Efficient Simulation and Integrated Likelihood Estimation in Non-Linear Non-Gaussian State Space Models

Joshua C. C. Chan*
Centre for Applied Macroeconomic Analysis,
Australian National University

Rodney W. Strachan†
Centre for Applied Macroeconomic Analysis,
University of Queensland and
Rimini Center for Economic Analysis

October 16, 2013

Abstract

We propose a generic approach to inference in the non-linear, non-Gaussian state space model. This approach builds on recent developments in precision-based algorithms to estimating general state space models with multivariate observations and states. The baseline algorithm approximates the conditional distribution of the states by a multivariate $t$ density, which is then used for integrated likelihood estimation via importance sampling or for posterior simulation using Markov chain Monte Carlo (MCMC). We build further upon this baseline approach to consider more sophisticated algorithms such as accept-reject Metropolis-Hasting and variational approximation. To illustrate the proposed approach, we estimate the risk of a liquidity trap in the US under a time-varying parameter vector autoregressive (TVP-VAR) model with stochastic volatility.

Keywords: non-Gaussian; state space; accept-reject Metropolis-Hastings; variational approximation; cross-entropy method; liquidity trap

Acknowledgment: we gratefully acknowledge financial support from the Australian Research Council through Discovery Grant DP0987170.

*Email: joshua.chan@anu.edu.au
†Email: r.strachan@uq.edu.au
1 Introduction

We propose a general approach to inference in the non-linear non-Gaussian state space model. 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 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. In addition, the approximating density can also be used for evaluating the integrated likelihood — the joint distribution of the observations given the model parameters but integrated over the states — via importance sampling. Furthermore, we build upon this baseline approach to consider two other more efficient algorithms for posterior simulation. The first is the accept-reject Metropolis-Hastings (ARMH) algorithm that combines the classic accept-reject sampling and the Metropolis-Hastings algorithm. The second 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.

Specifically, the 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 $p(y_t \mid \eta_t, \theta)$, where $\eta_t$ is an $m \times 1$ vector of latent states and $\theta$ denotes the set of model parameters. Note that the density $p(y_t \mid \eta_t, \theta)$ may depend on previous observations $y_{t-1}, y_{t-2}, \text{etc.}$ and other covariates; they are suppressed in the conditioning sets for notational convenience. The evolution of the states $\eta_t$, in turn, is specified by the state or transition equation summarized by the density function $p(\eta_t \mid \eta_{t-1}, \theta)$. 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 $\eta_{t-1}, \eta_{t-2}, \text{etc.}$; see the discussion in Section 5.

A vast collection of models can be written in the general state space form with measurement equation $p(y_t \mid \eta_t, \theta)$ and state equation $p(\eta_t \mid \eta_{t-1}, \theta)$. Popular models 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), Frühwirth-Schnatter (1994), de Jong and Shephard (1995) and Durbin and Koopman (2002); more recently, precision-based algorithms are proposed in Rue (2001), 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). One major drawback 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 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 filter (Doucet, De Freitas, and Gordon, 2001; Doucet and Johansen, 2011), which is a board class of techniques that involves sequential importance sampling and bootstrap resampling. In the state space setting, it is 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, it has been applied to estimating DSGE models in Rubio-Ramirez and Fernandez-Villaverde (2005) and Fernandez-Villaverde and Rubio-Ramirez (2007). Despite recent advances, particle filter is 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 filter is designed to evaluate expectation, not for efficient simulation of the states, i.e., 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 scheme to obtain posterior draws in a full Bayesian analysis. In fact, in posterior simulation with particle filter, it is a common practice to generate candidate draws for \( \theta \) via a random walk sampler, and then use 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 all the machinery in the MCMC literature to design efficient sampling scheme to generate draws from \( p(\theta | \eta, y) \).

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. In contrast to particle filter, this appears to be the only general approach for efficient simulation of the states. Of course, the main challenge is 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) and Shephard and Pitt (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 \( \eta \) 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 where the dimension of the states is large. It is therefore suggested dividing the states 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 introduces 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 models.

We pursue the second line of research and propose various improvements. In particular, the contributions of this paper are three-folds. First, building upon the recently proposed precision-based sampler in Chan and Jeliazkov (2009b) and McCausland, Miller, and Pelletier (2011) originally developed for linear Gaussian state space models, we present a quick method to
obtain a Gaussian or a t approximation for the conditional density of the states \( p(\eta \mid \theta, y) \). By exploiting the sparseness structure of the precision matrix for \( p(\eta \mid \theta, y) \), 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: one, we develop an accept-reject Metropolis Hastings (ARMH) algorithm for efficient simulation of the states. As mentioned previously, the acceptance rate 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 comes at a cost, however, as multiple draws from the proposal density might be required. That is why it is essential to have low marginal cost for additional draws. Two, we evaluate the integrated likelihood — an ingredient for maximum likelihood estimation and efficient MCMC design — via importance sampling that requires multiple draws from the proposal density.

Second, we develop a practical way to sample the model parameters \( \theta \) and the states \( \eta \) jointly. In performing a full Bayesian analysis, one often sequentially draws from the conditional densities \( p(\eta \mid y, \theta) \) and \( p(\theta \mid y, \eta) \). In typical situations where \( \theta \) contains parameters in the state equation, \( \eta \) and \( \theta \) are expected to be highly correlated. Consequently, the conventional sampling scheme might induce high autocorrelations for the samples, especially in high-dimensional settings. This motivates sampling \( \theta \) and \( \eta \) jointly by first drawing from \( p(\theta \mid y) \) marginally of the states \( \eta \) followed by a draw from \( p(\eta \mid y, \theta) \). The challenge, of course, is to locate a good proposal density for \( \theta \), denoted as \( q(\theta \mid y) \). We adopt the cross-entropy method (Rubinstein and Kroese, 2004) to obtain the optimal \( q(\theta \mid y) \) in a well-defined sense. Specifically, given a parametric family of densities \( \mathcal{P} \), we locate the member in \( \mathcal{P} \) that is the closest to the marginal density \( p(\eta \mid y) \) in the Kullback-Leibler divergence or the cross-entropy distance. By sampling \((\eta, \theta)\) 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 \( \theta \) 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 particle filter to evaluate the integrated likelihood instead of the algorithms discussed in this paper.

Third, to illustrate the proposed MCMC samplers, we...

The rest of this article is organized as follows. Section 2 first briefly discusses in the linear Gaussian case.

## 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 more sophisticated algorithms in later sections. For now, consider the following state space model:

\begin{align}
    y_t &= X_t \eta_t + \varepsilon_t, \\
    \eta_t &= \Gamma_t \eta_{t-1} + \zeta_t,
\end{align}

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

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

That is, \( \Sigma_t \) and \( \Omega_t \) are respectively the precision of \( \varepsilon_t \) and \( \zeta_t \). The initial state \( \eta_0 \) 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 \( \theta \) 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 \( p(y \mid \theta, \eta) \) is Gaussian. In fact, stacking (1) over the \( T \) time periods, we have

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

where

\[
    X = \begin{bmatrix} X_1 & \cdots & X_T \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix}, \quad \Sigma^{-1} = \begin{bmatrix} \Sigma_1^{-1} & \cdots \\ \vdots & \cdots \end{bmatrix}.
\]

A change of variable from \( \varepsilon \) to \( y \) implies that

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

It is important to realize that \( \Sigma \) is a banded matrix. In fact, we have

\[
    \Sigma = \begin{bmatrix} \Sigma_1 & \cdots \\ \vdots & \cdots \end{bmatrix}.
\]

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

\[
    K = \begin{pmatrix} I_m & I_m & I_m \\ -\Gamma_2 & I_m & I_m \\ -\Gamma_3 & \cdots & \cdots \\ \cdot & \cdots & \cdots \\ -\Gamma_T & \cdots & I_m \end{pmatrix}
    \quad \text{and} \quad
    \Omega = \begin{pmatrix} \Omega_1 & \Omega_2 \\ \Omega_2 & \Omega_3 \\ \vdots & \vdots \end{pmatrix},
\]

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

\[
    \gamma = \begin{bmatrix} \Gamma_1 \eta_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{and} \quad \zeta = \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_T \end{bmatrix} \sim \mathcal{N}(0, \Omega^{-1}).
\]
Noting that $|K| = 1$, by a simple change of variable from $\zeta$ to $\eta$, we have

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

(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

$$
\begin{pmatrix}
\Gamma_1' \Omega_2 \Gamma_2 + \Omega_1 & -\Gamma_1' \Omega_2 & -\Gamma_1' \Omega_3 & \cdots \\
-\Gamma_2' \Omega_2 & \Gamma_2' \Omega_3 + \Omega_2 & -\Gamma_2' \Omega_3 & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
-\Omega_{T-1} \Gamma_T & \Gamma_T \Omega_T \Gamma_T + \Omega_{T-1} & -\Gamma_T' \Omega_T & \Omega_T \\
\end{pmatrix}
$$

(6)

Since the likelihood function $p(y | \theta, \eta)$ in (4) and the prior $p(\eta | \theta)$ in (5) are both linear Gaussian in $\eta$, 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. In fact, we have

$$\log p(\eta | y, \theta) \propto \log p(y | \theta, \eta) + \log p(\eta | \theta)$$

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

In other words,

$$(\eta | y, \theta) \sim N(\hat{\eta}, H^{-1}),$$

(7)

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

$$H = K' \Omega K + X \Sigma X,$$

(8)

$$\hat{\eta} = H^{-1}(K^0 + X' \Sigma y).$$

(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 $\hat{\eta}$ in (9). More specifically, the mean $\hat{\eta}$ 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

$$C_H' C_H \hat{\eta} = K^0 + X' \Sigma y,$$

(10)

for $C_H \hat{\eta}$ by forward-substitution and then using the result to solve for $\hat{\eta}$ by back-substitution. Similarly, to obtain a random draw from $N(\hat{\eta}, H^{-1})$ efficiently, sample $u \sim N(0, I_{Tm})$, and solve $C_H x = u$ for $x$ by back-substitution. It follows that $x \sim N(0, H^{-1})$. Adding the mean $\hat{\eta}$ to $x$, one obtains a draw from $N(\hat{\eta}, H^{-1})$. 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 $\hat{\eta}$. 
3. Sample \( u \sim \mathcal{N}(0, I_{Tm}) \), and solve \( C_H x = u \) for \( x \) by back-substitution. Take \( \eta = \bar{\eta} + x \), so that \( \eta \sim \mathcal{N}(\bar{\eta}, H^{-1}) \).

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 \( b \) and \( C_H \), getting an additional draw from \( \mathcal{N}(\bar{\eta}, H_1) \) requires only (a) \( Tm \) iid draws from standard Gaussian \( u \sim \mathcal{N}(0, I_{Tm}) \); (b) perform a fast back-substitution to solve \( C_H x = u \) for \( x \); and (c) adding \( \bar{\eta} \) to \( x \). We will exploit this 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 \( p(y_t | \eta_t, \theta) \), 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; see the discussion in Section 5.

3.1 Gaussian Approximation

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

\[
\begin{align*}
  f_t & = \frac{\partial}{\partial \eta_t} \log p(y_t | \eta_t, \theta) \bigg|_{\eta_t = \bar{\eta}_t}, \\
  G_t & = -\frac{\partial^2}{\partial \eta_t \partial \eta_t'} \log p(y_t | \eta_t, \theta) \bigg|_{\eta_t = \bar{\eta}_t}.
\end{align*}
\]

Stacking

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

we can expand the log-likelihood \( \log p(y | \eta, \theta) = \sum_{t=1}^{T} \log p(y_t | \eta_t, \theta) \) around \( \bar{\eta} = (\bar{\eta}_1', \ldots, \bar{\eta}_T')' \):

\[
\log p(y | \eta, \theta) \approx \log p(y | \bar{\eta}, \theta) + (\eta - \bar{\eta})' f - \frac{1}{2} (\eta - \bar{\eta})' G (\eta - \bar{\eta}) + c_1, \tag{11}
\]

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

\[
\log p(\eta | y, \theta) \propto \log p(y | \eta, \theta) + \log p(\eta | \theta) \approx -\frac{1}{2} \left[ \eta' (G + K' \Omega K) \eta - 2 \eta' (f + G \bar{\eta} + K' \theta) \right] + c_2. \tag{12}
\]
where \( c_2 \) is some unimportant constant independent of \( \eta \). In other words, the approximating distribution is Gaussian with precision \( H \equiv G + K'\Omega K \) and mean vector \( H^{-1}(f + G\eta + K^0) \).

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

\[
\frac{\partial}{\partial \eta} \log p(\eta \mid y, \theta) \Bigg|_{\eta=\tilde{\eta}} = -H\tilde{\eta} + 2(f + G\tilde{\eta} + K^0).
\]

Hence, we can implement the Newton-Raphson method as follows: initialize with \( \eta = \eta^{(1)} \). For \( s = 1, 2, \ldots \), use \( \tilde{\eta} = \eta^{(s)} \) 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

\[
\eta^{(s+1)} = \eta^{(s)} + H(\eta^{(s)})^{-1} \frac{\partial}{\partial \eta} \log p(\eta \mid y, \theta) \Bigg|_{\eta=\eta^{(s)}} = H(\eta^{(s)})^{-1} \left( f(\eta^{(s)}) + G(\eta^{(s)})\eta^{(s)} + K^0 \right). \tag{13}
\]

If \( \|\eta^{(s+1)} - \eta^{(s)}\| > \epsilon \) for some pre-fixed tolerance level \( \epsilon \), then continue; otherwise stop and set \( \tilde{\eta} = \eta^{(s+1)} \). Again, it is important to note that (13) can be efficiently evaluated without inverting any high-dimensional matrix. This is because the precision \( H \) is banded, and its Cholesky decomposition \( C_H \) can be readily obtained. 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 \( C_H^T x = x \) for \( x \) by forward-substitution. Then given \( x \), solve \( C_H z = x \) for \( z \) by back-substitution. Finally, given the mode \( \tilde{\eta} \), the negative Hessian \( H \) at \( \tilde{\eta} \) 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 \( \theta \) but integrated over the states \( \eta \). More explicitly,

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

The need to evaluate the integrated likelihood efficiently arises in both frequentist and Bayesian estimation. In classical inference, one needs to maximize \( p(y \mid \theta) \) with respect to \( \theta \) to obtain the maximum likelihood estimator. For Bayesian estimation, if one can evaluate \( p(y \mid \theta) \) quickly, more efficient samplers can be developed to obtain draws from the posterior, such as sampling \( \eta \) and \( \theta \) jointly in a single step, as in the setting of the previous section.

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

\[
\bar{p}(y \mid \theta) = \frac{1}{M} \sum_{i=1}^{M} \frac{p(y \mid \theta, \eta^i)p(\eta^i \mid \theta)}{q(\eta^i \mid y, \theta)}.
\]
By the strong law of large number, \( \hat{p}(y | \theta) \) is a 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 \( \eta \), then the variance of the estimator is also finite (Geweke, 1989).

Again low marginal cost for extra draws. So perfect for importance sampling.

### 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

The most basic sampling scheme is to implement a Metropolis-Hastings step with proposal density \( N(\tilde{\eta}, H^{-1}) \). The mode \( \tilde{\eta} \) and the negative Hessian at \( \tilde{\eta} \) of the conditional density \( p(\eta | y, \theta) \) can be computed quickly as discussed in the previous section. Moreover, a draw from the proposal can be obtained as in Algorithm 1. Hence, we summarize this basic sampling scheme as follows:

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

1. Obtain \( \tilde{\eta} \) iteratively via (13). Given \( H \), compute its Cholesky decomposition \( C_H \) such that \( H = C_H' C_H \).
2. Sample \( u \sim N(0, I_{Tm}) \), and solve \( C_H x = u \) for \( x \) by back-substitution. Take \( \eta = \tilde{\eta} + x \), so that \( \eta \sim N(\tilde{\eta}, H^{-1}) \).

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 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 \( \eta \sim t(\nu, \tilde{\eta}, H^{-1}) \) with degree of freedom parameter \( \nu \), location vector \( \tilde{\eta} \) and scale matrix \( H^{-1} \). We emphasize that sampling from \( t(\nu, \tilde{\eta}, H^{-1}) \) involves only \( Tm \) iid standard Gaussian draws and a draw from the \( \text{Gamma}(\nu/2, \nu/2) \) distribution. We summarize the algorithm as follows:

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

1. Given the posterior mode \( \tilde{\eta} \) and negative Hessian \( H \), obtain the Cholesky decomposition \( C_H \) such that \( H = C_H' C_H \).
2. Sample \( u \sim N(0, I_{T_m}) \) and \( r \sim \text{Gamma}(\nu/2, \nu/2) \). Then \( v = u / \sqrt{r} \sim t(\nu, 0, I_{T_m}) \).

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

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 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

\[
p(y | \eta, \theta)p(\eta | \theta) \leq cq(\eta | y, \theta),
\]

for all \( \eta \) in the support of \( p(\eta | y, \theta) \). When \( \eta \) is a high-dimensional vector, as in the present case, such a constant \( c \), if exists, is difficult to obtain. To make matters worse, the target density \( p(\eta | y, \theta) \) depends on other model parameters \( \theta \) 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 \( \eta \), we resort to the MH algorithm. To present the algorithm, it is convenient to first define the set

\[
\mathcal{D} = \{ \eta : p(y | \eta, \theta)p(\eta | \theta) \leq cq(\eta | y, \theta) \},
\]

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 \( \eta^* \) with probability

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

Continue the process until a draw \( \eta^* \) is accepted.

2. **MH-step:** Given the current draw \( \eta \) and the proposal \( \eta^* \)

   (a)
   (b) if \( \eta \in \mathcal{D} \), set \( \alpha_{MH}(\eta, \eta^* | y, \theta) = 1 \);
   (c) if \( \eta \in \mathcal{D}^c \) and \( \eta^* \in \mathcal{D} \), set
   \[
   \alpha_{MH}(\eta, \eta^* | y, \theta) = \frac{cq(\eta | y, \theta)}{p(y | \eta, \theta)p(\eta | \theta)};
   \]
   (d) if \( \eta \in \mathcal{D}^c \) and \( \eta^* \in \mathcal{D}^c \), set
   \[
   \alpha_{MH}(\eta, \eta^* | y, \theta) = \min \left\{ 1, \frac{p(y | \eta^*, \theta)p(\eta^* | \theta)q(\eta | y, \theta)}{p(y | \eta, \theta)p(\eta | \theta)q(\eta^* | y, \theta)} \right\}.
   \]
As shown in Chib and Greenberg (1995), the draws produced at the completion of the AR step have the density

\[ q_{AR}(\eta | y, \theta) = \frac{p(y | \eta, \theta)p(\eta | \theta)cd}{q(\eta | y, \theta)}, \quad \eta \in \mathcal{D}, \]

\[ q_{AR}(\eta | y, \theta) = \frac{q(\eta | y, \theta)}{d}, \quad \eta \in \mathcal{D}^c, \]

i.e., the new proposal density coincides with the target density on the set \( \mathcal{D} \) (albeit normalizing constants), whereas on \( \mathcal{D}^c \) the new proposal is reduced to the original one (see Figure 1 for an illustration). The better approximation, of course, comes at a cost, because multiple draws from the proposal density \( q(\eta | y, \theta) \) 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.

Following Chib and Jeliazkov (2005), we now discuss a practical way to select the constant \( c \) and the trade-off in such a choice. Notice that if a bigger \( c \) is chosen, then the set \( \mathcal{D} \) is larger and we are more likely to accept the candidate \( \eta^* \). The cost, on the other hand, of selecting a larger \( c \) is that more draws from \( q(\eta | y, \theta) \) are required in the AR step. A practical way to strike a balance between these two conflicting considerations is to set \( c = r\mathbb{P}(y | \hat{\eta}, \theta)p(\hat{\eta} | \theta)/q(\hat{\eta} | y, \theta) \), where \( \hat{\eta} \) is the mode of the conditional density \( p(\eta | y, \theta) \) and \( r \) is, say, between 1 and 3. Such a choice would ensure that \( c \) is sufficiently small, while big enough so that the set \( \mathcal{D} \) contains the mode \( \hat{\eta} \) 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 \( p(\eta | y, \theta) \): the MH and the ARMH algorithms with a Gaussian or \( t \) proposal. In performing a full Bayesian analysis, one needs to sequentially draw from \( p(\eta | y, \theta) \) followed by sampling from \( p(\theta | y, \eta) \). In typical situations where \( \theta \) contains parameters in the state equation, \( \eta \) and \( \theta \) are expected to be highly correlated. Consequently, the conventional sampling scheme of sequentially drawing from \( p(\eta | y, \theta) \) and \( p(\theta | y, \eta) \) might induce high autocorrelation, especially in high-dimensional settings. For this reason, we seek to sample \((\theta, \eta)\) jointly by first drawing from \( p(\theta | y) \) marginally of the states \( \eta \) followed by a draw from \( p(\eta | y, \theta) \). This is an instant of the collapsed MCMC sampling scheme first introduced in Liu (1994). In order to sample \( \theta \)
and \( \eta \) jointly, we need two ingredients: (1) a quick routine to evaluate the integrated likelihood \( p(y \mid \theta) \) and (2) a way to generate candidate \( \theta \).

(1) can use the importance sampling estimator \( \bar{p}(y \mid \theta) \) in Section 3.2. Complications: we only have an unbiased estimator for \( p(y \mid \theta) \). Would that affect the convergence of the Markov chain? The answer is no, as per Andrieu, Berthelsen, Doucet, and Roberts (2007) and Flury and Shephard (2008).

(2) Use the cross-entropy to locate an optimal proposal density for \( \theta \), denoted as \( q(\theta \mid y) \). Something like Keith, Kroese, and Sofronov (2008). It would require a pre-run. Use ARHM.

## 4 Application

To illustrate the proposed approach, we estimate the risk of a liquidity trap in the US under a time-varying parameter vector autoregressive (TVP-VAR) model with stochastic volatility. (** Talk about why it is important to estimate the prob of liquidity trap).

### 4.1 The Model

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

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

where \( \mu_t \) is an \( n \times 1 \) vector of time-varying intercepts, \( A_{1t}, \ldots, A_{lt} \) are \( n \times n \) matrices of VAR coefficients at time \( t \), and \( \Sigma_t^{-1} \) is a time-varying precision matrix. For the purpose of estimation, we write the VAR system in the form of seemingly unrelated regressions:

\[
y_t = x_t \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma_t^{-1}),
\]

(16)

where \( x_t = I_n \otimes [y_{t-1}, \ldots, y_{t-l}] \) and \( \beta_t = \text{vec}([\mu_t : A_{t1} : \cdots : A_{t_l}]') \) is a \( k \times 1 \) vector of VAR coefficients with \( k = n^2l + n \). To model the time-varying precision matrix \( \Sigma_t \), we follow the approach proposed in Primiceri (2005) by first factoring the precision matrix as \( \Sigma_t = L_t D_t L_t' \), where \( D_t = \text{diag}(e^{h_{t1}}, \ldots, e^{h_{tn}}) \) is a diagonal matrix, and \( L_t \) is a lower triangular matrix with ones on the main diagonal, i.e.,

\[
D_t = \begin{pmatrix}
e^{h_{t1}} & 0 & \cdots & 0 \\
0 & e^{h_{t2}} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & e^{h_{tn}}
\end{pmatrix}, \quad
L_t = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
a_{t21} & 1 & 0 & \cdots & 0 \\
a_{t31} & a_{t32} & 1 & \cdots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
a_{tn1} & a_{tn2} & \cdots & a_{tn,n-1} & 1
\end{pmatrix}.
\]

This decomposition has been employed in various applications, especially in the context of efficient estimation of covariance matrices (see, e.g., Pourahmadi, 1999, 2000; Smith and Kohn, 2002; Chan and Jeliazkov, 2009a, among others). In the setting of VAR models with time-varying volatility, it is first considered in Cogley and Sargent (2005). The log-volatilities \( h_t = \)
$(h_{t1}, \ldots, h_{tn})'$ evolve according to the state equation
\[ h_t = h_{t-1} + \xi_t, \quad \xi_t \sim N(0, \Omega_h^{-1}), \quad (17) \]
for $t = 2, \ldots, T$, where $\Omega_h = \text{diag}(\omega_{h1}, \ldots, \omega_{hn})$ is a diagonal matrix. The process is initialized with $h_1 \sim N(0, V_h^{-1})$ for some known diagonal precision matrix $V_h$. Let $a_t$ denote the free elements in $L_t$ ordered by rows, i.e., $a_t = (a_{t21}, a_{t31}, a_{t32}, \ldots, a_{tn,n-1})'$, 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
\[ a_t = a_{t-1} + \zeta_t, \quad \zeta_t \sim N(0, \Omega_a^{-1}), \quad (18) \]
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 $a_1 \sim N(0, V_a^{-1})$ for some known diagonal precision matrix $V_a$. In what follows we use these two parameterizations, namely, $\Sigma_t$ and $(h_t, a_t)$, interchangeably. To complete the specification of the model, it remains to specify the evolution of the VAR coefficients $\beta_t$. We follow the standard approach of modeling the VAR coefficients $\beta_t$ as a random walk process:
\[ \beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \sim N(0, \Omega_\beta^{-1}), \quad (19) \]
for $t = 2, \ldots, T$, where $\Omega_\beta = \text{diag}(\omega_{\beta1}, \ldots, \omega_{\beta k})$ is a diagonal precision matrix. The process is initialized with $\beta_1 \sim N(0, V_\beta^{-1})$ for some known precision matrix $V_\beta$.

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 $y_{t1}$, the first element of $y_t$, is the nominal interest rate, and let $x_{t1}$ be the first row of $x_t$. We assume that $y_{t1} \geq 0$. Consequently, given $\beta_t$ and $\Sigma_t$, $y_t$ follows a multivariate Gaussian distribution with the first element restricted to be positive. To derive the likelihood function, first note that since only $y_{t1}$ is constrained while other elements of $y_t$ are not, the marginal distribution of $y_{t1}$ is a univariate Gaussian variable truncated below at 0. In fact, it can be easily shown that
\[ (y_{t1} \mid \beta_t, \Sigma_t) \sim N(x_{t1} \beta_t, e^{h_{t1}}) \mathbb{I}(y_{t1} \geq 0). \]

It follows that given $\beta_t$ and $\Sigma_t$, we have
\[ \mathbb{P}(y_{t1} \geq 0 \mid \beta_t, \Sigma_t) = 1 - \Phi \left( -x_{t1} \beta_t / e^{\frac{1}{2} h_{t1}} \right) = \Phi \left( x_{t1} \beta_t e^{-\frac{1}{2} h_{t1}} \right), \]
where $\Phi(\cdot)$ denotes the standard Gaussian cumulative distribution function. Letting $y = (y_1^T, \ldots, y_T^T)^T$, $\beta = (\beta_1^T, \ldots, \beta_T^T)^T$ and $\Sigma = (\Sigma_1, \ldots, \Sigma_T)$, the log-likelihood function is thus
\[ \log p(y \mid \beta, \Sigma) = \sum_{t=1}^T \log p(y_t \mid \beta_t, \Sigma_t), \quad (20) \]
where
\[ p(y_t \mid \beta_t, \Sigma_t) \propto -\frac{1}{2} \log |\Sigma_t| - \frac{1}{2} (y_t - x_t \beta_t)' \Sigma_t (y_t - x_t \beta_t) - \log \Phi \left( x_{t1} \beta_t e^{-\frac{1}{2} h_{t1}} \right). \]
Figure 2: Boxplots of the inefficiency factors for the three sampling schemes: MH (S1), ARMH (S2) and 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.

4.2 Prior and Estimation

Given the measurement equation (20) and the state equations (17)–(19), 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: \( \Omega_\beta = \text{diag}(\omega_\beta_1, \ldots, \omega_\beta_k), \Omega_h = \text{diag}(\omega_{h1}, \ldots, \omega_{hk}), \) and \( \Omega_a = \text{diag}(\omega_{a1}, \ldots, \omega_{am}). \) Specifically, the diagonal elements of \( \Omega_\beta, \Omega_h \) and \( \Omega_a \) follow independently Gamma distributions:

\[
\begin{align*}
\omega_\beta_i &\sim \text{Gamma}(r_{\beta_i}/2, s_{\beta_i}/2) \quad \text{for} \quad i = 1, \ldots, k, \\
\omega_{hi} &\sim \text{Gamma}(r_{hi}/2, s_{hi}/2) \quad \text{for} \quad i = 1, \ldots, n; \\
\omega_{ai} &\sim \text{Gamma}(r_{ai}/2, s_{ai}/2), \quad \text{for} \quad i = 1, \ldots, m.
\end{align*}
\]

For later reference, we stack \( h = (h_0, \ldots, h_T)' \) and \( a = (a_0, \ldots, a_T)' \), and let \( \theta \) denote the set of parameters except the latent states \( \beta, h \) and \( a \), i.e., \( \theta = (\Omega_\beta, \Omega_h, \Omega_a) \). Now, posterior draws can be obtained by sequentially drawing from (a) \( p(\beta \mid y, h, a, \theta) \); (b) \( p(h \mid y, \beta, a, \theta) \); (c) \( p(a \mid y, \beta, h, \theta) \); and (d) \( p(\theta \mid y, \beta, h, a) \).

In what follows, we briefly discuss the implementation of the proposed Markov sampler; we refer the readers to the Appendix for more details. Firstly, to efficiently sample the states \( \beta \) in the non-linear state space model (20) and (19), we consider implementing an independence-chain Metropolis-Hastings step by approximating the conditional distribution \( p(h \mid y, \beta, a, \theta) \) via a Gaussian distribution as discussed in Section 3.1. The next step is to sample from the conditional distribution \( p(a \mid y, h, \beta, \theta) \). Recall that \( h_{ii} \) is the \( i \)-th diagonal element in \( D_t \), and \( h_t = (h_{i1}, \ldots, h_{in})' \). Further let \( \bar{h}_i = (h_{1i}, \ldots, h_{Ti})' \). That is, \( h_t \) is the \( n \times 1 \) vector obtained by stacking \( h_{ii} \) by the second subscript, whereas \( \bar{h}_i \) is the \( T \times 1 \) vector obtained by stacking \( h_{ii} \) by the first subscript. In the Appendix we show that \( \log p(h \mid y, \alpha, \beta, \theta) = \sum_{i=1}^n \log p(h_{ii} \mid y, a, \beta, \theta) \).

In other words, to obtain a draw from \( p(h \mid y, a, \beta, \theta) \), we can instead sample from \( p(\bar{h}_i \mid y, a, \beta, \theta) \) sequentially without adversely affecting the efficiency of the sampler. Now, each a draw from \( p(\bar{h}_i \mid y, a, \beta, \theta) \) can be obtained via an independence-chain Metropolis-Hastings step with a Gaussian proposal density; more details are given in the Appendix. Thirdly, it can be easily shown that \( p(a \mid y, \beta, h, \theta) \) is a Gaussian distribution (see, e.g. Primiceri, 2005), and a draw from which can be obtained using Algorithm 1. Finally, \( p(\theta \mid y, \beta, h, a) \) is a product of Gamma densities, and a draw from which is standard (see Koop, 2003, p. 61-62).

4.3 Empirical Results

Put something here later. Remember to define inefficiency factor and discuss what it means.

5 Extensions

- non-Gaussian and non-Markovian state equation.
- measurement equation of the form \( p(y_t \mid \theta, \eta_t, \eta_{t-1}, \ldots, \eta_{t-l}). \)
Appendix: Efficient Simulation of the States $\beta$ 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 $\Sigma_t = L_t^t D_t^{-1} L_t$, where

$$D_t = \begin{pmatrix} e^{h_{t1}} & 0 & \cdots & 0 \\ 0 & e^{h_{t2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{h_{tn}} \end{pmatrix}, \quad L_t = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{t21} & 1 & 0 & \cdots & 0 \\ a_{t31} & a_{t32} & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{tn1} & a_{tn2} & \cdots & a_{tn,n-1} & 1 \end{pmatrix}.$$  

Recall that $h_{ti}$ denotes the $i$-th diagonal element in $D_t$, $h_t = (h_{t1}, \ldots, h_{tn})'$, and $\widetilde{h}_i = (h_{1i}, \ldots, h_{Ti})'$. That is, $h_t$ is the $n \times 1$ vector obtained by stacking $h_{ti}$ by the second subscript, whereas $\widetilde{h}_i$ is the $T \times 1$ vector obtained by stacking $h_{ti}$ by the first subscript. Also, $a_t$ denotes the free elements in $L_t$ ordered by rows, i.e., $a_t = (a_{t21}, a_{t31}, a_{t32}, \ldots, a_{tn,n-1})'$. In what follows we use the two parameterizations $\Sigma_t$ and $(h_t, a_t)$ interchangeably. Then the log-density for $y_t$ given $(\beta_t, \Sigma_t)$ is

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

where $a_t = x_{t1} \beta_t e^{-\frac{1}{2} h_{t1}}$. Using the notation in Section 3.1, we have

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

where

$$\frac{\partial}{\partial \beta_t} \log p(y_t | \beta_t, \Sigma_t) = x_t' \Sigma_t (y_t - x_t \beta_t) - \frac{\phi(a_t)}{\Phi(a_t)} e^{-\frac{1}{2} h_{t1}} x_{t1}' ,$$

$$\frac{\partial^2}{\partial \beta_t \partial \beta'_t} \log p(y_t | \beta_t, \Sigma_t) = -x_t' \Sigma_t x_t + \frac{\phi(a_t)}{\Phi(a_t)} e^{-h_{t1}}  \left( a_t + \frac{\phi(a_t)}{\Phi(a_t)} \right) x_{t1}' x_{t1} ,$$

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 $p(h | y, a, \beta, \theta)$. We first show that

$$\log p(h | y, a, \beta, \theta) = \sum_{i=1}^{n} \log p(\widetilde{h}_i | y, a, \beta, \theta).$$

Put differently, to obtain a draw from $p(h | y, a, \beta, \theta)$, we can instead sample from $p(\widetilde{h}_i | y, a, \beta, \theta)$ sequentially without adversely affecting the efficiency of the sampler. To this end, decompose $\Sigma_t^{-1} = L_t^{-1} D_t^{-1} L_t$ as before. Since $\log |\Sigma_t| = \log |D_t| =$
\[ \sum_{i=1}^{n} h_{ti} \text{ and } \sigma_{t,11}^2 = e^{h_{ti}}, \text{ it follows that the log-likelihood is given by} \]

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

\[
= \sum_{t=1}^{T} \left[ -\frac{1}{2} \sum_{i=1}^{n} h_{ti} - \frac{1}{2} \sum_{i=1}^{n} e^{-h_{ti}} s_{ti}^2 - \log \Phi \left( e^{-h_{ti}/2} \bar{x}_{ti} \beta_t \right) \right], \tag{21}
\]

where \( \epsilon_t = y_t - x_t \beta_t \) and \( s_{ti}^2 \) is the \( i \)-th diagonal element of \((L_t \epsilon_t)'(L_t \epsilon_t)'\). On the other hand, the state equation (17) implies that each \( \tilde{h}_i \) follows independently a Gaussian distribution. In fact, we have

\[
h_{ti} = h_{t-1,i} + \xi_{ti}, \quad \xi_{ti} \sim N(0, \omega_{hi}). \tag{22}
\]

Hence, it follows from (21) and (22) that \( \tilde{h}_i, i = 1, \ldots, n \) 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 \( p(\tilde{h}_1 \mid y, a, \beta, \theta) \) for \( i = 2, \ldots, n \), it cannot be used to draw from \( p(\tilde{h}_1 \mid y, a, \beta, \theta) \) due to the extra term \( \log \Phi \left( e^{-h_{ti}/2} \bar{x}_{ti} \beta_t \right) \) in the log-likelihood (21) that depends on \( \beta_t \). Instead, we sample each \( \tilde{h}_i \) 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 (21) around the posterior mode \( \tilde{h}_i = (\tilde{h}_{i1}, \ldots, \tilde{h}_{iT})' \). Define \( \gamma_t = e^{-h_{ti}/2} \bar{x}_{ti} \beta_t \)

\[
q_t = \frac{\partial}{\partial \tilde{h}_t} \log p(y \mid \beta, h, a, \theta) \bigg|_{\tilde{h}_t = \tilde{h}_i}, \quad r_t = -\frac{\partial^2}{\partial \tilde{h}_t^2} \log p(y \mid \beta, h, a, \theta) \bigg|_{\tilde{h}_t = \tilde{h}_i},
\]

\[ q_i = (q_{i1}, \ldots, q_{Ti})' \text{ and } R_i = \text{diag}(r_{i1}, \ldots, r_{Ti}), \]

where

\[
\frac{\partial}{\partial \tilde{h}_t} \log p(y \mid \beta, h, a, \theta) = \frac{1}{2} \left[ e^{-h_{ti}} s_{ti}^2 - 1 + \gamma_t \frac{\phi(\gamma_t)}{\Phi(\gamma_t)} \mathbf{1}(i = 1) \right],
\]

and

\[
\frac{\partial^2}{\partial \tilde{h}_t^2} \log p(y \mid \beta, h, a, \theta) = -\frac{1}{2} e^{-h_{ti}} s_{ti}^2 + \frac{1}{4} \gamma_t \frac{\phi(\gamma_t)}{\Phi(\gamma_t)} \left[ \gamma_t^2 + \gamma_t \frac{\phi(\gamma_t)}{\Phi(\gamma_t)} - 1 \right] \mathbf{1}(i = 1).
\]

If we expand the log-likelihood (21) around the mode \( \tilde{h}_i \), we have

\[
\log p(y \mid \beta, h, a, \theta) \approx -\frac{1}{2} \left[ \tilde{h}_i' R_i \tilde{h}_i - 2 \tilde{h}_i' (q_i + R_i \tilde{h}_i) \right] + c_3,
\]

where \( c_3 \) is some unimportant constant independent of \( \tilde{h}_i \). We again consider the proposal density \( t(v, \tilde{h}_i, (q_i + R_i \tilde{h}_i)^{-1}) \), and everything follows as before.

References


