Real Time Variational Density Forecasting

Nathaniel Tomasetti

March 1, 2017

Contents

1 Introduction

Electricity prices have been subject to an extensive forecasting literature and are known to be extremely volatile. For the 2011-2015 period in Victoria the mean price per megawatt hour (mWh) was \$42.06, with 99% of observations less than \$100 per mWh, however on several occasions the price spiked to greater than $10,000\%$ of the mean value. For this reason market participants are increasingly interested in a predictive density forecast, as the probability that the price will rise above a particular threshold in the next time period can be more important than the mean value. To calculate a suitably accurate predictive density the Bayesian methodology is to be adopted in this thesis, as it accounts for both parameter and model uncertainty in a relatively straightforward manner, which can be difficult in frequentist methods. See Geweke and Whiteman (2006) for a review on Bayesian forecasting and Gneiting and Katzfuss (2014) for density forecasting in general.

Bayesian methods are computationally complex as they require integration over the set of unknowns, whether parameters or latent stochastic variables. This set is typically quite large for models used to model short term electricity load and price as there are strong seasonal patterns in the data (Taylor, 2003). Analytical integration is typically not feasible to obtain the predictive distribution for such models, however there are a range of numerical and simulation-based techniques available for this purpose, notably Markov Chain Monte Carlo (MCMC) methods. While it is often possible to demonstrate theoretical convergence of MCMC algorithms, computation may be slow for even moderately complex models. This is problematic when attempting to forecast electricity load in Victoria, as realizations occur at five minute intervals, the predictive density for the next period's electricity load may not be available before it is observed.

Alternatives to MCMC include Variational Bayes (VB) methods, where the aim is to replace the unknown true posterior distribution for a vector of unknowns with a parametric approximation. The Variational Bayes approach will be explored in the thesis, to determine if it is a viable alternative computation strategy to MCMC, in the context of modelling and short term forecasting of Victorian electricity load. In particular, we seek to understand the benefits gained in terms of the computational time required to obtain a VB approximate forecast distribution against the loss in statistical accuracy associated with the use of an approximate posterior forecast distribution.

The literature on Variational Bayes is dominated by two major implementations: Mean Field Variational Bayes (MFVB) and what we call Stochastic Variational Bayes (SVB). For either implementation we emphasise that there are two seperate tasks: Choose a functional form for the approximating distribution, and then optimise the parameters of this distribution to minimise a well defined divergence between the approximation and the true posterior. A recent review of Variational Bayes is provided by Blei et al. (2017).

Mean Field Variational Bayes (see Bishop (2006) for an overview) can be applied to exponetial family models with conjugate priors, and restricts the class of approximating distributions to the factorisable family. MFVB provides both the optimal form of the approximating distribution within this class, and a computationally simple coordinate ascent algorithm is available to optimise parameters. Stochastic Variational Bayes (see Paisley et al. (2012) and Ranganath et al. (2014)) lifts both the restrictions for an exponential family model and the use of an approximating distribution that is factorisable. While parameter optimisation has a well developed gradient ascent algorithm, the determination of the functional form to be used is an active area of research.

We explore the use of a copula to augment the approximating distribution with a flexible dependence structure, which has seen success in Tran et al. (2015) and Ng et al. (2016). The time series context of our implementation implies that MCMC samples for out of date samples are available, and we emphasise the use of this information to guide the construction of a vine copula so that the approximation has a dependence structure similiar to the true posterior.

The discussion of MCMC and VB will be continued in Section 2, with the methods contrasted in the context of two simple models. Section 3 will introduce a planned empirical application of VB to Victorian electricity load using an exponential smoothing model and data from the 2012-2015 period. Section 4 outlines the major areas to which this thesis will make a contribution, and Section 5 outlines the planned timeline for these contributions.

2 Bayesian Inference

To facilitate the discussion, consideration is given to models where the unknowns may be summarized in a (finite) $k-$dimensional vector, denoted by $\boldsymbol{\theta}$. In this setting, given an observed time series, denoted by $y_{1:T} = \{y_t, t = 1, \dots T\}$, the Bayesian forecast distribution associated with time $t = T + 1$ is characterized by the conditional density

$$p(y_{T+1}|y_{1:T}) = \int p(y_{T+1}|y_{1:T}, \boldsymbol{\theta})p(\boldsymbol{\theta}|y_{1:T})d\boldsymbol{\theta}, \tag{2.1}$$

where $p(\boldsymbol{\theta}|y_{1:T})$ denotes the posterior density for $\boldsymbol{\theta}$, given by

$$p(\boldsymbol{\theta}|y_{1:T}) = \frac{p(y_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(y_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{2.2}$$

Generally the analytical solution to (2.2), and hence to (2.1) will be unavailable due to the complex functional form of the distributions involved. In this section two alternative methods for computing the desired posterior distribution in (2.2), MCMC and VB, will be reviewed. In brief, MCMC is used to create a sample from $p(\boldsymbol{\theta}|y_{1:T})$, with any function of $\boldsymbol{\theta}$ that is desired estimated from that sample. In contrast, VB replaces $p(\boldsymbol{\theta}|y_{1:T})$ with an approximation, denoted by $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a vector of auxiliary parameters associated with the approximation that may depend on the observations $y_{1:T}$. For example, if $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is Gaussian, then $\boldsymbol{\lambda}$ would denote the mean and variance parameters that characterise this distribution for $\boldsymbol{\theta}$. Variational Bayes aims to choose the distribution $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ within a restricted class such that the a measure of divergence, typically the Kullback-Leibler divergence from $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ to $p(\boldsymbol{\theta}|y_{1:T})$, is minimised. Once the posterior distribution in (2.2) is available, the predictive density forecast (2.1) can be found by simulation.

2.1 Markov Chain Monte Carlo

There are many types of MCMC algorithms, with arguably the simplest and most commonly used one being the Gibbs sampler. The Gibbs sampler algorithm iteratively samples the components of the $k-$dimensional parameter vector $\boldsymbol{\theta}$ via each of the so-called full conditional distributions as follows,

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

Under mild regularity conditions (see, e.g., Tierney (1994)) and with enough iterations of the Markov chain that results from the Gibbs sampler, these samples converge in distribution to $p(\boldsymbol{\theta}|y_{1:T})$. Samples taken before the MCMC converges to the posterior must be discarded, and the remaining samples may have strong dependence between consecutive draws of the same parameter due to the Markov nature of the algorithm. The computation time for each iteration and the overall number of iterations required to accruately summarise the posterior distribution is problem specific and typically increases with the number of parameters in the model. On occasion, the full conditional distribution cannot be recognised and a Metropolis-Hastings-within-Gibbs step can be used, see Gilks et al. (1995) for details.

For illustration, consider data y_t for $t = 1,\ldots,T$ generated independently from a $\mathcal{N}(\mu,\sigma^2)$ distribution with prior distributions given by $\mu|\sigma^2 \sim \mathcal{N}(\gamma,\sigma^2/\tau)$ and $\sigma^2 \sim IG(\text{shape} = \alpha, \text{scale} = \beta)$, where IG repesents the inverse of a Gamma distribution. It can be shown that

$$\mu|\sigma^2, y_{1:T} \sim \mathcal{N}\left(\frac{\tau\gamma + T\bar{y}}{\tau+T}, \frac{\sigma^2}{\tau+T}\right) \tag{2.3}$$

$$\sigma^2|\mu, y_{1:T} \sim IG\left(\text{shape} = \alpha + T/2, \text{scale} = \beta + \frac{1}{2}\sum_{t=1}^{T} y_t^2 + \tau\gamma^2 - \frac{(\tau\gamma+T\bar{y})^2}{\tau+T}\right) \tag{2.4}$$

where $\bar{y} = 1/T\sum_{t=1}^{T} y_t$. Posterior sampling can proceed by first simulating σ^2 from (2.4) and conditioning on this draw to simulate μ from (2.3).

2.2 Variational Bayes

As discussed in Section 1, a faster (albeit approximate) alternative to MCMC is VB. VB is an umbrella term for a broad collection of algorithms that introduce an approximating distribution $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ with the choice of the functional form q and subsequent auxiliary parameter vector $\boldsymbol{\lambda}$ determined so that an error function is minimised. The classical choice for this error function is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) from $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ the true posterior $p(\boldsymbol{\theta}|y_{1:T})$. The KL divergence is defined by

$$KL[q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \,||\, p(\boldsymbol{\theta}|y_{1:T})] = \int q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \ln\left(\frac{q(\boldsymbol{\theta}|\boldsymbol{\lambda})}{p(\boldsymbol{\theta}|y_{1:T})}\right) d\boldsymbol{\theta}. \tag{2.5}$$

The KL divergence is a non-negative, asymetric measure of the discrepancy between $p(\boldsymbol{\theta}|y_{1:T})$ and $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ that will be equal to zero if and only if $p(\boldsymbol{\theta}|y_{1:T}) = q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ almost everywhere (Bishop, 2006). Note that $KL[q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \,||\, p(\boldsymbol{\theta}|y_{1:T})]$ can be expressed as

$$KL[q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \,||\, p(\boldsymbol{\theta}|y_{1:T})] = \ln(p(y_{1:T})) - \mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T}) \tag{2.6}$$

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

$$\mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T}) = \int q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \ln(p(y_{1:T}, \boldsymbol{\theta}))d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \ln(q(\boldsymbol{\theta}|\boldsymbol{\lambda}))d\boldsymbol{\theta}. \tag{2.7}$$

Since $\ln(p(y_{1:T}))$ is constant with respect to $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$, maximising (2.7) with respect to $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is equivalent to minimising (2.5). Maximising the ELBO is much more convenient than minimising the KL Divergence, as will be shown in the following sections in the context of the two major implementations of Variational Bayes we consider, Mean Field Variational Bayes and Stochastic Variational Bayes. Each of these implementations take advantage of the functional form of the ELBO, and offer computationally efficient algorithms to find a $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ that is optimal within a particular class of distributions. The statistical properties of $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ in MFVB and SVB are largely driven by the choice of the KL divergence from $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ to $p(\boldsymbol{\theta}|y_{1:T})$ as an error function, and there is significant interest in other error functions, such as the KL divergence from $p(\boldsymbol{\theta}|y_{1:T})$ to $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ (Minka, 2001) and Stein's discrepency (Liu and Wang, 2016; Ranganath et al., 2016a).

2.2.1 Mean Field Variational Bayes

Mean Field Variational Bayes (MFVB) has origins in physics (Amari, 1985; Chandler, 1987) and has since developed a large literature in machine learning, notably developed in Jordan et al. (1999) and Gahramani and Beal (2001). MFVB restricts the class of distributions from which to select $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ to the set of factorisable distributions,

$$q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \prod_{i=1}^{k} q_i(\theta_i|\boldsymbol{\lambda_i}). \tag{2.8}$$

Each of the k components θ_i may be a vector, but in most implementations are scalars. Each θ_i has an associated vector $\boldsymbol{\lambda_i}$ which may have different dimension to θ_i. $\boldsymbol{\lambda_i}$ is an auxillary parameter vector for the relevant factor q_i, which will be used in this section as shorthand notation for the distribution $q_i(\theta_i|\boldsymbol{\lambda_i})$.

MFVB is widely used as it greatly simplifies maximisation of the ELBO, and has a straight-forward implementation in exponential family models (Wainwright and Jordan, 2008). Maximising the ELBO with respect to q_i is as analytically involved as deriving the conditional distributions used in Gibbs based MCMC schemes, but the resulting algorithm to maximise the ELBO is much simpler computationally than sampling from a Gibbs

MCMC scheme. MFVB factorises (2.7) into a product of k separate components that each contain only one q_i, and then maximises each of these factors with respect to the associated q_i. Using the notation $q_{\setminus i} = \prod_{j \neq i} q_j$, Attias (1999) shows that each component is maximised by

$$q_i(\theta_i|\boldsymbol{\lambda_i}) \propto \exp(\mathbb{E}_{q_{\setminus i}}[\ln(p(y_{1:T}, \boldsymbol{\theta}))]). \tag{2.9}$$

We illustrate the implementation of MFVB with the model considered in Section (2.1), where y_t for $t = 1, \dots, T$ is generated independently from a $\mathcal{N}(\mu, \sigma^2)$ distribution, and prior distributions are given by $\mu \sim \mathcal{N}(\gamma, \sigma^2/\tau)$ and $\sigma^2 \sim IG(\text{shape} = \alpha, \text{scale} = \beta)$. It can be shown that

$$q(\mu|\boldsymbol{\lambda}_1) \propto \exp\left\{ \frac{-(T+\tau)\mathbb{E}_{q(\sigma^2)}[\sigma^{-2}]}{2} \left(\mu - \frac{\tau\gamma + \sum_{t=1}^T y_t}{T+\tau} \right)^2 \right\} \tag{2.10}$$

and

$$\begin{aligned} q(\sigma^2|\boldsymbol{\lambda}_2) &\propto \sigma^{-2((T+1)/2+\alpha+1)} \\ &\times \exp\left\{ \frac{\beta + 1/2\left((\tau+T)\mathbb{E}_{q(\mu)}[\mu^2] - 2\mathbb{E}_{q(\mu)}[\mu]\left(\sum_{t=1}^T y_t + \tau\right) + \sum_{t=1}^T y_t^2 + \tau\gamma^2 \right)}{-\sigma^2} \right\}. \end{aligned} \tag{2.11}$$

As the prior distributions used were conjugate to the likelihood, the Variational Bayes optimal distributions for $q(\mu|\boldsymbol{\lambda}_1)$ and $q(\sigma^2|\boldsymbol{\lambda}_2)$ are the same family as the prior distributions, and are denoted by $q(\mu|\boldsymbol{\lambda}_1) \sim \mathcal{N}(\tilde{\gamma}, \tilde{\tau})$ and $q(\sigma^2|\boldsymbol{\lambda}_2) \sim IG(\tilde{\alpha}, \tilde{\beta})$. The vector $\boldsymbol{\lambda} = (\tilde{\gamma}, \tilde{\tau}, \tilde{\alpha}, \tilde{\beta})'$ is given by

$$\tilde{\gamma} = \frac{\tau\gamma + \sum_{t=1}^T y_t}{T+\tau} \tag{2.12}$$

$$\tilde{\tau} = \left((T+\tau)\mathbb{E}_{q(\sigma^2)}[\sigma^{-2}]\right)^{-1} \tag{2.13}$$

$$\tilde{\alpha} = (T+1)/2 + \alpha \tag{2.14}$$

$$\tilde{\beta} = \beta + 1/2\left((\tau+T)\mathbb{E}_{q(\mu)}[\mu^2] - 2\mathbb{E}_{q(\mu)}[\mu]\left(\sum_{t=1}^T y_t + \tau\right) + \sum_{t=1}^T y_t^2 + \tau\gamma^2 \right). \tag{2.15}$$

.

The expectations in (2.12) - (2.15) are available in a closed form for these distributions, and these are substituted in to derive the set of mean field equations:

$$\tilde{\tau} = \frac{\tilde{\beta}}{\tilde{\alpha}(T+\tau)} \tag{2.16}$$

$$\tilde{\beta} = \beta + 1/2\left((\tau+T)(\tilde{\gamma}^2 + \tilde{\tau}) - 2\tilde{\gamma}\left(\sum_{t=1}^T y_t + \tau\right) + \sum_{t=1}^T y_t^2 + \tau\gamma^2 \right) \tag{2.17}$$

$\tilde{\gamma}$ and $\tilde{\alpha}$ are omitted from the mean field equations as their values do not depend on other $\boldsymbol{\lambda}$ parameters.

As (2.16) and (2.17) contain a circular dependence, fitting the MFVB approximation requires an algorithm that cycles between these equations until the change in the ELBO is less than some pre-specified small constant ϵ. This algorithm is known as a coordinate ascent, and is described by Algorithm (1). At each step coordinate ascent will fix one parameter and then maximise the ELBO with respect to the other parameter, so that changes in parameters results in a non-decreasing sequence of ELBO values. Coordinate ascent converges only when neither parameter can be changed independently to increase the ELBO by more than ϵ, and hence reaches a local maximum. The current theory for Variational Bayes does not offer any insight to the shape of the ELBO so determining if the algorithm converges to a local or global maximum is difficult. One option to alleviate this problem is to start the algorithm at a range of initial values and choose parameter values with the maximum converged ELBO, and hence lowest KL divergence to the true posterior.

```
Input: Log Joint Density
Result: Mean Field Approximation
Use (2.9) to match $q(\mu|\boldsymbol{\lambda}_1)$ to a Gaussian distributon and $q(\sigma^2|\boldsymbol{\lambda}_2)$ to an Inverse
Gamma distribution;
Anaytically derive the set of mean field equations in (2.16) - (2.17);
Set $\tilde{\gamma}$ using 2.12.;
Set $\tilde{\alpha}$ using 2.14.;
Initialise $\tilde{\tau}^{(1)}$ and $\tilde{\beta}^{(1)}$ randomly.;
Evaluate $\mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T})^{(1)}$ using $\boldsymbol{\lambda}^{(1)}$.;
while $\mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T})^{(m)} - \mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T})^{(m-1)} > \epsilon$ do
    Set $\tilde{\tau}^{(m)}$ using 2.16.;
    Set $\tilde{\beta}^{(m)}$ using 2.17.;
    Evaluate $\mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T})^{(m)}$.;
    Set $m = m + 1$.;
end
```

Algorithm 1: Coordinate Ascent for MFVB

Data was simulated according to the Gaussian model introduced in Section (2.1) with $T = 100, \mu = 2$ and $\sigma^2 = 1$. The hyperparameters were set to $\gamma = 0, \tau = 1, \alpha = 1$ and $\beta = 1$. Figure (1) compares the analytically available true posterior marginal densities for both μ and σ^2 (black) with the densities generated by the Monte Carlo sampler (red) and the approximation from Mean Field Variational Bayes (blue). For both parameters all three densities are indistinguishable.

The set of mean field equations are only obtainable in the event that a prior that is conjugate to the model likelihood is used, restricting MFVB to exponential family models (Beal, 2003). In this case the optimal distributional family q_i is the same as the prior,

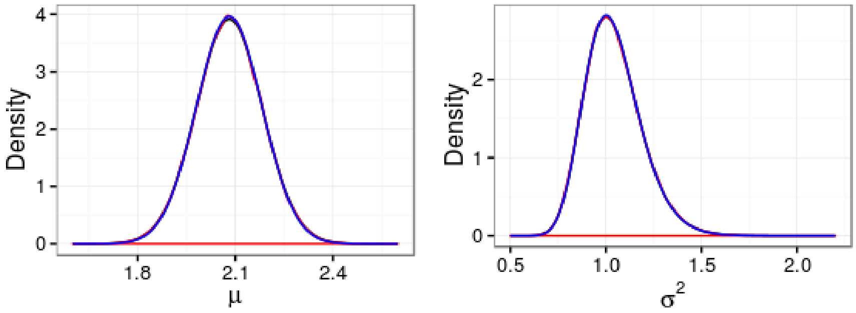

Figure 1: Marginal posterior densities for the model described in Section (2.1). The true density is in black, the Monte Carlo sampler is red and the MFVB approximation is blue.

analagous to the relationship between a prior-posterior conjugate pair. In the event that a non-conjugate prior is used, the form of the distribution in (2.9) is unrecognisable, a further approximation may be used by substituting in another distribution to replace the problematic $q_i(\theta_i|\boldsymbol{\lambda_i})$. A common option is the use of a Gaussian distribution through either a Laplace approximation (Friston et al., 2006) or a Delta Method inspired transform (Wang and Blei, 2013). This secondary level of approximation is analogous to the requirement of a Metropolis-Hastings-within-Gibbs step in MCMC to handle unrecognisable full conditional distributions.

2.2.2 Stochastic Variational Bayes

The requirement for an exponential family model without introducing further approximations, as well as restricting the approximating distribution to the factorisable family may be unsatisfactory in many applications. Paisley et al. (2012) and Ranganath et al. (2014) have adapted a gradient ascent algorithm for use in Variational Bayes, which selects optimal $\boldsymbol{\lambda}$ for a $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ from the class of distributions for which the score function can be evaluated, and the order of differentation of the ELBO with respect to $\boldsymbol{\lambda}$ and integration with respect to $\boldsymbol{\theta}$ are interchangable. The only model restriction is that the log likelihood for $y_{1:T}$ is able to be evaluated. We refer to the result as Stochastic Variational Bayes (SVB).

The application of gradient ascent in SVB iteratively takes the derivative of $\mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T})$ in (2.7) with respect to $\boldsymbol{\lambda}$ and the following updating step is applied until change in the ELBO is less than ϵ.

$$\boldsymbol{\lambda}^{(m+1)} = \boldsymbol{\lambda}^{(m)} + \rho^{(m)} \nabla_{\boldsymbol{\lambda}} \mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}^{(m)}), y_{1:T}), \tag{2.18}$$

where $\nabla_{\boldsymbol{\lambda}} \mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}^{(m)}), y_{1:T})$ is the vector of partial derivatives of $\mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}^{(m)}), y_{1:T})$ with respect to each element of $\boldsymbol{\lambda}$. This update requires some initial values $\boldsymbol{\lambda}^{(0)}$ and a sequence $\rho^{(m)}, m = 1, 2, \ldots$ 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).

$$\sum_{m=1}^{\infty} \rho^{(m)} = \infty \tag{2.19}$$

$$\sum_{m=1}^{\infty} (\rho^{(m)})^2 < \infty. \tag{2.20}$$

Ranganath et al. (2014) showed that a Monte Carlo estimate of the derivative of the ELBO can be given by

$$\nabla_{\lambda} \mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}^{(m)}), y_{1:T}) \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_{\boldsymbol{\lambda}} [\ln(q(\boldsymbol{\theta}_s|\boldsymbol{\lambda}^{(m)}))](\ln(p(y_{1:T}, \boldsymbol{\theta}_s)) - \ln(q(\boldsymbol{\theta}_s|\boldsymbol{\lambda}^{(m)}))) \tag{2.21}$$

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

Kingma and Welling (2014) introduces the re-parameterisation trick, where an auxillary noise variable $\boldsymbol{\epsilon}$ and differentiable transformation $f(\cdot, \cdot)$ are introduced such that $\boldsymbol{\theta} = f(\boldsymbol{\epsilon}, \boldsymbol{\lambda}$. Examples include a location scale transformation from a standard normal ϵ or an inverse-CDF transformation from a uniform$(0, 1)\epsilon$. In this case (2.21) can be replaced with a lower variance estimator,

$$\nabla_{\lambda} \mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}^{(m)}), y_{1:T}) \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_{\boldsymbol{\lambda}} \left(\ln(p(y_{1:T}, f(\boldsymbol{\epsilon}_s, \boldsymbol{\lambda}^{(m)})) - \ln(q(f(\boldsymbol{\epsilon}_s, \boldsymbol{\lambda}^{(m)}))) \right) \tag{2.22}$$

where $s = 1, \ldots, S$ now indicates simulations from $q(\boldsymbol{\epsilon})$. This transformation removes the restriction that $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ belong to a distribution where the score function can be evaluated.

Duchi et al. (2011) developed 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, so λ takes bigger steps in flat regions and smaller steps in steep regions.

Let

$$G_i^{(m)} = \sum_{j=1}^{m} \left(\nabla_{\lambda_i} \mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}^{(j)}), y_{1:T}) \right)^2, \tag{2.23}$$

then each component's learning rate is defined as

$$\rho_i^{(m)} = \eta \left(G_i^{(m)} \right)^{-1/2} \tag{2.24}$$

for some tuning parameter η.

The resulting Stochastic Gradient Ascent algorithm without reparameterisation proceeds below in Algorithm (2) with a p dimensional $\boldsymbol{\lambda}$ vector.

```
Input: Log Joint Density, Approximation family q
Result: Variational Approximation
Initialise $\boldsymbol{\lambda}^{(1)}$;
Evaluate $\mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T})^{(1)}$ using $\boldsymbol{\lambda}^{(1)}$.;
while $\mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T})^{(m)} - \mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T})^{(m-1)} > \epsilon$ do
    Simulate $\boldsymbol{\theta}^s$ for $s = 1, \dots S$ from $q(\boldsymbol{\theta}|\boldsymbol{\lambda}^{(m)})$;
    for $i = 1$ to $p$ do
        Calculate $\nabla_{\lambda_i}$ from (2.21);
        Update $G_i^{(m)}$ and $\rho_i^{(m)}$ from (2.23) and (2.24);
    end
    Update $\boldsymbol{\lambda}^{(m+1)}$ from (2.18);
    Evaluate $\mathcal{L}(q(\boldsymbol{\theta}|\boldsymbol{\lambda}), y_{1:T})^{(m)}$;
    Set $m = m + 1$;
end
```

Algorithm 2: Stochastic Gradient Ascent for SVB

2.2.3 Selection of the Approximating Distribution

Armed with an algorithm to find the Variational Bayes optimal parameters $\boldsymbol{\lambda}$ for a distribution q, there is a requirement to select a distribution q that well approximates the true posterior. If reparameterised aproximating distribution is used to calculate the gradients in (2.22) a function $f(\cdot, \cdot)$ that satisfies

$$f(\boldsymbol{\epsilon}, \boldsymbol{\lambda}) = p^{-1}(\boldsymbol{\epsilon}) \tag{2.25}$$

almost everywhere, where $\epsilon \sim U(0, 1)$ and p^{-1} denotes the inverse CDF of the true posterior, will yeild $q(\boldsymbol{\theta}|\boldsymbol{\lambda} = p(\boldsymbol{\theta}|y_{1:T}$ and thus minimise the KL divergence. Attempts to model a suitably flexible function $f(\cdot, \cdot)$ include heirarchical models (Ranganath et al., 2016b), neural networks (Kingma and Welling, 2014; Rezende et al., 2014; Rezende and Mohamed, 2015; Kingma et al., 2016), and Gaussian processes (Tran et al., 2016).

These approaches introduce a large number of parameters that would considerably slow computation of updates, so we follow Tran et al. (2015) and use copulas to flexibly model approximations to the true posterior distribution. The copula modelling class provides a flexible way to construct these distributions, as they allow the dependence structure between parameters to be fit independently from the marginal distributions. Sklar (1959) proves that any joint probability distribution can be written as the product of the marginals and a copula function,

$$p(\theta_1, \dots, \theta_k) = p(\theta_1) \dots p(\theta_k) c(P(\theta_1), \dots, P(\theta_k)) \tag{2.26}$$

where $p(\theta)$ is a pdf, $P(\theta)$ is a cdf, and $c(\cdot)$ is a copula.

The flexibility of a copula is extended by the vine copula, a structure that permits the factorisation of a high-dimensional copula into a set of bivariate copulas, see Joe (2014) and references within. A vine copula is represented by a series of trees, where the top-most tree is a graph that contains each variable as a node. The edges in each tree then form the nodes in the tree below, which are connected by an edge only if the associated edges in the previous tree shared the same node. This sequence is continued for a k dimensional $\boldsymbol{\theta}$ until the $(k-1)'th$ tree is reached that contains only one edge. Each of the $k(k-1)/2$ edges represents a bivariate copula between the two unique parameters in the connected nodes, conditioned on every parameter appearing in both nodes. Refer to Figure (2) for an example.

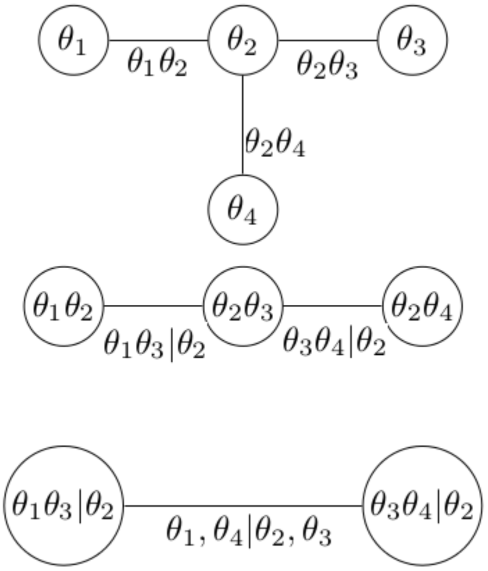

Figure 2: An example of a Vine Copula over four variables. The edges in the top tree represent unconditional bivariate copulas between the parameters on the connected nodes. The second tree edges represent the bivariate copulas for $\theta_1\theta_3|\theta_2$ and for $\theta_3\theta_4|\theta_2$, the conditioning on θ_2 caused by its appearance in each connected node. The edge in the final tree represents the bivariate copula for $\theta_1, \theta_4|\theta_2, \theta_3$, as both θ_2 and θ_3 appear in the connected nodes.

The flexibility in selecting a factorisation and set of bivariate copula familys in vine copula greatly increases the ability to model dependency in a distribution, but the cardinality of the set of possible vine structures and choices of family for each bivariate copula grows factorially with the number of parameters, so without detailed problem specific information the application of a vine copula to Stochastic Variational Bayes is difficult.

Given a sample of $\boldsymbol{\theta}$, Dißmann's algorithm (Dißmann et al., 2013) can be used to select the best vine factorisation by selecting a maximum spanning tree for the first tree that maximises the sum of Kendall's Tau for each pair of nodes connected by an edge. The algorithm then selects the best family for each of the first tree's bivariate copula by minimising an information criterion such as AIC. Edges and copula families in subsequent trees are selected sequentially. In the real-time forecasting models used for electricity load demand, MCMC samples of the posterior distribution are available, though often immediately out of date. The approach used in this thesis is to run an MCMC algorithm, thin the resulting draws until they are effectively independent, choose an optimal family for each marginal distribution by AIC, then run Dißmann's algorithm to infer a vine copula structure that fits the data well. This vine copula is used to construct an approximating distribution $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ that can then be used in an SVB algorithm to update $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ as more observations become available. MCMC can be ran simultaneously and the resulting sample can be used to check if the distribution q should be changed.

2.3 Time Series Application

In this section the techniques described throughout Section 2 are applied to a simple, but widely used, time-series model. Consider the second order auto-regressive time series model, denoted by AR(2) and described by

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t \tag{2.27}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ for $t = 3, \ldots, T$. The first two observations, y_1 and y_2 are distributed according to

$$y_1, y_2, \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{2.28}$$

where

$$\begin{aligned}
\Sigma &= \begin{bmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{bmatrix} \\
\gamma(0) &= \sigma^2 \frac{1-\phi_2}{(1+\phi_2)((1-\phi_2)^2 - \phi_1^2)} \\
\gamma(1) &= \sigma^2 \frac{\phi_1}{(1+\phi_2)((1-\phi_2)^2 - \phi_1^2)}.
\end{aligned}$$

The likelihood of parameters $\boldsymbol{\theta} = (\phi_1, \phi_2, \sigma^2)'$, that is, the density of the observations $y_{1:T}$, given $\boldsymbol{\theta}$, viewed as a function of $\boldsymbol{\theta}$, is given by

$$\begin{aligned}
L(\theta|y_{1:T}) &= p(y_1, y_2|\theta) \prod_{t=3}^{T} p(y_t|y_{1:t-1}, \theta) \\
&= \frac{1}{(2\pi)} |\Sigma|^{-1/2} \sigma^{-(T-2)} \exp\left\{ \frac{-1}{2} (y_1, y_2) \Sigma^{-1} (y_1, y_2)' \right\} \\
&\times \exp\left\{ \frac{-1}{2\sigma^2} \left(\sum_{t=3}^{T} (y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2})^2 \right) \right\}.
\end{aligned} \tag{2.29}$$

We assume that the prior distribution, is given by

$$p(\sigma^2, \phi_1, \phi_2) \propto \sigma^{-2}\mathbb{I}(\phi_2 > -1)\mathbb{I}(\phi_2 < 1 + \phi_1)\mathbb{I}(\phi_2 < 1 - \phi_1), \tag{2.30}$$

where $\mathbb{I}$ is the indicator function that equals one if the condition in the brackets is true and zero otherwise. This prior has positive mass if and only if a pair (ϕ_1, ϕ_2) is within the AR(2) stationary region. Note that this prior is not conjugate due to the form of ϕ_1 and ϕ_2 in Σ, and thus MFVB cannot be applied. Upon examination of the form of the product of the likelihood function in (2.29) and the prior density in (2.30) it can be seen that the full conditional distribution $p(\sigma^2|y, \phi_1, \phi_2)$ is an Inverse Gamma distribution, while the conditionals for both ϕ parameters are intractable but can be sampled using a Metropolis-Hastings step with truncated bivariate normal random walk proposal distribution. The truncation ensures that draws of (ϕ_1, ϕ_2) are only taken from the AR(2) stationary region.

Data was simulated with $T = 150, \sigma^2 = 1, \phi_1 = 0.7$ and $\phi_2 = 0.2$, then the MCMC algorithm described above is applied. Marginal distributions are fit to each parameter in $\boldsymbol{\theta}$ by minimising AIC, and an application of Dißmann's algorithm to the draws found an optimal Vine Copula structure. The functional form of $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ to be used in SVB is:

- $q(\phi_1|\boldsymbol{\lambda}_1)$ - Gaussian marginal
- $q(\phi_2|\boldsymbol{\lambda}_2)$ - Gaussian marginal
- $q(\sigma^2|\boldsymbol{\lambda}_3)$ - Inverse Gamma marginal
- $c(Q(\phi_1), Q(\phi_2)|\boldsymbol{\lambda}_4)$ - Gaussian copula
- $c(Q(\phi_1), Q(\sigma^2))$ - Independent
- $c(Q(\phi_2), Q(\sigma^2)|Q(\phi_1))$ - Independent

Figure 3 displays the marginal and bivariate posterior distributions that result from each of the Stochastic Variational Bayes (blue) and the MCMC (red) approaches. In the case of MCMC, a kernel estimate obtained from the relevant posterior sample is displayed, whereas for SVB the form of the relevant vine copula marginal determines the shape of the marginal posterior. As is clear from the figure, SVB appears to have selected values for $\boldsymbol{\lambda}$ that closely match the MCMC posterior distribution. SVB required approximately one hundredth of the computation time required for MCMC.

3 Electricity Load Forecasts

3.1 Motivation

The literature for electricity forecasting for both load and price is wide, and contains models based on many different fields, such as game theory, time-series modelling and neural networks, see Weron (2014) for a recent review. Density forecasting for short term electricity

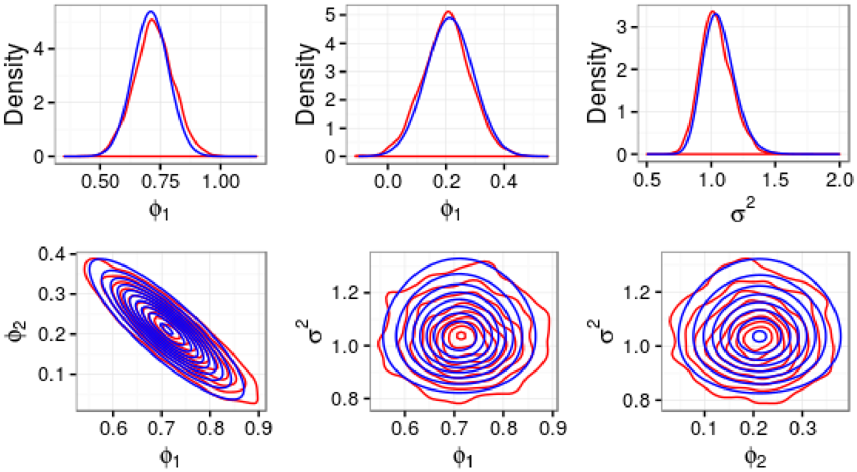

Figure 3: The fit of an SVB algorithm (blue) compared to MCMC (red) for the AR(2) model.

load has had less attention, with Fan and Hyndman (2012) and He et al. (2016) being notable examples, which used boot strapping and quantile regression respectively to generate forecast densities. Both of these approaches create forecast densities that are conditioned on the model and estimated parameters being correct, while the Bayesian approach used in this thesis can average over parameter and model uncertainty leading to a more accurate forecast density. This thesis also aims to re-estimate the model after each data point is observed to fully take advantage of any extra information on changes in parameters and thus further increase the accuracy of the forecast density.

Half hourly electricity load data for Victoria, Australia is provided by the Australian Energy Market Operator (AEMO) for the 2011-2015 period, and is plotted in Figure 4. The time series displays strong seasonal characteristics, with time of year, day of week, and time of day patterns. This yearly pattern is characterised by a low median load with high volatility in summer and a more consistent curve in winter. Electricity load is strongly dependent on temperature, with high load experienced on both very cold and very hot days. In summer periods, load is often low but shows extreme volatility, while load slowly rises then falls in winter and displays significantly less volatility.

The wholesale electricity market in Victoria operates by market participants sending electricity bid and offers for each five minute period to AEMO, which then matches electricity generators with retailers to determine the dispatch load and associated price. At the end of each half-hour block the previous six five-minute dispatch prices are averaged to obtain the final spot price that participants trade at. Market participants are allowed to revise bids and offers at any point before dispatch. The electricity supply curve is strongly hockey stick shaped, as renewables and coal generators are able to provide a large load cheaply but if demand exceeds this level the price spikes as other generators must be brought online expensively. These characteristics of the electricity market leads to a large amount of volatility in prices demonstrated by Figure 5. The seasonal effects of the load data are also present in prices, but these effects are dominated by irregular spikes as high as \$9,974/mWh, compared to a four year mean price of \$42/mWh.

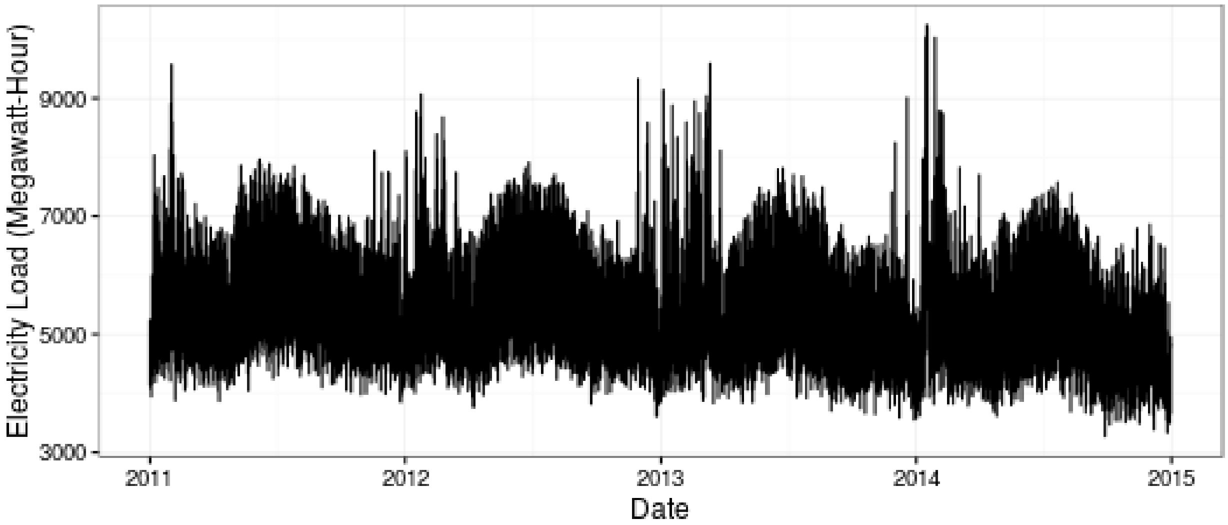

Figure 4: Half hourly load data measured in megawatt hours for Victoria from January 1 2011 to December 31 2015.

3.2 Exponential Smoothing

Taylor (2003) explores the seasonal properties of minute-by-minute electricity demand in the United Kingdom and introduces the double seasonal Holt-Winters exponential smoothing model with daily and weekly effects described by (3.1)-(3.4). Yearly seasonal effects are omitted as their inclusion requires a large number of latent states to be estimated, however Taylor (2008) later provides empirical support that this model is superior to other common time-series models such as SARIMA based models that include a yearly seasonal effect. The double seasonal Holt-Winters exponential smoothing model is described by

$$y_t = l_{t-1} + d_{t-m_1} + w_{t-m_2} + e_t \tag{3.1}$$
$$l_t = \alpha(y_t - d_{t-m_1} - w_{t-m_2}) + (1-\alpha)l_{t-1} \tag{3.2}$$
$$d_t = \delta(y_t - l_{t-1} - w_{t-m_2}) + (1-\delta)d_{t-m_1} \tag{3.3}$$
$$w_t = \omega(y_t - l_{t-1} - d_{t-m_1}) + (1-\omega)w_{t-m_2} \tag{3.4}$$

where m_1 and m_2 are the lengths of the daily and weekly cycle, and the smoothing parameters α, δ, ω are also restricted to to lie in $(0,1)^3$. This can be rewritten as a single source of error state-space model (Snyder, 1985),

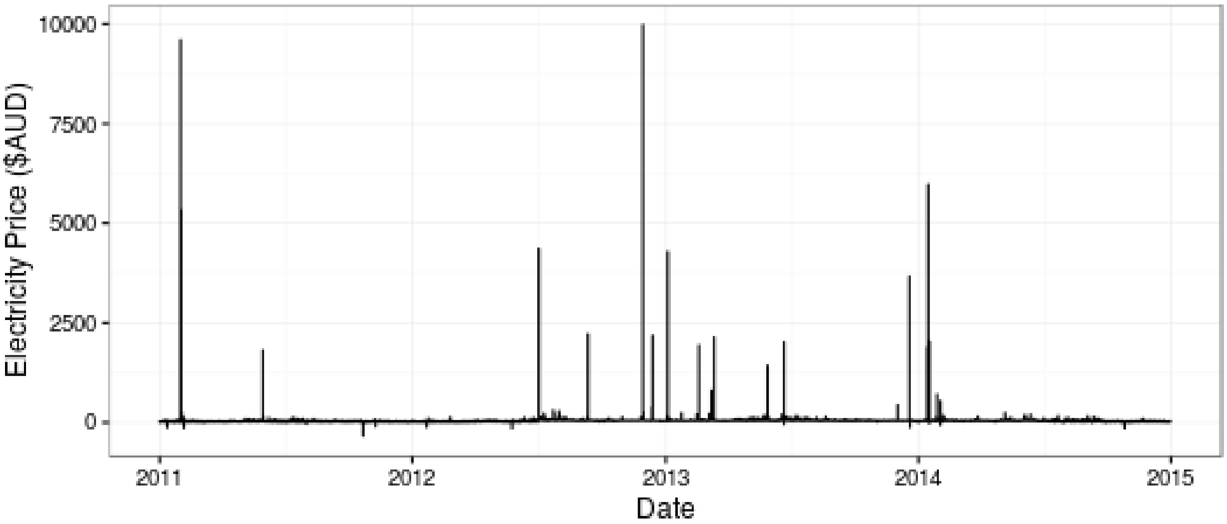

Figure 5: Half hourly price data for Victoria from January 1 2011 to December 31 2015.

$$y_t = l_{t-1} + d_{t-m_1} + w_{t-m_2} + e_t \tag{3.5}$$

$$l_t = l_{t-1} + \alpha e_t \tag{3.6}$$

$$d_t = d_{t-m_1} + \delta e_t \tag{3.7}$$

$$w_t = w_{t-m_2} + \omega e_t. \tag{3.8}$$

The unknown parameters are $\boldsymbol{\theta} = (\alpha, \delta, \omega)', \sigma^2$ and $\mathbf{b}_0 = (l_0, d_0, \ldots, d_{-(m_1-1)}, w_0, \ldots, w_{-(m_2-1)})'$ a $k = m_1 + m_2 + 1$ length vector of the initial states of the latent variables. The single source of error state-space model refers to the random error e_t being shared across the data each latent variable, and allows the likelihood to be expressed as a closed form function of the unknown parameters.

Forbes et al. (2000) provides a transformation to the model that allows a straight-forward Bayesian analysis in small samples, noting that the state space model can be expressed as

$$y_t = \mathbf{x}'\mathbf{b}_t + e_t \tag{3.9}$$

$$\mathbf{b}_t = T\mathbf{b}_{t-1} + \tilde{\boldsymbol{\theta}} e_{t-1}. \tag{3.10}$$

Using $D = T - \tilde{\boldsymbol{\theta}}\mathbf{x}'$, the transition equation in (3.10) can be written as

$$\mathbf{b}_t = D\mathbf{b}_{t-1} + \tilde{\boldsymbol{\theta}} y_t. \tag{3.11}$$

Setting $\bar{\mathbf{b}}_0 = 0$, the recursion in (3.11) allows the transformed observations $\tilde{y}_t$ to be obtained via

$$\tilde{y}_t = y_t - x'\bar{\mathbf{b}}_{t-1}. \tag{3.12}$$

Furthermore, transformed regression variables $\tilde{\mathbf{x}}_t$ can be obtained from $\tilde{\mathbf{x}}_t = D\tilde{\mathbf{x}}_{t-1}$ with $\tilde{\mathbf{x}}_0 = \mathbf{x}$. Collecting these as $\widetilde{X}' = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_T)$ and $\widetilde{Y}' = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_T)$, and assuming the errors e_t are Gaussian, the model likelihood is

$$L(\boldsymbol{\theta}, \sigma^2, \boldsymbol{b}_0 | y_{1:T}) \propto \sigma^{-T} \exp\left\{ \frac{-1}{2\sigma^2} (\widetilde{Y} - \widetilde{X}\boldsymbol{b}_0)'(\widetilde{Y} - \widetilde{X}\boldsymbol{b}_0) \right\}. \tag{3.13}$$

Hence once conditioned on a draw of $\boldsymbol{\theta}$, the remaining unknown parameters can be sampled as a Bayesian linear regression problem.

Forbes et al. (2000) provides the marginal distribution of $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta}|y_{1:T}) \propto \left|\widetilde{X}'\widetilde{X}\right|^{-1/2} \tilde{s}^{-(T-k)} p(\boldsymbol{\theta}), \tag{3.14}$$

where $\tilde{s}^2 = (\widetilde{Y} - \widetilde{X}\hat{\mathbf{b}}_0)'(\widetilde{Y} - \widetilde{X}\hat{\mathbf{b}}_0)/(T - (m_1 + m_2 + 1))$ is the sum of squared errors for $\hat{\mathbf{b}}_0 = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{Y}$. With the seasonality parameterisation in (3.5) - (3.8), the matrix $\widetilde{X}'\widetilde{X}$ is singular, but restricting the latent variables via

$$d_t = -\sum_{i=1}^{m_1-1} d_{t-i} \tag{3.15}$$

$$w_t = -\sum_{i=1}^{m_2-1} w_{t-i} \tag{3.16}$$

avoids this problem. The resulting model is described by

$$y_t = l_{t-1} - \sum_{i=1}^{m_1-1} d_{t-i} - \sum_{i=1}^{m_2-1} w_{t-i} + e_t \tag{3.17}$$

$$l_t = l_{t-1} + \alpha e_t \tag{3.18}$$

$$d_t = -\sum_{i=1}^{m_1-1} d_{t-i} + \delta e_t \tag{3.19}$$

$$w_t = -\sum_{i=1}^{m_2-1} w_{t-i} + \omega e_t. \tag{3.20}$$

and $\mathbf{b}_0 = (l_0, d_0, \ldots, d_{-(m_1-2)}, w, \ldots, w_{-(m_2-2)})'$ is a length $k = m_1 + m_2 - 1$ vector.

Forbes et al. (2000) recommends numerical integration of the marginal posterior density of $\boldsymbol{\theta}$ in (3.14), however the problems with using this method in our model are two-fold: repeating a three dimensional numerical integration each time a new data point is observed within the five minute window is not feasible, and the T in the exponent makes most evaluations of $p(\boldsymbol{\theta}|y_{1:T})$ computationally zero when T is large, as is the case with high frequency electricity load data. An efficient MCMC algorithm that scales to large datasets with models using multiple seasonality parameters will be required to sample the full posterior distribution $p(\boldsymbol{\theta}, \sigma^2, b_0|y_{1:T})$ and obtain an approximating distribution $q(\boldsymbol{\theta}, \sigma^2, b_0|\boldsymbol{\lambda})$ with a reasonable goodness of fit. From this point, Stochastic Variational Bayes will be find an approximation to the updated posterior distribution as more data is observed, and provide the predictive density $q(y_{T+J+1}|y_{T+J})$.

4 Discussion

One of the major aims of this thesis is to provide a methodology to forecast predictive densities with minimal calculation required to update parameter estimates as more data is observed, so an accurate forecast can be made even on a very short time-frame. However the short time-frame and high frequency of observating data implies that there is already a large number of past observations available, so any the marginal information from an additional observation on the parameters may be minimal.

The benefits of rapidly updating parameters will be increased in situations where there are major changes in the data generating process over time. One example is a state-space model where information on the next latent state is increased significantly after observing the most recent data point. However, unless the model has either a non-linearity or non-Gaussian error, the Kalman filter can be applied to obtain an accurate predictive density without difficult computation. Volatility forecasting for stock returns is a potentially useful application as data is observed on a short time-frame, and the stochastic volatility state-space models commonly used do not admit the use of the Kalman filter.

A contribution of this thesis is the use of information generated by MCMC to develop a Vine Copula used by the approximating distribution in Stochastic Variational Bayes, but updating state-space models by SVB requires an approximation for latent states associated with newly observed data that have not been included in the most recent MCMC sample. In this case, the dependence for the new latent state must be extrapolated from the copula structure used for previous latent states, which may be sub-optimal. The robustness of SVB approximations to choices of q that badly fit the true posterior should be investigated. Development of the theory of choosing the distribution q is currently extremely active, a methodology that relies on the form of the model likelihood and prior rather than MCMC samples to determine the form of a well-fitting approximation would be of interest.

Finally, there is a trade-off between the predictive accuracy of MCMC and the ability to update Variational Bayes posterior distributions. The statistical error associated with using an approximate distribution has not been investigated in the literature and we shall contribute by comparing this error in forecasting with VB to the statistical error associated with forecasting based on an MCMC generated posterior distribution that is not conditioned on the most recent data.

5 Timeline

June 2017	· Derive an efficient MCMC algorithm for exponential smoothing and apply SVB for electricity load forecasting. · Similarly apply SVB to competing models such as SARIMA. · Compare forecast accuracy to MCMC with and without updates to parameters.
December 2017	· Investigate state-space models, including time-varying parameters, non-linearities or non-Gaussian errors. · Apply this to Stochastic Volatility models. · Extend model with multivariate data and Markov Switching mechanisms. · Compare results to Particle Filter based methods.
June 2018	·Investigate impact of sub-optimal approximating distributions such as mis-specified copulas for new latent states. · Investigate other ways to construct the approximating distribution either with or without a posterior sample being available.
December 2018	· Allow extra time for updates in the active VB literature to develop, and adapt them to the benefit my research.
May 2019	· Put separate parts together and polish the final thesis for submission.

References

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

Attias, H. (1999), "A Variational Bayesian Framework for Graphical Models," in *Advances in Neural Information Processing Systems 12*, eds. Solla, S. A., Leen, T. K., and Müller, K., MIT Press, pp. 209–215.

Beal, M. J. (2003), "Variational Algorithms for Approximate Bayesian Inference," Ph.D. thesis, University College London.

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

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, to appear.

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

Dißmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013), "Selecting and Estimating Regular Vine Copulae and Application to Financial Returns," *Computional Statistics and Data Analysis*, 59, 52–69.

Duchi, J., Hazan, E., and Singer, Y. (2011), "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research*, 12, 2121–2159.

Fan, S. and Hyndman, R. J. (2012), "Short-term load forecasting based on a semi-parametric additive model," *IEEE Transactions on Power Systems*, 27, 134–141.

Forbes, C. S., Snyder, R. D., and Sharmi, R. G. (2000), "Bayesian Exponential Smoothing," Working paper.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburnder, J., and Penny, W. (2006), "Variational free energy and the Laplace approximation," *NeuroImage*, 34, 220–234.

Gahramani, Z. and Beal, M. J. (2001), "Propogation algorithms for Variational Bayesian Learning," in *Advances in Neural Information Processing Systems 13*, eds. Leen, T., K., Dietterich, T., and Tresp, V., MIT Press, pp. 507–513.

Geweke, J. and Whiteman, C. (2006), "Bayesian Forecasting," in *Handbook of Economic Forecasting, Volume 1*, eds. Elliott, G., Granger, C. W. J., and Timmermann, A., Elsevier, chap. 10.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995), "Adaptive Rejection Metropolis Sampling within Gibbs Sampling," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44, 445–472.

Gneiting, T. and Katzfuss, M. (2014), "Probabilistic Forecasting," *Annual Review of Statistics and Its Application*, 1, 125–151.

He, Y., Xu, Q., Wan, J., and Yang, S. (2016), "Short-term power load probability density forecasting based on quantile regression neural network and triangle kernel function," *Energy*, 114, 498–512.

Joe, H. (2014), *Dependence modeling with copulas*, CRC Press.

Jordan, M. I., Ghahramani, Z., Jaakola, T. S., and Saul, L. K. (1999), "An introduction to variational methods for graphical models," *Machine Learning*, 37, 183–233.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016), "Improved Variational Inference with Inverse Autoregressive Flow," in *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., pp. 4743–4751.

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.

Liu, Q. and Wang, D. (2016), “Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm,” in *Advances in Neural Information Processing Systems 29*, eds. Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., Curran Associates, Inc., pp. 2378–2386.

Minka, T. P. (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., Morgan Kaufmann Publishers Inc., pp. 362–369.

Ng, Yin, C., Chilinski, P., and Silva, R. (2016), “Scaling Factorial Hidden Markov Models: Stochastic Variational Inference without Messages,” in *Advances in Neural Information Processing Systems 29*, eds. Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., Curran Associates, Inc., pp. 4044–4052.

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*, eds. Langford, J. and Pineau, J., Omnipress, pp. 1367–1374.

Ranganath, R., Altosaar, J., Tran, D., and Blei, David, M. (2016a), “Operator Variational Inference,” in *Advances in Neural Information Processing Systems 29*, eds. Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., Curran Associates, Inc., pp. 496–504.

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.

Ranganath, R., Tran, D., and Blei, D. M. (2016b), “Heirarchical Variational Models,” in *Proceedings of the 33rd International Conference on Machine Learning*, eds. Balcan, M. and Weinberger, Kilian, Q., JMLR W&CP.

Rezende, D. J. and Mohamed, S. (2015), “Variational Inference with Normalizing Flows,” in *Proceedings of the 32nd International Conference on Machine Learning*, eds. Bach, F. and Blei, D., JMLR W&CP.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014), “Stochastic Backpropagation and Approximate Inference in Deep Generative Models,” in *Proceedings of the 31st International Conference on Machine Learning*, eds. Xing, E. P. and Jebara, T., JMLR W&CP.

Robbins, H. and Monro, S. (1951), “A Stochastic Approximation Method,” *The Annals of Mathematical Statistics*, 22, 400.

Sklar, A. (1959), “Fonctions de rpartition n dimensions et leurs marges,” *Publ. Inst. Statist. Univ. Paris*, 8, 229–231.

Snyder, R. D. (1985), "Recursive Estimation of Dynamic Linear Models," *Journal of the Royal Statisitcal Society, Series B*, 47, 272–276.

Taylor, J. W. (2003), "Short-Term Electricity Demand Forecasting Using Double Seasonal Exponential Smoothing," *The Journal of the Operational Research Society*, 54, 799–805.

— (2008), "An evaluation of methods for very short-term load forecasting using minute-by-minute British data," *International Journal of Forecasting*, 24, 645–658.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Annals of Statistics*, 22, 1701–1728.

Tran, D., Blei, D. M., and Airoldi, E. M. (2015), "Copula variational inference," in *Advances in Neural Information Processing Systems 28*, eds. Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., Curran Associates, Inc., pp. 3564–3572.

Tran, D., Ranganath, R., and Blei, D. M. (2016), "The Variational Gaussian Process," in *Proceedings of the International Conference on Learning Representations 2016.*

Wainwright, M. J. and Jordan, M. I. (2008), "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends in Machine Learning*, 1, 1–305.

Wang, C. and Blei, D. M. (2013), "Variational Inference in Nonconjugate Models," *Journal of Machine Learning Research*, 14, 1005–1031.

Weron, R. (2014), "Electricity price forecasting: A review of the start-of-the-art with a look into the future," *International Journal of Forecasting*, 30, 1030–1081.