Contents

1 Introduction

For many time-series applications a point forecast of the mean of the next observation in a series, y_{T+1}, can be easily obtained with frequentist methods, conditioning on the forecaster's observed series, $y_{1:T}$. Often the forecast is supplemented with a prediction interval backed by asymptotic theory, however there is a growing demand in the literature for density forecasts (find citations, probably Hyndman) which are typically much harder to obtain.

As a contrast to frequentism, the Bayesian methodology implicitly provides the entire probability density for variables of interest, in this case $p(y_{T+1}|y_{1:T})$. However the Bayesian approach requires solving integrals of the form

$$\int_\theta p(y_{T+1}|\theta)p(\theta|y_{1:T})d\theta, \tag{1.1}$$

where $p(\theta|y_{1:T})$ is known as the posterior distribution, with

$$p(\theta|y_{1:T}) = \frac{p(y_{1:T}|\theta)p(\theta)}{\int_\theta p(y_{1:T}|\theta)p(\theta)d\theta} \tag{1.2}$$

and $p(\theta)$ is some pre-specified prior distribution.

These often cannot be solved analytically, while numeric integration is computationally infeasible when the dimension of θ is large. To address this problem there is a wide range of techniques used to approximate the solution to the integral in (1.1), such as Markov Chain Monte Carlo (MCMC) and Variational

Bayes (VB). These approximations have an implicit trade-off: The better approximations are compuationaly intensive, and the forecaster must decide how much computation time and approximation error is acceptable. The focus of this research is in situtations where the time budget is too small for MCMC to be reliable, and uses VB as an alternative. This situtation is common in time-series forecasting, which must update the posterior distribution for each new data point observed via

$$p(\theta|y_{1:T+1}) = \frac{p(y_{T+1}|\theta)p(\theta|y_{1:T})}{\int_\theta p(y_{T+1}|\theta)p(\theta|y_{1:T})d\theta}. \tag{1.3}$$

2 Bayesian Inference

2.1 Exact Bayesian Computation

While it is technically an approximation method, MCMC algorithms result in what is often called an exact computation of the posterior. A Gibbs based MCMC iteratively samples from the conditional distributions

$$\begin{aligned}
&p(\theta_1|\theta_2, \ldots, \theta_p, y_{1:T}) \\
&p(\theta_2|\theta_1, \theta_3, \ldots, \theta_p, y_{1:T}) \\
&\vdots \\
&p(\theta_p|\theta_1, \ldots, \theta_{p-1}, y_{1:T})
\end{aligned}$$

where p is the dimension of θ. With enough iterations, the error in the approximation converges to zero and the algorithm can be ran for as much time as the forecaster desires to reduce error to a desired level. However, in the first iteration we must set arbitary starting values for each of $\theta_2, \ldots, \theta_p$ introducing a large amount of error in the early iterations. To avoid this error MCMC generally must be run for a large number of iterations and these early samples are discarded. The computation time per iteration and speed of convergence is extremely problem specific.

We illustrate this with an AR(2), a simple time series model used in the remainder of this section. The AR(2) is described by

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t \tag{2.1}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. We collect the unknown parameters as $\theta = (\mu, \rho_1, \rho_2, \sigma^2)'$ and set priors as

$$\begin{aligned}
p(\mu) &\propto 1 \\
p(\sigma^2) &\propto \sigma^{-2}
\end{aligned}$$

could make ρ a constant or add a uniform unit circle prior to polynomial roots to enforce stationarity - or maybe switch to an AR(1) or ARMA(1,1).

Describe conditional distributions, possible MH step, result in forecast distribution for y_{T+1}

2.2 Variational Bayes

Variational Bayes introduces some approximating distribution $q(\theta|\lambda)$ and aims to choose the family q and set of parameters λ so that $q(\theta|\lambda$ is as close as possible to the true posterior $p(\theta|y)$. It does this by minimising the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) from $q(\theta|\lambda)$ to $p(\theta|y)$. The KL divergence is defined by

$$KL[q(\theta|\lambda)||p(\theta|y)] = \int q(\theta|\lambda) \ln\left(\frac{q(\theta|\lambda)}{p(\theta|y)}\right) d\theta. \tag{2.2}$$

and is a non-negative, assymetric measure of the difference between $p(\theta|y)$ and $q(\theta|\lambda)$ that will equal zero if and only if $p(\theta|y) = q(\theta|\lambda)$ almost everywhere (Bishop, 2006). It has origins in information theory, and can be interpreted as: *'Given that I know $p(\theta|y)$ for some $\theta \in \Theta$, how much extra information is required, on average, to know the value of $q(\theta|\lambda)$?'* There are examples of the literature of approximations with other measures of divergence, such as Minka (2001) introducting Expectation Propogation (EP), which minimises the reverse measure $KL[p(\theta|y)||q(\theta|\lambda)]$, which was extended to Power-EP in Minka (2004), which aims to minimise the more general $\alpha-$divergence (Amari, 1985). It is shown in Bishop (2006) (note: find original proof) that minimising $KL[p(\theta|y)||q(\theta|\lambda)]$ used in in EP is equivalent to the MLE of $q(\theta|\lambda)$ given a sample of θ. However, we can write $KL[q(\theta|\lambda)||p(\theta|y)]$ as

$$KL[q(\theta|\lambda)||p(\theta|y)] = \ln(y) - \mathcal{L}(q, y) \tag{2.3}$$

where $\mathcal{L}(q, y)$ is known as the Evidence Lower BOund (ELBO), defined by

$$\mathcal{L}(q, y) = \int_\theta q(\theta|\lambda) \ln(p(y, \theta|\lambda))d\theta - \int_\theta q(\theta|\lambda) \ln(q(\theta|\lambda))d\theta. \tag{2.4}$$

From (2.4) it is clear that maximising $\mathcal{L}(q, y)$ with respect to q is equivalent to minimising (2.2). Maximising the ELBO is much more convenient than minimising either form of the KL Divergence, and has lead to Variational Bayes been used much more widely in the literature than alternatives such as EP.

2.2.1 The Mean Field Assumption

Mean Field Variational Bayes (MFVB) has origins in the physics literature (Chandler, 1987) and uses the assumption that the approximation dististribution factorises,

$$q(\theta|\lambda) = \prod_i q_i(\theta_i|\lambda_i). \tag{2.5}$$

This assumption is known as the Mean Field assumption and is has been widely used as it greatly simplifies maximisation of the ELBO (Jordan et al., 1999; Bishop, 2006). Each component θ_i may be a scalar or a vector, and λ_i is the component of the λ vector that parameterises the relevant factor $q_i(\theta_i|\lambda_i)$. From here, we will use the shorthand notation that $q_i = q_i(\theta_i|\lambda_i)$ and $q_{\setminus i} = \prod_{j \neq i} q_j$. Maximising the ELBO with respect to q_i is analytically involved but computationally simple, as (2.4) can be expressed as a function of q_i only and then maximised with respect to each q_i individually. Attias (1999) shows that

$$q(\theta_i|\lambda_i) \propto \exp(\mathbb{E}_{q_{\setminus i}}[\ln(p(y, \theta))]) \tag{2.6}$$

and maximisation can proceed by matching (2.6) to a known distribution. In the case that the likelihood and prior for θ_i form an exponential family conjugate pair, then the distributional family can be kept constant and all that is required is an update the parameters λ_i. If it does not match a known distribution, we may need to make a further approximation $q_i(\tilde{\theta_i}|\lambda_i)$ which has a recognizable distribution. One method as used in Friston et al. (2006) uses a Laplace approximation to and substitute in a Gaussian distribution for an otherwise unrecognizable $q_i(\theta_i|\lambda_i)$.

Matching distributions in this way is a very similar method to finding the posterior conditional distribution, $p(\theta_i|y, \theta_{\setminus i}$ in a Gibbs MCMC scheme, with dependence on other parameters replaced by their expectations. Hence, the optimal approximating family for θ_i, $q(\theta_i|\lambda_i)$, will come from the same distributional family as the conditional distribution $p(\theta_i|\theta_{j \neq i}, y)$, if it exists in a recognisable form such as the exponential family. The secondary Laplace approximation is analogous to a Metropolis-Hastings-within-Gibbs step in MCMC, as a way to handle unrecognisable distributions.

Given these optimal distributional families, a mean field updating equation for each λ_i can be found as a function of the data and other $\lambda_{j \neq i}$. This dependence requires an algorithm that continuously iterates between each λ_i and sets it to its maximising value until $\mathcal{L}(q(\theta|\lambda), y)$ converges within some pre-defined threshold. This is known as a coordinate ascent algorithm and follows below, where k as the dimension of λ.

Input: Log Joint Density
Result: Mean Field Approximation
Initialise λ randomly;
Use (**??**) to match each $q(\theta_i|\lambda_i)$ to a tractable distribution;
while Not converged **do**
  for $i = 1$ **to** k **do**
    Hold $\lambda_{j \neq i}$ fixed;
    Update λ_i using y and the most recent values of $\lambda_{j \neq i}$;
  end
end

Algorithm 1: Coordinate Ascent for MFVB

2.2.2 Stochastic Variational Bayes

In many cases, the posterior is too complex to be captured for a factorising approximation to be reasonabl as we desire our distribution $q(\theta|\lambda)$ to capture dependence between parameters. In this case, the easy maximisation in MFVB is unavailable, and we must resort to what is called Stochastic Variational Bayes (SVB) using a gradient ascent algorithm developed by Paisley et al. (2012) and **?**.

To maximise the function $\mathcal{L}(q(\theta|\lambda), y)$ we can take the derivative of $\mathcal{L}(q(\theta|\lambda), y)$ with respect to λ and use the updating step:

$$\lambda^{(m+1)} = \lambda^{(m)} + \rho^{(m)} \nabla_\lambda \mathcal{L}(q(\theta|\lambda^{(m)}), y), \tag{2.7}$$

where $\nabla_\lambda \mathcal{L}(q(\theta|\lambda^{(m)}), y)$ is the vector of partial derivatives of $\mathcal{L}(q(\theta|\lambda^{(m)}), y)$ with respect to each element of λ. This update requires initial values $\lambda^{(0)}$ and some learning rate sequence $\rho^{(m)}$. If $\rho^{(m)}$ is chosen to satisfy the following conditions, it is a Robbins-Monro sequence and the algorithm is guaranteed to converge to a local maximum.

$$\begin{aligned} \lim_{m \to \infty} \rho^{(m)} &= 0 \\ \sum_{m=1}^{\infty} \rho^{(m)} &= \infty \\ \sum_{m=1}^{\infty} (\rho^{(m)})^2 &< \infty. \end{aligned}$$

Whilst a global maximum is desired, the ELBO can have a problem specific shape that makes finding the global maximum extremely difficult, as we do not know how many stationary points exist. The dimension of the ELBO is the dimension of the λ vector, which is often much greater than the dimension of the parameters θ so a grid search is suspect to the curse of dimensionality. To alleiviate this problem, we can start the algorithm at a range of initial values choose the converged value with the highest value of the ELBO.

SVB does not provide the family of the the optimal approximating distribution, unlike under the mean field assumption. To run SVB we may need to try many approximating distributions and then choose the $q(\theta|\lambda)$ that has the highest ELBO, and hence lowest KL divergence to the true posterior. We must restrict the approximation to distributions that satisfy the condition that the order of differentation of the ELBO with respect to λ and integration with respect to θ are interchangable. In this case Ranganath et al. (2014) shows that a Monte Carlo estimate of the derivative of the ELBO can be taken by

$$\nabla_\lambda \mathcal{L}(q(\theta|\lambda^{(m)}) \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda [\ln(q(\theta_s|\lambda^{(m)}))](\ln(p(y, \theta_s)) - \ln(q(\theta_s|\lambda^{(m)}))) \tag{2.8}$$

where $s = 1, \ldots, S$ indicates simulations from $q(\theta|\lambda^{(m)})$.

As the distribution $q(\theta|\lambda)$ is specified by the user, the main restriction on the use of SVB is that the log-joint density $\ln(p(y, \theta))$ is able to be evaluated.

Duchi et al. (2011) introduces the AdaGrad algorithm which can be implemented within SVB to control $\rho^{(m)}$. AdaGrad allows each λ_i to have an independent $\rho_i^{(m)}$ that is inversely proportional to the gradient.

Let

$$G_i^{(m)} = \sum_{j=1}^{m} \left(\nabla_{\lambda_i} \mathcal{L}(q(\theta|\lambda^{(m)}) \right)^2, \tag{2.9}$$

then each component's learning rate is defined as

$$\rho_i^{(m)} = \eta \left(G_i^{(m)} \right)^{-1/2} \tag{2.10}$$

for some tuning parameter η.

The resulting Stochastic Gradient Ascent algorithm proceeds below.

Input: Log Joint Density, Approximation family q
Result: Variational Approximation
Initialise λ;
while Not converged **do**
  Simulate θ^s for $s = 1, \ldots S$ from $q(\theta|\lambda^{(m)})$;
  for $i = 1$ **to** k **do**
    Calculate $\nabla_{\lambda_i} =$
     $1/S \sum_{s=1}^{S} \nabla_{\lambda_i} [\log(q(\theta^s|\lambda^{(m)})](\log(p(\theta^s, y)) - \log(q(\theta^s|\lambda^{(m)})))$;
    Update $G_i^{(m)} = G_i^{(m-1)} + \left(\nabla_{\lambda_i} \mathcal{L}(q(\theta|\lambda^{(m)}) \right)^2$;
    Calculate $\rho_i^{(m)} = \left(G_i^{(m)} \right)^{-1/2}$;
  end
  Set $\lambda^{(m+1)} = \lambda^{(m)} + \rho^{(m)} \nabla_\lambda$;
  Set $m = m + 1$;
end

Algorithm 2: Stochastic Gradient Ascent for SVB

2.3 Variance Reduction Techniques

For many models the estimator in (2.8) has a large variance and computation time can be reduced with a more efficient estimator, and hence a lower value of S.

2.3.1 Rao Blackwellisation

Ranganath et al. (2014) considers the use of Rao Blackwellisation (Casella and Robert, 1996) -Rao Blackwellisation -Control Variates -Reparameterisation

2.4 AR2 model example (Revisited using VB)

3 Electricity Load Forecasts

3.1 Motivation

Will you include a subsection showing an overview of the actual electricity load data you have?? Real data has five minute updates Density is of interest due to rare price spikes Slow convergence of MCMC / Bad scaling to large dimensional models and datasets

3.2 Exponential Smoothing

Bayesian Version Difficulty with smoothing parameters in large models Infer model structure from draws where possible

3.3 Variational Bayes Implementation

Posterior Distributions should not change much as more data is observed, so we can keep approximating distribution family constant But we might want a model that does change to make updates worthwhile - Markov Switching mechanism to be added?

3.4 Forecasting

compare to MCMC when possible

4 Timeline

References

Amari, S. (1985), *Differential-geometrical methods in statistics*, Springer-Verlag.

Attias, H. (1999), "A Variational Bayesian Framework for Graphical Models," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pp. 209–215.

Bishop, C. (2006), *Pattern Recognition and Machine Learning*, Springer.

Casella, G. and Robert, C. (1996), "Rao-blackwellisation of sampling schemes," *Biometrika*, 83, 81–94.

Chandler, D. (1987), *Introduction to Modern Statistical Mechanics*, Oxford University Press.

Duchi, J., Hazan, E., and Singer, Y. (2011), "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research*, 2121–2159.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburnder, J., and Penny, W. (2006), "Variational free energy and the Laplace approximation," *NeuroImage*, 220–234.

Jordan, M., Ghahramani, Z., Jaakola, T., and Saul, L. (1999), "An introduction to variational methods for graphical models," *Machine Learning*, 183–233.

Kullback, S. and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86.

Minka, T. (2001), "Expectation Propagation for Approximate Bayesian Inference," in *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, eds. Breese, J. S. and Koller, D., pp. 362–369.

— (2004), "Power EP," Tech. rep., Microsoft Research, Cambridge.

Paisley, J., Blei, D. M., and Jordan, M. I. (2012), "Variational Bayesian Inference with Stochastic Search," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, eds. Langford, J. and Pineau, J., pp. 1367–1374.

Ranganath, R., Gerrish, S., and Blei, D. (2014), "Black Box Variational Inference," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, eds. Kaski, S. and Corander, J., pp. 814–822.