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 (**??**), 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

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

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 $\theta_2, \dots, \theta_p$ 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

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

add something for ρ to enforce stationarity - 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 (**?**) 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 (**?**). 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 **?** introducting Expectation Propogation (EP), which minimises the reverse measure $KL[p(\theta|y)||q(\theta|\lambda)]$, which was extended to Power-EP in **?**, which aims to minimise the more general $\alpha-$divergence (**?**). It is shown in **?** (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 (**??**) it is clear that maximising $\mathcal{L}(q, y)$ with respect to q is equivalent to minimising (**??**). 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

Note the change to subsubsection Mean Field Derivation gives optimal family of approximating distributions Lots of analytical work, fast algorithm, factorisiation is a weakness

2.2.2 Stochastic Variational Bayes

Note the change to subsubsection Mean Field Assumption not required, but needs a pre-defined approximating distribution family Same algorithm fits most problems Copula augmentation allows an easy way to model complex posterior dependencies

2.3 Variance Reduction Techniques

This is general stuff, not just for VB, so can be a subsection -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.

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

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,” pp. 362–369.

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