Estimation in Non-Linear Non-Gaussian State Space Models with Precision-Based Methods

Chan, Joshua and Strachan, Rodney

Australian National University

2012

Estimation in Non-Linear Non-Gaussian State Space Models with Precision-Based Methods

Joshua C.C. Chan[*]   Rodney Strachan[†]

Research School of Economics,
and Centre for Applied Macroeconomic Analysis,
Australian National University

March 2012

Abstract

In recent years state space models, particularly the linear Gaussian version, have become the standard framework for analyzing macroeconomic and financial data. However, many theoretically motivated models imply non-linear or non-Gaussian specifications — or both. Existing methods for estimating such models are computationally intensive, and often cannot be applied to models with more than a few states. Building upon recent developments in precision-based algorithms, we propose a general approach to estimating high-dimensional non-linear non-Gaussian state space models. The baseline algorithm approximates the conditional distribution of the states by a multivariate Gaussian or t density, which is then used for posterior simulation. We further develop this baseline algorithm to construct more sophisticated samplers with attractive properties: one based on the accept-reject Metropolis-Hastings (ARMH) algorithm, and another adaptive collapsed sampler inspired by the cross-entropy method. To illustrate the proposed approach, we investigate the effect of the zero lower bound of interest rate on monetary transmission mechanism.

[*]Corresponding author. Email: joshua.chan@anu.edu.au

[†]Email: rodney.strachan@anu.edu.au. Strachan is also a Fellow of the Rimini Centre for Economic Analysis.

1

1   Introduction

State space models have proved to be very useful for modeling a wide range of processes in economics, and the linear Gaussian version has been the standard specification. However, economists have become increasingly aware that many processes are better represented by specifications that are non-linear or non-Gaussian, or both. The current preferred technique for estimating these processes is sequential Monte Carlo methods. This paper proposes an alternative, general approach to inference in high-dimensional non-linear non-Gaussian state space models, which we believe are useful in many applications.

Building upon recent developments in precision-based algorithms for the linear Gaussian case, we present three fast sampling schemes for efficient simulation of the states in general state space models with multivariate observations and states. The first algorithm, the baseline algorithm, approximates the conditional distribution of the states by a multivariate Gaussian or t density, which is then used as a proposal density for posterior simulation using Markov chain Monte Carlo (MCMC) methods. This approximating density can also used for evaluating the integrated likelihood — the joint distribution of the observations given the model parameters but integrated over the states — via importance sampling. We then build upon this baseline approach to consider two other more efficient algorithms for posterior simulation. The first of these, our second algorithm, is the accept-reject Metropolis-Hastings (ARMH) algorithm that combines the classic accept-reject sampling and the Metropolis-Hastings algorithm. The third algorithm is a collapsed sampler used in conjunction with the cross-entropy method, where we sample the states and the model parameters jointly to reduce autocorrelations in the posterior simulator.

The specific framework we consider is a general state space model where the evolution of the $n \times 1$ vector of observations y_t is governed by the measurement or observation equation characterized by a generic density function $p(y_t \,|\, \eta_t, \theta)$, where η_t is an $m \times 1$ vector of latent states and θ denotes the set of model parameters. Note that the density $p(y_t \,|\, \eta_t, \theta)$ may depend on previous observations y_{t-1}, y_{t-2}, etc. and other covariates as is the case in the application in this paper. These observations are suppressed in the conditioning sets for notational convenience. The evolution of the states η_t, in turn, is specified by the state or transition equation summarized by

3

the density function $p(\eta_t \,|\, \eta_{t-1}, \theta)$. We note in passing that the proposed approach can be easily generalized to the case where the state equation is non-Markovian and the observation y_t depends on previous states η_{t-1}, η_{t-2}, etc.

A vast collection of models can be written in the general state space form with measurement equation $p(y_t \,|\, \eta_t, \theta)$ and state equation $p(\eta_t \,|\, \eta_{t-1}, \theta)$. Models that have proven popular among economists include time-varying parameter vector autoregressive (TVP-VAR) models, dynamic factor models, stochastic volatility models, and a large class of macroeconomic models generally known as dynamic stochastic general equilibrium (DSGE) models, among many others. Substantial progress has been made in the last two decades in estimating linear Gaussian state space models. For example, Kalman filter-based algorithms include Carter and Kohn (1994), Früwirth-Schnatter (1994), de Jong and Shephard (1995) and Durbin and Koopman (2002); more recently, precision-based algorithms are proposed in Rue, Martino, and Chopin (2009), Chan and Jeliazkov (2009b) and McCausland, Millera, and Pelletier (2011). Efficient simulation algorithms also exist for certain specific non-linear non-Gaussian state space models, the most notable example is the class of stochastic volatility models. Using data augmentation and finite Gaussian mixtures to approximate non-Gaussian errors, Kim, Shepherd, and Chib (1998) propose a Gibbs sampler for posterior simulation in a univariate stochastic volatility model. This approach is later applied to other univariate stochastic volatility models in Chib, Nardari, and Shephard (2002) and multivariate models in Cogley and Sargent (2005) and Primiceri (2005). Other successful applications of the auxiliary mixture sampling include state space models for Poisson counts in Frühwirth-Schnatter and Wagner (2006) and various logit models in Frühwirth-Schnatter and Frühwirth (2007). An important limitation of this approach, however, is that it is model-specific, and a sampler developed for one model is not generally applicable to other state space models.

There are two general approaches that are commonly used for estimating non-linear non-Gaussian state space models. The first is the so-called sequential Monte Carlo methods, or more popularly known as particle filtering (Doucet, De Freitas, and Gordon, 2001; Doucet and Johansen, 2011), which is a broad class of techniques that involves sequential importance sampling and bootstrap resampling. In the state space setting, particle filters are often used to evaluate the expected value of the states or functions of the states (such as the integrated likelihood) via sequential importance sampling

and resampling. For instance, particle filters have been applied to estimating non-linear DSGE models in Rubio-Ramirez and Fernandez-Villaverde (2005) and Fernandez-Villaverde and Rubio-Ramirez (2007). Despite recent advances, particle filters are still quite computationally intensive, especially when the dimension of the states is moderately high (e.g., when m is more than 5 or 6) or when the time series is long. For Bayesian estimation, it might take tens of hours to perform a full posterior analysis. In addition, particle filters are designed to evaluate expectations, not for efficient simulation of the states. That is, particle filters are not designed for generating draws from the conditional density $p(\eta \,|\, \theta, y)$, where $\eta = (\eta_1', \ldots, \eta_T')'$ and $y = (y_1', \ldots, y_T')'$. Without samples from $p(\eta \,|\, \theta, y)$, it is more difficult to design efficient MCMC sampling schemes to obtain posterior draws in a full Bayesian analysis. In fact, in posterior simulation with a particle filter, it is a common practice to generate candidate draws for θ via a random walk sampler, and then use a particle filter to compute the acceptance probability of the candidate draw. On the other hand, if one could generate efficiently from $p(\eta \,|\, \theta, y)$, one can use the machinery in the MCMC literature to design an efficient sampling scheme to generate draws from $p(\theta \,|\, \eta, y)$. More recently, there has been work on using particle filters to obtain candidate draws for the states (e.g., the particle Markov chain Monte Carlo methods of Andrieu, Doucet, and Holenstein, 2010). This direction looks promising, but most applications using these methods to date are limited to univariate state-space models due to computational limitations.

The alternative approach for estimating general state space models is based on fast approximations of the conditional density $p(\eta \,|\, \theta, y)$, where the approximating density is used for posterior simulation via the independence-chain Metropolis-Hastings algorithm (MH) algorithm. Of course, the main challenges are that it is crucial to have a fast routine to obtain a good approximation for $p(\eta \,|\, \theta, y)$, and it should be easy to generate candidate draws from the approximating density. Durbin and Koopman (1997) consider approximating the log target density $\log p(\eta \,|\, \theta, y)$ around the mode by a second order Taylor expansion. This approximation gives a Gaussian density, where its mean is the mode of the target density and its precision equals the negative Hessian evaluated at the mode. Candidate draws for the states are then generated via the Kalman filter and smoother. One problem with sampling η in a single block with the Gaussian proposal is that the acceptance rate in the MH step can be quite low, at least in the context of stochastic volatility models. It is therefore suggested that the states be divided into blocks, and each block is sampled sequentially via the MH

step. This methodology is implemented in, e.g., Shephard and Pitt (1997), Strickland *et al.* (2006) and Jungbacker and Koopman (2008). Departing from the obvious Gaussian approximation, McCausland (2008) recently introduced the HESSIAN method that provides an excellent approximation for $p(\eta \,|\, \theta, y)$, and it results in a highly efficient sampling algorithm. However, the HESSIAN method requires the first five derivatives of the log-likelihood with respect to the states, which places a substantial burden on the end-user. A more severe restriction is that currently it can only be applied to univariate state space models.[1]

We pursue the second line of research using approximations of the conditional density for the states, and propose various improvements. Specifically, the contributions of this paper are three-fold. In the first contribution we build upon the recently proposed precision-based sampler in Chan and Jeliazkov (2009b) and McCausland, Millera, and Pelletier (2011) originally developed for linear Gaussian state space models, and we present a quick method to obtain a Gaussian or a student t approximation for the conditional density of the states $p(\eta \,|\, \theta, y)$. By exploiting the sparseness structure of the precision matrix for $p(\eta \,|\, \theta, y)$, the precision-based algorithm is more efficient than Kalman filter-based methods in general. This feature is crucial as one needs to obtain the approximating density tens of thousands times in a full Bayesian analysis via MCMC. More importantly, the marginal cost of obtaining additional draws under the precision-based algorithm is much smaller compared to Kalman filter-based methods. We exploit this important feature for two purposes: efficient simulation of the states, and evaluation of the integrated likelihood. We develop an accept-reject Metropolis Hastings (ARMH) algorithm for efficient simulation of the states. As mentioned previously, the efficiency in the MH step with a Gaussian proposal can be quite low in certain settings, presumably because the Gaussian approximation is not sufficiently accurate. By using the ARMH algorithm, we construct a better approximation and consequently the acceptance rate is substantially higher compared to the baseline MH algorithm. This increased acceptance rate comes at a cost, however, as multiple draws from the proposal density might be required and this is why it is essential to have low marginal cost for additional draws. Next, we evaluate the integrated likelihood — an ingredient for maximum likelihood estimation and efficient

---

[1] One obvious way to get around this problem in multivariate-state settings is to draw each 'slice' of states one at a time by applying the HESSIAN method sequentially. However, when the (contemporary) states are expected to be highly correlated, this approach might induce high autocorrelations in the MCMC samples.

MCMC design — via importance sampling using multiple draws from the proposal density.[2]

The second contribution of the paper is to develop a practical way to sample the model parameters θ and the states η jointly. In performing a full Bayesian analysis, one often sequentially draws from the conditional densities $p(\eta \,|\, y, \theta)$ and $p(\theta \,|\, y, \eta)$. In typical situations where θ contains parameters in the state equation, η and θ are expected to be highly correlated. Consequently, the conventional sampling scheme might induce high auto-correlations for the samples, especially in high-dimensional settings. This motivates sampling θ and η jointly by first drawing from $p(\theta \,|\, y)$ marginally of the states η followed by a draw from $p(\eta \,|\, y, \theta)$. The challenge, of course, is to locate a good proposal density for θ, denoted as $q(\theta \,|\, y)$. We adopt the cross-entropy method (Rubinstein and Kroese, 2004) to obtain the optimal $q(\theta \,|\, y)$ in a well-defined sense. Specifically, given a parametric family of densities $\mathcal{G}$, we locate the member in $\mathcal{G}$ that is the closest to the marginal density $p(\theta \,|\, y)$ in the *Kullback-Leibler divergence* or the *cross-entropy distance*. By sampling (η, θ) jointly, we show via an empirical example that the efficiency of the sampling scheme is substantially improved. We note that the problem of locating a good proposal density for θ also arises in the particle filter literature as discussed above. Hence, the proposed cross-entropy approach is useful even if the researcher chooses to use a particle filter to evaluate the integrated likelihood instead of the algorithms discussed in this paper.

In the third contribution of the paper we demonstrate the overall approach with a topical application. We investigate the implications for transmission of monetary shocks of accounting for the zero lower bound (ZLB) on interest rates. Recent work using time-varying parameter vector autoregressive models (TVP-VARs) by Cogley and Sargent (2001 and 2005), Primiceri (2005), Sims and Zha (2006) and Koop, Leon-Gonzalez and Strachan (2009) demonstrated the importance of model specification for the evidence on changes in the transmission mechanism for monetary shocks. These studies considered periods of relatively high interest rates, with the exception of the period after the dot com bubble when interest rates fell as low as 1%. This latter event, and the history of Japan in the 1990s, led to an increase in interest in the effect of the ZLB on the conduct of monetary policy (see, for exam-

---

[2] A further advantage of the proposed method is that it can be applied to non-Markovian state equations, which arise in, e.g., various DSGE models, and they are more difficult to handle under other approaches.

ple, Iwata and Wu, 2006, Reifschneider and Williams, 2000, and Svensson, 2003). We study the effect of the ZLB on the transmission of a contractionary monetary shock in a low growth, low interest rate environment. We use a TVP-VAR with a censored interest rate variable and multivariate stochastic volatility. Modeling volatility has proven an important requirement for accurate inference in such models (see Cogley and Sargent, 2001, 2005, Primiceri, 2005 and Sims and Zha, 2006). Our interest is in how estimates of the impulse responses to positive monetary shocks change when we allow for the ZLB.

The rest of this article is organized as follows. Section 2 first briefly discusses the precision sampler in the linear Gaussian case. In Section 3 we consider a general form of the state space model for which we propose an approximation to the conditional density for the states, an approach to estimating the integrated likelihood in this case, and the three efficient simulation schemes for the states. In Section 4 we apply the sampler to estimate a standard model used a number of times in the literature to investigate the evolution of the transmission of monetary shocks, but we incorporate the zero lower bound on interest rates which results in a non-linear measurement equation. Section 5 concludes and discusses various future research directions.

2   The Linear Gaussian Case

In this section we present a precision-based sampler developed independently in Chan and Jeliazkov (2009b) and McCausland, Millera, and Pelletier (2011) for simulating the states in linear Gaussian state space models. By exploiting the sparseness structure of the precision matrix for the conditional density of the states, this new simulation algorithm is more efficient than Kalman filter-based methods in general. In addition, the marginal cost of obtaining additional draws using the precision-based algorithm is much smaller compared to Kalman filter-based methods, and we will take advantage of this fact to develop algorithms for more general model specifications in later sections.

Consider the following state space model:

$$y_t = X_t \eta_t + \varepsilon_t, \tag{1}$$
$$\eta_t = \Gamma_t \eta_{t-1} + \zeta_t, \tag{2}$$

for $t = 1, \ldots, T$, where y_t is an $n \times 1$ vector of observations, η_t is an $m \times 1$ latent state vector, and the disturbance terms are jointly Gaussian:

$$\begin{pmatrix} \varepsilon_t \\ \zeta_t \end{pmatrix} \sim \mathsf{N}\left(0, \begin{pmatrix} \Sigma_t^{-1} & 0 \\ 0 & \Omega_t^{-1} \end{pmatrix} \right). \tag{3}$$

That is, Σ_t and Ω_t are respectively the precision matrices of ε_t and ζ_t. The initial state η_0 can be assumed to be a known constant or treated as a model parameter. Define $y = (y_1', \ldots, y_T')'$ and $\eta = (\eta_1', \ldots, \eta_T')'$, and let θ represent the parameters in the state space model (i.e. $\eta_0, \{\Gamma_t\}, \{\Sigma_t\}$ and $\{\Omega_t\}$). The covariates $\{X_t\}$ are taken as given and will be suppressed in the conditioning sets below.

From (1) and (3) it is easily seen that the joint sampling density $p(y|\theta, \eta)$ is Gaussian. Stacking (1) over the T time periods, we have

$$y = X\eta + \varepsilon, \quad \varepsilon \sim \mathsf{N}(0, \Sigma^{-1}),$$

where

$$X = \begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_T \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix}, \quad \Sigma^{-1} = \begin{bmatrix} \Sigma_1^{-1} & & \\ & \ddots & \\ & & \Sigma_T^{-1} \end{bmatrix}.$$

A change of variable from ε to y implies that

$$\log p(y \mid \theta, \eta) \propto -\frac{1}{2} \log |\Sigma^{-1}| - \frac{1}{2} (y - X\eta)' \Sigma (y - X\eta). \tag{4}$$

It is important to realize that Σ is a banded matrix of the form

$$\Sigma = \begin{bmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_T \end{bmatrix}.$$

For the prior distribution of η, we note that the directed conditional structure for $p(\eta_t | \theta, \eta_{t-1})$ in (2) and the distributional assumption in (3) imply that the joint density for η is also Gaussian. To see this, define

$$K = \begin{pmatrix} I_m & & & & \\ -\Gamma_2 & I_m & & & \\ & -\Gamma_3 & I_m & & \\ & & \ddots & \ddots & \\ & & & -\Gamma_T & I_m \end{pmatrix} \quad \text{and} \quad \Omega = \begin{pmatrix} \Omega_1 & & & & \\ & \Omega_2 & & & \\ & & \Omega_3 & & \\ & & & \ddots & \\ & & & & \Omega_T \end{pmatrix},$$

9

so that (2) can be written as $K\eta = \gamma + \zeta$, where

$$
\gamma = \begin{bmatrix} \Gamma_1\eta_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{and} \quad \zeta = \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_T \end{bmatrix} \sim \mathsf{N}(0, \Omega^{-1}).
$$

Noting that $|K| = 1$, by a change of variable from ζ to η, we have

$$
\log p(\eta \,|\, \theta) \propto -\frac{1}{2}\log|\Omega^{-1}| - \frac{1}{2}(\eta - \eta^0)'K'\Omega K(\eta - \eta^0), \tag{5}
$$

where $\eta^0 = K^{-1}\gamma$ is the prior mean. Note that the $Tm \times Tm$ precision matrix $K'\Omega K$ is a banded matrix given by

$$
\begin{pmatrix}
\Gamma_2'\Omega_2\Gamma_2 + \Omega_1 & -\Gamma_2'\Omega_2 & & & \\
-\Omega_2\Gamma_2 & \Gamma_3'\Omega_3\Gamma_3 + \Omega_2 & -\Gamma_3'\Omega_3 & & \\
& \ddots & \ddots & \ddots & \\
& & -\Omega_{T-1}\Gamma_{T-1} & \Gamma_T'\Omega_T\Gamma_T + \Omega_{T-1} & -\Gamma_T'\Omega_T \\
& & & -\Omega_T\Gamma_T & \Omega_T
\end{pmatrix}. \tag{6}
$$

Since the likelihood function $p(y \,|\, \theta, \eta)$ in (4) and the prior $p(\eta \,|\, \theta)$ in (5) are both linear Gaussian in η, the standard update for Gaussian linear regression (see, e.g. Koop, 2003, p140–141) implies that the conditional posterior $p(\eta \,|\, y, \theta) \propto p(y \,|\, \theta, \eta)p(\eta \,|\, \theta)$ is also Gaussian. The log conditional density for the states can be written as

$$
\log p(\eta \,|\, y, \theta) \propto \log p(y \,|\, \theta, \eta) + \log p(\eta \,|\, \theta)
$$
$$
\propto -\frac{1}{2}\left[\eta'(X'\Sigma X + K'\Omega K)\eta - 2\eta'(X'\Sigma y + K'\Omega K\eta^0)\right].
$$

In other words,

$$
(\eta \,|\, y, \theta) \sim \mathsf{N}(\widehat{\eta}, H^{-1}), \tag{7}
$$

where the precision H and the mean $\widehat{\eta}$ are given by

$$
H = K'\Omega K + X'\Sigma X, \tag{8}
$$
$$
\widehat{\eta} = H^{-1}(K'\Omega K\eta^0 + X'\Sigma y). \tag{9}
$$

Since $X'\Sigma X$ is banded, it follows that H is also banded and contains a small number of non-zero elements on a narrow band around the main diagonal. An important consequence is that its Cholesky decomposition can

be obtained in $O(N)$ operations instead of $O(N^3)$ operations for full matrices, where N is the dimension of the matrix. By exploiting this fact, one can sample $(\eta \,|\, y, \theta)$ without the need to carry out an inversion to obtain H^{-1} and $\widehat{\eta}$ in (9). More specifically, the mean $\widehat{\eta}$ can be found in two steps. First, we compute the (banded) Cholesky decomposition C_H of H such that $C_H' C_H = H$. Second, we solve

$$C_H' C_H \widehat{\eta} = K'\Omega K \eta^0 + X'\Sigma y, \tag{10}$$

for $C_H \widehat{\eta}$ by forward-substitution and then using the result to solve for $\widehat{\eta}$ by back-substitution. Similarly, to obtain a random draw from $\mathsf{N}(\widehat{\eta}, H^{-1})$ efficiently, sample $u \sim \mathsf{N}(0, I_{Tm})$, and solve $C_H x = u$ for x by back-substitution. It follows that $x \sim \mathsf{N}(0, H^{-1})$. Adding the mean $\widehat{\eta}$ to x, one obtains a draw from $\mathsf{N}(\widehat{\eta}, H^{-1})$. We summarize the above procedures in the following algorithm.

Algorithm 1. *Efficient State Simulation for Linear Gaussian State Space Models*

1. *Compute H in (8) and obtain its Cholesky decomposition C_H such that $H = C_H' C_H$.*

2. *Solve (10) by forward- and back-substitution to obtain $\widehat{\eta}$.*

3. *Sample $u \sim \mathsf{N}(0, I_{Tm})$, and solve $C_H x = u$ for x by back-substitution. Take $\eta = \widehat{\eta} + x$, so that $\eta \sim \mathsf{N}(\widehat{\eta}, H^{-1})$.*

By counting the number of operations, McCausland, Millera, and Pelletier (2011) show that this precision-based algorithm for simulating the states is more efficient than conventional Kalman-filter based simulation methods when one draw is needed. When multiple samples are required, the marginal cost of obtaining an extra draw via the precision-based algorithm is substantially less compared with the latter methods. In fact, given $\widehat{\eta}$ and C_H, getting an additional draw from $\mathsf{N}(\widehat{\eta}, H^{-1})$ requires only (a) Tm independent standard Gaussian draws; (b) to perform a fast back-substitution to solve $C_H x = u$ for x; and (c) to add $\widehat{\eta}$ to x. We will exploit this important feature — low marginal costs for additional draws — to develop more sophisticated algorithms in Section 3.3.

3 General State Space Model

In this section we consider a general state space model where the measurement equation is characterized by a generic density function $p(y_t \mid \eta_t, \theta)$, whereas the state equation is linear Gaussian as in (2). We note that the proposed approach can be easily generalized to the case where the state equation is non-linear non-Gaussian or even non-Markovian.

3.1 Gaussian Approximation

We first discuss a quick method to obtain a Gaussian approximation for the conditional density $p(\eta \mid y, \theta)$. This approach builds upon the precision-based algorithm outlined in Section 2. To begin, let f_t and G_t denote respectively the gradient and negative Hessian of $\log p(y_t \mid \eta_t, \theta)$ evaluated at $\eta_t = \widetilde{\eta}_t$, i.e.,

$$f_t \equiv \frac{\partial}{\partial \eta_t} \log p(y_t \mid \eta_t, \theta)\Big|_{\eta_t = \widetilde{\eta}_t}, \quad G_t \equiv -\frac{\partial^2}{\partial \eta_t \eta_t'} \log p(y_t \mid \eta_t, \theta)\Big|_{\eta_t = \widetilde{\eta}_t}.$$

Stacking these terms and define the following vector and matrix:

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_T \end{bmatrix}, \quad G = \begin{bmatrix} G_1 & 0 & \cdots & 0 \\ 0 & G_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & G_T \end{bmatrix}.$$

We then expand the log-likelihood $\log p(y \mid \eta, \theta) = \sum_{t=1}^{T} \log p(y_t \mid \eta_t, \theta)$ around $\widetilde{\eta} = (\widetilde{\eta}_1', \ldots, \widetilde{\eta}_T')'$ to obtain the expression

$$\log p(y \mid \eta, \theta) \approx \log p(y \mid \widetilde{\eta}, \theta) + (\eta - \widetilde{\eta})'f - \frac{1}{2}(\eta - \widetilde{\eta})'G(\eta - \widetilde{\eta})$$
$$= -\frac{1}{2}\left[\eta'G\eta - 2\eta'(f + G\widetilde{\eta})\right] + c_1, \tag{11}$$

where c_1 is some unimportant constant independent of η. Combining (11) and the prior in (5), we have

$$\log p(\eta \mid y, \theta) \propto \log p(y \mid \eta, \theta) + \log p(\eta \mid \theta)$$
$$\approx -\frac{1}{2}\left[\eta'(G + K'\Omega K)\eta - 2\eta'(f + G\widetilde{\eta} + K'\Omega K\eta^0)\right] + c_2. \tag{12}$$

where c_2 is some unimportant constant independent of η. In other words, the approximating distribution is Gaussian with precision $H \equiv G + K'\Omega K$ and mean vector $H^{-1}\left(f + G\widetilde{\eta} + K'\Omega K\eta^0\right)$.

It remains to choose the point $\widetilde{\eta}$ around which to construct the Taylor expansion. One obvious choice is the posterior mode, denoted as $\widehat{\eta}$, which has the advantage that it can be easily obtained via the Newton-Raphson method. More specifically, it follows from (12) that the negative Hessian of $\log p(\eta \mid y, \theta)$ evaluated at $\eta = \widetilde{\eta}$ is H, while the gradient at $\eta = \widetilde{\eta}$ is given by

$$\left.\frac{\partial}{\partial \eta}\log p(\eta \mid y, \theta)\right|_{\eta=\widetilde{\eta}} = -H\widetilde{\eta} + 2(f + G\widetilde{\eta} + K'\Omega K\eta^0).$$

Hence, we can implement the Newton-Raphson method as follows: initialize with $\eta = \eta^{(1)}$. For $s = 1, 2, \ldots$, use $\widetilde{\eta} = \eta^{(s)}$ in the evaluation of f, G and H, and denote them as $f(\eta^{(s)})$, $G(\eta^{(s)})$ and $H(\eta^{(s)})$ respectively, where the dependence on $\eta^{(s)}$ is made explicit. Compute $\eta^{(s+1)}$ as

$$\begin{aligned}
\eta^{(s+1)} &= \eta^{(s)} + H(\eta^{(s)})^{-1}\left.\frac{\partial}{\partial \eta}\log p(\eta \mid y, \theta)\right|_{\eta=\eta^{(s)}} \\
&= H(\eta^{(s)})^{-1}\left(f(\eta^{(s)}) + G(\eta^{(s)})\eta^{(s)} + K'\Omega K\eta^0\right).
\end{aligned} \tag{13}$$

If $||\eta^{(s+1)} - \eta^{(s)}|| > \epsilon$ for some pre-fixed tolerance level ϵ, then continue; otherwise stop and set $\widehat{\eta} = \eta^{(s+1)}$. Again, it is important to note that because the precision H is banded, and its Cholesky decomposition C_H can be readily obtained. Hence, (13) can be efficiently evaluated without inverting any high-dimensional matrix. Following the approach discussed in Section 2, we compute $\eta^{(s+1)}$ as follows: given the Cholesky decomposition C_H for $H(\eta^{(s)})$, first solve $C_H'^{(s)}x = f\left(\eta^{(s)}\right) + G(\eta^{(s)})\eta^{(s)} + K'\Omega K\eta^0$ for x by forward-substitution. Then given x, solve $C_H\eta^{(s+1)} = x$ for $\eta^{(s+1)}$ by back-substitution. Finally, given the mode $\widehat{\eta}$, the negative Hessian H at $\widehat{\eta}$ can be easily computed.

3.2 Integrated Likelihood Evaluation

The integrated likelihood $p(y \mid \theta)$ is defined as the joint distribution of the data conditional on the parameter vector θ but integrated over the states η. More explicitly,

$$p(y \mid \theta) = \int p(y \mid \theta, \eta)p(\eta \mid \theta)\mathrm{d}\eta. \tag{14}$$

13

The need to evaluate the integrated likelihood efficiently arises in both frequentist and Bayesian estimation. In classical inference, one needs to maximize the integrated likelihood $p(y \,|\, \theta)$ with respect to θ to obtain the maximum likelihood estimator (MLE). In our context one has to compute the MLE numerically, and the maximization routine typically requires hundreds or even thousands of functional evaluations of $p(y \,|\, \theta)$. Hence, it is crucial to be able to evaluate the integrated likelihood efficiently. For Bayesian estimation, if one can evaluate $p(y \,|\, \theta)$ quickly, more efficient samplers can be developed to obtain draws from the posterior, such as the collapsed sampler that draws η and θ jointly in a single step which we discuss in Section 3.3.

Given the Gaussian approximation proposed in the previous section, one can estimate $p(y \,|\, \theta)$ via importance sampling (see, e.g., Geweke, 1989; Kroese *et al.*, 2011, ch. 9). To do this, sample M independent draws $\eta^1, \ldots, \eta^M$ from the proposal density $q(\eta \,|\, y, \theta)$, and compute the Monte Carlo average

$$\widehat{p}(y \,|\, \theta) = \frac{1}{M} \sum_{i=1}^{M} \frac{p(y \,|\, \theta, \eta^i) p(\eta^i \,|\, \theta)}{q(\eta^i \,|\, y, \theta)}.$$

It is easy to see that the Monte Carlo estimator $\widehat{p}(y \,|\, \theta)$ is an unbiased and consistent estimator for $p(y \,|\, \theta)$. In addition, if the likelihood ratio $p(y \,|\, \theta, \eta) p(\eta \,|\, \theta)/q(\eta \,|\, y, \theta)$ or equivalently $p(\eta \,|\, y, \theta)/q(\eta \,|\, y, \theta)$ is bounded for all η, then the variance of the estimator is also finite (Geweke, 1989). The proposed precision-based algorithms are especially fit for evaluating the integrated likelihood via importance sampling. This is because one needs multiple draws (often hundreds or thousands) from the proposal density $q(\eta \,|\, y, \theta)$ to compute the Monte Carlo average. As discussed earlier, one important and useful feature of the precision-based algorithms is that once we obtained the mean vector and precision matrix, additional draws can be obtained with little marginal cost.

3.3 Efficient Simulation for the States

Building upon the Gaussian approximation presented previously, we propose three different sampling schemes for drawing the states efficiently.

3.3.1 Metropolis-Hastings with Gaussian and t proposals

A simple sampling scheme is to implement a Metropolis-Hastings step with proposal density $\mathsf{N}(\widehat{\eta}, H^{-1})$. The mode $\widehat{\eta}$ and the negative Hessian at $\widehat{\eta}$ of the conditional density $p(\eta \,|\, y, \theta)$ can be computed as discussed in earlier. Moreover, a draw from the proposal can be obtained as in Algorithm 1. Using a Gaussian approximation would be adequate in models where either the measurement or the state equations is Gaussian, as the resulting conditional posterior in either case has exponentially decaying tails. We summarize this basic sampling scheme as follows:

Algorithm 2. *Metropolis-Hastings with the Gaussian Proposal $\mathsf{N}(\widehat{\eta}, H^{-1})$*

1. *Obtain $\widehat{\eta}$ iteratively via (13). Given H, compute its Cholesky decomposition C_H such that $H = C_H' C_H$.*

2. *Sample $u \sim \mathsf{N}(0, I_{Tm})$, and solve $C_H x = u$ for x by back-substitution. Take $\eta = \widehat{\eta} + x$, so that $\eta \sim \mathsf{N}(\widehat{\eta}, H^{-1})$.*

In implementing the Metropolis-Hastings algorithm, it is often suggested that the proposal density $q(\eta \,|\, y, \theta)$ should have heavier tails than the posterior distribution $p(\eta \,|\, y, \theta)$, so that the likelihood ratio $p(\eta \,|\, y, \theta)/q(\eta \,|\, y, \theta)$ is bounded. This is important because a bounded likelihood ratio ensures the geometric ergodicity of the Markov chain (Roberts and Rosenthal, 2004). In the context of estimating the integrated likelihood, this guarantees the estimator has finite variance. Thus, one concern of using a Gaussian proposal is that it has exponentially decaying tails, and consequently, the likelihood ratio might not be bounded. This motivates using a proposal density with heavier tails, such a t distribution. We note that one can easily modify the above Gaussian approximation to obtain a t proposal density instead. More explicitly, consider the t proposal $\eta \sim q(\eta \,|\, y, \theta) \equiv \mathsf{t}(\nu, \widehat{\eta}, H^{-1})$ with degree of freedom parameter ν, location vector $\widehat{\eta}$ and scale matrix H^{-1}. Note that sampling from $\mathsf{t}(\nu, \widehat{\eta}, H^{-1})$ involves only Tm iid standard Gaussian draws and a draw from the $\mathsf{Gamma}(\nu/2, \nu/2)$ distribution. We summarize the algorithm as follows:

Algorithm 3. *Metropolis-Hastings with the t proposal $\mathsf{t}(\nu, \widehat{\eta}, H^{-1})$*

1. *Given the posterior mode $\widehat{\eta}$ and negative Hessian H, obtain the Cholesky decomposition C_H such that $H = C_H' C_H$.*

2. *Sample $u \sim \mathsf{N}(0, I_{Tm})$ and $r \sim \mathsf{Gamma}(\nu/2, \nu/2)$. Then $v \equiv u/\sqrt{r} \sim \mathsf{t}(\nu, 0, I_{Tm})$.*

3. *Solve $C_H x = v$ for x by back-substitution and take $\eta = \widehat{\eta} + x$, so that $\eta \sim \mathsf{t}(\nu, \widehat{\eta}, H^{-1})$.*

3.3.2 Accept-Reject Metropolis-Hastings

As its name suggests, the accept-reject Metropolis-Hastings (ARMH) algorithm (Tierney, 1994; Chib and Greenberg, 1995) is an MCMC sampling procedure that combines classic accept-reject sampling with the Metropolis-Hastings algorithm. In the our setting the target density is the conditional density of the states $p(\eta \,|\, y, \theta) \propto p(y \,|\, \eta, \theta) p(\eta \,|\, \theta)$. Suppose we have a proposal density $q(\eta \,|\, y, \theta)$ from which we generate candidate draws (e.g. $q(\eta \,|\, y, \theta)$ can be the Gaussian or t density discussed in the previous section). In the classic accept-reject sampling a key requirement is that there exists a constant c such that

$$p(y \,|\, \eta, \theta) p(\eta \,|\, \theta) \leqslant c q(\eta \,|\, y, \theta), \tag{15}$$

for all η in the support of $p(\eta \,|\, y, \theta)$. When η is a high-dimensional vector, as in the present case, such a constant c, if it exists, is usually difficult to obtain. To make matters worse, the target density $p(\eta \,|\, y, \theta)$ depends on other model parameters θ that are revised at every iteration. Finding a new value of c for each new set of parameters might significantly increase the computational costs. The ARMH relaxes the domination condition (15) such that when it is not satisfied for some η, we resort to the MH algorithm. To present the algorithm, it is convenient to first define the set

$$\mathcal{D} = \{\eta : p(y \,|\, \eta, \theta) p(\eta \,|\, \theta) \leqslant c q(\eta \,|\, y, \theta)\},$$

and let $\mathcal{D}^c$ denote its complement. Then the ARMH algorithm proceeds as follows:

Algorithm 4. *Accept-Reject Metropolis-Hastings with Gaussian or t proposal*

1. *AR step: Generate a draw $\eta^* \sim q(\eta \,|\, y, \theta)$, where $q(\eta \,|\, y, \theta)$ is the Gaussian or t proposal obtained in Algorithms 2 or 3. Accept η^* with*

16

probability

$$\alpha_{AR}(\eta^* \,|\, y, \theta) = \min\left\{1, \frac{p(y \,|\, \eta^*, \theta)p(\eta^* \,|\, \theta)}{cq(\eta^* \,|\, y, \theta)}\right\}.$$

Continue the above process until a draw η^* is accepted.

2. *MH-step: Given the current draw η and the proposal η^**

   (a) *if $\eta \in \mathcal{D}$, set $\alpha_{MH}(\eta, \eta^* \,|\, y, \theta) = 1$;*

   (b) *if $\eta \in \mathcal{D}^c$ and $\eta^* \in \mathcal{D}$, set*

$$\alpha_{MH}(\eta, \eta^* \,|\, y, \theta) = \frac{cq(\eta \,|\, y, \theta)}{p(y \,|\, \eta, \theta)p(\eta \,|\, \theta)};$$

   (c) *if $\eta \in \mathcal{D}^c$ and $\eta^* \in \mathcal{D}^c$, set*

$$\alpha_{MH}(\eta, \eta^* \,|\, y, \theta) = \min\left\{1, \frac{p(y \,|\, \eta^*, \theta)p(\eta^* \,|\, \theta)q(\eta \,|\, y, \theta)}{p(y \,|\, \eta, \theta)p(\eta \,|\, \theta)q(\eta^* \,|\, y, \theta)}\right\}.$$

Return η^* with probability $\alpha_{MH}(\eta, \eta^* \,|\, y, \theta)$; otherwise return η.

As shown in Chib and Greenberg (1995), the draws produced at the completion of the AR step have the density

$$q_{\mathrm{AR}}(\eta \,|\, y, \theta) = d^{-1}\alpha_{\mathrm{AR}}(\eta \,|\, y, \theta)q(\eta \,|\, y, \theta),$$

where d is the normalizing constant (which needs not be known for implementing the algorithm). In other words, one might view the AR step as a means to sample from the density $q_{\mathrm{AR}}(\eta \,|\, y, \theta)$. By adjusting the original proposal density $q(\eta \,|\, y, \theta)$ by the function $\alpha_{\mathrm{AR}}(\eta \,|\, y, \theta)$, a better approximation of the target density is achieved. In fact, we have

$$q_{\mathrm{AR}}(\eta \,|\, y, \theta) = \begin{cases} p(y \,|\, \eta, \theta)p(\eta \,|\, \theta)/cd, & \eta \in \mathcal{D}, \\ q(\eta \,|\, y, \theta)/d, & \eta \in \mathcal{D}^c, \end{cases}$$

i.e., the new proposal density coincides with the target density on the set $\mathcal{D}$ (albeit with different normalizing constants), whereas on $\mathcal{D}^c$ the new proposal density is reduced to the original one. To give a feeling for the improvement this approach brings, consider Figure 1. In this figure, the true density is shown as a grey shaded area and the Gaussian candidate by the dotted line. The candidate density $q_{\mathrm{AR}}(\eta \,|\, y, \theta)$ is shown as the solid line which fits

the true density for values less than 1.6, then differs above this point but still fits better than the Gaussian density. The better approximation, of course, comes at a cost, because multiple draws from the proposal density $q(\eta \mid y, \theta)$ might be required in the AR step. This is where the precision-based method (as in Algorithms 2 or 3) comes in. As we have emphasized before, the marginal cost of generating additional draws using the precision-based method is low, and is substantially lower than generating candidate draws via Kalman filter-based algorithms. In fact, as demonstrated in the application, the gain in efficiency under the ARMH sampling scheme more than justifies its additional cost compared to a plain MH step.
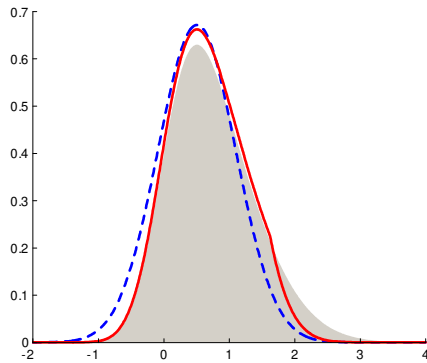

Figure 1: Illustration of the two approximations for a skew normal distribution (shaded): Gaussian (dotted line) and the same Gaussian with AR adjustment (solid line).

Chib and Jeliazkov (2005) present a practical way to select the constant c and the trade-off in such a choice which we outline here. Notice that if a bigger c is chosen, then the set $\mathcal{D}$ is larger and we are more likely to accept the candidate η^*. The cost, on the other hand, of selecting a larger c is that more draws from $q(\eta \mid y, \theta)$ are required in the AR step. A practical way to strike a balance between these two conflicting considerations is to set $c = rp(y \mid \widehat{\eta}, \theta)p(\widehat{\eta} \mid \theta)/q(\widehat{\eta} \mid y, \theta)$, where $\widehat{\eta}$ is the mode of the conditional density $p(\eta \mid y, \theta)$ and r is, say, between 1 and 5. Such a choice would ensure that c is sufficiently small to reduce the required number of draws from $q(\eta \mid y, \theta)$, while big enough so that the set $\mathcal{D}$ contains the mode $\widehat{\eta}$ and its neighboring points.

3.3.3 Collapsed Sampling with the Cross-entropy Method

We have so far discussed two sampling schemes for efficient simulation from the conditional density $p(\eta \mid y, \theta)$: the MH and the ARMH algorithms with either a Gaussian or a t proposal. In performing a full Bayesian analysis, one often sequentially draws from $p(\eta \mid y, \theta)$ followed by sampling from $p(\theta \mid y, \eta)$. In typical situations where θ contains parameters in the state equation, η and θ are expected to be highly correlated. Consequently, the conventional sampling scheme might induce high autocorrelation and slow mixing in the Markov chain, especially in high-dimensional settings. For this reason, we seek to sample (θ, η) jointly by first drawing from $p(\theta \mid y)$ marginally of the states η followed by a draw from $p(\eta \mid y, \theta)$, where the latter step can be accomplished by either the MH or ARMH algorithm previously discussed. To sample from $p(\theta \mid y)$, we again implement a MH step: we first generate a candidate draw θ^* from the proposal density $g(\theta)$, then we decide whether to accept θ^* or not according to the acceptance probability. Hence, we need two ingredients: (1) a quick routine to evaluate the integrated likelihood $p(y \mid \theta)$, which arises in computing the acceptance probability; and (2) a good proposal density $g(\theta)$ for generating candidate draws for the MH step.

The first ingredient, an efficient method to evaluate the integrated likelihood, is provided by the importance sampling estimator $\widehat{p}(y \mid \theta)$ discussed in Section 3.2. And this in turn gives us an estimator for the acceptance probability

$$\alpha\left(\theta \mid y\right) = \min\left\{1, \frac{p(y \mid \theta^*)p\left(\theta^*\right)g(\theta)}{p(y \mid \theta)p(\theta)g(\theta^*)}\right\}.$$

One might raise the concern that the simulation error may affect the convergence properties of the Markov chain, as the candidate draws are accepted or rejected according to *estimated* acceptance probabilities rather than the actual values. However, since the importance sampling estimator $\widehat{p}(y \mid \theta)$ is unbiased, the results in Andrieu, Berthelsen, Doucet, and Roberts (2007) and Flury and Shephard (2008) show that the stationary distribution of the constructed Markov chain is the posterior distribution as desired.

The second ingredient is a proposal density for generating candidate draws for θ. Of course, one may generate candidates via a random walk, but this strategy is not recommended as a random walk chain is typically inefficient, and it would defeat the purpose — to improve the mixing properties of the Markov chain — of the whole exercise. Therefore, it is essential to locate a good proposal density $g(\theta)$ to implement an independence-chain MH step.

We adopt the so-called cross-entropy adaptive independence sampler introduced in Keith, Kroese, and Sofronov (2008). Specifically, the proposal density is chosen such that the *Kullback-Leibler divergence*, or the *cross-entropy* (CE)*distance* between the proposal density and the target (the posterior density) is minimal, where the CE distance between the densities g_1 and g_2 is defined as:

$$\mathcal{D}(g_1, g_2) = \int g_1(x) \log \frac{g_1(x)}{g_2(x)} \mathrm{d}x.$$

Let $\mathcal{G}$ be a parametric family of densities $g(\cdot; v)$ indexed by the parameter vector v. Minimizing the CE distance is equivalent to finding

$$v_{\mathrm{ce}} = \underset{v}{\mathrm{argmax}} \int p(\theta \,|\, y) \log g(\theta; v) \, \mathrm{d}\theta.$$

As in the CE method (Rubinstein and Kroese, 2004; Kroese, Taimre, and Botev, 2011, ch. 13), we can estimate the optimal solution v_{ce} by

$$\widehat{v}_{\mathrm{ce}} = \underset{v}{\mathrm{argmax}} \frac{1}{N} \sum_{i=1}^{N} \log g(\theta_i; v), \tag{16}$$

where $\theta_1, \ldots, \theta_N$ are draws from the marginal posterior density $p(\theta \,|\, y)$. The solution to the maximization problem in (16) is typically easy to obtain; in fact, analytic solutions are often available. On the other hand, finding $\widehat{v}_{\mathrm{ce}}$ requires a pre-run to obtain a small sample from $p(\theta \,|\, y)$. This can be achieved by sequentially drawing from $p(\eta \,|\, y, \theta)$ and $p(\theta \,|\, y, \eta)$, as discussed in the previous section. It is important to note that although the sample obtained in this pre-run may exhibit slow mixing, we only use it to obtain the proposal density, and thus it has little adverse effect on the main collapsed sampler. Once we find $\widehat{v}_{\mathrm{ce}}$, we then use the proposal density $g(\theta; \widehat{v}_{\mathrm{ce}})$ to implement the independence-chain MH step. We discuss in more details the implementation in Appendix B.

4   Application

To illustrate the proposed approach we estimate a VAR with a lower bound restriction on one of the variables; this restriction implies a measurement equation that is non-linear in the states. Specifically, we investigate the implications for the transmission of monetary shocks of accounting for the zero

lower bound (ZLB) on interest rates. With time varying parameters, incorporating the lower bound on interest rates introduces a non-linearity in the states into the measurement equation. Recent work using time-varying parameter vector autoregressive models (TVP-VARs) on changes in the transmission mechanism for monetary policy shocks (see for example, Cogley and Sargent, 2001, 2005, Primiceri, 2005, Sims and Zha, 2006, and Koop, Leon-Gonzalez and Strachan, 2009) has ignored the lower bound on interest rates. Not accounting for the ZLB is reasonable when interest rates are relatively high and far from zero. However, episodes of low interest rates have occurred often in recent history including, as examples, in the US just after the dot com bubble of 2001, during the 1990s in Japan, or since 2009 in much of the developed world. The prevalence of low interest rates suggests it is important to know whether transmission of monetary shocks is affected and, if so, to understand how the transmission mechanism is affected. Our focus is upon the effect of a contractionary monetary shock when interest rates are on the ZLB. Such a situation might arise for the US if several rating agencies were to downgrade the rating of US government debt and creditors then began to demand a premium to compensate for the risk of default, or if the cost of funds to banks increased independently of moves in the Federal Funds rate inducing an effective, unintended tightening of monetary policy.[3]

4.1   The Model

The framework we consider is the following time-varying parameter vector autoregressive (TVP-VAR) model with l lags:

$$y_t = \mu_t + A_{1t}y_{t-1} + \cdots + A_{lt}y_{t-l} + \epsilon_t, \quad \epsilon_t \sim \mathsf{N}(0, \Sigma_t^{-1}),$$

where μ_t is an $n \times 1$ vector of time-varying intercepts, $A_{t1}, \ldots, A_{tl}$ are $n \times n$ matrices of VAR lag coefficients at time t, and Σ_t^{-1} is a time-varying precision matrix. For the purpose of estimation, we write the VAR system in the form of seemingly unrelated regressions:

$$y_t = x_t \beta_t + \epsilon_t, \quad \epsilon_t \sim \mathsf{N}(0, \Sigma_t^{-1}), \tag{17}$$

---

[3]This effect was observed in February 2012 in Australia when, after the central bank kept its rate unchanged, all banks increased their lending rates in response to increased costs of wholesale funding costs.

It is for this reason we do not term the shock a monetary policy shock, but an unintended shock to monetary conditions.

where $x_t = I_n \otimes [1, y'_{t-1}, \ldots, y'_{t-l}]$ and $\beta_t = \text{vec}([\mu_t : A_{1t} : \cdots : A_{lt}]')$ is a $k \times 1$ vector of VAR coefficients with $k = n^2 l + n$. To model the time-varying precision matrix Σ_t, we follow the approach proposed in Primiceri (2005) by first factoring the precision matrix as $\Sigma_t = L'_t D_t^{-1} L_t$, where $D_t = \text{diag}(e^{h_{1t}}, \ldots, e^{h_{nt}})$ is a diagonal matrix, and L_t is a lower triangular matrix with ones on the main diagonal, i.e.,

$$
D_t = \begin{pmatrix} e^{h_{1t}} & 0 & \cdots & 0 \\ 0 & e^{h_{2t}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{h_{nt}} \end{pmatrix}, \quad L_t = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21,t} & 1 & 0 & \cdots & 0 \\ a_{31,t} & a_{32,t} & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1,t} & a_{n2,t} & \cdots & a_{n(n-1),t} & 1 \end{pmatrix}.
$$

This decomposition has been employed in various applications, especially in the context of efficient estimation of covariance matrices (Pourahmadi, 1999, 2000, Smith and Kohn, 2002, Chan and Jeliazkov, 2009a, among others). In the setting of VAR models with time-varying volatility, this approach is first considered in Cogley and Sargent (2005). For notational convenience, we let $h_{\boldsymbol{\cdot} t} = (h_{1t}, \ldots, h_{nt})'$ and $h_{i \boldsymbol{\cdot}} = (h_{i1}, \ldots, h_{iT})'$. That is, $h_{\boldsymbol{\cdot} t}$ is the $n \times 1$ vector obtained by stacking h_{it} by the first subscript, whereas $h_{i\boldsymbol{\cdot}}$ is the $T \times 1$ vector obtained by stacking h_{it} by the second subscript. The log-volatilities $h_{\boldsymbol{\cdot} t}$ evolve according to the state equation

$$
h_{\boldsymbol{\cdot} t} = h_{\boldsymbol{\cdot} t-1} + \xi_t, \quad \xi_t \sim \mathsf{N}(0, \Omega_h^{-1}), \tag{18}
$$

for $t = 2, \ldots, T$, where $\Omega_h = \text{diag}(\omega_{h1}, \ldots, \omega_{hn})$ is a diagonal matrix. The process is initialized with $h_{\boldsymbol{\cdot} 1} \sim \mathsf{N}(0, V_h^{-1})$ for some known diagonal precision matrix V_h. Let a_t denote the free elements in L_t ordered by rows, i.e., $a_t \equiv a_{\boldsymbol{\cdot} t} = (a_{21,t}, a_{31,t}, a_{32,t}, \ldots, a_{n(n-1),t})'$, so that a_t is an $m \times 1$ vector of parameters where $m = n(n-1)/2$. The evolution of a_t is modeled as a random walk

$$
a_t = a_{t-1} + \zeta_t, \quad \zeta_t \sim \mathsf{N}(0, \Omega_a^{-1}), \tag{19}
$$

for $t = 2, \ldots, T$, where $\Omega_a = \text{diag}(\omega_{a1}, \ldots, \omega_{am})$ is a diagonal precision matrix. The process is initialized with $a_1 \sim \mathsf{N}(0, V_a^{-1})$ for some known diagonal precision matrix V_a. In what follows we use these two parameterizations, namely, Σ_t and $(h_{\boldsymbol{\cdot} t}, a_t)$, interchangeably. To complete the specification of the model, it remains to specify the evolution of the VAR coefficients β_t. We follow the standard approach of modeling the VAR coefficients β_t as a random walk process:

$$
\beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \sim \mathsf{N}(0, \Omega_\beta^{-1}), \tag{20}
$$

for $t = 2, \ldots, T$, where $\Omega_\beta = \text{diag}(\omega_{\beta 1}, \ldots, \omega_{\beta k})$ is a diagonal precision matrix. The process is initialized with $\beta_1 \sim \mathsf{N}(0, V_\beta^{-1})$ for some known precision matrix V_β.

After presenting the basic setup of a TVP-VAR model with stochastic volatility, we now wish to impose the restriction that the nominal interest rate is always non-negative. For this purpose, arrange the data y_t so that y_{1t}, the first element of y_t, is the nominal interest rate, and let x_{1t} be the first row of x_t. We assume that $y_{1t} \geqslant 0$. Consequently, given β_t and Σ_t, y_t follows a multivariate Gaussian distribution with the first element restricted to be positive. To derive the likelihood function, first note that since only y_{1t} is constrained while other elements of y_t are not, the marginal distribution of y_{1t} is a univariate Gaussian variable truncated below at 0. In fact, it can be easily shown that

$$(y_{1t} \mid \beta_t, \Sigma_t) \sim \mathsf{N}(x_{1t}\beta_t, \mathrm{e}^{h_{1t}})\mathbb{1}(y_{1t} \geqslant 0).$$

It follows that given β_t and Σ_t, we have

$$\mathbb{P}(y_{1t} \geqslant 0 \mid \beta_t, \Sigma_t) = 1 - \Phi\left(-x_{1t}\beta_t / \mathrm{e}^{\frac{1}{2}h_{1t}}\right) = \Phi\left(x_{1t}\beta_t \mathrm{e}^{-\frac{1}{2}h_{1t}}\right),$$

where $\Phi(\cdot)$ denotes the standard Gaussian cumulative distribution function. Letting $y = (y_1', \ldots, y_T')'$, $\beta = (\beta_1', \ldots, \beta_T')'$ and $\Sigma = (\Sigma_1, \ldots, \Sigma_T)$, the log-likelihood function is thus

$$\log p(y \mid \beta, \Sigma) = \sum_{t=1}^{T} \log p(y_t \mid \beta_t, \Sigma_t), \tag{21}$$

where

$$p(y_t \mid \beta_t, \Sigma_t) \propto -\frac{1}{2} \log |\Sigma_t^{-1}| - \frac{1}{2}(y_t - x_t\beta_t)'\Sigma_t(y_t - x_t\beta_t) - \log \Phi\left(x_{1t}\beta_t \mathrm{e}^{-\frac{1}{2}h_{1t}}\right).$$

4.2   Prior and Estimation

Given the measurement equation (21) and the state equations (18)–(20), we present a Markov sampler that builds upon the approximation methods discussed in last section to obtain a sample from the posterior distribution. To this end, we first specify the priors for the remaining parameters: $\omega_\beta = (\omega_{\beta 1}, \ldots, \omega_{\beta k})'$, $\omega_h = (\omega_{h1}, \ldots, \omega_{hk})'$, and $\omega_a = (\omega_{a1}, \ldots, \omega_{am})'$.

Specifically, the elements of ω_β, ω_h and ω_a follow independently Gamma distributions: $\omega_{\beta i} \sim \mathsf{Gamma}(r_{\beta i}, s_{\beta i})$ for $i = 1, \ldots, k$, $\omega_{hi} \sim \mathsf{Gamma}(r_{hi}, s_{hi})$, for $i = 1, \ldots, n$, and $\omega_{ai} \sim \mathsf{Gamma}(r_{ai}, s_{ai})$, for $i = 1, \ldots, m$. For later reference, we stack $h = (h'_{\cdot 1}, \ldots, h'_{\cdot T})'$ and $a = (a'_1, \ldots, a'_T)'$, and let θ denote the set of parameters except the latent states β, h and a, i.e., $\theta = (\omega_\beta, \omega_h, \omega_a)$.

In what follows, we briefly discuss the implementation of the three samplers; we refer the readers to Appendices A and B for more details. The first sampling scheme is the baseline Metropolis-Hastings sampler that involves sequentially drawing from:

a. $p(\beta \,|\, y, h, a, \theta)$ via an MH step;

b. $p(h \,|\, y, \beta, a, \theta)$ via an MH step;

c. $p(a \,|\, y, \beta, h, \theta)$ via a Gibbs step;

d. $p(\theta \,|\, y, \beta, h, a)$ via a Gibbs step.

To efficiently sample the states β in the non-linear state space model (20) and (21), we consider implementing an independence-chain MH step by approximating the conditional distribution $p(\beta \,|\, y, h, a, \theta)$ via a Gaussian distribution as discussed in Section 3.1. The next step is to sample from the conditional distribution $p(h \,|\, y, \beta, a, \theta)$. Recall that h_{it} is the i-th diagonal element in D_t, $h_{\cdot t} = (h_{1t}, \ldots, h_{nt})'$ and $h_{i\cdot} = (h_{i1}, \ldots, h_{iT})'$. Note that we are able to write $\log p(h \,|\, y, a, \beta, \theta) = \sum_{i=1}^n \log p(h_{i\cdot} \,|\, y, a, \beta, \theta)$. What this means is that to obtain a draw from $p(h \,|\, y, a, \beta, \theta)$, we can instead sample from $p(h_{i\cdot} \,|\, y, a, \beta, \theta)$ sequentially without adversely affecting the efficiency of the sampler. Now, a draw from $p(h_{i\cdot} \,|\, y, a, \beta, \theta)$ can be obtained via an independence-chain Metropolis-Hastings step with a Gaussian proposal density; more details are given in Appendix A. Thirdly, it can be easily shown that $p(a \,|\, y, \beta, h, \theta)$ is a Gaussian distribution (see, e.g. Primiceri, 2005), and a draw from which can be obtained using Algorithm 1. Finally, $p(\theta \,|\, y, \beta, h, a)$ is a product of Gamma densities, and a draw from which is standard (see Koop, 2003, p. 61-62).

In the second sampling scheme, we also sequentially drawing from the four full conditional densities as before. The only difference is that instead of using the MH algorithm to sample $p(\beta \,|\, y, h, a, \theta)$ and $p(h \,|\, y, \beta, a, \theta)$, we use the ARMH algorithm described in Section 3.3.2. Finally, in the third sampling scheme, we sample

a. $p(\omega_\beta \,|\, y, h, a)$ marginally via an MH step, followed by $p(\beta \,|\, y, h, a, \theta)$ via an ARMH step;

b. $p(\omega_h \,|\, y, \beta, a)$ marginally via an MH step, followed by $p(h \,|\, y, \beta, a, \theta)$ via an ARMH step;

c. $p(\omega_a \,|\, y, \beta, h)$ marginally via an MH step, followed by $p(a \,|\, y, \beta, h, \theta)$ via a Gibb step.

The details for the collapsed sampler are given in Appendix B.

4.3 Empirical Results

We now present empirical results based on a set of U.S. macroeconomic variables commonly used in the study of the evolution of monetary policy transmission. We have an interest rate to capture effects of monetary conditions, a real growth rate variable to capture the state of the economy, and inflation. The dataset is obtained from the U.S. Federal Reserve Bank at St. Louis website that consists quarterly observations from 1947Q1 to the 2011Q2 on the following $n = 3$ U.S. macroeconomic series: U.S. 3-month Treasury bill rate, CPI inflation rate, and real GDP growth. Both the CPI inflation rate and real GDP growth are computed via the formula $400(\log(z_t) - \log(z_{t-1}))$, where z_t is the original quarterly CPI or GDP figures. The inclusion of the interest rate variable which is bounded below at zero, provides a useful example to demonstrate our methods. The plot the evolution of the interest rate, given in Figure 2, shows that since the start of the quantitative easing in late 2008, the 3-month Tbill rate has become essentially zero.

We allow one lag in the VAR as this seems sufficient to capture much of the dynamics. We begin with a comparison of performance of the three sampling schemes. To this end, we estimate the *restricted* model using the three different sampling schemes outlined in the previous section: the Metropolis-Hastings sampler (S1), the Accept-Reject Metropolis-Hastings sampler (S2), and the collapsed sampler with cross-entropy method (S3). We also include results from the *unrestricted* model (U), for which both the transition and measurement equations are linear Gaussian. As such, this
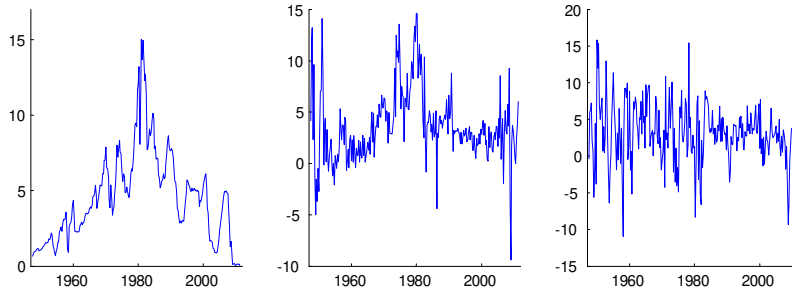
Figure 2: The U.S. 3-month Tbill rate (left), CPI inflation rate (middle), and GDP growth rate (right) from 1947 Q1 to 2011 Q2.

unrestricted model can be estimated using the standard outlined in Section 2.

One popular measure of MCMC efficiency is the inefficiency factor, defined as:

$$
1 + 2 \sum_{j=1}^{J} \phi_j,
$$

where ϕ_j is the sample autocorrelation at lag length j, and J is chosen large enough so that the autocorrelation tapers off. This statistic approximates the ratio of the numerical variance of the posterior mean from the MCMC output relative to that from hypothetical iid draws. As the posterior draws from the Markov chain become less serially correlated, the ratio will approach the ideal minimum value of 1. In the presence of inefficiency due to serial correlation in the draws, the ratio will be larger than 1. Figure 3 presents the boxplots of the inefficiency factors for the four sampling schemes.

Remember that the unrestricted model is linear Gaussian and can be estimated via standard Gibbs sampler. In contrast, the restricted model that incorporates the ZLB is non-linear, and the conditional densities of the states are non-standard. Since the proposed samplers need to approximate these conditional densities, we would generally expect that they would not perform as well compared to the standard Gibbs sampler used to estimate the *unrestricted* model. As evidenced by the plots in Figure 3, the proposed samplers do not perform substantially worse at estimating the non-linear
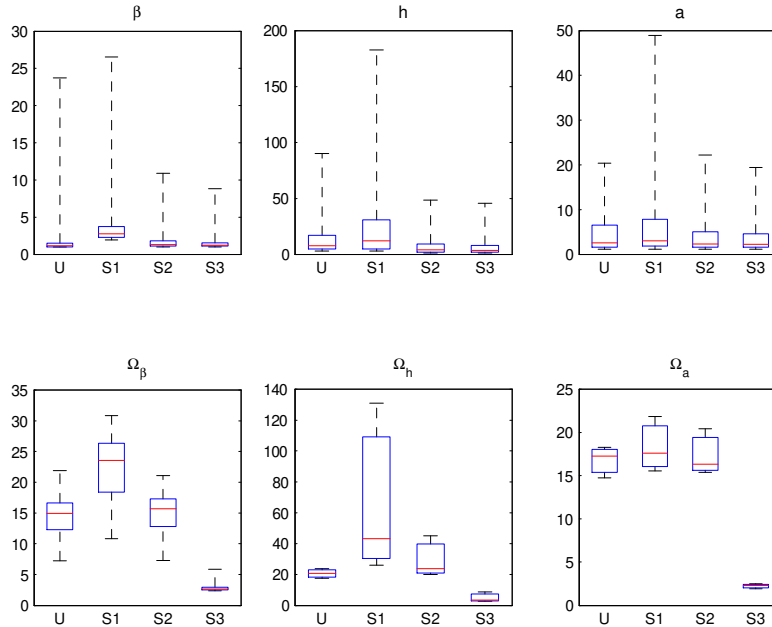
26

Figure 3: Boxplots of the inefficiency factors for the unrestricted linear Gaussian model (U), and the three sampling schemes: MH (S1), ARMH (S2) and the collapsed sampler (S3). The central mark of each box is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the maximum and minimum.

model than the standard precision sampler for fitting the linear Gaussian model. The collapsed sampler with cross-entropy, S3, has inefficiency factors as small as or smaller than the other samplers for most parameters, including the standard sampler for the unrestricted model. The improvement in efficiency for S3 is most substantial in the precision of the state parameters, which is significant as these are hyperparameters and are typically not as well estimated as parameters that appear in the measurement equation. These hyperparameters are also important as they play an important role in the estimation of the states. We see that S1 (the MH sampler) is generally not as efficient compared to the other samplers, although its performance is not greatly worse than the others. The efficiency of S2 (the ARMH sampler)

Table 1: Acceptance rate (in %) and the computing time (in minutes) of the three sampling schemes: MH (S1), ARMH (S2) and the collapsed sampler with CE (S3).

	β	$h_{1.}$	$h_{2.}$	$h_{3.}$	Ω_β	Ω_h	Ω_a	Time
S1	68	28	35	59	–	–	–	23
S2	95	71	79	97	–	–	–	27
S3	98	69	79	97	62	58	76	182

is as good as S3 for the states, but it is worse than S3 for the estimation of the state precisions.

In Table 1 we present the acceptance rates of draws from candidates, as well as the computation time for obtaining 50,000 draws. On the whole the three samplers are relatively fast and have reasonable acceptance rates. These results are more significant given our high-dimensional model: β has more than 3,000 elements and h has more than 750. To compare among the three sampling schemes, we see that although S1 is relatively fast, it can have low acceptance rates particularly for the log volatilities. By contrast, S2 has higher acceptance rates at the expense of only a little more computation time. Although S3 is more efficient relative to S2 in terms of lower inefficiency factors, its computation time is almost seven times compared to that of S2. For our model and dataset, it would seem that S2 is the best among the three.

We now present empirical results for the restricted model estimated using the ARMH sampler (S2). For comparison, we also report the corresponding results for the unrestricted model. This comparison is provided to demonstrate the implications for inference of neglecting the restriction of the ZLB. We begin with a discussion of the implications of the restriction for parameter estimation and then show the effect on impulse responses of not correctly accounting for the ZLB. These differences are significant and justify the new estimation methods presented in this paper.

The effect of neglecting the ZLB restriction shows up in all blocks of parameters: the variances; the correlations; and mean equation coefficients. Figure 4 shows the estimated log-volatilities and correlations for the restricted model and the unrestricted model. The figure for the log-volatilities of the monetary shock, $h_{1,t}$, shows that ignoring the restriction would lead to a significant underestimation of this parameter in the period since 2005.
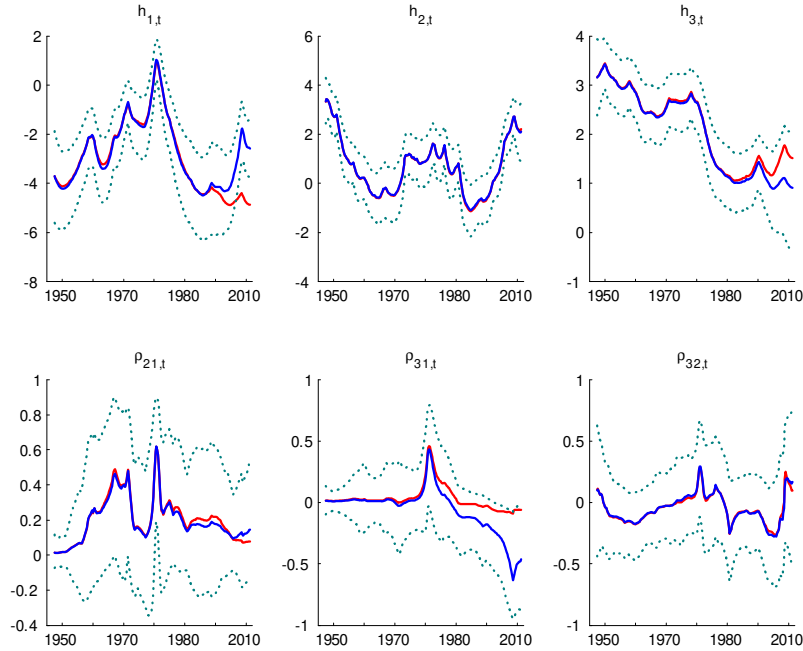
Figure 4: Evolution of the log-volatilities and correlations. The solid red line is the estimated posterior mean under the unrestricted model. The solid blue line is the estimated posterior mean and the dotted green lines are the 5%-tile and 95%-tile, respectively, under the model with the inequality restrictions imposed.

The monetary shock volatility is much higher than the unrestricted model suggests. Similarly the volatility of real activity, $h_{3,t}$, is over estimated when the ZLB is ignored. The volatility and correlations of the nominal variable shock does not show much influence from the ZLB. However, the correlation of the error from the interest rate equation with the error from the growth equation is strongly affected, this correlation would be estimated as being near zero rather than very negative. This effect has important implications for the impulse responses.

The plots in Figure 5 show the effect of the ZLB on the impulse response functions of the three variables to a monetary shock. We produce the im-

29

pulse responses as the differences in forecasts. However, due to the non-linear form of the model, these impulses are not the standard ones derived from the VMA representation in linear models. We forecast the variables using the parameter values at 2011 Q2, assuming the parameters cease evolving, but taking into account the ZLB into these forecasts. We then forecast again, but increase the error in the interest rate equation, the monetary shock, by 0.25%. Our impulse responses are the differences between these two forecasts. We see that there is a faster response of interest rates to the monetary shock, but the form of the response is similar.
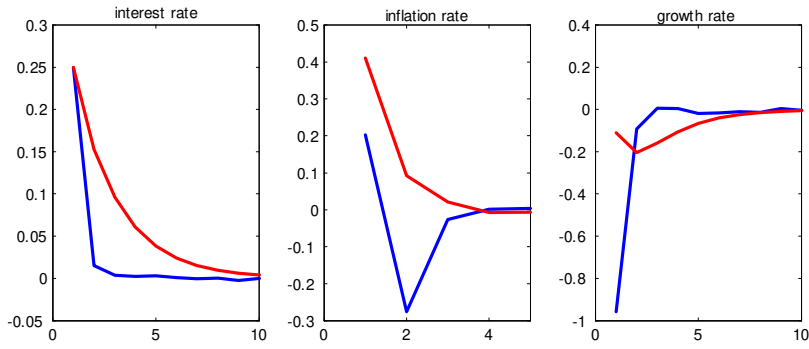

Figure 5: Impulse response of a 0.25% increase in interest rate under the unrestricted model (red solid line) and the model with the inequality restrictions imposed (blue solid line).

The responses of inflation and growth to this shock are very different with the ZLB imposed compared to that without the restriction. As these results come from a time varying parameter model, it is important to interpret these responses in the context of the economic environment at the time. In 2011 Q2 inflation and inflation volatility were increasing, while growth and volatility of the error in the growth equation were both low. The response of inflation with the ZLB is initially positive but then falls by a larger amount than the initial response. This pattern contrasts with the steady decline toward zero from the initial shock we would conclude was the response from the unrestricted model. The difference in the response of growth to this shock under the two specifications is even more stark. While the shock is negative but small for the unrestricted model, when we account for the ZLB we see the initial response is negative and very large. The differences of these responses would lead to very different assessments of the risks of

30

unanticipated monetary shocks.

5   Concluding Remarks and Further Research

In this paper we have proposed a new approach to efficiently estimate high-dimensional non-linear non-Gaussian state space models. Due to the general applicability of the proposed approach, it will prove useful in a wide range of applications. We extend the recently developed precision-based samplers (Chan and Jeliazkov, 2009b and McCausland, Millera, and Pelletier, 2011) and sparse matrix procedures to build fast, efficient samplers for these non-linear models. We develop a practical way to sample the model parameters θ and the states η jointly to circumvent the problem of high autocorrelations in high-dimensional settings. This approach uses the cross-entropy method (Rubinstein and Kroese, 2004) to obtain the optimal candidate densities $q(\theta \,|\, y)$. We show via an empirical example that the efficiency of the sampling scheme is substantially improved by drawing (θ, y) jointly. Three samplers are presented each with virtues in different circumstances. Finally, we apply these techniques in a TVP-VAR in which one of the variables is restricted to be strictly positive. Using this framework, we investigate the implications for transmission of monetary shocks of accounting for the zero lower bound (ZLB) on interest rates.

Another advantage of the proposed method is that it can be applied to non-Markovian state equations, which arise in, e.g., various non-linear DSGE models, and they are more difficult to handle under other approaches. Therefore in future work we will apply this approach to the estimation of non-linear DSGE models. Another direction will be in models with measurement equations that involving more than current, past or even future states.

Appendix A: Efficient Simulation of β and h

In this appendix we provide the details of the independence-chain Metropolis-Hastings step for sampling from $p(\beta \,|\, y, h, a, \theta)$ and $p(h \,|\, y, a, \beta, \theta)$. We use

the decomposition $\Sigma_t = L'_t D_t^{-1} L_t$, where

$$
D_t = \begin{pmatrix} e^{h_{1t}} & 0 & \cdots & 0 \\ 0 & e^{h_{2t}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{h_{nt}} \end{pmatrix}, \quad L_t = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21,t} & 1 & 0 & \cdots & 0 \\ a_{31,t} & a_{32,t} & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1,t} & a_{n2,t} & \cdots & a_{n(n-1),t} & 1 \end{pmatrix}.
$$

Recall that h_{it} is the i-th diagonal element in D_t, $h_{\cdot t} = (h_{1t}, \ldots, h_{nt})'$ and $h_{i\cdot} = (h_{i1}, \ldots, h_{iT})'$. That is, $h_{\cdot t}$ is the $n \times 1$ vector obtained by stacking h_{it} by the first subscript, whereas $h_{i\cdot}$ is the $T \times 1$ vector obtained by stacking h_{it} by the second subscript. Also, a_t denotes the free elements in L_t ordered by rows, i.e., $a_t = (a_{21,t}, a_{31,t}, a_{32,t}, \ldots, a_{n(n-1),t})'$. In what follows we use the two parameterizations Σ_t and $(h_{\cdot t}, a_t)$ interchangeably. Then the log-density for y_t given (β_t, Σ_t) is

$$
\log p(y_t \mid \beta_t, \Sigma_t) \propto -\frac{1}{2}(y_t - x_t \beta_t)' \Sigma_t (y_t - x_t \beta_t) - \log \Phi(\alpha_t),
$$

where $\alpha_t = x_{1t} \beta_t e^{-\frac{1}{2} h_{1t}}$. Using the notation in Section 3.1, we have

$$
f_t \equiv \frac{\partial}{\partial \beta_t} \log p(y_t \mid \beta_t, \Sigma_t) \Big|_{\beta_t = \widetilde{\beta}_t}, \quad G_t \equiv -\frac{\partial^2}{\partial \beta_t \beta'_t} \log p(y_t \mid \beta_t, \Sigma_t) \Big|_{\beta_t = \widetilde{\beta}_t},
$$

where

$$
\frac{\partial}{\partial \beta_t} \log p(y_t \mid \beta_t, \Sigma_t) = x'_t \Sigma_t (y_t - x_t \beta_t) - \frac{\phi(\alpha_t)}{\Phi(\alpha_t)} e^{-\frac{1}{2} h_{1t}} x'_{1t},
$$

$$
\frac{\partial^2}{\partial \beta_t \beta'_t} \log p(y_t \mid \beta_t, \Sigma_t) = -x'_t \Sigma_t x_t + \frac{\phi(\alpha_t)}{\Phi(\alpha_t)} e^{-h_{1t}} \left(\alpha_t + \frac{\phi(\alpha_t)}{\Phi(\alpha_t)} \right) x'_{1t} x_{1t},
$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard Gaussian probability density function and cumulative distribution function respectively. Given f_t and G_t, we can then use the Gaussian or t approximations in Section 3.1 as a proposal density.

We now discuss sampling from the conditional density $p(h \mid y, a, \beta, \theta)$. We first show that

$$
\log p(h \mid y, a, \beta, \theta) = \sum_{i=1}^{n} \log p(h_{i\cdot} \mid y, a, \beta, \theta).
$$

Put differently, to obtain a draw from $p(h \,|\, y, a, \beta, \theta)$, we can instead sample from $p(h_{i\cdot} \,|\, y, a, \beta, \theta)$ sequentially without adversely affecting the efficiency of the sampler. To this end, decompose $\Sigma_t = L_t' D_t^{-1} L_t$ as before. Since $\log |\Sigma_t| = \log |D_t| = \sum_{i=1}^{n} h_{it}$ and $\sigma_{t,11}^2 = e^{h_{1t}}$, it follows that the log-likelihood is given by

$$\log p(y \,|\, \beta, h, a, \theta) \propto \sum_{t=1}^{T} \left[-\frac{1}{2} \sum_{i=1}^{n} h_{it} - \frac{1}{2}(L_t \epsilon_t)' D_t^{-1} L_t \epsilon_t - \log \Phi \left(e^{-h_{1t}/2} x_{1t} \beta_t \right) \right],$$

$$= \sum_{t=1}^{T} \left[-\frac{1}{2} \sum_{i=1}^{n} h_{it} - \frac{1}{2} \sum_{i=1}^{n} e^{-h_{it}} s_{it}^2 - \log \Phi \left(e^{-h_{1t}/2} x_{1t} \beta_t \right) \right],$$

$$\text{(22)}$$

where $\epsilon_t = y_t - x_t \beta_t$ and s_{it}^2 is the i-th diagonal element of $(L_t \epsilon_t)(L_t \epsilon_t)'$. On the other hand, the state equation (18) implies that each $h_{i\cdot}$ follows independently a Gaussian distribution. In fact, we have

$$h_{it} = h_{i,t-1} + \xi_{it}, \quad \xi_{it} \sim \mathsf{N}(0, \omega_{hi}). \tag{23}$$

Hence, it follows from (22) and (23) that $h_{i\cdot}, i = 1, \ldots, n$ are conditionally independent given the data and other parameters.

We note that although one can apply the auxiliary variable approach in Kim, Shepherd, and Chib (1998) to sample from $p(h_{i\cdot} \,|\, y, a, \beta, \theta)$ for $i = 2, \ldots, n$, it cannot be used to draw from $p(h_{1\cdot} \,|\, y, a, \beta, \theta)$ due to the extra term $\log \Phi \left(e^{-h_{1t}/2} x_{1t} \beta_t \right)$ in the log-likelihood (22) that depends on β_t. Instead, we sample each $h_{i\cdot}$ sequentially via an independence-chain Metropolis-Hastings step. As before, we first derive an expression for a second order Taylor expansion of the log-likelihood (22) around the posterior mode $\widehat{h}_{i\cdot} = (\widehat{h}_{i1}, \ldots, \widehat{h}_{iT})'$. Define $\gamma_t = e^{-h_{1t}/2} x_{1t} \beta_t$

$$q_{it} = \left. \frac{\partial}{\partial h_{it}} \log p(y \,|\, \beta, h, a, \theta) \right|_{h_{it} = \widehat{h}_{it}}, \quad r_{it} = \left. -\frac{\partial^2}{\partial h_{it}^2} \log p(y \,|\, \beta, h, a, \theta) \right|_{h_{it} = \widehat{h}_{it}},$$

$q_i = (q_{i1}, \ldots, q_{iT})'$ and $R_i = \mathrm{diag}(r_{i1}, \ldots, r_{iT})$, where

$$\frac{\partial}{\partial h_{it}} \log p(y \,|\, \beta, h, a, \theta) = \frac{1}{2} \left[e^{-h_{it}} s_{it}^2 - 1 + \gamma_t \frac{\phi(\gamma_t)}{\Phi(\gamma_t)} \mathbb{1}(i = 1) \right],$$

and

$$\frac{\partial^2}{\partial h_{it}^2} \log p(y \,|\, \beta, h, \theta) = -\frac{1}{2} e^{-h_{it}} s_{it}^2 + \frac{1}{4} \gamma_t \frac{\phi(\gamma_t)}{\Phi(\gamma_t)} \left[\gamma_t^2 + \gamma_t \frac{\phi(\gamma_t)}{\Phi(\gamma_t)} - 1 \right] \mathbb{1}(i = 1).$$

If we expand the log-likelihood (22) around the mode $\widehat{h}_{i\boldsymbol{\cdot}}$, we have

$$\log p(y \mid \beta, h, a, \theta) \approx -\frac{1}{2}\left[\widehat{h}_{i\boldsymbol{\cdot}}' R_i \widehat{h}_{i\boldsymbol{\cdot}} - 2\widehat{h}_{i\boldsymbol{\cdot}}'(q_i + R_i \widehat{h}_{i\boldsymbol{\cdot}})\right] + c_3,$$

where c_3 is some unimportant constant independent of $\widehat{h}_{i\boldsymbol{\cdot}}$. We consider the proposal density $\mathsf{N}(\widehat{h}_{i\boldsymbol{\cdot}}, (q_i + R_i \widehat{h}_{i\boldsymbol{\cdot}})^{-1})$, and everything follows as before.

Appendix B: Collapsed Sampler with the Cross-Entropy Method

In this appendix we provide the details on the collapsed sampler used in the third sampling scheme. In a nutshell, we first use a small posterior sample from a pre-run and the cross-entropy method to locate an optimal proposal density within a given parametric family. Then given a candidate draw from the proposal, we implement a Metropolis-Hastings step to decide whether or not to accept the candidate, where the acceptance probability is computed using the importance sampling estimator for the integrated likelihood proposed in Section 3.2. We focus on discussing the approximation to $p(\omega_\beta \mid y, h, a)$, where $\omega_\beta = (\omega_{\beta 1}, \ldots, \omega_{\beta k})'$. The approximations to the other two marginal densities follow similarly. Recall that the elements of ω_β have an independent gamma prior: $\omega_{\beta i} \sim \mathsf{Gamma}(r_{\beta i}, s_{\beta i})$ for $i = 1, \ldots, k$. Therefore, a natural parametric family within which to locate the proposal density is the gamma family:

$$\mathcal{G} = \left\{\prod_{i=1}^{k} f_{\mathsf{G}}(\omega_{\beta i}; c_{\beta i}, d_{\beta i})\right\},$$

where $f_{\mathsf{G}}(\cdot; c, d)$ is the density of $\mathsf{Gamma}(c, d)$. Given the R posterior draws $\{\omega_{\beta 1}^{(j)}, \ldots, \omega_{\beta k}^{(j)}\}$, $j = 1, \ldots, R$, we solve the CE optimization problem in (16) to obtain $\widehat{v}_{\mathrm{ce}} = (\widehat{c}_{\beta 1}, \widehat{d}_{\beta 1}, \ldots, \widehat{c}_{\beta k}, \widehat{d}_{\beta k})$. Specifically, the optimal CE reference parameter vector $\widehat{v}_{\mathrm{ce}}$ can be obtained as follows. First note that $\widehat{d}_{\beta i}$ can be solved analytically given $c_{\beta i}$:

$$\widehat{d}_{\beta i} = \frac{Rc_{\beta i}}{\sum_{j=1}^{R} \omega_{\beta i}^{(j)}}.$$

Now by substituting $d_{\beta i} = \widehat{d}_{\beta i}$ into the density $f_{\mathsf{G}}(\cdot; c_{\beta i}, d_{\beta i})$, $\widehat{c}_{\beta i}$ can be obtained by any one-dimensional root-finding algorithm (e.g., Newton-Raphson

method). Hence, we can obtain $(\widehat{c}_{\beta 1}, \widehat{d}_{\beta 1}, \ldots, \widehat{c}_{\beta k}, \widehat{d}_{\beta k})$ easily. Finally, the proposal density is

$$f_\beta(\omega_\beta) = \prod_{i=1}^{k} f_{\mathsf{G}}(\omega_{\beta i}; \widehat{c}_{\beta i}, \widehat{d}_{\beta i}),$$

which is the member within $\mathcal{G}$ that is the closest in cross-entropy divergence to the marginal density $p(\omega_\beta \,|\, y, h, a)$.

References

Andrieu, C. , K. K. Berthelsen, A. Doucet, and G. O. Roberts. The expected auxiliary variable method for Monte Carlo simulation. Technical report, 2007.

Andrieu, C., A. Doucet and R. Holenstein. Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society Series B, 72:269–342, 2010.

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

Chan, J. C. C., and I. Jeliazkov. MCMC estimation of restricted covariance matrix. Journal of Computational and Graphical Statistics, 18:457–480, 2009a.

Chan, J. C. C., and I. Jeliazkov. Efficient simulation and integrated likelihood estimation in state space models. International Journal of Mathematical Modeling and Numerical Optimisation, 1:101–120, 2009b.

Chib, S., and E. Greenberg. Understanding the Metropolis-Hastings algorithm. The American Statistician, 49(4):327–335, 1995.

Chib, S., and I. Jeliazkov. Accept-reject Metropolis-Hastings sampling and marginal likelihood estimation. Statistica Neerlandica, 59:30–44, 2005.

Chib, S., F. Nardari, and N. Shephard. Markov chain Monte Carlo methods for stochastic volatility models. Journal of Econometrics, 108(2):281–316, 2002.

Cogley, T. and T. J. Sargent. Evolving post-World War II inflation dynamics, *NBER Macroeconomic Annual*, 16, 331-373, 2001.

Cogley, T., and T. J. Sargent. Drifts and volatilities: monetary policies and outcomes in the post WWII US. Review of Economic Dynamics, 8(2):262 – 302, 2005.

de Jong, P., and N. Shephard. The simulation smoother for time series models. Biometrika, 82:339–350, 1995.

Doucet, A., and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovskii, editors, The Oxford Handbook of Nonlinear Filtering. Oxford University Press, Oxford, 2011.

Doucet, A., N. De Freitas, and N.J. Gordon, editors. Sequential Monte Carlo Methods in Practice. Springer, New York, 2001.

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

Durbin, J., and S. J. Koopman. A simple and efficient simulation smoother for state space time series analysis. Biometrika, 89:603–615, 2002.

Fernandez-Villaverde, J., and J. F. Rubio-Ramirez. Estimating macroeconomic models: A likelihood approach. Review of Economic Studies, 74(4):1059–1087, 2007.

Flury, T., and N. Shephard. Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. Economics SeriesWorking Papers 413, University of Oxford, Department of Economics, 2008.

Früwirth-Schnatter, S. Data augmentation and dynamic linear models. Journal of Time Series Analysis, 15:183–202, 1994.

Frühwirth-Schnatter, S., and R. Frühwirth. Auxiliary mixture sampling with applications to logistic models. Computational Statistics & Data Analysis, 51(7):3509–3528, 2007.

Frühwirth-Schnatter, S., and H. Wagner. Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. Biometrika, 93:827–841, 2006.

Geweke, J. Bayesian inference in econometric models using Monte Carlo integration. Econometrica, 57 (6):1317–1339, 1989.

Iwata, S. and S. Wu. Estimating monetary policy effects when interest rates are close to zero. Journal of Monetary Economics, Elsevier, vol. 53(7), pages 1395-1408, October, 2006.

Jungbacker, B., and S. J. Koopman. Monte Carlo estimation for nonlinear non-Gaussian state space models. Biometrika, 94:827–839, 2008.

Keith, J. M., D. P. Kroese, and G. Y. Sofronov. Adaptive independence samplers. Statistics and Computing, 18:409–420, 2008.

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

Koop, G. Bayesian Econometrics. Wiley & Sons, New York, 2003.

Koop G., R. Léon-González and R. W. Strachan. On the Evolution of Monetary Policy. Journal of Economic Dynamics and Control 33, 997-1017, 2009.

Kroese, D. P., T. Taimre, and Z. I. Botev. Handbook of Monte Carlo Methods. John Wiley & Sons, New York, 2011.

McCausland, W. J.. The HESSIAN method (Highly Efficient State Smoothing, In A Nutshell). University of Montreal Department of Economics Working Paper Series, 2008-03, 2008.

McCausland, W. J., S. Millera, and D. Pelletier. Simulation smoothing for state-space models: A computational efficiency analysis. Computational Statistics and Data Analysis, 55:199–212, 2011.

Pourahmadi, M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. Biometrika, 86:677–690, 1999.

Pourahmadi, M. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. Biometrika, 87:425–435, 2000.

Primiceri, G. E. Time varying structural vector autoregressions and monetary policy. Review of Economic Studies, 72(3):821–852, 2005.

Reifschneider, D. and J. C. Williams. Three lessons for monetary policy in a low inflation era. Journal of Money, Credit and Banking 32:936-966, 2000.

Roberts, G. O., and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. Probability Surveys, 1:20–71, 2004.

Rubinstein, R. Y., and D. P. Kroese. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization Monte-Carlo Simulation, and Machine Learning. Springer-Verlag, New York, 2004.

Rubio-Ramirez, J. F., and J. Fernandez-Villaverde. Estimating dynamic equilibrium economies: linear versus nonlinear likelihood. Journal of Applied Econometrics, 20(7):891–910, 2005.

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

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

Sims, C. and T. Zha, Were there regime switches in macroeconomic policy? American Economic Review, 96, 54-81, 2006.

Smith, M., and R. Kohn. Parsimonious covariance matrix estimation for longitudinal data. Journal of the American Statistical Association, 97:1141–1153, 2002.

Strickland, C. M., C. S. Forbes, and G. M.Martin. Bayesian analysis of the stochastic conditional duration model. Computational Statistics and Data Analysis, 50:2247–2267, 2006.

Tierney, L. Markov chains for exploring posterior distributions. The Annals of Statistics, 22(4):1701– 1728, 1994.