

# 1 Introduction

Consider a high frequency time series with observed values  $y_1, \dots, y_{T+S}$ , collectively denoted by  $y_{1:T+S}$ , with a forecaster interested in the forecast distribution of the future value  $y_{T+S+h}$ , where  $S \geq 1$ . In many applications forecasting uncertainty can be reduced by conditioning on additional data, so knowledge of the distribution of  $p(y_{T+S+h}|y_{1:T+S})$  is preferable to that of  $p(y_{T+S+h}|y_{1:T})$ .

Many time series models of interest contain a set of global parameters  $\theta$  and observation specific latent variables  $x_{0:T}$ , such that  $x_t$  is assumed to follow a Markovian structure and  $y_t$  is conditionally independent of  $y_{1:t-1}$ , given  $\theta$  and  $x_t$ . One of the most common computational techniques for models with latent variables is Markov Chain Monte Carlo (MCMC), which involves iterative sampling from a Markov Chain designed to converge to the true posterior distribution, however these approaches can have slow convergence rates. In this paper it is assumed that an MCMC algorithm will require a time period of length  $S$  to converge, so at time  $T + S$  the posterior distribution of  $\theta$  and  $x_{0:T+S}$  can be conditioned only on  $y_{1:T}$  and hence only  $p(y_{1:T+S+h}|y_{1:T})$  is available.

The additional data  $y_{T+1:T+S}$  can be included in the forecast by using filtering techniques such as the Kalman filter or particle filtering, however this will only update knowledge on  $x_{0:T+S}$  but not  $\theta$ . As an alternative to this method, this paper will instead approximate the posterior distribution  $p(\theta, x_T|y_{1:T})$  with  $q_\lambda(\theta, x_T|y_{1:T})$  and the updated posterior distribution  $p(\theta, x_{T+S}|y_{1:T+S})$  with  $q_\lambda(\theta, x_{T+S}|y_{1:T+S})$ . The parameters of this approximation, denoted by  $\lambda$ , are optimised by gradient descent to minimise the Kullback-Leibler divergence from  $q$  to  $p$ , a technique known as Variational Inference, which has seen wide use in the literature (references).

Variational Bayes involves an implicit trade-off: reducing the forecast horizon of  $y_{T+S+h}$  from  $S + h$  to  $h$  reduces the uncertainty of the forecast, but the approximation will increase the statistical error. This paper additionally explores that trade-off for a Stochastic Volatility Model, a common latent state time series model.

- Some motivation about forecasting and updates, can probably be copied from the MIVB file
- Not focusing on MCMC sample
- VB estimate at time T, VB update to time T+S
- End goal: An approximation  $q(\theta, x_T)$  for use in forecasting.

## 2 MCMC and Particle Filtering

- MCMC background

- Stochastic Volatility Model
- PMCMC
- Data driven particle filters?
- Whatever else

### 3 Variational Inference

Variational Inference posits a divergence function between the true posterior distribution  $p(\boldsymbol{\theta}, x_{0:T}|y_{1:T})$  and some approximating distribution  $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, x_{0:T}|y_{1:T})$ , choosing the parameters  $\boldsymbol{\lambda}$  for a given functional form  $q$  that minimises the divergence function.

This paper will follow the traditional approach, where the divergence function is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) from  $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, x_{0:T}|y_{1:T})$  to the true posterior  $p(\boldsymbol{\theta}, x_{0:T}|y_{1:T})$ . The KL divergence is defined by

$$KL[q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, x_{0:T}|y_{1:T}) || p(\boldsymbol{\theta}, x_{0:T}|y_{1:T})] = \int q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, x_{0:T}|y_{1:T}) \ln \left( \frac{q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, x_{0:T}|y_{1:T})}{p(\boldsymbol{\theta}, x_{0:T}|y_{1:T})} \right) d\boldsymbol{\theta} dx_{0:T} \quad (3.1)$$

and can alternatively be expressed as

$$KL[q(\boldsymbol{\theta}, x_{0:T}|\boldsymbol{\lambda}) || p(\boldsymbol{\theta}, x_{0:T}|y_{1:T})] = \ln(p(y_{1:T})) - \mathcal{L}(\boldsymbol{\lambda}) \quad (3.2)$$

where  $\mathcal{L}(\boldsymbol{\lambda})$  is referred to as the Evidence Lower Bound (ELBO), as it provides a lower bound on the unknown constant  $\ln(p(y_{1:T}))$ .  $\mathcal{L}(\boldsymbol{\lambda})$  is defined by

$$\mathcal{L}(\boldsymbol{\lambda}) = \int q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, x_{0:T}|y_{1:T}) \ln \left( \frac{p(y_{1:T}, \boldsymbol{\theta}, x_{0:T})}{q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, x_{0:T}|y_{1:T})} \right) d\boldsymbol{\theta} dx_{0:T}, \quad (3.3)$$

and as  $\ln(p(y_{1:T}))$  is constant with respect to  $\boldsymbol{\lambda}$ , maximising (3.3) with respect to  $\boldsymbol{\lambda}$  is equivalent to minimising (3.1). Maximising (3.3) is more convenient than minimising (3.1) as it is a function of the known joint distribution  $p(y_{1:T}, \boldsymbol{\theta}, x_{0:T})$  instead of the unknown posterior  $p(\boldsymbol{\theta}, x_{0:T}|y_{1:T})$ .

#### 3.1 ELBO Optimisation

Equation (3.3) can be maximised using a gradient ascent approach, where the following update step is iteratively applied until (3.3) converges within some pre-specified tolerance:

$$\boldsymbol{\lambda}^{(m+1)} = \boldsymbol{\lambda}^{(m)} + \rho^{(m)} \frac{\delta}{\delta \boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^{(m)}), \quad (3.4)$$

where the derivative is evaluated at  $\boldsymbol{\lambda}^{(m)}$ . This update requires some initial values  $\boldsymbol{\lambda}^{(0)}$  and a sequence  $\rho^{(m)}, m = 1, 2, \dots$  known as the learning rate. If  $\rho^{(m)}$  is chosen to satisfy the following conditions the algorithm is guaranteed to converge to a local maximum (Robbins and Monro, 1951).

$$\sum_{m=1}^{\infty} \rho^{(m)} = \infty \quad (3.5)$$

$$\sum_{m=1}^{\infty} (\rho^{(m)})^2 < \infty. \quad (3.6)$$

This paper uses Adam (Kingma and Ba, 2015) to generate the sequence  $\rho^{(m)}$ . Ranganath et al. (2014) showed that a Monte Carlo estimate of the derivative of the ELBO can be given by

$$\frac{\delta}{\delta \boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^{(m)}) \approx \frac{1}{N} \sum_{i=1}^N \frac{\delta}{\delta \boldsymbol{\lambda}} [\ln(q_{\boldsymbol{\lambda}^{(m)}}(\boldsymbol{\theta}_i, x_{0:T,i} | y_{1:T})) \ln \left( \frac{p(y_{1:T}, \boldsymbol{\theta}_i, x_{1:T,i})}{q_{\boldsymbol{\lambda}^{(m)}}(\boldsymbol{\theta}_i, x_{0:T,i} | y_{1:T})} \right)] \quad (3.7)$$

where  $i = 1, \dots, N$  indicates independent simulations from  $q_{\boldsymbol{\lambda}^{(m)}}(\boldsymbol{\theta}, x_{0:T} | y_{1:T})$ . The terms in the sum in (3.7) can have large variances, and in practice a large value of  $N$  is required to ensure a precise estimate of the gradient of the ELBO is obtained, slowing computation. The variance can be reduced by the reparameterisation trick of Kingma and Welling (2014), introducing a random vector  $\boldsymbol{\epsilon}$  with a distribution  $q(\boldsymbol{\epsilon})$  that contains no free parameters, and a differentiable transform  $f$  such that

$$f(\boldsymbol{\epsilon}, \boldsymbol{\lambda}) = \{\boldsymbol{\theta}, x_{1:T}\}. \quad (3.8)$$

Kingma and Welling (2014) show that an  $f$  exists to transform  $\boldsymbol{\epsilon}$  to any continuous random variable, with examples including a location-scale transform of a standard normal  $\epsilon$  and an inverse-CDF transform of a uniform  $\epsilon$ . In this case, (3.3) becomes

$$\mathcal{L}(\boldsymbol{\lambda}) = \int q(\boldsymbol{\epsilon}) \ln \left( \frac{p(y_{1:T}, f(\boldsymbol{\epsilon}, \boldsymbol{\lambda}) | \det J(\boldsymbol{\epsilon}, \boldsymbol{\lambda}))}{q(\boldsymbol{\epsilon})} \right) d\boldsymbol{\epsilon}, \quad (3.9)$$

where  $J(\boldsymbol{\epsilon}, \boldsymbol{\lambda})$  is the Jacobian matrix of the transformation  $f$ . The derivative of (3.9) can be estimated by

$$\frac{\delta}{\delta \boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^{(m)}) \approx \frac{1}{M} \sum_{i=1}^M \frac{\delta f(\boldsymbol{\epsilon}_i, \boldsymbol{\lambda}^{(m)})}{\delta \boldsymbol{\lambda}} \frac{\delta \ln(p(y_{1:T}, f(\boldsymbol{\epsilon}_i, \boldsymbol{\lambda}^{(m)})))}{\delta f(\boldsymbol{\epsilon}_i, \boldsymbol{\lambda}^{(m)})} + \frac{\delta \ln(|\det J(\boldsymbol{\lambda}^{(m)}, \boldsymbol{\epsilon}_i)|)}{\delta \boldsymbol{\lambda}}, \quad (3.10)$$

where simulations of  $\boldsymbol{\theta}$  and  $x_{1:T+S}$  are replaced by simulations of  $\boldsymbol{\epsilon}$  from  $q(\boldsymbol{\epsilon})$ . The variance of this estimator is often orders of magnitude smaller than the estimator in (3.7) (citations), so  $M$  can be set much lower than  $N$ .

### 3.2 Randomised Quasi Monte Carlo

(Gunawan et al., 2017) (Matousek, 1998) (Bratley and Bennet, 1988) (Sobol, 1967)

### 3.3 Choice of $q$ distribution

## 4 Dimensionality Reduction

ADVI simulation results

As the latent variables are Markovian, once the distribution of  $x_{T+1}$  is available, the distributions of each previous  $x_t, t \leq T$  is irrelevant to forecasts of  $y_{T+S+h}$  and it may be more convenient to construct an approximation for  $p(\boldsymbol{\theta}, x_{T+1}|y_{1:T})$  than for  $p(\boldsymbol{\theta}, x_{0:T+1}|y_{1:T})$ , as they contain the same information about  $y_{T+S+h}$  but the dimensionality of the posterior is reduced from  $T+k+2$  to  $k+1$ , where  $k$  is the number of elements in  $\boldsymbol{\theta}$ . This in turn allows for more expressive approximating distributions to be used, as the size of  $\boldsymbol{\lambda}$  does not grow with  $T$ .

In this case, using  $f(\boldsymbol{\epsilon}, \boldsymbol{\lambda}) = \{\boldsymbol{\theta}, x_{T+1}\}$ ,

$$\ln(p(y_{1:T}, f(\boldsymbol{\epsilon}, \boldsymbol{\lambda}))) = \ln \left( \int_{X_T} p(x_{T+1}|\boldsymbol{\theta}, x_T) p(x_T|y_{1:T}, \boldsymbol{\theta}) d_{X_T} \right) + \ln(p(y_{1:T}|\boldsymbol{\theta})) + \ln(p(\boldsymbol{\theta})) \quad (4.1)$$

In many cases, the distributions  $p(x_T|y_{1:T}, \boldsymbol{\theta})$  and  $p(y_{1:T}|\boldsymbol{\theta})$  are intractable, but Tran et al. (2017) note that the particle filter estimators  $\hat{p}(x_T|y_{1:T}, \boldsymbol{\theta})$  and  $\hat{p}(y_{1:T}|\boldsymbol{\theta})$  are sufficient substitutes in what they refer to as Variational Bayes with Intractable Likelihood (VBIL). The particle filter estimation of the distribution  $\hat{p}(x_T|y_{1:T}, \boldsymbol{\theta})$  is expressed as a discrete set of point masses  $x_T^{(k)}$  and weights  $\pi_T^{(k)}$ , where  $k = 1, \dots, N$  represent the particles of the particle filter. In this case, the integral in (4.1) reduces to the sum

$$\int_{X_T} p(x_{T+1}|\boldsymbol{\theta}, x_T) p(x_T|y_{1:T}, \boldsymbol{\theta}) d_{X_T} = \sum_{k=1}^N \pi_T^{(k)} p(x_{T+1}|\boldsymbol{\theta}, x_T^{(k)}), \quad (4.2)$$

and hence

$$\ln(p(y_{1:T}, f(\boldsymbol{\epsilon}, \boldsymbol{\lambda}))) \approx \ln \left( \sum_{k=1}^N \pi_T^{(k)} p(x_{T+1}|\boldsymbol{\theta}, x_T^{(k)}) \right) + \ln(\hat{p}(y_{1:T}|\boldsymbol{\theta})) + \ln(p(\boldsymbol{\theta})) \quad (4.3)$$

can be substituted into (3.10). This derivative of (4.3) with respect to  $\boldsymbol{\lambda}$  can be obtained with automatic differentiation tools such as those provided by Stan (Carpenter et al., 2015).

### 4.1 Importance Sampling

(Sakaya and Klami, 2017)

## 5 Updating

$$p(y_{1:T+S}, \theta, x_T, x_{T+S}) = p(\theta)p(y_{1:T+S}|\theta)p(x_T|\theta, y_{1:T+S})p(x_{T+S}|x_T, \theta, y_{1:T+S})$$

- Proper updates with filter/smooth
- Secondary approximation - no smoothing
- Straight up particle filter on old approximation
- PF on old MCMC
- Simulation Results

## 6 Empirical

## 7 Discussion

## References

- Bratley, P. and Bennet, L. F. (1988), “Algorithm 659: Implementing Sobol’s quasirandom sequence generator,” *ACM Transactions on Mathematical Software*, 14, 88–100.
- Carpenter, B., Hoffman, M. D., Brubaker, M., Lee, D., Li, P., and Betancourt, M. (2015), “The Stan Math Library: Reverse-Mode Automatic Differentiation in C++,” *ArXiv e-prints*.
- Gunawan, D., Tran, M. N., and Kohn, R. (2017), “Fast Inference for Intractable Likelihood Problems using Variational Bayes,” *ArXiv e-prints*.
- Kingma, D. P. and Ba, J. L. (2015), “Adam: A method for stochastic optimization,” in *Proceedings of the International Conference on Learning Representations 2015*.
- Kingma, D. P. and Welling, M. (2014), “Auto-Encoding Variational Bayes,” *ArXiv e-prints*.
- Kullback, S. and Leibler, R. A. (1951), “On Information and Sufficiency,” *The Annals of Mathematical Statistics*, 22, 79–86.
- Matousek, J. (1998), “On the L2-discrepancy for anchored boxes,” *Journal of Complexity*, 14, 527–556.
- Ranganath, R., Gerrish, S., and Blei, David, M. (2014), “Black Box Variational Inference,” in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, eds. Kaski, S. and Corander, J., JMLR W&CP, pp. 814–822.

- Robbins, H. and Monro, S. (1951), “A Stochastic Approximation Method,” *The Annals of Mathematical Statistics*, 22, 400.
- Sakaya, J. and Klami, A. (2017), “Importance Sampled Stochastic Optimization for Variational Inference,” in *UAI '17: Proceedings of the 33rd Conference in Uncertainty in Artificial Intelligence*.
- Sobol, I. M. (1967), “On the distribution of points in a cube and the approximate evaluation of integrals,” *USSR Computational Mathematics and Mathematical Physics*, 7, 86–112.
- Tran, M. N., Nott, D. J., and Kohn, R. (2017), “Variational Bayes with Intractable Likelihood,” *Journal of Computational and Graphical Statistics*, to appear.