{z})$ are (conditional on $\boldsymbol{\theta}$) random variables as they depend via the EIS fixed-point repressions (34) on the CRNs $\mathbf{z}$. This calls for specific rules for implementing EIS which ensure that the resulting MC likelihood estimates meet the qualifications needed for their use within PM-HMC. In order to ensure that the EIS likelihood estimate (33) based on the random numbers $\mathbf{u}$ is unbiased the latter need to be a set of random draws different from the CRNs $\mathbf{z}$ used to find $\hat{\mathbf{a}}$ (Kleppe and Liesenfeld, 2014). Note also that since $\hat{\mathbf{a}}$ is an implicit function of $\boldsymbol{\theta}$, maximal accuracy requires us to rerun the EIS fixed-point regressions for any new value of $\boldsymbol{\theta}$. In order to ensure that the resulting EIS likelihood estimate (33) as a function of $\hat{\mathbf{a}}$ is smooth in $\boldsymbol{\theta}$, $\hat{\mathbf{a}}$ itself needs to be a smooth function of $\boldsymbol{\theta}$. This can be achieved by presetting the number of fixed-point iterations J across all $\boldsymbol{\theta}$-values to a fixed number, rather than using a stopping rule based on a relative-change threshold.

The EIS-specific tuning parameters are the number of $\mathbf{x}^{(i)}$-draws r used to run the EIS optimization process, the number of fixed-point iterations on the EIS regressions J, and the number of $\mathbf{x}^{(i)}$-draws n for the likelihood estimate (33). Those parameters should be selected to balance the trade-off between EIS computing time and the quality of the resulting EIS density with respect to the MC accuracy. In particular, for r it is recommended to select it as small as possible while retaining the EIS fixed-point regressions numerically stable and the parameter J should be set such that it is guaranteed that the fixed-point sequence $\{\mathbf{a}^{[j]}\}_j$ approximately converge for the $\boldsymbol{\theta}$ values in the relevant range of the parameter space. In our applications, where the selected class of kernels $\mathcal{K}$ imply that the EIS regressions are linear in the EIS parameters $\mathbf{a}_t$, we find that a J set equal to 1 or 2 and an r about 2 times the number of parameters in $(\mathbf{a}_t, c_t)$ suffice. We obtain EIS kernels k_t providing highly accurate approximations to the targeted product

$g_t f_t \chi_{t+1}$, with an R^2 of the EIS regressions in the final iteration typically larger than 0.95.

C Details related to the example models in Section 5

C.1 SV model

For the SV model, the standard prior assumptions for the parameters $\boldsymbol{\theta} = (\gamma, \delta, \nu)$ are the following: for γ we use a flat prior, for $(\delta + 1)/2$ a Beta prior $\mathcal{B}(\alpha, \beta)$ with $\alpha = 20$ and $\beta = 1.5$, and for ν^2 a scaled inverted-χ^2 prior $p_0 s_0 / \chi^2_{(p_0)}$ with $p_0 = 10$ and $s_0 = 0.01$. For numerical stability we use the parametrization $\boldsymbol{\theta}^* = (\gamma, \operatorname{arctanh} \delta, \log \nu^2)$ together with the priors for $\boldsymbol{\theta}^*$ to run the HMC algorithms, where the priors are derived from those on $\boldsymbol{\theta}$.

For the Laplace transport map, $\mathbf{G}_{\boldsymbol{\theta}}^{(0)}$ and $\mathbf{h}_{\boldsymbol{\theta}}^{(0)}$ are taken to be identical to (6,7). More refined solutions are found using Newton iterations;

$$\mathbf{h}_{\boldsymbol{\theta}}^{(k)} = \mathbf{h}_{\boldsymbol{\theta}}^{(k-1)} + \left[\nabla^2_{\mathbf{x}} \log \left[p(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{(k-1)}}\right]^{-1} \left\{\nabla_{\mathbf{x}} \log \left[p(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{(k-1)}}\right\},$$
$$\mathbf{G}_{\boldsymbol{\theta}}^{(k)} = \nabla^2_{\mathbf{x}} \log \left[p(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{(k-1)}}.$$

for $k = 1, 2, \ldots, K$. Further modifications, including changing to $\mathbf{G}_{\boldsymbol{\theta}}^{(k)} = \nabla^2_{\mathbf{x}} \log \left[p(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{(k)}}$ (at the cost of one additional Cholesky factorization), or keeping $\mathbf{G}_{\boldsymbol{\theta}}^{(k)} = \mathbf{G}_{\boldsymbol{\theta}}^{(0)}$ (costs only a single Cholesky factorization) both in the transport map and as the scaling matrix in the Newton iterations was tried, but did not produce better results.

It is straight forward to show that $\mathbf{G}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}} = 0.5\mathbf{I}_D$ is also the Fisher information of $p(\mathbf{y}|\mathbf{x})$ with respect to $\mathbf{x}$ (i.e. $p(\mathbf{y}|\mathbf{x})$ is a constant information parameterization). Hence also Stan-Laplace $K = 0$ may be interpreted as a special case of DRHMC (Kleppe, 2019).

C.2 Gamma model

For the Gamma model, the priors on the parameters $\boldsymbol{\theta} = (\tau, \beta, \delta, \nu)$ are as follows; we use flat priors for $\log \tau$ as well as $\log \beta$, a Beta $\mathcal{B}(\alpha, \beta)$ with $\alpha = 20$ and $\beta = 1.5$ for $(\delta + 1)/2$, and a scaled inverted-χ^2 for ν^2 with $p_0 s_0 / \chi^2_{(p_0)}$ and $p_0 = 10$, $s_0 = 0.01$. For the LD computations we use the parameterization $\boldsymbol{\theta}^* = (\log \tau, \log \beta, \operatorname{arctanh} \delta, \log \nu^2)$.

For this model, the same strategy for calculating the Laplace transport map as for the SV model was used. Notice that here $\mathbf{G}_{\boldsymbol{\theta},\mathbf{y}|\mathbf{x}} = \tau^{-1}\mathbf{I}_D$ is also the Fisher information of $p(\mathbf{y}|\mathbf{x})$ with respect to $\mathbf{x}$. Hence, Stan-Laplace, $K = 0$ may be interpreted as a DRHMC method.

	μ_1	μ_2	μ_3	μ_4	μ_5	δ_1	δ_2	δ_3	δ_4	δ_5
post. mean	4.16	4.12	3.72	4.11	3.53	0.97	0.98	0.96	0.94	0.96
post. std.	0.2	0.25	0.15	0.1	0.13	0.005	0.004	0.006	0.008	0.006
	σ_1	σ_2	σ_3	σ_4	σ_5	ν				
post. mean	0.31	0.26	0.29	0.28	0.25	33.61				
post. std.	0.009	0.008	0.009	0.009	0.009	0.283				
	$h_{2,1}$	$h_{3,1}$	$h_{4,1}$	$h_{5,1}$	$h_{3,2}$	$h_{4,2}$	$h_{5,2}$	$h_{4,3}$	$h_{5,3}$	$h_{5,4}$
post. mean	0.39	0.29	0.29	0.23	0.20	0.17	0.12	0.22	0.18	0.11
post. std.	0.003	0.003	0.003	0.002	0.003	0.003	0.002	0.004	0.003	0.002
	$x_{1,1}$	$x_{2,1}$	$x_{3,1}$	$x_{4,1}$	$x_{5,1}$	$u_{1,1}$	$u_{2,1}$	$u_{3,1}$	$u_{4,1}$	$u_{5,1}$
post. mean	5.23	5.28	4.27	5.46	5.11	-0.08	-0.05	-0.07	-0.10	-0.10
post. std.	0.206	0.195	0.198	0.205	0.202	1.022	0.993	0.99	1.031	1.049

Table 5: Posterior mean and standard deviations for the inverse Wishart model (14-16) based on Stan-Laplace, $K = 0$. All figures are means across 8 independent replica. Here, $u_{s,1}$ is the first element in $\mathbf{u}_s$, and should be close to standard normal when the transport map produces a sufficient de-coupling effect.

C.3 CEV model

For the CEV model, for α and β we assume Gaussian priors both with $N(0, 1000)$, for γ a uniform prior on the interval $[0, 4]$, and for σ_x^2 and σ_y^2 uninformative inverted-χ^2 priors with $p(\sigma_x^2) \propto 1/\sigma_x^2$ and $p(\sigma_y^2) \propto 1/\sigma_y^2$. The LD computations are conducted on the following transformed parameters: $\boldsymbol{\theta}^* = (\alpha, \beta, \gamma, \log \sigma_x^2, \log \sigma_y^2)$.

For the CEV model, the precision of the latent state prior is does not have closed-form, which precludes the application of (6,7). However, it is known that the measurement densities has a very small variance, hence $\mathbf{h}_{\boldsymbol{\theta}}^{(0)} = \mathbf{y}$ seems sensible. Subsequently, a full Newton iteration is performed:

$$\begin{aligned}
\mathbf{h}_{\boldsymbol{\theta}}^{(k)} &= \mathbf{h}_{\boldsymbol{\theta}}^{(k-1)} + \left[\nabla_{\mathbf{x}}^2 \log\left[p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{(k-1)}}\right]^{-1} \left\{\nabla_{\mathbf{x}} \log\left[p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{(k-1)}}\right\}, \\
\mathbf{G}_{\boldsymbol{\theta}}^{(k)} &= \nabla_{\mathbf{x}}^2 \log\left[p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{(k-1)}}.
\end{aligned}$$

for $k = 1, 2, \ldots, K$. Further modifications, including changing to $\mathbf{G}_{\boldsymbol{\theta}}^{(k)} = \nabla_{\mathbf{x}}^2 \log\left[p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{(k)}}$ (at the cost of one additional Cholesky factorization) did not improve the fit sufficiently to warrant the additional computation.

D Details related to the realized volatility model in Section 6

The (normalized) observation density is given by:

$$p(\mathbf{Y}_t|\Sigma_t, \nu) = \frac{|\boldsymbol{\Sigma}_t|^{\frac{\nu}{2}}}{2^{\frac{\nu r}{2}} \pi^{\frac{r(r-1)}{4}} \prod_{s=1}^{r} \Gamma\left([\nu + 1 - s]/2\right)} |\mathbf{Y}_t|^{-\frac{\nu+r+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_t \mathbf{Y}_t^{-1}\right]\right).$$

In the Stan implementation, $\prod_{t=1}^{D} |\mathbf{Y}_t|$ and $\mathbf{Y}_t^{-1}$, $t = 1, \ldots, D$ where precomputed.

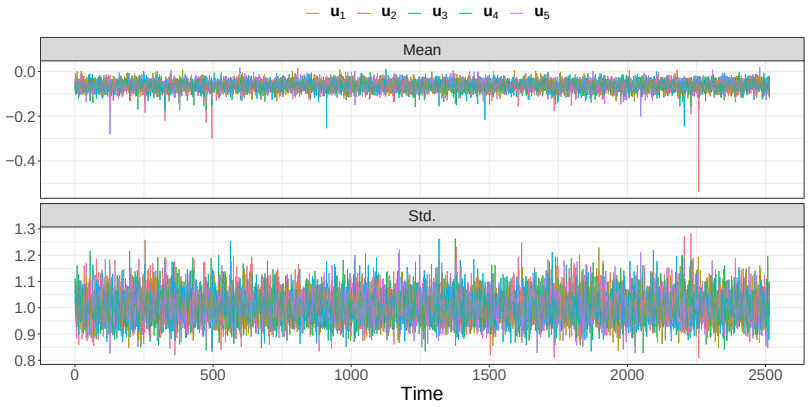

Figure 2: Posterior mean and standard deviation of $\mathbf{u}_s$, $s = 1, \ldots, 5$, for the inverse Wishart model (14-16) under Laplace transport map with $K = 2$.

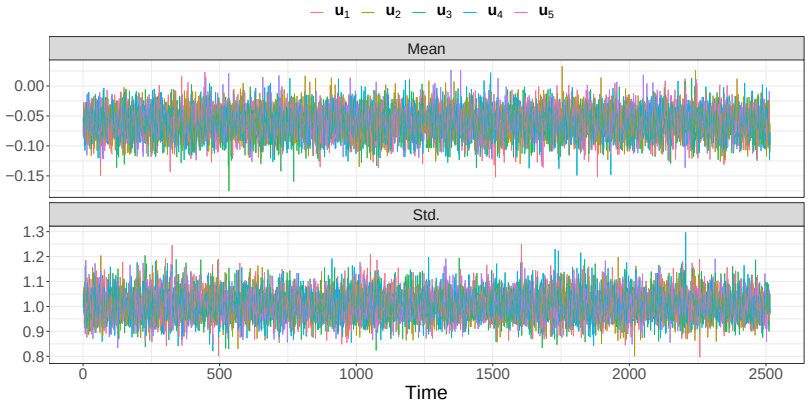

Figure 3: Posterior mean and standard deviation of $\mathbf{u}_s$, $s = 1, \ldots, 5$, for the inverse Wishart model (14-16) under Laplace transport with $K = 10$.

The (independent) priors used to complete the model specification in Section 6.1 are as follows: $\mu_s \sim N(0, 25)$, $\delta_s \sim$ uniform$(-1, 1)$, $\sigma_s^2 \sim p_0 s_0 / \chi^2_{p_0}$ where $p_0 = 4$ and $s_0 = 0.25$, $h_{i,j} \sim N(0, 100)$. Finally, a flat prior on $(6.0, \infty)$ was chosen for ν.

Posterior- means and standard deviations of the parameters and the first elements in $\mathbf{x}_s$ and $\mathbf{u}_s$ are given in Table 5. The results are very much in line with those of Grothe et al. (2019).

The Laplace transport maps for each of $\mathbf{x}_s$, $s = 1, \ldots, r$ are constructed as follows; the initial guesses for $\mathbf{h}_{\boldsymbol{\theta}}^{(0)}$ and $\mathbf{G}_{\boldsymbol{\theta}}^{(0)}$ are those given in (6,7), applied to (17). The mean is further refined via the following approximate Newton iteration

$$\mathbf{h}_{\boldsymbol{\theta}}^{(k)} = \mathbf{h}_{\boldsymbol{\theta}}^{(k-1)} + \left[\mathbf{G}_{\boldsymbol{\theta}}^{(0)}\right]^{-1} \left\{ \nabla_{\mathbf{x}} \log \left[p(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \right]_{\mathbf{x}=\mathbf{h}_{\boldsymbol{\theta}}^{(k-1)}} \right\},$$

whereas $\mathbf{G}_{\boldsymbol{\theta}}^{(k)} = \mathbf{G}_{\boldsymbol{\theta}}^{(0)}$ is kept fixed which result in that only a single Cholesky factorization is required. Figures 2,3 show the posterior mean and standard deviations of $\mathbf{u}_s$ over time t for Stan-Laplace, $K = 2$ and $K = 10$ respectively. It is seen that even with the approximate Newton iteration, the iteration makes $\mathbf{u}_s$ have a mean close to zero, where the remaining deviation from zero for $K = 10$ iterations in Figure 3 is presumably due to the non-quadratic nature of the "measurement density" in (17) (in addition to Monte Carlo variation).

References

Grothe, O., T. S. Kleppe, and R. Liesenfeld (2019). The Gibbs sampler with particle efficient importance sampling for state-space models. *Econometric Reviews 38*(10), 1152–1175.

Kleppe, T. S. (2019). Dynamically rescaled Hamiltonian Monte Carlo for Bayesian hierarchical models. *Journal of Computational and Graphical Statistics 28*(3), 493–507.

Kleppe, T. S. and R. Liesenfeld (2014). Efficient importance sampling in mixture frameworks. *Computational Statistics & Data Analysis 76*, 449 – 463.

Kleppe, T. S., J. Yu, and H. J. Skaug (2014). Maximum likelihood estimation of partially observed diffusion models. *Journal of Econometrics 180*(1), 73 – 80.

Liesenfeld, R. and J.-F. Richard (2003). Univariate and multivariate stochastic volatility models: estimation and diagnostics. *Journal of Empirical Finance 10*(4), 505–531.

Liesenfeld, R. and J.-F. Richard (2006). Classical and Bayesian analysis of univariate and multivariate stochastic volatility models. *Econometric Reviews 25*(2-3), 335–360.

Liesenfeld, R. and J.-F. Richard (2010). Efficient estimation of probit models with correlated errors. *Journal of Econometrics 156*(2), 367–376.

Liesenfeld, R., J.-F. Richard, and J. Vogler (2016). Likelihood evaluation of high-dimensional spatial latent gaussian models with non-gaussian response variables. In *Spatial Econometrics: Qualitative and Limited Dependent Variables*, pp. 35–77. Emerald Group Publishing Limited.

Liesenfeld, R., J.-F. Richard, and J. Vogler (2017). Likelihood-based inference and prediction in spatio-temporal panel count models for urban crimes. *Journal of Applied Econometrics 32*(3), 600–620.

Lindsten, F. and A. Doucet (2016). Pseudo-Marginal Hamiltonian Monte Carlo. arXiv preprint arXiv:1607.02516.

Moura, G. V. and D. E. Turatti (2014). Efficient estimation of conditionally linear and Gaussian state space models. *Economics Letters 124*(3), 494 – 499.

Richard, J.-F. and W. Zhang (2007). Efficient high-dimensional importance sampling. *Journal of Econometrics 141*(2), 1385–1411.

Scharth, M. and R. Kohn (2016). Particle efficient importance sampling. *Journal of Econometrics 190*(1), 133 – 147.

Paper IV

Estimating the Competitive Storage Model with Stochastic Trends in Commodity Prices

Estimating the Competitive Storage Model with Stochastic Trends in Commodity Prices

Kjartan Kloster Osmundsen*[1], Tore Selland Kleppe[1], Roman Liesenfeld[2], and Atle Oglend[3]

[1]Department of Mathematics and Physics, University of Stavanger, Norway
[2]Institute of Econometrics and Statistics, University of Cologne, Germany
[3]Department of Safety, Economics and Planning, University of Stavanger, Norway

January 14, 2020

Abstract

We propose a state-space model (SSM) for commodity prices that combines the competitive storage model with a stochastic trend. This approach fits into the economic rationality of storage decisions, and adds to previous deterministic trend specifications of the storage model. Parameters are estimated using a particle Markov chain Monte Carlo procedure. Empirical application to four commodity markets shows that the stochastic trend SSM is favored over deterministic trend specifications. The stochastic trend SSM identifies structural parameters that differ from those for deterministic trend specifications. In particular, the estimated price elasticities of demand are significantly larger under the stochastic trend SSM.

Keywords: Commodity price dynamics; Bayesian posterior analysis; Particle marginal Metropolis-Hastings; State-space model.

1 Introduction

Economic theories are often developed in a stationary context. However, the real world does not always correspond to stationarity. This potential mismatch creates a challenge when attempting to relate theory to historical data. This is a well-known problem in empirical macroeconomics, where structural parameters of business cycle models are often estimated on data that have been filtered in order to remove variation at frequencies that the model is not intended to explain, such as low-frequency trend variations and seasonal fluctuations (DeJong and Dave, 2011; Sala, 2015). For an overview of alternatives to the use of pre-filtered data in order to address this general problem, see Canova (2014).

*Corresponding author. Email: kjartan.osmundsen@gmail.com

In the competitive storage model for commodity prices introduced by Gustafson (1958), the situation is similar to that of business cycle models. The rational expectations equilibrium implied by the solution of this model is only known to exist in a stationary market. Accordingly, it is a model for describing dynamic price adjustments towards an exogenously given fixed steady-state equilibrium. However, it cannot explain low-frequency price movements due to persistent shocks. This is problematic when attempting to estimate the structural parameters of the model using commodity price data, since time series of commodity prices typically display a strongly persistent behavior in the price level, so that non-stationarity cannot be rejected when using conventional statistical tests (Wang and Tomek, 2007; Gouel and Legrand, 2017). As a result, the estimates for the structural parameters, which determine quantities like the price elasticity of demand and storage costs, are likely to be biased. This issue was recognized by Deaton and Laroque (1995) in one of the earliest attempts to directly estimate the structural parameters of the storage model.

This paper proposes an approach to estimate the structural parameters of the competitive commodity storage model using a state-space model (SSM) for commodity prices, which decomposes the observed price into a stationary component which is due to the storage model and a stochastic trend component included to capture low-frequency price variations the storage model is unable to explain. Using a stochastic trend specification to account for non-stationary price data, our empirical approach aims at fitting into the economic rationality of the stationary storage model so that it preserves theoretical coherence, promising meaningful estimates of the structural parameters. Such a fit results from the fact that a stochastic trend that scales equilibrium prices can be isolated in the storage model by assuming that the innovations to the trend do not interfere with the agents' equilibrium storage decisions. In the baseline storage model, unrestricted equilibrium storage decisions lead to an intertemporal pricing restriction of the form $P_t = \beta E_t(P_{t+1})$, where $E_t(P_{t+1})$ is the rational period-t expectation of the commodity price P_{t+1} and β represents some discount factor. Thus, a stochastic price scaling K_t will not impair the equilibrium storage decisions if $K_t P_t = \beta E_t(K_{t+1} P_{t+1})$. This generically identifies stochastic trends as shifts in the price levels that do not interfere with intertemporal stock allocations, allowing a coherent integration of the stationary rational expectations equilibrium into a non-stationary environment, thus providing the theoretical basis of our empirical SSM approach. The corresponding SSM, that jointly identifies the trend parameters and the structural parameters of the storage model, is non-linear in the latent states so that its likelihood function is not available in closed form. To overcome this difficulty, we propose to use a Bayesian posterior analysis based on a particle Markov chain Monte Carlo (PMCMC) procedure (Andrieu et al., 2010).

With our proposed approach we contribute to the literature concerned with the general problem of adapting stationary economic models to non-stationary data, and more specifically to the problem of estimating the structural parameters of the competitive storage model on non-stationary commodity price data. Legrand

(2019) identifies reliable estimation as one of the main issues of structural models for commodity prices. Early attempts of estimating the structural parameters revealed that fitted competitive storage models are not able to satisfactorily approximate the observed strong serial dependence in commodity price data, indicating misspecification of the empirical model and casting doubt on the reliability of the parameter estimates (Deaton and Laroque, 1995). Suggested solutions to this problem include ad-hoc enrichments of the dynamic structure of the storage model by including weakly dependent supply shocks (Deaton and Laroque, 1996; Kleppe and Oglend, 2017), or the tuning of the grid for the commodity stock state variable, used for approximating the policy function (Cafiero et al., 2011). Other approaches replace the estimation techniques applied in early empirical implementations of the storage model, like the pseudo maximum likelihood (ML) procedure of Deaton and Laroque (1996), by more sophisticated ones, such as the ML technique developed by Cafiero et al. (2015) or the particle filtering methods proposed in Kleppe and Oglend (2017).

Empirical approaches that, like ours, decompose the observed price into a component to be explained by the storage model and a trend component are those of Cafiero et al. (2011), Bobenrieth et al. (2013), Guerra et al. (2015) and Gouel and Legrand (2017). The first three of these studies propose to account for the strong persistence in the price data that the storage model is not able to approximate, by detrending the prices using a deterministic log-linear trend prior to the estimation of the structural parameters. Gouel and Legrand (2017) improves upon this procedure by jointly estimating the structural and deterministic trend parameters using the ML-estimator of Cafiero et al. (2015). The trend specifications Gouel and Legrand (2017) consider in their empirical application include log-linear trends as well as more flexible trends specified as restricted cubic splines. One of their main findings is that empirical models accounting for a properly specified trend component in the observed commodity price yield more plausible estimates of the structural parameters than models without a trend. However, the deterministic trends used in those studies inherently imply well predictable capital gains in the storage model, and so question the economic logic of separating the trend from structural economic pricing components. Moreover, the appropriate functional form of the deterministic trend needs to be tailored to the specific commodity market and the sampling frequency for which the storage models are applied. In contrast, the stochastic trend as used in our SSM approach represents, in Bayesian terms, a hierarchical prior for the low-frequency price component, which is not only consistent with the rationality of the economic model, but also flexible in its design to account for variation that the storage model is not intended to explain. This makes our approach applicable to a broad range of commodity markets and different sampling frequencies. The strategy of scaling prices to address non-stationarity was also done by Routledge et al. (2000) in their equilibrium term structure model of crude oil futures. However, they did not do so in a rigorous estimation framework.

A stochastic trend as used in our storage SSM allows a potentially large fraction of the observed variation

in commodity prices to be accounted for by the trend component. This risks miss-assigning price variation due to speculative storage to the trend component. Thus, if considered as an evaluation of the empirical relevance of the storage model, the use of a stochastic trend can be considered as a conservative test. To explore this issue further we perform a simulation experiment. The simulation results suggest that our proposed approach is able to accurately assign price variation to trend and model components. We further apply our storage SSM to monthly observations of nominal coffee, cotton, aluminum and natural gas prices. The results show, not surprisingly, that most of the observed price variation is due to the stochastic trend component. In order to assess the empirical relevance of the competitive storage model, we compare the storage SSM to the nested model that results in the absence of storage. The comparison reveals that the storage model predicting non-linear price dynamics with episodes of isolated price spikes and increased volatility adds significantly to explaining the observed commodity price behavior. We also compare the stochastic trend SSM to the deterministic trend models of Gouel and Legrand (2017) by using the Bayes factor and a model residual analysis. Results show that the SSM with stochastic trend fits the price data much better than models with deterministic trends. The estimates for the price elasticity of demand obtained from the stochastic trend SSM are substantially larger than for the deterministic trend models. Also, the estimated storage costs vary considerably depending on the commodity. This highlights the importance of properly accounting for the trend behavior when evaluating the role of speculative storage in commodity markets.

The rest of this paper is structured as follows. In the next section, we present the storage model used in the paper and the assumed price representation. We then present the estimation methodology (Section 3), simulation results (Section 4) and empirical results for historical data (Section 5). We discuss the findings before we offer some concluding remarks (Section 6).

2 Storage Model

2.1 State-Space Formulation with Stochastic Trend

Our approach relies upon the commodity storage model of Oglend and Kleppe (2017). It extends the Deaton and Laroque (1992) model by including an upper limit of storage capacity, $C \geq 0$, in addition to the conventional non-negativity constraint for stocks, so that the storage space is completely bounded. This upper limit takes into account possible congestion of the storage infrastructure, which can lead to negative price spikes in the event of substantial oversupply in the market. In addition, the assumption of a completely bounded storage space allows numerical solutions of the model that are more robust over a wider parameter

range than those for a model without this assumption (Oglend and Kleppe, 2017). This, in turn, simplifies estimation of the model parameters.

The economic model of commodity storage is a canonical dynamic stochastic partial equilibrium model in discrete time for a commodity market with risk neural storage agents and rational expectations. The rational expectations equilibrium is characterized by a price function, denoted by $f(x)$, which maps the stocks x to commodity prices. For empirical implementation, we assume that the observed commodity price can be decomposed into a component to be explained by the commodity storage model and a stochastic trend component. The corresponding time series model that we propose for the commodity log-price p_t, observed at time t $(t = 1, \ldots, T)$, has the form

$$p_t = k_t + \log f(x_t), \tag{1}$$

$$k_t = k_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \text{iid } N(0, v^2), \tag{2}$$

$$x_t = (1-\delta)\sigma(x_{t-1}) + z_t, \qquad z_t \sim \text{iid } N(0,1), \tag{3}$$

where the available quantity of commodity stocks x_t is treated as a latent state variable. Its dynamics are linear in the equilibrium storage policy $\sigma(\cdot)$, with stock depreciation rate δ and Gaussian supply shocks z_t. The latent trend component of the log-price k_t is specified as a driftless Gaussian random walk, so that it is allowed to vary gradually over time. The innovations of this stochastic trend ε_t and the supply shocks z_t are assumed to be serially and mutually independent.

The rational expectations equilibrium price function $f(x)$ satisfies for all x

$$f(x) = \min\left\{P(x-C), \max\left[\bar{f}(x), P(x)\right]\right\}, \tag{4}$$

$$\bar{f}(x) = \beta \int f\Big((1-\delta)\,\sigma(x) + z\Big)\phi(z)dz, \tag{5}$$

$$\sigma(x) = x - D(f(x)), \tag{6}$$

where $D(p)$ represents a continuous and monotonically decreasing aggregate demand function in the market, $P(x)$ is the corresponding inverse demand, and $\phi(z)$ is the probability density function of the supply shock z. The storage cost discount factor is given by $\beta = (1-\delta)/(1+r)$, where r is a relevant interest rate. According to Equation (4), the equilibrium pricing function exhibits three different pricing regimes: (i) a stock-out pricing regime, where $f(x) = P(x) \Leftrightarrow \sigma(x) = 0$, (ii) a no-arbitrage pricing regime, i.e. $f(x) = \bar{f}(x) \Leftrightarrow C > \sigma(x) > 0$, where $\bar{f}(x)$ is the expected next period commodity price, and (iii) a full capacity pricing regime, where $f(x) = P(x-C) \Leftrightarrow \sigma(x) = C$. The stock-out regime is characterized by positive price spiking and high price

volatility due to reduced shock buffering capabilities in the market. Under the no-arbitrage regime, prices evolve smoothly with a relatively low volatility. Full capacity pricing mirrors the stock-out regime but with negative price spikes. As the market transitions between regimes, prices move between periods of quiet and turmoil, generating non-linear dynamics in the price process. The rational expectations equilibrium $f(x)$ is stationary, having an associated globally stationary price density (Oglend and Kleppe, 2017).

Using $k_t = p_{t-1} - \log f(x_{t-1}) + \varepsilon_t$ in the price equation, the model as given in Equations (1)-(3) can be written as

$$p_t = p_{t-1} + \log\left(\frac{f(x_t)}{f(x_{t-1})}\right) + \varepsilon_t, \qquad \varepsilon_t \sim \text{iid } N(0, v^2), \tag{7}$$

$$x_t = (1-\delta)\sigma(x_{t-1}) + z_t, \qquad z_t \sim \text{iid } N(0,1). \tag{8}$$

This defines a non-linear Gaussian state-space model, with measurement equation (7) for the observed price and state-transition equation (8) for the latent stocks.

2.2 Stochastic Trends and Storage Decisions

Separating the trend from the storage model pricing component in a consistent way that does not compromise the rationality of storage agents in the market requires that the trend does not interfere with intertemporal allocation incentives. The martingale property of a trend component specified as a stochastic trend with innovations that are independent of supply shocks ensures that this requirement is met. By using a separable stochastic trend we are assuming that storage agents do not alter their storage decisions based on trend innovations. In other words, trend innovations are assumed perceived by agents as permanent scalings of price levels that do not warrant adjustments to storage allocations.

As an example, consider permanent shocks K to the inverse aggregate demand in the market, $P^* = KP$. The aggregate demand implied by P^* is D^*, and the resulting rational expectations equilibrium is given by

$$f^*(x) = \min\left\{P^*(x-C), \max\left[\beta \int f^*\Big((1-\delta)\left(x - D^*\left(f^*(x)\right)\right) + z\Big)\phi(z)dz, P^*(x)\right]\right\}. \tag{9}$$

Assume the scaling process is given by $K' = \gamma K + \epsilon$, where ϵ is a random variable with density ϕ_ϵ which is independent of the supply shock z. This scaling does not affect the optimal storage policy if $f^*(x) = Kf(x)$ solves the functional equation problem in Equation (9), where $f(x)$ is the rational expectations equilibrium

for the original non-scaled prices. Substituting the proposed solution for $f^*(x)$, we get

$$Kf(x) = \min\Bigg\{KP(x-C), \max\left[\beta \int (\gamma K + \epsilon) \int f\Big((1-\delta)(x - D^*(Kf(x))) + z\Big)\phi(z)\phi_\epsilon(\epsilon)dzd\epsilon, KP(x)\right]\Bigg\}. \tag{10}$$

Note that $D^*(Kf(x)) = D(f(x))$ by the definition of D^* as the inverse of the scaled inverse demand function $P^* = KP$, where $P = f(x)$. And so,

$$Kf(x) = \min\Bigg\{KP(x-C), \max\left[\beta \int (\gamma K + \epsilon) \int f\Big((1-\delta)(x - D(f(x))) + z\Big)\phi(z)\phi_\epsilon(\epsilon)dzd\epsilon, KP(x)\right]\Bigg\}. \tag{11}$$

If $\int (\gamma K + \epsilon)\, \phi_\epsilon(\epsilon)d\epsilon = K$ implying that $E(K') = K$, we obtain

$$Kf(x) = K \min\left\{P(x-C), \max\left[\beta \int f\Big((1-\delta)(x - D(f(x))) + z\Big)\phi(z)dz, P(x)\right]\right\}, \tag{12}$$

which establishes $f^*(x) = Kf(x)$ as the solution to the inverse demand scaled rational expectations equilibrium. Consequently, any observed commodity price can be represented as $P = Kf(x)$. Formally, the rational expectations equilibrium is linear homogeneous to the proportional scaling K of the inverse aggregate demand function when $E(K') = K$ and innovations to the trend are orthogonal to supply shocks.

Note that in our econometric model as given by Equations (1)-(3), it is the logarithm of the scaling process for the price levels $k = \log(K)$ and not K, for which we assume a stochastic trend. Hence, the martingale property for the scaling term process K will not apply exactly, and the assumed Gaussian process for $\log(K)$ implies that $E(K') = K \exp\left(v^2/2\right) > K$. By ignoring this bias in our econometric model we make the behavioral assumption that agents do not alter storage decision based on the capital gain due to the expected mark-up factor $\exp\left(v^2/2\right) > 1$. We consider this a reasonable trade-off to allow us to empirically analyze the storage model within a log-linear state and measurement space, which is comparatively convenient for statistical inference. In addition, the bias is small when v^2 is small, that is, when the trend is fairly smooth. In fact, the estimates we obtain for v in our empirical application discussed below imply that the factor $\exp\left(v^2/2\right)$ varies in a range between 1.001 and 1.004 so that it is essentially negligible. Ignoring this factor is essentially equivalent to transforming the probability space to a setting where agents ignore information from trend innovations, similar to a risk-neutral valuation setting where v^2 defines a required risk premium term or a nominal inflation term.

3 Statistical Inference

3.1 Preliminaries

In our empirical application of the storage model with stochastic trend based on its state-space representation as given in Equations (7) and (8), we use monthly commodity spot prices and rely on a Bayesian Markov chain Monte Carlo (MCMC) posterior analysis. For this application, we follow Kleppe and Oglend (2017) and use $P(x) = \exp(-bx)$ as the inverse demand function, where the parameter b measures the semi-elasticity of the demand price. In line with Gouel and Legrand (2017), we fix the yearly interest rate at 5%, so that the monthly storage cost discount factor is given by $\beta = (1-\delta)/(1+r)$, with $r = 1.05^{1/12} - 1$. The set of parameters then consists of the structural parameters (δ, b, C) and the trend parameter v.

In initial experiments to estimate the parameters, we found that the capacity limit C is empirically not well identified separately from the remaining parameters. This appears to be mainly due to the fairly small sample size of our data, ranging from 264 to 360 monthly spot price observations. Therefore, we decided to fix C at a positive predetermined value. Since C determines the full capacity threshold for equilibrium storage, it bounds the space for the unit-free latent state variable x, and by fixing its value (together with normalizing the mean of z to zero) we pin down the range of this space. The values of the remaining parameters (δ, b, v) and their implications for the price dynamics are then to be interpreted relative to this scale of x. In our empirical application below we set the capacity limit $C = 10$. This ensures that it is a fairly rare event for the market to be in the full-capacity regime. Suppose, for example, that all realizations of the supply shocks z for a sequence of periods are equal to one standard deviation and that nothing is consumed in those periods, so that $x_{t+1} = (1-\delta)x_t + 1$. Then a storage infrastructure with $C = 10$ and a notable depreciation rate of $\delta = 0.01$ can store those unconsumed supplies for about 10 months before reaching the capacity limit[1].

In order to solve the functional equation for the equilibrium price function $f(x)$ as defined by Equations (4)-(6), we use a numerical algorithm which is based on the method of Kleppe and Oglend (2019), detailed in Appendix A.1. This algorithm takes advantage of the fact that the storage space is completely bounded by the non-negative constraint and the capacity limit C, thus providing numerically robust and computationally fast solutions. This is critical for a Bayesian MCMC posterior analysis because it requires a significant number

[1] Our selection of $C = 10$ also corresponds to the lowest upper limit, which Deaton and Laroque (1995, Table I) use for their grids of x-values in the interpolation scheme to compute the equilibrium price function for a set of various yearly commodity prices. The upper grid boundaries for the different commodities have been chosen by the authors so that the calculations never generate x-values that exceed these maximum values for the grid.

of reruns of the algorithm to obtain a solution to the pricing function for each new parameter value.

3.2 Bayesian Inference Using Particle Markov Chain Monte Carlo

In the SSM model as given by Equations (7) and (8) the vector of parameters to be estimated is given by $\theta = (v, \delta, b)$ and the vector of latent state variables is $x_{1:T}$, where the notation $a_{s:s'}$ is used to denote $(a_s, a_{s+1}, \ldots, a_{s'})$. The posterior of the parameters is $\pi(\theta|p_{1:T}) \propto \pi_\theta(p_{1:T})\pi(\theta)$, where $\pi(\theta)$ denotes the prior density assigned to θ and $\pi_\theta(p_{1:T})$ represents the likelihood function, given by

$$\pi_\theta(p_{1:T}) = \int \left[\prod_{t=2}^{T} \pi_\theta(p_t|p_{t-1}, x_{t-1:t})\pi_\theta(x_t|x_{t-1})\right] \pi_\theta(p_1, x_1)dx_{1:T}, \tag{13}$$

with

$$\pi_\theta(p_t|p_{t-1}, x_{t-1:t}) = \mathcal{N}\left(p_t|p_{t-1} + \log\left(\frac{f(x_t)}{f(x_{t-1})}\right), v^2\right), \quad \pi_\theta(x_t|x_{t-1}) = \mathcal{N}\left(x_t|(1-\delta)\sigma(x_{t-1}), 1\right), \tag{14}$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes a normal density function with mean μ and variance σ^2. For the joint density of the price and state in the initial period $\pi_\theta(p_1, x_1)$ we assume that it factorizes into a uniform density on $(-2, C+2)$ for the state x_1, denoted by $\mathcal{U}(x_1|-2, C+2)$, and a dirac measure for the price p_1 located at its actually observed value (effectively conditioning the likelihood on the first price observation).

Due to the non-linear nature of the pricing function $f(x)$ and the storage function $\sigma(x)$ entering the measurement and state transition density as given in Equation (14), the likelihood (and hence the resulting posterior for θ) are not available in closed form, so that a Bayesian and likelihood-based inference requires approximation techniques. Several Monte Carlo (MC) approximation approaches have been developed for statistical inference in non-linear SSMs with analytically intractable likelihood functions. However, only a few of them are suited to the model considered here due to the discontinuous derivatives of $f(x)$. In particular, methods using MC estimators for the likelihood $\pi_\theta(p_{1:T})$ based on approximations to the conditional posterior of the states $\pi(x_{1:T}|\theta, p_{1:T})$, including second order/Laplace approximations (Shephard and Pitt, 1997; Durbin and Koopman, 2012) or global approximations as used by the efficient importance sampler (Liesenfeld and Richard, 2003; Richard and Zhang, 2007), perform poorly in such a context. The same applies to the Gibbs approach targeting the joint posterior distribution of the states and parameters $\pi(x_{1:T}, \theta|p_{1:T})$ and alternately simulating from the conditional posteriors $\pi(x_{1:T}|\theta, p_{1:T})$ and $\pi(\theta|x_{1:T}, p_{1:T})$. It is known that such a Gibbs procedure typically has problems in efficiently approximating the targeted joint posterior in non-linear SSMs due to a fairly slow mixing (Bos and Shephard, 2006). Moreover, the Gibbs procedure is also not very computationally attractive in the present context, since both the (joint)

conditional posterior of all the states $\pi(x_{1:T}|\theta, p_{1:T})$ and the single-site conditional posterior of the individual states $\pi(x_t|x_{1:t-1}, x_{t+1:T}, \theta, p_{1:T})$ are non-standard distributions.

Here, we propose to use the particle marginal Metropolis-Hastings (PMMH) approach as developed by Andrieu et al. (2010), which is well suited for a posterior analysis of our proposed storage SSM as it can cope with the discontinuity of the gradients of $f(x)$ and is very easy to implement. The PMMH uses unbiased MC estimates of the likelihood $\pi_\theta(p_{1:T})$ inside a standard Metropolis-Hastings (MH) algorithm targeting the posterior of the parameters $\pi(\theta|p_{1:T})$. The MC estimation error of the likelihood estimate does not affect the invariant distribution of the MH so that the PMMH allows for exact inference. The PMMH produces an MCMC sample $\{\theta_i\}_{i=1}^S$ from the target distribution by the following MH updating scheme: Given the previously sampled θ_{i-1} and the corresponding likelihood estimate $\hat{\pi}_{\theta_{i-1}}(p_{1:T})$, a candidate value θ_* is drawn from a proposal density $Q(\theta|\theta_{i-1})$, and the estimate of the associated likelihood $\hat{\pi}_{\theta_*}(p_{1:T})$ is computed. Then the candidate θ_* is accepted as the next simulated θ_i with probability

$$\alpha(\theta_*, \theta_{i-1}) = \min\left\{1, \frac{\hat{\pi}_{\theta_*}(p_{1:T})\pi(\theta_*)}{\hat{\pi}_{\theta_{i-1}}(p_{1:T})\pi(\theta_{i-1})}\frac{Q(\theta_{i-1}|\theta_*)}{Q(\theta_*|\theta_{i-1})}\right\}, \tag{15}$$

otherwise θ_i is set equal to θ_{i-1}. Under weak regularity conditions, the resulting sequence $\{\theta_i\}_{i=1}^S$ converges to samples from the target density $\pi(\theta|p_{1:T})$ as $S \to \infty$ (Andrieu et al., 2010, Theorem 4).

For the PMMH, we use a Gaussian random walk proposal density $Q(\theta|\theta_{i-1}) = \mathcal{N}(\theta|\theta_{i-1}, \Sigma)$ and follow the approach of Haario et al. (2001) to adaptively set the proposal covariance matrix Σ during the burn-in period of the MCMC iterations. After dropping the draws from the burn-in period, we use the θ draws from the next M PMMH iterations to represent the posterior $\pi(\theta|x_{1:T})$. The posterior mean of the parameters, used as point estimates, is approximated by the sample mean over the M PMMH draws. For numerical stability of the PMMH computations, we reparameterize the likelihood function using the transformed parameters $\theta = \big(\log(v), \operatorname{arctanh}(2\delta - 1), \log(b)\big)$ so that the resulting parameter space is unconstrained.

The prior densities for the parameters are selected as follows: For $\log(b)$ we assume a $N(0, 1)$ prior, and for v^2 an inverted chi-squared prior with $v^2 \sim 0.1/\chi^2_{(10)}$, where $\chi^2_{(10)}$ denotes a chi-squared distribution with 10 degrees of freedom. Under this prior for v^2, the mean is given by 0.01 and the standard deviation by 0.007. The prior density assigned to δ is a Beta with $\delta \sim \mathcal{B}(2, 20)$ so that the mean and standard deviation are given by 0.09 and 0.05, respectively.

3.3 Particle Filter Likelihood Evaluation

In order to obtain unbiased MC estimates for the likelihood in Equation (13), required as an input of the PMMH, we follow Andrieu et al. (2010) and Flury and Shephard (2011) and use a simple sampling importance

resampling (SIR) particle filter (PF). For given values of the parameters θ, it produces MC estimates for the sequence of period-t likelihood contributions $\pi_\theta(p_t|p_{1:t-1})$ by sequentially sampling and resampling using an importance sampling (IS) density $q(x_t|x_{1:t-1})$ for the states x_t (see, Doucet and Johansen, 2009; Cappé et al., 2007, for a detailed treatment of PFs). For the implementation of the PF we use the state-transition density $\pi_\theta(x_t|x_{t-1})$ as IS density (Gordon et al., 1993), and rely on a dynamic resampling scheme in which the particles are resampled only when their effective sample size falls below one half of the number of particles (Doucet and Johansen, 2009). This simple version of the PF (also known as the bootstrap PF, BPF) for approximating the likelihood as given by Equations (13) and (14) consists of the following steps:

For period $t = 1$ (initialization): Sample $x_1^k \sim \pi_1(x_1) = \mathcal{U}(x_1| - 2, C + 2)$ for $k = 1, \ldots, N$ and set the corresponding (normalized) IS weights to $W_1^k = 1/N$. For initialization set $\bar{x}_1^k = x_1^k$.

For periods $t = 2, \ldots, T$: Sample $x_t^k \sim \pi_\theta(x_t|\bar{x}_{t-1}^k) = \mathcal{N}\big(x_t|(1-\delta)\sigma(\bar{x}_{t-1}^k), 1\big)$ for $k = 1, \ldots, N$ and set $x_{1:t}^k = (x_t^k, \bar{x}_{1:t-1}^k)$. Compute the IS weights as

$$w_t^k = W_{t-1}^k \pi_\theta(p_t|p_{t-1}, x_{t-1:t}^k), \tag{16}$$

and their normalized versions $W_t^k = w_t^k/(\sum_{\ell=1}^N w_t^\ell)$. Then use the IS weights to obtain the period-t likelihood contribution as $\hat{\pi}_\theta(p_t|p_{1:t-1}) = (\sum_{k=1}^N w_t^k)/N$, and compute the effective particle sample size defined by $N_t^e = [\sum_{k=1}^N (W_t^k)^2]^{-1}$. If $N_t^e < N/2$, resample from the particles $\{x_{1:t}^k\}_{k=1}^N$ with replacement according to their IS weights W_t^k to obtain the resampled particles $\{\bar{x}_{1:t}^k\}_{k=1}^N$, and set their weights to $W_t^k = 1/N$. Otherwise, set $\bar{x}_{1:t}^k = x_{1:t}^k$.

The resulting BPF estimate for the likelihood (conditional on the first price observation) is given by $\hat{\pi}_\theta(p_{1:T}) = [\prod_{t=2}^T \hat{\pi}_\theta(p_t|p_{1:t-1})]$. The measurement density $\pi_\theta(p_t|p_{t-1}, x_{t-1:t})$ is not very informative about the states x_t for empirically relevant parameter values, resulting in a low signal-to-noise ratio. Thus, the simple BPF yields fairly precise MC estimates of the likelihood with a modest number of particles N (Cappé et al., 2007). High precision likelihood estimates are a critical requirement for the PMMH to produce a well mixing MCMC sample from the posterior of the parameters $\pi(\theta|x_{1:T})$ (Flury and Shephard, 2011). In our applications below, we use $N = 10,000$ particles. For a time series with $T = 360$, one BPF likelihood estimate requires approximately 2.5 seconds (on a computer with an Intel Core i5-6500 processor running at 3.20 GHz). The MC numerical standard deviation of the log-likelihood estimate $\log \hat{\pi}_\theta(p_{1:T})$, computed from reruns of the BPF for a fixed θ value under different seeds, is about 0.1 percent of the absolute value of the log-likelihood, illustrating the high accuracy of the BPF.

3.4 State Prediction and Diagnostics

The BPF outlined in the previous section, and used for the PMMH implementation, can also be used to produce MC estimates for the predicted values of the latent state vector $x_{1:t+1}$ and functions thereof, given the prices observed up to period t, $p_{1:t}$. MC estimates of such predictions can serve as the basis for diagnostic checks. Let $h(x_{1:t+1})$ be a function of interest in $x_{1:t+1}$. Its conditional mean given $p_{1:t}$ can be expressed as

$$E\big(h(x_{1:t+1})|p_{1:t}\big) = \int h(x_{1:t+1})\pi_\theta(x_{t+1}|x_t)\pi_\theta(x_{1:t}|p_{1:t})dx_{1:t+1}, \tag{17}$$

where $\pi_\theta(x_{t+1}|x_t)$ is the state-transition density as given by Equation (14) and $\pi_\theta(x_{1:t}|p_{1:t})$ is the filtering density for $x_{1:t}$. Since the particles and IS-weights $\{x^k_{1:t}, W^k_t\}_{k=1}^N$ produced by the BPF provide an MC approximation to this filtering density, the conditional mean in Equation (17) for a given value of θ can be easily estimated by

$$\hat{E}(h(x_{1:t+1})|p_{1:t}) = \sum_{k=1}^{N} h(x^k_{1:t+1})W^k_t, \tag{18}$$

with $x^k_{1:t+1} = (x^k_{1:t}, x^k_{t+1})$, where x^k_{t+1} is obtained by propagating the BPF particle $x^k_{1:t}$ via the state-transition density, i.e. $x^k_{t+1} \sim \pi_\theta(x_{t+1}|x^k_t)$. In practice, the parameters θ are set equal to their estimates.

This MC approximation of a predicted mean like that in Equation (17) enables us to compute several useful statistics, such as the filtered mean for the price function of the storage model $E(\log f(x_t)|p_{1:t})$ and the stochastic trend component $E(k_t|p_{1:t}) = p_t - E(\log f(x_t)|p_{1:t})$, for which the function h to be used is $h(x_{1:t+1}) = \log f(x_t)$. State predictions can also be used to compute standardized Pearson residuals defined as

$$\eta_{t+1} = [p_{t+1} - E(p_{t+1}|p_{1:t})]/\mathrm{Var}(p_{t+1}|p_{1:t})^{1/2}. \tag{19}$$

If the model is correctly specified, then η_{t+1} and η^2_{t+1} are serially uncorrelated so that they can be used for diagnostic checking of the assumed dynamic structure. The conditional moments of p_{t+1} for the storage SSM are given by $E(p_{t+1}|p_{1:t}) = p_t + E(\log[f(x_{t+1})/f(x_t)]|p_{1:t})$ and $\mathrm{Var}(p_{t+1}|p_{1:t}) = \mathrm{Var}(\log[f(x_{t+1})/f(x_t)]|p_{1:t}) + v^2$, which can be evaluated by Equation (18), using the functions $h(x_{1:t+1}) = \log[f(x_{t+1})/f(x_t)]$ and $h(x_{1:t+1}) = \{\log[f(x_{t+1})/f(x_t)] - \hat{E}(\log[f(x_{t+1})/f(x_t)]|p_{1:t})\}^2$.

In order to check the capability of the storage SSM to approximate the distributional properties of the

observed prices we use the probability integral transformed (PIT) residuals defined as

$$\xi_{t+1} = \Phi^{-1}(u_{t+1}), \qquad u_{t+1} = \Pr(p_{t+1} \leq p^o_{t+1}|p_{1:t}), \tag{20}$$

where $\Pr(p_{t+1} \leq p^o_{t+1}|p_{1:t})$ is the predicted probability that p_{t+1} is less or equal to the actually ex-post observed price p^o_{t+1}, and Φ denotes the cdf of a $N(0,1)$-distribution (Kim et al., 1998). The PIT residuals ξ_{t+1} follow a $N(0,1)$-distribution if the model is valid. For the storage SSM, the probability u_{t+1} can be calculated by setting the function $h(x_{1:t+1})$ equal to $\Phi(\{p^o_{t+1} - p_t - \log[f(x_{t+1})/f(x_t)]\}/v)$.

3.5 Marginal Likelihood for Model Comparison

Marginal likelihood is used to compare the storage SSM with alternative models and assess the empirical relevance of the structural storage model component in the SSM. In order to evaluate the marginal likelihood for the storage SSM we rely upon the procedure proposed by Chib and Jeliazkov (2001), which is specifically customized for Bayesian analyses implemented using MH algorithms targeting the posterior of the parameters. This procedure takes advantage of the fact that the marginal likelihood can be expressed as

$$\pi(p_{1:T}) = \frac{\pi_{\bar{\theta}}(p_{1:T})\pi(\bar{\theta})}{\pi(\bar{\theta}|p_{1:T})}, \tag{21}$$

where $\pi_{\bar{\theta}}(p_{1:T})$ is the likelihood function for the observed prices evaluated at some value of the parameters $\bar{\theta}$, and $\pi(\bar{\theta})$ and $\pi(\bar{\theta}|p_{1:T})$ are the corresponding ordinates of the prior and posterior of the parameters. Then it exploits that the posterior ordinate $\pi(\bar{\theta}|p_{1:T})$ can be expressed in terms of the MH acceptance probability $\alpha(\cdot,\cdot)$ and the proposal density $Q(\cdot|\cdot)$. Namely, as the ratio of the expectation of $\alpha(\bar{\theta},\theta)Q(\bar{\theta}|\theta)$ under the posterior $\pi(\theta|p_{1:T})$ relative to the expectation of $\alpha(\theta,\bar{\theta})$ under the proposal density $Q(\theta|\bar{\theta})$. This implies that a consistent MC estimate for $\pi(\bar{\theta}|p_{1:T})$ based on the MH acceptance probability defined in Equation (15) is given by

$$\hat{\pi}(\bar{\theta}|p_{1:T}) = \frac{M^{-1}\sum_{i=1}^{M}\alpha(\bar{\theta},\theta_i)Q(\bar{\theta}|\theta_i)}{L^{-1}\sum_{l=1}^{L}\alpha(\theta_l,\bar{\theta})}, \tag{22}$$

where $\{\theta_i\}_{i=1}^M$ are the M simulated draws from the posterior distribution $\pi(\theta|p_{1:T})$ and $\{\theta_l\}_{l=1}^L$ are draws from the proposal distribution $Q(\theta|\bar{\theta})$. For evaluating the likelihood $\pi_{\bar{\theta}}(p_{1:T})$ in Equation (21) we use the same BPF algorithm as applied for the computation of the MH acceptance probabilities in Equation (15) (and outlined in Section 3.3). The value of the point $\bar{\theta}$ is set equal to the posterior mean of θ.

4 Ability to Isolate the Trend and Storage Model Component

In order to illustrate the capability of our Bayesian storage SSM approach to empirically separate the variation in the observed prices into the variation generated by the structural storage model component and that of the stochastic trend, we conduct a simulation experiment. Prices are simulated from the storage SSM for parameters that are set equal to their posterior mean values found for the empirical application to natural gas prices, discussed further below (see Table 1). Prices are simulated for 800 periods, with the first 500 discarded as burn-in, so that the size of the simulated sample is $T = 300$. The storage SSM is then fitted to the time series of simulated prices by using the PMMH procedure, and the BPF is applied to produce estimates of the filtered mean for the storage model price component $E(f(x_t)|p_{1:t})$ and the stochastic trend $E(k_t|p_{1:t})$, evaluated at the posterior mean of the parameters.

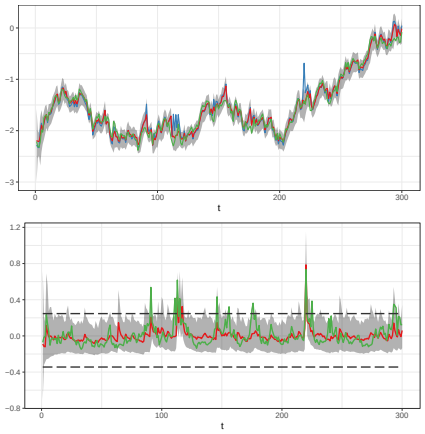

Figure 1: Filtered price components for simulated data. Upper panel: time series plot of the simulated log price $\log p_t$ (blue line), the actual stochastic trend component k_t (green line), and its estimated filtered mean $E(k_t|p_{1:t})$ (red line). Lower panel: time series plot of the actual storage model component $\log f(x_t)$ (green line) and its estimated filtered mean $E(\log f(x_t)|p_{1:t})$ (red line). The gray shaded areas indicate the 95% credible intervals under the filtering densities for k_t and $\log f(x_t)$, and the dashed lines in the lower panel mark the boundaries of the storage regimes. The prices are simulated using parameters set at $(v, \delta, b) = (0.097, 0.011, 0.420)$. The posterior mean of the parameters obtained by fitting the model to the simulated price data are $(\hat{v}, \hat{\delta}, \hat{b}) = (0.101, 0.013, 0.436)$.

Figure 1 shows the results of the simulation experiment. The upper panel displays the time series of the simulated log price $\log p_t$ and the actual simulated stochastic trend component k_t together with the estimate of its filtered mean, and the lower panel shows the time series of the actual price component which is generated by the competitive storage model $\log f(x_t)$ together with its estimated filtered mean. Also

plotted are the boundaries of the storage regimes. Values of the price component above the upper boundary correspond to the stock-out pricing regime, and values below the lower boundary correspond to the full storage capacity pricing regime. The plotted time series show that the estimated filtered means of the two price components track the time evolution of the true components quite well. We also observe that the estimated filtered means predict most of the stock-out events, which is critically important for identifying the structural parameters of the storage model. These simulation results illustrate that our approach appears to be well capable to empirically identify - from observed commodity prices - the fraction of the price variation which is due competitive storage decisions and to separate it from stochastic trend variation.

5 Empirical Application

In this section, we apply our Bayesian storage SSM approach to historical monthly price data for the following four commodities: Coffee (*Coffee, Other Mild Arabicas, New York cash price, ex-dock New York, US cents per pound*), cotton (*Average Spot Price in US cents per Pound for Upland cotton – color 41, leaf 4, staple 34*), aluminum (*Aluminum (LME) London Metal Exchange, unalloyed primary ingots, high grade, minimum 99.7% purity, USD per Metric Ton*), and natural gas (*Natural Gas (U.S.), spot price at Henry Hub, Louisiana, USD per MBtu*). The respective sample periods range from Jan 1989 until Dec 2018 ($T = 360$) for coffee, cotton and aluminum, and from Jan 1997 until Dec 2018 ($T = 264$) for natural gas. All prices are in nominal terms. We use monthly instead of annual prices to allow for more information about short-term price movements, as well as to avoid potentially spurious averaging effects of annual prices (Guerra et al., 2015).

5.1 Estimation Results for the Storage SSM with Stochastic Trend

For the Bayesian posterior analysis of the storage SSM, we run the PMMH algorithm for 12,000 iterations and discard the first 2,000 as burn-in. In order to evaluate the sampling efficiency of the PMMH for estimating the parameters, we compute the effective sample size (ESS) of their posterior PMMH samples (Geyer, 1992). The ESS measures the size of a hypothetical independent sample directly drawn from the posterior of the parameters which delivers the same numerical precision as the actual sample of M correlated PMMH parameter draws, so that large ESS values are to be preferred.

For each of the four commodities, the estimated posterior mean, standard deviation and ESS for the parameters are found in Table 1. The ESS values range from 376 to 1,028, indicating a satisfactory sampling efficiency with a fairly fast mixing rate of the PMMH algorithm. The estimates for the standard deviation of the trend innovations v imply that the stochastic trend accounts for 53% of the variation observed in the

		Natgas	Coffee	Cotton	Aluminum
v	Post. mean	0.0972	0.0607	0.0459	0.0448
	Post. std.	0.0083	0.0034	0.0027	0.0023
	ESS	634	442	1028	990
δ	Post. mean	0.0112	0.0023	0.0013	0.0011
	Post. std.	0.0048	0.0013	0.0008	0.0007
	ESS	580	376	761	847
b	Post. mean	0.4196	0.3849	0.3247	0.1987
	Post. std.	0.2594	0.0913	0.0598	0.0676
	ESS	515	386	972	917

Table 1: MCMC posterior analysis of the storage SSM with stochastic trend. The reported numbers are the posterior mean, posterior standard deviation and effective sample size (ESS) for the parameters. The results are based on $12,000$ PMMH iterations, discarding the first 2000 burn-in iterations.

monthly price changes for natural gas, 66% for coffee, 71% for cotton, and 81% for aluminum. As for the estimates of the depreciation rate δ, we observe that they are fully in line with the actual storage costs to be expected for the different types of commodities: For natural gas we find the largest estimated depreciation rate (1.1%), which implies that the monthly cost of storage amounts to 1.5% of the price. This relatively large estimated storage cost is in accordance with the fairly expensive storage technology for US natural gas, which is typically stored in underground salt caves and similar facilities. The second largest storage cost is found for coffee, with a monthly depreciation rate of 0.2% leading to estimated monthly costs of 0.6% of the price. The lowest storage costs are predicted for the non-food and non-energy products cotton and aluminum, for which the estimated depreciation rate is 0.1% resulting in storage costs of 0.5%. We also observe that the larger the estimated storage cost for a commodity, the larger the fraction of observed price variation which is captured by the storage decision behavior. This is in agreement with the rationality of the competitive storage model, where higher storage costs are associated with more frequent stock-out events, which in turn implies greater price volatility. The posterior mean values for the slope parameter b of the inverse demand function imply that a reduction in supply on the market by one standard deviation of production leads to a price increase of 42% for natural gas, 38% for coffee, 32% for cotton and 20% for aluminum. The size of these estimated price elasticities roughly corresponds to the size of the price peaks observed in these markets.

Figure 2 displays the time series of the log-prices for each of the four commodities, together with the filtered mean for their stochastic trend component k_t and their price component associated with the competitive storage model $f(x_t)$. We observe that the temporal evolution of the filtered estimates of the stochastic trend variable closely follows that of the observed prices. The filtered estimates for the storage model price component reveal that it predominantly captures periodically recurring price fluctuations with large price

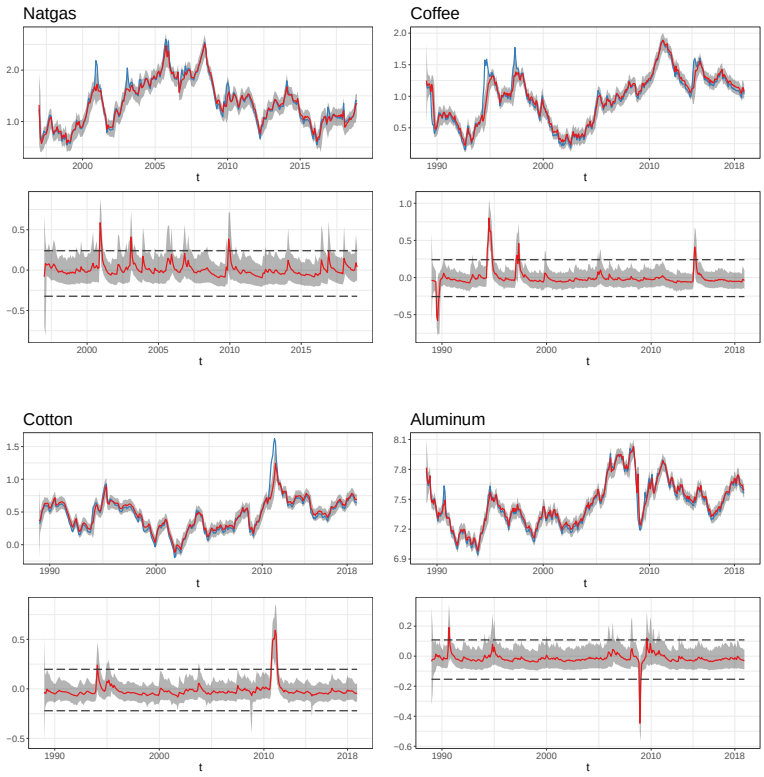

Figure 2: Commodity prices and filtered price components. Upper panels: time series plot of the log price $\log p_t$ (blue line) and the estimated filtered mean of the stochastic trend component $E(k_t|p_{1:t})$ (red line). Lower panels: time series plot of the estimated filtered mean of the storage model component $E(\log f(x_t)|p_{1:t})$ (red line). The gray shaded areas indicate the 95% credible intervals under the filtering densities for k_t and $\log f(x_t)$, and the dashed lines in the lower panels mark the boundaries of the storage regimes. The parameters are set to their posterior mean as given in Table 1.

peaks and drops. Beyond the periods with elevated price volatility, the contribution of this component to the price variation appears small. This reflects that when equilibrium storage is an inner solution (so that $0 < \sigma(x_t) < C$), the resulting price is subject to an intertemporal price restriction leading to prices which behave as a stationary Markov process. Accordingly, in this no-arbitrage pricing regime, the economic storage model provides little additional information about the price evolution that goes beyond the stochastic trend. However, storage becomes empirically relevant with a significant impact on the price behavior when the normal no-arbitrage pricing mechanism collapses in the stock-out and full-capacity regime, which occurs

in the storage model in periods of severe and prolonged commodity shortages or oversupply.

The limits-to-arbitrage regimes (stock-out or full-capacity) detected by the fitted storage model tend to coincide with known historical market events. For example, the time periods with peaks in the filtered storage price component for natural gas usually correspond to periods when the historical level of natural gas storage in the market was very low (Kleppe and Oglend, 2017). The sharp drop in the storage price component for coffee in 1989 coincides with the collapse of the International Coffee Agreement (a cartel of coffee-producing countries) and oversupply in the market due to World Bank subsidies, while the 1994 peak is consistent with a negative supply shock triggered by significant frost damage in much of the coffee-growing areas of Brazil. The cotton price peak detected by the storage model in 2011 was arguably due to the severe global shortages, which were caused, inter alia, by the tightening of Indian export restrictions on cotton. The early nineties spike in aluminum prices coincides with the collapse of the Soviet Union, and the 2008-2009 price drop is consistent with the sharp decline in global aluminum demand that created a large stock overhang during this period after the subprime crisis.

5.2 Model Comparisons

In this section, we assess the empirical relevance of the price component related to the competitive storage model for explaining the observed price variation, and compare the storage SSM model with stochastic trend to that with deterministic trend specifications. For this assessment, we rely on the marginal likelihood as well as diagnostic checks on Pearson and PIT residuals.

5.2.1 Alternative Models

For assessing the relevance of the storage model price component, we compare our SSM model to the restricted SSM that results in the absence of storage. The latter is obtained by letting $\delta \to 1$, making storage prohibitively costly, so that the stock process x_t collapses to that of the supply shocks z_t. In this case the SSM in Equations (1)-(3) with the assumed demand function $P(x) = \exp(-bx)$ reduces to

$$\begin{aligned} p_t &= k_t - bz_t, \qquad z_t \sim \text{iid}N(0,1), \\ k_t &= k_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \text{iid}N(0, v^2). \end{aligned}$$

This represents a standard linear Gaussian local level (LGLL) SSM (Durbin and Koopman, 2012) so that the Kalman filter can be applied for likelihood evaluation. As the Kalman filter provides exact values for the likelihood, the PMMH used for simulating from the posterior of the parameters for the unrestricted storage SSM can be replaced by a standard MH algorithm. The priors assigned to the two parameters (b, v) are the

	Natgas	Coffee	Cotton	Aluminum
Storage SSM	164.15	420.07	522.80	545.36
LGLL SSM	146.77	404.08	510.28	540.88
	(17.38)	(15.99)	(12.52)	(4.48)
Linear trend	109.57	309.64	427.00	496.77
	(54.58)	(110.43)	(95.80)	(48.59)
RCS3 trend	144.49	362.90	473.59	488.53
	(19.66)	(57.17)	(49.21)	(56.83)
RCS7 trend	132.03	375.50	488.44	517.23
	(32.12)	(44.57)	(34.36)	(28.13)

Table 2: Log marginal likelihood values with the log Bayes factor of the storage SSM relative to the alternative models in parentheses.

same as those we assume for the unrestricted storage SSM.

As deterministic trend specifications to be compared with the stochastic trend in the storage SSM, we consider those used in the study of Gouel and Legrand (2017). They use a linear time trend, for which k_t in Equation (2) is replaced by $k_t = \alpha + \beta t$. In addition, they consider restricted cubic spline trend specifications of the form $k_t = \sum_{g=1}^{G} = \gamma_g B_g(t)$, where $B_g(\cdot)$ are the basis functions of B-splines, G is the degree of freedom, and γ_g are the corresponding trend parameters to be estimated. For our comparison we consider restricted cubic splines with 3 knots (RSC3) and 5 trend parameters as well as 7 knots (RSC7) and 9 trend parameters[2]. For these deterministic trends the SSM in Equations (1)-(3) reduces to a univariate, non-linear autoregression for the log-price:

$$p_t = k_t + \log f\left[(1-\delta)\sigma(x_{t-1}) + z_t\right], \qquad x_{t-1} = f^{-1}\left[\exp(p_{t-1} - k_{t-1})\right], \qquad z_t \sim \text{iid}N(0,1). \tag{23}$$

Analogously to the LGLL SSM, we can simulate from the posterior for the parameters of the deterministic trend models by using a standard MH algorithm. For the structural parameters (δ, b) we assume the same priors as used in the storage SSM, and to the deterministic trend parameters $(\alpha, \beta, \gamma_g)$ we assign independent $N(0, 20^2)$ priors. For details on the computation and derivation of the Pearson and PIT residuals of the deterministic trend models, see Appendix A.2.

5.2.2 Marginal Likelihood Model Comparisons and Diagnostics Checks

Table 2 provides the log marginal likelihood values $\log \pi(p_{1:T}|\text{model}_s)$ for the storage SSM together with those of the LGLL SSM and the storage model combined with the deterministic trend specifications. Also reported are the resulting values for the log Bayes factor of the storage SSM relative to the four alternative models

[2]The knots for the RSC3 specification are located at the 25%, 50% and 75% quantiles of the time index and for the RSC7 at the 12.5%, 25%, 37.5%, 50%, 67.5%, 75% and 87.5% quantiles.

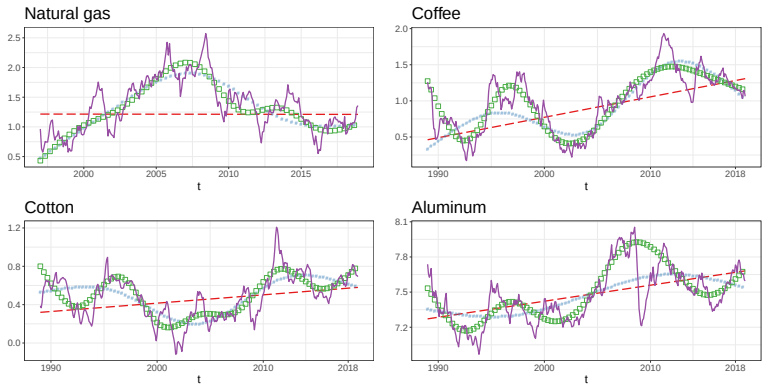

Figure 3: Fitted stochastic and deterministic trends. Smoothed stochastic trend (purple solid line), linear trend (red dashed line), RSC3 trend (blue number sign), RCS7 trend (green square).

$\log[\pi(p_{1:T}|\text{storage SSM})/\pi(p_{1:T}|\text{model}_\ell)]$. The results reveal that the storage SSM is strongly preferred over the LGLL SSM for all commodities, which suggests that the structural storage component in the SSM substantially contributes to the model fit. Hence, the non-linear price dynamics with periodically recurring increases in price volatility and price spiking, as predicted by the competitive storage model, adds significantly to explaining the price behavior. For all commodities, we also observe that the storage SSM is clearly favored over all deterministic trend specifications. Thus, the storage SSM has a trend component that is not only consistent with the rationality of the economic model, but is also much more supported by the data than the deterministic trends, such as those used by Gouel and Legrand (2017) for the estimation of the structural parameters of the competitive storage model. Our estimates of the structural parameters for the deterministic trend models are found in Appendix A.3. Figure 3 shows the time series plots of the fitted deterministic trends $\hat{k}_t$ and the smoothed mean of the stochastic trend $E(k_t|p_{1:T})$, all computed by setting the parameters to their posterior mean values[3]. Unsurprisingly, we find that the stochastic trend captures a substantially larger fraction of the observed price variations than the deterministic trends.

Table 3 provides the results of diagnostic checks on the PIT residuals ξ_t and the Pearson residuals η_t for the storage SSM and the four alternative models considered. The PIT residuals of the storage SSM suggest that this model accounts well for the observed distributional properties of the prices for all commodities. The skewness and kurtosis of its PIT residuals are close to their benchmark values for a normal distribution and they all pass the Jarque-Bera normality test at the 5% significance level. In contrast, the LGLL SSM

[3]The smoothed mean $E(k_t|p_{1:T}) = p_t - E(\log f(x_t)|p_{1:T})$ is computed using the particle smoothing algorithm, which adds to the BPF as outlined in Section 3.3 a backward sampling step (Doucet and Johansen, 2009, Section 5).

	Skew(ξ_t)	Kurt(ξ_t)	JB(ξ_t)	$\rho_1(\eta_t)$	$LB_{12}(\eta_t)$	$LB_{12}(\eta_t^2)$
			Storage SSM			
Natgas	0.053	3.069	0.915	0.075	0.027	0.297
Coffee	0.255	3.333	0.062	0.359	<0.001	<0.001
Cotton	-0.064	3.445	0.201	0.518	<0.001	<0.001
Aluminum	-0.214	3.25	0.159	0.291	<0.001	<0.001
			LGLL SSM			
Natgas	0.033	4.298	<0.001	0.084	0.055	0.452
Coffee	0.801	7.681	<0.001	0.258	<0.001	<0.001
Cotton	-0.23	6.325	<0.001	0.502	<0.001	<0.001
Aluminum	-0.381	4.652	<0.001	0.268	<0.001	<0.001
			Linear trend			
Natgas	0.049	5.368	<0.001	0.075	0.065	0.427
Coffee	-0.532	4.795	<0.001	0.253	<0.001	0.053
Cotton	0.137	4.904	<0.001	0.497	<0.001	<0.001
Aluminum	0.625	6.502	<0.001	0.243	<0.001	<0.001
			RCS3 trend			
Natgas	-0.099	3.962	0.005	<0.001	0.003	0.068
Coffee	-0.508	5.231	<0.001	0.185	<0.001	0.004
Cotton	0.094	4.669	<0.001	0.449	<0.001	<0.001
Aluminum	0.369	5.407	<0.001	0.205	<0.001	0.002
			RCS7 trend			
Natgas	-0.258	3.835	0.005	0.022	<0.001	0.004
Coffee	-0.472	4.938	<0.001	0.184	<0.001	0.022
Cotton	0.181	4.437	<0.001	0.454	<0.001	<0.001
Aluminum	-0.059	3.495	0.144	0.198	<0.001	<0.001

Table 3: Diagnostics on the PIT and Pearson residuals. Skewness, Kurtosis, and p-value of the Jarque-Bera test (JB) for the PIT residuals. Lag-1 autocorrelation (ρ_1) and p-value of the Ljung-Box test (LB) for the Pearson residuals and their squared values, including 12 lags.

as well as the storage models with deterministic trends have difficulties approximating the distributional properties of the prices. Only the PIT residuals of the storage model with an RSC7 trend for aluminum pass the Jarque-Bera normality test at a conventional significance level.

The first-order serial correlation of the Pearson residuals η_t and the p-values of the Ljung-Box test for η_t and η_t^2 including 12 lags reported in Table 3 show that the storage SSM successfully accounts for the observed autocorrelation in the level and volatility of the gas price, while they point towards significant residual correlation in price level and volatility for coffee, cotton and aluminum. However, all competing models cannot fully capture the serial correlation in the price levels of those three commodities either. Only the volatility dynamics for coffee is better approximated by the linear and RSC7 trend model than by the storage SSM. Clearly, based on these results, we can not identify whether the failure of the storage

SSM and the deterministic trend models to explain all of the observed dynamics in the coffee, cotton and aluminum prices is due to a potential misspecification of the trend or the competitive storage model itself, since the diagnostic tests are, as any specification test in this context, joint tests for the validity of both price components.

In sum, the results show that the storage SSM outperforms the deterministic trend models in explaining the observed distributional properties of commodity prices, and that its ability to account for the dynamics in the price levels is not worse. Only in the approximation of the volatility dynamics, the deterministic trend specifications appear to have a slight advantage.

5.2.3 Structural Parameter Estimates Under Stochastic and Deterministic Trends

As it is evident from Figure 6, the dynamic and distributional characteristics of the de-trended prices substantially differ depending on whether a stochastic or deterministic trend is assumed. Therefore, it can be expected that the nature of the trend has a critical impact on the estimates of the parameters that determine the storage costs (δ) and the price elasticity of demand (b), since these parameters are identified by the strength of the serial correlation and the size of the spikes in the trend-adjusted prices. The lower the storage costs in the competitive storage model are, the stronger the predicted serial correlation, while the more inelastic the demand is, the larger the resulting price spikes. As larger price spikes also imply more speculative storage activity, an inelastic demand also contributes to the strength of the predicted serial correlation in the prices.

Table 4 summarizes the estimates for the annualized storage costs (net of interest costs) in percent of the average price, and the price elasticities of demand obtained from the fitted storage SSM and the deterministic RCS trend models. The annual storage costs are computed as $-[(1-\delta)^{12}-1]$ and the price elasticity is given by $[-(b\bar{x})^{-1}]$, where $\bar{x}$ is the mean supply. We observe that the SSM with stochastic trend predicts substantially larger elasticities (in absolute values) than the deterministic trend models for all commodities and, except for natural gas, lower storage costs. The larger elasticities found under the storage SSM reflect that the stochastic trend produces, due to its greater flexibility to track the observed price, trend-adjusted prices that have spikes that are smaller than those obtained under a deterministic trend. Hence, in contrast to the deterministic trend specifications, the stochastic trend SSM is not forced to match the large spikes observed in the actual prices by small estimated values for the elasticity. For natural gas, the residual serial correlation in the prices adjusted by the stochastic trend component also appears to be relatively low, which indicates relatively high storage costs. However, for the other commodities, this residual serial correlation is larger leading to substantially lower estimated storage costs.

Gouel and Legrand (2017) provide estimates of storage costs and price elasticities of demand based on

	Natgas		Coffee		Cotton		Aluminum	
	costs	elast.	costs	elast.	costs	elast.	costs	elast.
Storage SSM	12.6	-1.03	2.7	-0.65	1.6	-0.69	1.3	-1.46
RCS3 trend	8.9	-0.11	4.8	-0.20	2.1	-0.25	6.5	-0.27
RCS7 trend	11.1	-0.10	4.1	-0.19	4.7	-0.26	1.7	-0.30

Table 4: Estimates for the annual storage costs (net of interest costs) in percent of the average price and price elasticities of demand.

deterministic trend models used for annual data on various commodities, including coffee and cotton. This allows for some comparisons with our results for those two commodities. The annual storage costs estimates they report for their preferred trend model for coffee and cotton are, respectively, 1.4% and 0.3% of the average price. These estimates based on annual data are much lower than those we found for the storage SSM as well as the deterministic trend models fitted to monthly data. However, they argue that their estimated annual costs are possibly too small - an assessment that is consistent with our estimates for the storage costs. For the annual price elasticity of demand, the estimates of Gouel and Legrand (2017) are -0.04% for coffee and -0.03% for cotton. These estimates imply a demand for those commodities which is substantially more inelastic than that implied from our estimates. One can argue which elasticities better reflect the markets. Mehta and Chavas (2008) assume a range of plausible values for the annual elasticity of demand for coffee between -0.2% and -0.4%, while Duffy et al. (1990) argue that the annual export demand for cotton is likely fairly elastic. Hence, our elasticity estimates are more in line with these assessments than those found by Gouel and Legrand (2017).

6 Conclusion

In this paper, we have proposed a stochastic trend competitive storage model for commodity prices, which defines a non-linear state-space model (SSM). For the Bayesian posterior analysis of the proposed stochastic trend SSM, we use an efficient MCMC procedure. This adds to existing empirical commodity storage models based on deterministic trend specifications. Our stochastic trend approach fits into the economic rationality of the competitive storage model and is also sufficiently flexible to account for the variation in the observed prices that the competitive storage model is not intended to explain. The obvious benefit is that it makes the storage model applicable to markets with highly persistent unit root-like prices, which appears relevant for many commodity markets. Our approach aims at increasing the empirical relevance and applicability of the competitive storage model.

The MCMC procedure we propose for jointly estimating the structural and trend parameters in the SSM

is a particle marginal Metropolis-Hastings algorithm based on the bootstrap particle filter. A Monte Carlo simulation experiment shows that this approach is able to disentangle the stochastic trend from the price variation due to speculative storage. The SSM is applied to monthly price data for natural gas, cotton, coffee and aluminum. Not surprisingly, the stochastic trend explains a large part of the observed variation in the commodity prices. More importantly, the competitive storage component adds short-run price volatility and price spiking, and becomes periodically relevant to explain non-linear pricing behavior related to states of market turmoil. A formal empirical comparison of the SSM to the corresponding model that results in the absence of storage suggests that the speculative storage price component significantly contributes to commodity price variation.

Which trend to apply will depend on the specific market under consideration. If a stochastic trend is not appropriate, fitting a highly flexible stochastic trend model risks overfitting the price variation and downplaying the contribution of the storage model. Consequently, the price elasticity of demand will tend to be overestimated and the estimates of the storage costs can be expected to be correspondingly biased. On the other hand, failing to account for a stochastic trend when it is appropriate will tend to underestimate the elasticity of demand. Our empirical results show that the stochastic trend model consistently estimates a higher elasticity of demand and a different amount of storage costs than existing deterministic trend models for the commodity markets investigated in this paper. Pre-testing of price characteristics can guide trend choice. For instance, unit root tests can be applied to evaluate whether a stochastic trend specification is suitable.

The empirical comparison of the stochastic trend SSM to existing deterministic trend models using the Bayes factor and model residual analysis shows that the stochastic trend fits the price data for the investigated commodity markets much better than the deterministic trends. In particular, in contrast to the deterministic trend specifications, the stochastic trend SSM captures the observed distributional properties of the prices, such as their skewness and kurtosis, quite well. While the stochastic trend SSM also accounts for the price dynamics in the observed prices on the natural gas market it is not able to fully capture all the serial dependence of the coffee, cotton and aluminum prices. This is similar to the results of financial approaches on modeling commodity term structures, showing the relevance of additional pricing factors beyond the traditional ones for the spot price and the convenience yield (Miltersen and Schwartz 1998; Schwartz 1997; Tang 2012). The stochastic trend SSM is essentially a two-factor model with one reduced-form random walk component orthogonally appended to a factor restricted by economic constraints. Increasing the flexibility in the economic model will arguably improve the explanatory power of the model, although with additional statistical challenges in separately identifying the trend behavior from the price component related to the competitive storage model.

References

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.

Bobenrieth, E., B. Wright, and D. Zeng (2013). Stocks-to-use ratios and prices as indicators of vulnerability to spikes in global cereal markets. *Agricultural Economics 44*(s1), 43–52.

Bos, C. S. and N. Shephard (2006). Inference for adaptive time series models: Stochastic volatility and conditionally gaussian state space form. *Econometric Reviews 25*(2-3), 219–244.

Cafiero, C., E. Bobenrieth H., and J. Bobenrieth H. (2011). Storage arbitrage and commodity price volatility. *Safeguarding food security in volatile global markets*, 301–326.

Cafiero, C., E. Bobenrieth H., J. Bobenrieth H., and B. D. Wright (2011). The empirical relevance of the competitive storage model. *Journal of Econometrics 162*(1), 44–54.

Cafiero, C., E. Bobenrieth H., J. Bobenrieth H., and B. D. Wright (2015). Maximum likelihood estimation of the standard commodity storage model: Evidence from sugar prices. *American Journal of Agricultural Economics 97*(1), 122–136.

Canova, F. (2014). Bridging dsge models and the raw data. *Journal of Monetary Economics 67*, 1–15.

Cappé, O., S. J. Godsill, and E. Moulines (2007). An overview of existing methods and recent advances in sequential monte carlo. *Proceedings of the IEEE 95*(5), 899–924.

Chib, S. and I. Jeliazkov (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association 96*(453), 270–281.

Deaton, A. and G. Laroque (1992). On the behaviour of commodity prices. *The Review of Economic Studies 59*(1), 1–23.

Deaton, A. and G. Laroque (1995). Estimating a nonlinear rational expectations commodity price model with unobservable state variables. *Journal of Applied Econometrics 10*(S1), S9–S40.

Deaton, A. and G. Laroque (1996). Competitive Storage and Commodity Price Dynamics. *The Journal of Political Economy 104(5)*, 896–923.

DeJong, D. N. and C. Dave (2011). *Structural Macroeconometrics* (2 ed.). Princeton University Press.

Doucet, A. and A. M. Johansen (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering 12*(656-704), 3.

Duffy, P. A., M. K. Wohlgenant, and J. W. Richardson (1990). The elasticity of export demand for us cotton. *American Journal of Agricultural Economics 72*(2), 468–474.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods* (2 ed.). Number 38 in Oxford Statistical Science. Oxford University Press.

Flury, T. and N. Shephard (2011). Bayesian inference based only on simulated likelihood: Particle filter analysis of dynamic economic models. *Econometric Theory 27*(Special Issue 05), 933–956.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science 7*(4), 473–483.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing 140*(2), 107–113.

Gouel, C. and N. Legrand (2017). Estimating the competitive storage model with trending commodity prices. *Journal of Applied Econometrics 32*(4), 744–763.

Guerra, V., E. Bobenrieth H., J. Bobenrieth H., and C. Cafiero (2015). Empirical commodity storage model: the challenge of matching data and theory. *European Review of Agricultural Economics 42*(4), 607–623.

Gustafson, R. L. (1958). *Carryover levels for grains: a method for determining amounts that are optimal under specified conditions*. Number 1178. US Department of Agriculture.

Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive Metropolis algorithm. *Bernoulli 7*(2), 223–242.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: Likelihood inference and comparison with arch models. *The Review of Economic Studies 65*(3), 361–393.

Kleppe, T. S. and A. Oglend (2017). Estimating the competitive storage model: A simulated likelihood approach. *Econometrics and Statistics 4*, 39–56.

Kleppe, T. S. and A. Oglend (2019). Can limits-to-arbitrage from bounded storage improve commodity term-structure modeling? *Journal of Futures Markets 39*(7), 865–889.

Legrand, N. (2019). The empirical merit of structural explanations of commodity price volatility: Review and perspectives. *Journal of Economic Surveys 33*(2), 639–664.

Liesenfeld, R. and J.-F. Richard (2003). Univariate and multivariate stochastic volatility models: estimation and diagnostics. *Journal of Empirical Finance 10*(4), 505–531.

Mehta, A. and J.-P. Chavas (2008). Responding to the coffee crisis: What can we learn from price dynamics? *Journal of Development Economics 85*(1-2), 282–311.

Miltersen, K. R. and E. S. Schwartz (1998). Pricing of options on commodity futures with stochastic term structures of convenience yields and interest rates. *Journal of Financial and Quantitative Analysis 33*(1), 33–59.

Oglend, A. and T. S. Kleppe (2017). On the behavior of commodity prices when speculative storage is bounded. *Journal of Economic Dynamics and Control 75*, 52–69.

Richard, J.-F. and W. Zhang (2007). Efficient high-dimensional importance sampling. *Journal of Econometrics 127*(2), 1385–1411.

Routledge, B. R., D. J. Seppi, and C. S. Spatt (2000). Equilibrium forward curves for commodities. *The Journal of Finance 55*(3), 1297–1337.

Sala, L. (2015). Dsge models in the frequency domains. *Journal of Applied Econometrics 30*(2), 219–240.

Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implications for valuation and hedging. *The Journal of Finance 52*(3), 923–973.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*, 653–667.

Tang, K. (2012). Time-varying long-run mean of commodity prices and the modeling of futures term structures. *Quantitative Finance 12*(5), 781–790.

Wang, D. and W. G. Tomek (2007). Commodity prices and unit root tests. *American Journal of Agricultural Economics 89*(4), 873–889.

A Appendix

A.1 Numerical Solution of the Price Function

The numerical algorithm we use to solve the functional equation for the price function $f(x)$ as defined by Equations (4)-(6) is based on that used by Kleppe and Oglend (2019) for a model with autocorrelated supply shocks. The algorithm is based on solving the storage policy function $\sigma(x)$ and then recover $f(x)$ via Equation (6), which implies that

$$f(x) = P(x - \sigma(x)). \tag{A-1}$$

Let x^* be defined such that $P(x^*) = \bar{f}(x^*)$ and x^{**} such that $P(x^{**} - C) = \bar{f}(x^{**})$. For $x < x^*$, so that $f(x) = P(x)$, it follows that $\sigma(x) = 0$ (stock-out regime); for $x^* \leq x \leq x^{**}$, so that $f(x) = \bar{f}(x)$, it follows that $\sigma(x) \in [0, C]$ (storage regime); and for $x > x^{**}$, so that $f(x) = P(x - C)$, it follows that $\sigma(x) = C$ (full capacity storage regime).

The numerical representation of $\sigma(x)$ is given by $\mathcal{S} = \{\hat{x}^*, \hat{x}^{**}, s(x)\}$, where the function $s(x)$ (with $s(x) \simeq \sigma(x)$ for $x \in [\hat{x}^*, \hat{x}^{**}]$) is represented on a (comparatively sparse) grid on $[\hat{x}^*, \hat{x}^{**}]$ and is evaluated using a suitable interpolation method (e.g. linear). The resulting approximation is given by

$$\sigma_{\mathcal{S}}(x) = \begin{cases} 0 & \text{if } x < \hat{x}^* \\ s(x) & \text{if } \hat{x}^* \leq x \leq \hat{x}^{**} \\ C & \text{if } x > \hat{x}^{**} \end{cases},$$

and correspondingly $f_{\mathcal{S}}(x) = P(x - \sigma_{\mathcal{S}}(x))$.

The iteration to find $\hat{\sigma}_{\mathcal{S}}(x) \simeq \sigma(x)$ consists of the following steps:

1. Select an initial guess, e.g. $\mathcal{S}_1 = \{\hat{x}_1^*, \hat{x}_1^{**}, s_1(x)\} = \{0, C, s_1(x)\}$, where $s_1(x)$ is the linear function such that $s(0) = 0$, $s(C) = C$. Set $n = 1$.

2. Update the left kink point $\hat{x}_{n+1}^*$ according to

$$\hat{x}_{n+1}^* = D\left(\beta \int f_{\mathcal{S}_n}(z)\phi(z)dz\right). \tag{A-2}$$

3. Update the right kink point $\hat{x}_{n+1}^{**}$ according to

$$\hat{x}_{n+1}^{**} = D\left(\beta \int f_{\mathcal{S}_n}((1-\delta)C + z)\phi(z)dz\right) + C. \tag{A-3}$$

4. Update the grid $\{x_{n+1}^{(j)}\}$ to be on $[\hat{x}_{n+1}^{*}, \hat{x}_{n+1}^{**}]$.

5. For each grid point j, find the update $s_{n+1}(x_{n+1}^{(j)})$ as the solution in s to

$$s = x_{n+1}^{(j)} - D\left(\beta \int f_{\mathcal{S}_n}((1-\delta)s + z)\phi(z)dz\right), \tag{A-4}$$

   using a univariate non-linear root-finding algorithm. Notice that the solution $s = s_{n+1}(x_{n+1}^{(j)})$ is constrained to be in $[0, C]$, and that $s_{n+1}(\hat{x}_{n+1}^{*}) = 0$, $s_{n+1}(\hat{x}_{n+1}^{**}) = C$.

6. Until convergence, set $n \leftarrow n + 1$ and go back to step 2.

The integrals in Equations (A-2)-(A-4) are approximated using the trapezoidal quadrature rule with 128 subintervals, over the interval $[-4, 4]$, and the non-linear equation (A-4) is solved using Brent's method. Allowing the grid space to adjust to the updated functional solutions ensures that the grid can dynamically concentrate in the region of the state-space where a high precision is needed, namely the region defining the storage regime. This provides both efficient and precise numerical solutions to the pricing function.

A.2 Residuals for the Deterministic Trend Models

For the deterministic trend models as given by Equation (23) the conditional expectation $E(p_{t+1}|p_{1:t})$ and variance $\mathrm{Var}(p_{t+1}|p_{1:t})$ defining the Pearson residuals in Equation (19) can be evaluated by MC integration as the sample mean and variance of the simulated prices

$$p_{t+1}^{k} = k_{t+1} + \log f[(1-\delta)\sigma(x_t) + z_{t+1}^{k}], \qquad k = 1, \ldots, N, \tag{A-5}$$

where $\{z_{t+1}^{k}\}_{k=1}^{N}$ are iid draws from a $N(0,1)$ distribution.

The PIT residuals in Equation (20) obtain as follows: The probability $u_{t+1} = \Pr(p_{t+1} \leq p_{t+1}^{o}|p_{1:t})$, which follows for a correctly specified model a uniform distribution $U_{[0,1]}$ on the unit interval, results as

$$\begin{aligned} u_{t+1} &= \Pr(\exp(p_{t+1} - k_{t+1}) \leq \exp(p_{t+1}^{o} - k_{t+1}) \,|\, p_{1:t}) && \text{(A-6)} \\ &= \Pr(f^{-1}[\exp(p_{t+1} - k_{t+1})] \geq f^{-1}[\exp(p_{t+1}^{o} - k_{t+1})] \,|\, p_{1:t}) && \text{(A-7)} \\ &= \Pr(z_{t+1} \geq z_{t+1}^{o} \,|\, p_{1:t}), && \text{(A-8)} \end{aligned}$$

where $z_{t+1}^{o} = f^{-1}[\exp(p_{t+1}^{o} - k_{t+1})] - (1-\delta)\sigma(x_t)$. Equation (A-7) follows from the fact that the inverse of the rational expectations equilibrium price function f^{-1} is monotonically non-increasing. Since z_t is a $N(0,1)$ random variate with cdf denoted by Φ, Equation (A-8) implies that $1 - u_{t+1} = \Phi(z_{t+1}^{(o)})$. Since for

$u_{t+1} \sim U_{[0,1]}$ it holds that $1 - u_{t+1} \sim U_{[0,1]}$, the PIT residuals are given by

$$\xi_{t+1} = \Phi^{-1}(1 - u_{t+1}) \tag{A-9}$$

$$= \Phi^{-1}(\Phi(z_{t+1}^{(o)})) \tag{A-10}$$

$$= z_{t+1}^{(o)}. \tag{A-11}$$

A.3 Additional Estimation Results

Table A-1 provides the posterior mean and standard deviation for the structural parameters of the competitive storage model combined with deterministic trend specifications.

			Natgas	Coffee	Cotton	Aluminum
δ	Linear	Post. mean	0.0090	0.0025	0.0023	0.0087
		Post. std.	0.0053	0.0016	0.0015	0.0030
	RCS3	Post. mean	0.0077	0.0041	0.0017	0.0056
		Post. std.	0.0044	0.0023	0.0012	0.0021
	RCS7	Post. mean	0.0098	0.0035	0.0040	0.0014
		Post. std.	0.0051	0.0024	0.0022	0.0010
b	Linear	Post. mean	2.05	1.33	1.02	0.75
		Post. std.	0.122	0.058	0.045	0.036
	RCS3	Post. mean	1.83	1.12	0.84	0.74
		Post. std.	0.110	0.056	0.045	0.046
	RCS7	Post. mean	1.82	1.08	0.77	0.67
		Post. std.	0.118	0.054	0.043	0.032

Table A-1: Estimates for the storage model parameters under the storage model with deterministic trends.