A Bayesian framework for sparse estimation in high-dimensional mixed frequency Vector Autoregressive models

George Michailidis

gmichail@ucla.edu

CRETE 2023

Motivating Application: Forecasting GDP - I

- Gross Domestic Product forecasts available at quarterly frequency
- For the US, the first quarterly estimate becomes available four weeks after the end of the corresponding quarter
  For the Euro zone the first quarterly estimate comes at around six weeks
- A number of leading and coincident economic indicators are available at a monthly or higher frequency with no or little publishing delays
- Beneficial to diverse stakeholders to build forecasting models that leverage such data and also provide near term (one or two months ahead) forecasts

Motivating Application: Forecasting GDP - II

- The GDPNow project by the Atlanta Fed aims to provide frequent updates for the US GDP
- The GDPNow methodology segments the GDP contributors to 13 multivariate components (e.g., personal exenditures for goods and services, federal and state consumptions and investments, etc.) and builds a Bayesian Vector Autoregressive (BVAR) model for each of them
- The BVAR produces quarterly forecasts that are combined with monthly time series autoreregressive based forecasts through regression models
- Forecasting problems with mixed frequency data motivated work in Mixed Data Sampling (MIDAS) regression models (more details in the sequel)

Illustration of Mixed Frequency Data Structure

Consider one high frequency (x) and one low frequency (y) variable

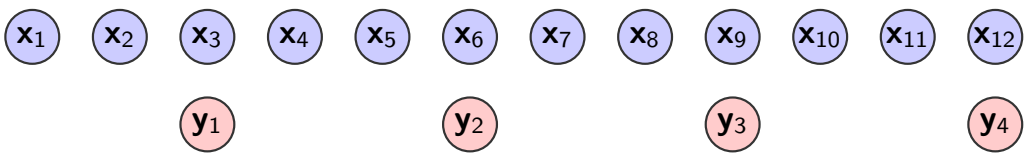


Figure 1: A diagram for monthly-quarterly; freq ratio $= 3$

Prediction Task:

Build a model to predict y (and possibly x) using past lags of y and x.

Illustration of Forecasting/Nowcasting Concepts

Forecast Task:

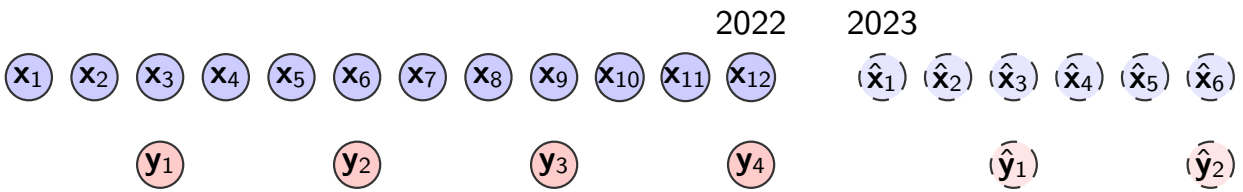


Figure 2: A diagram for the case where freq ratio = 3, forecast vintage, horizon=2 periods ahead

Nowcast Task:

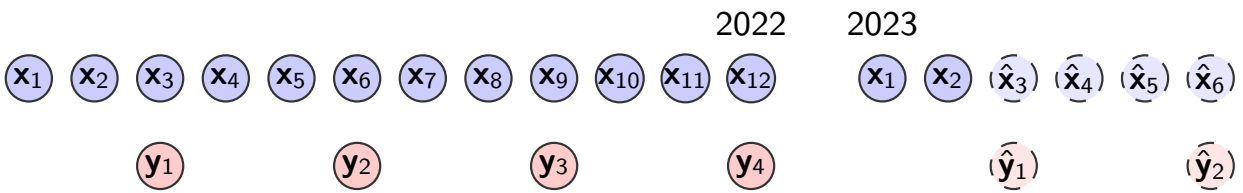


Figure 3: A diagram for the case where freq ratio = 3, nowcast vintage 2, horizon=2 periods ahead

Challenges

- Given the lack of frequency adherence in the data, how to organize the data to construct statistical/ML models
- In regression settings $y = f(x)$, forecasting $h > 1$ periods ahead, require a mechanism to obtain forecasts of the x predictors;
  in the GDPnow project, some of them are provided through the BVAR models, but for the remaining monthly indicators this remains a challenge
- How to obtain nowcasts?

Two Data Organizing Frameworks

1. Impute missing values of low frequency variables $\rightarrow$
   automatically addresses challenges regarding nowcasting - standard modeling at the high frequency cadence
2. Reorganize/align the high frequency variables $\rightarrow$
   model at the low frequency cadence

A Broad Classification of Existing Methods

- Existing methods largely fall into the following 3 categories, depending on the assumed data generating process:
  1. Autoregressive distributed lag (ADL) models:
     - Univariate low frequency response variables, predictors are lag values of itself and those of others
     - Model respecification required for nowcasting tasks
  2. Vector autoregressive (VAR) models:
     - Low and high frequency variables are modeled jointly as a multivariate system
     - Frequency alignment needed, all aligned variables evolve at the low frequency
     - Model re-specification across vintages required
  3. State-space models:
     - Latent high-frequency states drive the underlying dynamics
     - Both the low and the high-frequency blocks are functions of the governing states

- Models predominantly linear in nature, many equipped with theoretical guarantees, e.g.,
  - Statistical consistency of the estimated parameters
  - Prediction consistency

Vector Autoregressive (VAR) models

Consider the standard VAR model for p variables and d lags

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_d y_{t-d} + \varepsilon_t,$$

where $A_j, j = 1, \cdots, d$ are $p \times p$ transition matrices that capture lead-lag cross-autocorrelations

How to incorporate mixed frequency data in VAR models - the Frequency Alignment Approach

- Each high frequency variable is converted to F new time series evolving at the low frequency. For example, for monthly-quarterly data, each monthly time series is colnverted to 3 quarterly ones, the first containing January, April, July and October data, the second February, May, August, November data and soon
- See, e.g., Ghysels, 2016; Foroni et al., 2021, Chakraborty et al, 2022
- Quickly leads to very large VAR models that require regularization and care to train

Modeling Framework - I

- Consider 2 monthly variables; e.g. INDPRO (Industrial Production), UNRATE (Unemployment Rate) and,
- 2 quarterly variables; e.g. GDP (Gross Domestic Product), GPDI (Gross Private Domestic Investment)
- expand each monthly series into 3 new variables for the 1^{st}, 2^{nd} and 3^{rd} month in each quarter (observation-driven viewpoint)

$$\begin{pmatrix} y_{i,H}^{t-2/3} \\ y_{i,H}^{t-1/3} \\ y_{i,H}^{t} \end{pmatrix}$$

- value at the end of the first month of 't'-th quarter
- value at the end of the second month of 't'-th quarter
- value at the end of the last month of 't'-th quarter

- $\{y_{j,\mathrm{L}}^{t}\}_{t=1}^{T}$: quarterly series for each quarterly (low-frequency) variable at quarter 't'

Modeling Framework - II

Focus on a VAR(1) model, after implementing frequency alignment

$$y^t = \mathbf{W} y^{t-1} + \varepsilon^t \tag{1}$$

i.e.
$$\begin{bmatrix} y_{1,H}^{t-2/3} \\ y_{1,H}^{t-1/3} \\ y_{1,H}^{t} \\ y_{2,H}^{t-2/3} \\ y_{2,H}^{t-1/3} \\ y_{2,H}^{t} \\ y_{1,L}^{t} \\ y_{2,L}^{t} \end{bmatrix} = \underbrace{\begin{bmatrix} \underbrace{\mathbf{W}_{H,H}}_{\text{monthly block}} & \underbrace{\mathbf{W}_{H,L}}_{\text{quarterly on monthly}} \\ \underbrace{\mathbf{W}_{L,H}}_{\text{monthly on quarterly}} & \underbrace{\mathbf{W}_{L,L}}_{\text{quarterly block}} \end{bmatrix}}_{\mathbf{W}} \begin{bmatrix} y_{1,H}^{t-5/3} \\ y_{1,H}^{t-4/3} \\ y_{1,H}^{t-1} \\ y_{2,H}^{t-5/3} \\ y_{2,H}^{t-4/3} \\ y_{2,H}^{t-1} \\ y_{1,L}^{t-1} \\ y_{2,L}^{t-1} \end{bmatrix} + \varepsilon^t$$

Modeling Framework - III

- The temporal dependence structure of the model is characterized by the transition matrix $\mathbf{W}$.
- $\varepsilon^t \overset{iid}{\sim} \mathcal{N}(0, \Sigma_\varepsilon)$
- General setup: k_1 monthly and k_2 quarterly variables
- Total no. of time series: $p = (3 * k_1 + k_2)$

Modeling Framework - IV

Challenges:

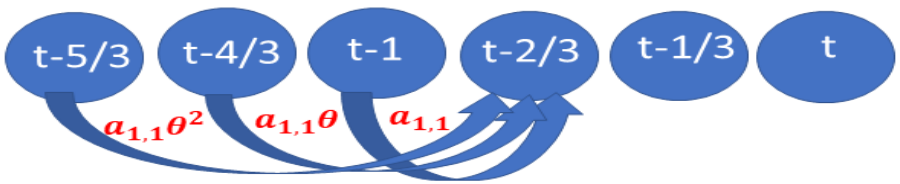
- High dimensional $\Longrightarrow$ when both k_1 and/or k_2 are comparable to or larger than the sample size.
- For VAR models, no. of parameters grows *quadratically* in the no. of time series. Parameter space is even larger for mixed-frequency data under this set-up.

We adopt a novel approach for parameter reduction through

- structural relationships between appropriate VAR coefficients
- regularization (sparsity)

Modeling Framework - V

- Structural Assumptions: the effect is dampened by a factor θ when moving back one month in time

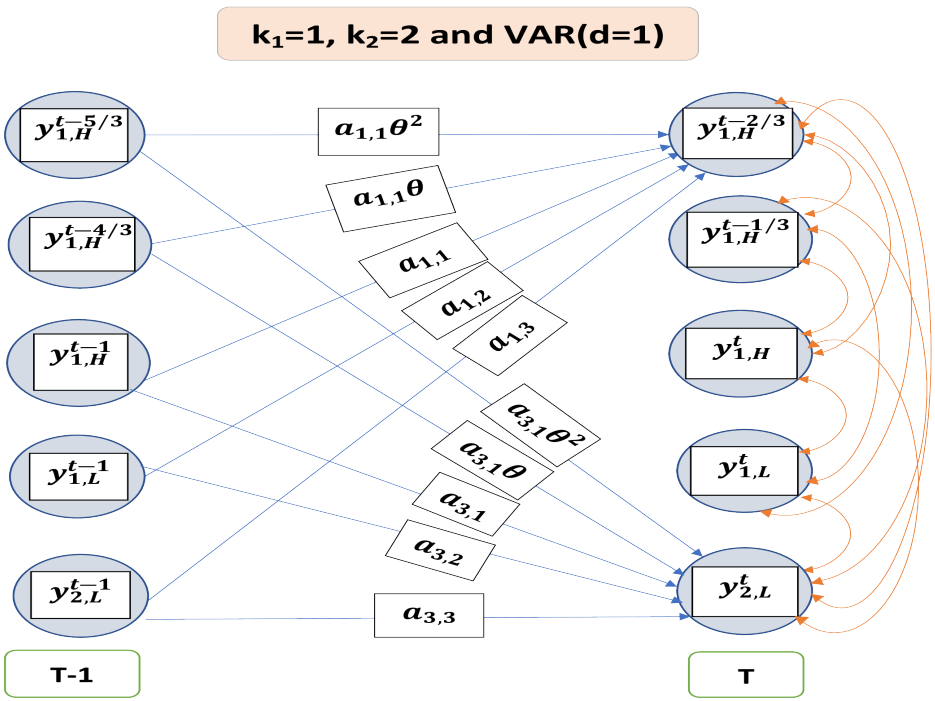


- Impose following reducing assumptions on $\mathbf{W}_{H,H}$ and similarly on other blocks.

$$\mathbf{W}_{H,H} = \begin{bmatrix} a_{1,1}\theta^2 & a_{1,1}\theta & a_{1,1} & \text{dampening by } \theta \leftarrow & a_{1,2} \\ a_{1,1}\theta^3 & a_{1,1}\theta^2 & a_{1,1}\theta & & \downarrow \\ a_{1,1}\theta^4 & a_{1,1}\theta^3 & a_{1,1}\theta^2 & & \text{dampening by } \theta \\ \text{dampening by } \theta & \leftarrow & a_{2,1} & \text{dampening by } \theta \leftarrow & a_{2,2} \\ & & \downarrow & & \downarrow \\ & & \text{dampening by } \theta & & \text{dampening by } \theta \end{bmatrix}$$

Modeling Framework - VI

- VAR(1) model with $k_1 = 1$ and $k_2 = 2$.



- Temporal dependence of all the lagged values on $y_{1,\mathrm{H}}^{t-2/3}$ and $y_{2,\mathrm{L}}^{t}$ (blue directed edges)
- Contemporaneous dependence (red bi-directed edges)

Modeling Framework - VII

- Let $\mathbf{A} = ((a_{i,j}))_{1 \leq i,j \leq k_1+k_2}$, such that

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11\, k_1 \times k_1} & \mathbf{A}_{12\, k_1 \times k_2} \\ \mathbf{A}_{21\, k_2 \times k_1} & \mathbf{A}_{22\, k_2 \times k_2} \end{bmatrix},$$

- $\mathbf{A}_{11}$: HF block of cross-autoregressive effects
- $\mathbf{A}_{12}$: LF on HF block of cross-autoregressive effects
- $\mathbf{A}_{12}$: HF on LF block of cross-autoregressive effects
- $\mathbf{A}_{22}$: LF block of cross-autoregressive effects

Modeling Framework - VII

- Let $\mathbf{A} = ((a_{i,j}))_{1 \leq i,j \leq k_1+k_2}$, such that

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11\, k_1 \times k_1} & \mathbf{A}_{12\, k_1 \times k_2} \\ \mathbf{A}_{21\, k_2 \times k_1} & \mathbf{A}_{22\, k_2 \times k_2} \end{bmatrix},$$

- $\mathbf{A}_{11}$: HF block of cross-autoregressive effects
- $\mathbf{A}_{12}$: LF on HF block of cross-autoregressive effects
- $\mathbf{A}_{12}$: HF on LF block of cross-autoregressive effects
- $\mathbf{A}_{22}$: LF block of cross-autoregressive effects

- Can we express $\mathbf{W}$ in terms of $\mathbf{A}$ and θ using the structural relationships between the coefficients?

Modeling Framework - VIII

- Form of **W**:

$$\mathbf{W} = \left[\begin{array}{ccc|c} \theta^2 & \theta^4 & \theta^3 & \\ 1 & \theta^2 & \theta & \left(\begin{array}{c}\theta^2\\1\\\theta\end{array}\right) \otimes \mathbf{A}_{12} \\ \theta & \theta^3 & \theta^2 & \\ \hline \multicolumn{3}{c|}{\left(\begin{array}{ccc} 1 & \theta^2 & \theta \end{array}\right) \otimes \mathbf{A}_{21}} & \mathbf{A}_{22} \end{array}\right] \left(\begin{array}{ccc}\theta^2 & \theta^4 & \theta^3\\ 1 & \theta^2 & \theta\\ \theta & \theta^3 & \theta^2\end{array}\right) \otimes \mathbf{A}_{11}$$

- Each entry of **A** has a very clear and specific interpretation.

- Parameter reduction through structural relationships:
  $(3k_1 + k_2)^2 \rightarrow (k_1 + k_2)^2 + 1$ no. of parameters
- Illustrative example: if $k_1 = 10, k_2 = 10$, then the number of parameters is reduced from 1,600 to $400 + \theta$

Modeling Framework - VIII

- Form of **W**:

$$\mathbf{W} = \left[\begin{array}{ccc|c} \theta^2 & \theta^4 & \theta^3 & \\ 1 & \theta^2 & \theta & \left(\begin{array}{c}\theta^2\\1\\\theta\end{array}\right) \otimes \mathbf{A}_{12} \\ \theta & \theta^3 & \theta^2 & \\ \hline \multicolumn{3}{c|}{\left(\begin{array}{ccc} 1 & \theta^2 & \theta \end{array}\right) \otimes \mathbf{A}_{21}} & \mathbf{A}_{22} \end{array}\right] \otimes \mathbf{A}_{11}$$

- Each entry of **A** has a very clear and specific interpretation.

- Parameter reduction through structural relationships:
  $(3k_1 + k_2)^2 \rightarrow (k_1 + k_2)^2 + 1$ no. of parameters
- Illustrative example: if $k_1 = 10, k_2 = 10$, then the number of parameters is reduced from 1,600 to $400 + \theta$

  Nevertheless, need for further reduction through regularization

Further Parameter Reduction through Regularization and a Pseudo-Likelihood Approach

- The main parameter of interest is $\mathbf{W}$, which due to frequency alignment can be very large
- Σ_ε is a nuisance parameter
- Assume,

$$\Sigma_\varepsilon = \text{diag}\left(\mathbf{\Sigma_{1,H}}, \cdots, \mathbf{\Sigma_{k_1,H}}, \sigma_{\mathbf{1}}^{\mathbf{2}}, \cdots, \sigma_{\mathbf{k_2}}^{\mathbf{2}}\right)$$

- Ignore cross-correlations between the high and low frequency variables.

Pros and Cons of a Pseudo-likelihood Approach

- $6k_1 + k_2$ parameters for Σ_ε, as opposed to $(3k_1 + k_2)(3k_1 + k_2 + 1)/2$ parameters
- Computationally simpler, but comes at a cost of some efficiency loss
- But, under high-dimensional scaling does it lead to consistent estimates?

Specification of Prior Distribution - I

- Define binary variables

$$\gamma_{ij} = 1_{a_{ij} \neq 0} \quad \forall\, 1 \leq i, j \leq k_1 + k_2,$$

- Define the activity graph $\mathcal{G}$ with edge set:

$$\mathcal{G} = \left\{\gamma_{ij}\right\}_{1 \leq i,j \leq k_1+k_2} : \quad \text{captures the pattern of active entries in } \mathbf{A}$$

Gaussian mixture prior distribution for entries of $\mathbf{A}$: (Spike-and-Slab)

- For first k_1 rows of $\mathbf{A}$: $(1 \leq i \leq k_1,\ 1 \leq j \leq k_1 + k_2)$

$$a_{ij} \mid \Sigma_{i,\mathrm{H}}, \quad \sim \quad (1 - \gamma_{ij})\, 1_{\{a_{ij}=0\}} + \gamma_{ij} \mathcal{N}(0, \tilde{\sigma}_i^2 \tau^2)$$

  $\tilde{\sigma}_i^2$: a suitable function of $\Sigma_{i,\mathrm{H}}^{-1}$

- For last k_2 rows of $\mathbf{A}$: $(k_1 + 1 \leq i \leq k_1 + k_2,\ 1 \leq j \leq k_1 + k_2)$

$$a_{ij} \mid \sigma_{i-k_1}^2, \quad \sim \quad (1 - \gamma_{ij})\, 1_{\{a_{ij}=0\}} + \gamma_{ij} \mathcal{N}(0, \sigma_{i-k_1}^2 \tau^2)$$

Specification of Prior Distribution - II

- $\gamma_{ij} \sim Bernoulli(q) \; \forall \; 1 \leq i, j \leq k_1 + k_2$
- $\Sigma_{i,\mathrm{H}} \overset{iid}{\sim}$ Inverse-Wishart$(\omega, \mathbf{V})$
- $\sigma_j^2 \overset{iid}{\sim}$ Inverse-Gamma(α, β)
- $\theta \sim U(0, 1)$

Posterior distribution

- Joint posterior of $\mathbf{A}, \theta, \{\Sigma_{i,\mathrm{H}}\}_{i=1}^{k_1}, \{\sigma_j^2\}_{j=1}^{k_2}$ is intractable
- The full conditional posterior distributions of all the parameters (except θ) are standard: for the entries of $\mathbf{A}$ spike and slab mixture distributions with normal slab components, for the variance parameters Wishart and Inverse-Gamma
  For θ: sample based on an efficient discrete approximation
- Hence, use a Gibbs sampling algorithm to generate approximate samples from the joint posterior density

Theoretical Considerations - I

Impose assumptions on

- Stability of the VAR data generating process
- Bounded eigenvalues of the Gram data matrix and the error covariance matrix
- Control the rate of decay as number of variable $(k_1 + k_2)$ grow with the sample size
- A minimum signal strenth condition on the entries of $\mathbf{W}$

Then, we can establish strong selection consistency; i.e.,

$$\Pi_{\text{pseudo}}(\mathcal{G} = \mathcal{G}_0 \mid \text{data}) \xrightarrow{\mathbb{P}_0} 1 \qquad \text{as} \quad n \to \infty.$$

In words, the posterior distribution of the activity graph concentrates around the

true activity graph

Theoretical Considerations - II

Further, the entries of $\mathbf{A}$ (and thus $\mathbf{W}$) are estimated consistently; i.e.,

$$\mathbb{E}_0 \left[\Pi_{\text{pseudo}} \left(\|\mathbf{A} - \mathbf{A}_0\|_F > C\mathcal{M}\sqrt{\frac{\delta_n \log p}{n}} \middle| \mathcal{Y} \right) \right] \to 0 \text{ as } n \to \infty.$$

where

- δ_n: total number of non-zero entries in the data underlying $\mathbf{A}_{0n}$,
- $p = 3k_1 + k_2$
- $\mathcal{M}$ that depends on the temporal and cross-sectional dependence of the underlying VAR process

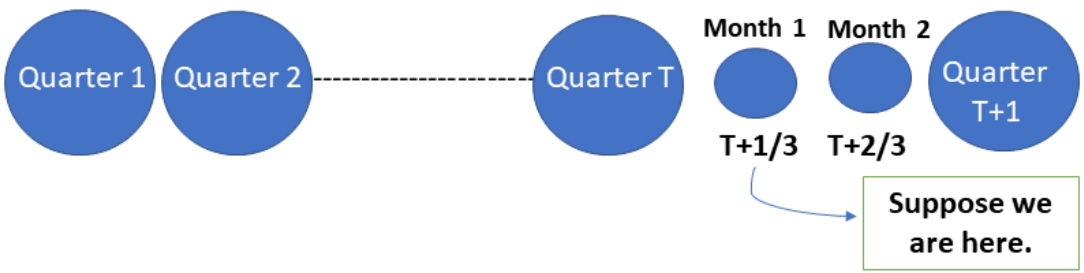
How to obtain Forecasts

- Data until quarter 1 to quarter T can be used to obtain forecasts for monthly variables at time $T + 1/3$, $T + 2/3$ and $T + 1$ and quarterly variables at time $T + 1$ and so on

- These forecasts corresponds to forecast horizon $h = 1, 2, ..$ etc.

How to obtain Nowcasts - I

- Suppose we are at the end of the first month of quarter $T+1$, when fresh monthly data becomes available
- **Q**: How can we incorporate in the model such fresh information to obtain nowcasts (for $h = 2/3$) at time $T+1$?



How to obtain Nowcasts - II

- Replace the oldest month of quarter T with the most recent monthly value at time $T + 1/3$.
  $\Rightarrow$ the most recent 'quarter' covers months $T + 1/3, T, T - 1/3$

- Define new 'y' with this new 'grouping' of monthly variables (standard approach under frequency alignment, although not very satisfactory)

- Similar when new monthly data available until the 2^{nd} month of quarter $T + 1$.

- Then, the most recent 'quarter' covers months $T + 2/3, T + 1/3, T$
  $\Rightarrow$ provides nowcast for $h = 1/3$

Application to Macroeconomic Indicators

- 77 monthly and 9 quarterly key macroeconomic variables for the period January 1960 to June 2019.
- Data collected from FRED database.
- Fit Bayesian MF with lag $d = 1$.
- The estimated connectivity pattern between the time series depicts cross correlations between the monthly and quarterly block.

Competing Models

- Use this data to investigate the nowcasting/forecasting performance of the proposed Bayesian MF VAR model (BMF)
- Modify BMF to also obtain nowcasts for $h = 1/3, 2/3$, etc
- Compare the performance with
  - MFBVAR model
  - MIDAS regression model
  - a quarterly VAR model (aggregate monthly data to quarterly level using equal/skewed weights)
- Compute relative RMSE values for the quarterly forecasts w.r.t a benchmark random walk model, for all the competing models
- Relative RMSE < 1 implies that a method outperforms the benchmark

Forecasting/Nowcasting Performance - I

Table 1: Relative RMSE values for comparing the nowcasting/forecasting performance of competing models

	BMF	MFBVAR(Minn)	MFBVAR(SS)	MFBVAR(Hier. SS)	U-MIDAS	MIDAS(Res.)	QVAR(Equal wt.)	QVAR(Skewed)
$h = 1/3$	**0.66**	1.01	1.22	1.04	1.37	1.21	0.80	0.92
$h = 2/3$	**0.68**	0.94	1.01	0.95	1.12	0.89	0.78	0.74
$h = 1$	**0.82**	1.00	1.07	1.01	1.28	1.10	0.97	1.04
$h = 4/3$	**0.73**	0.81	0.87	0.83	1.24	1.04	0.78	0.78
$h = 5/3$	**0.80**	0.89	0.97	0.91	1.04	0.81	0.78	0.78
$h = 2$	**0.80**	0.89	0.99	0.93	1.15	0.94	0.94	1.08

Forecasting/Nowcasting Performance

Table 2: Relative RMSE values for overall comparison of nowcasting/forecasting performance of competing models, using initial releases of all the variables

	BMF	MFBVAR(Minn)	MFBVAR(SS)	MFBVAR(Hier. SS)	U-MIDAS	MIDAS(Res.)	QVAR(Equal wt.)	QVAR(Skewed)
$h = 1/3$	0.50	1.41	1.42	1.46	1.85	1.54	1.19	1.33
$h = 2/3$	0.53	1.62	1.52	1.47	2.21	1.51	1.10	1.28
$h = 1$	0.55	1.37	1.41	1.49	2.74	2.05	1.91	1.60
$h = 4/3$	0.45	0.85	0.88	0.86	1.86	1.41	0.64	0.59
$h = 5/3$	0.47	1.14	1.12	1.26	2.11	1.43	0.55	0.63
$h = 2$	0.46	0.84	0.99	1.14	2.68	1.80	0.91	0.98

BMF still outperforms competing methods, but all of them exhibit an inferior performance compared to the model trained on the final vintages

Using the output of the BMF model

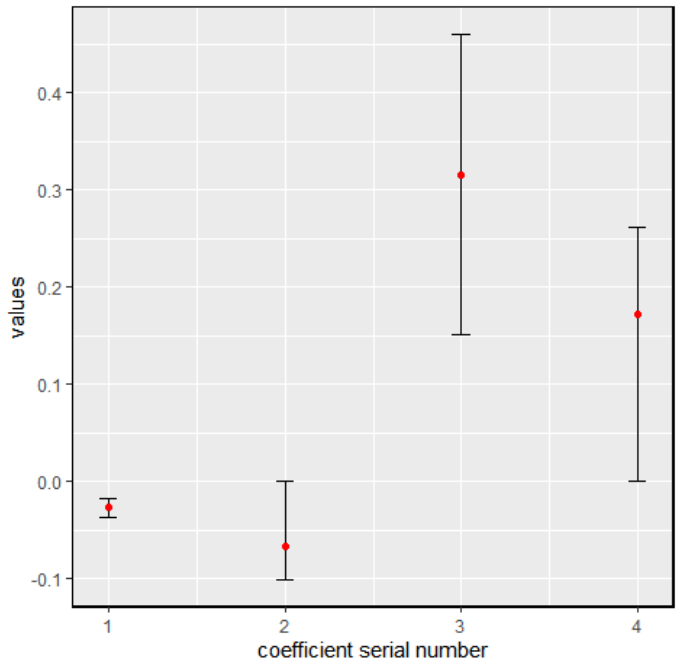


Figure 4: 95% posterior credible intervals using the BMF model for the following four coefficients (i) total reserves of depository institutions on the consumer price index, (ii) total assets for households and nonprofit organizations on securities in bank credit, (iii) real personal consumption expenditure on GDP, and (iv) real government consumption expenditure and gross investment on real hourly compensation for all employed persons

Concluding Remarks

- Mixed frequency data pose a number of interesting modeling challenges
- Bayesian methodology offers uncertainty quantification through inclusion probabilities, credible intervals
- Straightforward computations - competitive/improved performance vis-a-vis MIDAS variants
- In linear regression models, careful lag selection is important to forecasting/nowcasting performance (see, e.g., Ghosh, Khare & Michailidis, 2022)
- Neural network architectures unexplored
  offer very promising performance and overcome issues of model respecification (see Lin & Michailidis, 2023)

References

1. Ghosh, Khare, Michailidis (2022), The Bayesian Nested Lasso for Mixed Frequency Data, forthcoming in the *Annals of Applied Statistics*
2. Chakraborty, Khare, Michailidis (2023), A Bayesian Framework for Sparse Estimation in High-Dimensional Mixed Frequency Vector Autoregressive Models, *Statistica Sinica*, 33, 1629-1652
3. Lin, Michailidis (2022), A Multi-Task Encoder-Decoder Framework for Mixed Frequency Data Prediction, tentatively accepted in *International Journal of Forecasting*