Nowcasting by the BSTS-U-MIDAS Model

by

Jun Duan
B.A., East China Normal University, 1998
M.A., East China Normal University, 2001

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF ARTS

in the Department of Economics

Nowcasting by the BSTS-U-MIDAS Model

by

Jun Duan
B.A., East China Normal University, 1998
M.A., East China Normal University, 2001

Supervisory Committee

Dr. David E.A. Giles, Supervisor
(Department of Economics)

Dr. Kenneth G. Stewart, Departmental Member
(Department of Economics)

Supervisory Committee

Dr. David E.A. Giles, Supervisor
(Department of Economics)

Dr. Kenneth G. Stewart, Departmental Member
(Department of Economics)

ABSTRACT

Using high frequency data for forecasting or nowcasting, we have to deal with three major problems: the mixed frequency problem, the high dimensionality (fat regression, parameter proliferation) problem, and the unbalanced data problem (missing observations, ragged edge data). We propose a BSTS-U-MIDAS model (Bayesian Structural Time Series-Unlimited-Mixed-Data Sampling model) to handle these problem. This model consists of four parts. First of all, a structural time series with regressors model (STM) is used to capture the dynamics of target variable, and the regressors are chosen to boost the forecast accuracy. Second, a MIDAS model is adopted to handle the mixed frequency of the regressors in the STM. Third, spike-and-slab regression is used to implement variable selection. Fourth, Bayesian model averaging (BMA) is used for nowcasting. We use this model to nowcast quarterly GDP for Canada, and find that this model outperform benchmark models: ARIMA model and Boosting model, in terms of MAE (mean absolute error) and MAPE (mean absolute percentage error).

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

My parents for supporting me.

Supervisor Dr. David Giles for his encouragement, patience, and advice.

Dr. Kenneth Stewart and Dr. Farouk Nathoo for their support and advice.

My wife Sheri Liu for helping me.

The department of Economics for funding me and my graduate research with a Graduate Scholarship.

> *Have the courage to use your own reason- That is the motto of enlightenment.*
> Immanuel Kant

DEDICATION HIN

This thesis is dedicated to my parents.

Chapter 1

Introduction

Utilizing the abundant high frequency data such as financial daily data is of great interest in forecasting or nowcasting low frequency macroeconomic variables such as (quarterly) GDP. The goal of this thesis is to provide a comprehensive method for short-term forecasting using mixed-frequency data sets. Our BSTS-U-MIDAS (Bayesian Structural Time Series - Unrestricted - Mixed-Frequency Data Sampling) Model incorporates four advanced econometric techniques.

Those four methods are the mixed-data sampling (MIDAS) model dealing with mixed frequency data; the Kalman filter (state-space model) for filtering the data; a spike-and-slab regression for variable selection, and Bayesian model averaging (BMA) for forecasting. The BSTS approach estimates the states and parameters of a model by using Markov Chain Monte Carlo (MCMC) simulation. Our BSTS-U-MIDAS Model outperforms the ARIMA and Boosting models in terms of MAE (mean absolute error) and MAPE (mean absolute percentage error) in an empirical application of forecasting GDP for Canada.

The BSTS-U-MIDAS model shows strong capability to capture the structural breaks in the economy and an ability to extract signals from high dimensional data sets with high frequency data. The BSTS-U-MIDAS has those advantages because it is flexible in terms of capturing the dynamics of stochastic processes and incorporating leading variables. Firstly, structural time series model decomposes a time series into several stochastic processes such as trend, seasonality, cycle, and irregular component. Those stochastic is helpful to avoid the influence of structural breaks. Secondly, a MIDAS model is able to transfer high-frequency leading variables to a low-frequency data matrix, therefore enhancing the ability to detect structural breaks. Furthermore, the BSTS-U-MIDAS model also show strong robustness to over-fitting even though it also suffers slightly from noisy data. The robustness of BSTS-U-MIDAS comes from a spike-and-slab regression and Bayesian model average. Firstly, a spike-and-slab prior helps to select important variables and maintain the sparsity of the model. Secondly, BMA helps to reduce the model uncertainty and instability.

With the advance of statistical technique and computational power, this kind of practice is much more easily implemented than before. The mixed-data sampling (MIDAS) model is now a widely used technique in economic forecasting (Ghysels, Sinko, & Valkanov, 2007). However, incorporating high frequency data causes some problems such as parameters proliferation and ragged edge data.

Parameters proliferation ("curse of dimensionality",or, fat regression) refers to the case where N is greater than T, where N is the number of regressors, and T is the number of observations on the time series. In this situation, the number of regressors is such large that there is not enough data available for the computation in terms of degrees of freedom. Many solutions have been developed to reduce the dimension such as Bridge models, MIDAS with weighting schema (Ghysels, 2012), factor MIDAS(Stock & Watson, 2006), PCA (Boivin & Ng, 2006), ridge regression, and lasso regression(De Mol, Giannone, & Reichlin, 2008) . We incorporate a Bayesian structural time series model (BSTS) into a MIDAS model to solve the "curse of dimensionality" problem. In BSTS model, a spike-and-slab regression is used to do variable selection and Bayesian model averaging is used to combine possible models for improving forecasts.

Ragged edge data means that different data have different publication lags and therefore different availability. At a specific time point, differences in the availability of data make the data set unbalanced, which makes forecasting more difficult. For example, in a specific day, the quarterly, monthly, and weekly data are not available yet although some daily data such as a stock price index are already available. In practice, realigning the data by filling the missing data is common. The missing data are usually estimated by an AR process or the Kalman filter (Foroni & Marcellino, 2013). Alternatively, in a MIDAS model, available predictors are chosen in terms of the forecasting horizon. Available data, lags or leads, are included in the regression directly. Incorporating MIDAS into a BSTS model makes the model more flexible in its handling of the ragged edge data.

Scott and Varian (2014a) propose a Bayesian structural time series model(BSTS) for nowcasting an economic time series variable by using Google search engine query data as a predictor. We combine their BSTS and U-MIDAS model for nowcasting low frequency macroeconomic variable using high frequency data such as financial data.

Incorporating the BSTS approach in a MIDAS model has many good features. BSTS-U-MIDAS is flexible. The BSTS component can handle regular and irregular data using a state-space form. The MIDAS components can address the unbalanced data set problem by transforming data in different frequencies. BSTS-U-MIDAS is robust. With the state-space form and the Kalman filter, we can relax the assumption about stationarity and normality. With the MIDAS setup, the model is more robust to specification errors than is a full system model (Bai, Ghysels, & Wright, 2013) . BSTS-U-MIDAS improves forecast accuracy. The Kalman filter can remove noise and extract signals, BMA ensembles estimations from many small regressions to overcome over-fitting and model instability, and the use of spike-and-slab prior picks up those regressors with high influence on the target variable to achieve model sparsity. In the Bayesian framework, BSTS-U-MIDAS provides not only point estimates but also density predictions.

Our empirical application of the BSTS-U-MIDAS model to forecast quarterly GDP for Canada shows an improvement in accuracy in terms of mean absolute error (MAE) than a benchmark ARIMA model and a Boosting model.

Our BSTS-U-MIDAS model can be used in many fields. In the macroeconomics variable forecasting, many high frequency financial data such as daily interest rates, or weekly or monthly labor market data can be utilized by using a BSTS-U-MIDAS model.

This thesis is organized by follows. Chapter 2 provides a literature review. The model and theory are discussed in Chapter 3. Chapter 4 focuses on the implementation of the model in an empirical application of forecasting quarterly GDP using monthly unemployment rate, monthly interest rate spread, monthly housing starts, the daily oil price and the daily S&P/TSX Composite Index. Our conclusions and some suggestions for further study, are given in Chapter 5.

Chapter 2

Literature Review

2.1 Research Question

The availability of abundant data in some areas of economics provides an opportunity for practitioners to include more relevant data into models to improve forecasting. As long as new data and the target variable have co-movement, including new data should help to improve the forecast accuracy. However, there is a trade-off between bias and variance of fitted model. Including more data into model might reduce the bias, but increase the variance. How many predictors should we put into a model and what form should these predictors take?

In macroeconomic variable forecasting, mixed frequency data are very common. Many data are published at a quarterly frequency, such as GDP (gross domestic product). Other data are measured at a monthly or weekly frequency, like the unemployment rate. Finally, many financial data such interest rates or stock prices are available daily, or even intra-daily. This situation gives practitioners a challenge to combine all of these data to enhance forecast accuracy.

One problem induced by mixed frequency data is that the data set is not "balanced" any more. Due to the different publication lag structures, at a specific time period, different variables have different availability. For example, in the middle of April, the daily financial data is available up to current day, and some monthly variables might be available up to March. However, other monthly variables will only be available up to February.

There are many macroeconomic time series available for analysis. The parameters proliferation problem gets worse when we take the lags and leads of variables into account. For example, daily or intra-day trading financial data are abundant. Summarizing those data or selecting representative variables in a certain period presents a challenging task.

High frequency data are becoming more available in the era of big data. With the correct choice of methods, utilizing the high-frequency data to improve forecasts of low-frequency macroeconomics variables is an important topic (Ghysels et al., 2007) . Many researchers have shown already that MIDAS model can incorporate the high-frequency data into the forecast of low-frequency variables and have a positive effect in the improvement of forecast accuracy.

2.2 Aggregation and Interpolation

Traditionally, temporal aggregation is one common way to deal mixed frequency data. The standard aggregation methods, for a stock variable, are to average the observations or to take the latest available observation of high frequency data to match up low frequency data. For a flow variable, the normal way is to sum up the latest available observations in the current period of interest.

On the other hand, interpolation is usually implemented to the low-frequency data to match data frequencies. A two-step procedure is commonly taken: first of all, interpolate the missing data, then estimate the model using the new augmented data.

However, aggregation and interpolation cannot handle the parameter proliferation problem, so a factor model or Bayesian model selection are often developed, as will be discussed below.

A bridge equation is an early econometric method that is used to link or bridge high frequency data to low frequency data:

$$y_{t_q} = \alpha + \sum_{i=1}^{n} \beta_i(L) x_{it_q} + u_{t_q} \ ,$$

where y_{t_q} is the target variable, for example quarterly GDP growth, x_{it_q} are high-frequency predictors but aggregated in quarterly frequency, n is the number of predictors, $\beta_i(L)$ is a lag polynomial of length p, and u_{t_q} is white noise.

In a manner similar to aggregation and interpolation, a bridge equation can handle mixed data frequencies and the ragged edge data problem. For example, bridge models relate the period t value of the quarterly target variable, such as GDP growth, to period t quarterly average of key monthly indicators. At period t, the average of each monthly indicator is obtained with data available within the quarter and forecasts for the rest of the months' values in that quarter (obtained typically from an autoregressive model for the monthly indicator) (Ghysels, Santa-Clara, & Valkanov, 2004).

A bridge equation does not provide the solution for dimension reduction. The variable selection in a bridge equation model is usually taken in a "general to specific" fashion, but it does not search the entire model space and not pick up the best solution necessarily. BMA and other methods can be incorporated in a bridge equation model (Bencivelli, Marcellino, & Moretti, 2012).

2.3 MIDAS Models

Instead of indirectly averaging high frequency data in a bridge equation, the MIDAS model directly introduces high frequency data into the equation of interest (Ghysels et al., 2004) . The MIDAS approach is a form of ADL (additive distribution lag) regression. By transforming high frequency data into low frequency data, a monthly series is converted into 3 quarterly series, each of which collect observations from the same month across the quarters:

$$\begin{bmatrix} y_{2nd\,quarter} & & ; & y_{1st\,quarter} & & \\ x_{June} & x_{May} & x_{Apr}; & x_{Mar} & x_{Feb} & x_{Jan}\cdots \end{bmatrix}$$

$$\begin{bmatrix} y_{2nd\,quarter} & | & x_{June} & x_{May} & x_{Apr} \\ y_{1st\,quarter} & | & x_{Mar} & x_{Feb} & x_{Jan} \\ \cdots & | & \cdots & \cdots & \cdots \end{bmatrix}$$

MIDAS is flexible. The MIDAS approach has been applied successfully to the forecasting of quarterly macroeconomic series using both monthly (Clements & Galvão, 2008, 2009; Kuzin, Marcellino, & Schumacher, 2011) and daily (Andreou, Ghysels, & Kourtellos, 2013; Ghysels & Wright, 2009) data.

One of the informational advantage provided by MIDAS regressions is the use of leads space(Clements & Galvão, 2008). Due to the ragged edge of the data, some high frequency data are leading in terms of the low frequency data. For example, the first quarter GDP usually is published in June. At the beginning of June, two monthly leading variables, 8 weekly leading variables and 44 daily leading variables might be available in terms of forecasting first quarter GDP. The observed information in the leading high frequency variable can help us to infer the information in low frequency data which is not observed yet. Research shows that the gains from the use of leads is significant (Andreou et al., 2013).

2.3.1 Specification of MIDAS

In the MIDAS model, the parameters depend on the forecast horizon , and forecasts are computed directly without requiring forecasts of predictors.

A basic model specification for nowcasting is:

$$y_t = \beta_0 + \beta_1 B(L^{1/m}; \theta) x^{(m)}_{t-(h+1)/m} + \varepsilon_t \, .$$

Here h determines what is the lagging or leading structure. If h is positive, lags of x are included in model. If h is negative, leads of x are used. $1/m$ indicates it is in frenquency-m space. In this way, ragged edge data are directly modeled in the regression. For a forecast for a low-frequency variable using high frequency data, leads are more important. MIDAS regressions with leads, were introduced by Clements and Galvão (2008) and Kuzin et al. (2011) in the context of monthly quarterly data mixtures.

y is the target variable which is the macroeconomic variable, we are interested in; x is the set of predictors which we use for forecasting; m denotes the frequency difference. For instance, if y is annual data, $x_t^{(4)}$ is quarterly. ε is the disturbance and $B(L^{1/m}; \theta)$ is a lag distribution. For instance, this may be the Beta density function or the Almon Lag function, which is used to avoid parameter proliferation. These weighting scheme can be chosen by cross validation to capture the impacts of lags of the variables.

2.3.2 Weighting schemes for parameter proliferation

In the MIDAS model, we can treat the coefficients associated with the high frequency variables as weights. To prevent parameter proliferation, a MIDAS model uses a weighting scheme to reduce the number of parameters. The weighting schemes are assumed to take some functional form such as Beta density or an Almon lag polynomial.

For example, by using a constrained distributed lag of the predictors, MIDAS can relate the value of the quarterly variable of interest to a small number of linear or nonlinear combinations of monthly, weekly, or daily predictors.

Normalized exponential Almon lag restricts

In the econometrics literature, weight functions can be either linear or nonlinear specifications. The normalized exponential Almon lag function is particular flexible (Ghysels et al., 2007).

Consider a two-parameter exponential Almon lag function with the following h-period ahead predictive regression:

$$y_t = \beta_0 + \beta_1 W(L^{1/m}, \theta) x^{(m)}_{1,t-(h+1)/m} + \varepsilon_t \, ,$$

where the weighting scheme is

$$W(L^{1/m}, \theta) = \sum_{k=1}^{K} w(k; \theta) L^{(k-1)/m} \, ,$$

the lag structure is

$$L^{s/m} x^{(m)}_{1,t} = x^{(m)}_{1,t-s/m} \, .$$

Here, t denotes the basic time unit for the lower frequency data (from 1 to T), m is the frequency difference, and $x^{(m)}$ indicates higher sampling frequency observations. The lag structure indexes from 1 to K (where K can be chosen by some criteria such as BIC). $L^{1/m}$ is the lag operator in frequency-m space, and $w(k; \theta)$ is the weight on each of the K lagged higher frequency predictors. ε_t is white noise.

A two-parameters normalized exponential Almon lag restricts the coefficients θ_h in the following way:

$$w(k; \theta_1; \theta_2) = \frac{\exp(\theta_1(k+1) + \theta_2(k+1)^2)}{\sum_{k=0}^{K} \exp(\theta_1(k+1) + \theta_2(k+1)^2)} \, .$$

In this manner, the parameters associated with x_1 are reduced from K to 2, and all lagged x_1 have different weights, $w(k; \theta_1; \theta_2)$, and share the same coefficient β_1.

The Almon lag structure can capture the evolution of lag effects smoothly. Usually, the more recent lags have a larger effect on the target variable, and the less recent lags have smaller impact on the target variable, and this can be represented by an Almon lag weighting schema. For example, different parameters of Almon lag weighting schema can represent different forms of structure for the impact of lags on the target variable.

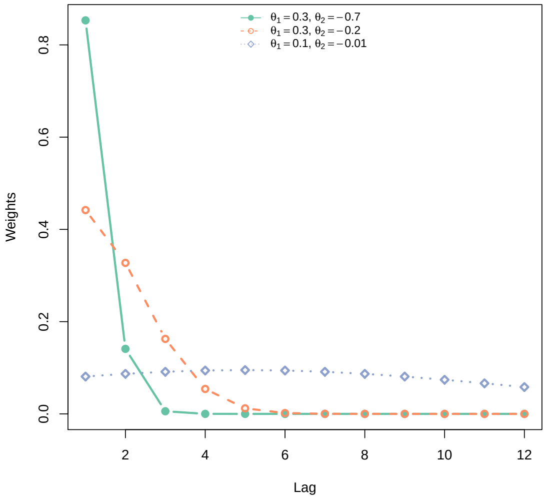


Figure 2.1: Weighting schema of exponential Almon Lag function

The resulting model can be estimated by non-linear least squares and used to forecast the target variable from constrained distributed lags of the available data. (Guérin & Marcellino, 2013).

One advantage of such a weighting scheme is that it reduces the dimension of the parameter space and it extracts information from higher frequency data efficiently.

One disadvantage of such a weighting scheme comes from imposing arbitrary restrictions on the higher frequency data. For example, exponential Almon lag put positive restriction on the weights for the higher frequency data, which is not necessarily correct (Foroni & Marcellino, 2013). The model is susceptible to misspecification.

2.3.3 U-MIDAS weight specifications

Without a weight scheme, an unrestricted MIDAS regression is similar to an ordinary least squares regressions model which including unconstrained distributed lags of the high frequency predictors.

Consider a two predictors model:

$$y_t = \beta y_{t-1} + \sum_{h=0}^{d} \theta_h x_{tm-h} + \sum_{j=0}^{p} \phi_j z_{tn-j} + u_t$$

where m and n show x and z have different frequencies. h and j show x and j have different publication lags. Researchers find that an U-MIDAS model can be easily and fast estimated by OLS. Foroni and Marcellino (2013) find that if the frequency mismatch is small, U-MIDAS outperforms MIDAS, for example, when mixing monthly and quarterly data.

2.4 State-Space Approach

Banbura, Giannone, Modugno, and Reichlin (2013) distinguish two types of models for macroeconomic forecasting: partial model and full system methods.

A bridge equation and a MIDAS model only have a single equation in the model, and can be categorized as a partial model. On the other hand, state-space approach can be categorized as full system model since it involves a system of equations. The system of equations models not only the dynamics of the target variable, but also the dynamics of the predictors.

Bai et al. (2013) compare the state-space system model and MIDAS single equation model. They find MIDAS actually can be represented as a reduced form of a state-space model under some conditions. In most cases, the state-space model gives slight accurate forecast, but state-space model is more computation intensive when the number of predictors is large. On the other hand, the computation of MIDAS model is straightforward and it can handle a large number of predictors.

The full system approach includes mixed-frequency VAR, dynamic factor model and factor MIDAS model. This approach uses state space form to handle data with different frequencies. The low frequency target variable is treated as a high frequency one with missing observations. The missing or unobserved latent data can be estimated in a state space model using Kalman filter.

To deal with the ragged edge data and prediction problem, this approach uses Kalman filter to model the dynamic of the factors and predict the missing value of factors (VAR is a special case of Kalman filter/ state space model). At the same time, in order to solve the parameter proliferation problem, some model in this approach use a few factors to capture the information among many predictors.

2.4.1 Mixed frequency VAR

The vector autoregressive model (VAR) is a great tool to capture co-movement of a set of variables. Instead of a univariate equation, Mixed frequency VAR characterize the co-movements in a system of equations.

Classical approach

A classical approach to estimate the VAR model is maximum-likelihood estimation (MLE) with the expectation maximization algorithm(EM).

Following the notation of Foroni and Marcellino (2014), a mixed frequency VAR model with quarterly target variable and monthly predictors can be represented by a state-space model:

Transition equation:

$$s_{t_m} = F s_{t_m - 1} + G v_{t_m} \; .$$

s_{t_m} is the state variable at monthly frequency, which can be constructed as follows:

$$s_{t_m} = \begin{bmatrix} z_{t_m} \\ \cdot \\ \cdot \\ \cdot \\ z_{t_m - 4} \end{bmatrix}, \qquad z_{t_m} = \begin{bmatrix} y^*_{t_m} - \mu^*_y \\ x_{t_m} - \mu_x \end{bmatrix}$$

where $y^*_{t_m}$ is an unobserved latent monthly target variable, $\mu_y = E(y)$, $\mu_y = 3\mu^*_y$ and $v_{t_m} \sim N(0, I_2)$. F is a transition matrix.

Measurement equation:

$$\begin{bmatrix} y_{t_q} - \mu_y \\ x_{t_m} - \mu_x \end{bmatrix} = H s_{t_m} + \varepsilon_t$$

where y_{t_q} and x_{t_m} are observed data at quarterly and monthly frequencies. ε_t is a measurement error.

With this state-space representation, the MLE is used to estimation, and Kalman filter is used to forecast.

Bayesian approach

Instead of MLE with EM algorithm, the Mixed frequency VAR also can be estimated by Bayesian methods. Schorfheide and Song (2015) develop a MCMC method for mixed frequency VARs. They use a Minnesota prior for the parameters matrix to deal with the parameters proliferation problem.

Banbura and Modugno (2014) and Foroni and Marcellino (2013) provide a detailed survey of full system methods, including the Bayesian approach.

2.4.2 Mixed-frequency factor model

In order to avoid parameters proliferation, factor models are broadly used to summarize information among a large number of predictors.

Factor models decompose time series into a common component, driven by a few factors that represent the key economic driving forces.

Factor model

Consider a regression model with a large number of predictors,

$$y_t = \beta_0 + \beta \mathbf{x}_t + \epsilon_t, \quad t = 1, \cdots T$$

where $\mathbf{x}$ is a $N \times 1$ vector and $N > T$ for a fat regression scenario. N is the number of predictors, and T is the number of observations. This will be impossible to estimate due to the high dimension of the model. With the help of principal component analysis, the dimension can be reduced from N to q. q is the number of the factors. Let:

$$\mathbf{x}_t = \Lambda \mathbf{f}_t + \eta_t$$

$$y_t = \beta_0 + \beta \mathbf{f}_t + \epsilon_t \,,$$

where $\mathbf{f}$ is $q \times 1$ vector, Λ is the factor loading matrix. Traditionally, common factors are estimated by principal components.

In a static factor model, the relationship be tween $\mathbf{x}$ and $\mathbf{f}$ is static , although $\mathbf{f}$ itself can be a dynamic process.

Dynamic factor model

In a dynamic factor model, the relationship between **x** and **f** is dynamic:

$$X_t = \lambda(L) f_t + e_t$$

$$f_t = \Psi(L) f_{t-1} + \eta_t$$

where X_t and e_t are $N \times 1$, f_t and η_t are $q \times 1$, L is the lag or lead operator depending on what data we have. The lag polynomial dynamic factor loading matrix $\lambda(L)$ is $N \times q$, and the lag polynomial transition updating matrix $\Psi(L)$ is $q \times q$. With this setup, the factors can be used to forecast the target variable using a regression of the target variable on the factors and lagged target variable.

If q low dimension latent dynamic factors represent the co-movement of a high dimension vector of N time series X_t, we can get an efficient forecast for y_{t+h}:

$$E(y_{t+h}) = \alpha(L)\mathbf{f_t} + \delta(L) y_t$$

By using aggregated factors as predictors, the dimension of the model is reduced from N to q. When we include new variables in X, it will not add dimension to f. A dynamic factor model can be represented by a state-space formulation and estimated using the Kalman filter and MLE. The missing values in the ragged edged data can be handled by an EM algorithm (Stock & Watson, 2006) .

An alternative solution is a full state-space model approach, which treat low-frequency target variable as a latent monthly data.

$$X_{t_m} = \Lambda F_{t_m} + \xi_{t_m}$$

$$\Psi(L_m) F_{t_m} = B\eta_{t_m} ,$$

where η_{t_m} is a q-dimensional vector that represents the dynamic factor shocks. The target variable also is treated as a high-frequency variable with missing observations. For example, say, y_{tm} is latent monthly GDP growth. The observed quarterly GDP growth y_q will be treated as the observation in the third month of each quarter, and the other monthly observations are taken as missing.

One disadvantage of the MLE and the Kalman filter for dynamic factor models is when the dimension of X gets very large and there are many parameters needed to be estimated, the MLE becomes unfeasible since it does not have unique solution and the model is not identified. Doz, Giannone, and Reichlin (2012) propose a quasi-ML algorithm to estimate the factors, and then use the Kalman filter to estimate the unobserved latent monthly factor. Jungbacker, Koopman, and van der Wel (2011) show that a transformation of X into low dimension can speed up the calculation of Kalman filter.

2.4.3 Factor-MIDAS

Given the limitations of a state-space model for dealing with high dimensionality, another popular option is to combine the factor model and MIDAS model.

Suppose the current information set has high dimensional N predictors X_{t_m}. X_{t_m} can be summarized by low dimensional r factors F_{t_m}. The most recent available estimated factors F_{t_m} can be used in a MIDAS model.

The basic Factor-MIDAS approach can be represented as follows: (for simplicity, r is set to equal 1; the target variable is quarterly,the factor is monthly, and the forecast horizon is h_q quarters with $h_q = h_m/3$)

$$y_{t_q+h_q} = y_{t_m+h_m} = \beta_0 + \beta_1 b(L_m;\theta)\hat{f}^{(3)}_{t_m+w} + \varepsilon_{t_m+h_m}$$

$$b(L_m;\theta) = \sum_{k=0}^{K} c(k;\theta)L_m^k$$

Here, $c(k;\theta)$ is a weighting function, which could be an exponential lag function, or it can be a constant if it is an unlimited MIDAS model. $\hat{f}^{(3)}_{t_m}$ is skip-sampled from the monthly factor $\hat{f}_{t_m}$. (3) is the frequency difference. For example, $\hat{f}^{(3)}_{t_m} = \hat{f}_{t_m}, \forall t_m + w = ..., T_m + w - 6, T_m + w - 3, T_m + w$. With the proper weighting function, the model can be solved by non-linear least squares.

2.5 Shrinkage Methods

Stock and Watson (2012) discuss generalized shrinkage methods and show shrinkage methods can handle the high dimension problem just like a factor model. Shrinkage methods include high-dimensional Bayesian vector autoregression (Andersson & Karlsson, 2008; Banbura, Giannone, & Reichlin, 2010; Korobilis, 2013; Carriero, Kapetanios, & Marcellino, 2011), Bayesian model averaging (Koop & Potter, 2004; Wright, 2009), bagging (Inoue & Kilian, 2008) , Lasso (De Mol et al., 2008; Bai & Ng, 2009) , boosting (Bai & Ng, 2007), and forecast combination (Timmermann, 2006).

2.5.1 The LASSO augmented MIDAS

To reduce the dimension of the parameter space, factor models decrease the number of regressors directly. In contrast, penalized regression puts a penalty on the parameters. Penalized regression such as Ridge regression and the LASSO (Least Absolute Shrinkage and Selection Operator) method (Tibshirani, 1994) have been exploited in the MIDAS approach literature for several years (De Mol et al., 2008; Ferrara & Marsilli, 2013).

De Mol et al. (2008) show that for the forecasting of macroeconomic panel data, a Bayesian shrinkage method is a valid alternative to principal components.

The LASSO augmented MIDAS model can be specified as follows:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{t=1}^{T} (y_t - \beta_0 - \sum_{i}^{N} \beta_i x_{t,i})^2 + \lambda_{lasso} \sum_{i}^{N} |\beta_i|$$

or in matrix form as:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||Y - X\beta||_2^2 + \lambda_{lasso} ||\beta_i||_1 \,,$$

where y_t is target variable, $x_{t,i}$ is predictors, β is the vector of the regression parameters, N is the size of the model, T is number of observations, and λ_{lasso} is the tuning parameter which decides how large the penalty is and how sparse the regularization is. Similarly, a Ridge regression comes with a penalty $\lambda_{ridge}||\beta_i||_2$ based on the norm ℓ_2, instead of ℓ_1.

The idea behind the LASSO regression is that the penalty term in optimization process will drive the coefficients β_i associated with non-informative predictors x_i to zero and result in a parsimonious model (Ridge regression pushes β_i towards to zero, not exactly to zero). The λ can be chosen by cross validation. Both LASSO and Ridge regression have Bayesian interpretation. If we assume β_i with a prior of a Gaussian distribution with zero mean and standard deviation a function of λ, then the posterior mode for β_i is the Ridge regression solution. If we assume β_i with a prior of double-exponential (Laplace) distribution with zero mean and scale parameter a function of λ, then the the posterior mode for β_i is given by the Lasso. (James, Witten, Hastie, & Tibshirani, 2013)

2.5.2 Model averaging

Three popular solutions for "curse of dimensionality" problem in the literature are the factor model with principal components; penalized regression with shrinkage; and Bayesian model averaging with variable selection (De Mol et al., 2008; Ouysse, 2013) . These three methods are highly correlated. In some sense, the approach of the factor model and MIDAS model are to aggregate data before forecasting. However, the approach with model averaging is to aggregate the data after forecasting (Hendry & Hubrich, 2011) .

Consider a model space consists of r models $M_1, ..., M_r$, and we have the observation data D, the prior of ith model $P(M_i)$the prior of the parameter vector in the ith model $P(\theta_i|M_i)$, and likelihood $P(D|\theta_i, M_i)$. The posterior of ith model is follows:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{i=1}^{r} P(D|M_i)P(M_i)},$$

where

$$P(D|M_i) = \int P(D|\theta_i, M_i)P(\theta_i|M_i)d\theta_i$$

which is marginal likelihood of the ith model. On each ith model, we can have a forecast $E(y_{t+1}|D, M_i)$. The weighted average forecast is follow:

$$y_{t+1|t} = E(y_{t+1}|D) = \sum_{i=1}^{r} E(y_{t+1}|D, M_i)P(M_i|D)$$

Timmermann (2006) and Stock and Watson (2009) provide a survey of forecast combination methods. When there are n possible regressors, there are 2^n possible models in the model space. Searching through the entire model space is very demanding. Furthermore, the parameters might not be time invariant. These issue causes uncertainty and instability of the model. Forecast combinations extract information from all possible forecasting models rather than from a single specific model, which helps to overcome the the problem of model uncertainty (Andreou et al., 2013). A model ignoring the uncertainty and variability, will not be reliable.Forecast combinations can produce more stable forecasts than individual forecasts, which is useful for dealing with model instability and structural breaks.

2.5.3 Bagging and boosting

Along with shrinkage method such as Ridge and LASSO, other statistical learning techniques are also used in macroeconomic data analysis. Those methods include ensemble methods like bagging, random forests, and boosting. In the macroeconomics data analysis, Bootstrap aggregation or bagging (Breiman, 1996) smooths the hard threshold in pre-test estimators by averaging over a bootstrap sample of pre-test estimators.

In terms of the trade-off between bias and variance, bagging can reduce the variance of the estimators significantly without introducing a lot of bias. Inoue and Kilian (2008) apply bagging to a forecasting situation and report promising results. Bai and Ng (2009) discuss three methods: LASSO (least absolute shrinkage and selection operator), LARS (least angle regression), and the elastic net for macroeconomic forecasting.

Bai and Ng (2009) suggest a way of using boosting to select variables in a factor model setting, and show that some forms of boosting outperform the standard factor-augmented forecasts. Buchen and Wohlrabe (2011) show that boosting is a serious competitor for forecasting US industrial production. Carriero, Clark, and Marcellino (2015) show that compared to factor models and Bayesian VAR, multivariate boosting performs best when forecasting CPI inflation. Wohlrabe and Buchen (2014) also show that boosting outperforms the autoregressive benchmark when forecasting macroeconomic variables for the United States.

Chapter 3

Model: BSTS-U-MIDAS

3.1 Motivation

We propose a hybrid model between a partial model and a full system model for macroeconomic data forecasting. BSTS-U-MIDAS consists of two parts: a Bayesian Structural Time Series (BSTS) model, and U-MIDAS model. The first part captures the dynamic feature of the target variable. The second part includes a large number of macroeconomic panel data as predictors. To specific, we adopt spike-and-slab regression for variable selection to handle the high dimensionality problem, and use BMA to deal with model uncertainty and instability.

Currently, in the full system approach, the dynamic factor model is a good method for forecasting macroeconomic variables with mixed frequency data. On the other hand, in the partial model approach, MIDAS is easily implemented and is flexible enough to handle the ragged edge data problem. A combination such as a dynamic factor MIDAS is a good choice. However, the factor model has its limitation. Indeed, factors can summarize and extract useful information from predictors; however, factors may extract noise too (Bai & Ng, 2009). Choosing factors based on the ordering of the eigenvalues does not guarantee that we get factors with high predictability. Evidence shows that model averaging outperforms modeling based on factors (Ouysse, 2013). If most macroeconomic variables are highly correlated, a few selected variables are able to capture the important information just like factors can do. Research shows that forecasts from variables are highly correlated with the forecasts from factors (Bai & Ng, 2009). Moreover, with the factors as predictors,a forecasting model is less useful for interpreting the relationship between macroeconomic variables that we are interested in. We choose spike-and-slab regression and BMA to deal with parameter proliferation instead of a factor model.

We adopt U-MIDAS instead of regular MIDAS because we do not want to impose arbitrary restrictions on the parameters. Since we introduce spike and slab regression and BMA in the model, the weighting scheme for handling the parameter proliferation problem is not necessary. An U-MIDAS model is good enough for addressing the ragged edge data problem. In our model, the BSTS component deals only with the dynamics of the target variable since we only focus on forecast performance and in such way BSTS has less computational burden. In our model, in order to avoid spurious regression, all predictors are filtered by detrending, deseasonalization, and scaling, which is done by using 'forecast' package in R (Hyndman, 2015).

Our contribution is that we are the first to combine BSTS and MIDAS for macroeconomic nowcasting. We also adopt a vertically aligned method (Altissimo, Cristadoro, Forni, Lippi, & Veronese, 2010) to include leading variables into the MIDAS model directly. This method is very flexible and easy to implement, and the only disadvantage is that it needs to recalculate the model every time when a new variable comes out.

3.2 Model Specification

The stages involved in specifying the model are as follows: **Bayesian Structural Time Series**

- Use Kalman filter to whiten time series;
- Control for trend and auto-aggression;
- Build a model for the predictable part of time series.

Spike and slab regression for variable selection

- Find regressors that predict the target.

Bayesian model averaging for final forecast.

- Find forecasts across many possible forecast models.

The first part of our model is a structural time series model(STM). Harvey (2006) reviews structural time series models for forecasting. Structural time series models decompose time series into different stochastic components like trend, seasonality, and cycle. The flexible structure allows the models to handle irregular data such as missing data and mixed frequency data. The flexibility also makes forecasting and nowcasting more feasible.

According to Harvey, the STM approach is different from the Box-Jenkins ARIMA approach which dominated the 1970's and 1980's time series forecasting (Harvey, 2006; Frale, Marcellino, Mazzi, & Proietti, 2011). An ARIMA model can be represented in state-space form. With the development of computing power and new algorithms, the STM has been given more attention due to its good forecasting performance.

3.2.1 Local linear trend model with regression

Harvey (1990), Durbin and Koopman (2012), Petris, Petrone, and Campagnoli (2008) and many others have advocated the use of the Kalman filter for time series forecasting.

In our model, we only include three components: trend, autoregression and regression components. According to our experiment, detrended and deseasonalized predictors and a deseasonalized target variable have better forecast performance. Since our goal is forecasting, we don't include seasonal terms in our model, but it can be done within our model (Scott & Varian, 2014b).

The "basic structural model" decomposes the time series into four components: a level, a local trend, seasonal effects and an error term. In contrast to the Box-Jenkins ARIMA approach, STM is able to handle some irregular data such as being nonstationary, heterosedastic, nonlinear, and non-Gaussian. The Kalman filter is used to formulate the likelihood function and conduct inference.

Local linear trend model with regression

- Observation equation(level + regression):

$$y_t = \mu_t + z_t + v_t, \quad v_t \sim N(0, V)$$

- State equation 1 (random walk + trend):

$$\mu_t = \mu_{t-1} + b_{t-1} + w_{1t}, \quad w_{1t} \sim N(0, W_1)$$

- State equation 2 (random walk for trend):

$$b_t = b_{t-1} + w_{2t}, \quad w_{2t} \sim N(0, W_2)$$

- Regression component:

$$z_t = \beta x_t,$$

  By appending a constant 1 ot state vector α and appending βx_t to observation matrix F, we only increase the dimension of the state vector α by one (Scott & Varian, 2014a).

- Parameters to estimate:

$$\theta : \beta, V, W_1, W_2$$

- States to estimate:

$$\alpha : \mu_t, b_t$$

The Kalman filter and is used to estimate the parameters and states. Those estimations are used to forecast in the Kalman filter framework.

Optimal Kalman forecast:

$$\hat{y}_t = \mu_t + z_t + K_t(\hat{y}_{t-1} - y_{t-1})$$

where K is optimal Kalman gain which is a function of the variance terms v_{1t}, w_{1t}, w_{2t}. In general, the optimal forecast will be a weighted average of past observations and the current observation. The weight K depends on variances of the error terms (See Appendix A for detail) .

Advantages of the Kalman filter:

The Kalman filter has several advantages. First of all, it has no problem with mixed frequency. Second, it has no problem with unit roots or other kinds of non-stationarity as well. Since it does not require stationary data, it does not need a differencing or identification stage, so it is easy to automate. Third, it has a nice Bayesian interpretation. It updates its estimates when new observation arrives. Fourth, it has good forecast performance. It decomposes time series into several stochastic processes which are able to capture the dynamics of the time series.

Disadvantages of the Kalman filter:

The Kalman filter model assumes linearity, and is based upon noise terms that are normally distributed. This strong assumption is not always met in practice. Furthermore, the computational complexity increases linearly in the number of observations and quadratically in the size of the model. This is the reason why we prefer model only the dynamics of the target variable by the Kalman filter and only take regression component $z_t = \beta x_t$ as one state.

3.2.2 MIDAS for mixed frequency data

Though the Kalman filter can handle the mixed frequency data problem and therefore the ragged edge data problem, it becomes computationally difficult when high dimensionality presents itself in the state vectors. We tried to model a dynamic regression component in our model. It models the dynamics of the predictors by setting the $\beta's$ of the predictors as the states in the state-space specification, but this did not improve the forecast accuracy. Since our goal is to forecast, and not to fit the model, we adopt the MIDAS model to handle the mixed frequency problem. MIDAS includes the high-frequency data directly in the model, which simplifies the computation and improves the efficiency. We choose U-MIDAS since it does not impose restrictions on the parameters by weighting schemes. We adopt a flexible and easy way to align the ragged edge data which is similar to Altissimo et al. (2010).

We realign each high frequency time series in the sample to match the low frequency target variable according to the forecast horizon vertically. For example, to match a quarterly target variable from a daily predictor y_{q+h}, (q is the quarter and h is the forecast horizon), we realign all of the predictors to the current date to form a balanced data set. For daily data, the alignment is as follows:

$$\tilde{x}_{i,q} = x_{i,q+k_i} \,,$$

where k_i is number of days of the publication lags or leads for ith variable. For financial data, the number of trading days in one month is 22, so we have $k = 0, ..., 22 \times 3$. Since we focus on nowcast, we tend to choose those variables with leads. Specifically, GDP publication lag usually is two months, so for a daily data, we may have $k \in 0, ..., 22 \times 2$ leads, and for monthly data, we may have $k = 0, ..., 3 \times 2$.

The $\tilde{x}_{i,q}$ is skip-sampled data; for example, if k is 25, to match up quarterly GDP data, the $\tilde{x}_{i,q}$ is the 25th daily observation in the each quarter. Applying this method to each predictor, we get a balanced dataset $\tilde{X}_q$.

Here is a example of an alignment of quarterly and monthly data when we want to forecast second quarter target variable but we only have monthly data up to May.

$$\text{Vertical alignment: } \begin{bmatrix} y_{2nd\,quarter} & & ; & y_{1st\,quarter} & & ;\cdots \\ x_{May} & x_{Apr} & x_{Mar}; & x_{Feb} & x_{Jan} & x_{Dec};\cdots \end{bmatrix}$$

$$\left[\begin{array}{c|ccc} y_{2nd\,quarter} & x_{May} & x_{Apr} & x_{Mar} \\ y_{1st\,quarter} & x_{Feb} & x_{Jan} & x_{Dec} \\ \cdots & \cdots & \cdots & \cdots \end{array}\right]$$

The main advantages of this approach is its simplicity. Whenever new data come out, we can reestimate the model and update forecasts.

The disadvantage is every time when new data come out, the balanced dataset $\tilde{X}_q$ changes, the correlations between variables change, and the stability of model may change as well (Foroni & Marcellino, 2013) .

3.3 Variable Selection: Spike-and-Slab Regression

In the Bayesian framework, the *spike and slab prior* is an effective variable selection device to achieve sparsity (George & Mcculloch, 1997; Madigan & Raftery, 1994) . It has been used in macroeconomic forecasting (Rodriguez & Puggioni, 2010; Kaufmann & Schumacher, 2012; Ferrara, Marsilli, & Ortega, 2014) .

If there are N predictors, then there are at least 2^N possible linear models in the model space. The spike-and-slab prior is a way to carry out the variable selection. One of the approaches is a stochastic search variable selection strategy (SVSS), which takes the prior to be a mixture of two normal distributions:

$$\beta_i \sim \gamma_i N(b_i, \varphi^2) + (1 - \gamma_i) N(0, c\varphi^2)$$

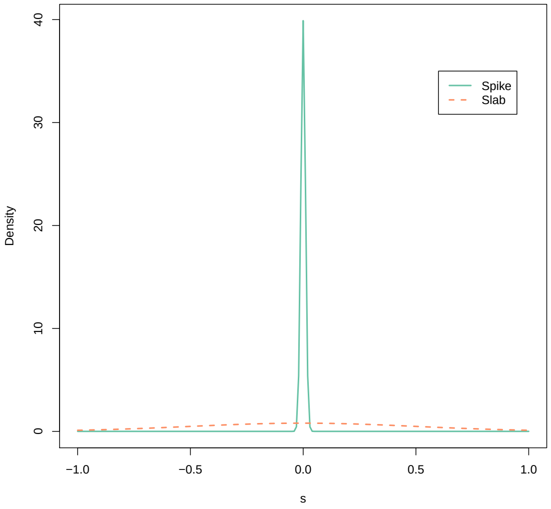


Figure 3.1: Spike and slab prior

where γ is a vector of indicator variable. The marginal distribution $p(\gamma)$ is the "spike" since it has positive probability mass at zero. When $\gamma_i = 1$, variable x_i has a non-zero coefficient in regression. c is a very small positive number, so when $\gamma_i = 0$, variable β_i could be 'safely' estimated by zero and excluded from the regression. γ_i is a binary random variable which can be modeled by a product of independent Bernoulli distribution.

$$\gamma_i \sim \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}$$

and we have:

$$\gamma \sim \prod_i \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}\,.$$

Given $p(\gamma)$, a spike-and-slab prior can be factored as:

$$p(\beta, \sigma^{-2}, \gamma) = p(\beta \mid \gamma, \sigma^{-2})p(\sigma^{-2} \mid \gamma)p(\gamma)\,,$$

where π_i is predictor x_i's probability of inclusion in the regression. When detailed prior information is unavailable, it is convenient to set all π_i equal to the same number, π. The common informative prior inclusion probability can easily be elicited from the expected number of nonzero coefficients. For example, if n out of N coefficients are expected to be nonzero, then we can set $\pi = n/N$ in the prior. For the expected number of predictors n, we choose 4 in our model since in a cross validation, the performance of model with 4 is better than performances of models with other numbers.

In such a hierarchical structure, a variable x_i with $\gamma_i = 1$ has a *conditionally conjugate* “slab” prior:

$$\beta_\gamma \mid \sigma^2, \gamma \sim N(b_\gamma, \sigma^2(\Omega_\gamma^{-1})^{-1}) \, ,$$

$$\frac{1}{\sigma^2} \mid \gamma \sim \Gamma(\frac{df}{2}, \frac{ss}{2}) \, .$$

where b_γ is a vector of prior guesses for the regression coefficients, σ^2 is the overall variance level of error, and Ω^{-1} is a prior precision matrix, and Ω_γ^{-1} is rows and columns of Ω^{-1} when $\gamma_i = 1$.

Conventionally, b_γ is set to equal 0, and $\Omega^{-1} \propto X^T X$, which is known as Zellner's informative g-prior (Chipman, George, & McCulloch, 2001). If we have specific information about b_γ, a informative prior will helpful for the estimation. The matrix $X^T X$ is the total Fisher information matrix in the full data, and we can parametrize $\Omega^{-1} = \kappa(X^T X)/T$, which representing the average information available from κ observations. And it means we places κ observations worth of weight on the prior mean b_γ (Scott & Varian, 2014a).

It shows that given γ, the prior for the precision follows a Gamma distribution with parameters $\frac{df}{2}$ and $\frac{ss}{2}$. Thus, the reciprocal of the mean of the Gamma distribution ss/df is a prior estimate of σ^2.

The “Slab” prior is a very weakly informative prior which is close to being flat. In some sense, ss can be interpreted as a prior sum of squared error, and the df can be interpreted as a prior sample size. The bigger the sample size is , the more information about the parameters we have, and then we have a better precision and a smaller variance for estimation of the parameters. Further details of elicitation of the priors are discussed in Appendix C.4.

It turns out that there is a connection between SSVS and LASSO (Yuan & Lin, 2005). A spike and slab prior for β_i can be formulated as follows:

$$\beta_i = (1 - \gamma_i)\delta(0) + \gamma_i DE(0, \lambda)$$

$\delta(0)$ is the point mass distribution centered at zero for predictors which is excluded in model. $DE(0, \lambda)$ is the double exponential (Laplace) distribution with density $\lambda e^{\lambda|x|}$ which is for the predictors which is included in model.

The prior for indicator γ is

$$p(\lambda) \propto \pi^{|\gamma|}(1-\pi)^{p-|\gamma|}|X_\gamma^T X_\gamma|^{\frac{1}{2}}.$$

$|\gamma| = \sum_{i=1}^{p} \gamma_i$ is the number of predictors in the model. $|X_\gamma^T X_\gamma|^{\frac{1}{2}}$ penalizes models with correlated predictors. It can be shown that the posterior mode of γ selects the same parameters as the LASSO.

3.4 Model Estimation Using MCMC

We follow Scott and Varian (2014a) and implement Markov Chain Monte Carlo (MCMC) to obtain the posterior distribution of the coefficients. For a MCMC process, we need to know the full set of conditional distributions of the parameters. Scott and Varian (2014a) give a detailed algorithm and 'bsts' package in R to implement the algorithm (Scott, 2015).

The conditional posterior of β and σ^2 given γ

Suppose we have a local linear trend model, the observation equation is

$$y_t = \mu_t + \beta \mathbf{x}_t + v_t$$

We subtract the target time series component μ from y_t, and get an axillary variable $y^* = y_t - \mu_t$. Then we have

$$y_t^* = y_t - \mu_t = \beta x_t + \epsilon_t \sim N(\beta x_t, \sigma^2)$$

And we are left with a standard spike-and-slab regression. We can use "stochastic search variable selection"(SSVS) algorithm to draw from $p(\beta_\gamma, \sigma^2|\gamma, \alpha, \mathbf{y}^*)$, where vector $\mathbf{y}^* = y_{1:T}^*$ is all the information about y^* up to time T.

Then conditional on γ, the joint posterior distribution for β and σ^2 can be estimated from standard conjugacy formula (Gelman et al., 2013) :

$$\beta_\gamma \mid \sigma, \gamma, \mathbf{y}^* \sim N(b_\gamma, \sigma^2(V_\gamma^{-1}))$$

$$\frac{1}{\sigma^2} \mid \gamma, \mathbf{y}^* \sim \Gamma(\frac{df+T}{2}, \frac{ss+\tilde{S}}{2})$$

More discussion can be found in Scott and Varian (2014b) and Ferrara et al. (2014). The details are discussed in Appendix C.5.

The marginal posterior of γ

Due to the conjugacy, we can get analytical expression for the marginal posterior of γ by marginalizing over β_γ and $1/\sigma^2$:

$$\gamma \mid \mathbf{y}^* \sim C(\mathbf{y}^*) \frac{|\Omega^{-1}|^{1/2}}{|V_\gamma^{-1}|^{1/2}} \frac{p(\gamma)}{SS_\gamma^{\frac{T}{2}-1}} ,$$

where $C(\mathbf{y}^*)$ is a normalizing constant.

MCMC process

Since we have a state-space model and a spike-and-slab regression, the parameters we need to estimate are as follows:

Parameters to estimate:

$$\theta_1 : V_{1t}, W_{1t}, W_{2t}$$

and

$$\theta_2 : \beta, \sigma^2, \gamma$$

States to estimate:

$$\alpha : \mu_t, b_t$$

Since the state α has Markov property, we assume that conditional on state α, the time series components which relate to v_{1t}, w_{1t}, w_{2t}, and the regression components related to β, γ are independent. Then we can follow three steps as follows:

1. Simulate the state α from

$$\alpha \sim p(\alpha \mid \mathbf{y}, \theta_1, \theta_2) .$$

2. Simulate the parameters for the time series components:

$$\theta_1 \sim p(\theta_1 \mid \mathbf{y}, \alpha, \theta_2) .$$

3. Simulate the parameters for the regression components:

$$\theta_2 \sim p(\theta_1 \mid \mathbf{y}, \alpha, \theta_1)\,.$$

Iterating the three steps, we can get the posterior distribution of state α and parameters θ_1, θ_2 given $\mathbf{y}$. Under a Gaussianity assumption, the $p(\alpha|y)$ is solved by a stochastic version of the Kalman smoother (Durbin & Koopman, 2002). The first set of parameters θ_1 are obtained via conjugacy prior. For the variance parameters W_t, W_{2t}, they have independent inverse Gamma priors, and their full conditional posterior distributions are independent inverse Gamma distributions as well. The second set of parameters θ_2 can be estimated by the stochastic search variable selection (SSVS) algorithm from George and Mcculloch (1997), which uses a Gibbs sampling algorithm. Further details are discussed in Appendix C.5.

One of issues of Bayesian variable selection is computational difficulty. It can be shown that the some form of the joint posterior of parameters may have more than one mode (Park & Casella, 2008). The multiple modality could cause conceptual and computational problems. Conceptually, it is hard to summarize the posterior distribution. Computationally, the Gibbs sampler will have difficulty to converge or stuck at one corner and not explore the entire model space enough. There are many methods developed to tackle these issues(Ishwaran & Rao, 2005). We are aware these issues, and try to increase the number of iteration to improve the estimation.

3.5 Bayesian Model Averaging

Let ϕ be $\alpha, \theta_1, \theta_2$, then with the draw of state and parameters from their posterior distribution, we can get the posterior predictive distribution for the target variable y:

$$p(\tilde{y} \mid \mathbf{y}) = \int p(\tilde{y} \mid \phi) p(\phi \mid \mathbf{y})\; d\phi\,.$$

For each draw of $\phi^{(i)}$ from $p(\phi \mid \mathbf{y})$, we can sample a $\tilde{y}$ from $p(\tilde{y} \mid \phi)$. The sample of draws from the posterior predictive distribution $p(\tilde{y} \mid \mathbf{y})$ can give us the information that we need for forecasting. To summarize the information, we can use the mean, median, or mode as a point forecast of the target variable y.

To consider the model uncertainty, we also can get a forecast interval or use a histogram or density to summarize the forecast. To understand the impact of the predictors on th target variable, we can get the inclusion probability of specific predictors by summary of γ_k. For example, we can take the average over draws of γ to see which predictors have high probability of being in the regression.

Chapter 4

Empirical Application

In this chapter, we report our results from an empirical application. We use five economic variables to predict GDP for Canada.

4.1 Why Choose Financial Data?

Andreou et al. (2013) demonstrate that daily financial data can help forecast and nowcast the quarterly real GDP growth. Banbura and Modugno (2014) show that real time daily financial data can improve the nowcasting of low frequency macroeconomic data. Ferrara and Marsilli (2013) have recently shown that daily financial series have a significant forecasting power when it comes to US economic growth, specially when the volatility of those signals is considered as a explaining factor and not only its returns.

We choose real gross domestic product(GDP) for Canada as the target variable for forecasting. GDP is a very important indicator for economic activity and has been used for economic decisions and public policy in many institutions. Forecasting GDP is very challenging since it is related to and affects many factors such as employment, consumption, and exports. In this chapter, we show BSTS-U-MIDAS can help to improve forecast accuracy, compared to ARIMA model. In practice, AR(1) and AR(4) models perform very well, and a few models can outperform them (Koopman & van der Wel, 2013). Our empirical evidence shows that a BSTS-U-MIDAS model performs very well during the 2007-2011 financial crisis period. It shows that BSTS-U-MIDAS model is more capable to predict the change points and structural breaks.

4.2 Data

We choose 5 macroeconomic variables in the regression component of our model. They are daily Toronto Stock Exchange (TSX) index and West Texas crude oil prices, monthly unemployment rate, spread between interest rates of ten years government bonds and three month treasure bill , and housing starts data. Those variable are high correlated with Canadian GDP. (Details of the data are discussed in Appendix B)

The data of real GDP for Canada and housing starts are taken from Statistics Canada. The data of the monthly unemployment rate, spread between interest rates of 10 years government bonds and treasure bill , and West Texas crude oil prices are taken from FRED database. TSX composite index is taken from Yahoo finance.

We estimate three models with different sets of data. The first model has monthly unemployment rate, spread between interest rates of 10 years government bonds and treasure bill, and TSX composite index. The second model has housing start and the data in first model. The third model has West Texas crude oil prices and the data in second model. In this section we are going to discuss the details for third model and some remarkable differences between the three models. (See the details for the first and second models in Appendix C.)

4.2.1 Target: GDP growth rate for Canada

Our target variable is the growth rate of total gross domestic product for Canada. It is a quarterly ,seasonally adjusted data which is measured by expenditure in constant prices. Figure 4.1 shows the dynamics of GDP growth for Canada between 1980 and 2015.

In order to compare performance of BSTS, ARIMA and Boosting models, we need a stationary time series data for the ARIMA, so we calculate GDP growth rate by differencing log GDP. The log differenced GDP passes the unit root and stationary test at 5% significance level.

4.2.2 Pre-process of the data: whiten data

As we have mentioned, all predictors' data are realigned to match up to the quarterly frequency by the skip-sampled method that we discuss in the previous chapter. All predictors' information is given in Appendix B.

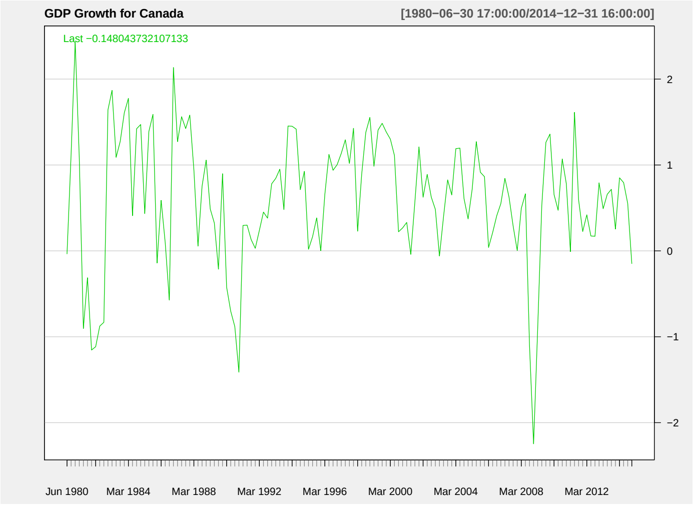


Figure 4.1: Real GDP Growth Rate (percent) for Canada 1980/07-2015/01 Quarterly

Due to the data availability, we match up all data at a window between 1980 third quarter and 2015 first quarter by using the MIDAS (mixed-data sampling) method in first and second models. The third model has the data at a window between 1986 fourth quarter and 2015 first quarter since we only have West Texas crude oil prices data down to 1986 fourth quarter. After a MIDAS skip-sampled transformation, each monthly time series data becomes a matrix with 3 quarterly time series data components, and each daily time series data becomes a matrix with 22×3 time series data components.

The data set is ragged. Different time series data have different the most recent data point. For example, for the daily data such as the TSX composite index, we choose the data up to May 28. Since Statistics Canada publishes the first quarter's GDP data on May 29, 2015, we are able to keep updating our forecasting practice using daily data until May 28, 2015. In terms of monthly data, we are able to include data up to March 2015.

For all predictors, we include leads and lags in the regression. For monthly data, we keep 23 month lags (two years). For daily data, we keep $12 \times 22 - 1$ lags (one year). Then in the third model, the total number of predictors in the regression component is 600, but we only have 114 observations. Our model is a fat regression type model. Our BSTS-U-MIDAS model does not select some specific variables. Instead, it calculates the probability of inclusion of variables and compute the posterior distribution of the prediction.

In order to avoid spurious regression, we detrend, deseasonalize, and scale the predictors using the “decomp” command in R package “forecast” (Hyndman, 2015). The target variable is seasonally adjusted. We also deseasonalize the target variable since in our case, deseasonalizing target variable improve the forecasting performance. For daily data such as TSX index and oil price, we also take first order log difference to obtain stationarity, so in some sense, the returns of TSX index and oil price are included in the models.

The rationale behind pre-processing data is to avoid the spurious regression (Scott & Varian, 2014a). Without the common trend and seasonality, the variation of the target variable can be more accurately estimated by the variation of the predictors.

4.3 Results: One Period Ahead Forecasting / Nowcasting

In our application, we use a generalized local trend model with AR(4) component and a regression component with 10000 MCMC iterations and 2000 burn-in. (See Appendix C for details). We add AR terms of target variable in the model in order to utilize the information of lagged target variables for forecasting. For simplicity, we include AR terms up to AR(4) since our target variable is quarterly data.

In a state-space presentation of BSTS-U-MIDAS model, the expected filtered “observation” is one period ahead forecast, which is estimated by the predicted state plus regression component.

$$gdp_t = F_t\alpha_t + z_t + v_t, \quad v_t \sim \mathcal{N}ID(0, V_t)$$

$$\alpha_t = G_t\alpha_{t-1} + w_t, \quad w_t \sim \mathcal{N}ID(0, W_t)$$

Let $GDP^t = (gdp_1, ..., gdp_t)$ be the vector of observation up to time t. The *filtering* distributions, $p(\alpha_t|GDP^t)$ can be computed recursively as follows, where α denotes the states which includes trend and AR term in our time structural model.:

First, we start with initial value $\alpha_0 \sim N(m_0, C_0)$ at time 0. Then we enter a prediction stage and estimate the one step forecast for the *state*:

$$\hat{\alpha_t}|GDP^{t-1} \sim N(a_t, R_t)$$

where $a_t = G_t \cdot m_{t-1}$, and $R_t = G_tC_{t-1}G_t' + W_t$. In our BSTS model, G_t is the transition or system matrix. (The details are discussed in Appendices A and C)

4.3.1 Distribution of one step ahead forecast for observations

One step forecast for the *observation* can be estimated as follows:

$$gdp_t|GDP^{t-1}, X^{1:t-1} \sim N(f_t, Q_t)$$

where $f_t = F_t \cdot a_t + z_t$, and $Q_t = F_t R_{t-1} F_t' + V_t$. F_t is the observation matrix, and $z_t = \beta \times x_t$ is regression component, and $X^{1:t-1}$ is the design matrix which includes all of the covariates. R_t is the covariance matrix for state α_t, and V_t is the covariance matrix for observation error.

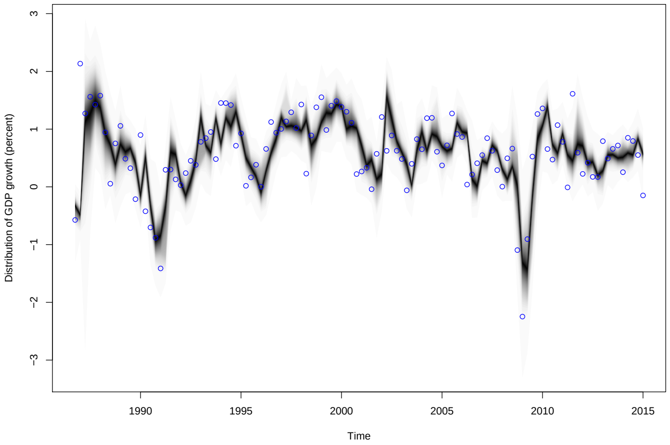


Figure 4.2: Distribution of one step ahead forecasts for the observations

Figure 4.2 shows the distribution of the one step ahead forecasts for the observations. The distribution of the one step ahead forecast for period t is made by using the value of GDP growth at $t-1$ and the observed covariates in regression component at time t. In Figure 4.2, the blue circle dots are observed GDP growth rate. The median of the distribution of forecast is colored black. To show the distribution, each 1% quantile away from the median is shaded slightly lighter, until the 99th and 1st percentiles are shaded white. Figure 4.2 shows the distribution of forecast is close to the observation except in some peculiar period such as the 2007 -2008 financial crisis period.

When we get the observation of GDP growth rate at time t, we can update the *posterior* states at time t;

$$\alpha_t | GDP^t \sim N(m_t, C_t)$$

where $m_t = a_t + R_t f_t' Q_t^{-1}(y_t - f_t)$, $y_t - f_t$ is the forecast error, and $C_t = R_t - R_t F_t' Q_t^{-1} F_t R_t$ (Petris et al., 2008).

Figure 4.3 shows that the posterior distribution of the state of the target variable GDP growth rate has much less variance than one step ahead forecast after the correction of forecast error. This is because that after we get the information of current observation for target variable, the state estimation is corrected through a weighting scheme. The weight of the correction term is given by the gain matrix $K_t = R_t f_t' Q_t^{-1}$. A close look at Figure 4.3 shows that after the correction, the posterior distribution of the state is much closer to the observation than the one step ahead forecast.

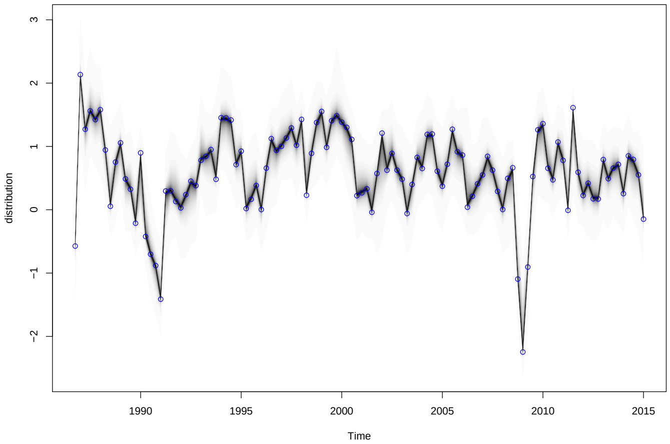


Figure 4.3: Posterior distribution of states

4.3.2 Diagnostic testing of the model

The difference also shows between the residual and forecast error. The residual is the difference between the posterior state and observation, and the forecast error is the difference between the one step ahead forecast and observation.

In BSTS model, we assume the irregular error ϵ follows normal distribution. Figure 4.4 shows that the variation of the forecast errors is wider than the variation of the residuals. Both are centered at zero.

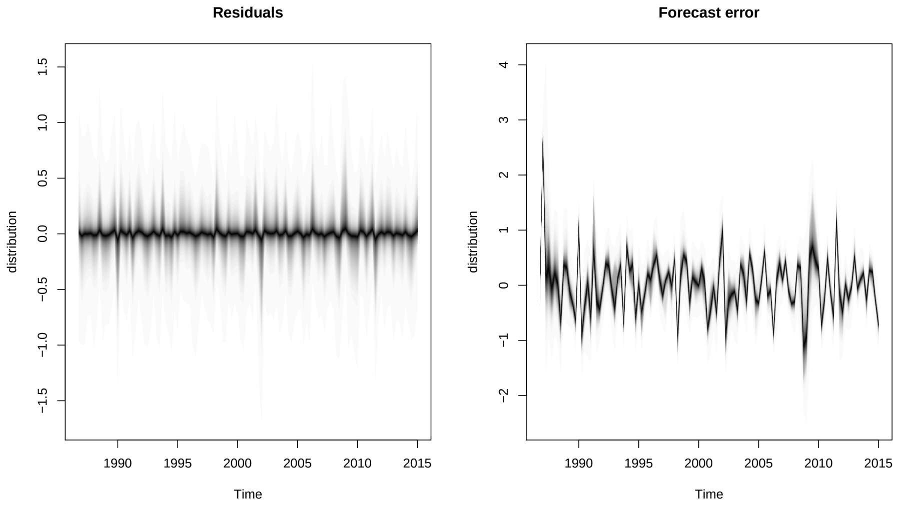


Figure 4.4: The distribution of the residual and forecast error

Figure 4.5 shows that the distribution of one step ahead forecast errors is very close to normal distribution; however, it does not pass the Shapiro-Wilk normality test. The forecast errors do pass the Box-Ljung test, and do not show strong autocorrelation, which also can be verified by the ACF diagram in Figure 4.5. Of course, it is not necessary that forecast errors follow certain normal distributions, and we report these attributes of forecast errors just to illustrate the difference between forecast errors and residuals. In a state-space representation, the forecast errors and residuals are separately calculated in observation equation and state equation.

Figure 4.6 shows that the distribution of residuals is almost normal, and it does pass Shapiro-Wilk normality test. The residuals do not pass the Box-Ljung test, and show some evidences of autocorrelation, which also can be verified by the ACF diagram in Figure 4.6.

We also plot the trace plot for forecasts in the MCMC iterations. Most plot are stable after 2000 iteration, so we choose 2000 as brun-in parameter.

4.3.3 Contributions of components for GDP growth

In our model, trend, AR terms and regression are three components besides the irregular error term. Figure 4.7 shows how much variation in GDP growth can be explained by the trend and regression components. In Figure 4.7, the AR term is relatively stable, and the trend component is more volatile. The regression component

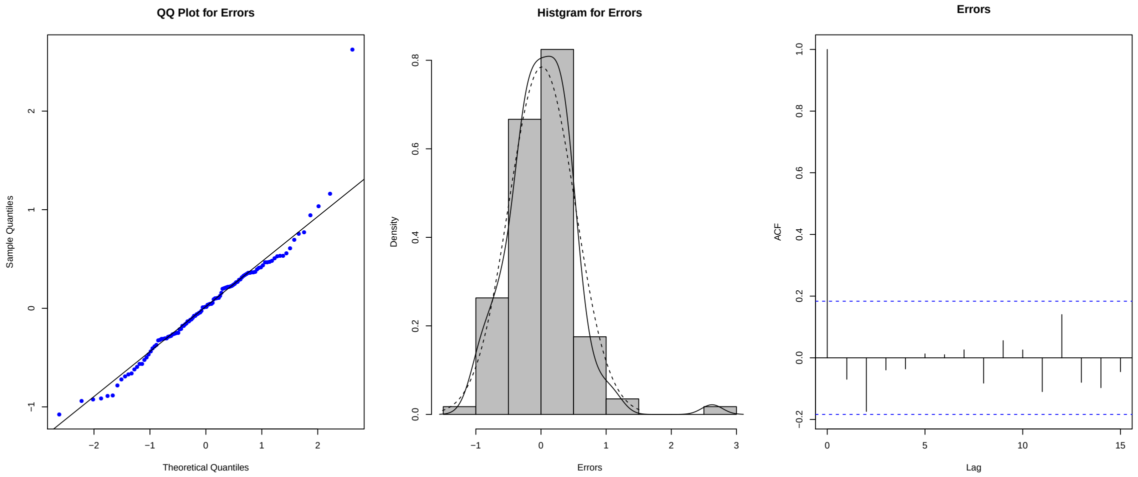


Figure 4.5: QQ plot, histogram, and ACF for one step ahead forecast errors

exhibits more variation than the trend, which can help to capture the structural break or turning points in the dynamics of the GDP growth.

4.3.4 The cumulative one step ahead forecast error for two models

In order to illustrate the value of leading macroeconomic predictors, we fit a generalized local trend and AR(4) model without regression component and compare it with our model 3. The cumulative sum of the absolute values of the one step ahead forecast errors for the two models is shown in Figure 4.8. The bottom part of Figure 4.8 shows the scaled value of the target variable GPD growth. The model 3 performs better, especially during the 2008 to 2009 financial crisis, which emphasizes one of the advantages of BSTS-U-MIDAS: it is robust to change points and structural break.

The cumulative sum of the forecast errors for model without regression component jumps when the recession starts at 2008. However, the cumulative sum of the forecast errors for model 3 increases at a constant rate. In BSTS with regression model, the signal among those leading predictors helps to capture the economic forces which drive the dramatic change in the target variable.

The robustness to structural break of the BSTS-U-MIDAS could be very helpful for us to predict recessions or boom periods.

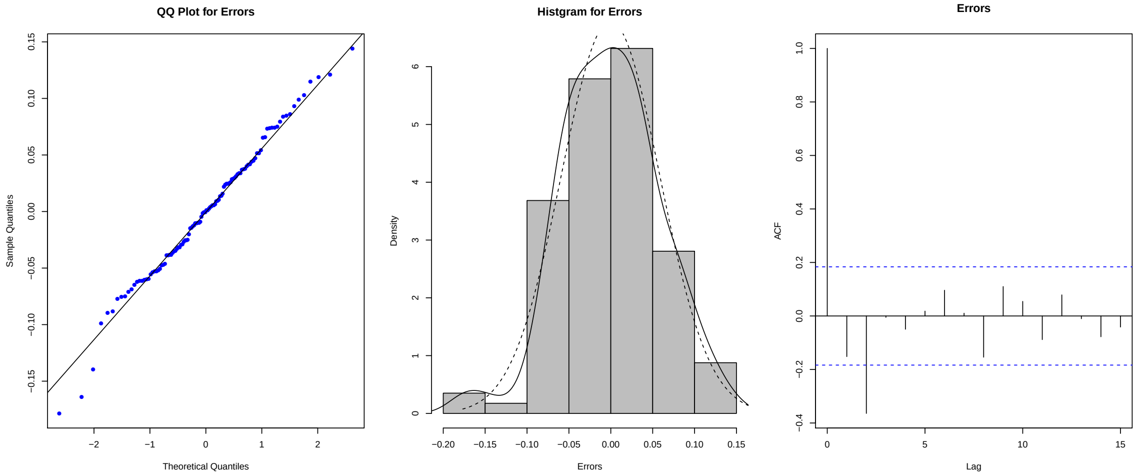


Figure 4.6: QQ plot, histogram, and ACF for residuals

4.3.5 The predictors with high inclusion probability

Another thing of interest in our forecast is to find out which predictor is significant according to its ability for helping predict the variation of GDP growth in Canada.

Figure 4.9 shows the predictors with inclusion probability in MCMC iterations greater than five percent. A white bar indicates that the predictor has a positive relationship with target variable, and a black bar indicates a negative relationship. The size of the bar represents the inclusion probability.

The unemployment rate is labeled as “labor”, TSX stock market index return is labeled as “tsx”, and oil price return is labeled as “oil”. The “labor.0.q” represent the most recent observation for the unemployment rate. In our model, “labor.0.q” is the unemployment rate for the last month in the reference quarter. “labor.6.q” is fourth month before the reference quarter.For example, the last target quarter is first quarter 2015, and then the last “labor.0.q” observation is the unemployment rate for March 2015. The last “labor.6.q” value is the unemployment rate for September 2014.

Likewise, for daily data, “oil.3.q” is oil price return (log difference of oil price) in May 25, 2015, three days before May 28, 2015. “tsx.61.q” is stock market return sixty one business days before May 28, 2015, one day in March 2015. The locations of predictors with high probability on time table are illustrated in Figure 4.10. The background color shows the availability of the data. A white background represents the observation in that period is not available, and a yellow background shows that

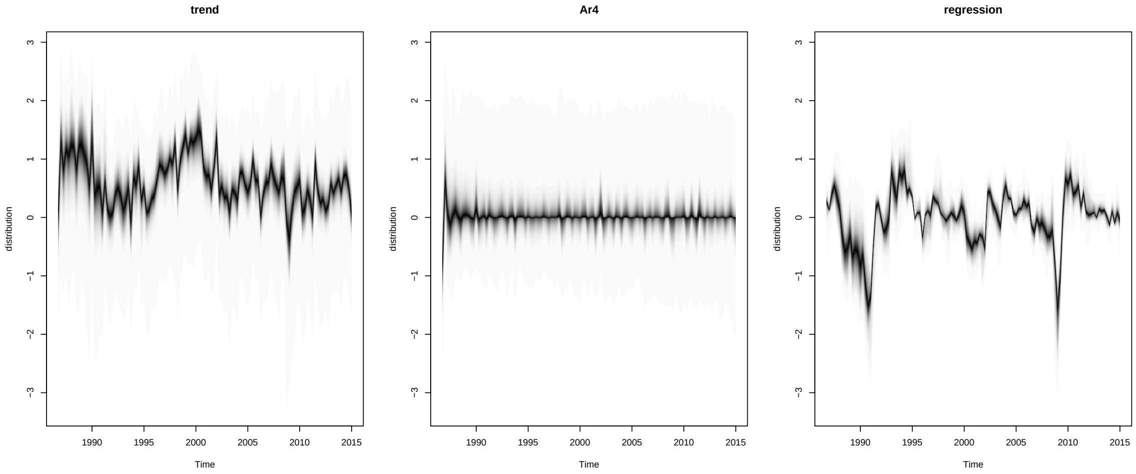


Figure 4.7: Contributions to state for GDP growth

observation is available.

The ragged yellow cells demonstrate the different publication lags for those variables. Since the publication lag for GDP in Canada is about two months, we get chance to utilize the heading high frequency data extending to two monthes. For example, Statistics Canada released first quarter's GDP on May 30, so we can practice forecasting for first quarter's GDP till May 29. That is why the daily data are available on time table till the end of May. In this sense, our practice can be called now-cast since we forecast a figure which is already there but has not been observed by using more recent high frequency data.

The boldface number in green cell means the lagged numbers of predictors and this predictor has a positive relationship with GDP growth. The italicized and underlined number in shaded cell means the lagged numbers and this predictor has a negative relationship with GDP growth. For example, the “labor.0” and “labor.1” have a negative relation with GDP growth.

The signs of the coefficients of the predictors mostly meet our anticipation based on economic theory. For example, the stock market return is positively related to GDP growth. However, some others do not. For example, the top predictor “labor.6.q” is positively related to GDP growth. If we take the MIDAS into account, and we will anticipate a certain combination of high frequency data should work as a predictor for a low frequency counter part. As we see in Figure 4.9 and Figure 4.10, the “labor.0” and “labor.1” in March and February 2015 combined with “labor.6” and “labor.7” in September and August 2014 help us predict GDP growth. Specifically, the scenario,

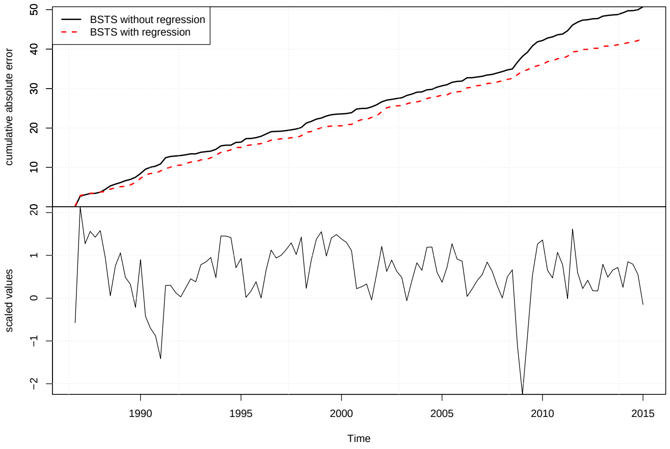


Figure 4.8: The cumulative absolute one step ahead forecast error

that unemployment rates were low two quarters ago and high in this quarter, means the GDP growth in this quarter is going to be low. It turns out that one single high frequency time series is not sufficient for forecasting. Only the combination of several time series works. This phenomenon also shows in Figure 4.11.

Figure 4.11 shows the contribution of each state component and predictor to the fit of model. Variables are ordered by probability of inclusion. The mean absolute error in terms of the posterior distribution of the state of GDP growth is given in the title. The faint line in each panel is the previous fit, and the residuals are shown at the bottom of each panel.

It is interesting that after including "labor.6" in model, the MAE and residual increase. The MAE and residuals continue decreasing after "labor.1" included. This phenomenon shows that individually including one specific predictor may not improve forecasting, but jointly including a set of predictors might improve forecasting a lot.

Another thing we can take away from Figure 4.9 is that BSTS-U-MIDAS gives us a sparse model. The top two predictors have over 80 percent inclusion probability in MCMC iterations. Others have less probability. This pattern also is verified by the distribution of the size of model in Figure 4.12

Figure 4.12 shows the posterior distribution of the number of predictors in the model. We have 600 variables to choose from in model 3. The median number of the distribution of model size in model 3 is 3. The largest model in the MCMC sample has 6 predictors. The BSTS-U-MIDAS model indeed gives us sparsity.

In the top 100 predictors with high inclusion probability, there is no housing

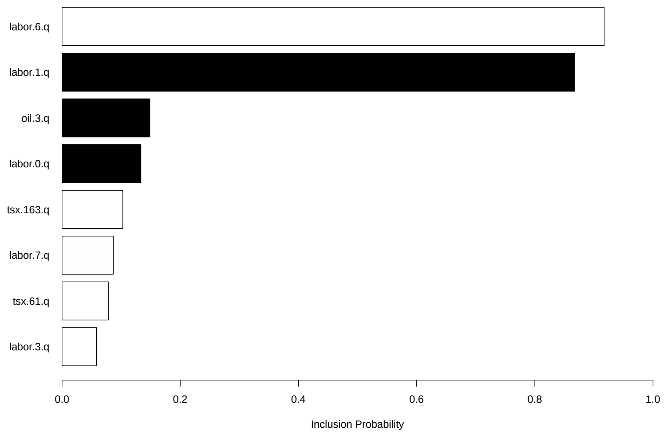
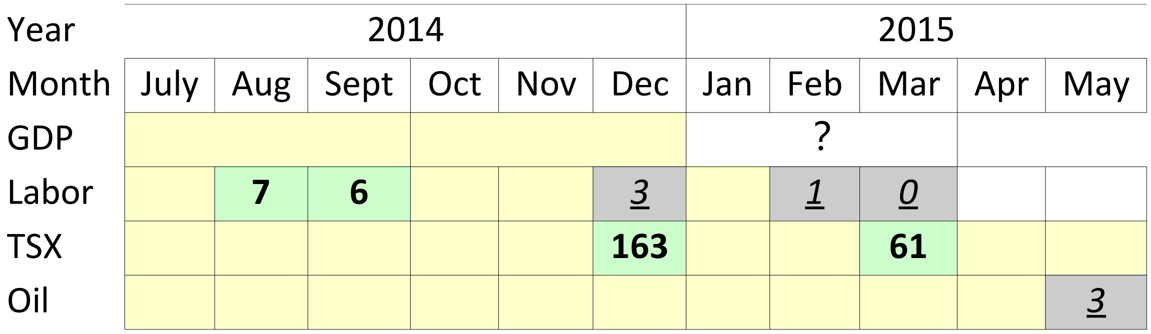


Figure 4.9: Predictors with inclusion probability greater than 5%



Year	2014						2015				
Month	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
GDP							?				
Labor		**7**	**6**			*3*		*1*	*0*		
TSX						**163**			**61**		
Oil											*3*

Figure 4.10: Predictors with high inclusion probability on time table

start data. One possible reason is that the housing data we have collected are not representative, so they are not included in many MCMC iterations. However, they are not absolutely excluded from the model. Some variables from “house” family still have 0.01 percent inclusion probability. Through the Bayesian model averaging, the information in “spread” and “house” family still will be conveyed into the final model, so they still have impact on the final forecasts.

Figure 4.13 shows the predictors with inclusion probability greater than 5% in model 1 and model 2. The composition of top predictors changes after the housing starts data is included in model 2. Even the model size also changes, which is shown in Figure 4.14.

Suppose the housing data is noisy in some sense. After the housing data is included in the model, the inclusion probability of the top predictors decreases, and the size of the model decreases. Every single model in the MCMC iterations becomes more

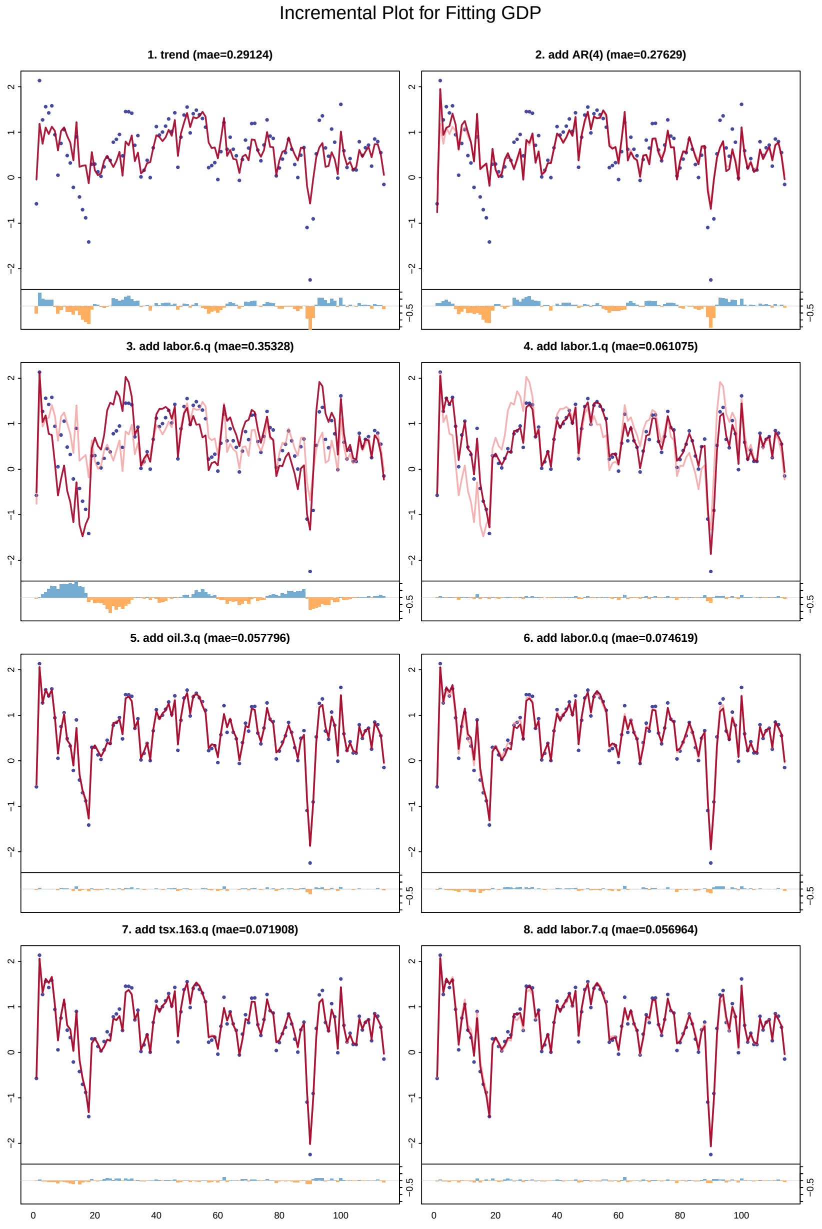


Figure 4.11: Decomposition of forecast for GDP growth

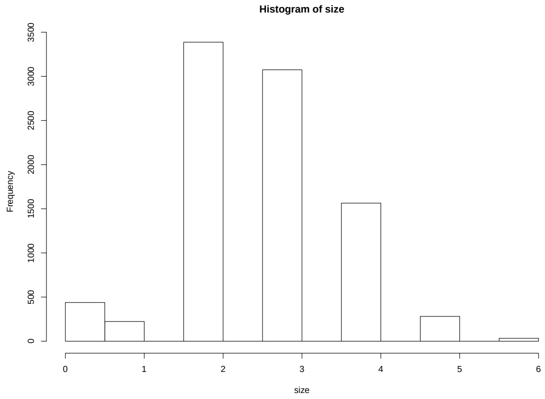


Figure 4.12: The posterior distribution of model size

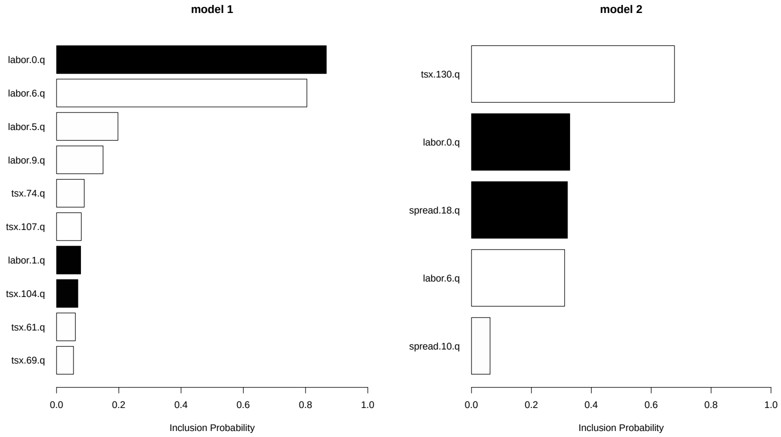


Figure 4.13: Predictors with inclusion probability greater than 5%

sparse and weaker in terms of forecast ability. It shows BSTS-U-MIDAS model can suffer from noisy or redundant data. However, when we include the oil return data into model 3, the performance of model 3 gets better, even better than model 1. The reason might be because when more relevant predictors are included in the model, the inclusion probability of less relevant predictors are pushed towards zero.

We select second quarter in 2003 to first quarter in 2015 as our forecast horizon. We compare the performance of three models in terms of GDP, which is shown in Table 4.1. Model 2 with housing starts data performs slightly worse than model 1. Model 3 with housing starts and oil return data performs best.

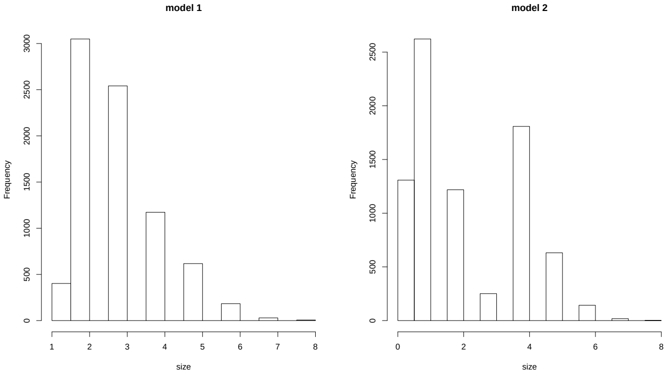


Figure 4.14: The distributions of the size of model 1 and model 2

	ME	RMSE	MAE	MPE	MAPE
Model 1	-219.327	7548.331	6100.694	-0.012	0.386
Model 2	-256.496	8161.589	6625.706	-0.013	0.419
Model 3	-159.853	6952.102	5641.339	-0.008	0.356

Table 4.1: Comparison of forecast error 2003-2015 in model 1, 2, and 3

4.4 Comparison Between BSTS, ARIMA, and Boosting Model

We also compare one step ahead forecast performance between BSTS-U-MIDAS, ARIMA and Boosting models. ARIMA is a benchmark which is a normal practice in the forecasting business. The Boosting model has many advantage similar to BSTS-U-MIDAS. It can deal with a large numbers of predictors and is robust to over-fitting. It ensembles many weak classifiers or regression trees to get a good prediction, which is similar to Bayesian model averaging in the BSTS-U-MIDAS model. Boosting model allows us to model complex relationships among the large data set by using more flexible model (non-linear) instead of simple linear models(Varian, 2014).

The ARIMA model is implemented by automatic ‘arima’ algorithm of the Forecast package in R (Hyndman & Khandakar, 2008). The Boosting model is implemented by the GBM package in R (Ridgeway, 2015). All forecast are one step ahead forecasts which are estimated in a recursive way. We refit the model every time when the forecast horizon extends to the next period. (See appendix D for details.)

We take data from third quarter in 1980 to first quarter in 2003 as the training period and second quarter in 2003 to first quarter in 2015 as the test period. We predict the GDP growth rate and the level of GDP for Canada from second quarter 2003 to first quarter 2015. For visual comparison, we use GDP growth rate since the level of GDP changes a lot from the second quarter in 2003 to first quarter in 2015.

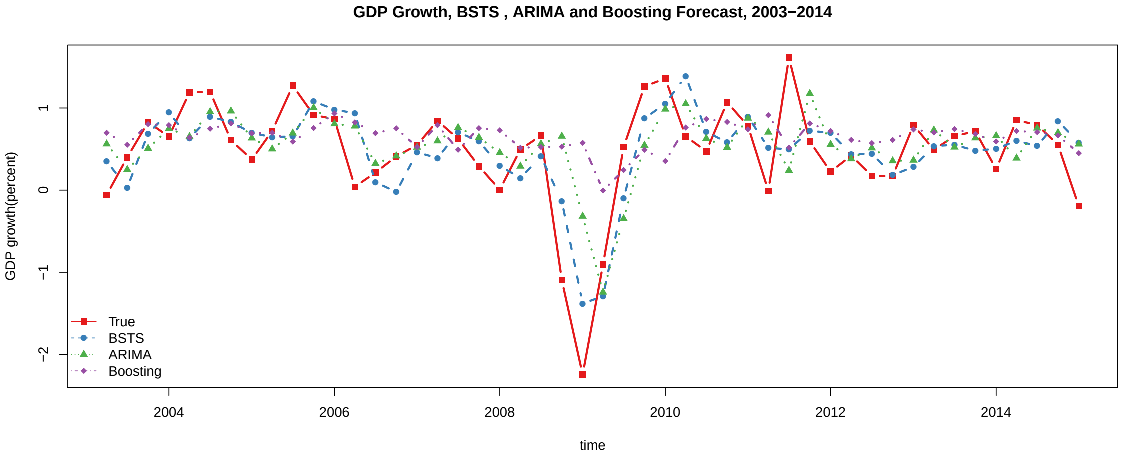


Figure 4.15: GDP Growth, BSTS, ARIMA and Boosting Forecasts 2003-2015

Figure 4.15 shows that the forecasts are very similar except during the 2008 to 2009 financial crisis. The one step ahead forecasts of ARIMA look like first order lag of target time series. The one step ahead forecasts of Boosting are very stable and look more like smoothed target time series. The one step ahead forecasts of BSTS are more volatile and do a better job at capturing the structural break during the 2008 to 2009 financial crisis. As we mentioned before, the signal among those leading predictors helps BSTS to predict the economic force which drives dramatic change in the target time series.

We also calculate the forecast error in terms of the level of GDP. Table 4.2 shows that BSTS model is slightly better than ARIMA and the Boosting model according to the regular criteria.

	ME	RMSE	MAE	MPE	MAPE
BSTS	-159.853	6952.102	5641.339	-0.008	0.356
ARIMA	-1089.482	8996.830	6347.325	-0.067	0.401
Boosting	-792.281	8843.832	6349.097	-0.049	0.405

Table 4.2: Comparison of forecast error 2003-2015

We also conduct the Diebold-Mariano test in which BSTS vs. ARIMA with the null hypothesis that the two methods have the same forecast accuracy and alternative hypothesis that ARIMA is less accurate than BSTS. The p-value is 0.061. We fail to reject the null hypothesis at the 5% significance level. Similarly, we conduct the Diebold-Mariano Test in which BSTS vs. Boosting with the null hypothesis that the two methods have the same forecast accuracy and alternative hypothesis that Boosting is less accurate than BSTS. The p-value is 0.084. We fail to reject null hypothesis at the 5% significance level. In both cases, at the 10% significance level, we will reject null hypothesis in favor of the alternatives that the forecasts of the ARIMA and Boosting models are less accurate than forecasts of the BSTS model.

In order to inquire as to the performance of the three models in different scenarios, we separate the forecast horizon into three periods. We call the first period 2003 - 2007 the pre-crisis period; second period 2007-2011 the crisis period; and third period 2011-2015 the post-crisis period. The evidence shows that the performances of these three methods change during those periods.

4.4.1 Comparison of one step ahead forecasts, 2003 to 2007

During 2003 to 2007, the ARIMA model outperforms the other models. In Table 4.3, we can see that the performances of BSTS and Boosting are very close.

	ME	RMSE	MAE	MPE	MAPE
BSTS	228.032	5748.218	4713.560	0.016	0.318
ARIMA	-163.452	5199.532	4023.557	-0.010	0.273
Boosting	-1263.461	6078.649	4748.083	-0.084	0.320

Table 4.3: Comparison of forecast error 2003-2007

Figure 4.16 shows that the ARIMA model outperforms the other two models. The BSTS model does not show anyadvantage over ARIMA

4.4.2 Comparison of one step ahead forecasts, 2007 to 2011

During the crisis period, BSTS performs much better than ARIMA and Boosting. It confirms Petris et al. (2008)'s statement that a state-space model can capture the structural break or change point better than an ARIMA model.

During 2007 to 2011, the financial crisis hindered economic development in many ways, Canadian GDP plunged as did other countries' GDP. In Figure 4.17, the ARIMA and Boosting models detect the decreasing in GDP later than BSTS. BSTS

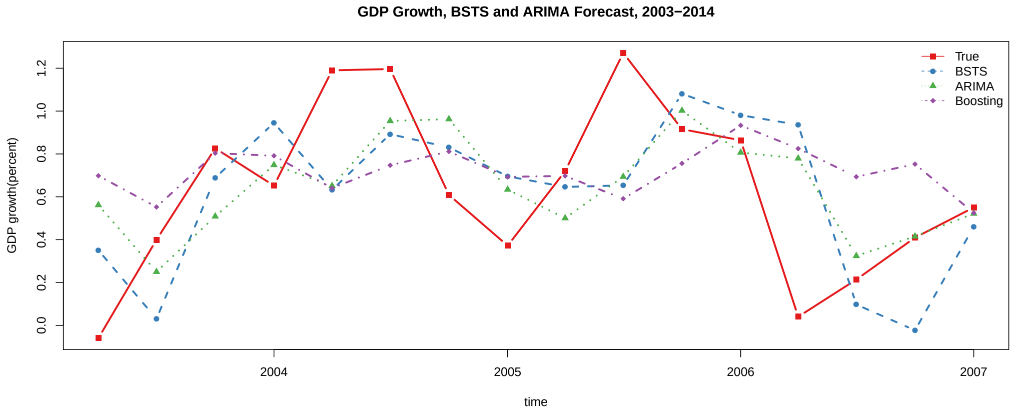


Figure 4.16: GDP Growth, BSTS, ARIMA and Boosting Forecast, 2003-2007

	ME	RMSE	MAE	MPE	MAPE
BSTS	-384.114	7754.361	6719.998	-0.023	0.427
ARIMA	-1909.907	12000.140	8508.274	-0.121	0.543
Boosting	-4373.490	14798.320	9544.599	-0.281	0.610

Table 4.4: Comparison of forecast error 2007-2011

does not only use a state-space model to capture the dynamics of GDP, but also gets the signal of the structural change from the leading predictor such as stock market return or unemployment rate.

Figure 4.17 shows that the BSTS model outperforms other two models. Especially, during 2008, the GDP growth rate decreases abruptly, but ARIMA and Boosting models still predict GDP will keep stable. Only BSTS model forecasts the direction of change in GDP growth rate correctly, and the forecasted magnitude of the change in GDP growth rate is close to the true value as well.

4.4.3 Comparison of one step ahead forecasts, 2011 to 2015

During 2011 to 2015, the ARIMA model performs a little worse than the BSTS and Boosting models. Table 4.5 shows that the performance of BSTS just is a little better than the Boosting model.

Figure 4.18 shows that the forecast of ARIMA model is more like a first order lag of the target variable. The ARIMA model shows this pattern during the whole test period. The forecast of THE BSTS model is more volatile than during 2003 to 2007,

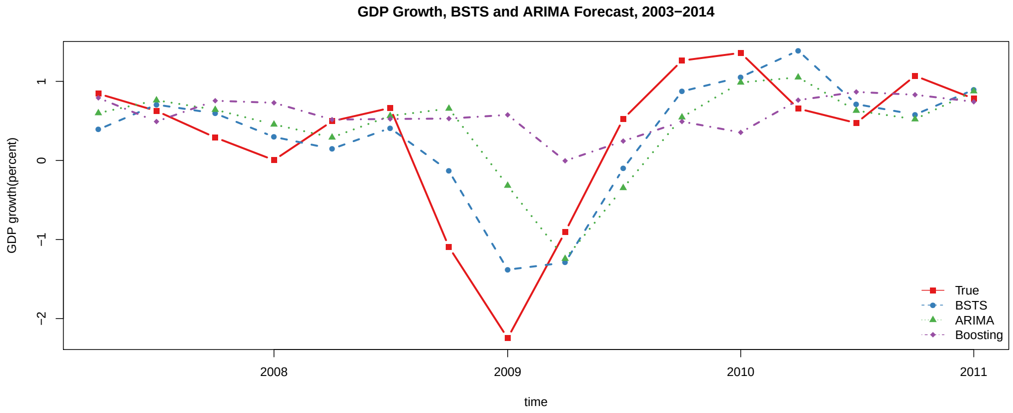


Figure 4.17: GDP Growth, BSTS, ARIMA and Boosting Forecasts, 2007-2011

	ME	RMSE	MAE	MPE	MAPE
BSTS	-321.159	7198.822	5490.460	-0.017	0.324
ARIMA	-1195.086	8472.912	6510.144	-0.069	0.385
Boosting	-2773.160	7512.228	5656.435	-0.164	0.336

Table 4.5: Comparison of forecast error 2011-2015

and its dynamics are closer to the dynamics of the target variable. The forecast of the Boosting model is more steady as it does in entire test period.

The performances of the three model change during the three different test periods. ARIMA does best in the first period and worst in the last period. BSTS does best in crisis period. Boosting does worst in the first period and does well in the last period. One potential reason is that including more data in the model when the time dimension expands helps improve the performance of BSTS and the Boosting model. The Boosting model ensembles many weak classifiers and is robust to over-fitting. Based on similar reasons, the BSTS model is getting better in the last test period. The ARIMA model is a univariate model, and when the time coverage extends, the increase in number of data points is very limited. Then the performance improves very slowly.

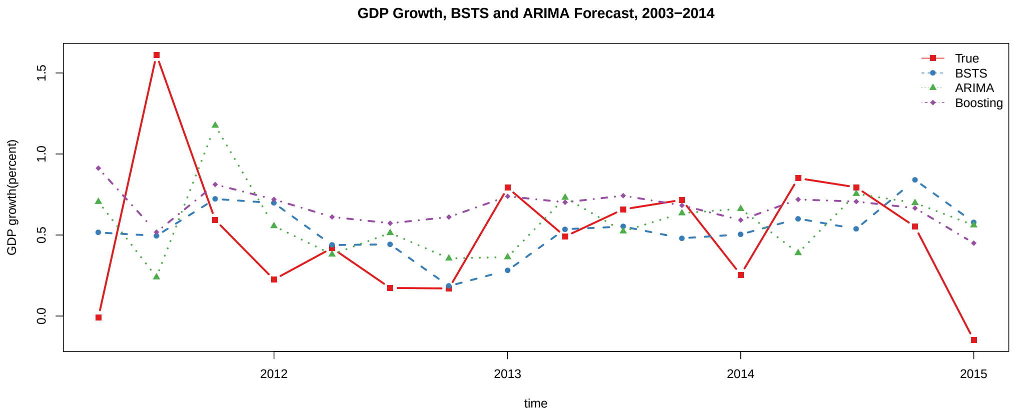


Figure 4.18: GDP Growth, BSTS, ARIMA and Boosting Forecasts, 2011-2015

Chapter 5

Conclusions

We discuss a new method to exploit the signals in high frequency data to improve the forecasts for low frequency data. This BSTS-U-MIDAS model merges a structural time series model, spike-and-slab prior, Bayesian model averaging, and the Mixed-data sampling method. It utilizes the high frequency data by mixed-data sampling, and reduces the parameter proliferation problem by a spike-and-slab prior. Then a Bayesian structural time series model decomposes the target time series into trend, AR process, regression components and irregular term. By using the Kalman filter, we estimate the states and parameters in a state-space representation of the model. Last, we estimate the posterior distribution of the forecasts by drawing from the posterior distribution of the parameters many time.

Our empirical application shows that BSTS-U-MIDAS model not only improves the all over forecast accuracy, but also is capable of capturing the structural breaks or turning points. It is good at dealing with high dimension data, and especially at incorporating high frequency data. Furthermore, it is robust to irrelevant or redundant variables even though it does suffer somewhat from noisy data.

In conclusion, BSTS-U-MIDAS is flexible for handling high frequency and ragged data. It does not require stationarity of the time series. It does not require pre-processing of the data although the detrending and deseasonalizing do improve the forecast performance. It is easy to implement on a daily or weekly basis until the observation of the target variable is published. And due to the recursive algorithm, it is easy to incorporate new information for predictors to update the forecast.

Our understanding of the BSTS-U-MIDAS model still has a long way to go. The most important question is simulation based evidence. Since we only have empirical evidence, to test the model with a simulated data set will be very valuable. In addition, there are many other things to be investigated.

First, it is possible to use data-based prior for estimation of the state and observation variances. For example, we can use a two-step strategy to get the prior for estimation. We can use the longer history data for GDP to run univariate state-space model; and we can use the posterior distribution as the prior in the model with regression in the second stage.

Second, since the BSTS-U-MIDAS model still suffers from noisy data, we can choose better predictors based on previous research. For example, we can include monthly GDP, industrial production, consumption, PPI, or CPI. Also instead of stock market index return, we can model the market volatility of the financial time series instead of level of the time series by using a GARCH model.

Third, we suppose that the coefficients of the regression components are time invariant. We may be better off to set a time-varying model even though it will increase the complexity of the computation.

Appendix A

State-Space Model with Kalman Filter

We adopt the notation from (Petris et al., 2008) and modify it for convenience.

A.1 State-Space Model

The linear Gaussian state-space model is defined in three parts:

Observation/measurement equation

The observations uncertainty $p(y_t \mid \alpha_t, \theta)$ described by the observation equation

$$y_t = F_t \alpha_t + v_t, \quad v_t \sim \mathcal{N}ID(0, V_t)\,.$$

State/transition/process/model equation

The process uncertainty of the unknown states α_t and their evolution given by the state equations as $p(\alpha_t \mid y_t \theta)$.

$$\alpha_t = G_t \alpha_{t-1} + B_t u_t + w_t, \quad w_t \sim \mathcal{N}ID(0, W_t)\,.$$

Parameters to estimated $p(\theta)$

$$\theta : \alpha_t, F_t, G_t, V_t, W_t\,.$$

The initial state distribution

$$\alpha_0 \sim \mathcal{N}(m_0, C_0)\,,$$

And then we have: **The transition probabilities of state from time $t-1$ to t**

$$\alpha_t|\alpha_{t-1} \sim \mathcal{N}(G_t \cdot \alpha_{t-1}; W_t)$$

The conditional probability of the variable y at time t; given that the state of the system, α at time t

$$y_t|\alpha_t \sim \mathcal{N}(F_t \cdot \alpha_t; V_t)$$

where:

- y_t is the observation of target variable at time t , with $t = 1, ..., T$.

- α_t is state vector (length m), the unobserved / hidden states of the system. In time series setting, α_t will have various components such as trend, seasonality, etc.

- G_t is state transition matrix, an $(m \times m)$ matrix, which is applied to the previous state α_{t-1}. The unobserved state x_t evolves in time according to the G_t matrix.

- F_t is the observation model an ($m \times 1$) matrix/vector which transforms the true state space into the observed space in our structural time series model.

- v_t is the measurement errors (observation noise) which are assumed to be zero mean Gaussian white noise with covariance matrix V_t. In our structural time series model, it is a scalar/constant ;

- w_t is the state innovations (the process noise) which are assumed to be drawn from a zero mean multivariate normal distribution with covariance matrix W_t.

- B_t is the model that predicts what changes based on control/commands.

- u_t is the control / commands/ input in time t. In our model, we have not included any control or intervention policy, so no B and u terms are in our model.

- m_0 is the mean of the initial state; C_0 is the variance of the initial state.

The initial state, and the noise vectors at each step $\alpha_0, w_1, ..., w_t, v_1 ... v_t$ are all assumed to be mutually *independent*.

A.2 Kalman Filter

The Kalman Filter has two stages:

1. Predicting the new state and its uncertainty.

2. Correcting with the new measurement.

A.2.1 The state of Kalman Filter

The state of the filter is represented by two variables:conditional mean(expected value) and covariance.

$$p(\alpha_t \mid y_t) \sim \mathcal{N}(m_t, C_t)$$

1. m_t, the posterior state estimate at time t given observations up to and including those at time t;

2. C_t, the posterior error covariance matrix (a measure of the estimated accuracy of the state estimate). It reflects the variance of the state distribution.

$$C_t = \mathrm{cov}(\alpha_t - m_t)$$

A.2.2 Two phases: Predict and Correct

We start with $\theta_0 \sim N(m_0, C_0)$ at time 0.

Predict/(state, error)stage/ (time update)

The Predict/time update projects the current state estimate ahead in time.

$$a_{t|t-1} = \mathbf{G}_t a_{t-1|t-1} + \mathbf{B}_t \mathbf{u}_t$$

$$\mathbf{R}_{t|t-1} = \mathbf{G}_t \mathbf{C}_{t-1|t-1} \mathbf{G}_t^{\mathrm{T}} + \mathbf{W}_t$$

where

$a_{t|t-1}$ is one step ahead prediction for (a prior) state. B and u are control or intervention policies and can be chosen by people to change the state α. $a_{t|t-1}$ is the estimate in next state α_t; $m_{t-1|t-1}$ is the estimated state in the last state α_{t-1}. The initial state α_0 is known.

$\mathbf{R}_{t|t-1}$ is covariance of one step ahead prediction for (a prior) state. $\mathbf{C}_{t-1|t-1}$ is the previous error covariance at time $t-1$.

$\mathbf{W}$: the covariance matrix of the error noise.

Then we have:

$$\alpha_t | y^{t-1} \sim N(a_t, R_t)$$

where y^{t-1} is $\{y_0, y_1, y_2, ..., y_t, ...\}$ up to time $t-1$.

We also can get one step ahead prediction for the observation f_t:

$$f_t = F_t \cdot a_t$$

$$Q_t = F_t R_{t-1} F_t' + V_t$$

And we have

$$y_t | y^{t-1} \sim N(f_t, Q_t)$$

Correct /Update stage/ measurement update

The Correct/measurement update adjusts the projected estimate by an actual measurement at that time.

Updated (a posteriori) state estimate $m_{t|t}$ with new observation y_t. $m_{t|t}$ is a weighted average of latest estimate and gain from observation.

$$m_{t|t} = a_{t|t-1} + \mathbf{K}_t \tilde{v}_t$$

Updated (a posteriori) estimate covariance $\mathbf{C}_{t|t}$

$$\mathbf{C}_{t|t} = (I - \mathbf{K}_t \mathbf{F}_t) \mathbf{R}_{t|t-1}$$

Then we have posterior of state at time t:

$$\alpha_t | y^t \sim N(m_t, C_t)$$

Where $\tilde{v}_t$ is **prediction error**, and K is **Optimal Kalman gain**.

When the **prediction error** $\tilde{v}_t$ is non-zero, there is new information about the system, so the state α should be modified. $\tilde{v}_t$ is also called Measurement innovation or the predictive residual.

The prediction error $\tilde{v}_t$ reflects the discrepancy between the predicted measurement $E(y_{t|t-1})$ and the actual measurement y_t. In our case, the prediction error is a scalar $\tilde{v}_t$:

$$\begin{aligned}\tilde{v}_t &= y_t - E(y_{t|t-1})\\ &= y_t - F_t E(\alpha_{t|t-1})\\ &= y_t - f_t\,,\end{aligned}$$

and the covariance matrix of $\tilde{v}_t$ is

$$\mathbf{S}_t = \mathrm{cov}(\tilde{v}_t) = \mathbf{F}_t \mathbf{R}_{t|t-1} \mathbf{F}_t^{\mathrm{T}} + \mathbf{V}_t\,,$$

where V_t describes the noise in the observation $\mathbf{y}_t$.

The contribution of $\tilde{v}_t$ to the state vector is weighted by the **Optimal Kalman gain $\mathbf{K}_t$**. $\mathbf{K}_t$ measures how much to trust a new observation y_t. It can be a matrix, and in our case it is a vector.

An **Optimal Kalman gain $\mathbf{K}_t$** is chosen to be the gain or blending factor so that we minimize the a posterior estimation error covariance $\mathbf{C}$:

$$\mathbf{K}_t = \mathbf{R}_{t|t-1} \mathbf{F}_t^T \mathbf{S}_t^{-1}$$

where $\mathbf{S}_t$ is prediction error covariance. $\mathbf{F}_t$ describes how the observation reflects the state in the model (a function of how much influence goes from observation to state vector).

A.2.3 Optimal Kalman gain

The Kalman filter is a minimum mean-square error estimator. The error in the posterior state estimation is

$$\alpha_t - m_{t|t}\,.$$

We seek to minimize the expected value of the square of the magnitude of this vector, $\mathrm{E}[\|\alpha_t - m_{t|t}\|^2]$. This is equivalent to minimizing the trace of the a posteriori estimate of the covariance matrix $\mathbf{C}_{t|t}$. Solving this optimization problem for $\mathbf{C}_{t|t}$ yields the optimal Kalman gain $\mathbf{K}_t$:

$$\mathbf{K}_t \mathbf{S}_t = (\mathbf{F}_t \mathbf{R}_{t|t-1})^{\mathrm{T}} = \mathbf{R}_{t|t-1} \mathbf{F}_k^{\mathrm{T}}$$

$$\mathbf{K}_t = \mathbf{R}_{t|t-1} \mathbf{F}_t^{\mathrm{T}} \mathbf{S}_t^{-1}\,.$$

This gain is the optimal Kalman gain and yields MMSE (minimum mean square error) estimates.

Appendix B

Data of Predictors

Predictors are all seasonally adjusted, detrended, and scaled. They are all stationary, as determined by the ADF and KPSS tests, at the 5% significance level.

B.1 Unemployment rate: Monthly, 1980/07-2015/03

The unemployment rate data is monthly seasonally adjusted data which covers all aged 15 and over in the labor force in Canada. The data is taken from St. Louis Fed. website. The data has been detrended, deseasonalized and scaled, and then it is skipping sampled as was described in section 3.2.2.

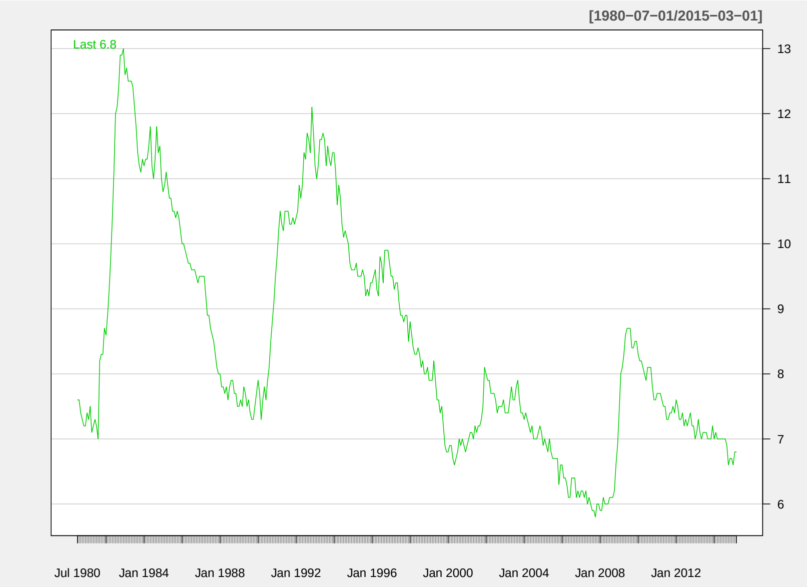


Figure B.1: Unemployment Rate for Canada

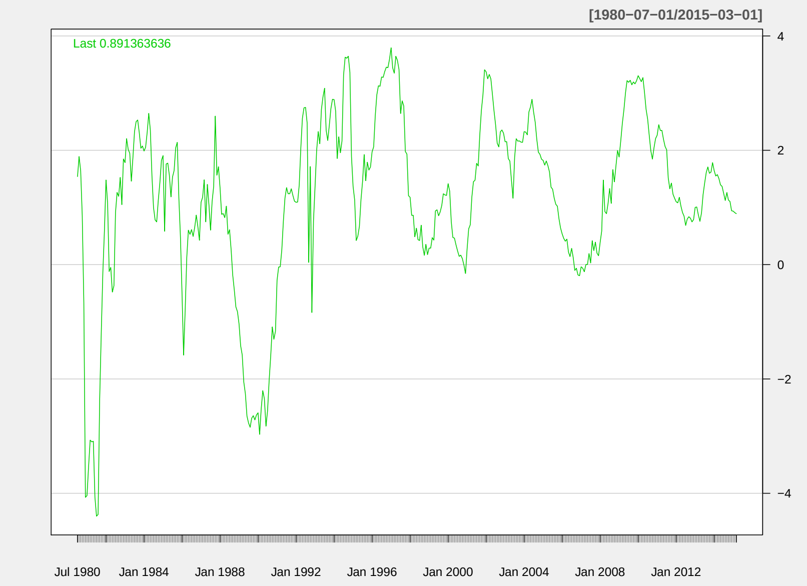


Figure B.2: Spread between 3 month and 10 year government bond interest rate

B.2 Spread between 3 months treasury bills yields and 10 year's government bond yields. Monthly, 1980/07 - 2015/03

The spread is calculated by subtracting interest rates of 3 month treasury bills for Canada from interest rate of 10 year's government bond for Canada. Both data is taken from St Louis Fed website.

It is detrended, deseasonalized and scaled. And then it is skipping sampled was described in section 3.2.2.

B.3 Toronto Stock Exchange (S&P/TSX) Composite index 1980/07/01-2015/05/28

S&P/TSX Composite index is daily and is taken from St. Louis Fed website.

Before included in our model, it is transformed to log difference and is detrended, deseasonalized and scaled to achieve stationary.

In order to match up the quarterly target variable, it is skipping sampled was described in section 3.2.2.

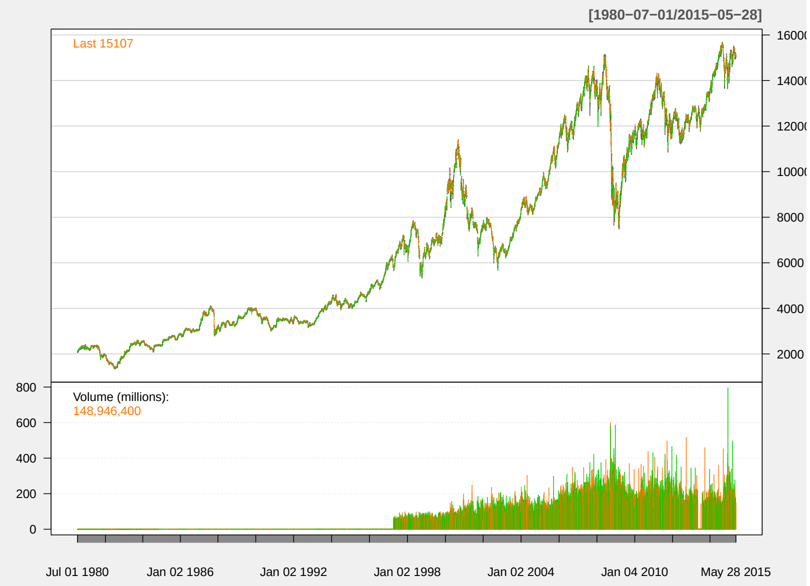


Figure B.3: TSX stock market index

B.4 Housing starts, monthly, 1980/07-2015/03

The housing starts data is from Canada Mortgage and Housing Corporation. It covers housing under construction and completions in centres with 10,000 and over population in selected census metropolitan areas.

The data is detrended, deseasonalized and scaled, and then it is skipping sampled was described in section 3.2.2.

B.5 Crude Oil Price: West Texas Intermediate (WTI), daily, 1980/07/01-2015/05/28

The oil price data is daily and is taken from St Louis Fed website.

Before included in our model, it is transformed to log difference and is detrended, deseasonalized and scaled to achieve stationary.

In order to match up the quarterly target variable, it is skipping sampled was described in section 3.2.2.

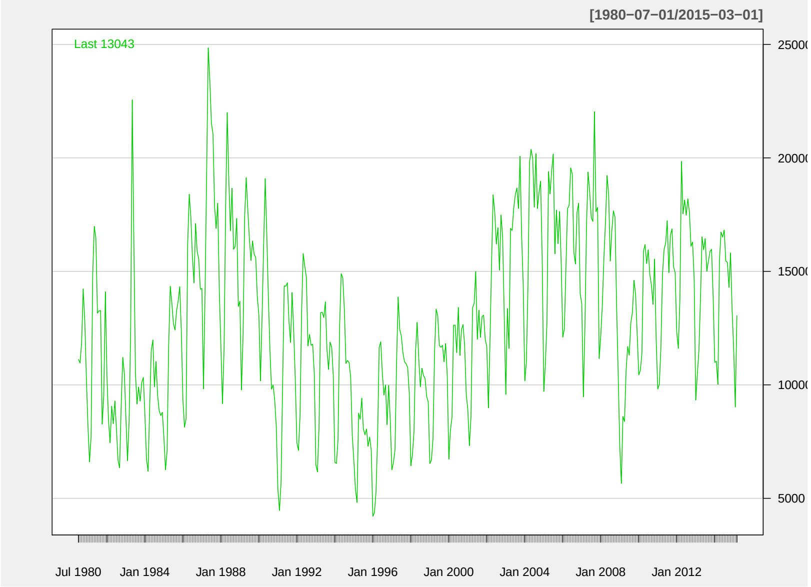


Figure B.4: Housing starts

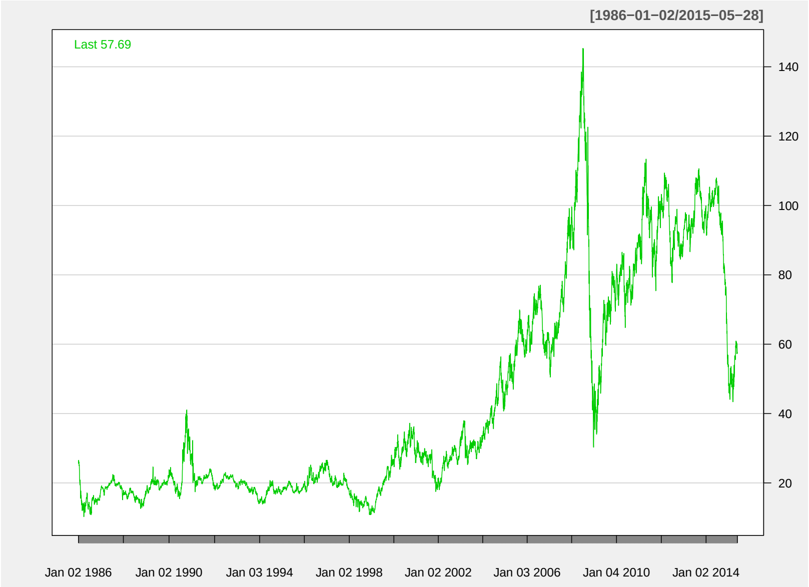


Figure B.5: Crude Oil Price

Appendix C

Model Specification

C.1 Generalized local trend model with regression

The generalized local linear trend model is more complicated than a regular local trend model. In our practice, we also add a ar(t) term in the equation.

Observation equation (level + regression):

$$y_t = \mu_t + c_t + z_t + v_t, \quad v_t \sim N(0, V)\,.$$

It assumes the level moves according to a random walk, but the slope moves according to an AR(1) process centered on some potentially nonzero value D. The equation for the mean is

State equation 1 (random walk + trend):

$$\mu_{t+1} = \mu_t + b_t + w_{1t}, \quad w_{1t} \sim N(0, W_1)\,.$$

The equation for the slope b is

State equation 2 (AR(1) for trend):

$$b_{t+1} = D + \phi(b_t - D) + w_{2t}, \quad w_{2t} \sim N(0, W_2)\,.$$

The prior distribution for this model has four independent components. There is an inverse gamma prior on the level standard deviation, sigma.level; an inverse gamma prior on the slope standard deviation sigma.slope; a Gaussian prior on the long run slope parameter D; and a potentially truncated Gaussian prior on the AR(1) coefficient ϕ.

The slope exhibits short term stationary variation around the long run slope D when absolute value of the prior on ϕ is less than 1 (Scott, 2015). The parameter ϕ represents the learning rate at which the local trend is updated. Hence, the model balances short-term information $(b_t - D)$ with information D from the distant past.

State equation 3 (dynamics of ar(4) process):

$$c_t = \psi_1 c_{t-1} + \psi_2 c_{t-2} + \psi_3 c_{t-3} + \psi_4 c_{t-4}, \quad w_{3t} \sim N(0, W_3)\,.$$

Regression component:

$$z_t = \beta x_t$$

we can append a constant 1 of the state vector α and append z_t to observation matrix F. By doing so, we only increase the dimension of state vector α by one.

C.2 ARIMA model

We use Hyndman's "forecast" package in R to fit ARIMA models and estimate one step ahead forecast (Hyndman, 2015). The ARIMA models are chosen according to either AIC, AICc or BIC value in the "auto.arima" function in "forecast" package. In the test period 2003 to 2015, there are 48 models. The first 35 models selected by "auto.arima" are AR(1) with an intercept. The last 13 models are AR(2) with an intercept.

Table C.1 shows that the fitted models chosen by "auto.arima" function do not change much when the we move forward during the test period 2003 to 2015. Since an ARIMA model is univariate model, one extra data point usually does not provide much information for refitting a model.

C.3 Boosting model

We fit the Boosting model by using the "GBM" package in R (Ridgeway, 2015). In the Boosting model, we also include the AR terms of target variables up to 7 th lag. For a Boosting model, there are many hyper-parameters for tuning. We choose the number of trees in the model as 400, which decides how many iterations we do in the model. We choose 0.01 as the shrinkage parameter, which a regularization parameter to determine how fast the algorithm moves across the gradient. Decreasing shrinkage usually improves performance, but requires more trees for the ensemble. We choose

Year/Quarter	AR1	AR2	Intercept
2003/02	0.55		0.658
2003/03	0.556		0.638
2003/04	0.552		0.643
2004/01	0.552		0.643
2004/02	0.549		0.650
2004/03	0.548		0.666
2004/04	0.552		0.673
2005/01	0.552		0.673
.			
.			
.			
2013/02	0.580	-0.052	0.593
2013/03	0.576	-0.047	0.589
2013/04	0.575	-0.046	0.591
2014/01	0.576	-0.046	0.593
2014/02	0.575	-0.046	0.585
2014/03	0.569	-0.041	0.594
2014/04	0.569	-0.042	0.595
2015/01	0.569	-0.043	0.592

Table C.1: ARIMA models 2003-2015

20 as the depth of the model, which means each tree will evaluate 20 decisions, and each decision will yield 20 nodes from the prior node. We choose 5 as the minimum number of observations in the terminal node.

C.4 Prior distributions and prior elicitation

Given γ, the prior for the precision follows a Gamma distribution with parameters $\frac{df}{2}$ and $\frac{ss}{2}$. Thus, the reciprocal of the mean of the Gamma distribution ss/df is a prior estimate of σ^2.

To elicit priors for ss and df, it could be done by using a device $ss = df(1-R^2)s^2_{y^*}$. R^2 is expected R^2, and $s^2_{y^*}$ is the sample variance of the modified target variable, which is $s^2_{y^*} = \sum_{t=1}^{T} \frac{({y^*}_t - \bar{y^*})^2}{T-1}$.

The “Slab” prior is a very weakly informative prior which is close to being flat. In some sense, ss can be interpreted as a prior sum of squared error, and the df can be interpreted as a prior sample size. (which decides the weight given to the guess at R^2)

Using our previous knowledge of parameters, we can set our values for the parameters π, b_γ, Ω^{-1}, df, and ss. For simplicity, we also can just specify an expected model size n, κ, and expected R^2, and a sample size df. In our case, we use the default values $R^2 = 0.5$ and $df = 0.01$ in the R package “bsts” (Scott, 2015). We set “expected model size” to 4, so the $\pi = 4/602$ is the average probability to included in the model.

For elicitation of the prior of Ω^{-1}, we have $\Omega^{-1} \propto X^T X$, but it is not feasible when $X^T X$ is singular or not full rank, which is possible in our design matrix.

First, we have a fat regression, the number of predictors is larger than the number of observation. Second, we have many macroeconomic variables as predictors that are closely correlated. There are possible strong multi-collinearity in design matrix. When $X^T X$ is rank deficient, $p(\beta, \sigma|\gamma)$ is improper for some value of γ. Scott and Varian (2014b) propose that we can averaging $X^T X$ by its diagonal to restore propriety.

$$\Omega^{-1} = \frac{\kappa}{T}[wX^T X + (1-w)diag(X^T X)].$$

In their research, Brodersen, Gallusser, Koehler, Remy, and Scott (2014) set $\kappa = 1$ and $w = 1/2$ as default values. The matrix $X^T X/\sigma^2$ is the total Fisher information matrix in the full data, and $\frac{1}{T}X^T X$ is the average information in a single observation.

C.5 Estimating the model using Markov Chain Monte Carlo

We follow Scott and Varian (2014b) in that we draw samples using Markov Chain Monte Carlo.

Say y is the target series, α is vector of the states, and θ is vector of the parameters. The complete data posterior distribution is

$$p(\theta, \alpha|y) \propto p(\theta_0)p(\alpha_0)\prod_{t=1}^{T} p(y_t|\alpha_t, \theta)p(\alpha_t|\alpha_{t-1}, \theta).$$

We can use Gibbs sampling to draw $p(\alpha|\theta, y)$ and $p(\theta|\alpha, y)$ alternatively. And we can get a sequence of samples from a Markov chain with stationary distribution $p(\theta, \alpha|y)$. This chain consists of $(\theta, \alpha)_0$,$(\theta, \alpha)_1$, ... a sequence of samples.

Since α is a Markov Chain, the time series components and regression components are independent conditional on α. Let ψ is vector of the parameters associated with α. θ includes β, σ^{-2}, and ψ. The $p(\theta|\alpha, y)$ can be decomposed into several independent conditional posterior distribution of the α. Then we have

$$p(\psi, \theta, \sigma^{-2}|\alpha, y) = P(\psi|\alpha, y)p(\theta, \sigma^{-2}|\alpha, y)$$

C.5.1 Sampling ψ

Suppose we have a local linear trend model with two states.

- State equation 1 (random walk + trend):

$$\mu_t = \mu_{t-1} + b_{t-1} + w_{1t}, \quad w_{1t} \sim N(0, W_1)$$

- State equation 2 (random walk for trend):

$$b_t = b_{t-1} + w_{2t}, \quad w_{2t} \sim N(0, W_2)$$

Assume independent Gamma priors for state variances:

$$\frac{1}{W_1} \sim \Gamma(df_1/2, ss_1/2),$$

$$\frac{1}{W_2} \sim \Gamma(df_2/2, ss_2/2),$$

In the Gamma distribution, the reciprocal of the expectation ss/df is a prior estimate of W, and df/ss is a prior estimate of precision $1/W$. The prior parameter ss can be interpreted as a prior sum of squared error, and df is the weight assigned to the prior estimate of precision $1/W$.

If we do not have enough information about the prior distribution of the state variance W, we can choose a small value for df and small value for df/ss. Brodersen et al. (2014) choose $1/W \sim \Gamma(10^{-2}, 10^{-2}s_y^2)$ as their default priors for a seasonal and local linear trend model, where $s_y^2 = \sum_t^T \frac{(y_t-\bar{y})^2}{(T-1)}$ is the sample variance of the target variable. In this way, they scaling the sample variance to elicit a prior for W. Scaling by the sample variance is similar to scaling the data before the analysis. By scaling in the prior, they can model the data on its original scale (Scott & Varian, 2014a; Brodersen et al., 2014).

The full conditional posterior distribution is the product of two independent Gamma distributions:

$$p(1/W_1, 1/W_2|\alpha) = \Gamma(\frac{df_1 + T - 1}{2}, \frac{SS_1}{2})\Gamma(\frac{df_2 + T - 1}{2}, \frac{SS_2}{2}),$$

where

$$SS_1 = ss_1 + \sum_{t=2}^{T}(\mu_t - \mu_{t-1} - b_{t-1})^2$$

$$SS_2 = ss_2 + \sum_{t=2}^{T}(b_t - b_{t-1})^2$$

Given the $\alpha : \mu, b$, we can draw W_1 and W_2 from their full conditional distributions.

C.5.2 Sampling θ, σ

The full distribution for θ, σ is independent conditional on α.

Suppose observation equation is:

$$y_t = \mu_t + \beta \mathbf{x}_t + v_t, \quad v_t \sim N(0, V)$$

We subtract the target time series component μ_t from y_t, and get an axillary variable $y^* = y_t - \mu_t$. Then we have

$$y_t^* = y_t - \mu_t = \beta x_t + \epsilon_t \sim N(\beta x_t, \sigma^2)$$

And we are left with a standard spike-and-slab regression. σ^2 is the overall variance level. We can use “stochastic search variable selection” (SSVS) algorithm to draw from $p(\beta_\gamma, \sigma^2|\gamma, \alpha, \mathbf{y}^*)$, where vector $\mathbf{y}^* = y_{1:T}^*$ is all the information about y^* up to time T.

A “slab” prior includes:

$$p(\beta, \sigma^{-2}, \gamma) = p(\beta \mid \gamma, \sigma^{-2})p(\sigma^{-2} \mid \gamma)p(\gamma)\,,$$

$$\beta_\gamma \mid \sigma^2, \gamma \sim N(b_\gamma, \sigma^2(\Omega_\gamma^{-1})^{-1})\,,$$

$$\frac{1}{\sigma^2} \mid \gamma \sim \Gamma(\frac{df}{2}, \frac{ss}{2}),$$

Then conditional on γ, the joint posterior distribution for β and σ^2 can be estimated from standard conjugacy formula (Gelman et al., 2013) :

$$\beta_\gamma \mid \sigma, \gamma, \mathbf{y}^*, \alpha \sim N(\tilde{\beta_\gamma}, \sigma^2(V_\gamma^{-1})),$$

$$\frac{1}{\sigma^2} \mid \gamma, \mathbf{y}^*, \alpha \sim \Gamma(\frac{df+T}{2}, \frac{ss+\tilde{S}}{2}),$$

where

$$V_\gamma^{-1} = X_\gamma^T X_\gamma + \Omega_\gamma^{-1},$$

$$\tilde{\beta_\gamma} = (V_\gamma^{-1})^{-1}(X_\gamma^T y_\gamma^* + \Omega_\gamma^{-1} b_\gamma),$$

$$\tilde{S} = \sum_{t=1}^{T} (y_\gamma^* - \mathbf{x}^T \tilde{\beta_\gamma})^2 + (\tilde{\beta_\gamma} - b_\gamma)^T \Omega_\gamma^{-1} (\tilde{\beta_\gamma} - b_\gamma) = y_\gamma^{*T} y_\gamma^* + b_\gamma^T \Omega_\gamma^{-1} b_\gamma - \tilde{\beta_\gamma}^T V_\gamma^{-1} \tilde{\beta_\gamma},$$

C.5.3 Sampling γ

The prior distribution for $p(\gamma|\mathbf{y}^*, \alpha)$ is :

$$p(\gamma \mid \mathbf{y}^*, \alpha) \propto \frac{|\Omega^{-1}|^{1/2}}{|V_\gamma^{-1}|^{1/2}} \tilde{S}^{-\frac{df+T}{2}},$$

Under Zellner's g-prior,

$$\frac{|\Omega^{-1}|}{|V_\gamma^{-1}|} = \left(\frac{\kappa/T}{1+\kappa/T}\right)^{|\gamma|}$$

where $|\gamma|$ is the number of the included predictors. In general, $|\Omega^{-1}| \leq |V_\gamma^{-1}|$. This implies that $p(\gamma|\mathbf{y}^*, \alpha)$ is tends to pick models with few predictors and small residual variation.

Due to the conjugacy, we can get analytical expression for the marginal posterior of γ by marginalizing over β_γ and $1/\sigma^2$:

$$\gamma \mid \mathbf{y}^*, \alpha \sim C(\mathbf{y}^*) \frac{|\Omega^{-1}|^{1/2}}{|V_\gamma^{-1}|^{1/2}} \frac{p(\gamma)}{(ss + \tilde{S})^{\frac{T}{2}-1}} \,,$$

where $C(\mathbf{y}^*)$ is a normalizing constant.

We can use a Gibbs sampling to draw γ_i given γ_{-i} and $p(\gamma|\mathbf{y}^*, \alpha)$. γ_{-i} is other γ_j in γ where $j \neq i$.

Since the calculation of $p(\gamma|\mathbf{y}^*, \alpha)$ only need to compute those matrices associated with $\gamma = 1$, the SSVS is tractable when computating a system with many predictors (Scott & Varian, 2014b).

References

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., & Veronese, G. (2010). New Eurocoin: Tracking Economic Growth in Real Time. *The Review of Economics and Statistics*, *92*(4), 1024–1034.

Andersson, M. K., & Karlsson, S. (2008). Bayesian forecast combination for VAR models. In *Bayesian econometrics* (Vol. 23, pp. 501–524). Emerald Group Publishing Limited.

Andreou, E., Ghysels, E., & Kourtellos, A. (2013). Should Macroeconomic Forecasters Use Daily Financial Data and How? *Journal of Business & Economic Statistics*, *31*(2), 240–251.

Bai, J., Ghysels, E., & Wright, J. H. (2013). State space models and MIDAS regressions. *Econometric Reviews*, *32*, 779–813.

Bai, J., & Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, *25*, 52–60.

Bai, J., & Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, *24*, 607–629.

Banbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. In G. Elliott & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 195–233). Elsevier.

Banbura, M., Giannone, D., & Reichlin, L. (2010). Nowcasting. In M. Clements & D. Hendry (Eds.), *Oxford handbook on economic forecasting.* Oxford: Oxford University Press.

Banbura, M., & Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, *29*(4), 133–160.

Bencivelli, L., Marcellino, M., & Moretti, G. (2012). *Selecting predictors by using Bayesian model averaging in bridge models.* Bank of Italy.

Boivin, J., & Ng, S. (2006, May). Are more data always better for factor analysis? *Journal of Econometrics*, *132*, 169–194.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Brodersen, B. K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2014). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*.

Buchen, T., & Wohlrabe, K. (2011). Forecasting with many predictors: Is boosting a viable alternative? *Economics Letters*, *113*, 16–18.

Carriero, A., Clark, T. E., & Marcellino, M. (2015). Real-time nowcasting with a Bayesian mixed frequency model with stochastic bolatility. *Journal of the Royal Statistical Society: Series A (Statistics in Society).*

Carriero, A., Kapetanios, G., & Marcellino, M. (2011). Forecasting large datasets with Bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, *26*, 735–761.

Chipman, H., George, E. I., & McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. *IMS Lecture Notes - Monograph Series*, *38*, 65–134.

Clements, M. P., & Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data. *Journal of Business & Economic Statistics*, *26*, 546–554.

Clements, M. P., & Galvão, A. B. (2009). Forecasting US output growth using leading indicators: An appraisal using MIDAS models. *Journal of Applied Econometrics*, *24*, 1187–1206.

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*(2), 318–328.

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi maximum likelihood approach for large approximate dynamic factor models. *The Review of Economics and Statistics*, *94*, 1014–1024.

Durbin, J., & Koopman, S. (2012). *Time series analysis by state space methods.* Oxford: Oxford University Press.

Durbin, J., & Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*(3), 603–616. doi: 10.1093/biomet/89.3.603

Ferrara, L., & Marsilli, C. (2013). *Variable selection with mixed frequencies : An assessment based on macroeconomic forecasting.* Working paper, Banque de France.

Ferrara, L., Marsilli, C., & Ortega, J. P. (2014). Forecasting growth during the Great Recession: Is financial volatility the missing ingredient? *Economic Modelling*, *36*, 44–50.

Foroni, C., & Marcellino, M. (2013). *A Survey of Econometric Methods for Mixed-Frequency Data.* Working paper, Norges Bank.

Foroni, C., & Marcellino, M. (2014). A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates. *International Journal of Forecasting*.

Frale, C., Marcellino, M., Mazzi, G. L., & Proietti, T. (2011). EUROMIND: A monthly indicator of the euro area economic conditions. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *174*, 439–470.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition.* London: Chapman & Hall/CRC.

George, E. I., & Mcculloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*, 339–373.

Ghysels, E. (2012). *Mixed frequency vector autoregressive models.* Working paper, University of North Carolina.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). *The MIDAS touch: mixed data sampling regression models.* Discussion Paper, University of California and University of North Carolina.

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*, 53–90.

Ghysels, E., & Wright, J. H. (2009). Forecasting professional forecasters. *Journal of Business & Economic Statistics*, *27*, 504–516.

Guérin, P., & Marcellino, M. (2013). Markov-switching MIDAS models. *Journal of Business and Economic Statistics*, *31*, 45–56.

Harvey, A. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge, UK: Cambridge University Press.

Harvey, A. (2006). Forecasting with unobserved components time series models. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1, pp. 327–412). Amsterdam: North Holland.

Hendry, D., & Hubrich, K. (2011). *Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate.*

Hyndman, R. J. (2015). *forecast: Forecasting functions for time series and linear models.*

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for {R}. *Journal of Statistical Software*, *26*, 1–22.

Inoue, A., & Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association*, *103*, 511–522.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *Annals of Statistics*(2), 730–773.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer-Verlag.

Jungbacker, B., Koopman, S. J., & van der Wel, M. (2011). Maximum likelihood estimation for dynamic factor models with missing data. *Journal of Economic Dynamics and Control*, *35*, 1358–1368.

Kaufmann, S., & Schumacher, C. (2012). *Finding relevant variables in sparse Bayesian factor models: Economic applications and simulation results.* Discussion Paper 29, Deutsche Bundesbank.

Koop, G., & Potter, S. (2004). Forecasting in dynamic factor models using Bayesian model averaging. *Econometrics Journal*, *7*, 550–565.

Koopman, S. J., & van der Wel, M. (2013). Forecasting the US term structure of interest rates using a macroeconomic smooth dynamic factor model. *International Journal of Forecasting*, *29*(4), 676–694.

Korobilis, D. (2013). Hierarchical shrinkage priors for dynamic regressions with many predictors. *International Journal of Forecasting*, *29*, 43–59.

Kuzin, V., Marcellino, M., & Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting*, *27*, 529–542.

Madigan, D., & Raftery, A. E. (1994). *Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window* (Vol. 89).

Ouysse, R. (2013). *Forecasting using a large number of predictors: Bayesian model averaging versus principal components regression.* Australian School of Business Research Paper,University of New South Wales.

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*(482), 681–686.

Petris, G., Petrone, S., & Campagnoli, P. (2008). *Dynamic Linear Models with R.* New York: Springer-Verlag.

Ridgeway, G. (2015). gbm: Generalized Boosted Regression Models [Computer software manual]. R package version 2.1.1.

Rodriguez, A., & Puggioni, G. (2010). Mixed frequency models: Bayesian approaches to estimation and prediction. *International Journal of Forecasting*, *26*, 293–311.

Schorfheide, F., & Song, D. (2015). Real-time forecasting with a mixed-frequency

VAR. *Journal of Business & Economic Statistics*, *33*, 366–380.

Scott, S. L. (2015). bsts: Bayesian Structural Time Series [Computer software manual].

Scott, S. L., & Varian, H. R. (2014a). Bayesian Variable Selection for Nowcasting Economic Time Series. *Economics of Digitization*, 1–22.

Scott, S. L., & Varian, H. R. (2014b). Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, *5*, 4.

Stock, J., & Watson, M. (2006). Dynamic factor models. *Oxford Handbook of Economic Forecasting*(January), 1–43.

Stock, J., & Watson, M. (2009). Forecasting in Dynamic Factor Models Subject To Structural Instability. *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, 173–205.

Stock, J., & Watson, M. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business and Economic Statistics*, *30*(4), 481–493.

Tibshirani, R. (1994). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society*, *58*, 267–288.

Timmermann, A. (2006). Forecast Combinations. In G. Elliott, C. W. Granger., & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1, pp. 135–196). Amsterdam: North Holland.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics Tools to Manipulate Big Data. *Journal of Economic Perspectives*, *28*(2), 3–28.

Wohlrabe, K., & Buchen, T. (2014). Assessing the macroeconomic forecasting performance of boosting: Evidence for the United States, the Euro area and Germany. *Journal of Forecasting*, *33*, 231–242.

Wright, J. H. (2009). Forecasting US inflation by Bayesian model averaging. *Journal of Forecasting*, *28*, 131–144.

Yuan, M., & Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*(472), 1215-1225.