Joint sampling of states and parameters in state space models

William J. McCausland *[1] and Samuel Gingras †[1]

[1]Département de sciences économiques, Université de Montréal

November 14, 2019

Last version available here

Abstract

We consider the problem of sampling the posterior distribution in univariate non-linear, non-Gaussian state space models. We propose a new method that updates the parameter vector θ and the state vector x together as a single block. The proposal (θ^*, x^*) is drawn in two steps. The marginal proposal distribution for θ^* is constructed using approximations of the gradient and Hessian of the log posterior density of θ. The conditional proposal distribution for x^* given θ^* is that described in McCausland (2012). Computation of the approximate gradient and Hessian requires no simulation. Rather, it combines computational by-products of the x^* draw with a modest amount of additional computation. We compare the numerical efficiency of our posterior simulation with that of the Ancillarity-Sufficiency Interweaving Strategy (ASIS) described in Kastner & Frühwirth-Schnatter (2014), using the stochastic volatility model and the panel of 23 daily exchange rates from that paper. For computing the posterior mean of the volatility persistence parameter, our numerical efficiency is 4-19 times higher; for the volatility of volatility parameter, 15-36 times higher.

1 Introduction

We consider stationary univariate state space models with Gaussian states x_t,

$$x_1 \sim N(\mu, \sigma^2(1-\phi^2)^{-1}), \quad x_t \sim N(\mu(1-\phi)+\phi x_{t-1}, \sigma^2), \quad t = 2, \ldots, n,$$

and observations y_t with the following conditional independence structure:

$$p(y_1, \ldots, y_n | x_1, \ldots, x_n) = \prod_{t=1}^{n} p(y_t | x_t).$$

*william.j.mccausland@umontreal.ca

†samuel.gingras@umontreal.ca

The y_t may be scalars or vectors, and the dimension of y_t may vary with t, to accommodate mixed-frequency observations or missing data. Any element of y_t may be a discrete, continuous or mixed random variable. There may be functions of y_t that are non-random.

The most common examples are variations on the stochastic volatility (SV) model introduced by Taylor (1982). Some of these variations add flexibility to self-contained SV models, typically by allowing excess kurtosis in the measurement equation, jumps, or (negative) correlation between returns and the innovation of the state equation. In some variations, SV models are embedded in more complicated models.

Other examples of state space models include models with time-varying counts or durations. Durbin & Koopman (1997) study counts of deaths of van drivers in Britain; Frühwirth-Schnatter & Wagner (2006), casualties of pedestrians in Linz; and Jung et al. (2006), admissions for asthma to a hospital in Sydney. Bauwens & Veredas (2004), Strickland et al. (2006) and Strickland et al. (2008) study durations between transactions in financial markets.

Several methods for Bayesian posterior simulation in such state state models have been proposed. Direct methods sample latent states from their conditional posterior distribution. Sampling may be done one-at-a-time as in Jacquier et al. (1994); in blocks, as in Shephard & Pitt (1997), Watanabe & Omori (2004), Strickland et al. (2006) or Omori & Watanabe (2008); or all at once, as in McCausland (2012) or Djegnene & McCausland (2015). Auxiliary mixture methods involve transforming the model into a linear Gaussian model, approximating any non-Gaussian distributions in the transformed model by finite Gaussian mixtures. Kim et al. (1998), Chib et al. (2002) and Omori et al. (2007) use auxiliary mixture sampling for various SV models. Stroud et al. (2003) use it for Gaussian, but non-linear, state space models with state dependent variances; Frühwirth-Schnatter & Wagner (2006) for state space models with Poisson counts; and Frühwirth-Schnatter & Frühwirth (2007) for logit and multinomial logit models.

Numerical efficiency varies greatly across posterior simulation methods. Since there is often a great deal of posterior autocorrelation in x, and much posterior dependence between θ and x, it helps to update sequences of state values together in the same Gibbs block and to update parameters and states together. At the same time, the larger the block, the more difficult it is to approximate non-standard distributions. So, for example, a multivariate normal distribution is adequate for direct sampling of blocks of 20-50 state values, but not for the complete observed sequence x. McCausland (2012) and Djegnene & McCausland (2015) provide an approximation of $p(x|\theta, y)$ for generic state space models that is not multivariate normal and that proved highly efficient for drawing x in a single block. Auxiliary mixture models yield Gaussian x when one conditions on discrete mixture component indicators, and x can be drawn as a single block here too.

There have been previous attempts to draw θ and x together in a single block. For the Taylor SV model, Kim et al. (1998) draw θ and x together, conditional on mixture

component indicators, in what they call an "integration sampler" because x is marginalized out to draw parameters. Chib et al. (2002) analyse several SV models using a sampler in which x is marginalized out to draw μ, ϕ, σ and other parameters. McCausland (2012) draws parameters and x together, directly, as as single block in many state space models. One application replicates the analysis in Chib et al. (2002) of a 6-parameter Student's t SV model, the model from that paper with the highest Bayes factor for a sample of size $n = 8851$ of S&P 500 stock index returns. For the posterior sample mean of ϕ, Chib et al. (2002) achieve a numerical efficiency of 0.19 and McCausland (2012), 0.61. For the posterior sample mean of σ, the efficiencies are 0.10 and 0.87, respectively.

In all these cases, the SV model is a self-contained model for a single return series, not embedded in a larger model. This makes it easy to precompute the shape of the posterior distribution of parameters when x is marginalized out.

When the SV model is embedded in a larger model, the shape of this posterior distribution is a moving target. In part because of this issue, it is often desirable to update parameters and x—and mixture component indicators, if any—in separate Gibbs blocks. Kastner & Frühwirth-Schnatter (2014) do this as efficiently as possible by interleaving draws of two different parameterizations of θ to ensure that the update of θ is close to a pure Gibbs draw.

Our contribution is a method to draw parameters and states together, but in a way that does not rely on any pre-computation. Instead, it is based on computations of the local shape of the posterior distribution of parameters, with x marginalized out. The result is a sampler that is much more numerically efficient than that of Kastner & Frühwirth-Schnatter (2014) and not much less efficient than that of McCausland (2012), which does rely on pre-computing the shape of the posterior distribution of parameters.

We describe our new method in Section 2. We demonstrate it in Section 3, comparing its numerical efficiency with those of competing methods. We conclude in Section 4. The long and tedious work required to derive good approximations of the gradient and Hessian of the log posterior density of θ is relegated to the appendices. Once this is out of the way, the rest is quite simple.

2 Joint sampling of states and parameters

Our new simulation method takes advantage of close approximations of the gradient and Hessian of the log posterior density $\log p(\theta|y)$, where x has been marginalized out. The gradient and Hessian give a good idea of the shape of $p(\theta|y)$ at any point θ, allowing us to draw well targeted proposals of θ.

We describe our new methods in four parts. First, we describe a parameterization θ of the latent state process x that makes $p(\theta|y)$ well enough approximated by a Gaussian distribution. We then describe how we use the approximate gradient and Hessian of $\log p(\theta|y)$

to propose a candidate value θ^*. We then describe how we use the HESSIAN method to propose a candidate value x^*. Finally, we describe the Metropolis Hastings accept/reject decision that is applied to the joint proposal (θ^*, x^*).

2.1 Parameterizations

We will use two different parameterizations of the latent state process. One, ψ, is useful for computing analytic expressions. The other, θ, is useful for posterior simulation.

Let $\psi \equiv (\omega, \phi, \mu)$, where $\omega = \sigma^{-2}$, the precision of the state innovation. Since ω appears linearly in the quadratic (in x) term of $\log p(x|\theta)$, it is more convenient to work with than σ^2. However, the log posterior density $\ln p(\psi|x)$ is not well approximated by a quadratic in ψ. For one, its support is not $\mathbb{R}^3$; ψ lies in the space $\Psi \equiv [0, \infty] \times (-1, 1) \times \mathbb{R}$, and the ϕ parameter often has considerable posterior mass near the boundary of its support.

We prefer a parameterization whose log posterior density more closely resembles a positive definite quadratic function. This is for numerical efficiency, as it lets us better target the posterior density with a Gaussian approximation. With this in mind, let $\theta \equiv (\ln \omega, \tanh^{-1} \phi, \mu)$. The transformation $\theta(\psi)$ maps the parameter space Ψ to $\mathbb{R}^3$. We will use θ for posterior simulation, but we will need to transform θ back to ψ in order to evaluate $p(x|\omega, \phi, \mu)$. The reverse transformation is $(\omega, \phi, \mu) = \psi(\theta) = (e^{\theta_1}, \tanh \theta_2, \theta_3)$.

2.2 Drawing the proposal θ^*

The marginal proposal distribution for θ^* is based on a second order Taylor expansion of the log target density $\log p(\theta|y)$. The approach here is similar to that described in Robert & Casella (2010).

Let $g_{\theta|y}(\theta)$ and $H_{\theta|y}(\theta)$ be the gradient and Hessian of the log target density. We can write

$$g_{\theta|y}(\theta) = g_\theta(\theta) + g_{y|\theta}(\theta) \quad \text{and} \quad H_{\theta|y}(\theta) = H_\theta(\theta) + H_{y|\theta}(\theta),$$

where $g_\theta(\theta)$ and $H_\theta(\theta)$ are the gradient and Hessian of the log prior density $p(\theta)$; and $g_{y|\theta}(\theta)$ and $H_{y|\theta}(\theta)$ are the gradient and Hessian of $\log p(y|\theta)$ with respect to θ.

Unfortunately, $g_{y|\theta}(\theta)$ and $H_{y|\theta}(\theta)$ are not available. Instead, we use approximations $\tilde{g}_{y|\theta}(\theta)$ and $\tilde{H}_{y|\theta}(\theta)$, described in Appendix A. This gives the following approximations of the gradient and Hessian of the log target distribution:

$$\tilde{g}_{\theta|y}(\theta) = g_\theta(\theta) + \tilde{g}_{y|\theta}(\theta) \quad \text{and} \quad \tilde{H}_{\theta|y}(\theta) = H_\theta(\theta) + \tilde{H}_{y|\theta}(\theta).$$

Now let (θ_c, x_c) denote the current value of the parameter and state vectors in the MCMC chain. The values $\tilde{g}_{y|\theta}(\theta_c)$ and $\tilde{H}_{y|\theta}(\theta_c)$ will already be available from previous computations, as we will see.

Given the current value θ_c, we draw the proposal θ^* according to

$$\theta^*|\theta_c \sim N(\theta_c - \frac{1}{2}\gamma\tilde{H}_{\theta|y}(\theta_c)\tilde{g}_{\theta|y}(\theta_c), -\gamma\tilde{H}_{\theta|y}(\theta_c)^{-1}),$$

and evaluate the prior density $p(\theta^*)$ and the log proposal density $q(\theta^*|\theta_c)$. Since $\tilde{H}_{\theta|y}(\theta)$ depends on θ, we have to be careful not to ignore the normalization factor of $q(\theta^*|\theta_c)$.

The factor γ is present for robustness. Without it, numerical efficiency would be much lower: when θ_c is very far from its posterior mode, the gradient $\tilde{g}_{\theta|y}(\theta_c)$ is large, the mean of the proposal is far from θ_c, and the Hessian $\tilde{H}_{y|\theta}(\theta^*)$ may be quite different from $\tilde{H}_{y|\theta}(\theta_c)$ A consequence is that the acceptance probability may be extremely low. This situation is most likely to occur during a burn-in period where values of various variables may in regions of low posterior density.

Multiplying the Hessian matrix by γ reins in excessively large proposals. We use the following multiplier

$$\gamma = \coth\left(\frac{3+2\sqrt{6}}{-\tilde{g}_{\theta|y}(\theta)\tilde{H}^{-1}_{\theta|y}(\theta)\tilde{g}_{\theta|y}(\theta)}\right),$$

where coth is the hyperbolic cotangent. The function $\coth(1/x)$ is a "soft" $\min(x, 1)$, without a kink. The maximum value of $\gamma(-\tilde{g}_{\theta|y}(\theta)\tilde{H}^{-1}_{\theta|y}(\theta)\tilde{g}_{\theta|y}(\theta))$ is $3 + 2\sqrt{6}$, two standard deviations above the mean of a $\chi^2(3)$ random variable.

2.3 Drawing the proposal $x^*|\theta^*$

We use the HESSIAN method to draw x^* from a close approximation $q(x^*|\theta^*, y)$ to the conditional posterior distribution $p(x^*|\theta^*, y)$ and to compute the proposal density $q(x^*|\theta^*, y)$ We also compute $p(x^*|\theta^*)$ and $p(y|x^*)$,

Using a modest amount of additional computation, described in the appendices, we can also compute $\tilde{g}_{y|\theta}(\theta^*)$ and $\tilde{H}_{y|\theta}(\theta^*)$ at the same time. We need these to evaluate the reverse proposal density $q(\theta_c|\theta^*)$. In the event that the proposal (θ^*, x^*) is accepted, the values of $\tilde{g}_{y|\theta}(\theta^*)$ and $\tilde{H}_{y|\theta}(\theta^*)$ can be kept for the next iteration.

2.4 Joint accept-reject

We then compute the reverse proposal density $q(\theta_c|\theta^*)$ using $\tilde{g}_{y|\theta}(\theta^*)$ and $\tilde{H}_{y|\theta}(\theta^*)$, and accept (θ^*, x^*) with probability

$$\min\left[1, \frac{p(\theta^*)p(x^*|\theta^*)p(y|x^*)q(\theta_c|\theta^*)q(x_c|\theta_c, y)}{p(\theta_c)p(x_c|\theta_c)p(y|x_c)q(\theta^*|\theta_c)q(x^*|\theta^*, y)}\right].$$

3 Results

3.1 Stochastic volatility

We apply our methods to daily exchange rate data for 23 currencies, from January 3rd, 2000 to April 4, 2012. The data, from the European Central Bank, are those used by Kastner & Frühwirth-Schnatter (2014) in their empirical application. We compute 3139 log returns using exchange rates against the Euro observed on 3140 consecutive trading days. The stochastic volatility model, without leverage, is given by

$$\begin{aligned} y_t &\sim \mathcal{N}\Big(\exp(x_t)\Big) \\ x_t &= m_t + \phi\big(x_{t-1} - m_{t-1}\big) + \sigma u_t \\ x_1 &= m_1 + \sigma(1-\phi^2)^{-1/2} u_1, \end{aligned}$$

where $u_t \sim_{iid} \mathcal{N}(0,1)$. We use the following prior for $\theta = (\ln\omega, \tanh^{-1}\phi, \mu)$:

$$\theta \sim N\left(\begin{bmatrix} 3.6 \\ 2.5 \\ -10.5 \end{bmatrix}, \begin{bmatrix} 1.25 & 0.5 & 0 \\ 0.5 & 0.25 & 0 \\ 0 & 0 & 0.25 \end{bmatrix}\right),$$

where $\omega = \sigma^{-2}$. The prior is based on independent priors for $\tanh^{-1}\phi$ and $\log\omega(1-\phi^2)$. We do this because in practice, the unconditional precision $\omega(1-\phi^2)$ of x_t covaries less with ϕ than does the conditional precision ω, across financial return series. Results (including efficiency) are fairly robust to setting the covariance $\mathrm{Cov}[\theta_1, \theta_2]$ to zero.

Table 1 illustrates the results. For each currency, and the three parameters σ, ϕ and μ, we report the posterior sample mean and standard deviation, as well as the relative numerical efficiency for the posterior sample mean. The relative numerical efficiency is the ratio of the numerical variance of the mean of an iid sample to the numerical variance of the posterior sample mean. It is the reciprocal of the inefficiency factor used by Kastner & Frühwirth-Schnatter (2014) and others. We estimate the numerical variance of our posterior sample means using the overlapping batch means method—see Flegal & Jones (2010).

For σ and ϕ parameters, we also report the number of times more efficient the posterior sample means are, compared to those reported by Kastner & Frühwirth-Schnatter (2014). For computing the posterior mean of σ, our numerical efficiency is 15-36 times higher; for ϕ, 4-19 times higher.

For the μ parameter, numerical efficiency is high, although slightly lower than that reported by Kastner & Frühwirth-Schnatter (2014). Given that the conditional posterior distribution of μ, (i.e. $\mu|\phi, \sigma, x, y$) is Gaussian, the numerical efficiency for μ and higher moments could easily be greatly improved through antithetic sampling or Rao-Blackwellization.

Currency	$E[\sigma\|y]$	sd$[\sigma\|y]$	rne	×	$E[\phi\|y]$	sd$[\phi\|y]$	rne	×	$E[\mu\|y]$	sd$[\mu\|y]$	rne
Australian dollar	0.155	0.021	0.33	31.9	0.981	0.006	0.18	12.5	-10.26	0.17	0.18
Canadian dollar	0.077	0.015	0.25	29.5	0.993	0.004	0.09	8.4	-10.12	0.25	0.15
Swiss franc	0.202	0.019	0.23	16.8	0.986	0.004	0.18	5.9	-12.01	0.27	0.26
Czech koruna	0.260	0.032	0.21	20.6	0.960	0.010	0.18	12.8	-11.50	0.12	0.23
Danish krone	0.409	0.040	0.23	16.5	0.912	0.017	0.19	11.0	-18.07	0.09	0.21
UK pound sterling	0.098	0.012	0.37	32.3	0.993	0.002	0.20	7.9	-10.84	0.28	0.20
Hong Kong dollar	0.064	0.010	0.35	26.2	0.996	0.002	0.13	4.7	-10.16	0.32	0.18
Indonesian rupiah	0.208	0.032	0.24	34.7	0.974	0.009	0.18	20.9	-9.86	0.16	0.15
Japanese yen	0.114	0.015	0.30	27.1	0.991	0.003	0.16	7.7	-9.95	0.26	0.22
Korean won	0.135	0.016	0.26	20.8	0.989	0.004	0.19	7.7	-10.03	0.24	0.25
Mexican peso	0.153	0.020	0.34	29.3	0.982	0.006	0.18	11.0	-9.76	0.16	0.20
Malaysian ringgit	0.074	0.012	0.33	29.3	0.994	0.003	0.11	6.3	-10.28	0.27	0.17
Norwegian krone	0.165	0.021	0.25	19.2	0.976	0.007	0.16	8.4	-11.14	0.14	0.22
New Zealand dollar	0.155	0.027	0.30	40.0	0.974	0.010	0.15	17.5	-10.01	0.13	0.13
Philippine peso	0.133	0.021	0.27	43.4	0.983	0.007	0.12	15.1	-10.11	0.16	0.10
Polish zloty	0.181	0.020	0.37	25.3	0.979	0.006	0.22	9.5	-10.42	0.17	0.22
Romanian leu	0.299	0.025	0.30	18.0	0.972	0.006	0.29	9.3	-11.08	0.20	0.22
Russian rouble	0.143	0.016	0.24	20.0	0.990	0.003	0.17	6.4	-10.62	0.27	0.23
Swedish krona	0.108	0.012	0.27	16.4	0.992	0.002	0.16	3.6	-11.33	0.28	0.22
Singapore dollar	0.067	0.010	0.32	31.9	0.996	0.002	0.17	8.2	-10.58	0.35	0.22
Thai baht	0.115	0.018	0.29	26.4	0.987	0.005	0.12	7.8	-10.17	0.18	0.15
Turkish lira	0.302	0.025	0.25	17.2	0.962	0.008	0.21	8.9	-9.78	0.15	0.25
US dollar	0.064	0.010	0.39	29.2	0.996	0.002	0.16	6.0	-10.14	0.32	0.19

Table 1: Posterior mean, standard deviation and numerical efficiency for the SV model and ECB exchange rate data. Results are based on 45,000 posterior draws recorded after a burn-in period of 5,000 draws. The columns "×" give the ratio of the numerical efficiency obtained with our method and the ones reported in Kastner & Frühwirth-Schnatter (2014) for the same data.

3.2 High frequency counts with diurnal patterns.

This next example illustrate the use of our method for dynamic counts data that exhibit diurnal pattern. We use a data set with 2730 observations of transaction counts for IBM stock. There are 78 observations for each, corresponding to 5 min interval from 9:30 am to 4:00 pm. There are 35 trading days, from November 1, 1990 to December 21, 1990, where we removed the observations for November 23 because of halt in trading, (see Engle & Russell 1998). Details are in Chapter 5 of Tsay (2002), and the raw data are kindly provided by Ruey Tsay at the website for his book.[1]

Dynamic count models have been successfully used in many applications where count intensity varies over time. A simple parameter-driven model is given by

$$\begin{aligned} y_t &\sim \text{Poisson}\Big(\exp(x_t)\Big) \\ x_t &= m_t + \phi\big(x_{t-1} - m_{t-1}\big) + \sigma u_t \\ x_1 &= m_1 + \sigma(1-\phi^2)^{-1/2} u_1, \end{aligned}$$

where m_t is the value of a function describing a diurnal pattern and $u_t \sim_{iid} \mathcal{N}(0,1)$. To model intraday seasonality in transaction counts, we use a cubic B-spline function defined on a set of equally spaced knots set each half-hour between the opening and closing times of the market. This gives an expansion on $K = 16$ piecewise polynomials,

$$m_t = \sum_{k=1}^{K} \beta_k B_k(\tau_t)$$

where B_k denotes the k-th basis cubic polynomial and β_k is coefficient. Basis polynomials are evaluated at τ_t, the middle of the 5 min interval corresponding to observation y_t measured in second since the opening of the market. Lets $b_t = \big(B_1(\tau_t), \ldots, B_K(\tau_t)\big)$ be the $K \times 1$ vector of basis polynomials evaluated at τ_t.

We use a first order Gaussian random walk prior for the adjacent coefficients of the B-spline function,

$$\beta_k \,|\, \beta_{k-1}, \tau \sim \mathcal{N}(\beta_{k-1}, \tau^{-1}) \quad k = 2, \ldots, K,$$

where τ is a precision parameter having a chi-square distribution, $\bar{s}\tau \sim \chi^2(\bar{\nu})$. This prior enforces smoothness and mitigates overfitting. The additional parameter τ controls the degree of smoothness of the B-spline function and will be estimated with the coefficients. We choose a Gaussian distribution for the initial coefficient, $\beta_1 \sim \mathcal{N}(\bar{\beta}, \bar{h}^{-1})$, which gives a proper prior distribution for β. We choose the values $(\bar{\beta}, \bar{h}, \bar{s}, \bar{\nu}) = (2.5, 2.0, 1.0, 200)$ for the

[1] https://faculty.chicagobooth.edu/ruey.tsay/teaching/fts/

hyperparameters and use the following prior for $\theta = (\log \sigma^{-2}, \tanh^{-1} \phi)$,

$$\theta \sim N\left(\begin{bmatrix}4.0\\1.5\end{bmatrix}, \begin{bmatrix}2.0 & 0.5\\0.5 & 0.65\end{bmatrix}\right).$$

We sample the joint posterior distribution of parameters and state variables by iteratively sampling from $p(\theta, x \,|\, \beta, y)$ using our joint proposal, $p(\tau \,|\, \beta)$ with a direct Gibbs draw, and $p(\beta \,|\, \theta, y)$ using a Metropolis-Hastings algorithm. More precisely, we draw τ from $\bar{\bar{s}}\tau \,|\, \beta \sim \chi^2(\bar{\bar{\nu}})$ where $\bar{\bar{s}} = \bar{s} + \beta' R \beta$ and $\bar{\bar{\nu}} = \bar{\nu} + L - 1$. Here, $R = \Delta'\Delta$ where Δ is the matrix of first order backward operator, giving restriction the adjacent coefficients imply by the random walk prior.

$$z_t = w_t' \beta + \sigma u_t$$

where $z_t = x_t - \phi x_{t-1}$ and $w_t = b_t - \phi b_{t-1}$ for $t = 2, \ldots, n$ and $z_1 = x_1(1-\phi^2)^{1/2}$ and $w_1 = b_1(1-\phi^2)^{1/2}$. Hence, a Gaussian proposal based on Bayesian linear regression theory provides a good approximation to the conditional posterior distribution. Thus the proposal distribution is $\beta^* \sim \mathcal{N}(\bar{\bar{\beta}}, \bar{\bar{\Omega}}^{-1})$, where $\bar{\bar{\Omega}} = \tau R + \sigma^{-2} W'W$ and $\bar{\bar{\beta}} = \sigma^{-2}\bar{\bar{\Omega}}^{-1} W'z$, and we update the current value with probability

$$\min\left\{1, \exp\left(-\frac{\bar{h}}{2}\left[(\beta_1^* - \bar{\beta})^2 - (\beta_1 - \bar{\beta})^2\right]\right)\right\}.$$

Table 2 shows the results. Again, we obtain very high numerical efficiencies for all parameters; the efficiency for ϕ and σ are comparable with the ones obtained in the exchange rates example and the efficiency for the coefficients of the diurnal expansion are all in the interval $[0.590, 0.997]$.

4 Conclusions

With modest additional computation, we can compute approximations of the gradient and Hessian of $\log p(\theta|y)$ in univariate non-linear non-Gaussian state space models. This allows us to construct a one-block posterior sampler for (θ, x). In an empirical application, we show that the approximation is good enough to achieve high numerical efficiency for the 23 return series we investigate and for the transaction counts for IBM stock.

References

Bauwens, L. & Veredas, D. (2004), 'The stochastic conditional duration model: a latent variable model for the analysis of financial durations', *Journal of Econometrics* **119**, 381–412.

	$q_{0.01}$	$q_{0.05}$	$q_{0.5}$	$q_{0.95}$	$q_{0.99}$	Mean	Std	NSE	RNE
ϕ	0.563	0.579	0.618	0.655	0.670	0.6173	0.0229	0.00022	0.252
σ	0.338	0.344	0.360	0.377	0.384	0.3602	0.0098	0.00010	0.217
β_1	2.707	2.754	2.872	2.990	3.041	2.8714	0.0719	0.00043	0.625
β_2	2.612	2.652	2.751	2.850	2.893	2.7507	0.0605	0.00033	0.766
β_3	2.445	2.482	2.570	2.659	2.696	2.5702	0.0541	0.00027	0.903
β_4	2.303	2.339	2.422	2.505	2.539	2.4219	0.0505	0.00026	0.835
β_5	2.212	2.247	2.330	2.413	2.448	2.3302	0.0503	0.00024	0.997
β_6	2.160	2.195	2.277	2.360	2.394	2.2771	0.0505	0.00025	0.913
β_7	2.116	2.148	2.232	2.315	2.350	2.2317	0.0504	0.00025	0.902
β_8	1.984	2.020	2.103	2.186	2.221	2.1030	0.0507	0.00027	0.805
β_9	1.923	1.957	2.043	2.126	2.162	2.0424	0.0511	0.00027	0.821
β_{10}	1.956	1.990	2.074	2.158	2.193	2.0745	0.0510	0.00028	0.732
β_{11}	2.041	2.074	2.157	2.240	2.276	2.1570	0.0504	0.00026	0.819
β_{12}	2.131	2.165	2.248	2.331	2.365	2.2480	0.0502	0.00024	0.934
β_{13}	2.224	2.259	2.342	2.425	2.460	2.3420	0.0508	0.00028	0.746
β_{14}	2.285	2.322	2.412	2.503	2.540	2.4125	0.0548	0.00030	0.758
β_{15}	2.301	2.344	2.444	2.543	2.586	2.4440	0.0606	0.00033	0.750
β_{16}	2.283	2.333	2.453	2.572	2.621	2.4526	0.0728	0.00045	0.590

Table 2: Posterior quantiles and moments for a dynamic Poisson count model, with a diurnal pattern modelled as a cubic B-spline function, applied to transaction counts for IBM stock. Results are based on 45,000 posterior draws recorded after a burn-in period of 5,000 draws. The columns give five posterior quantiles, posterior mean and standard deviation, and the numerical standard error (NSE) and relative numerical efficiency (RNE) for the posterior mean. The first two rows are for the autocorrelation coefficient and innovation standard deviation of the latent state process. The other sixteen rows are for the coefficients of the B-spline function that gives the diurnal pattern.

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

Djegnene, B. & McCausland, W. J. (2015), 'The hessian method with conditional dependence', *Journal of Financial Econometrics* **13**(722–755).

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

Engle, R. & Russell, J. R. (1998), 'Autoregressive conditional duration: A new model for irregularly spaced transaction data', *Econometrica* **66**, 1127–1162.

Flegal, J. M. & Jones, G. L. (2010), 'Batch means and spectral variance estimators in Markov chain Monte Carlo', *Annals of Statistics* **38**(2), 1034–1070.

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

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

Jacquier, E., Polson, N. & Rossi, P. (1994), 'Bayesian analysis of stochastic volatility models', *Journal of Business and Economic Statistics* **12**(4), 371–388.

Jung, R. C., Kukuk, M. & Liesenfeld, R. (2006), 'Time series of count data: modeling, estimation and diagnostics', *Computational Statistics and Data Analysis* **51**, 2350–2364.

Kastner, G. & Frühwirth-Schnatter, S. (2014), 'Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models', *Computational Statistics and Data Analysis* **76**, 408–423.

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

McCausland, W. J. (2012), 'The HESSIAN method: Highly efficient state smoothing, in a nutshell', *Journal of Econometrics* **168**, 189–206.

Omori, Y., Chib, S., Shephard, N. & Nakajima, J. (2007), 'Stochastic volatility with leverage: fast and efficient likelihood inference', *Journal of Econometrics* **140**, 425–449.

Omori, Y. & Watanabe, T. (2008), 'Block sampler and posterior mode estimation for asymmetric stochastic volatility models', *Computational Statistics and Data Analysis* **52**, 2892–2910.

Robert, C. P. & Casella, G. (2010), *Monte Carlo Statistical Methods*, Springer.

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

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

Strickland, C. M., Forbes, C. S. & Martin, G. M. (2008), 'Parametrisation and efficient MCMC estimation of non-Gaussian state space models', *Computational Statistics and Data Analysis* **52**, 2911–2930.

Stroud, J. R., Müller, P. & Polson, N. G. (2003), 'Nonlinear state-space models with state-dependent variances', *Journal of the American Statistical Association* **98**, 377–386.

Taylor, S. J. (1982), Financial returns modelled by the product of two stochastic processes– a study of daily sugar prices 1691–1679., *in* O. D. Anderson, ed., 'Time Series Analysis: Theory and Practice 1.', North–Holland, pp. 203–226.

Tsay, R. (2002), *Analysis of Financial Time Series*, Wiley.

Watanabe, T. & Omori, Y. (2004), 'A multi-move sampler for estimating non-Gaussian time series models: Comments on Shephard and Pitt (1997)', *Biometrika* **91**, 246–248.

A Computing approximate gradient and Hessian of $\log p(y|\theta)$

Our objective is to compute approximations to the following gradient and Hessian:

$$g_{y|\theta}(\theta) \equiv \frac{\partial \log p(y|\theta)}{\partial \theta}, \quad H_{y|\theta}(\theta) \equiv \frac{\partial^2 \log p(y|\theta)}{\partial \theta \partial \theta^\top}.$$

We will use the exact result, proved in Appendix B, that

$$g_{y|\theta}(\theta) = E_{x|\theta,y}[g_{x|\theta}(\theta)] \quad \text{and} \quad H_{y|\theta}(\theta) = E_{x|\theta,y}[H_{x|\theta}(\theta)] + \text{Var}_{x|\theta,y}[g_{x|\theta}(\theta)],$$

where

$$g_{x|\theta}(\theta) \equiv \frac{\partial \log p(x|\theta)}{\partial \theta}, \quad H_{x|\theta}(\theta) \equiv \frac{\partial^2 \log p(x|\theta)}{\partial \theta \partial \theta^\top}.$$

We first compute exact expressions for $g_{x|\theta}(\theta)$ and $H_{x|\theta}(\theta)$. Then we compute exact expressions for their means. Then we use approximate distributions for $x|\theta, y$ to approximate these means. Finally, we compute an approximation of the variance of $g_{x|\theta}$.

A.1 Computing $g_{x|\theta}(\theta)$ and $H_{x|\theta}(\theta)$

First decompose $\log p(x|\psi) = k + L(\psi)$, where

$$k \equiv \frac{n}{2}(\ln \omega - \ln 2\pi) + \frac{1}{2}\ln(1-\phi^2)$$

and

$$L(\psi) = -\frac{\omega}{2}\left[(1-\phi^2)(x_n-\mu)^2 + \sum_{t=1}^{n-1}[(x_t-\mu)-\phi(x_{t+1}-\mu)]^2\right].$$

We will compute all first and second order partial derivatives of $L(\psi)$ with respect to ω, ϕ and μ.

First order partial derivatives:

$$\frac{\partial L(\psi)}{\partial \omega} = -\frac{1}{2}\left[(1-\phi^2)(x_n-\mu)^2 + \sum_{t=1}^{n-1}[(x_t-\mu)-\phi(x_{t+1}-\mu)]^2\right].$$

$$\frac{\partial L(\psi)}{\partial \phi} = \omega\left[\sum_{t=1}^{n-1}(x_t-\mu)(x_{t+1}-\mu) - \phi\sum_{t=2}^{n-1}(x_t-\mu)^2\right]$$

$$\frac{\partial L(\psi)}{\partial \mu} = \omega(1-\phi)\left[(x_1+x_n-2\mu) + (1-\phi)\sum_{t=2}^{n-1}(x_t-\mu)\right]$$

Second order partial derivatives:

$$\frac{\partial^2 L(\psi)}{\partial \omega^2} = 0, \qquad \frac{\partial^2 L(\psi)}{\partial \phi^2} = -\omega \sum_{t=2}^{n-1} (x_t - \mu)^2, \qquad \frac{\partial^2 L(\psi)}{\partial \mu^2} = -\omega \left[2(1-\phi) + (n-2)(1-\phi)^2\right],$$

$$\frac{\partial^2 L(\psi)}{\partial \omega \partial \phi} = \frac{1}{\omega}\frac{\partial L}{\partial \phi}, \qquad \frac{\partial^2 L(\psi)}{\partial \omega \partial \mu} = \frac{1}{\omega}\frac{\partial L}{\partial \mu},$$

$$\frac{\partial^2 L(\psi)}{\partial \phi \partial \mu} = -\omega \left[(x_1 + x_n - 2\mu) + 2(1-\phi)\sum_{t=2}^{n-1}(x_t - \mu)\right]$$

Now we compute gradient and Hessian of $L(\psi(\theta))$ with respect to θ, the vector of transformed parameter values. We can write

$$\frac{\partial L(\psi(\theta))}{\partial \theta} = \frac{\partial L(\psi)}{\partial \psi}\frac{\partial \psi}{\partial \theta}$$

$$\frac{\partial^2 L(\psi(\theta))}{\partial \theta_i^2} = \frac{\partial^2 L(\psi)}{\partial \psi_i^2}\left(\psi_i'(\theta_i)\right)^2 + \frac{\partial L(\psi)}{\partial \psi_i}\psi_i''(\theta_i)$$

For $i \neq j$,

$$\frac{\partial^2 L(\psi(\theta))}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 L(\psi(\theta))}{\partial \psi_i \partial \psi_j}\psi_i'(\theta_i)\psi_j'(\theta_j)$$

The first two derivatives of $\omega = \exp(\theta_1)$ with respect to θ_1 are

$$\psi_1'(\theta_1) = \frac{\partial \omega}{\partial \theta_1} = \exp(\theta_1) = \omega, \qquad \psi_1''(\theta_1) = \frac{\partial^2 \omega}{\partial \theta_1^2} = \exp(\theta_1) = \omega,$$

and the first two derivatives of $\phi = \tanh\theta_2$ with respect to θ_2 are

$$\psi_2'(\theta_2) = \frac{\partial \phi}{\partial \theta_2} = 1 - \tanh^2\theta_2 = 1 - \phi^2, \qquad \psi_2''(\theta_2) = \frac{\partial^2 \phi}{\partial \theta_2^2} = -2\tanh\theta_2(1 - \tanh^2\theta_2) = -2\phi(1-\phi^2).$$

The gradient with respect to θ is

$$\frac{\partial L(\psi(\theta))}{\partial \theta} = \begin{bmatrix} -\frac{\omega}{2}\left[(1-\phi^2)(x_n - \mu)^2 + \sum_{t=1}^{n-1}[(x_t - \mu) - \phi(x_{t+1} - \mu)]^2\right] \\ \omega(1-\phi^2)\left[\sum_{t=1}^{n-1}(x_t - \mu)(x_{t+1} - \mu) - \phi\sum_{t=2}^{n-1}(x_t - \mu)^2\right] \\ \omega(1-\phi)\left[(x_1 + x_n - 2\mu) + (1-\phi)\sum_{t=2}^{n-1}(x_t - \mu)\right] \end{bmatrix}$$

The elements of the Hessian are

$$\frac{\partial^2 L(\psi(\theta))}{\partial \theta_1^2} = \frac{\partial L(\psi(\theta))}{\partial \theta_1},$$

$$\frac{\partial^2 L(\psi(\theta))}{\partial \theta_2^2} = (1-\phi^2)^2 \frac{\partial^2 L(\psi)}{\partial \phi^2} - 2\phi(1-\phi^2)\frac{\partial L(\psi)}{\partial \phi} = (1-\phi^2)\left[(1-\phi^2)\frac{\partial^2 L(\psi)}{\partial \phi^2} - 2\phi\frac{\partial L}{\partial \phi}\right]$$

$$\frac{\partial^2 L(\psi(\theta))}{\partial \theta_3^2} = \frac{\partial^2 L(\psi)}{\partial \mu^2}.$$

$$\frac{\partial^2 L(\psi(\theta))}{\partial \theta_1 \partial \theta_2} = (1-\phi^2)\frac{\partial L(\psi)}{\partial \phi} = \frac{\partial L(\psi(\theta))}{\partial \theta_2}.$$

$$\frac{\partial^2 L(\psi(\theta))}{\partial \theta_1 \partial \theta_3} = \frac{\partial L(\psi)}{\partial \mu} = \frac{\partial L(\psi(\theta))}{\partial \theta_3}.$$

$$\frac{\partial^2 L(\psi(\theta))}{\partial \theta_2 \partial \theta_3} = (1-\phi^2)\frac{\partial^2 L(\psi)}{\partial \phi \partial \mu}.$$

The log normalizing factor $k = \frac{n}{2}(\ln\omega - \ln 2\pi) + \frac{1}{2}\ln(1-\phi^2)$ also depends on θ. Noting that $\theta_1 = \ln\omega$, we have

$$\frac{\partial k}{\partial \theta_1} = \frac{n}{2}, \qquad \frac{\partial^2 k}{\partial \theta_1^2} = 0.$$

Then

$$\frac{\partial k}{\partial \theta_2} = (1-\phi^2)\frac{\partial k}{\partial \phi} = (1-\phi^2)\frac{1}{2}\frac{-2\phi}{(1-\phi^2)} = -\phi,$$

$$\frac{\partial^2 k}{\partial \theta_2^2} = (1-\phi^2)\frac{\partial}{\partial \phi}\frac{\partial k}{\partial \theta_2} = -(1-\phi^2).$$

A.2 Computing the mean of $g_{x|\theta}(\theta)$ and $H_{x|\theta}(\theta)$

We express the mean of L and its derivatives in terms of the decomposition $x_t \equiv c_t + e_t$, where $c_t \equiv E[x_t|\theta, y]$ and $e_t \equiv x_t - c_t$.

Here we just take the mean of $g_{x|\theta}(\theta)$ and $H_{x|\theta}(\theta)$ with respect to the distribution $x|\theta, y$, component by component. For the rest of this section, we will suppress conditioning on θ and y so that, for example, $E[x]$ denotes $E[x|\theta, y]$.

This gives

$$E\left[\frac{\partial L(\psi(\theta))}{\partial \theta_1}\right] = -\frac{\omega}{2}\left[(1-\phi^2)[(c_n-\mu)^2 + E[e_n^2]] + \sum_{t=1}^{n-1}(c_t - \phi c_{t+1} - \mu(1-\phi))^2 + E[(e_t - \phi e_{t+1})^2]\right]$$

$$E\left[\frac{\partial^2 L(\psi(\theta))}{\partial \theta_1^2}\right] = E\left[\frac{\partial L(\psi(\theta))}{\partial \theta_1}\right]$$

$$E\left[\frac{\partial L(\psi(\theta))}{\partial \theta_2}\right] = \omega(1-\phi^2)\left[\sum_{t=1}^{n-1}(c_t-\mu)(c_{t+1}-\mu) + E[e_t e_{t+1}] - \phi\sum_{t=2}^{n-2}(c_t-\mu)^2 + E[e_t^2]\right]$$

$$E\left[\frac{\partial^2 L(\psi(\theta))}{\partial \theta_2^2}\right] = -\omega(1-\phi^2)^2\left[\sum_{t=2}^{n-1}(c_t-\mu)^2 + E[e_t^2]\right] - 2\phi E\left[\frac{\partial L(\psi(\theta))}{\partial \theta_2}\right]$$

$$E\left[\frac{\partial^2 L(\psi(\theta))}{\partial\theta_1\partial\theta_2}\right] = E\left[\frac{\partial L(\psi(\theta))}{\partial\theta_2}\right].$$

$$E\left[\frac{\partial L(\psi(\theta))}{\partial\theta_3}\right] = \omega(1-\phi)\left[(c_1+c_n-2\mu)+(1-\phi)\sum_{t=2}^{n-1}(c_t-\mu)\right]$$

$$E\left[\frac{\partial^2 L(\psi(\theta))}{\partial\theta_3^2}\right] = -\omega[2(1-\phi)+(n-2)(1-\phi)^2]$$

$$E\left[\frac{\partial^2 L(\psi(\theta))}{\partial\theta_1\partial\theta_3}\right] = E\left[\frac{\partial L(\psi(\theta))}{\partial\theta_3}\right]$$

$$E\left[\frac{\partial^2 L(\psi(\theta))}{\partial\theta_2\partial\theta_3}\right] = -\omega(1-\phi^2)\left[(c_1+c_n-2\mu)+2(1-\phi)\sum_{t=1}^{n-1}(c_t-\mu)\right]$$

To compute exact values of these expressions, we would need to compute all the $c_t = E[x_t]$ and the $E[e_t^2]$ and the $E[(e_t-\phi e_{t+1})^2]$. In the next section, we develop approximations.

A.3 Computing approximate moments of $g_{x|\theta}(\theta)$ and $H_{x|\theta}(\theta)$

First, let $x^\circ = (x_1^\circ, \ldots, x_n^\circ)$ be the mode of $x|\theta, y$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \equiv x - x^\circ$. Recall that we previously defined $c \equiv E[x]$ and $e \equiv x - c$. This implies that $e = \epsilon - E[\epsilon]$.

A.3.1 Introduce $\mu_{t|t+1}(\epsilon_{t+1})$ and $b_{t|t+1}(\epsilon_{t+1})$

Recall that $\epsilon_t = x_t - x_t^\circ$. All distributions in this appendix are conditional on y and θ, and the notation for this conditioning will be suppressed. Let $\mu_{t|t+1}(\epsilon_{t+1}) \equiv E[\epsilon_t|\epsilon_{t+1}]$. Let $b_{t|t+1}(\epsilon_{t+1})$ be the conditional mode of ϵ_t given ϵ_{t+1}. Let $v_t(\epsilon_{t+1}) = \epsilon_t - b_{t|t+1}(\epsilon_{t+1})$.

We have the following approximate Taylor expansions of $\mu_{t|t+1}(\epsilon_{t+1})$ and $b_{t|t+1}(\epsilon_{t+1})$. Definitions of coefficients are in McCausland (2012), and are already computed as by-products.

$$\mu_{t|t+1}(\epsilon_{t+1}) \approx \mu_t + \dot{\mu}_t\epsilon_{t+1} + \frac{1}{2}\ddot{\mu}_t\epsilon_{t+1}^2 + \frac{1}{6}\dddot{\mu}_t\epsilon_{t+1}^3$$

$$b_{t|t+1}(\epsilon_{t+1}) \approx b_t + \dot{b}_t\epsilon_{t+1} + \frac{1}{2}\ddot{b}_t\epsilon_{t+1}^2 + \frac{1}{6}\dddot{b}_t\epsilon_{t+1}^3$$

Centering to put these functions as polynomials in e_{t+1} gives, for example,

$$\begin{aligned}\mu_{t|t+1}(\epsilon_{t+1}) = &\left(\mu_t + \dot{\mu}_t E[\epsilon_{t+1}] + \frac{1}{2}\ddot{\mu}_t E[\epsilon_{t+1}]^2 + \frac{1}{6}\dddot{\mu}_t E[\epsilon_{t+1}]^3\right) \\ &+ \left(\dot{\mu}_t + \ddot{\mu}_t E[\epsilon_{t+1}] + \frac{1}{2}E[\epsilon_{t+1}]^2\right)e_{t+1} \\ &+ \frac{1}{2}\left(\ddot{\mu}_t + \dddot{\mu}_t E[\epsilon_{t+1}]\right)e_{t+1}^2 + \frac{1}{6}\dddot{\mu}_t e_{t+1}^3.\end{aligned} \tag{1}$$

A.3.2 Computing $E[\epsilon_t]$, $E[e_t e_{t+1}]$ and $\text{Var}[e_t]$

To compute $E[\epsilon_t]$, $E[e_t e_{t+1}]$ and $\text{Var}[e_t]$, we first use the laws of total expectation and total variance to obtain:

$$E[\epsilon_t] = E[E[\epsilon_t|x_{t+1}]], \tag{2}$$

$$E[e_t e_{t+1}] = E[e_{t+1} E[e_t|x_{t+1}]], \tag{3}$$

$$\text{Var}[e_t] = E[\text{Var}[e_t|x_{t+1}]] + \text{Var}[E[e_t|x_{t+1}]]. \tag{4}$$

By definition, $E[\epsilon_t|x_{t+1}] = \mu_{t|t+1}$, so we have, taking expectations of both sides of (1),

$$\begin{aligned} E[\epsilon_t] = &\left(\mu_t + \dot{\mu}_t E[\epsilon_{t+1}] + \frac{1}{2}\ddot{\mu}_t E[\epsilon_{t+1}]^2 + \frac{1}{6}\dddot{\mu}_t E[\epsilon_{t+1}]^3\right) \\ &+ \frac{1}{2}\left(\ddot{\mu}_t + \dddot{\mu}_t E[\epsilon_{t+1}]\right)\text{Var}\, e_{t+1} + \frac{1}{6}\dddot{\mu}_t E[e_{t+1}^3]. \end{aligned} \tag{5}$$

Now express the conditional expectation $E[e_t|x_{t+1}]$ as a third order polynomial in e_{t+1}:

$$\begin{aligned} E[e_t|x_{t+1}] &= E[\epsilon_t|x_{t+1}] - E[\epsilon_t] \\ &= (\dot{\mu}_t + \ddot{\mu}_t E[\epsilon_{t+1}])e_{t+1} + \frac{1}{2}\ddot{\mu}_t(e_{t+1}^2 - E[e_{t+1}^2]) + \frac{1}{6}\dddot{\mu}_t(e_{t+1}^3 - E[e_{t+1}^3]) \end{aligned} \tag{6}$$

Then using (3),

$$E[e_t e_{t+1}] = (\dot{\mu}_t + \ddot{\mu}_t E[\epsilon_{t+1}])\,\text{Var}[e_{t+1}] + \frac{1}{2}\ddot{\mu}_t E[e_{t+1}^3] + \frac{1}{6}\dddot{\mu}_t E[e_{t+1}^4]. \tag{7}$$

Similarly,

$$\begin{aligned} \text{Var}[E[e_t|x_{t+1}]] = &\left(\dot{\mu}_t + \ddot{\mu}_t E[\epsilon_{t+1}] + \frac{1}{2}\dddot{\mu}_t E[\epsilon_{t+1}]^2\right)^2 \text{Var}[e_{t+1}] \\ &+ \left(\dot{\mu}_t + \ddot{\mu}_t E[\epsilon_{t+1}] + \frac{1}{2}\dddot{\mu}_t E[\epsilon_{t+1}]^2\right)(\ddot{\mu}_t + \dddot{\mu}_t E[\epsilon_{t+1}])E[e_{t+1}^3] \\ &+ \frac{1}{4}(\ddot{\mu}_t + \dddot{\mu}_t E[\epsilon_{t+1}])^2\,\text{Var}[e_{t+1}^2]. \end{aligned} \tag{8}$$

A.3.3 Derivatives of $h_t \equiv \log p(x_t|x_{t+1})$ with respect to e_t

Result from McCausland (2012):

$$h_1''(x_1) = -\omega(1-\phi^2) + \psi_1''(x_1).$$

$$h_t''(x_t) = \omega\phi\mu_{t-1|t}'(x_t) - \omega + \psi_t''(x_t).$$

$$h_n''(x_n) = \omega\phi\mu'_{n-1|n}(x_t) - \omega(1-\phi^2) + \psi_t''(x_n),$$

where

$$\psi_t''(x_t) = \frac{\partial \log p(y_t|x_t,\theta)}{\partial x_t}$$

is a computational byproduct of the HESSIAN method. We can write

$$h_t''(x_t^\circ + b_{t|t+1}) = -\Sigma_t^{-1} + \omega\phi(\mu_{t-1|t}(b_{t|t+1}) - \dot{a}_{t-1}) + \psi_t''(x_t^\circ + b_{t|t+1}),$$

where $\Sigma_t \equiv -(\omega\phi\dot{a}_{t-1} - \omega + \psi_t''(x_t^\circ))^{-1}$ is a by-product of Hessian method computations.

Using the approximations

$$\mu_{t-1|t}(b_{t|t+1}) = \dot{\mu}_{t-1} + \ddot{\mu}_{t-1}b_{t|t+1}, \qquad \psi_t''(x_t^\circ + b_{t|t+1}) = \psi_t''(x_t^\circ) + \psi_t'''(x_t^\circ)b_{t|t+1},$$

we have

$$h_t''(x_t^\circ + b_{t|t+1}) \approx -\Sigma_t^{-1} + \omega\phi(\dot{\mu}_{t-1} - \dot{a}_{t-1}) + (\omega\phi\ddot{\mu}_{t-1} + \psi_t'''(x_t^\circ))b_{t|t+1}.$$

Then approximate

$$\begin{aligned} -h_t''(x_t^\circ + b_{t|t+1})^{-1} &\approx \frac{\Sigma_t}{1 - \dot{a}_t(\dot{\mu}_{t-1} - \dot{a}_{t-1}) - (\dot{a}_t\ddot{\mu}_{t-1} + \psi_t'''(x_t^\circ)\Sigma_t)b_{t|t+1}} \\ &\approx \Sigma_t[1 + \dot{a}_t(\dot{\mu}_{t-1} - \dot{a}_{t-1}) + (\dot{a}_t\ddot{\mu}_{t-1} + \psi_t'''(x_t^\circ)\Sigma_t)b_{t|t+1}]. \end{aligned} \tag{9}$$

We get the following special cases using a similar development, for $t = 1$ and $t = n$:

$$-h_1''(x_1^\circ + b_{1|2})^{-1} \approx \Sigma_1[1 + \psi_1'''(x_1^\circ)\Sigma_1 b_{1|2}],$$

$$-h_n''(x_n^\circ + b_n)^{-1} \approx \Sigma_n[1 + \dot{a}_n(\dot{\mu}_{n-1} - \dot{a}_{n-1}) + (\dot{a}_n\ddot{\mu}_{n-1} + \psi_n'''(x_n^\circ)\Sigma_n)b_n].$$

A.3.4 Computing $E[\epsilon_t]$, $E[\mathrm{Var}[e_t|x_{t+1}]]$ and $E[e_t^3]$

First, we compute the conditional expectations $E[\epsilon_t|x_{t+1}]$, $\mathrm{Var}[e_t|x_{t+1}]$ and $E[e_t^3|x_{t+1}]$, then we will take expectations again to get the unconditional expectations.

Now let $v_t \equiv \epsilon_t - b_{t|t+1}$ and $\delta_{t|t+1} = \mu_{t|t+1} - b_{t|t+1}$.

We can write $\epsilon_t = v_t - (\mu_{t|t+1} - b_{t|t+1}) = v_t - \delta_{t|t+1}$, and $e_t = v_t - (E[\epsilon_t] - b_{t|t+1})$.

We will base calculations on

$$E[v_t^2|x_{t+1}] = -h_t''(x_t^\circ + b_{t|t+1})^{-1}, \qquad E[v_t^3|x_{t+1}] = 0.$$

This implies (note that $E[v_t|x_{t+1}] = \delta_{t|t+1}$)

$$\mathrm{Var}[e_t|x_{t+1}] = E[(\epsilon_t - \mu_{t|t+1})^2|x_{t+1}] = E[(v_t - \delta_{t|t+1})^2|x_{t+1}] = -h_t''(x_t^\circ + b_{t|t+1})^{-1} - \delta_{t|t+1}^2. \tag{10}$$

$$\begin{aligned} E[e_t^3|x_{t+1}] &= E[(v_t - (E[\epsilon_t] - b_{t|t+1})^3)|x_{t+1}] \\ &= E[-3v_t^2(E[\epsilon_t] - b_{t|t+1}) + 3v_t(E[\epsilon_t] - b_{t|t+1})^2 - (E[\epsilon_t] - b_{t|t+1})^3|x_{t+1}] \\ &= 3h_t''(x_t^\circ + b_{t|t+1})^{-1}(E[\epsilon_t] - b_{t|t+1}) + 3\delta_{t|t+1}(E[\epsilon_t] - b_{t|t+1})^2 - (E[\epsilon_t] - b_{t|t+1})^3. \end{aligned} \tag{11}$$

A.3.5 Computing the gradient

Start-up, $t = n$:

1. Set $E[\epsilon_n] = \mu_n$.
2. Compute $E[e_n^2] = -h_n''(b_n)^{-1}$.
3. Compute $E[e_n^3] = (\mu_n - b_n)^3$.

Iterative step $t = n-1, \ldots, 1$, where we have $E[\epsilon_t]$, $E[e_t^2]$ and $E[e_t^3]$ from the previous step:

1. Compute $E[\epsilon_t]$ using (5), $E[e_t e_{t+1}]$ using (7), and $\mathrm{Var}[E[e_t|x_{t+1}]]$ using (8).
2. Compute coefficients giving $E[e_t^3|x_{t+1}]$ in (11) as a polynomial in e_{t+1}, take expectations to get $E[e_t^3]$.
3. Compute coefficient giving $\mathrm{Var}[e_t|x_{t+1}]$ in (10) as a polynomial in e_{t+1}, take expectations to get $E[\mathrm{Var}[e_t|x_{t+1}]]$. Combine this quantity with $\mathrm{Var}[E[e_t|x_{t+1}]]$ using (**??**) to obtain $E[e_t^2]$.

B Derivation of gradient and Hessian of $\log p(y|\theta)$

We can write, for any function $q(x)$ with the same support as $p(x|\theta)p(y|x)$,

$$p(y|\theta) = \int \frac{p(x|\theta)p(y|x)}{q(x)} q(x)\, dx.$$

For convenience, let

$$w_\theta(x) = \frac{p(x|\theta)p(y|x)}{q(x)},$$

so that $p(y|\theta) = \int w_\theta(x)q(x)\, dx$. The notation is meant to evoke an importance weight, where $p(x|\theta)p(y|x)$ is the target and $q(x)$ is the importance distribution. Note that for $q(x) = p(x|\theta, y)$, $w_\theta(x) = p(\theta|y)$ for all values of x.

We can write the gradient of $\log p(y|\psi)$ as

$$\frac{\partial \log p(y|\theta)}{\partial \theta} = \frac{1}{p(y|\theta)} \frac{\partial p(y|\theta)}{\partial \theta} = \frac{1}{p(y|\theta)} \int \frac{\partial p(x|\theta)}{\partial \theta} \frac{p(y|x)}{q(x)} q(x)\, dx = \frac{\int \frac{\partial \log p(x|\theta)}{\partial \theta} w_\theta(x) q(x)\, dx}{\int w_\theta(x) q(x)\, dx}.$$

Taking the derivative of the second expression, with respect to ψ, yields the Hessian matrix,

$$\frac{\partial^2 \log p(y|\theta)}{\partial \theta \partial \theta^\top} = \frac{1}{p(y|\theta)} \frac{\partial^2 p(y|\theta)}{\partial \theta \partial \theta^\top} - \frac{1}{p(y|\theta)^2} \frac{\partial p(y|\theta)}{\partial \theta} \frac{\partial p(y|\theta)}{\partial \theta^\top}.$$

We can write the first term as

$$\begin{aligned}
\frac{1}{p(y|\theta)} \frac{\partial^2 p(y|\theta)}{\partial \theta \partial \theta^\top} &= \frac{1}{p(y|\theta)} \int \frac{\partial^2 p(x|\theta)}{\partial \theta \partial \theta^\top} \frac{p(y|x)}{q(x)} q(x)\, dx \\
&= \frac{1}{p(y|\theta)} \int \left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta \partial \theta^\top} p(x|\theta) + \frac{1}{p(x|\theta)} \frac{\partial p(x|\theta)}{\partial \theta} \frac{\partial p(x|\theta)}{\partial \theta^\top} \right] \frac{p(y|x)}{q(x)} q(x)\, dx \\
&= \frac{1}{p(y|\theta)} \int \left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta \partial \theta^\top} + \frac{\partial \log p(x|\theta)}{\partial \theta} \frac{\partial \log p(x|\theta)}{\partial \theta^\top} \right] w_\theta(x) q(x)\, dx
\end{aligned}$$

We can write the second term as

$$-\frac{1}{p(y|\theta)^2} \frac{\partial p(y|\theta)}{\partial \theta} \frac{\partial p(y|\theta)}{\partial \theta^\top} = -\frac{\partial \log p(y|\theta)}{\partial \theta} \frac{\partial \log p(y|\theta)}{\partial \theta^\top}.$$

Putting it together, we have

$$\frac{\partial^2 \log p(y|\theta)}{\partial \theta \partial \theta^\top} = \frac{\int \left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta \partial \theta^\top} + \frac{\partial \log p(x|\theta)}{\partial \theta} \frac{\partial \log p(x|\theta)}{\partial \theta^\top} \right] w_\theta(x) q(x)\, dx}{\int w_\theta(x) q(x)\, dx} - \frac{\partial \log p(y|\theta)}{\partial \theta} \frac{\partial \log p(y|\theta)}{\partial \theta^\top}.$$

Note that if we could use $q(x) = p(x|\theta, y)$, the weights $w_\theta(x)$ would not depend on x, and we would have

$$\frac{\partial \log p(y|\theta)}{\partial \theta} = E_{x|\theta,y} \left[\frac{\partial \log p(x|\theta)}{\partial \theta} \right],$$

$$\frac{\partial^2 \log p(y|\theta)}{\partial \theta \partial \theta^\top} = E_{x|\theta,y} \left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta \partial \theta^\top} \right] + \mathrm{Var}_{x|\theta,y} \left[\frac{\partial \log p(x|\theta)}{\partial \theta} \right].$$

Or, in terms of the notation in Section A,

$$g_{y|\theta} = E_{x|\theta,y}[g_{x|\theta}], \qquad H_{y|\theta} = E_{x|\theta,y}[H_{x|\theta}] + V_{x|\theta,y}[g_{x|\theta}].$$

C Variance of $\partial L(\psi(\theta))/\partial\theta_1$

Let

$$m \equiv \begin{bmatrix} (c_1-\mu)-\phi(c_2-\mu) \\ (c_2-\mu)-\phi(c_3-\mu) \\ \vdots \\ (c_{n-1}-\mu)-\phi(c_n-\mu) \\ \sqrt{1-\phi^2}(c_n-\mu) \end{bmatrix} \quad \text{and} \quad v \equiv \begin{bmatrix} e_1-\phi e_2 \\ e_2-\phi e_3 \\ \vdots \\ e_{n-1}-\phi e_n \\ \sqrt{1-\phi^2}e_n \end{bmatrix}.$$

We want to find the variance of $(v+m)^\top\Lambda(v+m)$, where $\Lambda = I$. We can use the formula (assuming "Gaussianity")

$$\begin{aligned} \operatorname{Var}[(v+m)^\top\Lambda(v+m)] &= 2\operatorname{tr}[\Lambda\Sigma\Lambda\Sigma] + 4m^\top\Lambda\Sigma\Lambda m \\ &= 2\operatorname{tr}[\Sigma^2] + 4m^\top\Sigma m \\ &\equiv 2k_1 + 4k_2, \end{aligned}$$

where $\Sigma = \operatorname{Var}[v]$.

Note that $k_1 = \operatorname{tr}[\Sigma^2]$ is the sum of the squared elements of Σ. Now let's compute the elements of Σ.

The diagonal elements we already have from (**??**) and (**??**). The upper triangular elements are as follows. For $1 \le s < t < n$,

$$\begin{aligned} \Sigma_{st} &= E[(e_s-\phi e_{s+1})(e_t-\phi e_{t+1})] \\ &= E[((\dot\mu_s-\phi)e_{s+1}+u_s)(e_t-\phi e_{t+1})] \\ &= (\dot\mu_s-\phi)\dot\mu_{s+1}\dot\mu_{s+2}\cdots\dot\mu_{t-1}(E[e_t^2]-\phi\dot\mu_t E[e_{t+1}^2]). \end{aligned}$$

and

$$\begin{aligned} \Sigma_{sn} &= E[((\dot\mu_s-\phi)e_{s+1}+u_s)\sqrt{1-\phi^2}e_n] \\ &= (\dot\mu_s-\phi)\dot\mu_{s+1}\dot\mu_{s+2}\cdots\dot\mu_{n-1}\sqrt{1-\phi^2}E[e_n^2]. \end{aligned}$$

The sum of squared upper triangular elements—and by symmetry the sum of squared lower triangular elements—is

$$\sum_{s=1}^{n-1}\sum_{t=s+1}^{n}\Sigma_{st}^2 = \iota_{n-1}^\top DAz, \tag{12}$$

where $\iota_{n-1} = (1, 1, \ldots, 1)$, $D = \operatorname{diag}((\dot{\mu}_1 - \phi)^2, \ldots, (\dot{\mu}_{n-1} - \phi)^2)$,

$$A = \begin{bmatrix} 1 & \dot{\mu}_2^2 & \dot{\mu}_2^2\dot{\mu}_3^2 & \dot{\mu}_2^2\dot{\mu}_3^2\dot{\mu}_4^2 & \cdots \\ & 1 & \dot{\mu}_3^2 & \dot{\mu}_3^2\dot{\mu}_4^2 & \cdots \\ & & \ddots & & \\ & & & 1 & \dot{\mu}_{n-1}^2 \\ & & & & 1 \end{bmatrix} = \begin{bmatrix} 1 & -\dot{\mu}_2^2 & & & \\ & 1 & -\dot{\mu}_3^2 & & \\ & & 1 & \ddots & \\ & & & \ddots & -\dot{\mu}_{n-1}^2 \\ & & & & 1 \end{bmatrix}^{-1},$$

$$z = \begin{bmatrix} (E[e_2^2] - \phi\dot{\mu}_2 E[e_3^2])^2 \\ (E[e_3^2] - \phi\dot{\mu}_3 E[e_4^2])^2 \\ \vdots \\ (E[e_{n-1}^2] - \phi\dot{\mu}_2 E[e_n^2])^2 \\ (1 - \phi^2)(E[e_n^2])^2 \end{bmatrix}.$$

Now $v_2 = m^\top \Sigma m = m\Phi M^{-1} \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2) M^{-1} \Phi^\top m$, where

$$\Phi = \begin{bmatrix} 1 & -\phi & & & \\ & 1 & -\phi & & \\ & & 1 & \ddots & \\ & & & \ddots & -\phi \\ & & & & 1 \end{bmatrix}, \quad \text{and} \quad M = \begin{bmatrix} 1 & -\dot{\mu}_2 & & & \\ & 1 & -\dot{\mu}_3 & & \\ & & 1 & \ddots & \\ & & & \ddots & -\dot{\mu}_{n-1} \\ & & & & 1 \end{bmatrix}.$$