Bayesian Bi-level Sparse Group Regression for Macroeconomic Forecasting

Matteo Mogliani[1] and **Anna Simoni**[2]

[1]Banque de France
[2]CREST-CNRS, Ensae and Ecole Polytechnique

Conference on Real-Time Data Analysis, Methods, and Applications, October 19-20, 2023

The views expressed in this paper are those of the authors and do not necessarily reflect those of the Banque de France or the Eurosystem.

Optimal forecast of y_t, h horizons ahead, based on a set of predictors $\mathbf{x}_t := (\mathbf{x}'_{1,t}, \ldots, \mathbf{x}'_{N,t})'$ to track economic conditions in real-time.

- Large datasets with predictors organized into N groups (each with a possibly infinite number of elements, strong covariation, common characteristics).
- The forecasting / nowcasting model we consider is: $\forall t = 1, \ldots, T$, $\forall h \geq 0$,

$$y_t = \sum_{j=1}^{N} \varphi_j(x_{j,t-h,1}, x_{j,t-h,2}, \ldots) + \varepsilon_t, \qquad \mathbf{E}[\varepsilon_t | \mathbf{x}_{1,t-h-\ell}, \ldots, \mathbf{x}_{N,t-h-\ell}, \ell \geq 0] = 0, \tag{1}$$

  where:
  - $j = 1, \ldots, N$ is the group index,
  - $\varphi_j(\cdot)$ denotes a j-specific unknown function of $\mathbf{x}_{j,t-h}$ taking values in $\mathbb{R}$,
  - $\mathbf{x}_{j,t} := \{x_{j,t,i}\}_{i \geq 1}$ for every $j \in \{1, \ldots, N\}$,
  - If $\mathbf{x}_{1,t}$ contains the lagged values of y_t, then $\mathbf{x}_{1,t} = (y_{t-1}, y_{t-2}, \ldots)'$.
- Unify high-dimensional and nonparametric regression settings.

The Mixed Data Sampling (MIDAS) regression model (*e.g.* Ghysels *et al.* 2006, 2007) can be written as: $\forall t = 1, \ldots, T, \forall h = 0, 1/m, 2/m, \ldots$

$$y_t^L = \sum_{u=1}^{p_y} \beta_u L^u y_t^L + \sum_{j=2}^{N} \Psi(L^{1/m}; \boldsymbol{\theta}_j) x_{j,t-h}^H + \varepsilon_t^L, \tag{2}$$

where $\Psi(L^{1/m}; \boldsymbol{\theta}_j)$ is the high-frequency lag polynomial

$$\Psi(L^{1/m}; \boldsymbol{\theta}_j) = \sum_{u=0}^{p_x} \psi(u; \boldsymbol{\theta}_j) L^{u/m}. \tag{3}$$

- This model can be cast in model (1) with N groups:
  - the first group is $\varphi_1(\mathbf{x}_{1,t}) = \boldsymbol{\theta}_1'(y_{t-1}, \ldots, y_{t-p_y})'$ with $\boldsymbol{\theta}_1 := (\beta_1, \ldots, \beta_{p_y})'$,
  - the remaining $N-1$ groups are given by each high-frequency predictor: $\forall j = 2, \ldots, N$,

$$\varphi_j(x_{j,t-h}^H, \ldots, x_{j,t-h-p_x/m}^H) = \Psi(L^{1/m}; \boldsymbol{\theta}_j) x_{j,t-h}^H.$$

- Possible parameterizations of the weighting function $\psi(u;\boldsymbol{\theta}_j)$:
  - unrestricted MIDAS (Foroni, Marcellino & Schumacher, 2015): $\psi(u;\boldsymbol{\theta}_j) = \theta_{u,j}$;
  - Almon lag polynomials (power polynomials): $\psi(u;\boldsymbol{\theta}_j) = \sum_{i=0}^{C} \theta_{j,i} u^i$;
  - more generally, by using orthogonal basis functions $\{\phi_i(u)\}_i$ on $\mathbb{R}_+$ we obtain:

$$\psi(u;\boldsymbol{\theta}_j) = \sum_{i=1}^{\infty} \theta_{j,i}\phi_i(u). \tag{4}$$

Hence,

$$\begin{aligned}\varphi_j(\mathbf{x}_{j,t-h}) &= \Psi(L^{1/m};\boldsymbol{\theta}_j)x_{j,t-h}^H \\ &= \sum_{u=0}^{p_x}\sum_{i=1}^{\infty}\theta_{j,i}\phi_i(u)x_{j,t-h-u/m}^H = \sum_{i=1}^{\infty}\theta_{j,i}\Phi_i'\mathbf{x}_{j,t-h},\end{aligned}$$

where $\Phi_i := (\phi_i(0), \phi_i(1), \ldots, \phi_i(p_x))'$.

- Related literature: Babii et al. (2022, JBES, Grouped Lasso estimator), Mogliani & Simoni (2021, JoE – without sparsity within groups).

Additional motivating examples encompassed in model (1).

1). Nonlinear predictive model for $y_t : \forall t = 1, \ldots, T$,

$$y_t = \sum_{j=1}^{N} \varphi_j(x_{j,t-h}) + \varepsilon_t, \qquad \mathbf{E}[\varepsilon_t | x_{j,t-h-\ell},\, j = 1, \ldots, N,\, \ell \geq 0] = 0,$$

where $\varphi_j(\cdot)$ is an unknown function of one covariate. For a set of approximating functions $\{\phi_{j1}, \phi_{j2}, \ldots\}$,

$$\varphi_j(x_{j,t-h}) = \sum_{i=1}^{\infty} \phi_{ji}(x_{j,t-h})\theta_{j,i}, \qquad j \in \{1, \ldots, N\}.$$

2). Data-poor environment with many lags per each predictor and for y_t.

3). Grouped predictors in a data-rich environment where each group of covariates $\mathbf{x}_{j,t}$ contains $\leq g$ elements. By assuming a linear model: $\forall t = 1, \ldots, T$,

$$y_t = \sum_{j=1}^{N} \mathbf{x}'_{j,t-h}\boldsymbol{\theta}_j + \varepsilon_t, \qquad \mathbf{E}[\varepsilon_t | \mathbf{x}_{1,t-h-\ell}, \ldots, \mathbf{x}_{N,t-h-\ell}, \ell \geq 0] = 0 \qquad (5)$$

and $\varphi_j(\mathbf{x}_{j,t-h}) = \mathbf{x}'_{j,t-h}\boldsymbol{\theta}_j$.

Contributions.

1) Propose a Bayesian approach to deal with / exploit the sparse group structure:

- we construct a hierarchical prior that:
  - induces a bi-level sparsity: some groups and some predictors inside a group can be irrelevant for forecasting the target variable, conditional on the remaining predictors;
  - treats the coefficients of each block independently but, after marginalization, imposes a correlation among the coefficients in each block
- appealing because:
  - it allows assessment of the uncertainty;
  - it has a build-in prediction with optimal properties;
  - easy to introduce stochastic volatility (to robustify the forecasting accuracy in volatile periods exhibiting large fluctuations)

2) Establish frequentist asymptotic properties.

3) Gibbs sampler with a one step of Metropolis-Hasting.

4) Monte Carlo exercise to study finite sample properties.

5) Empirical application: nowcast of US GDP with grouped predictors.

Outline:

By assuming $\varepsilon_t \sim^{i.i.d.} \mathcal{N}(0, \sigma^2)$ then the sampling model is: $\forall h \geq 0$

$$y_t | \mathbf{x}_{t-h}, \varphi, \sigma^2 \sim \mathcal{N}\left(\sum_{j=1}^{N} \varphi_j(\mathbf{x}_{j,t-h}), \sigma^2 \right),$$

where:

- $\mathbf{x}_t := (\mathbf{x}'_{1,t}, \ldots, \mathbf{x}'_{N,t})'$ is a vector of potentially infinite dimension,
- $\mathbf{X} := (\mathbf{x}_{-h+1}, \ldots, \mathbf{x}_{T-h})'$ is a matrix with T rows,
- $\varphi := (\varphi_1, \ldots, \varphi_N)'$, $\varphi \in \mathcal{H}$, Hilbert space.

To reduce the dimension of the model, we assume there exist

- a vector $\mathbf{z}_{j,t-h} := (z_{j,t-h,1}, \ldots, z_{j,t-h,g})'$ of transformations of $\mathbf{x}_{j,t-h}$
- and parameters $\{\theta_{j,i}\}_{i=1}^{g}$

such that for every $j \in \{1, \ldots, N\}$, the function $\varphi_j(\mathbf{x}_{j,t-h})$ is well approximated by

$$\varphi_j(\mathbf{x}_{j,t-h}) \approx \sum_{i=1}^{g} z_{j,t-h,i}\theta_{j,i} = \mathbf{z}'_{t-h}\boldsymbol{\theta},$$

where $g \geq 1$ is a truncation parameter.

We introduce:

- for every $j = 1, \ldots, N$, define $\boldsymbol{\theta}_j := (\theta_{j,1}, \ldots, \theta_{j,g})' \in \mathbb{R}^g$,
- $\boldsymbol{\theta} := (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_N)' \in \Theta \subset \mathbb{R}^{Ng}$,
- $\mathbf{z}_t := (\mathbf{z}'_{1,t}, \ldots, \mathbf{z}'_{N,t})'$ is a $(gN \times 1)$ vector,
- and $\mathbf{Z} := (\mathbf{z}_{1-h}, \ldots, \mathbf{z}_{T-h})'$ is a $(T \times gN)$ matrix.

Examples:

1. MIDAS:

$$\varphi_j(\mathbf{x}_{j,t-h}) = \Psi(L^{1/m};\boldsymbol{\theta}_j)x^H_{j,t-h} = \sum_{u=0}^{p_x}\sum_{i=1}^{\infty}\theta_{j,i}\phi_i(u)x^H_{j,t-h-u/m}$$

$$\approx \sum_{i=1}^{g}\theta_{j,i}\Phi_i'\mathbf{x}_{j,t-h} = \sum_{i=1}^{g}\theta_{j,i}z_{j,t-h,i},$$

where $\Phi_i := (\phi_i(0), \phi_i(1), \ldots, \phi_i(p_x))'$, $z_{j,t-h,i} := \Phi_i'\mathbf{x}_{j,t-h}$, and $\mathbf{z}_{j,t-h} := (\mathbf{x}'_{j,t-h}\Phi_1, \ldots, \mathbf{x}'_{j,t-h}\Phi_g)'$.

2. Grouped predictors: $\mathbf{z}_{j,t} = \mathbf{x}_{j,t}$, no approximation.

3. Nonlinear predictive models: we approximate $\varphi_j(x_{j,t-h})$ as

$$\varphi_j(x_{j,t-h}) \approx \sum_{i=1}^{g}\phi_{ji}(x_{j,t-h})\theta_{j,i} =: \mathbf{z}'_{j,t-h}\boldsymbol{\theta}_j.$$

Let (φ_0, σ_0^2) be the true value of (φ, σ^2) that generates the data.

$$y_t|\mathbf{x}_{t-h}, \varphi_0, \sigma_0^2 \sim \mathcal{N}\left(\sum_{j=1}^{N} \varphi_{0,j}(\mathbf{x}_{j,t-h}), \sigma_0^2 \mathbf{I}_T\right).$$

The approximation bias in the mean is

$$\begin{aligned} B_{0,t}(g) &:= \mathbf{E}[y_t|\{\mathbf{x}_{t-h}\}_{t=1,\ldots,T}] - \mathbf{z}'_{t-h}\boldsymbol{\theta}_0 \\ &= \sum_{j=1}^{N} \left(\varphi_{0,j}(\mathbf{x}_{j,t-h}) - \mathbf{z}'_{j,t-h}\boldsymbol{\theta}_{0,j}\right), \qquad \forall t = 1, \ldots, T \end{aligned}$$

and $B_0(g) := (B_{0,1}(g), \ldots, B_{0,T}(g))'$ is a T-vector.

- In this paper we adopt a Bayesian approach and specify a convenient prior that is degenerate at zero for the quantity $B_t(g)$.

Sparsity structure:

We assume bi-level sparsity:

Bi-level sparsity is the feature of the model that guarantees $\exists$ *an approximation*

$$\mathbf{z}'_{t-h}\boldsymbol{\theta}_0 \equiv \sum_{j=1}^{N} \mathbf{z}'_{j,t-h}\boldsymbol{\theta}_{0,j}$$

to $\sum_{j=1}^{N} \varphi_{0,j}(\mathbf{x}_{j,t-h})$ *in* (1) *with a small number of active groups and of non-zero coefficients for each active group such that the approximation bias* $B_{0,t}(g)$ *is small relative to the estimation error.*

We specify a prior that induces exact sparsity both at the group level and within groups.

- This prior puts all its mass on the approximation $\mathbf{z}_{t-h}'\boldsymbol{\theta}$ conditional on $\mathbf{z}_{t-h}$.
- For every group $j = 1, \ldots, N$, $\boldsymbol{\theta}_j = V_j^{1/2}\mathbf{b}_j$, $\mathbf{b}_j := (b_{j,1}, \ldots, b_{j,g})'$, $V_j^{1/2} := \text{diag}(v_{j1}, \ldots, v_{jg})$ and $v_{ji} \geq 0$ for $i = 1, \ldots, g$.
- We treat the truncation parameter g as deterministic and, under Assumption 6.1, it might depend on s_0
- The double spike-and-slab prior is inspired from Xu & Ghosh (2015).

Sparse Group Selection with spike-and-slab prior (cont.)

Prior distributions inducing **bi-level sparsity** (hard spike-and-slab): $\forall j = 1, \ldots, N$

$$B_{t,j}(g)|\mathbf{x}_{t-h}, g \sim \delta_0, \qquad \forall t = 1, \ldots, T, \tag{6}$$

$$\mathbf{b}_j|g, \pi_0 \overset{ind.}{\sim} (1-\pi_0)\mathcal{N}_g(0, I_g) + \pi_0\delta_0(\mathbf{b}_j), \tag{7}$$

$$v_{ji}|\pi_1, \tau_j \overset{ind.}{\sim} (1-\pi_1)\mathcal{N}^+(0, \tau_j^2) + \pi_1\delta_0(v_{ji}), \qquad i = 1, \ldots, g, \tag{8}$$

where $\mathcal{N}^+(0, \tau_j^2)$ denotes a truncated $\mathcal{N}(0, \tau_j^2)$ distribution truncated below at 0.

Prior on the hyperparameters and model variance:

$$\tau_j \overset{ind.}{\sim} \Gamma\left(\frac{1}{2}, \lambda_{1,j}\right),$$
$$\pi_0 \sim \mathcal{B}eta(c_0, d_0),$$
$$\pi_1 \sim \mathcal{B}eta(c_1, d_1),$$
$$\sigma^2 \sim \mathcal{I}\Gamma(a, b).$$

Outline:

Outline:

Outline:

We consider the MIDAS model:

$$y_t^L = 0.5 + 0.3 y_{t-1}^L + \sum_{j=1}^{N} \sum_{i=0}^{p_x=11} \psi\left(i; \widetilde{\boldsymbol{\theta}}\right) x_{j,t-i/3}^H + \varepsilon_t^L$$

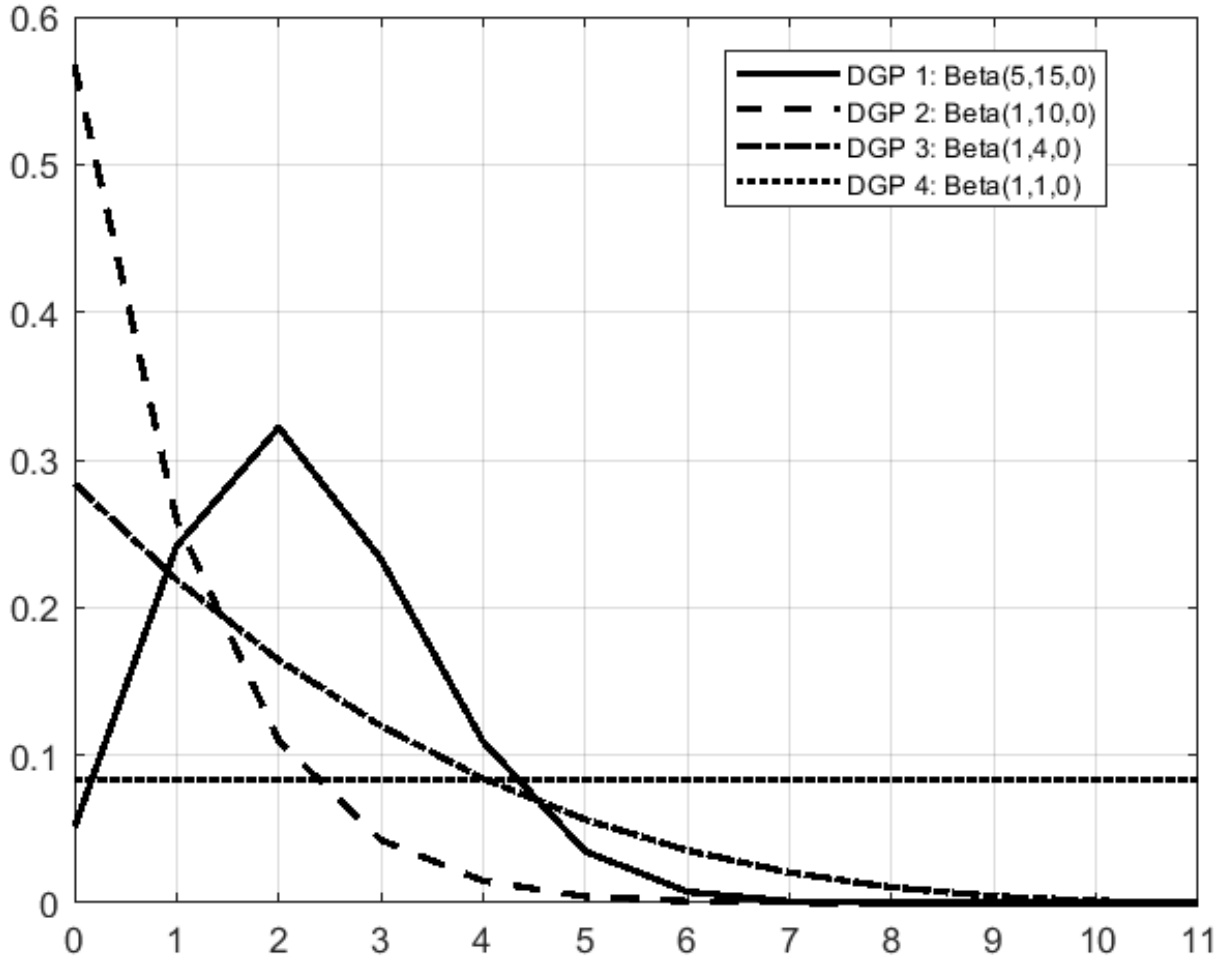
$$x_{j,t}^H = 0.9 x_{j,t-1/3}^H + \epsilon_{j,t}^H$$

$$\psi\left(i; \boldsymbol{\theta}_j\right) = \left(\frac{i+1}{p_x+1}\right)^{\theta_1 - 1} \left(1 - \frac{i+1}{p_x+1}\right)^{\theta_2 - 1} \frac{\Gamma(\theta_1+\theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} + \theta_3$$

$$\begin{pmatrix} \varepsilon_t^L \\ \boldsymbol{\epsilon}_t^H \end{pmatrix} \sim \text{i.i.d. } \mathcal{N}\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\epsilon \end{pmatrix}\right]$$

- $T = 200$
- $N = \{50, 100\}$
- $s_0^{gr} = \{5, 10\}$
- $\boldsymbol{\Sigma}_\epsilon = \mathcal{S}_\epsilon \mathcal{R}_\epsilon \mathcal{S}_\epsilon$, with $\mathcal{S}_\epsilon$ diagonal matrix with elements σ_ϵ and $\mathcal{R}_\epsilon$ a Toeplitz correlation matrix with off-diagonal elements $\rho_\epsilon^{|j-j'|}$ for all $j \neq j'$
- $\sigma = 0.5$ and σ_ϵ fixed such that NSR $= 0.2$.
- $\rho_\epsilon = 0.5$

Weighting function $\psi\left(i;\widetilde{\boldsymbol{\theta}}\right)$



We estimate the model by approximating the true weighting function through a set of polynomials:

- $\psi(i;\boldsymbol{\theta}) = \theta_{i,j} \Rightarrow$ linear lag polynomials (Unrestricted MIDAS)
- $\psi(i;\boldsymbol{\theta}) = \sum_{p=1}^{g} \theta_{p,j} i^p \Rightarrow$ algebraic power lag polynomials (Almon, w and w/o end-point restrictions; Mogliani & Simoni, 2021)
- $\psi(i;\boldsymbol{\theta}) = \sum_{p=1}^{g} \theta_{p,j} \phi_p(i) \Rightarrow \phi_p(i)$ orthogonal lag polynomials:
  - Legendre
  - Bernstein
  - Chebyshev first-kind (T)

We set $g = 5$ (=3 for restricted Almon). Orthogonal polynomials are normalized and shifted over the interval [0,1].

We iterate the Gibbs sampler for 50000 sweeps (+10000 burn-in) and we perform 100 MC simulations.

Monte Carlo simulations: selection and predictive accuracy

			DGP 1 fast-decaying		DGP 2 bell-shaped		DGP 3 slow-decaying		DGP 4 flat	
N	s_0^{gr}	Polynomial	TPR	CRPS	TPR	CRPS	TPR	CRPS	TPR	CRPS
50	5	Unrestricted	99.8	0.71	99.8	0.67	98.8	0.74	45.8	0.93
		Almon	65.4	0.82	38.3	0.92	48.3	0.90	85.4	0.81
		Restr. Almon	98.0	0.70	83.6	0.71	96.2	0.71	93.3	0.81
		Legendre	60.3	0.86	92.0	0.72	60.9	0.86	81.1	0.81
		Bernstein	98.2	0.73	99.7	0.66	98.4	0.73	57.4	0.89
		Chebychev U	60.3	0.86	95.3	0.71	62.9	0.86	81.9	0.81
100	5	Unrestricted	99.6	0.70	99.7	0.68	95.9	0.75	30.5	0.96
		Almon	44.4	0.89	23.7	0.97	31.3	0.95	65.2	0.85
		Restr. Almon	95.9	0.69	83.4	0.71	94.3	0.71	80.7	0.83
		Legendre	40.7	0.92	71.9	0.78	42.2	0.91	57.1	0.88
		Bernstein	91.6	0.74	98.8	0.66	96.4	0.73	37.2	0.94
		Chebychev U	41.7	0.91	80.1	0.76	43.1	0.91	58.8	0.87
100	10	Unrestricted	12.6	0.99	15.1	0.98	11.7	0.99	10.2	1.00
		Almon	9.5	1.00	9.2	1.00	9.2	1.00	9.5	1.00
		Restr. Almon	51.7	0.86	36.5	0.91	39.3	0.91	12.7	0.99
		Legendre	10.8	0.99	11.2	0.99	10.6	0.99	9.5	1.00
		Bernstein	11.8	0.99	17.3	0.98	12.5	0.99	10.2	1.00
		Chebychev U	10.8	0.99	11.6	0.99	10.7	0.99	9.5	1.00

Table: TPR and CRPS denote respectively the true positive rate and the continuously ranked probability score, the latter in relative terms with respect to the AR(1) benchmark.

▸ Extension: Stochastic Volatility

Outline:

Empirical application: nowcasting US GDP in a mixed-frequency framework.

Nowcasting exercise of US GDP in the following mixed-frequency framework:

$$y_t = \alpha + \beta y_{t-1} + \sum_{j=2}^{N} \sum_{u=0}^{p_x} \psi(u; \boldsymbol{\theta}_j) x_{j,t-h-u/m} + \varepsilon_t,$$

where

- $y_t = 400 \log(Y_t / Y_{t-1}) =$ annualized quarterly growth rate of GDP,
- $\mathbf{x}_t =$ vector of $N = 122$ macroeconomic series sampled at monthly frequency and extracted from the FRED-MD database (McCracken & Ng, 2016).
- The data sample starts in 1980Q1, while the pseudo out-of-sample analysis spans 2013Q1 to 2022Q4.
- Rolling window of $T = 132$ quarterly observations, and h-step-ahead posterior predictive densities for $y_\tau | x_{\tau-h}, \tau > T$ are generated from:

$$f(y_\tau | x_{\tau-h}, y, \mathbf{X}) = \int f_0(y_\tau | \varphi, \sigma^2, x_{\tau-h}) \Pi(\varphi, \sigma^2 | y, \mathbf{X}) d\varphi d\sigma^2. \qquad (9)$$

Empirical application: nowcasting US GDP in a mixed-frequency framework. (cont.)

- 3 nowcasting horizons: $h = 0, 1/3, 2/3$ and two lag polynomials: restricted Almon and the orthonormal Bernstein polynomials.
- We allow for time-varying volatility with heavy tails and occasional outliers in the regression errors (to account for the Great Moderation, the Great Recession, and the Covid crisis).

We consider two modelling strategies to exploit our bi-level sparsity prior approach:

1. First, we estimate the forecasting model on the whole set of 122 indicators. The total number of parameters is either 244 (restricted Almon) or 732 (Bernstein).
2. Alternative strategy: estimating the model on separate groups of indicators, where the groups are set according to partition of indicators defined in McCracken & Ng (2016).
   - We have a total number of 8 groups (output and income; labour market; housing; consumption, orders, and inventories; money and credit; interest and exchange rates; prices; stock market), each one including between 5 and 31 indicators.

Empirical application: nowcasting US GDP in a mixed-frequency framework. (cont.)

Summing up, we estimate a large set of alternative specifications, according to:

- the 2 lag polynomials (Almon and Bernstein),
- the 5 volatility process (homoskedastic, SV, SV with Student-t shocks, SV with outliers, SV with Student-t shocks and outliers),
- and the 2 partition strategies (whole dataset *vs* 8 groups).

To process this large amounts of results, we combine the set of obtained individual density forecasts.

Empirical application: nowcasting US GDP - RESULTS.

	BSGS-SS			BSGL		
	h=0	h=1/3	h=2/3	h=0	h=1/3	h=2/3
Panel A. RMSFE						
Groups - Almon	0.77	0.75	0.74	0.82	0.72	0.75
Groups - Bernstein	0.70	0.62	0.81	0.84	0.74	0.74
Groups - all	0.67	0.71	0.77	0.91	0.72	0.74
Whole dataset - Almon	0.93	0.82	0.71	0.97	0.75	0.83
Whole dataset - Bernstein	0.69	0.96	0.87	3.00	1.00	0.87
Whole dataset - all	0.69	0.96	0.75	1.20	0.88	0.88
Panel B. LogS						
Groups - Almon	10.04	6.16	8.56	9.28	5.72	6.67
Groups - Bernstein	9.83	12.92	3.63	7.07	2.12	6.65
Groups - all	10.05	10.98	6.25	9.02	10.56	7.56
Whole dataset - Almon	9.56	5.18	8.05	4.22	9.81	6.55
Whole dataset - Bernstein	11.33	6.34	5.95	5.96	-15.46	6.20
Whole dataset - all	12.84	4.03	6.77	7.34	-12.46	6.27
Panel C. CRPS						
Groups - Almon	0.79	0.77	0.75	0.81	0.72	0.78
Groups - Bernstein	0.75	0.67	0.85	0.85	0.77	0.77
Groups - all	0.74	0.71	0.79	0.85	0.72	0.76
Whole dataset - Almon	0.92	0.82	0.74	4.23	3.75	3.86
Whole dataset - Bernstein	0.73	0.92	0.86	6.29	4.34	4.45
Whole dataset - all	0.72	0.93	0.76	0.98	0.80	0.86

Table: BSGS-SS denotes the proposed bi-level sparsity prior. BSGL denotes the Bayesian Sparse Group Lasso prior (Xu & Ghosh, 2015). RMSFE, LogS, and CRPS denote respectively the root mean squared forecast error, the log-score, and the continuously ranked probability score, in relative terms with respect to the AR(1) benchmark.

Outline:

- $(\varphi_0, \sigma_0^2, \boldsymbol{\theta}_0)$ denotes the true value of $(\varphi, \sigma^2, \boldsymbol{\theta})$ that generates the data.
- $\mathbf{E}_0[\cdot]$ denotes the expectation taken with respect to the true data distribution conditional on $(\mathbf{X}, \varphi_0, \sigma_0^2)$.
- Our asymptotic analysis is for $T \to \infty$. We allow N, s_0^{gr}, s_0 and $g \to \infty$ with T.

$\boldsymbol{\theta}_0$ is (s_0, s_0^{gr})-sparse, where

- $s_0^{gr} := |S_0^{gr}| \ll N$, $S_0^{gr} := \{j \in \{1, \ldots, N\}; \|\boldsymbol{\theta}_{0,j}\|_2 > 0\}$ is the group support and s_0^{gr} is the number of active groups.
- If $S_0^{gr} \neq \varnothing$, for every $j \in S_0^{gr}$ let $S_{0,j}$ be the set of the indices of the nonzero elements in $\boldsymbol{\theta}_{0,j}$.
- So, $S_0 := \bigcup_{j \in S_0^{gr}} S_{0,j}$ is the support of $\boldsymbol{\theta}$.
- Number of active coefficients: $s_0 := \sum_{j \in S_0^{gr}} |S_{0,j}| \ll Ng$ and $|S_{0,j}| \ll g$.

Rate of contraction of the posterior distribution:

$$\epsilon := \max\left\{\sqrt{\frac{s_0^{gr}\log(N)}{T}}, \sqrt{\frac{s_0\log(T)}{T}}, \sqrt{\frac{s_0\log(s_0^{gr}g)}{T}}\right\}$$

Define $\|\mathbf{Z}\|_o := \max\{\|\mathbf{Z}_j\|_{op}; 1 \leq j \leq N\}$, where $\mathbf{Z}_j$ is the $(T \times g)$-submatrix of $\mathbf{Z}$ made of all the rows and the columns corresponding to the indices in the j-th group.

Theorem 1

Suppose Assumptions 6.1, 6.2, 6.3 and 6.4 hold. Let $\epsilon \to 0$. *Then, for a sufficiently large* $M > 0$*:*

$$\sup_{(\varphi_0,\sigma_0^2)\in\mathcal{F}_0(s_0,s_0^{gr};\mathbf{Z})} \mathbf{E}_0\left[\Pi\left(\varphi; \left\|\sum_{j=1}^{N}\left(\varphi_j^{(T)}(\mathbf{X}) - \varphi_{0,j}^{(T)}(\mathbf{X})\right)\right\|_2^2 \leq MT\epsilon^2 \middle| y, \mathbf{X}\right)\right] \to 0. \tag{10}$$

Remarks:

- In the grouped predictors example:

$$\left\|\sum_{j=1}^{N}\left(\varphi_j^{(T)}(\mathbf{X}) - \varphi_{0,j}^{(T)}(\mathbf{X})\right)\right\|_2^2 = \|\mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_2^2.$$

- Similarly, in the MIDAS example:

$$\left\| \sum_{j=1}^{N} \left(\varphi_j^{(T)}(\mathbf{X}) - \varphi_{0,j}^{(T)}(\mathbf{X}) \right) \right\|_2^2 = \|\mathbf{Z}^{\infty}(\boldsymbol{\theta}^{\infty} - \boldsymbol{\theta}_0^{\infty})\|_2^2$$

with $\boldsymbol{\theta}^{\infty} = \{\theta_{j1}, \theta_{j2}, \ldots\}_{j=1}^{N}$ an infinite dimensional vector, $\mathbf{z}_t^{\infty}$ is defined similarly and $\mathbf{Z}^{\infty} = (\mathbf{z}_{1-h}^{\infty}, \ldots, \mathbf{z}_{T-h}^{\infty})'$ is a matrix with T rows and an infinite number of columns.

We now look at parameter recovery of our procedure (*i.e.* consistency of the marginal posterior of $\boldsymbol{\theta}$ – coefficients of the approximation of φ).

Definition 1 (Smallest scaled sparse singular value.)

For every $s, r > 0$, the smallest scaled sparse singular value of dimension (s, r) is defined as

$$\widetilde{\phi}(s,r) := \inf \left\{ \frac{\|\mathbf{Z}\boldsymbol{\theta}\|_2^2}{\|\mathbf{Z}\|_o^2 \|\boldsymbol{\theta}\|_2^2},\ 0 \leq s_{\boldsymbol{\theta}}^{gr} \leq s \ \textit{and}\ 0 \leq s_{\boldsymbol{\theta}} \leq r \right\}. \tag{11}$$

- The double sparse eigenvalue condition requires that for every $s, r > 0$, $\exists$ a constant $\kappa > 0$ such that $\widetilde{\phi}(s,r) > \kappa$. Under this assumption:

$$\|\mathbf{Z}\boldsymbol{\theta}\|_2^2 \geq \kappa \|\mathbf{Z}\|_o^2 \|\boldsymbol{\theta}\|_2^2.$$

- We use the notation $\widetilde{\phi}_0 := \widetilde{\phi}(M_0 \widetilde{s}_0^{gr} + s_0^{gr}, M_1 \widetilde{s}_0 + s_0)$ for two positive constants M_0 and M_1.

Theorem 2

Suppose Assumptions 6.1, 6.2, 6.3 and 6.4 hold. Let $\epsilon \to 0$. *Then, for every constant* $M_3 \geq 2M + \overline{\sigma}^2/8$ *where* M *is as in Theorem 3 we have:*

$$\sup_{(\varphi_0,\sigma_0^2)\in\mathcal{F}_0(s_0,s_0^{gr};\mathbf{Z})} \mathbf{E}_0\left[\Pi\left(\boldsymbol{\theta}\in\Theta;\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2 \geq \frac{M_3 T\epsilon^2}{\widetilde{\phi}_0\|\mathbf{Z}\|_o^2}\middle| y,\mathbf{X}\right)\right]\to 0. \tag{12}$$

If there exists two constants $\kappa_\ell, \kappa_z > 0$ *such that* $\widetilde{\phi}(s,r) > \kappa_\ell$ *and* $\|\mathbf{Z}\|_o \leq \sqrt{\kappa_z}\sqrt{T}$ *w.p.a.* 1, *then*

$$\sup_{(\varphi_0,\sigma_0^2)\in\mathcal{F}_0(s_0,s_0^{gr};\mathbf{Z})} \mathbf{E}_0\left[\Pi\left(\boldsymbol{\theta}\in\Theta;\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2 \geq \frac{M_3\epsilon^2}{\kappa_\ell\kappa_z}\middle| y,\mathbf{X}\right)\right]\to 0. \tag{13}$$

Conclusions

- Optimal forecast of y_t, h horizons ahead, based on a set of grouped-predictors to track economic conditions in real-time.
- We propose a Bayesian approach (assessment of the uncertainty, introduce stochastic volatility).
- We exploit the group structure and the sparsity, and construct a prior that induces bi-level sparsity.
- Demonstrate good asymptotic properties for this prior.
- Good performance to nowcast US GDP growth.

Bayesian Bi-level Sparse Group Regression for Macroeconomic Forecasting

Matteo Mogliani[1] and **Anna Simoni**[2]

[1]Banque de France
[2]CREST-CNRS, Ensae and Ecole Polytechnique

Conference on Real-Time Data Analysis, Methods, and Applications, October 19-20, 2023

The views expressed in this paper are those of the authors and do not necessarily reflect those of the Banque de France or the Eurosystem.

References

Azzalini, A. & Capitanio, A. (2014). *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge (UK): Cambridge University Press.

Babii, A., Ghysels, E., & Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, *40*(3), 1094–1106.

Cai, T. T., Zhang, A. R., & Zhou, Y. (2022). Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *IEEE Transactions on Information Theory*, *68*(9), 5975–6002.

Foroni, C., Marcellino, M., & Schumacher, C. (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 57–82.

Kastner, G. & Fruhwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics and Data Analysis*, *76*, 408–423.

Li, Z., Zhang, Y., & Yin, J. (2022). Minimax rates for high-dimensional double sparse structure over ℓ_q-balls.

McCracken, M. W. & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

Mogliani, M. & Simoni, A. (2021). Bayesian midas penalized regressions: Estimation, selection, and prediction. *Journal of Econometrics*, *222*(1, Part C), 833–860.

Omori, Y., Chib, S., Shephard, N., & Nakajima, J. (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics*, *140*(2), 425–449.

Xu, X. & Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, *10*(4), 909–936.

Modify the model as follows:

$$
\begin{aligned}
y_t &= \sum_{j=1}^{N} \varphi_j(x_{j,t-h,1}, x_{j,t-h,2}, \ldots) + e^{\sigma_t/2}\varepsilon_t, \quad \mathbf{E}[\varepsilon_t | \mathbf{x}_{1,t-h-\ell}, \ldots, \mathbf{x}_{N,t-h-\ell}, \ \ell \geq 0] = 0, \\
\sigma_t &= \mu_1 + \mu_2(\sigma_{t-1} - \mu_1) + u_t, \qquad u_t \sim \mathcal{N}(0, \xi^2)
\end{aligned}
$$

with

- μ_2 being between -1 and 1 (for stationarity);
- $\sigma_0 \sim \mathcal{N}(\mu_1, \xi^2/(1-\mu_2^2))$;
- $\varepsilon_t \sim \mathcal{N}(0,1)$ or $\varepsilon_t|\tau_t \sim \mathcal{N}(0,\tau_t)$ with $\tau_t \sim$ Inv-Gamma $\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ (which gives $\exp\{\sigma_t/2\}\varepsilon_t|\sigma_t \sim t_\nu(0, \exp\{\sigma_t\})$).

Solve the intractability of the SV's likelihood function by treating the latent volatilities as unknown parameters (augmentation).

So, replace the inverse-gamma prior for σ^2 with the above AR(1) model and

$$\begin{aligned} \mu_1 &\sim \mathcal{N}(\underline{\mu_1}, \underline{V_1}) \\ (\mu_2 + 1)/2 &\sim \mathcal{B}eta(a_2, b_2) \\ \xi^2 &\sim \text{Gamma}(0.5, 0.5/V_\xi) \\ \nu &\sim \mathcal{U}(0, \underline{\nu}). \end{aligned}$$

We consider the same DGP as above, but we now include SV:

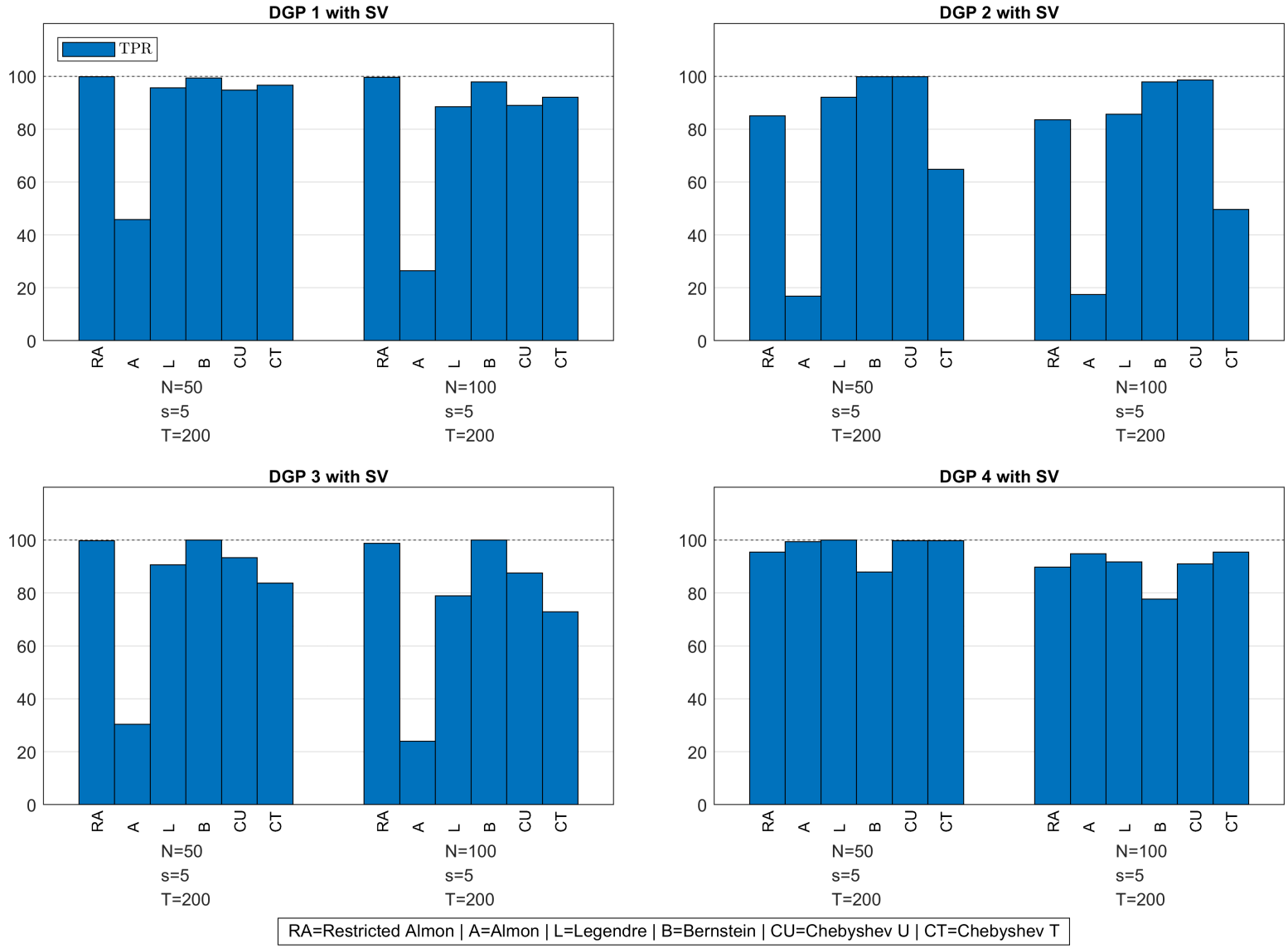
$$y_t^L = 0.5 + 0.3 y_{t-1}^L + \sum_{j=1}^{N} \sum_{i=0}^{p_x=11} \psi\left(i; \widetilde{\boldsymbol{\theta}}\right) x_{j,t-i/3}^H + e^{\sigma_t/2} \varepsilon_t^L$$

$$\sigma_t = \mu_1 + \mu_2(\sigma_{t-1} - \mu_1) + u_t, \qquad u_t \sim \mathcal{N}(0, \xi^2)$$

- $\mu_1 = 2\log(0.5)$
- $\mu_2 = 0.90$
- $\xi = \sqrt{0.05}$

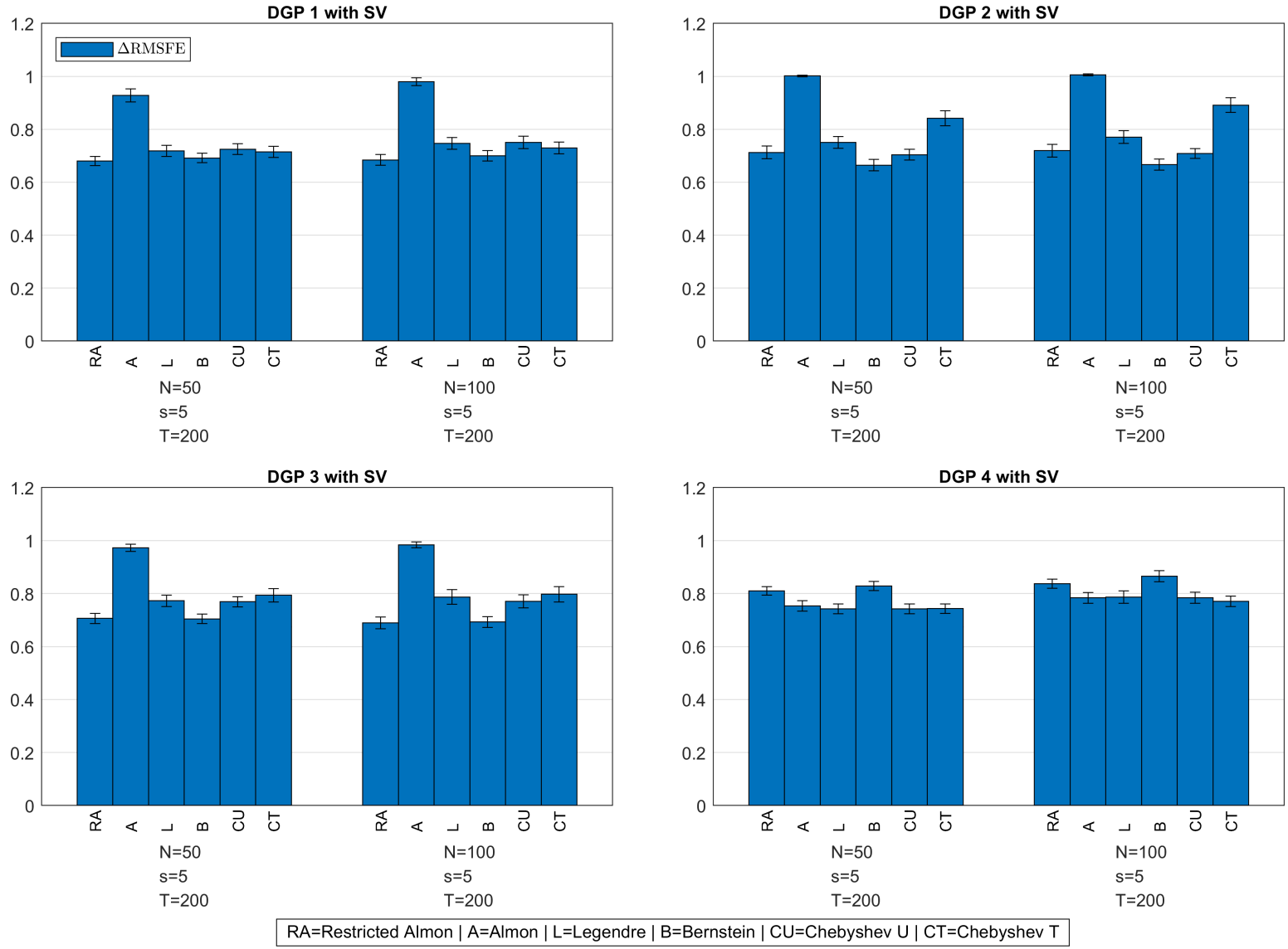
We employ standard samplers for SV (Omori et al., 2007), but we consider an interweaving-strategy between centered and non-centered parameterization (Kastner & Fruhwirth-Schnatter, 2014).

Simulation results (SV): True Positive Rate



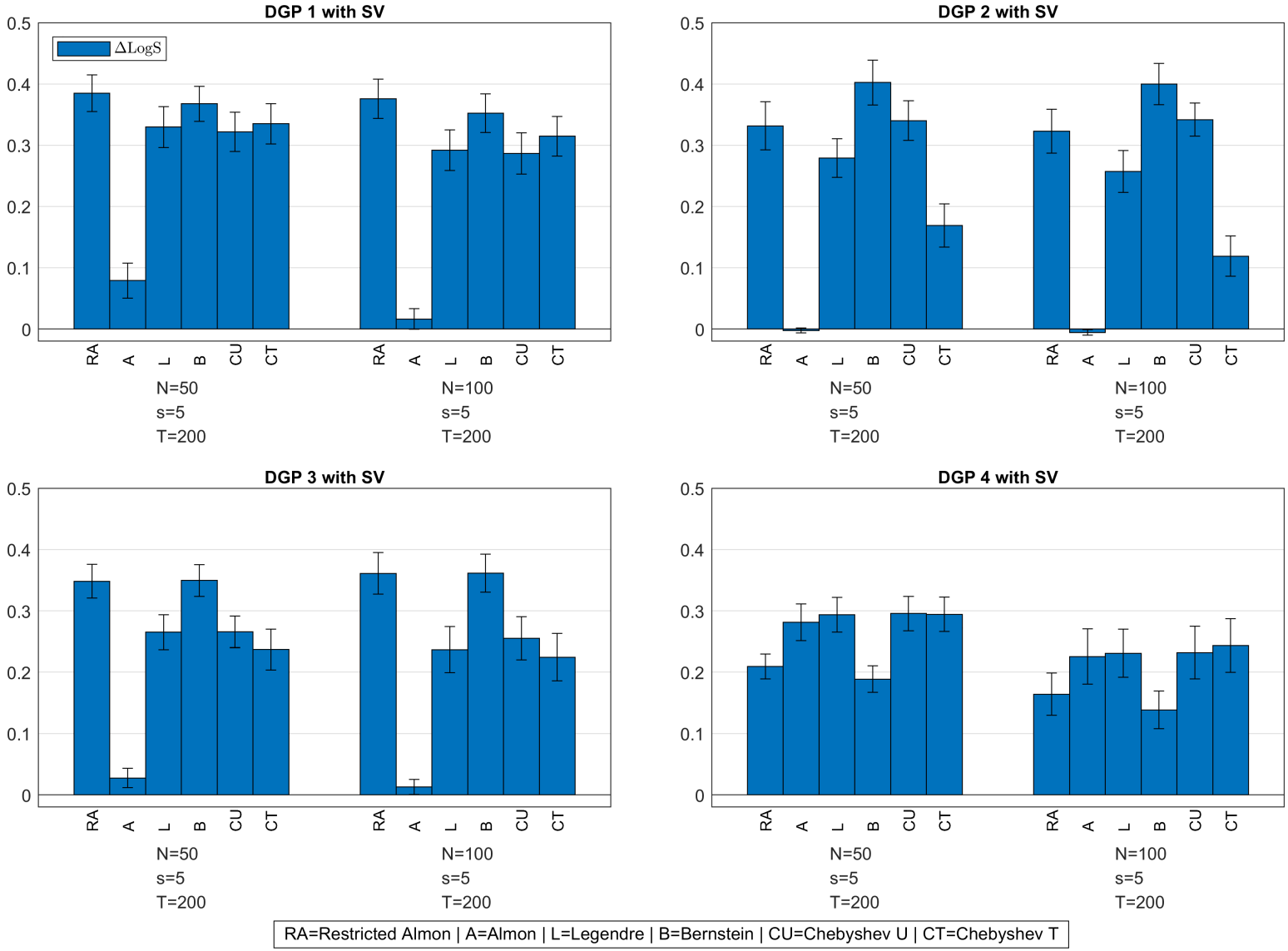
Notes: True Positive Rate = True Positive/(True Positive+False Negative).

Simulation results (SV): relative RMSFE



Notes: relative RMSFE w.r.t. AR(1)-SV, computed over 50 out-of-sample observations. Error bars denote $\pm$2SE computed through bootstrap.

Simulation results (SV): relative Log Score



Notes: relative Log Score w.r.t. AR(1)-SV, computed over 50 out-of-sample observations. Error bars denote $\pm$2SE computed through bootstrap.

The next assumption restricts the size of the approximation bias $B_{0,t}(g)$ and guarantees that it is small relative to the estimation error (similar to Belloni et al. 2014).

Assumption 6.1

Let s_0^{gr}, s_0 be positive integers satisfying $s_0^{gr} \leq N$ and $s_0^{gr} \leq s_0 \leq g s_0^{gr}$. The functions $\{\varphi_{0,j}\}_{j=1,\ldots,N}$ admit the following sparse approximation form: for every $j = 1, \ldots, N$,

$$\varphi_{0,j}(\mathbf{x}_{j,t-h}) = \mathbf{z}'_{j,t-h}\boldsymbol{\theta}_{0,j} + B_{0,t,j}(g), \qquad \sum_{j=1}^{N} \mathbb{1}\{\|\boldsymbol{\theta}_{0,j}\|_2 > 0\} \leq s_0^{gr},$$

$$\sum_{j=1}^{N}\sum_{i=1}^{g} \mathbb{1}\{|\theta_{0,ji}| > 0\} \leq s_0, \qquad \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{N} B_{0,t,j}(g)\right)^2 \leq \frac{s_0}{16T}\sigma_0^2.$$

For the asymptotic analysis we will let N, g, s_0 and s_0^{gr} to $\nearrow$ with T.

This together with the Assumption 6.1 allows the size of the approximation model to grow with the sample size T.

Assumption 6.2 (Hyperparameters)

Let $\lambda_{\max} := \max\{\lambda_{1,j};\, j \leq N\}$ *and assume that* $\sqrt{T}/(\|\mathbf{Z}\|_o \min\{\log(gs_0^{gr}), \log(T)\}) < C$ *with probability* 1 *for some* $C > 0$. *The scale parameters* $\lambda_{1,j}$ *are allowed to change with T and belong to the range:*

$$\max\left\{\frac{1}{|S_{0,j}|}, \frac{\sqrt{T}}{\|\mathbf{Z}\|_o}\right\} \underline{c} \leq \lambda_{1,j} \leq \lambda_{\max} \leq \overline{C} \min\{\log(s_0^{gr} g), \log(T)\}$$

for two positive constants $1 < \underline{c} < \overline{C} < \infty$ *and where*

$$\|\mathbf{Z}\|_o := \max\{\|\mathbf{Z}_j\|_{op};\, 1 \leq j \leq N\},$$

where $\mathbf{Z}_j$ *is the* $(T \times g)$*-submatrix of* $\mathbf{Z}$ *made of all the rows and the columns corresponding to the indices in the j-th group.*

Conditional on π_0 and $\{v_{ji}\}_{i=1}^g$, the prior (6)-(7) can be better understood as a mixture of a degenerate Gaussian process and a Dirac distribution at zero. To see this:

- denote $z_{j,t-h,i} := z_{j,i}(\mathbf{x}_{j,t-h})$;
- let $\Omega_{0,j} : \mathcal{H} \to \mathcal{H}$ be a covariance operator:

$$\forall h \in \mathcal{H}, \qquad (\Omega_{0,j}h)(\cdot) := \sum_{i=1}^{g} v_{j,i}^2 \langle z_{j,i}, h\rangle z_{j,i}(\cdot),$$

  where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$.

If $\mathcal{H}$ is an infinite dimensional space (or it has dimension $> g$), then $\Omega_{0,j}$ is not injective and has a nontrivial null space that contains $B_{t,j}(g)$.

Hence, (6)-(7) induce the following conditional mixture prior on the random function φ_j: for every $j \in \{1, \ldots, N\}$,

$$\varphi_j | \pi_0, \{v_{j,i}\}_{i=1}^g, g \sim (1-\pi_0)\mathcal{GP}(0, \Omega_{0,j}) + \pi_0 \delta_0(\varphi_j). \tag{14}$$

The **induced prior** on $\{\theta_{j,i}, j = 1, \ldots, N, i = 1, \ldots, g\}$ (conditional on π_0, π_1) is as follows.

- $\theta_{j,i} = 0$ with probability:

$$\Pi(\theta_{j,i} = 0|\pi_0, \pi_1) =$$
$$\Pi(\theta_{j,i} = 0|j - \text{th group is active}, \pi_0, \pi_1)\Pi(j - \text{th group is active}|\pi_0, \pi_1)$$
$$+\Pi(\theta_{j,i} = 0|j-\text{th group is not active}, \pi_0, \pi_1)\Pi(j-\text{th group is not active}|\pi_0, \pi_1)$$
$$= \pi_1(1 - \pi_0)(1 - \pi_1^g) + 1(\pi_0 + \pi_1^g - \pi_0\pi_1^g),$$

where $\Pi(j - \text{th group is active}|\pi_0, \pi_1)$ is equal to

$$\Pi\left(\|\mathbf{b}_j\|_2 > 0\right) \Pi\left(\exists i \in \{1, \ldots, g\}; v_{ji} > 0\right).$$

- Conditionally on $\{\theta_{j,i} \neq 0\}$: the Lebesgue density of $\theta_{j,i}$ is

$$f_{\theta_{j,i}}(\theta_{j,i}|\tau_j) = \int_{\mathbb{R}_+} \frac{1}{\pi\tau_j} \underbrace{\frac{1}{t} \exp\left\{-\frac{1}{2}\left(\frac{\theta_{j,i}^2}{t\tau_j^2} + t\right)\right\}}_{=GIG(t;1,\theta_{j,i}^2/\tau_j^2,0)2K_0(|\theta_{j,i}|/\tau_j)} dt = \frac{2}{\pi\tau_j} K_0(|\theta_{j,i}|/\tau_j),$$

  where $GIG(t;a,b,p)$ denotes the pdf of a Generalized Inverse Gaussian distribution with parameters a, b and p, and $K_0(\cdot)$ is the modified Bessel function of the second kind. We remark that

$$\sqrt{\pi/2}e^{-|\theta_{j,i}|/\tau_j}(|\theta_{j,i}|/\tau_j+a)^{-1/2} < K_0(|\theta_{j,i}|/\tau_j) < \sqrt{\pi/2}e^{-|\theta_{j,i}|/\tau_j}(|\theta_{j,i}|/\tau_j)^{-1/2}$$

  for every $a \geq 1/4$.

- $f_{\theta_{j,i}}(\theta_{j,i}|\tau_j)$ is upper bounded by the density of a $Gamma(1/2, \tau_j)$.

- The induced conditional prior on $\theta_{j,i}$ is:

$$\begin{aligned} \theta_{j,i}|\{\text{group } j \text{ is active}\}, \pi_0, \pi_1, \tau_j &\sim (1-\pi_1) f_{\theta_{j,i}}(\theta_{j,i}|\tau_j) + \pi_1 \delta_0(\theta_{j,i}) \\ \theta_{ji}|\{\text{group } j \text{ is not active}\}, \pi_0, \pi_1, \tau_j &\sim \delta_0(\theta_{j,i}). \end{aligned} \quad (15)$$

We consider the linear model with grouped predictors:

$$y_t = 0.2 + 0.3y_{t-1} + \sum_{j=1}^{N}\sum_{p=1}^{g} z_{j,t,p}\theta_{p,j} + \varepsilon_t$$

$$z_{j,t,p} = 0.9z_{j,t-1,p} + \epsilon_{j,t,p}$$

$$\begin{pmatrix} \varepsilon_t \\ \boldsymbol{\epsilon}_t \end{pmatrix} \sim \text{i.i.d. } \mathcal{N}\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\epsilon \end{pmatrix}\right],$$

- $\boldsymbol{\Sigma}_\epsilon = \mathcal{S}_\epsilon \mathcal{R}_\epsilon \mathcal{S}_\epsilon$ block-diagonal matrix,
  - $\mathcal{S}_\epsilon$ is a $(Ng \times Ng)$ diagonal matrix with elements σ_ϵ,
  - $\mathcal{R}_\varepsilon$ is a block-diagonal Toeplitz correlation matrix with N blocks – each of size $(g \times g)$ – and featuring diagonal elements equal to one and off-diagonal elements $\rho_{j,\epsilon}^{|p-p'|}$ for all $p \neq p'$.
- S_0^{gr} and $S_{0,j}$ randomly set.

Example 1: Design of the experiments (cont.)

- $|\theta_{p,j}| = 0.5$, for each $j \in S_0^{gr}$ and $p \in S_{0,j}$, and 0 otherwise.
- The sign of $\theta_{p,j}$ is a fixed realization of random draws with replacement from $\{-1, 1\}$.
- $\sigma = 0.50$ and σ_ϵ fixed such that NSR $= 0.2$.
- $\rho_{j,\epsilon} = 0.50$.

Table: Monte Carlo simulations: estimation and selection accuracy

Ng	N	g	s_0^{gr}	MSE_θ	VAR_θ	BIAS^2_θ	TPR_N	TPR_g	MCC_N	MCC_g
100	5	20	1	0.01	0.00	0.00	99.7	99.8	1.00	1.00
	10	10	1	0.00	0.00	0.00	99.7	99.7	1.00	1.00
	10	10	5	0.04	0.04	0.00	99.5	99.4	0.99	0.99
	20	5	5	0.03	0.03	0.00	99.8	99.7	1.00	0.99
	20	5	10	1.90	0.74	1.17	41.6	39.6	0.49	0.54
300	5	60	1	0.01	0.01	0.00	99.5	99.8	1.00	1.00
	10	30	1	0.01	0.00	0.00	99.7	100.0	1.00	1.00
	10	30	5	0.05	0.05	0.00	98.8	98.8	0.99	0.99
	20	15	5	0.06	0.06	0.00	97.9	98.0	0.98	0.98
	20	15	10	2.54	0.32	2.22	18.3	16.9	0.29	0.37

Table: $T = 200$, $s_{0,j} = 1$, $s_0 = s_0^{gr}$. MSE, VAR, and BIAS^2 denote the Mean Squared Error, the Variance, and the Squared Bias, respectively. TPR and MCC denote the True Positive Rate and the Matthews Correlation Coefficient, respectively, computed at the groups level (subscript N) and at the variables level (subscript g).

- Selection deteriorates only with s_0^{gr} and/or $s_0 \uparrow$ while T fixed (consistently with the theoretical rate).
- BSGS-SS largely outperforms the Sparse Group Lasso.

Table: Monte Carlo simulations: modified DGP

				$\rho_{j,\epsilon} = 0.75$		$\rho_{j,\epsilon} = 0.75$ $\mathcal{R}_\epsilon$ full		$\rho_{j,\epsilon} = 0.75$ $\mathcal{R}_\epsilon$ full $\epsilon_t \sim$ Skew-$\mathcal{N}$		NSR $= 0.5$	
Ng	N	g	s_0^{gr}	TPR_N	TPR_g	TPR_N	TPR_g	TPR_N	TPR_g	TPR_N	TPR_g
100	5	20	1	99.8	99.8	100.0	100.0	100.0	100.0	97.5	98.2
	10	10	1	99.8	100.0	99.8	100.0	99.3	99.7	97.5	98.3
	10	10	5	99.5	98.7	98.6	97.2	99.2	98.7	70.9	67.2
	20	5	5	99.8	99.1	94.9	92.4	95.8	93.7	74.0	73.0
	20	5	10	48.9	40.2	51.8	42.3	54.0	41.8	17.4	16.3
300	5	60	1	99.7	99.7	99.7	99.7	99.7	99.8	97.3	98.3
	10	30	1	99.7	99.8	99.5	99.7	99.5	99.7	95.0	97.3
	10	30	5	98.2	96.7	98.5	97.6	98.1	96.8	53.8	50.3
	20	15	5	98.8	98.1	98.7	98.2	98.0	96.7	51.0	49.8
	20	15	10	22.9	17.4	24.6	18.6	25.6	19.2	13.0	11.6

Table: See Table 15. The Skew-$\mathcal{N}$ is parameterized as in Azzalini & Capitanio (2014), with skew parameter set at -5.

- Results overall robust to changes in some key calibration parameters.

Table: Monte Carlo simulations: predictive accuracy

				BSGS-SS			BSGL		
Ng	N	g	s_0^{gr}	RMSFE	LogS	CRPS	RMSFE	LogS	CRPS
100	5	20	1	0.71	0.35	0.71	0.77	0.27	0.77
	10	10	1	0.71	0.35	0.71	0.75	0.29	0.75
	10	10	5	0.73	0.32	0.73	0.94	0.06	0.95
	20	5	5	0.72	0.34	0.72	0.86	0.15	0.87
	20	5	10	0.95	0.05	0.95	0.98	0.01	0.99
300	5	60	1	0.71	0.34	0.71	0.87	0.14	0.88
	10	30	1	0.71	0.35	0.71	0.82	0.20	0.83
	10	30	5	0.74	0.31	0.74	1.11	-0.11	1.12
	20	15	5	0.73	0.32	0.73	1.03	-0.03	1.04
	20	15	10	1.02	-0.03	1.03	1.07	-0.07	1.08

Table: See Table 15. RMSFE, LogS, and CRPS denote respectively the root mean squared forecast error, the log-score, and the continuously ranked probability score, in relative terms with respect to the AR(1) benchmark.

- We adopt a frequentist point of view: (φ_0, σ_0^2) denotes the true value of (φ, σ^2) that generates the data.
- $\mathbf{E}_0[\cdot]$ denotes the expectation taken with respect to the true data distribution $\mathcal{N}_T\left(\sum_{j=1}^{N} \varphi_{0,j}^{(T)}, \sigma_0^2 \mathbf{I}_T\right)$, conditional on $(\mathbf{X}, \varphi_0, \sigma_0^2)$.
- $\boldsymbol{\theta}_0$ = true value of the approximation.
- Our asymptotic analysis is for $T \to \infty$. We allow N, s_0^{gr}, s_0 and $g \to \infty$ with T.

Rate of contraction of the posterior distribution:

$$\epsilon := \max\left\{\sqrt{\frac{s_0^{gr}\log(N)}{T}}, \sqrt{\frac{s_0\log(T)}{T}}, \sqrt{\frac{s_0\log(s_0^{gr}g)}{T}}\right\}$$

which is equal to

$$\epsilon := \max\left\{\sqrt{\frac{s_0^{gr}\log(N)}{T}}, \sqrt{\frac{s_0\log(s_0^{gr}g)}{T}}\right\}$$

if $\log(T) \leq \max\{s_0^{gr}\log(N), s_0\log(s_0^{gr}g)\}$.

If in addition, $\log(N) \asymp \log(N) - \log(s_0^{gr})$ and $\log(s_0^{gr}g) \asymp \log(s_0^{gr}g) - \log(s_0)$ then ϵ corresponds to the minimax rate for recovering φ

$$\max\left\{\sqrt{\frac{s_0^{gr}\log(N/s_0^{gr})}{T}}, \sqrt{\frac{s_0\log(s_0^{gr}g/s_0)}{T}}\right\}.$$

given in Cai et al. (2022) and in Li et al. (2022).

- If $N = 1$, then $s_0^{gr} = 1$, $s_0 = |S_{0,1}|$ and $\epsilon := \underbrace{\sqrt{\frac{|S_{0,1}| \log(g)}{T}}}_{\text{rate for recovery of sparse vectors over } \ell_0\text{-balls}}$.

- If only 1 element per group (♯ of groups = ♯ of parameters), then $g = 1$, $s_0 = s_0^{gr}$ and $\epsilon := \underbrace{\sqrt{\frac{s_0^{gr} \log(N)}{T}}}_{\text{rate for recovery of sparse vectors over } \ell_0\text{-balls}}$.

- The required sample size to achieve $\epsilon \to 0$:

$$T > C \max \{s_0^{gr} \log(N), s_0 \log(s_0^{gr} g)\} .$$

  - $s_0^{gr} \log(N)$ corresponds to the complexity of capturing s_0^{gr} non-zero groups,
  - $s_0 \log(s_0^{gr} g)$ corresponds to the complexity of estimating s non-zero elements of $\boldsymbol{\theta}$ in s_0^{gr} known groups (estimation over ℓ_0-balls).

Assumption 6.3

For positive and bounded constants $\underline{\sigma}^2$, $\overline{\sigma}^2$, c_g *and* $c_{\boldsymbol{\theta}}$, *suppose that:*

(i) $0 < \underline{\sigma}^2 \leq \sigma_0^2 \leq \overline{\sigma}^2 < \infty$;

(ii) $\max\{\log(N), \log(T)\} \leq s_0^{gr} g$;

(iii) $\max_{j \in S_0^{gr}} \max_{i \in S_{0,j}} |\theta_{0,j,i}| \leq \log(s_0^{gr} g)$.

Define $\|\mathbf{Z}\|_o := \max\{\|\mathbf{Z}_j\|_{op}; 1 \leq j \leq N\}$, where $\mathbf{Z}_j$ is the $(T \times g)$-submatrix of $\mathbf{Z}$ made of all the rows and the columns corresponding to the indices in the j-th group.

Assumption 6.4 (Hyperparameters of the prior for (π_0, π_1))

$\exists$ *constants* $\kappa_0, \kappa_1 > 0$ *such that the hyper-parameters* c_0, d_0, c_1, d_1 *of the Beta priors for* π_0 *and* π_1 *satisfy:*

(i) $\frac{d_0+j-1}{(c_0+N-j)} \leq \kappa_0 \frac{j}{[N^{u_0}(N-j+1)]}$ *for every* $\frac{\log(2)}{\log(N)} < u_0 < s_0^{gr}$ *and* $\forall j \in \{1, \ldots, N\} \subseteq \mathbb{N}$,

(ii) $\frac{d_1+j-1}{(c_1+Ng-j)} \leq \kappa_1 \frac{j}{[(Ng)^{u_1}(Ng-j+1)]}$ *for every* $\frac{\log(2)}{\log(Ng)} < u_1 < s_0$ *and* $\forall j \in \{1, \ldots, g\} \subseteq \mathbb{N}$.

- Assumptions (6.4) *(i)* and *(ii)* demand: $c_0, c_1 \uparrow$ together with N, s_0^{gr} and g and control their rate.
- To satisfy the assumption, if $d_0 = cst.$ and $d_1 = cst.$ then, $c_0 \gtrsim N^{u_0}$ and $c_1 \gtrsim (s_0^g g)^{u_1}$, for u_0, u_1 in the range of values given in the assumption and up to a constant.
- In practice, in finite samples one can choose the constants κ_0 and κ_1 very small as long as they are fixed and do not increase with T.

For positive integers s_0^{gr}, s_0 satisfying $s_0^{gr} \leq N$ and $s_0^{gr} \leq s_0 \leq g s_0^{gr}$, we define

$$\mathcal{F}(s_0, s_0^{gr}; \mathbf{Z}) :=$$

$$\left\{ (\varphi, \sigma^2); \|B(g)\|_2^2 \leq \frac{s_0 \sigma^2}{16}, s_{\boldsymbol{\theta}}^{gr} \leq s_0^{gr},\ s_{\boldsymbol{\theta}} \leq s_0,\ \|\theta\|_\infty \leq \log(s_0^{gr} g),\ \text{and } \sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2] \right\},$$

where for every vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{Ng}$:

- there is an associated group structure - by using the inverse of the $Vec(\cdot)$ operator we obtain a $(g \times N)$ matrix $\Upsilon(\boldsymbol{\theta})$ whose j-th column is equal to $(\theta_{g(j-1)+1}, \ldots, \theta_{gj})' \in \mathbb{R}^g$;
- the columns of this matrix are the groups in $\boldsymbol{\theta}$;
- $S_{\boldsymbol{\theta}}^{gr} \subseteq \{1, 2, \ldots, N\}$ is the set of indices of the active groups in $\boldsymbol{\theta}$ (the non-zero columns of $\Upsilon(\boldsymbol{\theta})$).
- $S_{\boldsymbol{\theta}} \subseteq \{1, 2, \ldots, Ng\}$ the set of nonzero elements in $\boldsymbol{\theta}$.
- For given positive integers s_0^{gr}, s_0 satisfying $s_0^{gr} \leq N$ and $s_0^{gr} \leq s_0 \leq g s_0^{gr}$, all vectors $\boldsymbol{\theta} \in \Theta$ such that $|S_{\boldsymbol{\theta}}^{gr}| \leq s_0^{gr}$ and $|S_{\boldsymbol{\theta}}| \leq s_0$ are said to be (s_0, s_0^{gr})-sparse.

Theorem 3

Suppose Assumptions 6.1, 6.2, 6.3 and 6.4 hold. Let $\epsilon \to 0$. *Then, for a sufficiently large* $M > 0$*:*

$$\sup_{(\varphi_0,\sigma_0^2)\in\mathcal{F}_0(s_0,s_0^{gr};\mathbf{Z})} \mathbf{E}_0\left[\Pi\left(\varphi; \left\|\sum_{j=1}^{N}\left(\varphi_j^{(T)}(\mathbf{X}) - \varphi_{0,j}^{(T)}(\mathbf{X})\right)\right\|_2^2 \leq MT\epsilon^2 \middle| y, \mathbf{X}\right)\right] \to 0. \tag{16}$$

Remarks:

- In the grouped predictors example:

$$\left\|\sum_{j=1}^{N}\left(\varphi_j^{(T)}(\mathbf{X}) - \varphi_{0,j}^{(T)}(\mathbf{X})\right)\right\|_2^2 = \|\mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_2^2.$$

- Similarly, in the MIDAS example:

$$\left\| \sum_{j=1}^{N} \left(\varphi_j^{(T)}(\mathbf{X}) - \varphi_{0,j}^{(T)}(\mathbf{X}) \right) \right\|_2^2 = \|\mathbf{Z}^{\infty}(\boldsymbol{\theta}^{\infty} - \boldsymbol{\theta}_0^{\infty})\|_2^2$$

with $\boldsymbol{\theta}^{\infty} = \{\theta_{j1}, \theta_{j2}, \ldots\}_{j=1}^{N}$ an infinite dimensional vector, $\mathbf{z}_t^{\infty}$ is defined similarly and $\mathbf{Z}^{\infty} = (\mathbf{z}_{1-h}^{\infty}, \ldots, \mathbf{z}_{T-h}^{\infty})'$ is a matrix with T rows and an infinite number of columns.

Sketch of the proof: posterior consistency for the Rényi divergence of order $\frac{1}{2}$,

$$d(f_0, f) := -\frac{1}{T} \log \int \sqrt{f_0 f},$$

where $f_0 = \mathcal{N}_T\left(\sum_{j=1}^N \varphi_{0,j}^{(T)}, \sigma_0^2 \mathbf{I}_T\right)$ and $f = \mathcal{N}_T(\sum_{j=1}^N \varphi_j^{(T)}, \sigma^2 \mathbf{I}_T)$.

[1]. f_0 **belongs to the Kullback-Leibler support of the prior distribution**.

Let f^g be the Lebesgue density of $\mathcal{N}_T(\mathbf{Z}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_T)$.
We show that, for large T:

$$\Pi\left((\boldsymbol{\theta}, \sigma^2);\, K(f_0, f^g) \leq T\epsilon^2, V(f_0, f^g) \leq T\epsilon^2\right) \geq e^{-C_1 T \epsilon^2}. \tag{17}$$

for a constant $C_1 = C_1(b, C_{cd}, \underline{c}, \bar{C}, \underline{\sigma}^2) > 0$ and where, for two probability densities f_1 and f_2,

$$K(f_1, f_2) := \int f_1 \log(f_1/f_2)$$

and

$$V(f_1, f_2) := \int f_1 (\log(f_1/f_2) - K(f_1, f_2))^2.$$

[2]. Let

$$\Theta(\widetilde{s}_0^{gr}, \widetilde{s}_0) := \left\{ \boldsymbol{\theta} \in \Theta; s_{\boldsymbol{\theta}}^{gr} < M_0 \frac{T\epsilon^2}{\log(N)} \text{ and } s_{\boldsymbol{\theta}} < M_1 \frac{T\epsilon^2}{\log(s_0^{gr} g)} \right\}$$

for two positive constants M_0, M_1.

Lemma 1 (Dimensionality)

Let us consider the prior in (7) *and* (8) *with* c_0, c_1, d_0, d_1 *satisfying Assumption* (6.4) *(i) - (ii). Let* $C_1, M_0, M_1 > 0$ *be some constants such that* $C_1 < \min\{u_0(M_0 - 1), u_1(M_1 - 1)\} - 3$ *that do not depend on* $(\boldsymbol{\theta}_0, \sigma_0^2)$. *Then, it holds that:*

$$\sup_{\boldsymbol{\theta}_0 \in \overline{\Theta}_0, \sigma_0^2 \in [\underline{\sigma}^2, \overline{\sigma}^2]} \mathbf{E}_0 \Pi \left(\boldsymbol{\theta}; s_{\boldsymbol{\theta}}^g \geq M_0 \frac{T\epsilon^2}{\log(N)}, \, s_{\boldsymbol{\theta}} \geq M_1 \frac{T\epsilon^2}{\log(s_0^{gr} g)} \middle| y, \mathbf{Z} \right)$$
$$\leq e^{-T\epsilon^2 \left(-2C_1 + \min\{u_0(M_0-1), u_1(M_1-1)\} - 3 \right)} + \frac{1}{C_1^2 T\epsilon^2}.$$

The support of the posterior can overshoot the true dimension s_0^{gr}, s_0 since

$$\frac{T\epsilon^2}{\log(N)} = \max\left\{ s_0^{gr}, \frac{s_0 \log(T)}{\log(N)}, \frac{s_0 \log(s_0^{gr} g)}{\log(N)} \right\} \qquad \text{and}$$

$$\frac{T\epsilon^2}{\log(s_0^{gr} g)} = \max\left\{ \frac{s_0^{gr} \log(N)}{\log(s_0^{gr} g)}, \frac{s_0 \log(T)}{\log(s_0^{gr} g)}, s_0 \right\}.$$

[3]. Define the sieves:

$$\mathcal{F}_T(C_2) := \left\{ (\boldsymbol{\theta}, \sigma^2) \in \Theta(\widetilde{s}_0^{gr}, \widetilde{s}_0) \times \mathbb{R}_+;\ \max_{1 \le j \le N} \|\boldsymbol{\theta}_j\|_2 \le \frac{C_2 + 1}{\underline{c}} \xi,\ T^{-1} \le \sigma^2 \le e^{C_2 T \epsilon^2} \right\}, \tag{18}$$

where $\xi := (T\epsilon^2)^2 \log(s_0^{gr} g)$.

Lemma 2 (Testing)

(i) There exists a constant C_2 such that for T large:

$$\Pi((\Theta(\widetilde{s}_0^{gr}, \widetilde{s}_0) \times \mathbb{R}_+) \backslash \mathcal{F}_T(C_2)) \lesssim \exp\left\{ -T\epsilon^2 C_2 \right\} \left(2 + \frac{b}{a-1} \right), \tag{19}$$

and (ii) there exists a test ϕ_T such that

$$\mathbf{E}_0 \phi_T \le e^{-M_2 T \epsilon^2 / 2}, \qquad \sup_{f^g \in \mathcal{F}_T(C_2); d(f_0, f^g) > M_1 T \epsilon^2} \mathbf{E}_{f^g}(1 - \phi_T) \le e^{-M_2 T \epsilon^2} \tag{20}$$

for some M_0 that does not depend on $(\boldsymbol{\theta}_0, \sigma_0^2)$ and where: $d(f_0, f) := -\frac{1}{T} \log \int \sqrt{f_0 f}$ (Rényi divergence of order $\frac{1}{2}$).

We now look at parameter recovery of our procedure, that is, consistency of the marginal posterior of $\boldsymbol{\theta}$ (coefficients of the approximation of φ).

Definition 2 (Smallest scaled sparse singular value.)

For every $s, r > 0$, the smallest scaled sparse singular value of dimension (s, r) is defined as

$$\widetilde{\phi}(s,r) := \inf \left\{ \frac{\|\mathbf{Z}\boldsymbol{\theta}\|_2^2}{\|\mathbf{Z}\|_o^2 \|\boldsymbol{\theta}\|_2^2},\ 0 \leq s_{\boldsymbol{\theta}}^{gr} \leq s \ \textit{and}\ 0 \leq s_{\boldsymbol{\theta}} \leq r \right\}. \tag{21}$$

- The double sparse eigenvalue condition requires that for every $s, r > 0$, $\exists$ a constant $\kappa > 0$ such that $\widetilde{\phi}(s,r) > \kappa$. Under this assumption:

$$\|\mathbf{Z}\boldsymbol{\theta}\|_2^2 \geq \kappa \|\mathbf{Z}\|_o^2 \|\boldsymbol{\theta}\|_2^2.$$

- This is the same assumption as in Li et al. (2022). In addition, they assume the columns of $\mathbf{Z}$ are normalized: $\sum_{t=1}^{T} z_{j,t-h,i}^2 = \sqrt{T}$.

- We use the notation $\widetilde{\phi}_0 := \widetilde{\phi}(M_0\widetilde{s}_0^{gr} + s_0^{gr}, M_1\widetilde{s}_0 + s_0)$ for two positive constants M_0 and M_1.

Theorem 4

Suppose Assumptions 6.1, 6.2, 6.3 and 6.4 hold. Let $\epsilon \to 0$. Then, for every constant $M_3 \geq 2M + \overline{\sigma}^2/8$ where M is as in Theorem 3 we have:

$$\sup_{(\varphi_0,\sigma_0^2)\in\mathcal{F}_0(s_0,s_0^{gr};\mathbf{Z})} \mathbf{E}_0\left[\Pi\left(\boldsymbol{\theta}\in\Theta; \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2 \geq \frac{M_3 T\epsilon^2}{\widetilde{\phi}_0\|\mathbf{Z}\|_o^2}\middle| y,\mathbf{X}\right)\right] \to 0. \qquad (22)$$

If there exists two constants $\kappa_\ell, \kappa_z > 0$ such that $\widetilde{\phi}(s,r) > \kappa_\ell$ and $\|\mathbf{Z}\|_o \leq \sqrt{\kappa_z}\sqrt{T}$ w.p.a. 1, then

$$\sup_{(\varphi_0,\sigma_0^2)\in\mathcal{F}_0(s_0,s_0^{gr};\mathbf{Z})} \mathbf{E}_0\left[\Pi\left(\boldsymbol{\theta}\in\Theta; \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2 \geq \frac{M_3\epsilon^2}{\kappa_\ell\kappa_z}\middle| y,\mathbf{X}\right)\right] \to 0. \qquad (23)$$

Let us consider the assumption $\|\mathbf{Z}\|_o \leq \sqrt{\kappa_z}\sqrt{T}$, where $\|\mathbf{Z}\|_o := \max\{\|Z_j\|_{op}; 1 \leq j \leq N\}$.

- MIDAS: by using the inequality $\|\cdot\|_{op} \leq \|\cdot\|_F$

$$\|\mathbf{Z}_j\|_{op} \leq \sqrt{T}\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{j,t-h}\mathbf{x}'_{j,t-h}\right\|_{op}\|\Phi'\Phi\|_F,$$

  where $\Phi' := (\Phi_1, \ldots, \Phi_g)$ is $p_x \times g$ and recall $\mathbf{x}_{j,t-h} = (x^H_{j,t-h}, \ldots, x^H_{j,t-h-p_x/m})'$.

- Grouped predictors: $\|\mathbf{Z}_j\|_{op} = \left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{j,t-h}\mathbf{x}'_{j,t-h}\right\|_{op}$.

- Nonlinear predictive models:

$$\|\mathbf{Z}_j\|_{op} = \left\|\sum_{t=1}^{T}\begin{pmatrix}\phi_{j1}(x_{j,t-h})\\ \vdots \\ \phi_{jg}(x_{j,t-h})\end{pmatrix}(\phi_{j1}(x_{j,t-h}), \ldots, \phi_{jg}(x_{j,t-h}))\right\| = \mathcal{O}_p(\sqrt{T}).$$

h steps-ahead forecasts are obtained from the posterior predictive density for $y_\tau | x_{\tau-h}, \tau > T$:

$$f(y_\tau | x_{\tau-h}, y, \mathbf{X}) = \int f_0(y_\tau | \varphi, \sigma^2, x_{\tau-h}) \Pi(\varphi, \sigma^2 | y, \mathbf{X}) d\varphi d\sigma^2 \tag{24}$$

where

- Draws from the predictive distribution (24) can be obtained directly from the Gibbs sampler.
- Point and density forecasts are evaluated through standard metrics, such as the root mean squared forecast error (RMSFE), the log-score (LogS), and the continuously ranked probability score (CRPS), averaged over $T_{oos} = 50$ out-of-sample observations.

Evaluate it by using the mean KL-divergence:

$$\mathbf{E}_{x_{\tau-h}} KL(f_0(y_\tau|x_{\tau-h},\varphi_0,\sigma_0^2), f(y_{\tau|\tau-h}|y,\mathbf{X}))$$
$$= \int\int \log\left(\frac{f_0(y_\tau|x_{\tau-h},\varphi_0,\sigma_0^2)}{f(y_\tau|x_{\tau-h},y,\mathbf{X})}\right) f_0(y_\tau|x_{\tau-h},\varphi_0,\sigma_0^2) dy P(dx_{\tau-h}).$$

Theorem 5

Suppose Assumptions 6.1, 6.2, 6.3 and 6.4 hold. Let $\epsilon \to 0$. Then,

$$\sup_{(\varphi_0,\sigma_0^2)\in\mathcal{F}_0(s_0,s_0^{gr};\mathbf{Z})} \mathbf{E}_{x_{\tau-h}}\mathbf{E}_0 KL(f_0(y_\tau|x_{\tau-h},\varphi_0,\sigma_0^2), f(y_\tau|x_{\tau-h},y,\mathbf{X})) \to 0. \quad (25)$$

▸ Extension: Stochastic Volatility