Forecasting with Many Predictors

Steve Scott
Hal Varian

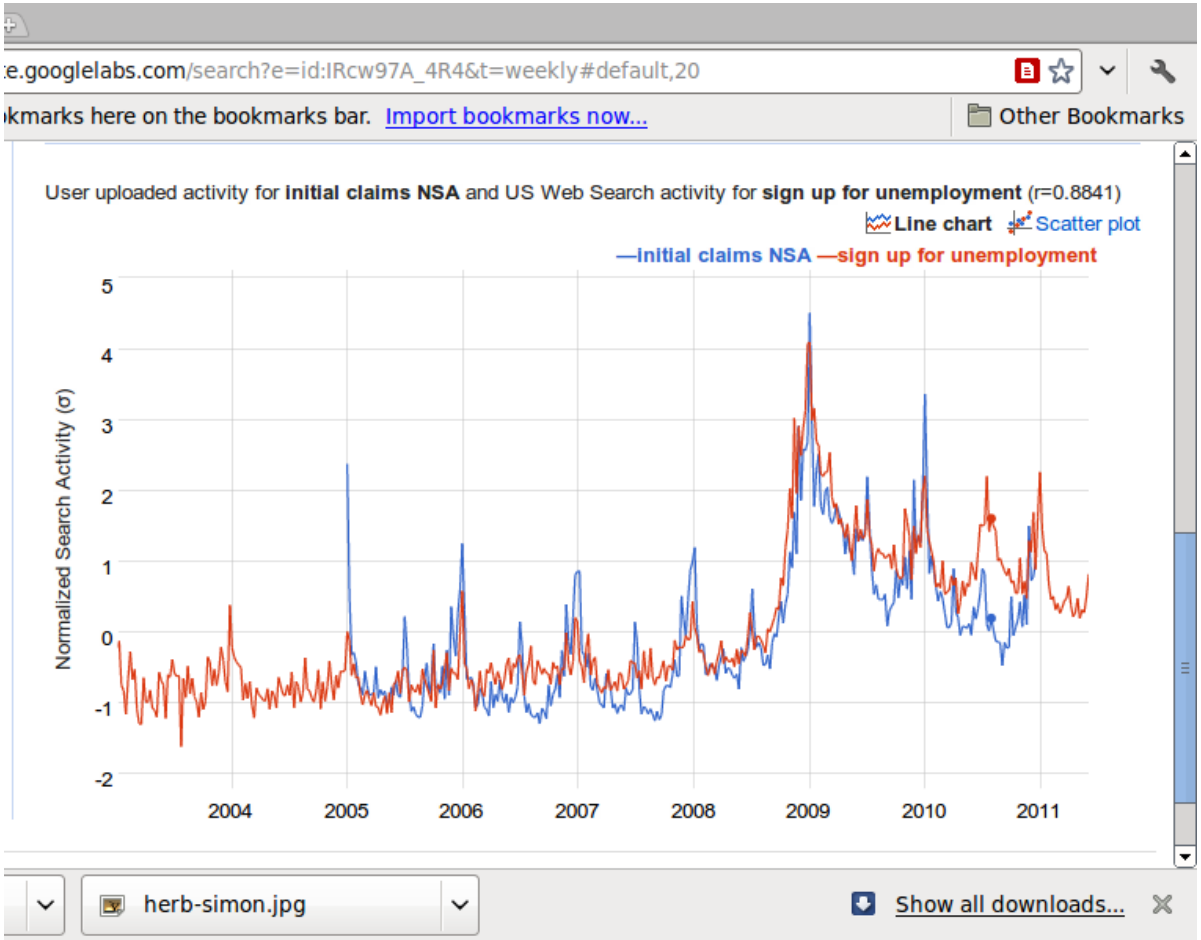
June 30, 2012

Problem motivation

- Want to use Google query data for nowcasting economic series
- Fat regression problem: there are many more predictors than observations
- Millions of queries, hundreds of categories
  - $n \sim 100$ for monthly economic data
  - $k \sim 150$ for "economic" categories in I4S
- How do we choose which variables to include?

Example: unemployment

- Sometimes simple correlation works
- Initial claims for unemployment benefits (NSA)
- Query "sign up for unemployment"



Approaches to variable selection

- Human judgment
- Significance testing (forward and backward stepwise regression)
- Information criteria (AIC, BIC, SIC)
- Principle component and factor models
- Lasso, penalized regression models

Our approach

- Original approach
  - forecast y_t using its own past values and human-chosen contemporaneous regressors from I4S
  - non-seasonal AR1: $y_t = a_1 y_{t-1} + b x_t + e_t$
  - seasonal AR1: $y_t = a_1 y_{t-1} + a_{12} y_{t-12} + b x_t + e_t$
- Current approach
  - Local linear trend and Kalman filter for time series
  - Spike and slab regression for variable (model) selection
  - Bayesian model averaging over all models for final forecast

Basic structural model with regression

- Classic time series model with constant level, linear time trend, regressor
  - $y_t = \mu + bt + \beta x_t + e_t$
- "Local linear trend" is a stochastic generalization of this
  - Observation: $y_t = \mu_t + b_t + z_t + e_{1t}$
  - State 1: $\mu_t = \mu_{t-1} + e_{2t}$
  - State 2: $b_t = b_{t-1} + e_{3t}$
  - State 3: $z_t = \beta x_t + e_{4t}$
  - In simple case where $b_t = \beta = 0$
    - $e_{2t} = 0$ is constant mean model where best estimate is avg
    - $e_{1t} = 0$ is random walk where best estimate is current value
- Parameters to estimate: regression coefficients β and variances of (e_{it}) for $i = 1, \ldots, 4$
- Use these variances to construct optimal Kalman forecast: $\hat{y}_t = y_{t-1} + \beta x_t + k_t(\text{variances}) \times \text{forecast error at } t-1$

Advantages of Kalman

- No problem with unit roots or other kinds of nonstationarity
- No problem with missing observations
- No problem with mixed frequency
- No differencing or identification stage (easy to automate)
- Nice Bayesian interpretation
- Easy to compute estimates (particuarly in Bayesian case)
- Nice interpretation of structural components
- Adaptive estimates (good for recession)
- Good forecasts

Spike and slab regression for variable choice

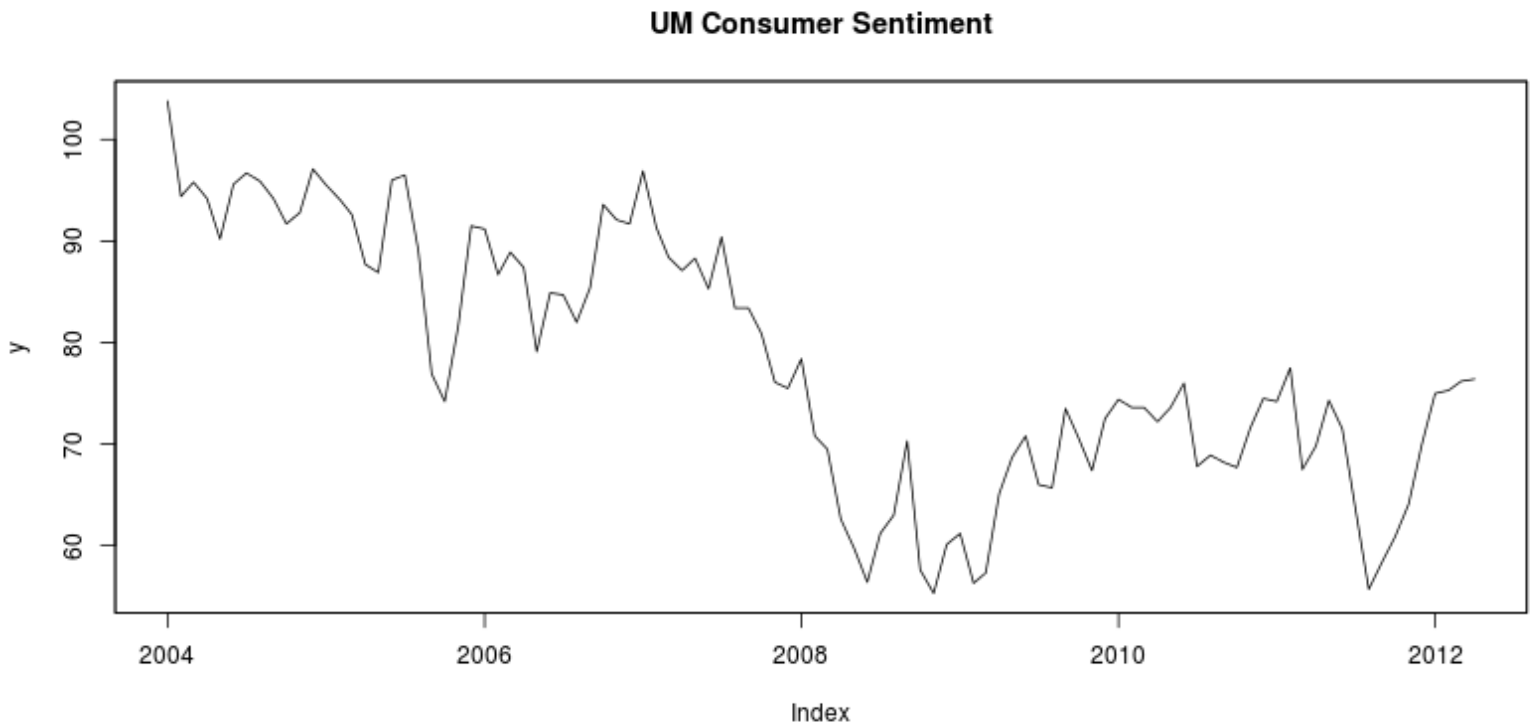
- Spike
  - Define vector γ that indicates variable inclusion
  - $\gamma_i = 1$ if variable i has non-zero coefficient in regression, 0 otherwise
  - Put a diffuse beta prior distribution, $p(\gamma)$, on γ
  - Can use an informative prior; e.g., expected number of predictors
- Slab
  - Conditional on being in regression ($\gamma_i = 1$) put a diffuse prior on β_i
- Estimate posterior distribution of (γ, β) using MCMC

Bayesian model averaging

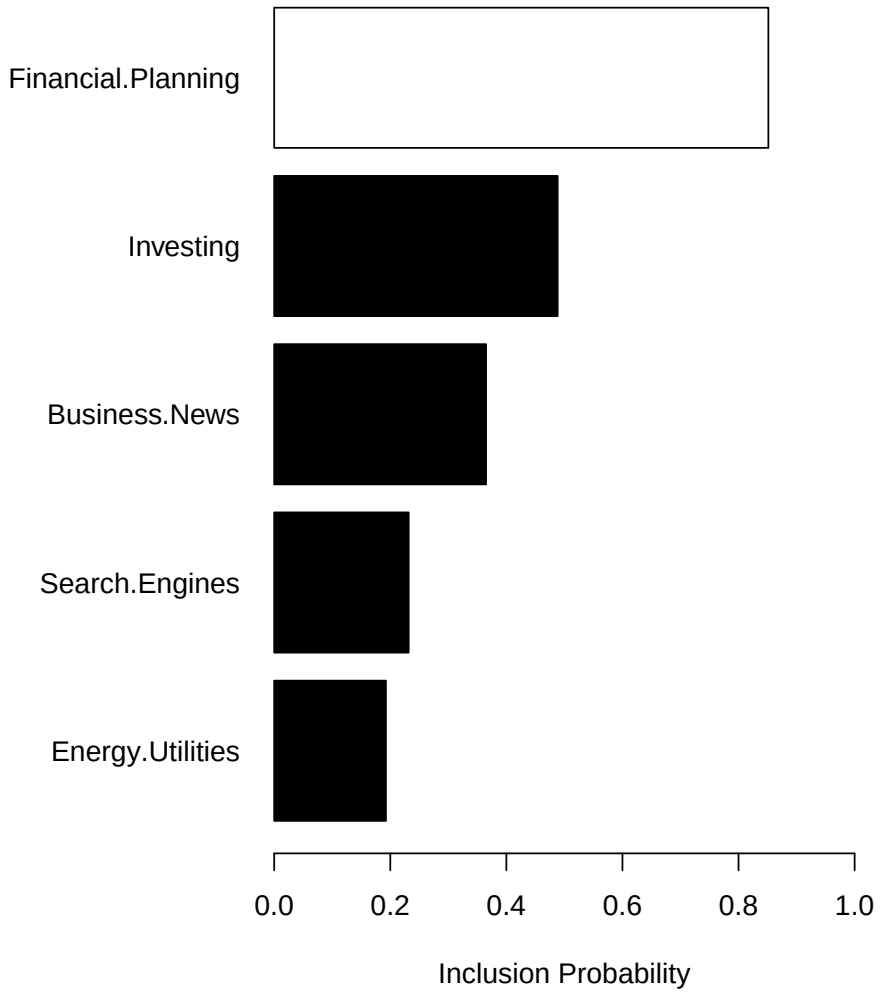
- We simulate draws from posterior using MCMC
- Each draw has a set of variables in the regression (γ) and a set of regression coefficients (β)
- Make a forecast of y_t using these cofficiency
- Take average over all the forecasts for final prediction
- Take average over draws of γ to see which predictors have high probability of being in regression

Example: Consumer Sentiment

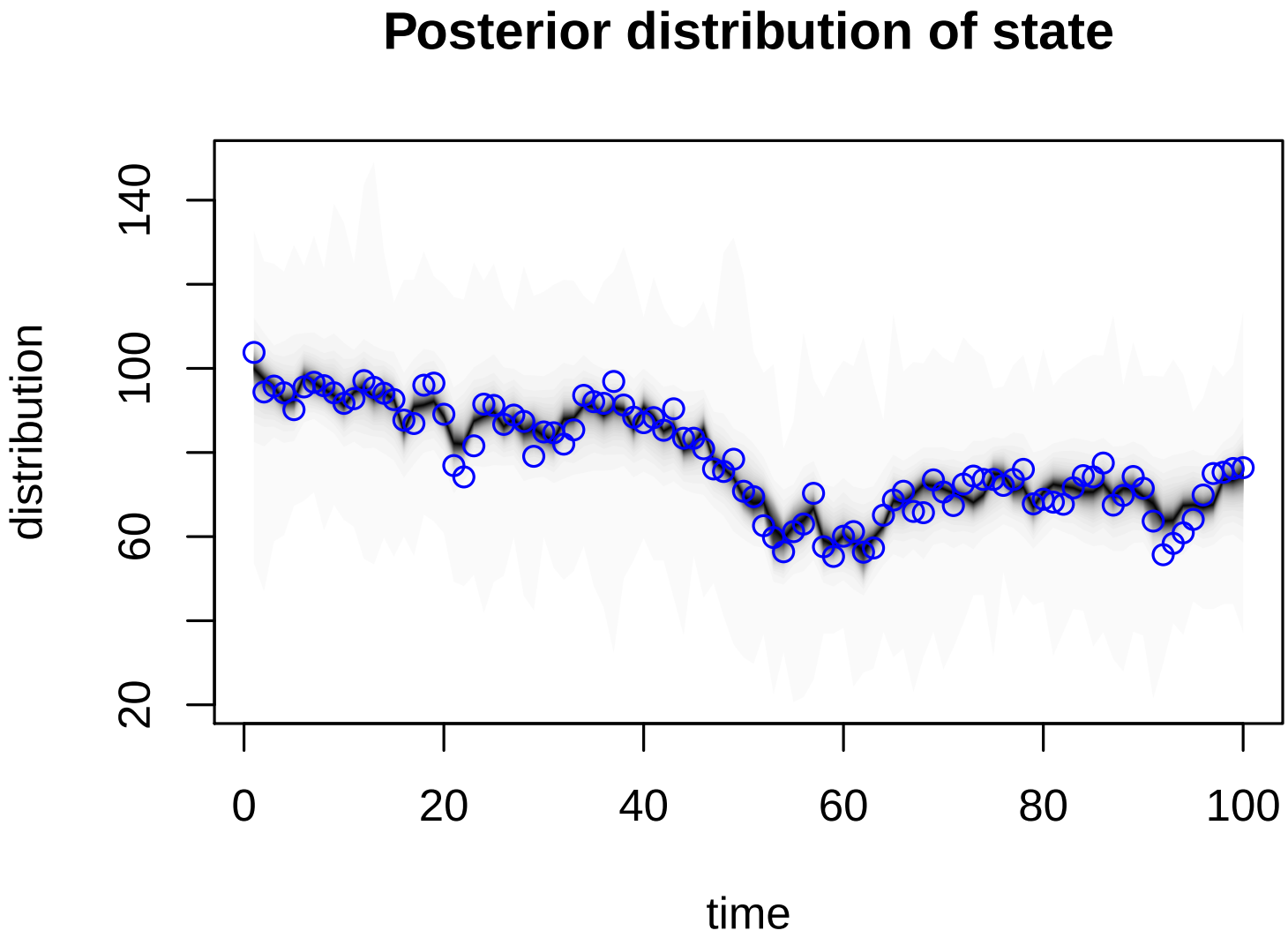
- Monthly UM Consumer sentiment from Jan 2004 to Apr 2012 ($n = 100$)
- Google Insights for Search categories related to economics ($k = 150$)
- No compelling intuition about what predictors should be



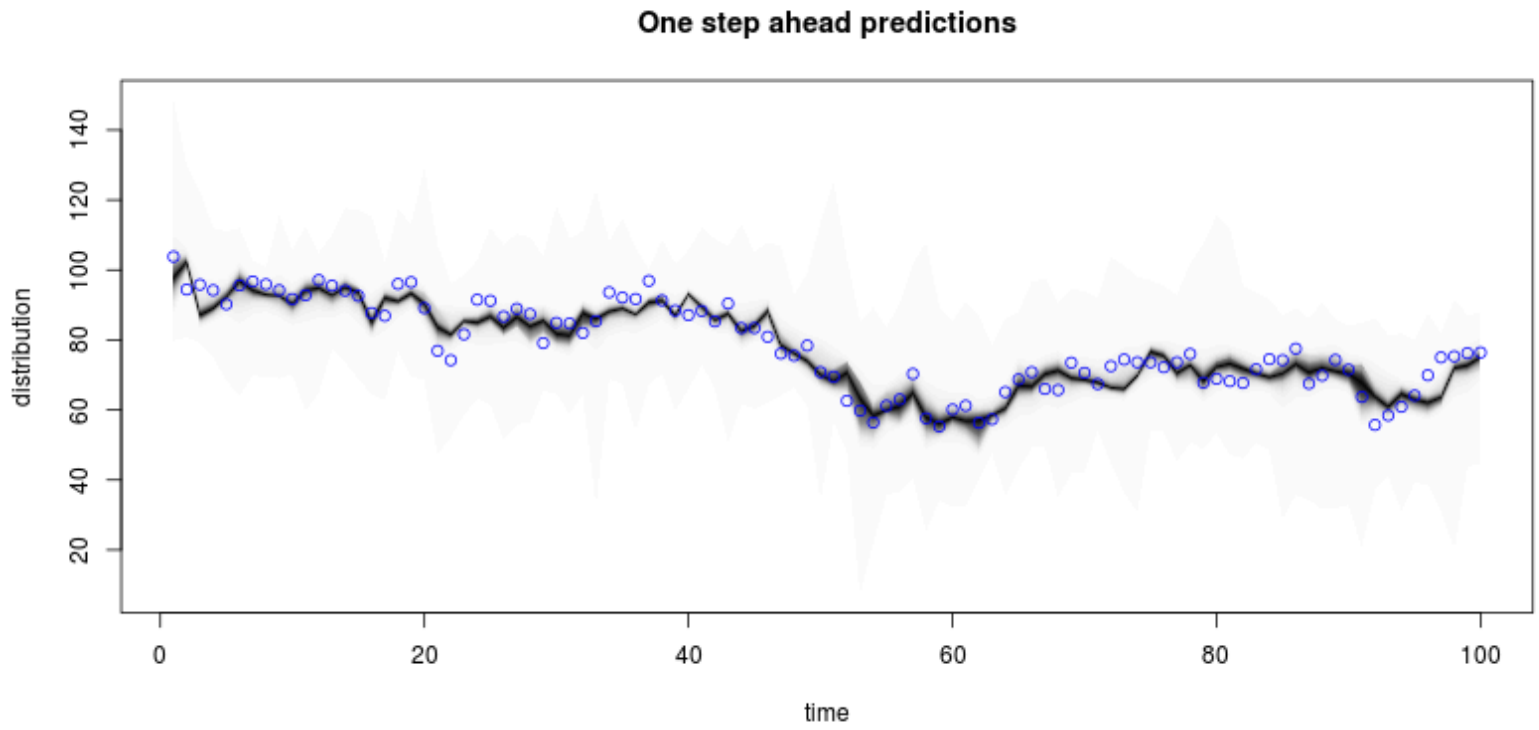
UM Consumer Sentiment



Posterior distribution of state and observations



Posterior distribution of one-step ahead forecast



Compare Actual, AR(1), reg and BSTS predictions

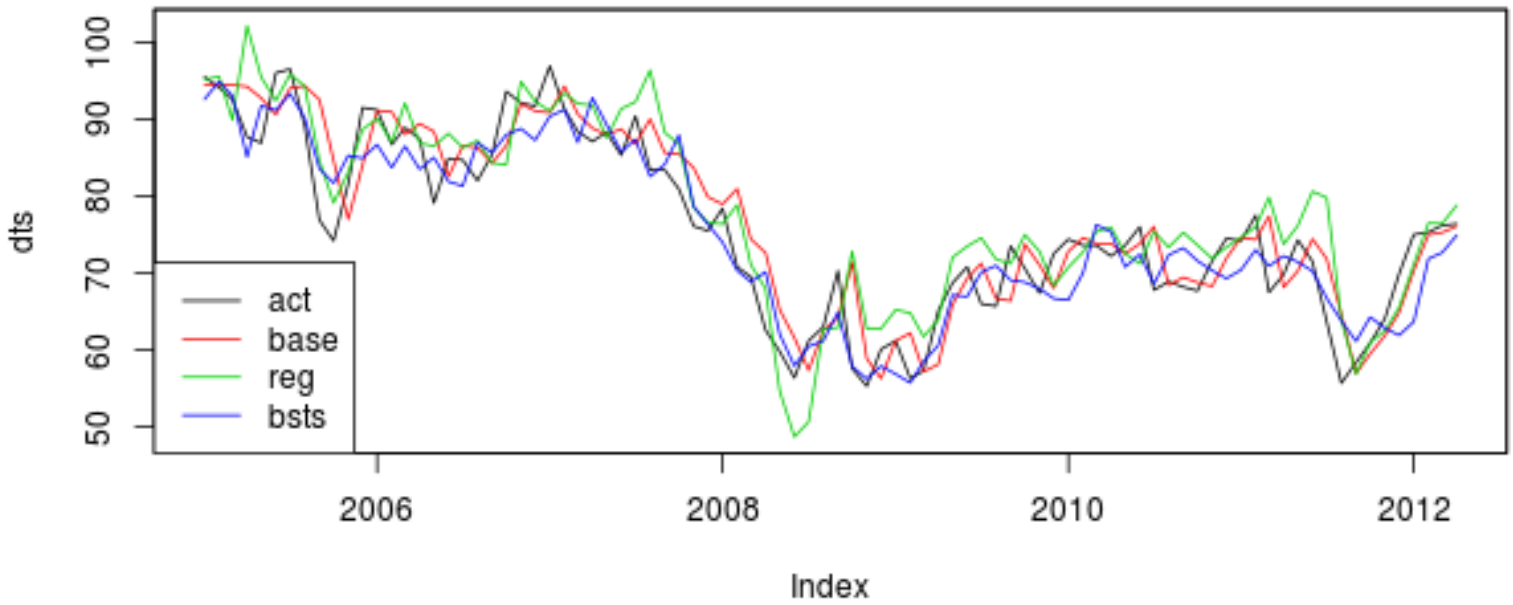


Figure: Actual, AR(1), regression, and BSTS one-step ahead predictions.

Decomposition of forecast into trend and regression

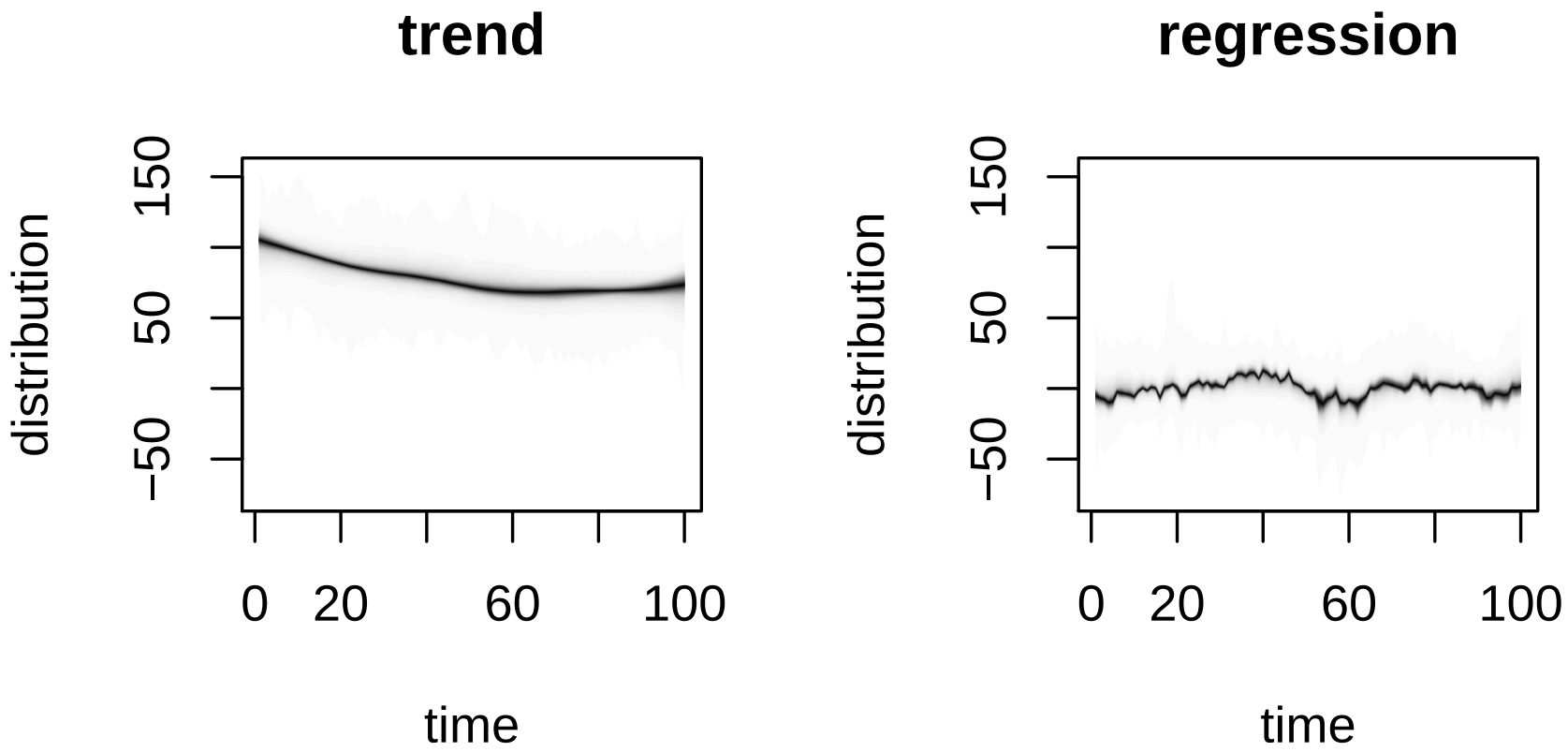


Figure: Decomposition of forecast.

Future work

- Seasonality — done
- Mixed frequency forecasting — almost done
- Fat tail distributions
- Parallel MCMC
- Automate the whole thing

Nowcasting with Fat Regressions

Steve Scott
Hal Varian

July 2012

Abstract

We consider the problem of short-term time series forecasting when there are more possible predictors than observations. Our approach combines three Bayesian techniques: Kalman filter for state space models, spike-and-slab regression, and model averaging. We illustrate the method using search engine query data to nowcast consumer sentiment.

1 Introduction

Choi and Varian [2009a,b, 2011] described how to use search engine data to forecast contemporaneous values of macroeconomic indicators. This type of contemporaneous forecasting, or "nowcasting," is of particular interest to central banks, and there have been several subsequent research studies from researches at these institutions. See, for example, Arola and Galan [2012], McLaren and Shanbhoge [2011], Hellerstein and Middeldorp [2012], Suhoy [2009], Carrière-Swallow and Labbé [2011].

In these studies, the researchers selected predictors using their judgment of relevance to the particular prediction problem. For example, it seems natural that search engine queries in the "Vehicle Shopping" category would be a good candidate for forecasting automobile sales while queries such as "file for unemployment" could be useful in forecasting initial claims for unemployment benefits.

One difficulty with using human judgment is that it does not easily scale to models where the number of possible predictors exceeds the number of observations—the so-called "fat regression" problem. For example, Google's Insights for Search (I4S) service provides data for millions of search queries and hundreds of search categories extending back to January

1, 2004. Even if we restrict ourselves to using only categories of queries, we will have several hundred possible possible predictors for 100 months of data. In this paper we describe a scalable approach to time series prediction for fat regressions of this sort.

2 Approaches to variable selection

Castle et al. [2009, 2010] describes and compares 21 techniques for variable selection for time-series forecasting. These techniques fall into 4 major categories.

- Significance testing (forward and backward stepwise regression, Gets)
- Information criteria (AIC, BIC)
- Principle component and factor models (e.g. Stock and Watson [2010])
- Lasso and penalized regression models (e.g., Hastie et al. [2009])

Our approach combines 3 statistical methods into an integrated system we call *Bayesian Structural Time Series* or BSTS for short.

- A "basic structural model" for trend and seasonality, estimated using Kalman filters;
- Spike and slab regression for variable selection;
- Bayesian model averaging over the best performing models for the final forecast.

We briefly review each of these methods and how they fit into our framework.

2.1 Structural time series and the Kalman filter

Harvey [1991], Durbin and Koopman [2001], Petris et al. [2009] and many others have advocated the use of Kalman filters for time series forecasting. The "local linear trend" model decomposes the time series into three components: a level, a local linear trend, and an error term. The parameters of the model can be estimated using maximum likelihood or Bayesian techniques and forecasts can be made using a Kalman filter.

This "local linear trend model" is a stochastic generalization of the classic constant-trend regression model,

$$y_t = \mu + bt + \beta x_t + e_t$$

In this classic case the level (μ) and trend (b) parameters are constant, (x_t) is a vector or contemporaneous regressors, β is a vector of regression coefficients, and e_t is an error term.

In local linear trend model each of these structural components is stochastic:

$$
\begin{aligned}
y_t &= \mu_t + b_t + z_t v_t \quad v_t \sim N(0, V) && (1)\\
\mu_t &= \mu_{t-1} + b_{t-1} + w_{1t} \quad w_{1t} \sim N(0, W_1) && (2)\\
b_t &= b_{t-1} + w_{2t} \quad w_{2t} \sim N(0, W_2) && (3)\\
z_t &= \beta x_t + w_{2t} \quad w_{3t} \sim N(0, W_3) && (4)
\end{aligned}
$$

The parameters to be estimated are the variance terms (V, W_1, W_2, W_3) and the regression coefficients, β.

If we drop the trend and regression coefficients by setting $b_t = 0$ and $\beta = 0$, the "local trend model" becomes the "local level" model. When $V = 0$, the local level model is a random walk, so the best forecast of y_{t+1} is y_t; when $W_1 = 0$, the local level model is a constant mean model, where the best forecast of y_{t+1} is the average of all previously observed values of y_t. Hence, this model yields two popular time series models as special cases.

It is easy to add a seasonal component to the local linear trend model, in which case it is referred to as the "basic structural model." In the Appendix we describe a general structural time series model that contains these and other models in the literature as special cases.

2.2 Spike and slab prior distributions

The spike-and-slab approach to model selection was developed by George and McCulloch [2007]) and Madigan and Raftery [1994].

Let γ denote a vector the same length as the list of possible regressors that indicates where or not a particular regressor is included in the regression. More precisely, γ is a vector the same length as β, where $\gamma_i = 1$ indicates $\beta_i \neq 0$ and $\gamma_i = 0$ indicates $\beta_i = 0$. Let β_γ indicate the subset of β for which $\gamma_i = 1$, and let σ^2 be the variance of the prior distribution on γ.

A spike and slab prior for the joint distribution of $(\beta, \gamma, \sigma^{-2})$ can be written

$$p(\beta, \gamma, \sigma^{-2}) = p(\beta_\gamma | \gamma, \sigma^{-2}) p(\sigma^{-2} | \gamma) p(\gamma). \quad (5)$$

There are several possibilities for the component distributions in equation (5), but a convenient choice is described below.

The "spike" part of a spike-and-slab prior refers to the point mass at zero, for which we assume

$$\gamma \sim \prod_i \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}. \tag{6}$$

When detailed prior information is unavailable, it is convenient to set all π_i equal to the same number, π. The common prior inclusion probability can easily be elicited from the expected number of nonzero coefficients. If k out of k coefficients are expected to be nonzero then set $\pi = k/k$ in the prior. More complex choices of $p(\gamma)$ can be made as well. For example, a non-binomial model could be used to encode rules such as the hierarchical principle (no high order interactions without lower order interactions). The MCMC methods described below are robust to the specific choice of the prior.

The "slab" component is a prior for the values of the nonzero coefficients, conditional on knowledge of which coefficients are nonzero. Let b be a vector of prior guesses for regression coefficients, let Ω^{-1} be an priori precision matrix, and let Ω_γ^{-1} denote rows and columns of Ω^{-1} for which $\gamma_i = 1$. A conditionally conjugate "slab" prior is

$$\begin{aligned} \beta_\gamma|\gamma,\sigma^2 &\sim \mathcal{N}\left(b_\gamma, \sigma^2\left(\Omega_\gamma^{-1}\right)^{-1}\right), \\ \frac{1}{\sigma^2} &\sim \Gamma\left(\frac{df}{2}, \frac{ss}{2}\right). \end{aligned} \tag{7}$$

It is conventional to assume $b = 0$ (with the possible exception of the intercept term) and $\Omega^{-1} \propto \mathbf{X}^T\mathbf{X}$, in which case equation (7) is known as Zellner's $g-$prior Chipman et al. [2001]. Because $\mathbf{X}^T\mathbf{X}/\sigma^2$ is the total Fisher information in the full data, it is reasonable to parametrize $\Omega^{-1} = \kappa(\mathbf{X}^T\mathbf{X})/n$, the average information available from κ observations.

One issue with Zellner's $g-$prior is that when the design matrix contains truly redundant predictors (as is the case when the number of possible predictors exceeds the number of observations), then $\mathbf{X}^T\mathbf{X}$ is rank deficient, which means that for some values of γ, $p(\beta,\sigma|\gamma)$ is improper. We can restore propriety by averaging $\mathbf{X}^T\mathbf{X}$ with its diagonal, so that

$$\Omega^{-1} = \frac{\kappa}{n}\left[w\mathbf{X}^T\mathbf{X} + (1-w)\text{diag}(\mathbf{X}^T\mathbf{X})\right].$$

The final values that need to be chosen are df and ss. These can be elicited by asking the modeler for the R^2 statistic he expects to obtain from the regression, and the weight he would like to assign to that guess, measured in terms of the equivalent number of observations. The df parameter is the equivalent number of observations, and $ss = df(1-R^2)s_y^2$.

Software implementing the spike-and-slab prior can make reasonable default choices for expected model size, κ, expected R^2, and df, giving the modeler the option to accept the defaults, or provide his own inputs.

2.3 Bayesian model averaging

Bayesian inference with spike-and-slab priors is an effective way to implement Bayesian model averaging over the space of time series regression models. Each combination of predictors is associated with a posterior probability $p(\gamma|\mathbf{y})$ that these predictors "belong in" the regression. Hence we can compute a weighted average over all the regression forecasts, using the posterior probabilities as weights. This approach is motivated by the Madigan and Raftery [1994] proof that averaging over an ensemble of models does no worse than using the best single model in the ensemble. See **?** for links to tools and applications of Bayesian model averaging.

3 Estimating the model

The Kalman filter, spike-and-slab regression, and model averaging all have natural Bayesian interpretations and tend to play well together. The basic parameters we need to estimate are γ (which variables are in the regression), β (the regression coefficients), and the variances of the error terms (V, W_1, W_2, W_3).

As the appendix describes in detail, we specific priors for each of these parameters and then sample from the posterior distribution using Markov Chain Monte Carlo techniques. There are a number of attractive short cuts available that make this sampling process quite efficient.

These techniques yield a sample from the posterior distribution for the parameters that can be then used to construct a posterior distribution for forecasts of time series of interest.

4 Nowcasting consumer sentiment

To illustrate the use of BSTS for nowcasting, we use the University of Michigan monthly survey of Consumer Sentiment from January 2004-April 2012. We focus on "nowcasting" since we expect that queries at time t could be related to sentiment at time t but are not necessarily predictive of future sentiment.

Our data from Google Search Insights starts at January 2004, and our sample ends in April 2012, giving us about 100 observations. For predictors, we use 151 categories from Google Search Insights that have some connection with economics. These potential predictors were chosen from the roughly 300 query categories using a using human judgment.

Our problem is to find a good set of predictors for 100 observations chosen from a set of 151 possible predictors. This qualifies as a qualifies as a mildly obese, if not quite fat, regression.

The Consumer Sentiment index is not highly seasonal but many of the potential predictors are seasonal so we first deseasonalize the data by using the R command **stl**. We then detrend the predictors by regressing each predictor on a simple time trend. A visual inspection of the predictors indicated that these techniques were sufficient to "whiten" the data.

We then applied the BSTS estimation procedure described earlier. Figure 1 shows the inclusion probability for the top k predictors. A white bar indicates that the predictor has a positive relationship with consumer sentiment and a black bar indicates a negative relationship. The size of the bar measures the proportion of the estimated models in which that predictor was present.

The top predictor is Financial Planning which is included in almost all of the models explored. The top queries in this category in the US can be found on the Google Search Insights web page. They are schwab, 401k, charles schwab, ira, smith barney, fidelity 401k, john hancock, 403b, 401k withdrawl, and roth ira.

The second most probable predictor is Investing, which tends to have a negative relationship with confidence. The top queries in this category are stock, gold, fidelity, stocks, stock market, silver, gold price, mutual, scottrade, and finance.

The inclusion of the Energy and Hybrid Vehicle categories are likely due to gasoline prices, which are known to have a negative impact on consumer sentiment in the US.

Figure 2 shows the posterior distribution of the one-step ahead forecast along with the actual observations.

Note that the regression parameters are estimated using the entire sample of data, but the forecasts for period t are made using the value of consumer sentiment at $t-1$ and the observed query categories at time t (for the included categories).

The model predicts reasonably well with a mean absolute one-step-ahead prediction error of about 4.5%. A naive AR(1) model has a mean absolute one-step-ahead prediction error of 5.2%, indicating an improvement of about 14%. See 3 for a time series plot of the actual, AR(1), and BSTS one-step-ahead predictions.

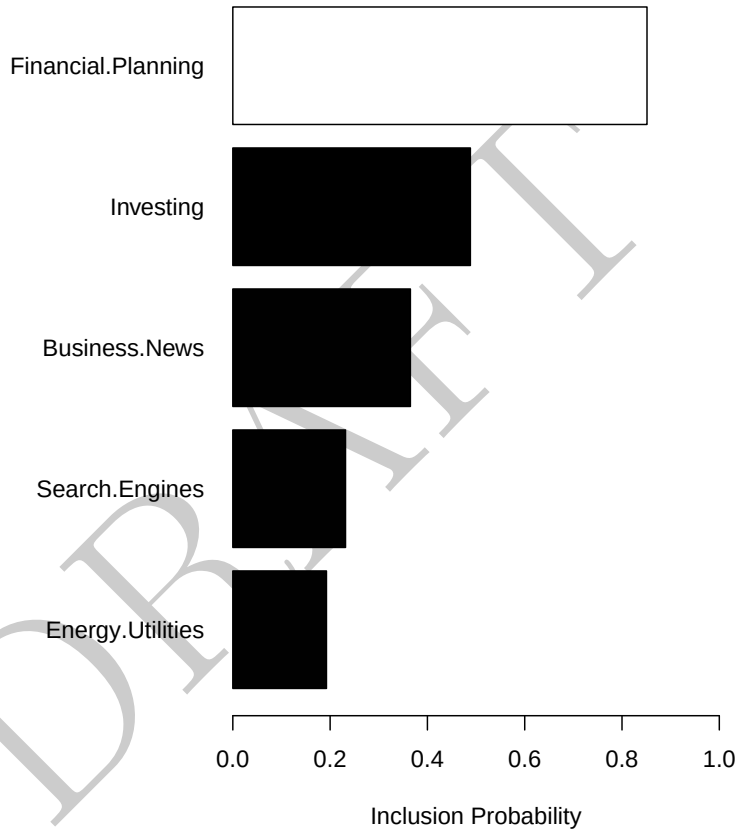


Figure 1: Top 5 predictors for consumer sentiment. Bars show the probability of inclusion. Shading indicates the sign of the coefficient.

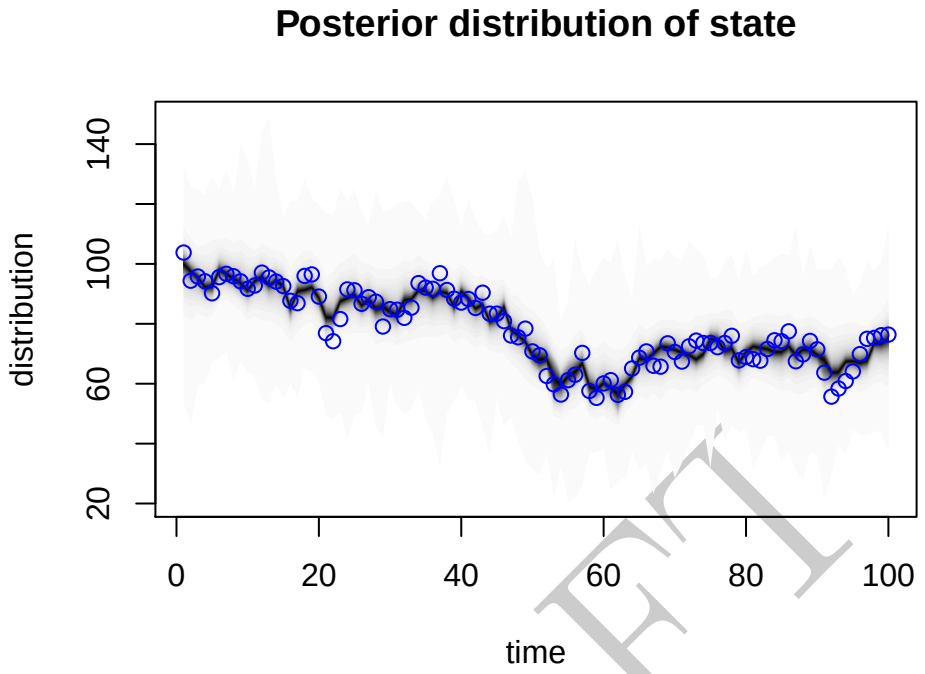


Figure 2: Posterior distribution of forecast and the observations.

Figure 3: Actual, base AR(1), regression, and BSTS one-step ahead predictions.

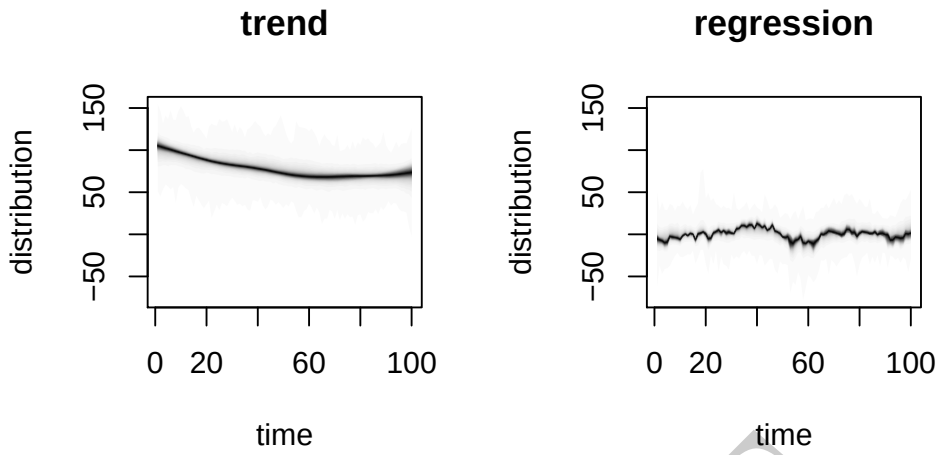


Figure 4: Decomposition of forecast.

As we have seen BSTS system can decompose the forecast into the trend and regression components. The trend component is basically the univariate Kalman filter forecast, while the regression component uses the predictors from the query categories. This decomposition, shown in This is shown in Figure 4, illustrates how much of the forecast is due to trend extrapolation and how much is due to the predictors.

5 Appendices

A Structural time series models

We focus on structural time series models of the standard form

$$\begin{aligned} y_t &= Z_t^T \alpha_t + \epsilon_t & \epsilon_t &\sim \mathcal{N}(0, H_t) \\ \alpha_{t+1} &= T_t \alpha_t + R_t \eta_t & \eta_t &\sim \mathcal{N}(0, Q_t)\,. \end{aligned} \tag{8}$$

Here y_t is time series to be modeled and the vector α_t is a latent variable indicating the state of the model; it contains any trend, seasonal, or other components deemed necessary by the modeler.

Z_t is a vector of coefficients applied to the state variables, ϵ_t is a Normally distributed error term with mean zero and H_t is its variance. Each state component contributes to the

block diagonal transition matrix T_t, the rectangular block diagonal residual matrix R_t, and the observation vector Z_t. The error term η_t has covariance matrix Q_t.

The model matrices (Z, T, R, H, Q) can be used to construct the Kalman filter, which can then be used to forecast future values $y_{t+\tau}$ from current observations $(y_1, \ldots, y_t)$. One attractive feature of the Kalman filter is that it has a natural Bayesian interpretation and can easily be combined with the variable selection and model averaging techniques we have chosen.

A.1 Regression

Regressors can be included in a structural time series model in either a static framework (where the regression coefficients are fixed) or dynamic framework (where the regression coefficients can change over time).

In a dynamic regression the coefficients are a component of the state vector which evolve over time according to some stochastic process. In a static regression, by contrast, the coefficients are fixed, unknown parameters. A convenient way to include a static regression component in the model is to set $\alpha_t = 1$, $t_t = 1$, $q_t = 0$, and $z_t = \beta^t \mathbf{x}_t$. This specification adds $\beta^t \mathbf{x}_t$ to the contributions of the other state components in a computationally efficient way, because it only adds one additional state to the model. A small dimension is helpful because the kalman recursions are quadratic in the dimension of the state space.

B Estimating the model using Markov Chain Monte Carlo

We estimate the posterior distribution of the model parameters using Markov Chain Monte Carlo. Let θ denote the collection of model parameters (β, σ, ψ) where ψ is the collection of all model parameters associated with state components other than the static regression. Then the complete data posterior distribution is

$$p(\theta, \boldsymbol{\alpha}|\mathbf{y}) \propto p(\theta)p(\alpha_0)\prod_{t=1}^{n} p(y_t|\alpha_t, \theta)p(\alpha_t|\alpha_{t-1}, \theta). \tag{9}$$

In order to sample from the posterior distribution we use an efficient Gibbs sampling algorithm that alternates between draws of $p(\boldsymbol{\alpha}|\theta, \mathbf{y})$ and $p(\theta|\boldsymbol{\alpha}, \mathbf{y})$, which produces a sequence $(\theta, \boldsymbol{\alpha})_0, (\theta, \boldsymbol{\alpha})_1, \ldots$ from a Markov chain with stationary distribution $p(\theta, \boldsymbol{\alpha}|\mathbf{y})$.

The key point is that, conditional on $\boldsymbol{\alpha}$, the time series and regression components of the model are independent. Thus the draw from $p(\theta|\boldsymbol{\alpha}, \mathbf{y})$ decomposes into several independent draws from the different conditional posterior distributions of the state components. In particular, $p(\psi, \beta, \sigma^{-2}|\boldsymbol{\alpha}, \mathbf{y}) = p(\psi|\boldsymbol{\alpha}, \mathbf{y})p(\beta, \sigma^{-2}|\boldsymbol{\alpha}, \mathbf{y})$.

B.1 Sampling $\boldsymbol{\alpha}$

The idea of using Kalman filtering to sample the state in a linear Gaussian structural time series model was independently proposed by [Carter and Kohn, 1994] and [Frühwirth-Schnatter, 1994]. Various improvements to the early algorithms have been made by [de Jong and Shepard, 1995] [Rue, 2001], and others. We use the method proposed by [Durbin and Koopman, 2002], who observed that the variance of $p(\alpha|\theta, \mathbf{y})$ does not depend on the numerical value of $\mathbf{y}$. Durbin and Koopman [2001] describes a fast smoothing method for computing $E(\boldsymbol{\alpha}|\mathbf{y}, \theta)$ using the Kalman filter.

Thus one may simulate a fake data set $(\mathbf{y}^*, \boldsymbol{\alpha}^*) \sim p(\mathbf{y}, \boldsymbol{\alpha}|\theta)$ by simply iterating equation (8). Then the fast mean smoother can be used to subtract the conditional mean $E(\boldsymbol{\alpha}^*|\theta, \mathbf{y}^*)$ from $\boldsymbol{\alpha}^*$, which is now mean zero with the correct variance. A second fast smoother can be used to add in $E(\boldsymbol{\alpha}|\mathbf{y}, \theta)$, yielding a draw of $\boldsymbol{\alpha}$ with the correct moments. Because $p(\boldsymbol{\alpha}|\mathbf{y}, \theta)$ is Gaussian, the correct moments imply the correct distribution.

B.2 Sampling θ

Many of the usual models for state components are simple random walks, whose variance parameters are trivial to sample conditional on $\boldsymbol{\alpha}$. For example, consider the state variables for the local linear trend model described in 1

$$\begin{aligned}\mu_{t+1} &= \mu_t + \delta_t + \eta_{0t} \\ \delta_{t+1} &= \delta_t + \eta_{1t},\end{aligned}$$

where η_0 and η_1 are independent Gaussian error terms with variances ψ_0^2 and ψ_1^2. With independent Gamma priors on $\psi_0^{-2} \sim \Gamma(df_0/2, ss_0/2)$ and $\psi_1^{-2} \sim \Gamma(df_1/2, ss_1/2)$, their full conditional is the product of two independent Gamma distributions

$$p(\psi_0^{-2}, \psi_1^2|\boldsymbol{\alpha}) = \Gamma\left(\frac{df_0 + n - 1}{2}, \frac{SS_0}{2}\right)\Gamma\left(\frac{df_1 + n - 1}{2}, \frac{SS_1}{2}\right),$$

where $SS_0 = ss_0 + \sum_{t=2}^{n}(\mu_t - \mu_{t-1} - \delta_{t-1})^2$ and $SS_1 = ss_1 + \sum_{t=2}^{n}(\delta_t - \delta_{t-1})^2$. These complete data sufficient statistics are observed given $\boldsymbol{\alpha}$, so drawing ψ_0^{-2} and ψ_1^{-2} from their full conditional distribution is trivial. Most of the traditional state models can be handled similarly, including the seasonal component of the BSM and dynamic regression coefficients.

The full conditional for (β, σ^{-2}) is likewise independent of the other state components, with $\tilde{y}_t = y_t - Z_t^T\alpha_t + \beta^T\mathbf{x}_t \sim \mathcal{N}\left(\beta^T\mathbf{x}_t, \sigma^2\right)$. Thus, by subtracting the contributions from the other state components from each y_t we are left with a standard spike and slab regression. The posterior distribution can be simulated efficiently by drawing from $p(\gamma|\boldsymbol{\alpha}, \mathbf{y})$ using a sequence of Gibbs sampling steps, and then drawing from the well known closed form $p(\beta_\gamma, \sigma^{-2}|\gamma, \boldsymbol{\alpha}, \mathbf{y})$. This technique is known as "stochastic search variable selection" [George and McCulloch, 1997]. There have been many suggested improvements to the SSVS algorithm (notably [Ghosh and Clyde, 2011]), but we have obtained satisfactory results with the basic algorithm.

The conditional posteriors for β_γ and σ^{-2} can be found in standard texts [e.g. Gelman et al., 2002]. They are

$$\begin{aligned} p(\beta|\mathbf{y}, \boldsymbol{\alpha}, \gamma, \sigma^{-2}) &= \mathcal{N}\left(\tilde{\beta}_\gamma, \sigma^2 V_\gamma\right), \text{ and} \\ p(\sigma^{-2}|\mathbf{y}, \boldsymbol{\alpha}, \gamma) &= \Gamma\left(\frac{df + n}{2}, ss + \tilde{S}\right), \end{aligned} \tag{10}$$

where the complete data sufficient statistics are $V_\gamma^{-1} = \mathbf{X}^T\mathbf{X}_\gamma + \Omega_\gamma^{-1}$, $\tilde{\beta}_\gamma = V_\gamma(\mathbf{X}^T\tilde{\mathbf{y}}_\gamma + \Omega_\gamma^{-1}b_\gamma)$, and $\tilde{S} = \sum_{t=1}^{n}(\tilde{y}_t - \mathbf{x}_t^T\tilde{\beta}_\gamma)^2 + (\tilde{\beta}_\gamma - b_\gamma)^T\Omega_\gamma^{-1}(\tilde{\beta}_\gamma - b_\gamma)$. The distribution for $p(\gamma|\boldsymbol{\alpha}, \mathbf{y})$ can be shown to be

$$p(\gamma|\mathbf{y}, \boldsymbol{\alpha}) \propto \frac{|\Omega^{-1}|^{-1/2}}{|V_\gamma^{-1}|^{-1/2}}\tilde{S}^{-(df+n)/2}. \tag{11}$$

Let $|\gamma|$ denote the number of included components. Under Zellner's $g-$prior it is easy to see that

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

is decreasing in $|\gamma|$. It is true in general that $|\Omega^{-1}| \leq |\Omega^{-1} + \mathbf{X}^T\mathbf{X}_\gamma|$ which implies that $p(\gamma|\mathbf{y}, \boldsymbol{\alpha})$ prefers models with few predictors and small residual variation.

Equation (11) can be used in a Gibbs sampling algorithm that draws each γ_i given γ_{-i} (the elements of γ other than γ_i). Each full conditional distribution is proportional to equation (11), and γ_i can only assume two possible values. Notice that $p(\gamma|\mathbf{y}, \boldsymbol{\alpha})$ only requires matrix computations for those variables that are actually included in the model.

Thus if the model is sparse the Gibbs sampler involves many inexpensive decompositions of small matrices, which makes SSVS computationally tractable even for problems with a relatively large number of predictors.

References

Concha Arola and Enrique Galan. Tracking the future on the web: Construction of leading indicators using internet searches. Technical report, Bank of Spain, 2012. URL http://www.bde.es/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/Fich/do1203e.pdf.

Yan Carrière-Swallow and Felipe Labbé. Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 2011. doi: 10.1002/for.1252. URL http://ideas.repec.org/p/chb/bcchwp/588.html. Working Papers Central Bank of Chile 588.

C. K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3): 541–553, 1994.

Jennifer L. Castle, Xiaochuan Qin, and W. Robert Reed. How to pick the best regression equation: A review and comparison of model selection algorithms. Technical Report 13/2009, Department of Economics, University of Canterbury, 2009. URL http://www.econ.canterbury.ac.nz/RePEc/cbt/econwp/0913.pdf.

Jennifer L. Castle, Nicholas W. P. Fawcett, and David F. Hendry. Evaluating automatic model selection. Technical Report 474, Department of Economics, University of Oxford, 2010. URL http://economics.ouls.ox.ac.uk/14734/1/paper474.pdf.

H. Chipman, E.I. George, R.E. McCulloch, M. Clyde, D.P. Foster, and R.A. Stine. The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.

Hyunyoung Choi and Hal Varian. Predicting the present with Google Trends. Technical report, Google, 2009a. URL http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Hyunyoung Choi and Hal Varian. Predicting initial claims for unemployment insurance using Google Trends. Technical report, Google, 2009b. URL http://research.google.com/archive/papers/initialclaimsUS.pdf.

Hyunyoung Choi and Hal Varian. Using search engine data for nowcasting—an illustration. In *Actes des Rencontrees Ećonomiques*, pages 535–538, Aix-en-Provence, FRANCE, 2011. Recontres Ećonomiques d′Aix-en-Provence, Le Cercle des ećonomistes. URL `http://www.lecercledeseconomistes.asso.fr/IMG/pdf/Actes_Rencontres_Economiques_d_Aix-en-Provence_2011.pdf`.

Piet de Jong and Neil Shepard. The simulation smoother for time series models. *Biometrika*, 82(2):339–350, 1995.

J. Durbin and S. J. Koopman. A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–616, 2002.

James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.

S. Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15(2):183–202, 1994.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis (2nd ed)*. Chapman & Hall, 2002.

E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–374, 1997.

Edward I. George and Robert E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373, 2007. URL `http://www3.stat.sinica.edu.tw/statistica/oldpdf/A7n26.pdf`.

Joyee Ghosh and Merlise A. Clyde. Rao-blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *Journal of the American Statistical Association*, 106(495):1041–1052, 2011.

Andrew Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1991.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2 edition, 2009.

Rebecca Hellerstein and Menno Middeldorp. Forecasting with internet search data. *Liberty Street Economics Blog of the Federal Reserve Bank of New York*,

Jan 4 2012. URL http://libertystreeteconomics.newyorkfed.org/2012/01/forecasting-with-internet-search-data.html.

D. M. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1335–1346, 1994.

Nick McLaren and Rachana Shanbhoge. Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, June 2011. URL http://www.bankofengland.co.uk/publications/quarterlybulletin/qb110206.pdf.

Giovanni Petris, Sonia Petrone, and Patrizia Campagnoli. *Dynamic Linar Models with R*. Springer, 2009.

H. Rue. Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338, 2001. ISSN 1467-9868.

James Stock and Mark Watson. Dynamic factor models. In M. Clements and D. Hendry, editors, *Oxford Handbook of Economic Forecasting*. Oxford University Press, 2010. URL http://www.economics.harvard.edu/faculty/stock/files/DynamicFactorModels.pdf.

Tanya Suhoy. Query indices and a 2008 downturn: Israeli data. Technical report, Bank of Israel, 2009. URL http://www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf.