Robust Bayesian Forecasting With State-Space Model

Ziyi Xiong, Xingyi Sun, Steven Xu

Abstract

Gaussian distribution has always been the go-to model for White Noise(WN) Process in the analysis of time series. In practice however, assumption of normality might be too conservative to capture any abrupt change of the latent signal because of its rather light-tailed characteristic. In a Bayesian setting, a relaxed model on WN is also beneficial because of its robustness against any potential adverse effect of outliers. The Exponential Power distribution is a family of symmetric density with flexible tail characteristic that is sometimes used as a robust prior. In this paper we explore the possibility of using it on the likelihood level and compared its performance with Gaussian likelihood with simulated and real data. A state-space model is fitted to forecast the 1 year ahead residential water consumption in London with the practical importance of aiding planning and management of water resources.

Keywords:
Exponential Power distribution, state-space model, Bayesian forecasting, latent process

The Honor Pledge: We have neither given nor received unauthorized aid on this assignment

Contents

1. Introduction

1.1. Background

Time series, gaining its name for the naturally ordered-by-time characteristic, has always been a widely researched topic deal to its abundant occurrence in disciplines such as econometrics, meteorology, engineering etc. Common interests of time series data include analysis of its trend over time, seasonality and the inherited autocorrelation. Like other types of data, the possibility of a time series' historical trace providing information towards its future path is of most interest for statisticians and researchers in relative areas mentioned above.

The prediction of a time series' future trend, or more formally called forecasting, has provided rich information for reasonable decision-making in almost every area. Most commonly, predictions of a response in some future time points are generally conditional expectations based on observed data, but the available information offered by a point forecast is limited as we often want to have some knowledge on the range or extreme values of the future response.

Within the time series settings, Bayesian treatment has become straightforward with the help of *state-space* form and *structural* representation. The Bayesian method offers several advantages over traditional methods. The best thing is that objectives such as fitting, imputation and forecasting can be all done in same time with the help of Gibbs sampling MCMC. Also, the availability of predictive posterior distribution allow us to perform probabilistic forecast naturally.

1.2. Relaxation of WN assumption

Fortunately data scientists have developed many R packages which can well solved these time series forecasting problems. However, most of them are based on the models which assume a Gaussian error distribution, which is the simplest case in time series analysis. Although Gaussian error distribution performs well in ideal case where assumptions of normality is met, it remains unknown whether this is the best model for all cases, especially when there might be non-linear trend in the transition of historical time points to future time points.

Instead of using Gaussian error distribution, we would like to explore the possibility of using a generalized normal distribution for the random error.

This symmetric family contains a variety of either heavy-tailed and light-tailed distribution with Gaussian distribution being a special case. In our project, we would like to adopt this more flexible model and compare the results with the outputs from based on Gaussian assumptions in terms of well known metrics in forecasting by applying them to the London Water dataset. If indeed this model performs better more generally, the future hope is to implement it into existing packages.

The rest of the paper is structured as follow: we briefly introduced the London Water dataset in Section 2 with proposed objectives. In Section 3 we proposed a *state-space* to model the given time series with discussion on possible distribution assumption for WN. Section 4 focused on a detailed explanation of the characteristic of Exponential Power density and possible generation methods. In Section 5, Bayesian approach is taken and the model is represented using hiearchical forms. Section 6 gives numerical results and finally, a conclusion as well as possible future extension are given in Section 7.

2. London Water Data

The water consumption records for every month is published on Time Series Data Library, which is created by Rob Hyndman, Professor of Statistics at Monash University, Australia. The database is provided by Hipel and McLeod (1994), which contains monthly water usage (ml/day) in London Ontario from year 1966 to 1988. This is a univariate time series with time stamp as the only covariate.

As shown in Figure 1, the two patterns one could easily point out by first look are the overall increasing trend and the strong yearly seasonality. Within each year, it seems like the water consumption is peaked at every summer and low at spring and winter. The peaks are of similar height except for the last two years, where water usage in summer is tremendously larger than all previous years. Without knowing future water usage (later than 1988), we probably need to transform the series for better fitting and forecasting but also admitting that the seasonal effect is evolving through time. The book written by Hipel and Mcleod (1994) provided a detailed traditional time series analysis of the data. Our objective is to perform a probabilistic imputation based on the idea of forecasting on one year water consumption using historical data. Since the original data does not contain missing value, we subjectively picked an interval and treat it as missing. In particular, we

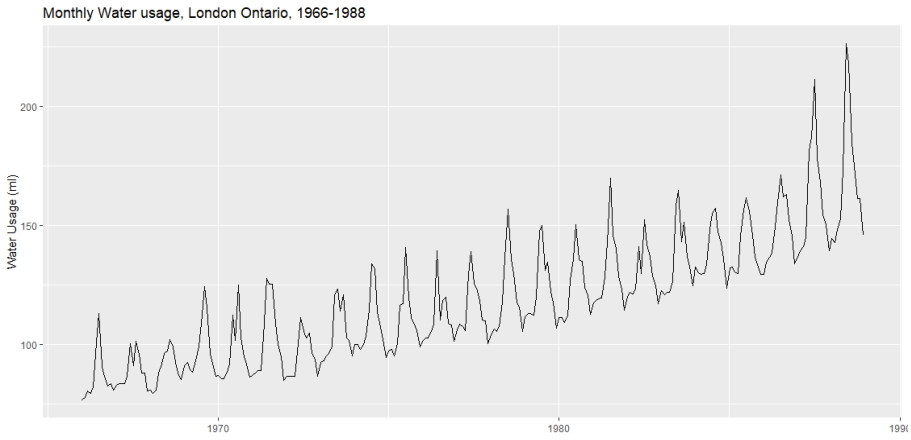

Figure 1: London Water Usage from 1966 to 1988

picked one interval before year 1987 as "easy" task and one interval on year 1987 as "hard" task because the clear difference of peak value. We believe that Gaussian and EP should perform similar on the "easy" task while EP potentially better on the "hard" task.

3. Proposed Model

3.1. State-Space Model

A time series model is said to be in *state-space* form if it can be represented by an *observation* equation and a *transition* equation.

$$y_t = \mathbf{Z}_t^T \alpha_t + \epsilon_t \quad \epsilon_t \sim N(0, \mathbf{H}_t) \tag{1}$$

$$\alpha_{t+1} = \mathbf{T}_t \alpha_t + \mathbf{R}_t^T \eta_t \quad \eta_t \sim N(0, \mathbf{Q}_t) \tag{2}$$

Here $\mathbf{Z}_t$, $\mathbf{H}_t$, $\mathbf{T}_t$, $\mathbf{R}_t$ are matrices consist of scalars and parameters. The *observation* equation (1) describes the relation between the observed value y_t and the latent states α_t while the *transition* equation (2) describes how the latent states evolve through time. The *state-space* form provides convenient representation to almost all of the commonly used time series models, for example ARIMA. It also gave rise to the application of smoothing methods such as Kalman filter and a direct relationship with Bayesian analysis. For more detail, Durbin and Koopman (2001) provided a comprehensive treatment of the state-space approach to time series analysis. A structural time

series is a more specific representation of the *state-space* form as we usually think that the time series can be explained by an aggregate effect of different components, for example trend, seasonality, regression predictors, etc. Scott and Varian (2013) used a basic structural time series with regression component to forecast Google trend data. In our paper, we considered a similar model without the regression component and some variation on the transition as well.

$$y_t = \mu_t + \tau_t + \epsilon_t \tag{3}$$
$$\mu_t = \mu_{t-1} + \delta_{t-1} + u_t \tag{4}$$
$$\delta_t = \delta_{t-1} + v_t \tag{5}$$
$$\tau_t = \tau_S - \theta\epsilon_{\tau_{t-12}} + w_t \tag{6}$$

Here $\mu_t,\ \tau_t$ denotes the stochastic trend effect and seasonality effect at time t. The *observation* equation (5) says that the observed y_t is determined by the underlying trend and seasonality with some measurement noise. The δ_t denotes the stochastic slope of the trend, so that increasing/decreasing trend at previous times can be taken into account. The seasonality effects are parametrized by a set of S dummy variables, and in order for them to be identifiable and stochastic, we constrain that the expectation of the summed effect over a full cycle S is 0. In Scott and Varian's paper the *transition* equation was $\tau_t = \tau_S - \theta\epsilon_{\tau_{t-12}} + w_t$, which can be seen as a constrained random walk. We modified it so that each season has a base effect and can stochastically change based on one cycle ahead moving average. The *transition* equations (4) and (5) are the sole component of a *local linear trend* model. Notice that if we set u_t to be zero, y_t will have a *integrated random walk* trend, or so called *smooth trend* (Harvey 2010). Also for this model we are assuming that the components have a additive effect. In the case where the components might have a multiplicative effect i.e. $y_t = \mu_t\tau_t$, one can simply take the logarithm of y_t and therefore have a additive model on $log(y_t)$.

Within the domain of time series, the random error term $\epsilon_t,\ u_t,\ v_t,\ w_t$ are assumed to come from a WN. In general, the white noise process is defined to have mean zero, constant variance and no correlation over time. For analyzing and computation purpose, people usually assume it to follow a zero mean Gaussian distribution at each time point. The reason of choosing

Gaussian family is it works well with the assumption of generalized linear transformation of signal from one time point to the next, so that the observations through out time can be jointly view as samples from a multivariate normal distribution. However, in almost every statistical related field, the assumption of Gaussian comes with some limitation, as it constrains the likelihood of observing outliers because of its light-tail characteristic. In time series this would mean that it is unlikely to observe abrupt change from y_t to y_{t+1}, which is clearly not guaranteed at least in some cases. Therefore, relaxation is necessary on the attached distribution of WN. In a Bayesian setting, a relaxed model on WN is also beneficial because of its robustness with the hope that it might capture the latent states more accurately when heavy-tailed noises exist. A common alternative of Gaussian distribution is the Student-t distribution, which has already been used in problems such as random effects model (Choy & Smith 1997). Another less well-known family is the Exponential Power distribution or Generalized Gaussian distribution. In the next section we will give a brief introduction and explains why this distribution might be useful in modelling WN process.

4. Exponential Power Distribution

The Exponential Power distribution can be parametrized in many forms, in our paper we adopt a particular form suggested by Choy and Chan (2008) because it is relevant to the simulation technique that we will use when performing Gibbs sampler.

$$EP(x|\mu,\sigma,\beta) = \frac{c_1}{\sigma} exp\{-|c_0^{1/2}\sigma^{-1}(x-\mu)|^{2/\beta}\},\ -\infty < x < \infty \qquad (7)$$

where

$$c_0 = \frac{\Gamma(3\beta/2)}{\Gamma(\beta/2)},\quad c_1 = \frac{c_0^{1/2}}{\beta\Gamma(\beta/2)}$$

The expectation and variance are

$$E(X) = \mu$$
$$Var(X) = A(\sigma,\beta)$$

Here $A(\sigma,\beta)$ is a function involving the two parameters, therefore if we fix β the variance depends solely on σ.

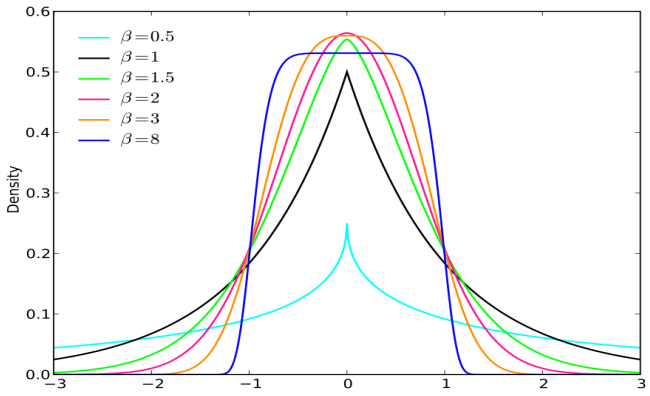

Figure 2: Generalized Gaussian Distribution

It is easy to identify that μ is a location parameter and σ is a scale parameter, which is similar as that of Gaussian. This particular feature of the EP density conveniently ensures that any linear transformation of a EP random variable will also follow an EP distribution. The flexibility of Exponential Power distribution is offered by the extra shape parameter $\beta \in (0, 2]$. In particular by fixing β the Exponential Power distribution will turn into some familiar distribution. When $\beta = 2$ we get the Laplace (Double Exponential) distribution. When $\beta = 1$ we have the Gaussian distribution. In general, the shape parameter controls the tail characteristic. When $0 < \beta < 1$ the distribution is more platykurtic than Gaussian and when $1 < \beta < 2$ it is more leptokurtic than Gaussian. Thus, the Exponential Power density can be used as a Robust likelihood by setting β to be greater than 1.

The biggest obstacle of using Exponential Power distribution in Bayesian treatment is the computation part. In order to perform meaningful analysis one needs to find methods of generating posterior based on a Exponential Power likelihood. There have been a lot of proposed methods on this part. Examples are representing Exponential Power family as a scale mixture of normal (Choy et al. 1997), scale mixture of uniform (Walker et al. 1999), etc. The first one is particularly difficult because it involves generating random samples from a Stable distribution which is itself computational burdensome, also only a range of β can be considered for reliable sampling. The second one represents EP distribution as a mixture of uniform and gamma and provide reliable simulation using only two steps. Also under the uniform mixture all

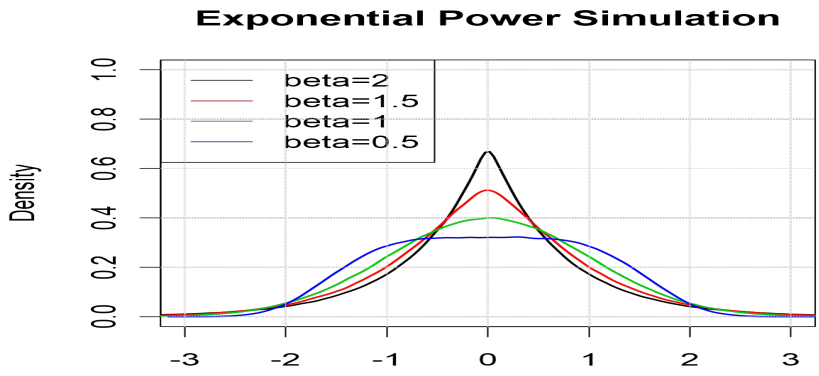

Figure 3: EP Simulation with SMU

values of β can be considered.

$$EP(x|\mu,\sigma,\beta) = \int_0^\infty U(x|\theta - \frac{\sigma}{\sqrt{2c_0}}u^{\beta/2}, \theta + \frac{\sigma}{\sqrt{2c_0}}u^{\beta/2,})G(u|1+\frac{\beta}{2}, 2^{-1/\beta}\ du) \tag{8}$$

$U(\cdot|a,b)$ and $G(\cdot|\alpha,\beta)$ in (8) represents uniform density with support a, b and gamma density with shape α, scale β respectively. This mixture form gives rise to the following two stage sampling for EP random variable x.

Algorithm 1 EP random variable generation

For given β and σ
Calculate $c_0 = \frac{\Gamma(3\beta/2)}{\beta/2}$
loop:
For i in 1:N **do**
Simulate $u_i \sim Ga(\alpha,\beta)$, where $\alpha = 1+\frac{\beta}{2}$ and $\beta = 2^{-1/\beta}$
Simulate $X \sim U(a,b)$, where $a = \mu - \frac{\sigma}{\sqrt{2c_0}}u^{\beta/2}$ and $b = \mu + \frac{\sigma}{\sqrt{2c_0}}u^{\beta/2}$
$(x1, x2, ..., x_N)$ is a sample of size N from $EP(\mu,\beta,\sigma)$ distribution.

5. Bayesian Hierarchy

Gaussian error was used in observation equation under the assumption that observed state should be close to latent state. Generalized Gaussian

error was used in transition equation for trend to allow abrupt change from the previous state to current state. Standard normal prior was given to the MA coefficients. Flat priors were given to the parameters as well as initial values for the latent states. In particular, initial values for τ are standardized to ensure their expectation sum to 0 for for a cycle of S for identifiability. The advantage of Bayesian treatment is that the shape parameter β can be given a prior, so that we could monitor the posterior distribution of it to check how far away our data is from normality assumption (Portela et al. 2004). However, this could cause potential issues when using a uniform mixture representation as we might not get the ideally truncated sample of β. Therefore in this paper we treat β as fixed and perform a exploratory analysis by varying its value.

$$
\begin{aligned}
y_t &\sim N(\mu_t + \tau_t, \frac{1}{\sigma_\epsilon^2}) \\
\mu_t &\sim EP(\mu_{t-1} + \delta_{t-1}, \sigma_\mu, \beta_\mu) \\
\delta_t &\sim N(\delta_{t-1}, \sigma_\delta) \\
\tau_t &\sim N(\tau_S - \theta\epsilon_{\tau_{t-12}}, \sigma_\tau)
\end{aligned}
$$

We used the scale mixture uniform representation to generate random samples from EP density with a two stage hierarchy, and all the priors generated as follows.

$$
\begin{aligned}
&y_t \sim N(\mu_t + \tau_t, \frac{1}{\sigma_\epsilon^2}),\\
&\mu_t \sim U(\mu_{t-1} + \delta_{t-1} \pm \frac{\sigma_\mu}{\sqrt{2c_0}} u_{1,t}^{\frac{\beta_\mu}{2}}),\ u_{1,t} \sim Ga(1 + \frac{\beta_\mu}{2}, 2^{-\frac{1}{\beta_\mu}})\\
&\delta_t \sim N(\delta_{t-1}, \sigma_\delta)\\
&\tau_t \sim N(\tau_S - \theta\epsilon_{\tau_{t-12}}, \frac{1}{\sigma_\tau})\\
&\sigma_\epsilon^2 \sim IG(0.01, 0.01),\ \sigma_\mu \sim IG(0.01, 0.01)\\
&\sigma_\delta \sim IG(0.01, 0.01),\ \sigma_\tau \sim IG(0.01, 0.01)\\
&\mu_0 \sim N(0, 0.01),\ \delta_0 \sim N(0, 0.01), \theta \sim N(0, 1)\\
&\tau_i \overset{i.i.d}{\sim} N(0, 0.01),\ i = 1, ..., S-1,\ \tau_S = \sum_{i=1}^{S-1} \tau_i
\end{aligned}
$$

6. Numerical Results

For the "easy" task, we used data from 1966 to 1985, and 1987 to generate posterior samples of all the latent state and the forecast value with a Markov Chain Monte Carlo (MCMC) method using Gibbs sampler of year 1986. The easy task was imputed really well because the peak value was relatively the same as the previous years. For the "hard" task, we used data from 1966 to 1986, and 1988 to generate posterior samples of all the latent state and the forecast value with a Markov Chain Monte Carlo (MCMC) method using Gibbs sampler of year 1987. The "hard" task was not imputed very well because there is an abrupt change of peak value, possibly due to multiplicative seasonal effect.

6.1. Posterior forecasting

For all the figures shown, the posterior median and 11 posterior quantiles (0.01, 0.1, 0.2, ... , 0.9, 0.99) are plotted. The observed y is plotted in a red line in 6 and 7. The trace plots and acf plots were not shown but convergence will be discussed in below.

Seasonality: The posterior of τ_t (Figure 4) for the "easy" task describes a constant seasonal effect over years, which is desirable because the seasonal effects of years before 1987 are not differed that much according to Figure 1.

Trend: The posterior of trend (Figure 5) is increasing over time which matches our prior knowledge. Predictive intervals for observed states are tight while loose for the missing year, which means that the fitting was really good but some improvement can still be brought to the forecasting stage.

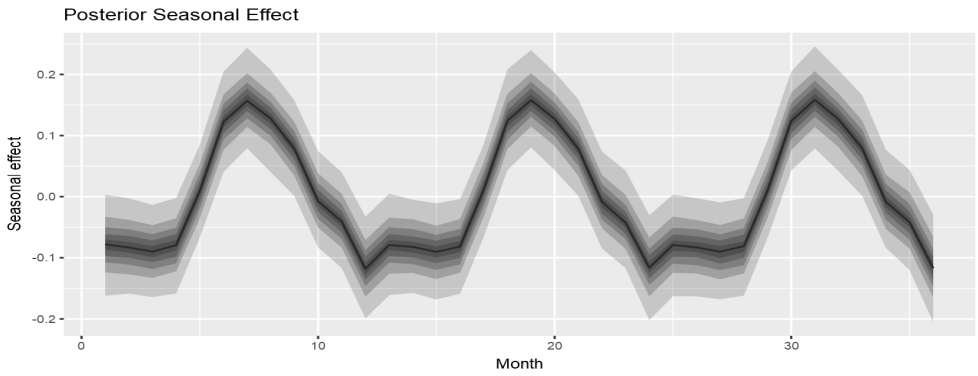

Figure 4: Posterior Seasonal Effects

Figure 5: Posterior Trend Effects

6.2. Forecasting and Comparison

From Figures 6 and 7, we could see that these two models give similar forecasting results. An careful observation tells that the overall range of predictions provided by Exponential Power model is narrower which is satisfactory. In Figures 8 and 9, the two models again show comparable results with EP density having narrower interval. This suggests that our *state-space* model might need improvement to more accurately capture the increasing seasonal effect at years 1987 and 1988.

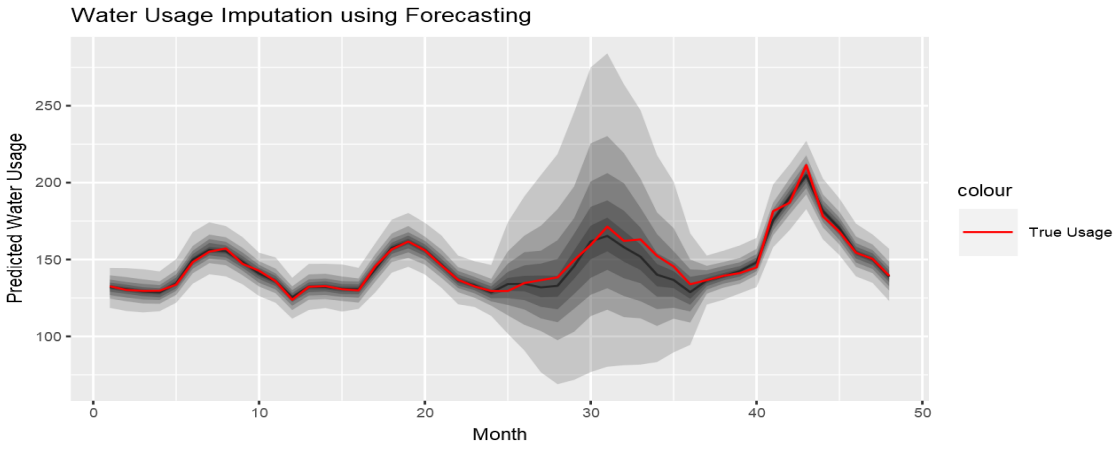

Figure 6: "Easy" Imputation from Generalized Gaussian error model

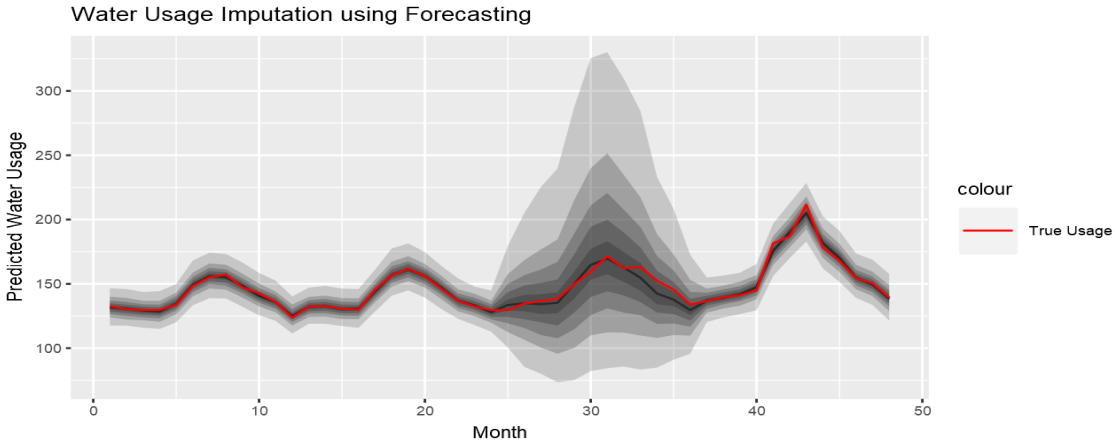

Figure 7: "Easy " Imputation from Gaussian error model

7. Discussion

In our proposed model, we relax the most common assumption of normality on WN and used an exponential power distribution model. We overcome the biggest challenge on computation involved in using Exponential Power distribution by adopting the scale mixture of uniform representation to perform Gibbs sampler. The implementation result of our Bayesian forecasting method with *state-space* model shows the advantage of easy construction and estimation. And it also suggests its potential use in practice, such as dynamic prediction for financial market, longitudinal data analysis, signal processing

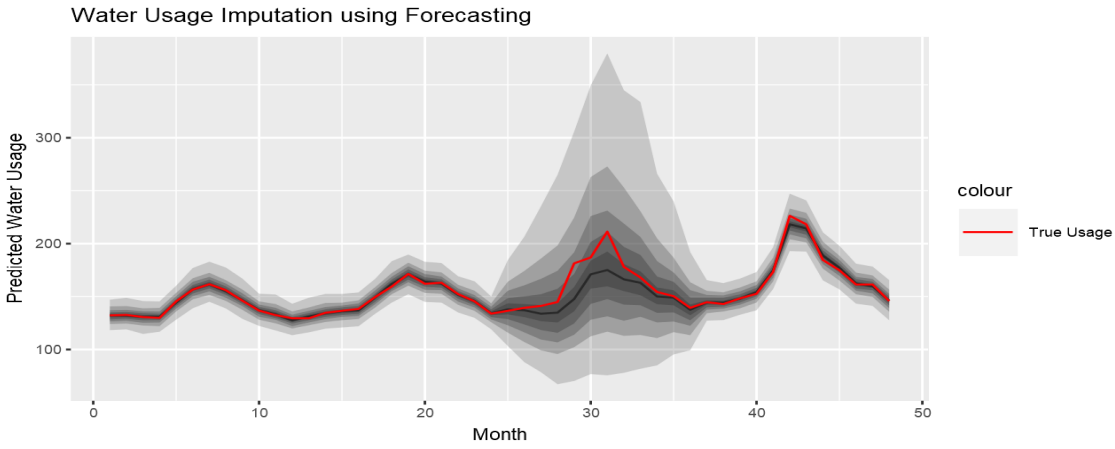

Figure 8: "Hard" Imputation from Generalized Gaussian error model

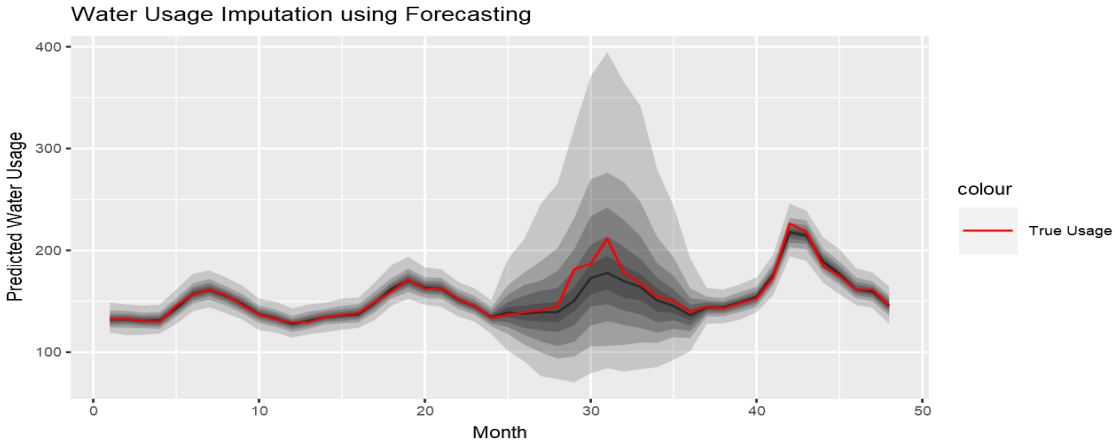

Figure 9: "Hard" Imputation from Gaussian error model

and so on.

All frameworks that account for imputation are not without limitations, and our model is no exception. From the forecasting results, the present model does not indicate a significant improvement over the model with normal random error. Moreover while it does not possess significant mixing issues regarding the MCMC Gibbs sampling, more complex data with multiple missing covariates associated with a single observation would introduce additional dependency to the probabilistic framework of the model and may prevent efficient convergence towards the stationary distribution. For large

data-sets, this has potential to significantly increase computational time and reduce efficiency.

Future research can be undertaken to address these limitations and provide additional insight into the present model. One potential improvement of interest is to obtain the full posterior and develop a more efficient method to implement the model instead of relying on the algorithms based on JAGS. Besides, an immediate extension to the present Bayesian State Space model can be considered in presence of more covariates. One could also use Bayesian model averaging technique to weight forecasts from different models for better performance.

8. Appendices

Choy, S., Chan, C. (2003). Scale Mixtures Distributions in Insurance Applications. ASTIN Bulletin, 33(1), 93-104.

Durbin, J. and Koopman, S. J. (2001). Time Series Analysis by State Space Methods. Oxford University Press.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Appl. Statist. 41, 337-348.

Harvey, A.C. (2009). The local quadratic trend model. Journal of Forecasting, 29, 94-108

Javier Portela M. A. Gómez-Villegas. (2004). Implementation of a robust bayesian method, Journal of Statistical Computation and Simulation, 74:4, 235-248

McCausland, W. J., Miller, S., and Pelletier, D. (2011). Simulation smoothing for state-space models: A computational e.ciency analysis. Computational Statistics and Data Analysis 55, 199-212.

S. Kalke, W.-D. Richter. (2013). Simulation of the p-generalized Gaussian distribution. Journal of Statistical Computation and Simulation 83:4, 641-667.

S. T. B. Choy, A. F. M. Smith. (1997). Hierarchical models with scale mixtures of normal distributions. Test, 1997, Volume 6, Number 1, 205.

Scott, S.L., & Varian, H.R. (2014). Predicting the present with Bayesian structural time series. IJMNO, 5, 4-23.