Big Data and the Macro Economy: A Bayesian Mixed Frequency Estimator

David Kohns & Arnab Bhattacharjee

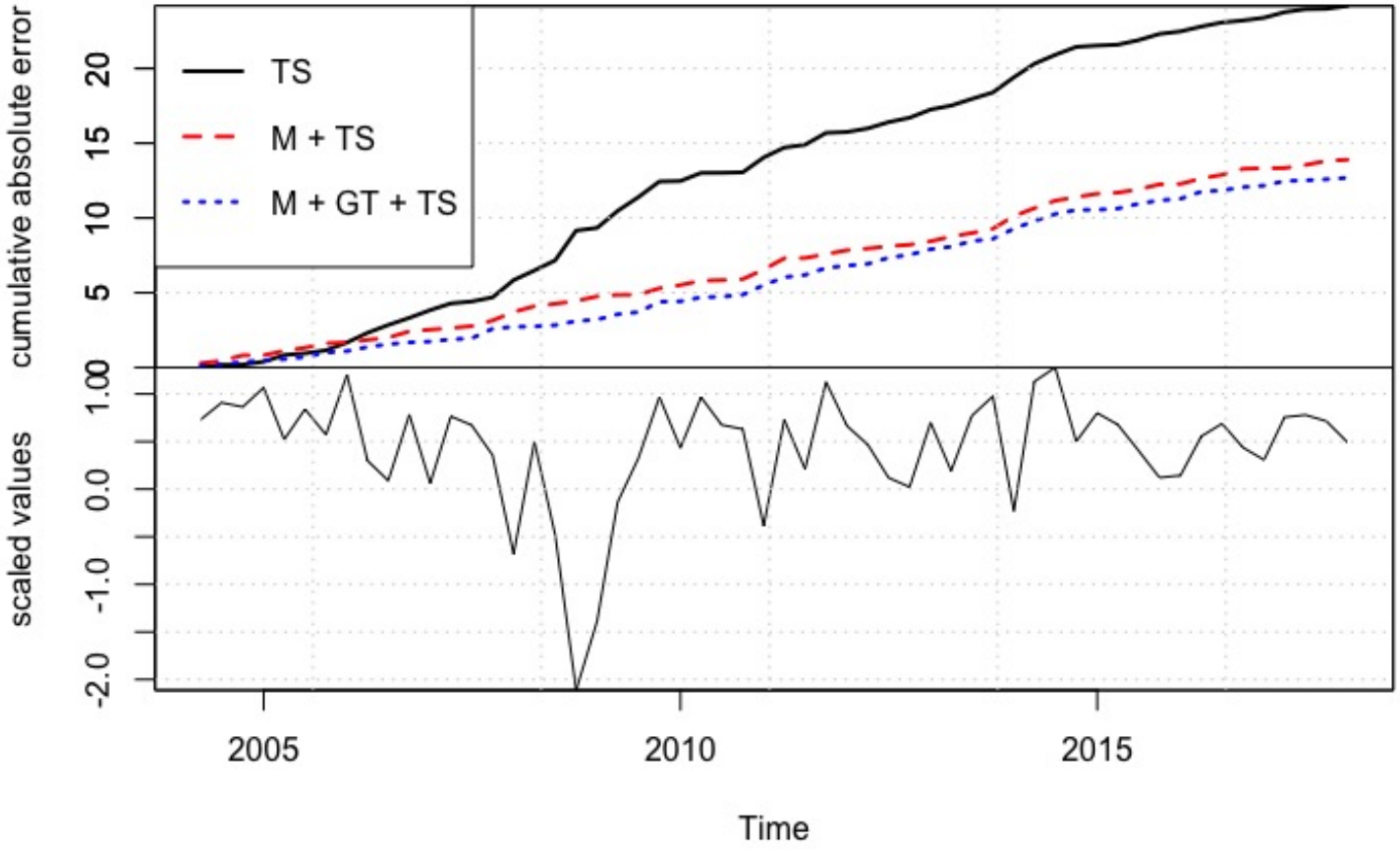
Heriot-Watt University

November 04, 2019

Outline

1. Introduction
2. Mixed Frequency Estimators - Previous Findings
3. MF-BSTS
4. Data
5. Results & Discussion
6. Way Forward
7. References

Cumulative Absolute Errors



Introduction

- Institutional analysts and policy makers demand early estimates of key economic variables (GDP, Inflation, IP) due to publication lags
- Early estimates are based more and more on information of Big Data sources such as Google Trends (GT)
- Especially GT have received a lot of attention in nowcasting literature (see Choi and Varian, 2012)
- Big data issues in forecasting:
  i) Mixed Frequency (MF)
  ii) Asynchronous data publications (ragged edges)
  iii) High dimensionality (HD)
  iv) Parameter changes
  v) Degree of Sparsity

Purpose and Contribution

- This paper: offers methodology which handles econometric issues (i)-(v) while remaining computationally simple and allowing for tractability of high frequency Big Data through
  - Marginal inclusion probabilities
  - Unrestricted intra-period dynamics
- We use a prior which stays agnostic about the underlying sparsity representation by separating variable selection from shrinkage
- Methodology is showcased by using Google Trends to augment nowcasts of real US GDP growth

Mixed Frequency Estimators

- Bańbura et al. (2013) classify mixed frequency estimators into (i) full system and (ii) partial system estimators
- Main difference: full system models jointly estimate all regressors
- Findings from the mixed frequency nowcasting literature:

  1. Direct partial system methods with Bayesian shrinkage can perform as well as full-system models (Carriero et al., 2015)
  2. When the frequency mismatch is low, U-MIDAS performs better than MIDAS methods (Foroni et al., 2014) Detail MIDAS

- Findings from broader forecasting literature:

  1. There is uncertainty over the degree sparsity (Giannone et al., 2018)
     - Traditional in macroeconomics: variants of factor estimation (e.g. Stock & Watson, 2002)
     - Common for Big Data sources: variants of sparse estimation (e.g. Tibshirani, 1996)
  2. Allowing for model uncertainty in correlated designs can improve forecasts (Steel, 2018)

Bayesian Structural Time Series augmented U-MIDAS

- The BSTS-U-MIDAS estimator seeks to combine these 4 findings into 1 model

- The proposed estimator consists itself of three distinct parts:
  1. "Structural" time series model (Harvey, 2006) augmented with mixed frequency regressors
     - Can flexibly handle time series features (seasonality, long term trends, unobserved components)
  2. Spike and Slab priors (George & McCulloch, 1997)
     - Two-component prior which allows for parsimony and controls model complexity
  3. Bayesian model averaging (Raftery et al., 2005)
     - Allows for model uncertainty and makes predictive inference more robust

Bayesian Structural Time Series augmented U-MIDAS

$$y_t = \mu_t + z_t + v_t, v_t \sim N(0, \sigma^2) \tag{1}$$

$$\mu_t = \mu_{t-1} + w_t, w_t \sim N(0, W) \tag{2}$$

$$z_t = \beta' X_{t,m} \tag{3}$$

- $\boldsymbol{y} = (y_1, \cdots, y_T)$ are the observations of the dependent variable at quarterly frequency
- $X_{t,m} = [x_{1,t,m}, \cdots, x_{K,t,m}]$, $m \in \{1, 2, 3\}$ are a monthly variables, transformed to quarterly frequency
- $\boldsymbol{X}$ is a $T \times K$ matrix, where the original dimension of $\boldsymbol{X}$ before skip-sampling is $TM \times K/M$ Detail
- Adding regressors in (3) increases dimensionality of state system only by 1
- States are estimated by Kohn & Carter (1994) algorithm

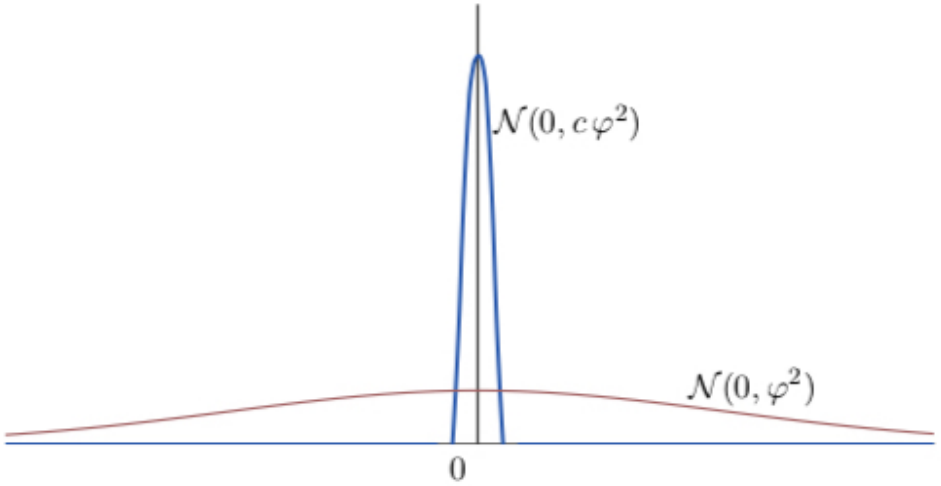
Spike and Slab Priors I

- We apply Spike and Slab priors to: $y_t - \mu_t$

$$p(\beta, \sigma^{-2}, \gamma) = p(\beta_\gamma | \gamma, \sigma^{-2}) p(\sigma^{-2} | \gamma) p(\gamma) \tag{4}$$

- Prior for β is a mixture of two normals with c $= 0$

$$\beta_j | \gamma \sim \gamma_j N(0, g(X_\gamma' X_\gamma)^{-1}) + (1 - \gamma_j) N(0, cg(X_\gamma' X_\gamma)^{-1}) \tag{5}$$



Spike and Slab Priors II

- The latent indicator variable γ is modelled as a Bernoulli random variable

$$\gamma \sim \prod_j \pi_j^{\gamma_j} (1 - \pi_j)^{1-\gamma_j} \tag{6}$$

- Conditional on knowing γ, we can apply conjugate priors which will significantly speed up computation
- We use a non-informative prior for the standard deviation of the measurement σ

$$p(\sigma) \propto 1/\sigma \tag{7}$$

Posterior Simulation

- Parameters to estimate $\theta = \{\beta, \sigma, \phi, W_1, W_2\}$
- States to estimate $\alpha = \{\mu_t\}$
- Using Bayes' rule, the full posterior is:

$$p(\theta, \boldsymbol{\alpha}|\mathbf{y}) = p(\theta)p(\alpha_0)\prod_{t=1}^{T} p(y_t|\alpha_t, \theta)p(\alpha_t|\alpha_{t-1}, \theta) \tag{8}$$

- Since the conditional posteriors all take well known standard forms, we can rely on standard Gibbs MCMC sampling

Posterior Marginal Inclusion Probability

- The dirac prior creates a reducible Markov chain, so we need to marginalise over the regression parameters to conduct variable selection
- Conditional on the states and state variances, the marginal posterior is analytically available:

$$p(\gamma|y) \propto (1+g)^{-q_\gamma/2} S(\gamma)^{-T/2} \prod_{j}^{K} \pi_j^{\gamma_j} (1-\pi_j)^{1-\gamma_j} \tag{9}$$

- q_γ is defined as the length of a given β_γ and $S(\gamma)$ is a simple function of y and X_γ
- All information needed for variable selection is in (9).

Bayesian Model Averaging

- Sampling algorithm explores the posterior model space which produces with each iteration a different model
- Bayes' Law provides a formal probabilistic way of conducting predictive inference over the drawn models
- Define $\phi = (\theta, \alpha)$, y^* as the forecast of y_t where M_j is a model selected from a set of K possible models and $\phi_j \in \Phi_j$ groups the unknown parameters in M_j :

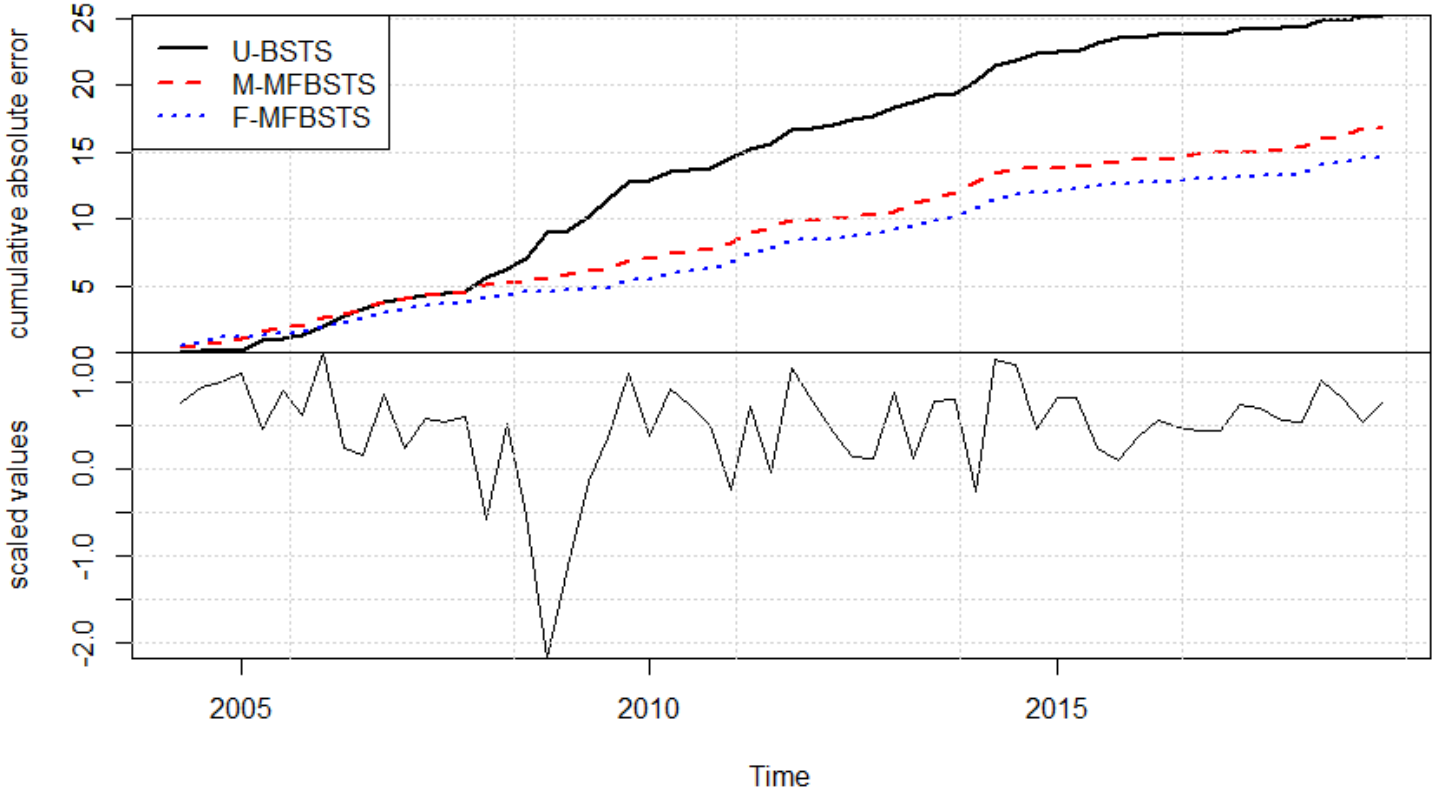
$$p(y^*|\boldsymbol{y}) = \sum_{j=1}^{K} \left[\int_{\Phi_j} p(\boldsymbol{y}^*|\phi_j, \boldsymbol{y}, M_j) p(\phi_j|\boldsymbol{y}, M_j) d\phi_j \right] P(M_j|\boldsymbol{y}) \tag{10}$$

- Each forecast is weighted by its respective marginal likelihood

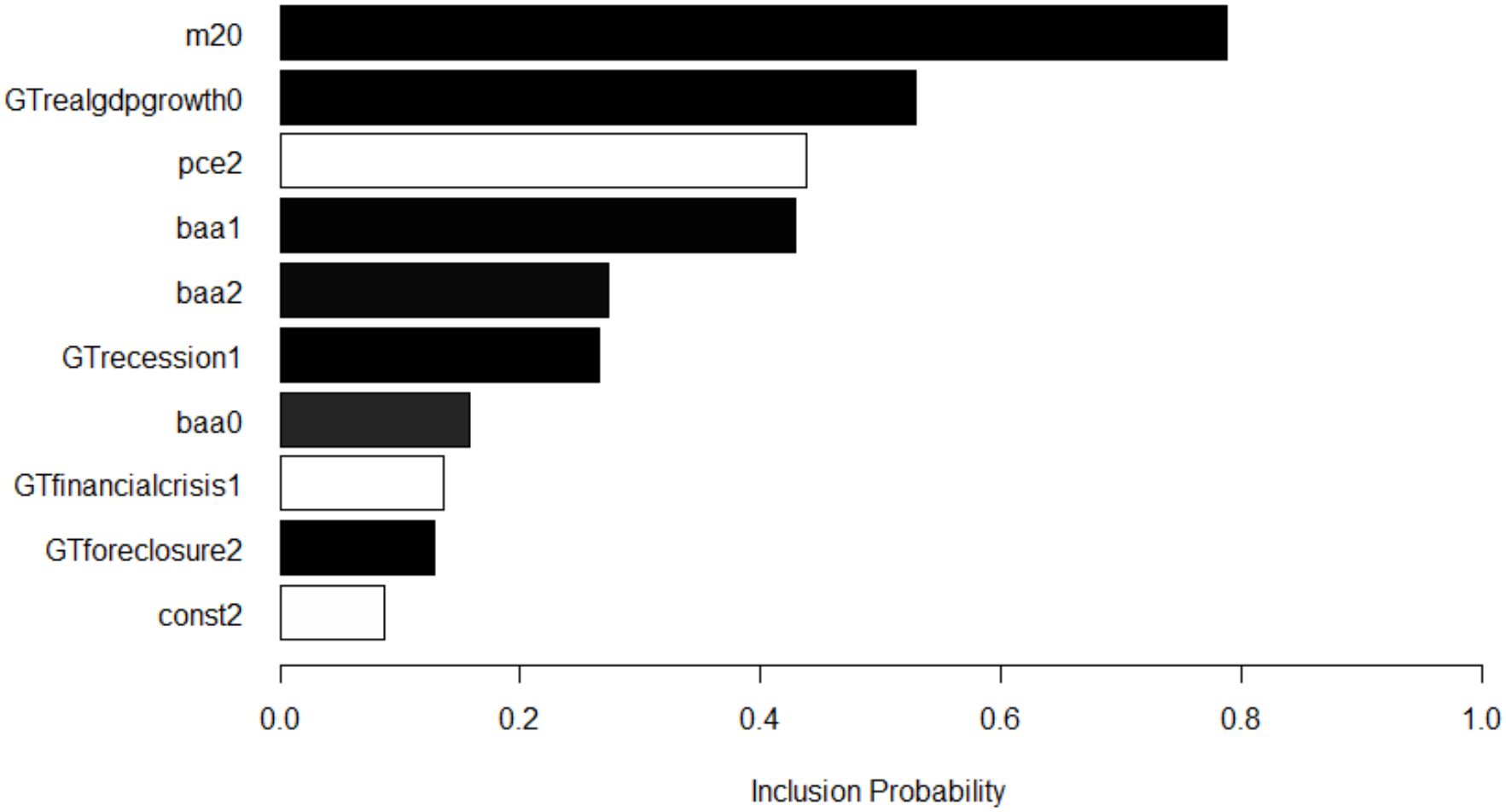
Data

- The dependent variable is real quarterly US GDP growth (second vintage as in Croushore, 2006)
- The independent variable set contains 13 macroeconomic indicators based on Giannone et al. (2016) and a selection of 27 GT
- GT are selected based on root term methodology
- In total 120 variables and 60 real growth observations
- Time frame: 01/04-03/19
- Training sample time frame: 01/04-07/14
- Pseudo Real Time calendar used for nowcasting includes 30 vintages

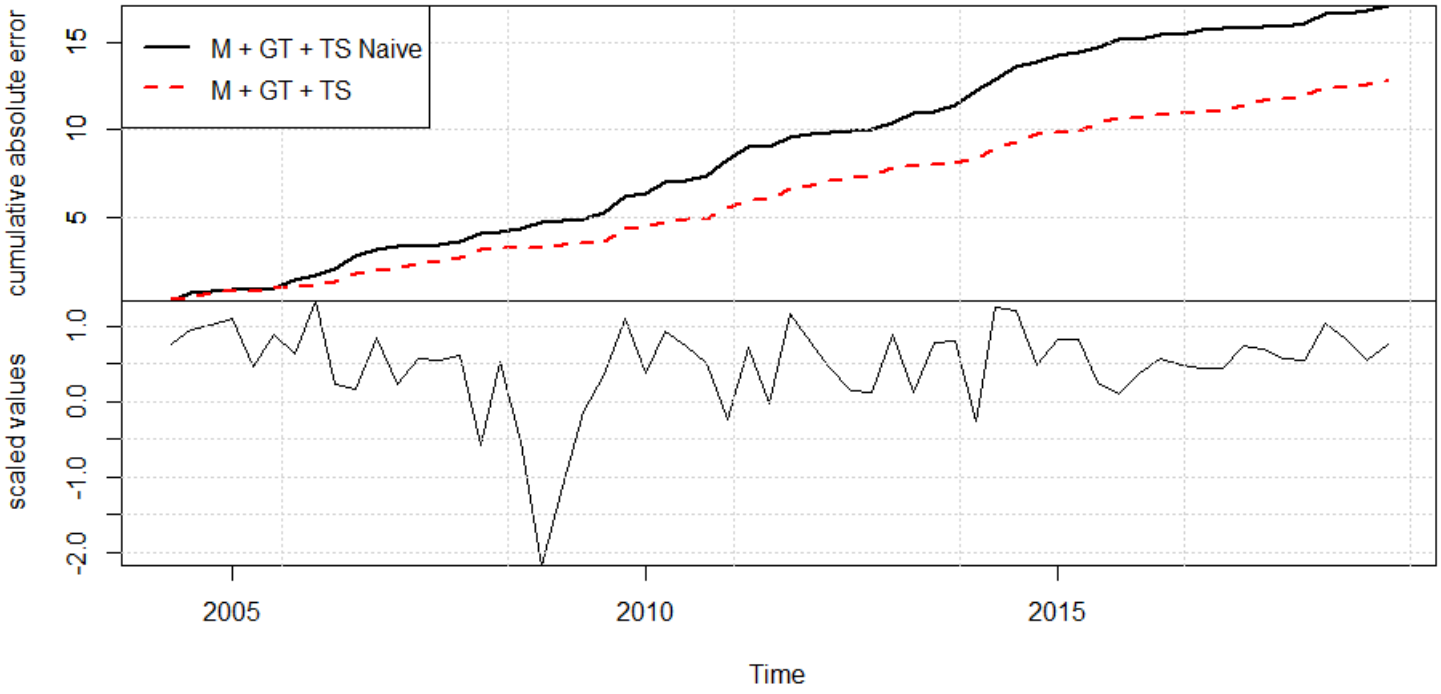
Cumulative Absolute Error



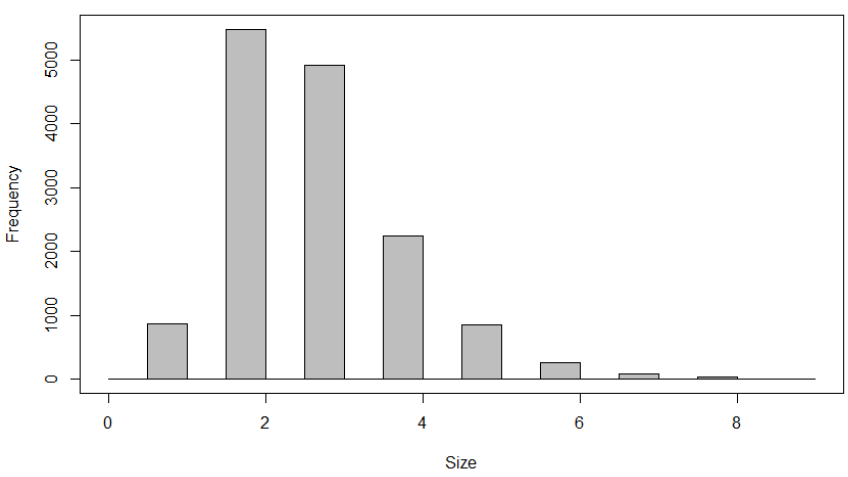
Marginal Inclusion Probabilities



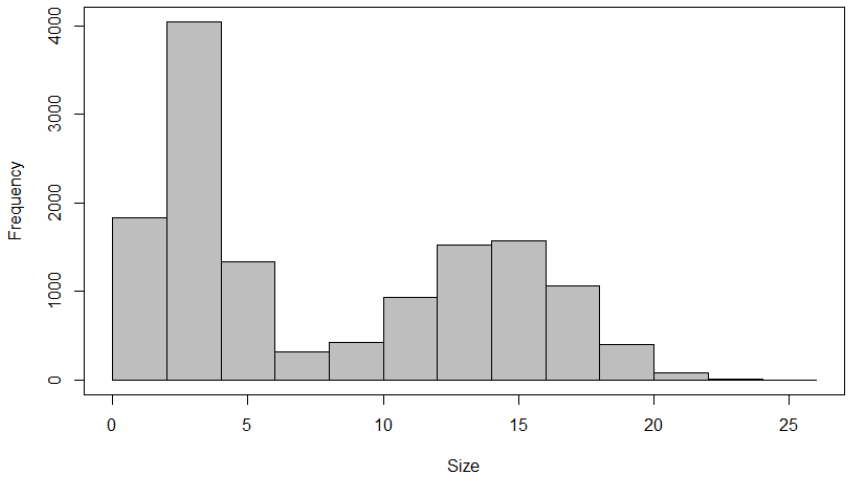
Naive vs. Unrestricted Mixed Frequency



Posterior Model Sizes



(a) $\pi = 12$



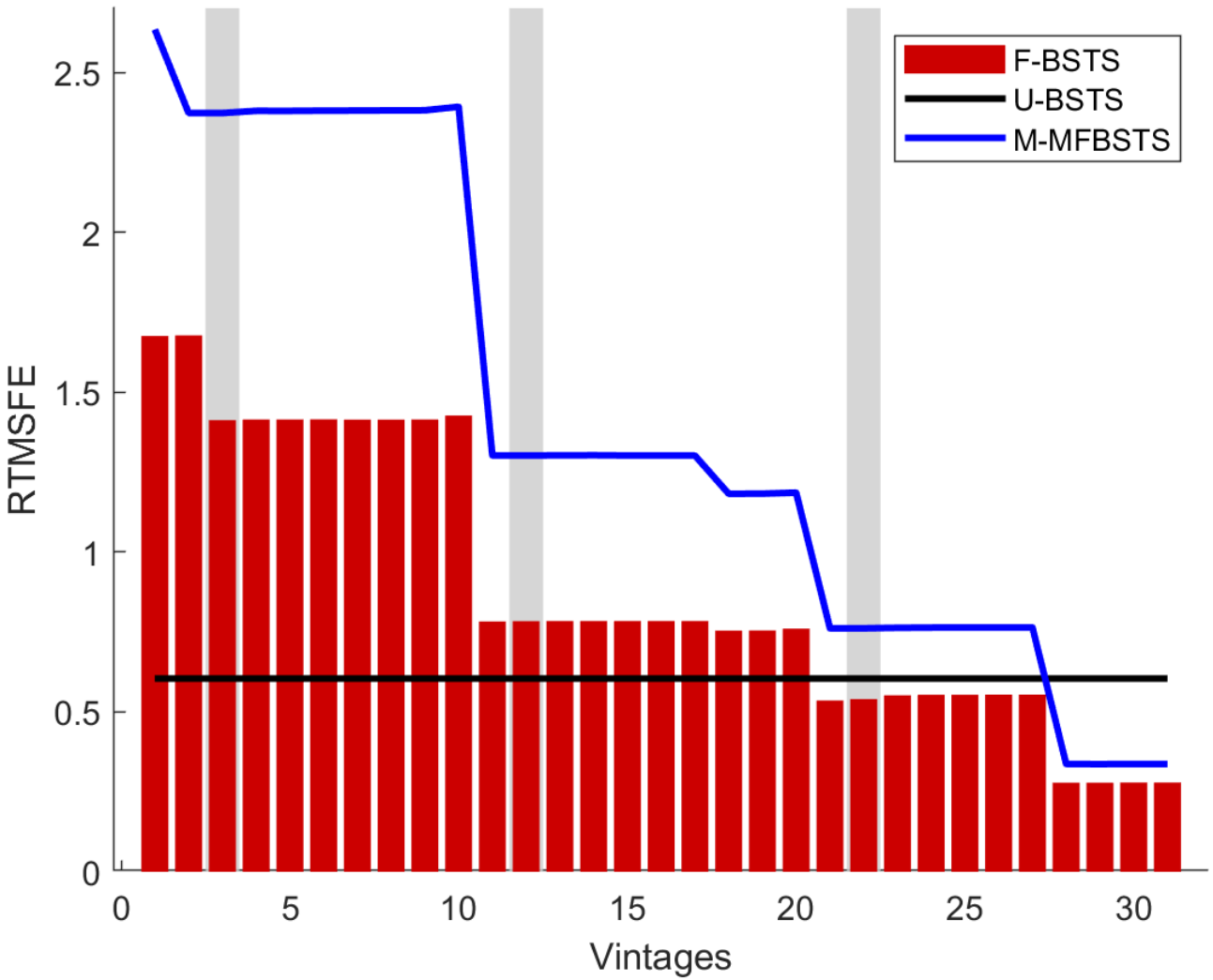
(b) $\pi = 18$

- Prior expected model sizes below 12 result in heavily shrunk models
- Prior expected model sizes above 18 result in bi-modal posterior
- Larger models include far larger proportion of macro variables, while Google Trends remain largely the same
- Completely hierarchical prior does not explore large models due to fixed global shrinkage

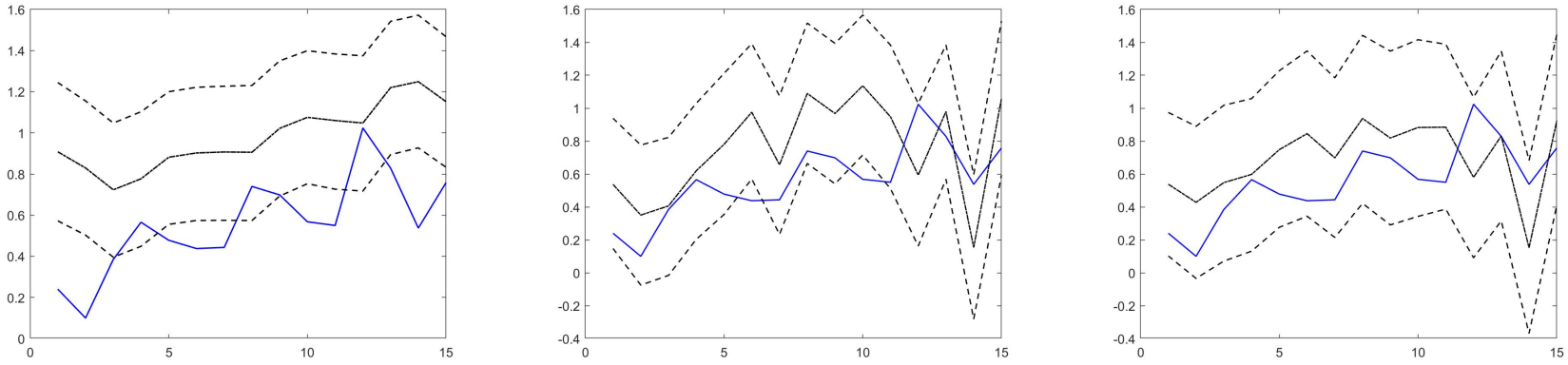
Pseudo Real Time Calendar

Vintage	Timing	Release	Variable Name	Pub. lag	Transformation	FRED Code
1	First day of month 1	-	-	-	-	-
2	Last day of month 1	Fed fund & BAA	fedfunds & baa	m	3	FEDFUNDS & BAAY10
3	Last day of month 1	Google Trends		m	4	-
4	1st bus. day of month 2	Economic Policy Index	uncertainty	m-1	1	USEPUINDXM
5	1st Friday of month 2	Employment situation	hours & unrate	m-1	2	AWHNONAG & UNRATE
6	Middle of month 2	CPI	cpi	m-1	2	CPI
7	15th-17th of month 2	Industrial Production	indpro	m-1	2	INDPRO
8	3rd week of month 2	Credit & M2	loans & m2	m-1	2	LOANS & M2
9	Later part of month 2	Housing starts	housst	m-1	1	HOUST
10	Last week of month 2	PCE & PCEPI	pce & pce2	m-1	2	PCE & PCEPI
11	Last day of month 2	Fed fund & BAA	fedfunds & baa	m	3	FEDFUNDS & BAAY10
12	Last day of month 2	Google Trends		m	4	-
13	1st bus. day of month 3	Economic Policy Index	uncertainty	m-1	1	USEPUINDXM
14	1st bus. day of month 3	Construction starts	construction	m-2	1	TTLCONS
15	1st Friday of month 3	Employment situation	hours & unrate	m-1	2	AWHNONAG & UNRATE
16	Middle of month 3	CPI	cpi	m-1	2	CPI
17	15th-17th of month 3	Industrial Production	indpro	m-1	2	INDPRO
18	3rd week of month 3	Credit & M2	loans & m2	m-1	2	LOANS & M2
19	Later part of month 3	Housing starts	housst	m-1	1	HOUST
20	Last week of month 3	PCE & PCEPI	pce & pce2	m-1	2	PCE & PCEPI
21	Last day of month 3	Fed fund & BAA	fedfunds & baa	m	3	FEDFUNDS & BAAY10
22	Last day of month 3	Google Trends		m	4	-
23	1st bus. day of month 4	Economic Policy Index	uncertainty	m-1	1	USEPUINDXM
24	1st bus. day of month 4	Construction starts	construction	m-2	1	TTLCONS
25	1st Friday of month 4	Employment situation	hours & unrate	m-1	2	AWHNONAG & UNRATE
26	Middle of month 4	CPI	cpi	m-1	2	CPI
27	15th-17th of month 4	Industrial Production	indpro	m-1	2	INDPRO
28	3rd week of month 4	Credit & M2	loans & m2	m-1	2	LOANS & M2
29	Later part of month 4	Housing starts	housst	m-1	1	HOUST
30	Last week of month 4	PCE & PCEPI	pce & pce2	m-1	2	PCE & PCEPI
31	Later part of month 5	Construction starts	const	m-2	1	TTLCONS

Real Time MSFE



Density Nowcasts: Last Vintage



(a) U-BSTS (b) M-BSTS (c) F-BSTS

- Full Model provides best density forecasts
- Noise in Google search terms results in greater uncertainty around the the full model forecasts

Discussion

1. Google Trends improve nowcasts of GDP. Act as an early warning signal & uncertainty index akin to financial variables
2. GT improve nowcasts early in the quarter prior to release of official information
3. Marginal inclusion probabilities & unrestricted MIDAS sampling help to uncover the mechanisms through which GT improve forecasts
4. Marginal inclusion probabilities also show limitations of the model: Reverse causality
5. Posterior model space includes regions of smaller & larger models when allowing for informative priors. However, large models suffer from estimation uncertainty.
6. Uncertainty over sparsity calls for priors which are able to handle dense and sparse model components

Way Forward

- The prior does not give us direct answers to whether regressors belong into a dense or sparse model part
  - Using a multi-layer Spike and Slab algorithm could sort into dedicated dense and sparse model parts (under active investigation of the co-authors)
- Further extensions to improve fit:
  - Tighter prior on state component so as to reduce noise in early vintages: the inverse-Gamma distribution biases the state variance away from zero.
  - Stochastic volatility
  - Prior which allows for varying degrees of shrinkage: e.g. global local priors
  - Pre-screening approach akin to Fan & Lv (2008) could help with noise reduction in ultra high dimensional settings such as using larger unstructured Big Data sources

References I

Andreou et al. (2013). Should Macroeconomic Forecasters Use Daily Financial Data and How? *Journal of Business & Economic Statistics*, 31, 240-251.

Banbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. *Handbook of economic forecasting*, 195-233.

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

Scott, S., Varian, H. (2014). Predicting the Present with Bayesian Structural Time Series. *International Journal of Mathematical Modelling and Numerical Optimisation*, Vol. 5, pp. 4-23.

MIDAS Style Skip-Sampling

- Suppose y_t is observed quarterly, and x_M is observed monthly

$$\left(\begin{array}{c|ccc} y_{1stquarter} & x_{Mar} & x_{Feb} & x_{Jan} \\ y_{2ndquarter} & x_{Jun} & x_{May} & x_{Apr} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \end{array}\right) \tag{11}$$

- Skip sampling is flexible enough to handle any kind of frequency mismatch and ragged edges
- ragged egdes are handles either through setting values to zero prior to release, or by re-aligning the vectors
- Higher frequency mismatch merely expands the column space of the design

Back to main.

Partial System Estimators I

- The rationale of Ghysels' et al. (2004) MIDAS style estimators is to take data driven averages of high frequency observations
- Kernel functions of these averages are usually parsimonious
- Define y_t as the dependent low frequency variable, where $t \in \{1, \cdots, T\}$, $x_{m,t}$ as the higher frequent regressors observed M times more frequent than t, where $m \in \{1, \cdots, M\}$
- Additive Distributed Lag (ADL)-MIDAS:

$$y_{t+1} = \alpha + \sum_{i=0}^{p_y-1} \rho_{i+1} y_{t-i} + \beta \sum_{i=0}^{p_x-1} \sum_{j=0}^{M-1} w_{M-j+i*M}(\theta) x_{M-j,t-i} + \epsilon_t$$

- $w_m(\theta)$ is a low-polynomial weighting function such as the Almon lag kernel:

$$w_m(\theta) = \frac{exp(\theta_1 m + \theta_2 m^2)}{\sum_{j=1}^{M} exp(\theta_1 j + \theta_2 j^2)}$$

Back to main.

Partial System Estimators II

- The Unrestricted (U-)-MIDAS estimator instead estimates each higher frequent observation separately:

$$y_{t+1} = \alpha + \beta_1 x_{1,t} + \cdots + \beta_M x_{m,t} + \epsilon_t$$

- Factor-ADL-MIDAS extracts first principal components of the higher frequent cross-section and applies a low-polynomial weighting function to the resulting higher frequent factors:

$$y_{t+1} = \alpha + \sum_{i=0}^{\rho_y - 1} \rho_{i+1} y_{t-i} + \sum_{i=0}^{\rho_f - 1} \phi_{i+1} F^q_{t-i} + \epsilon_t$$

- F^q_{t-i} is aggregated according to $F^q_t = \sum_{j=0}^{L-1} w_{L-j}(\theta) f_{L-j,t}$.

Back to main.

Full System Estimators

- In the Full-system approach of Banbura et al. (2013) and Koop et al. (2018) is to treat the low frequency variable as being higher frequent with missing observations
- Missing observations are then treated as latent states
- Banbura et al. (2013) estimate a mixed frequency dynamic factor model (MF-DFM) via Mariano & Murasawa's (2010) EM algorithm:

$$Y_t^m = \Lambda^m F_t^m + E_t^m$$

$$F_t^m = A(L) F_{t-1}^m + U_t^m$$

- Define $y_t^q, y^{\{}m, t\}, z_t$ as the quarterly, monthly unobserved observations and a stacked vector respectively Koop et al. (2018) estimate a mixed frequency VAR as:

$$y_t = \phi_0 + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + u_t$$

$$y_t^q = M_t^q \Lambda^q z_t$$

Back to main.

Top 10 Models

Variables	1	2	3	4	5	6	7	8	9	10
Constant	3.8943	3.8943	3.8943	3.8943	3.8943	3.8943	3.8943	3.8943	3.8943	3.8943
Baa0	0	-0.086643	0	0	0	0	-0.086643	-0.086643	0	0
Baa1	0	0	0	0	0	-0.20575	0	0	0	-0.20575
Loans2	0.012742	0.012742	0.012742	0	0	0	0	0.012742	0	0
M20	-0.011817	-0.011817	-0.01181	-0.011817	-0.011817	-0.011817	-0.011817	-0.011817	-0.011817	-0.011817
M21	0	-0.00030418	0	0	0	0	0	0	0	0
Const0	0	0	1.03E-07	0	0	0	0	0	0	0
Const2	0	0	0	0	0	0	8.32E-07	0	0	0
Uncertainty2	0	0	0	0	0	0	9.23E-06	0	0	0
Bankruptcy0	0	0	0	0	0	0.00080785	0	0	0	0
GDPgrowth2	0	0	0	-0.00082407	0	-0.00082407	0	0	-0.00082407	0
Industrialproduction1	0.012386	0.012386	0.012386	0.012386	0.012386	0	0.012386	0.012386	0.012386	0.012386
Industrialproduction2	-0.00032733	0	-0.00032733	0	-0.00032733	0	-0.00032733	0	0	-0.00032733
Recession2	0	0	0	0	-0.000656	-0.000656	0	0	-0.000656	-0.000656
Stockmarket0	0	-0.0264	-0.0264	-0.0264	0	0	-0.0264	-0.0264	0	0
USdefault1	0	0	0	-2.33E-05	0	0	0	0	0	0

Table 1: Posterior coefficients of the top 10 models by marginal likelihood