CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

WORKING PAPER SERIES

BAYESIAN BI-LEVEL SPARSE GROUP REGRESSIONS FOR MACROECONOMIC DENSITY FORECASTING

Matteo Mogliani, Anna Simoni

N°12/Septembre 2025

CREST
Center for Research in Economics and Statistics
UMR 9194

5 Avenue Henry Le Chatelier
TSA 96642
91764 Palaiseau Cedex
FRANCE

Phone: +33 (0)1 70 26 67 00
Email: info@crest.science
https://crest.science/

Bayesian Bi-level Sparse Group Regressions for Macroeconomic Density Forecasting*

Matteo Mogliani† Anna Simoni‡

May 14, 2025

Abstract

We construct optimal macroeconomic density forecasts in a high-dimensional setting where the underlying model exhibits a known group structure. Our approach is designed for a general model which encompasses forecasting models featuring either many covariates, or unknown nonlinearities, or series sampled at different frequencies. By relying on the novel concept of bi-level sparsity in time-series econometrics, our density forecasts are based on a prior that induces sparsity both at the group level and within groups. We demonstrate the consistency of both the predictive density and the posterior distribution of the model parameter. For the posterior distribution, we show that it contracts at the minimax-optimal rate and, asymptotically, puts mass on a set that includes the support of the model. While predictors in the same group are usually characterized by strong covariation and common characteristics and patterns, our theory allows for correlation even between groups. Finite sample performance is illustrated through Monte Carlo experiments and a real-data nowcasting exercise of the US GDP growth rate.

Keywords: Density forecasts, Data-rich environment, nonlinearities, posterior predictive distribution, Posterior contraction, Spike-and-slab prior

*An earlier version of this paper circulated under the title "Bayesian Bi-level Sparse Group Regressions for Macroeconomic Forecasting".

†Banque de France - International Macroeconomics Division, 46-1374 DGSEI-DECI-SEMSI, 31 Rue Croix des Petits Champs, 75049 Paris CEDEX 01 (France). Phone: +33(0)142925939. e-mail: matteo.mogliani@banque-france.fr

‡CREST, CNRS, École Polytechnique, ENSAE, 5 Avenue Henry Le Chatelier, 91120 Palaiseau (France). Phone: +33(0)170266837. e-mail: anna.simoni@polytechnique.edu

1 Introduction

Density forecasts of macroeconomic and financial time-series are of crucial importance for policymakers, as they provide an assessment of future (upside or downside) tail risks (Garratt et al., 2003; Adrian et al., 2019; Adams et al., 2021; Carriero et al., 2024a). Nowadays, researchers dispose of rich datasets, released by official or alternative sources, that potentially contain valuable information for predicting tail events. On the one hand, handling these datasets might be challenging, mainly because of their large dimension (*i.e.* the number of variables is large compared to the number of available observations in the time dimension). On the other hand, the variables in these datasets are (or can be) often organised in groups, a feature that may help the researcher in reducing the dimensionality. The research question of interest is then: how can we optimally exploit group-structures to construct accurate density forecast? To the best of our knowledge, there is no clear answer to this question in the literature for general macroeconomic forecasting models. The contribution of this paper is to fill this gap by constructing optimal density forecasts that take advantage of the group-structure for simultaneously *i)* reducing the potentially large dimension in the dataset, and *ii)* evaluating the tail risks probabilities. By detecting, in a data-driven way, the relevant driving factors and groups, our procedure leads to economically interpretable forecasting models and results, which is a crucial feature for decision-taking by policymakers.

Groups might be formed by predictors that present strong covariation, as well as common characteristics and patterns. Some datasets may be organised in groups by the researcher according to economic criteria. For example, real economic activity can be analysed by using a large number of official series arranged in homogeneous blocks of indicators, such as production, employment, consumption, and housing (see *e.g.* McCracken and Ng, 2016). Alternatively, datasets may be organised in a group-structure directly by the data provider. This is the case, for instance, of the

Google Search data used in Ferrara and Simoni (2022). In this paper, we widen the relevance of group-structures beyond datasets organised in groups, and we point out that specific forecasting models – like models with unknown nonlinearities or with covariates observed at mixed frequencies – might present a group-structure.

To allow for different sources of group-structure, we consider a general model flexible enough to include, among others, *i)* linear regression models with many predictors organised in groups, *ii)* mixed-frequency regression models with non-parametric weighting functions, and *iii)* nonlinear forecasting models as special cases. The nonparametric models *ii)* and *iii)* are suitable for capturing complex relationships between the target variable and the predictors, and the group-structure arises there from generalized Fourier expansions of these relationships.

While the true model is allowed to have an infinite number of elements in each group, we approximate the model by restricting each group to have no more than $g \geq 1$ components. This introduces a bias in finite samples, but this bias vanishes asymptotically as g is allowed to increase with the sample size T. The approximate model is assumed to be bi-level sparse: only s_0^{gr} among the N groups are active, and within each of the s_0^{gr} active groups, only a few elements have a non-zero impact on the target variable. This means that some groups and some predictors within an active group can be irrelevant for modelling and forecasting the target variable, conditional on the remaining predictors.

Our Bayesian procedure is based on a hierarchical prior that generates a group structure with bi-level sparsity and charges only the approximate model. To induce exact bi-level sparsity, we use $(g+1)N$ spike-and-slab priors specified as mixtures between a continuous distribution and a Dirac distribution at zero, the latter inducing exact zeros with non-negligible probability. Density forecasts are obtained from the posterior predictive distribution, which is well-known to dominate plug-in predictive distributions when there is prior information. We establish frequentist asymptotic optimality of our Bayesian procedure in a setting where the true data

generation process (DGP) has a bi-level sparse group-structure. Importantly, optimality does not require orthogonality between the groups of predictors, such that cross-correlation between covariates in different groups is allowed by our theory. Asymptotic properties are established for the sample size T increasing to infinity, as well as for the number of groups N, active groups s_0^{gr}, components per groups g, and active predictors in the model s_0 increasing to infinity with T. As a by-product, we provide parameter recovery and point forecasts, as well as the contraction rate of the posterior distribution of the in-sample prediction error. Our posterior contraction rate attains the minimax rate established in Cai et al. (2022) and Li et al. (2024) for bi-level sparse settings.

The bi-level sparse group-structure that we consider in this paper has remarkable advantages over the sparse group-structure considered *e.g.* in Mogliani and Simoni (2021), which only accounts for sparsity among groups but not within groups. First, if the true model is bi-level sparse, then the posterior of Mogliani and Simoni (2021) obtained under group sparsity would not put zero mass asymptotically on models with dense groups. Therefore, the true sparsity pattern is not recovered with positive probability, implying a worsening of both selection of relevant driving factors and out-of-sample predictive accuracy. This issue is illustrated in our first Monte Carlo simulation. Second, the contraction rate with bi-level sparsity is faster than the rate with group sparsity if $g \gg \log(N)/\log(T)$ and if $s_0^{gr} g \gg \log(s_0^{gr} g) s_0$. It follows that, in this case, the required sample size for consistency would be large. This is not appealing for macroeconomic forecasting, where the time-series dimension T is usually not large.

Finally, our framework can be easily extended to account for stochastic volatility and ARMA errors, which are important features in several macroeconomic applications and have been proven to improve predictive results (Chan and Hsiao, 2014; Zhang et al., 2020; Carriero et al., 2024b). We use this extended setting in an em-

pirical application on density nowcasting the US quarter-on-quarter GDP growth rate.

The paper is structured as follows. In Sections 2 and 3 we present respectively the model and the prior. Conditional posterior and predictive distributions are provided in Section 4. The asymptotic properties of our procedure are analysed in Section 5, while Monte Carlo experiments are discussed in Section 6. Section 7 provides an empirical application. Section 8 concludes. Additional theoretical results and all the proofs are reported in the Online Appendix.

2 The model

2.1 The sampling model

Let y_t be the series of interest to be predicted h-steps ahead. At each time t, a large number of predictors, organized into N groups and denoted by $\mathbf{x}_{j,t} := \{x_{j,t,i}\}_{i\geq 1}$ for every $j \in \{1,\dots,N\}$, are available to the forecaster. A group j might be formed, for instance, by indicators belonging to a given sector or category, by lagged values of one predictors or the dependent variable, or by functional transformations of one predictor. Density forecasts of y_T, conditional on information available up to $T-h$, are based on the model:

$$y_t = \sum_{j=1}^{N} \varphi_j(x_{j,t-h,1}, x_{j,t-h,2}, \dots) + \varepsilon_t, \qquad \mathbb{E}\left(\varepsilon_t | \{\mathbf{x}_{j,t-h-\ell}\}_\ell,\, j = 1,\dots,N\right) = 0 \quad (1)$$

for every $h \geq 0$ and $t = 1,\dots,T$. Examples of model (1) will be presented in Section 2.3 and an extension to the case where one group contains lagged values of y_t is considered in Appendix A.7. For every $j \in \{1,\dots,N\}$, $\varphi_j(\cdot)$ denotes a j-specific unknown function of $\mathbf{x}_{j,t-h}$, belonging to a separable Hilbert space $\mathcal{H}_j$ and taking values in $\mathbb{R}$. Depending on the dimension of $\mathbf{x}_{j,t-h}$, each function φ_j might depend on a potentially infinite number of arguments. We assume that ε_t are independent and identically distributed according to a $\mathcal{N}(0,\sigma^2)$ distribution.

Therefore, by introducing the vector $\mathbf{x}_t := (\mathbf{x}'_{1,t}, \ldots, \mathbf{x}'_{N,t})'$ of potentially infinite dimension, the matrix $\mathbf{X} := (\mathbf{x}_{1-h}, \ldots, \mathbf{x}_{T-h})'$ with T rows, and the N-vectors $\varphi(\mathbf{x}_t) := (\varphi_1(\mathbf{x}_{1,t}), \ldots, \varphi_N(\mathbf{x}_{N,t}))'$ and $\varphi := (\varphi_1, \ldots, \varphi_N)'$, we write the sampling model as:

$$y_t|\mathbf{x}_{t-h}, \varphi, \sigma^2 \sim \mathcal{N}\left(\sum_{j=1}^{N} \varphi_j(\mathbf{x}_{j,t-h}), \sigma^2\right), \qquad \forall h \geq 0, \quad \forall t = 1, \ldots, T \tag{2}$$

and the joint sampling distribution of $y := (y_1, \ldots, y_T)'$ conditional on $\mathbf{X}$ is $\prod_{t=1}^{T} \mathcal{N}\left(\sum_{j=1}^{N} \varphi_j(\mathbf{x}_{j,t-h}), \sigma^2\right)$.

For $j = 1, \ldots, N$, let $\{z_{j,t-h,i}\}_{i\geq 1}$ be transformations of the elements of $\mathbf{x}_{j,t-h}$ such that the function $\varphi_j(\mathbf{x}_{j,t-h})$ writes as $\varphi_j(\mathbf{x}_{j,t-h}) = \sum_{i=1}^{\infty} \theta_{j,i} z_{j,t-h,i}$. In the remainder of this paper, we shall use this expression for the function $\varphi_j(\cdot)$. To reduce the dimension of the model, each function $\varphi_j(\mathbf{x}_{j,t-h})$ is then approximated by $\sum_{i=1}^{g} \theta_{j,i} z_{j,t-h,i}$, where $g \geq 1$ is a truncation parameter. We introduce the following notation associated with this approximation: for every $j = 1, \ldots, N$, define $\boldsymbol{\theta}_j := (\theta_{j,1}, \ldots, \theta_{j,g})' \in \mathbb{R}^g$, $\boldsymbol{\theta} := (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_N)' \in \Theta \subset \mathbb{R}^{Ng}$, $\mathbf{z}_{j,t-h} := (z_{j,t-h,1}, \ldots, z_{j,t-h,g})'$, $\mathbf{z}_t := (\mathbf{z}'_{1,t}, \ldots, \mathbf{z}'_{N,t})'$ a $(Ng \times 1)$ vector, and $\mathbf{Z} := (\mathbf{z}_{1-h}, \ldots, \mathbf{z}_{T-h})'$ a $(T \times Ng)$ matrix. By using this notation, the approximation bias in the mean is:

$$\begin{aligned} \forall t = 1, \ldots, T, \quad B_t(g) \equiv B_{t|t-h}(g) &:= \mathbb{E}[y_t|\mathbf{x}_{t-h}] - \mathbf{z}'_{t-h}\boldsymbol{\theta} \\ &= \sum_{j=1}^{N} \left(\varphi_j(\mathbf{x}_{j,t-h}) - \mathbf{z}'_{j,t-h}\boldsymbol{\theta}_j\right) = \sum_{j=1}^{N} \sum_{i>g} \theta_{j,i} z_{j,t-h,i} =: \sum_{j=1}^{N} B_{t,j}(g), \end{aligned}$$

and $B(g) := (B_1(g), \ldots, B_T(g))'$ is a T-vector. Therefore, $B_{t,j}(g) := \sum_{i>g} \theta_{j,i} z_{j,t-h,i}$. In this paper we adopt a Bayesian approach and specify a convenient prior for $\boldsymbol{\theta}$ such that the induced prior for $B_t(g)$ is degenerate at zero (see Section 3).

2.2 The bi-level sparsity structure

Let (φ_0, σ_0^2) be the true value of (φ, σ^2) that generates the data. Under the gaussianity assumption of the error term, the true conditional distribution of y given $\mathbf{X}$ is $\prod_{t=1}^{T} \mathcal{N}\left(\sum_{j=1}^{N} \varphi_{0,j}, \sigma_0^2\right)$, with Lebesgue density denoted by f_0. The approximation bias of the true sampling mean is denoted by $B_0(g) = (B_{0,1}(g), \ldots, B_{0,T}(g))$ and the true value of $\boldsymbol{\theta}$ in the approximation of $\varphi_0(\mathbf{x}_t)$ by $\boldsymbol{\theta}_0 := (\boldsymbol{\theta}_{0,1}', \ldots, \boldsymbol{\theta}_{0,N}')'$. The latter is assumed to be bi-level group sparse in the sense explained below.

Exact bi-level group sparsity is the feature of the model that guarantees the existence of 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. To the best of our knowledge, the assumption of exact bi-level sparsity dates back to the recent contributions of Huang et al. (2012), Simon et al. (2013), and Breheny (2015), which propose algorithms for estimation and sparse recovery in cross-section models. Our approach is instead designed for time-series models, with an explicit predictive density motivation and a theoretical foundation.

To clarify the concept of exact bi-level sparsity of $\boldsymbol{\theta}_0$, let $S_0^{gr} := \{1 \leq j \leq N; \|\boldsymbol{\theta}_{0,j}\|_2 > 0\} \subset \{1, 2, \ldots, N\}$ be the set of indices of the groups (subvectors) in $\boldsymbol{\theta}_0$ with at least one nonzero component (active groups), and let $s_0^{gr} := |S_0^{gr}|$ denote the cardinality of S_0^{gr}. If $S_0^{gr} \neq \varnothing$, then for every $j \in S_0^{gr}$ let $S_{0,j}$ be the set of the indices of the nonzero elements in $\boldsymbol{\theta}_{0,j}$, such that $S_0 = \bigcup_{j \in S_0^{gr}} |S_{0,j}|$ and $s_0 := |S_0| = \sum_{j \in S_0^{gr}} |S_{0,j}|$. Remark that $s_0 \geq 1$ since there is at least one active group under the assumption $S_0^{gr} \neq \varnothing$. If $S_0^{gr} = \varnothing$, then $s_0 = 0$. Moreover, $s_0 \equiv s_0(g)$ is a non-decreasing function of g. Hence, we say that $\boldsymbol{\theta}_0$ is (s_0, s_0^{gr})-sparse.

The next assumption supposes exact bi-level sparsity of $\boldsymbol{\theta}_0$ and controls the approximation bias $B_{0,t}(g)$ relative to the estimation error. The latter is similar to Belloni et al. (2014, Condition ASM).

Assumption 2.1 (Exact bi-level sparsity and bias) *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 $j = 1, \ldots, N$ and every $t = 1, \ldots, T$,*

$$\begin{aligned} \varphi_{0,j}(\mathbf{x}_{j,t-h}) &= \mathbf{z}'_{j,t-h}\boldsymbol{\theta}_{0,j} + B_{0,t,j}(g), \\ \sum_{j=1}^{N} \mathbb{1}\{\|\boldsymbol{\theta}_{0,j}\|_2 > 0\} &= s_0^{gr}, \qquad \sum_{j=1}^{N}\sum_{i=1}^{g} \mathbb{1}\{|\theta_{0,ji}| > 0\} = s_0, \qquad \|B_0(g)\|_2^2 \leq \frac{1}{16} s_0 \sigma_0^2. \end{aligned}$$

In Section 5, we will let N, g, s_0 and s_0^{gr} to increase with T. This, together with Assumption 2.1, will allow the size of the approximation model to grow with the sample size T. The constant value in the denominator of the upper bound of the approximation bias can be replaced by any constant larger or equal than 16.

2.3 Examples

2.3.1 Linear regression models with grouped predictors

Many datasets used for macroeconomic forecasting display a large panel of real and financial indicators. Theses series can be often organized in homogeneous groups by following either economic information or statistical procedures. For example, a nominal group for different price indicators, an output group for supply and production indicators, a financial group for interest rates and stock prices, etc. In this paper, we propose a procedure for density forecasting that takes into account the group structure in large datasets.

We suppose that each group of covariates contains at most g elements. Let y_t be the target variable to be predicted h-steps ahead, $j \in \{1, \ldots, N\}$ be the group index, and $\mathbf{x}_{j,t}$ be the g-vector of variables in each group j. Then, by assuming a linear model: $\forall t = 1, \ldots, T$ and $h > 0$,

$$y_t = \sum_{j=1}^{N} \mathbf{x}'_{j,t-h}\boldsymbol{\theta}_j + \varepsilon_t, \qquad \mathbb{E}\left(\varepsilon_t | \{\mathbf{x}_{j,t-h-\ell}\}_\ell,\, j = 1, \ldots, N\right) = 0 \tag{3}$$

which can be cast in model (1) with $\mathbf{z}_{j,t} = \mathbf{x}_{j,t}$, $\varphi_j(\mathbf{x}_{j,t-h}) = \mathbf{x}'_{j,t-h}\boldsymbol{\theta}_j$ and zero approximation bias.

2.3.2 Nonparametric mixed-frequency regression models

Consider a low-frequency variable y_t^L, where $t = 1, \dots, T$ indexes the low frequency time unit, and consider $(N-1)$ high-frequency variables $x_{j,t}^H$ for $j = 1, \dots, N-1$. By denoting with m the number of times the higher sampling frequency appears in the low-frequency time unit t, then $t - k/m$ denotes the k-th past high frequency period for $k = 0, 1, 2, \dots$. To simplify the notation, we set the same m for all the groups, but our framework can accommodate the case with a m_j specific to each group. Let us define the high-frequency lag operator $L^{1/m}$, such that $L^{1/m} x_{j,t}^H = x_{j,t-1/m}^H$. Further, let $h = 0, 1/m, 2/m, 3/m, \dots$ be an (arbitrary) forecast horizon. For given orders $p_y, p_x \geq 0$, the general Mixed Data Sampling (MIDAS) regression model can be written as follows: $\forall t = 1, \dots, T$,

$$y_t^L = \sum_{j=1}^{N} \Psi_j(L^{1/m}) x_{j,t-h}^H + \varepsilon_t^L, \qquad \mathbb{E}\left(\varepsilon_t^L | \{x_{j,t-h-\ell}\}_{\ell,\, j=1,\dots,N}\right) = 0, \tag{4}$$

where $\Psi_j(L^{1/m})$ is the high-frequency lag polynomial

$$\Psi_j(L^{1/m}) = \sum_{u=0}^{p_x} \psi_j(u) L^{u/m}, \qquad \psi_j(\cdot) : \mathbb{R}_+ \to \mathbb{R}. \tag{5}$$

This model can be cast in model (1) with N groups, where $\varphi_j(\mathbf{x}_{j,t-h}) = \Psi_j(L^{1/m}) x_{j,t-h}^H$ for every $j = 1, \dots, N$ with $\mathbf{x}_{j,t-h} = (x_{j,t-h}^H, \dots, x_{j,t-h-p_x/m}^H)'$.[1] Previous literature has considered different parameterizations of the weighting function $\psi_j(u)$ in (5). Parametric specifications have been considered, for instance, by Foroni et al. (2015) and Ghysels et al. (2007). Mogliani and Simoni (2021) use (non-orthogonalized) algebraic power polynomials which correspond to restricted Almon lag polynomi-

[1] Model (4) can be generalized to accommodate groups of indicators sampled at the same frequency as the target variable y_t.

als, while Babii et al. (2022) use shifted orthogonal polynomials such as Jacobi and Legendre. With the method developed in the present paper, we can allow $\psi_j(\cdot)$ to belong to a separable Hilbert space $\mathcal{H}_j$ with a countable orthonormal basis $\{\phi_1(\cdot), \phi_2(\cdot), \ldots\}$ and construct density forecasts for MIDAS models. Therefore, for any $\psi_j(\cdot) \in \mathcal{H}_j$, we can write

$$\psi_j(u) = \sum_{i=1}^{\infty} \theta_{ji}\phi_i(u), \tag{6}$$

where $\theta_{ji} := \langle \psi_j, \phi_i \rangle$ is the i-th Fourier coefficient. Hence,

$$\Psi_j(L^{1/m})x_{j,t-h}^H = \sum_{u=0}^{p_x}\sum_{i=1}^{\infty} \theta_{ji}\phi_i(u)x_{j,t-h-u/m}^H = \sum_{i=1}^{\infty} \theta_{ji}\Phi_i'\mathbf{x}_{j,t-h} = \varphi_j(\mathbf{x}_{j,t-h}),$$

where $\Phi_i := (\phi_i(0), \phi_i(1), \ldots, \phi_i(p_x))'$. By cutting the sum in i at some $g > 0$, we get $\varphi_j(\mathbf{x}_{j,t-h}) = \Psi_j(L^{1/m})x_{j,t-h}^H = \sum_{i=1}^{g} \theta_{ji}\Phi_i'\mathbf{x}_{j,t-h} + B_{t,j}(g)$, which yields a mixed-frequency regression model with approximated high-frequency polynomial $\sum_{i=1}^{g} \theta_{ji}\Phi_i'\mathbf{x}_{j,t-h}$, $z_{j,t-h,i} = \Phi_i'\mathbf{x}_{j,t-h}$, $\mathbf{z}_{j,t-h} = (\mathbf{x}_{j,t-h}'\Phi_1, \ldots, \mathbf{x}_{j,t-h}'\Phi_g)'$, and an approximation bias at time t given by $B_{t,j}(g) = \sum_{i>g} \theta_{ji}\Phi_i'\mathbf{x}_{j,t-h}$.

2.3.3 Nonlinear predictive regression models with interaction effects

In many settings, it may be desirable to account for possible nonlinearities in the conditional mean function of the target variable y_t. Model (1) can accommodate different types of nonlinearities. The simplest case is when, given p covariates, the effect of each covariate on y_t can be separated in p nonlinear functions and interaction effects are taken into account by using an additive partially linear model: $\forall t = 1, \ldots, T$ and $h > 0$,

$$y_t = \sum_{j=1}^{p} \varphi_j(x_{j,t-h}) + z_{t-h}'\boldsymbol{\theta}_{p+1} + \varepsilon_t, \qquad \mathbb{E}\left(\varepsilon_t | \{x_{j,t-h-\ell}\}_{\ell,\, j=1,\ldots,N}\right) = 0, \tag{7}$$

where $z_{t-h} := \{x_{j,t-h}x_{k,t-h}\}_{k>j}$ is a $p(p-1)/2$-vector that includes the interactions between covariates. Here, we have $N = p + 1$ groups, with the last group having a very high number of components (of the order p^2), and φ_j is a function of only one covariate taking values in a separable Hilbert space $\mathcal{H}_j$ with a countable orthonormal basis $\{\phi_{j,1}(\cdot), \phi_{j,2}(\cdot), \ldots\}$: $\varphi_j \in \mathcal{H}_j$. Then, for every $j \in \{1, \ldots, p\}$, $\varphi_j(x_{j,t-h}) = \sum_{i=1}^{\infty} \theta_{ji}\phi_{j,i}(x_{j,t-h})$, where $\theta_{ji} := \langle \varphi_j, \phi_{j,i} \rangle$ is the i-th Fourier coefficient and $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathcal{H}_j$. For some $g > 0$ we can approximate $\varphi_j(x_{j,t-h})$ as

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

which yields an approximation bias at time t given by $B_{t,j}(g) = \sum_{i>g} \theta_{ji}\phi_{ji}(x_{j,t-h})$. Alternatively, we can replace $z'_{t-h}\boldsymbol{\theta}_{p+1}$ with $\sum_{j=1}^{p} z'_{j,t-h}\boldsymbol{\theta}_{p+j}$ in (7), where $z_{j,t-h} := \{x_{j,t-h}x_{k,t-h}\}_{k\neq j}$ is a $(p-1)$-vector. The total number of groups would hence be $N = 2p$, with the last p groups of dimension $p - 1$ each, which is much lower than p^2. It follows that the choice of the group structure may imply a trade-off between the number of groups and the number of components of each group.

3 The BSGS-SS prior

With the approximation model in Section 2.1 and the Assumption 2.1 in mind, we elicit a prior that puts all its mass on the approximation $\mathbf{z}'_{t-h}\boldsymbol{\theta}$, conditional on $\mathbf{z}_{t-h}$, and that induces sparsity at the group level and within groups. Let us first define, for every group $j = 1, \ldots, N$, the following quantities: $\boldsymbol{\theta}_j = \mathbf{V}_j^{1/2}\mathbf{b}_j$, $\mathbf{b}_j := (b_{j1}, \ldots, b_{jg})'$, $\mathbf{V}_j^{1/2} := \text{diag}(v_{j1}, \ldots, v_{jg})$ and $v_{ji} \geq 0$ for $i = 1, \ldots, g$.

We treat the truncation parameter g as deterministic and, under Assumption 2.1, it may depend on s_0. Our proposed Bayesian Sparse Group Selection with Spike-and-Slab prior (BSGS-SS henceforth) is specified as: $\forall j = 1, \ldots, N$,

$$\theta_{ji} \sim \delta_0, \qquad \forall i > g, \tag{8}$$

$$\mathbf{b}_j|g,\pi_0 \overset{ind.}{\sim} (1-\pi_0)\mathcal{N}_g(0,I_g)+\pi_0\delta_0(\mathbf{b}_j), \tag{9}$$

$$v_{ji}|\pi_1,\tau_j \overset{ind.}{\sim} (1-\pi_1)\mathcal{N}^+(0,\tau_j^2)+\pi_1\delta_0(v_{ji}), \qquad i=1,\ldots,g, \tag{10}$$

$$\tau_j \overset{ind.}{\sim} \mathcal{G}\left(\lambda_0,\lambda_{1,j}\right), \qquad \lambda_0=\frac{1}{2}, \tag{11}$$

$$\pi_0\sim\mathcal{B}(c_0,d_0), \qquad \pi_1\sim\mathcal{B}(c_1,d_1), \qquad \sigma^2\sim\mathcal{G}^{-1}(a_0,a_1), \tag{12}$$

where $\mathcal{N}^+(0,\tau_j^2)$ denotes a $\mathcal{N}(0,\tau_j^2)$ distribution truncated below at zero, $\delta_0(\cdot)$ denotes a Dirac distribution at zero, $\mathcal{G}$ (resp. $\mathcal{G}^{-1}$) denotes the Gamma (resp. Inverse-Gamma) distribution, $\lambda_{1,j}$ in the prior for τ_j is the scale parameter, and $\mathcal{B}$ denotes a Beta distribution. The conditional priors for $\mathbf{b}_j|g,\pi_0$ and for $v_{ji}|\pi_1,\tau_j^2$ are specified as hard spike-and-slab priors. A similar prior has been considered in Xu and Ghosh (2015, Section 3.2). However, a subtle difference here stems from the non-conjugate Gamma prior for the standard deviation hyperparameter τ_j. We prove in Section 5 that this specification is crucial to ensure good asymptotic properties for the corresponding posterior. Compared with the priors in Chen et al., 2016, the proposed prior is computationally convenient as we can simulate from the posterior distribution by using a relatively simple Gibbs sampler with one Metropolis-Hastings step.

By integrating out $\{v_{ji}\}_{i=1}^g$, the induced prior on $\theta_{ji}|\pi_0,\pi_1,\tau_j$ is: $\forall j=1,\ldots,N$,

$$\begin{aligned}\theta_{ji}|\pi_0,\pi_1,\tau_j,g \sim\ & (1-\pi_1)(1-\pi_0)(1-\pi_1^g)f_{\theta_{ji}}(\theta_{ji}|\tau_j)\\ &+\ \delta_0(\theta_{ji})\left(\pi_1+[\pi_1^g(\pi_0-1)-\pi_0](\pi_1-1)\right), \qquad \forall i=1,\ldots,g,\end{aligned} \tag{13}$$

where the Lebesgue density $f_{\theta_{ji}}$ is upper bounded by the Lebesgue density of a $\mathcal{G}(1/2,\tau_j)$ distribution. Its explicit expression is given in the Online Appendix A.1, together with a proof of this result. In the appendix we also discuss the conditional prior on the random function φ_j, given $\{v_{ji}\}_{i=1}^g$ and g, induced by the prior (8)-(9).

4 The BSGS-SS predictive density and the MCMC algorithm

Let $\Pi(\cdot|y, \mathbf{X})$ denote the posterior of the model's parameters (φ, σ^2) obtained from the sampling model (2) and the BSGS-SS prior in Section 3. We construct density forecasts based on the posterior predictive density of $y_\tau|\mathbf{x}_{\tau-h}$, denoted by $\widehat{f}(y_\tau|\mathbf{x}_{\tau-h}, y, \mathbf{X})$ and conditional on a new value of the covariates $\mathbf{x}_{\tau-h}$, for $\tau > T$:

$$\widehat{f}(y_\tau|\mathbf{x}_{\tau-h}, y, \mathbf{X}) := \int f_0(y_\tau|\mathbf{x}_{\tau-h}, \varphi, \sigma^2)\Pi(d\varphi, d\sigma^2|y, \mathbf{X}) \tag{14}$$

where $f_0(y_\tau|\mathbf{x}_{\tau-h}, \varphi, \sigma^2)$ denotes the Lebesgue density associated with the sampling distribution in (2). Draws from the predictive distribution (14) can be obtained directly from the MCMC sampler detailed in Algorithm 1, where we denote by "rest" the conditioning parameters and data in each conditional distribution.

The hyperparameters of the priors on π_0 and π_1 must be chosen very carefully as they control the overall amount of between-group and within-group prior sparsity. In this paper, we propose to select c_0 and c_1 by using a data-driven approach based on the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), defined as:

$$\text{DIC}(\mathbf{c}|y, \mathbf{X}) := -4\mathbb{E}_{\boldsymbol{\theta},\sigma^2}\big[\log f(y|\boldsymbol{\theta}, \sigma^2, \mathbf{X})|y, \mathbf{X}, \mathbf{c}\big] + 2\log f(y|\widehat{\boldsymbol{\theta}}_\mathbf{c}, \widehat{\sigma}^2_\mathbf{c}, \mathbf{X}), \tag{15}$$

where $(\boldsymbol{\theta}'_\mathbf{c}, \sigma^2_\mathbf{c})'$ denotes the model parameters for a given set of hyperparameters $\mathbf{c} = (c_0, c_1)$, and $(\widehat{\boldsymbol{\theta}}_\mathbf{c}, \widehat{\sigma}^2_\mathbf{c})$ is a point estimate of $(\boldsymbol{\theta}'_\mathbf{c}, \sigma^2_\mathbf{c})'$.[2] This metrics of fit presents the advantage of being easily available from the MCMC output, with no additional costs in terms of computational burden.[3]

[2]The first part of (15) is the posterior mean deviance, which is here estimated by averaging the log-likelihood function, $\log f(y|\boldsymbol{\theta}, \sigma^2, \mathbf{X})$, over the posterior draws of $(\boldsymbol{\theta}'_\mathbf{c}, \sigma^2_\mathbf{c})'$. $(\widehat{\boldsymbol{\theta}}_\mathbf{c}, \widehat{\sigma}^2_\mathbf{c})$ is computed using the posterior median for $\boldsymbol{\theta}$ and the posterior mean for σ^2.

[3]We select the c_0 and c_1 that minimize the DIC by performing a random search on a fine 2-dimensional grid, with lower-bounds set as described in (A.1.3). In empirical applications, s_0^{gr} and s_0 are unknown, but we recommend to tune them at some arbitrary values such that the inequalities in (A.1.3) are satisfied for low values of c_0 and c_1. Therefore, the price to pay for not knowing s_0^{gr} and s_0 is simply the burden of searching on a larger grid.

Algorithm 1: MCMC sampling scheme

Define: $\eta_{ji}^2 := \left(b_{ji}^2 \mathbf{Z}_{ji}' \mathbf{Z}_{ji} \sigma^{-2} + \tau_j^{-2}\right)^{-1}$, $\nu_{ji} := \sigma^{-2} \eta_{ji}^2 b_{ji} \mathbf{Z}_{ji}' \left[y - \mathbf{Z}_{j\setminus i} \boldsymbol{\theta}_{j\setminus i}\right]$,
$\boldsymbol{\Sigma}_j := (\sigma^{-2} \mathbf{V}_j^{1/2} \mathbf{Z}_j' \mathbf{Z}_j \mathbf{V}_j^{1/2} + \mathbf{I}_{m_j})^{-1}$, $\boldsymbol{\mu}_j := \sigma^{-2} \boldsymbol{\Sigma}_j \mathbf{V}_j^{1/2} \mathbf{Z}_j' \left[y - \mathbf{Z}_{\setminus j} \boldsymbol{\theta}_{\setminus j}\right]$, where
$\mathbf{Z}_{j\setminus i} \boldsymbol{\theta}_{j\setminus i} = \mathbf{Z}_{\setminus j} \boldsymbol{\theta}_{\setminus j} + \sum_{k \neq i} b_{jk} v_{jk} \mathbf{Z}_{jk}$ and $\mathbf{Z}_{\setminus j} \boldsymbol{\theta}_{\setminus j} = \sum_{\kappa \neq j} b_{\kappa i} v_{\kappa i} \mathbf{Z}_{\kappa i}$.

Data: y, $\mathbf{Z}$.

Input: $\mathbf{b}^{(0)}$, $\sigma^{2(0)}$, $\{\mathbf{V}_j^{1/2,(0)}\}_{j=1,\ldots,N}$, $\pi_0^{(0)}$, $\pi_1^{(0)}$, $\boldsymbol{\tau}^{(0)}$,a_0, a_1, c_0, c_1, d_0, d_1.

for $\ell = 1, \ldots, MC$ **do**

(1) Sample $\sigma^{2(\ell)}|\text{rest} \sim \mathcal{G}^{-1}\left(\frac{T}{2} + a_0, \frac{1}{2}\|y - \mathbf{Z}\boldsymbol{\theta}^{(\ell-1)}\|_2^2 + a_1\right)$.

(2) $\forall j = 1, \ldots, N$ and $i = 1, \ldots, g$:

(2-1) Update $\eta_{ji}^{2(\ell)}$ and $\nu_{ji}^{(\ell)}$, and sample $\gamma_{1,ji}^{(\ell)}|rest \sim \text{Bernoulli}(\widetilde{\pi}_{1,ji}^{(\ell)})$, where

$$\widetilde{\pi}_{1,ji}^{(\ell)} = \frac{\pi_1^{(\ell-1)}}{\pi_1^{(\ell-1)} + 2(1 - \pi_1^{(\ell-1)}) \frac{\eta_{ji}^{(\ell)}}{\tau_j^{(\ell)}} \exp\left(\frac{\nu_{ji}^{2(\ell)}}{2\eta_{ji}^{2(\ell)}}\right) \Phi\left(\frac{\nu_{ji}^{(\ell)}}{\eta_{ji}^{(\ell)}}\right)}$$

and $\Phi(\cdot)$ denotes the cdf of the Normal distribution.

(2-2) Sample $v_{ji}^{(\ell)}|\text{rest} \sim (1 - \gamma_{1,ji}^{(\ell)}) \mathcal{N}^+\left(\nu_{ji}^{(\ell)}, \eta_{ji}^{2(\ell)}\right) + \gamma_{1,ji}^{(\ell)} \delta_0(v_{ji}^{(\ell)})$ and set $\mathbf{V}_j^{1/2,(\ell)} = \text{diag}(v_{j1}^{(\ell)}, \ldots, v_{jg}^{(\ell)})$.

(3) $\forall j = 1, \ldots, N$:

(3-1) Update $\boldsymbol{\mu}_j^{(\ell)}$, $\boldsymbol{\Sigma}_j^{(\ell)}$, and sample $\gamma_{0,j}^{(\ell)}|rest \sim \text{Bernoulli}(\widetilde{\pi}_{0,j}^{(\ell)})$, where

$$\widetilde{\pi}_{0,j}^{(\ell)} = \frac{\pi_0^{(\ell-1)}}{\pi_0^{(\ell-1)} + (1 - \pi_0^{(\ell-1)}) |\boldsymbol{\Sigma}_j^{(\ell)}|^{\frac{1}{2}} \exp\left(\frac{1}{2} \boldsymbol{\mu}_j^{(\ell)\prime} (\boldsymbol{\Sigma}_j^{(\ell)})^{-1} \boldsymbol{\mu}_j^{(\ell)}\right)}.$$

(3-2) Sample $\mathbf{b}_j^{(\ell)}|\text{rest} \sim (1 - \gamma_{0,j}^{(\ell)}) \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) + \gamma_{0,j}^{(\ell)} \delta_0(\mathbf{b}_j)$.

(3-3) Update $\boldsymbol{\theta}_j^{(\ell)} = \mathbf{V}_j^{1/2,(\ell)} \mathbf{b}_j^{(\ell)}$.

(4) $\forall j = 1, \ldots, N$, sample $\tau_j^{(\ell)}$ conditional on $\mathbf{V}_j, \lambda_{1,j}$ using a Metropolis-Hastings step with proposal density an exponential distribution. Draw $\tau_j^{new} \sim \mathcal{E}xp(\widetilde{\lambda})$, where $\widetilde{\lambda}^{-1} = \tau_j^{(\ell-1)}$. Set $\tau_j^{(\ell)} = \tau_j^{new}$ with probability $\min(1, R)$, where

$$R = \left(\frac{\tau_j^{(\ell-1)}}{\tau_j^{new}}\right)^{(\xi_j + \frac{3}{2})} \exp\left\{-\frac{1}{2} \sum_{i; v_{ji}^{(\ell)} > 0} v_{ji}^{2(\ell)} \left[\frac{1}{(\tau_j^{new})^2} - \frac{1}{\tau_j^{2(\ell-1)}}\right] - \frac{(\tau_j^{new} - \tau_j^{(\ell-1)})}{\lambda_{1,j}}\right\}$$
$$\times \exp\left\{-\frac{\tau_j^{(\ell-1)}}{\tau_j^{new}} + \frac{\tau_j^{new}}{\tau_j^{(\ell-1)}}\right\}$$

and $\xi_j := \sharp\{i; v_{ji}^{(\ell)} > 0\}$.

(5) Sample $\pi_0^{(\ell)}|\text{rest} \sim \mathcal{B}\left(N - \sum_{j=1}^N \gamma_{0,j}^{(\ell-1)} + c_0, \sum_{j=1}^N \gamma_{0,j}^{(\ell-1)} + d_0\right)$ and $\pi_1^{(\ell)}|\text{rest} \sim \mathcal{B}\left(Ng - \sum_{j=1}^N \sum_{i=1}^g \gamma_{1,ji}^{(\ell-1)} + c_1, \sum_{j=1}^N \sum_{i=1}^g \gamma_{1,ji}^{(\ell-1)} + d_1\right)$.

4.1 Extending the model to a general error process

One of the main advantage of the proposed prior is its flexibility on the treatment of the error process. Unlike other shrinkage priors (Group Lasso, Sparse Group Lasso), it is quite straightforward to extend the prior in Section 3 to account, for instance, for stochastic volatility and ARMA errors. The relevance of these features in macroeconomic applications has been extensively stressed by the literature (e.g., Clark, 2011; Chan and Hsiao, 2014; Carriero et al., 2019). Further, recent works (Carriero et al., 2024b; Lenza and Primiceri, 2022) have suggested that time-varying volatility could attenuate inferential issues (*e.g.*, explosive patterns and widening uncertainty for forecasts and Impulse Response Functions) when the estimation sample includes extreme observations such as those recorded during the COVID-19 pandemic.

We then modify the homoskedastic model in (1) to account for a general error process as follows:

$$y_t = \sum_{j=1}^{N} \varphi_j(x_{j,t-h,1}, x_{j,t-h,2}, \ldots) + \varepsilon_t \tag{16}$$

$$\phi_\varepsilon(L)\varepsilon_t = \psi_u(L)u_t, \qquad u_t \sim t_\nu\left(0, \omega_t \exp(\zeta_t)\right), \tag{17}$$

where we assume a stationary and invertible ARMA(p,q) specification for the errors ε_t and heavy-tails innovations u_t, the latter featuring time-varying log-volatility ζ_t and outliers component ω_t in the variance. Since the Student-t distribution t_ν in (17) can be expressed as a scale mixture of Normal distributions with scale term $\tau_t \sim \mathcal{G}^{-1}(0.5\nu, 0.5\nu)$, then, conditional on $(\tau_t, \omega_t, \zeta_t)$, the innovations have distribution $u_t \sim \mathcal{N}\left(0, \tau_t\omega_t \exp(\zeta_t)\right)$. The log-volatility ζ_t is assumed to evolve according to a stationary AR(1) process: $\zeta_t \sim \mathcal{N}\left(\mu_\zeta + \phi_\zeta(\zeta_{t-1} - \mu_\zeta), \sigma_\zeta^2\right)$, with initial conditions $\zeta_0 \sim \mathcal{N}\left(\mu_\zeta, \sigma_\zeta^2/(1-\phi_\zeta^2)\right)$ and ϕ_ζ restricted to lie in the stationary region. While the scale term τ_t is intended to capture frequent, although limited,

jumps in the volatility, the additional scale term ω_t is instead intended to capture infrequent, but extreme, jumps. Following Stock and Watson (2016) and Carriero et al. (2024b), the prior for ω_t has a two-part distribution, depending on whether an observation at period t is considered either regular or outlier: $\omega_t = 1$ with probability $(1 - p_\omega)$ and $\omega_t \sim \mathcal{U}(2, \bar{\omega})$ with probability p_ω for a constant $\bar{\omega} > 2$, where $\mathcal{U}(2, \bar{\omega})$ denotes the Uniform distribution with support $(2, \bar{\omega})$.

The implementation of these features implies only minor modifications to the MCMC Algorithm 1 (see the Online Appendix A.3 and Section 7 for an application). Conditional posteriors for stochastic volatility and ARMA parameters are the same as in Kastner and Fruhwirth-Schnatter (2014), Chan (2013), and Zhang et al. (2020), and we refer the reader to these contributions for more details.

5 Theoretical properties

This section provides the theoretical validation of our procedure. In Section 5.1 we establish consistency of the predictive density for density forecast. As a by-product, in Section 5.2 we demonstrate consistency of the posterior distribution of the model's parameters (φ, σ^2), denoted by $\Pi(\cdot|y, \mathbf{X})$. We adopt a frequentist point of view, in the sense that we admit the existence of a true value (φ_0, σ_0^2) that generates the data. We denote by $\mathbb{E}_0[\cdot]$ the expectation taken with respect to the true data distribution $P_0 = \prod_{t=1}^{T} \mathcal{N}\left(\sum_{j=1}^{N} \varphi_{0,j}(\mathbf{x}_{j,t-h}), \sigma_0^2\right)$ of y conditional on (φ_0, σ_0^2) and $\mathbf{X}$ with Lebesgue density denoted by f_0. All the analysis is conditional on $\mathbf{Z}$. In the following, we denote by $\|\mathbf{Z}\|_{op}$ the operator norm and $\|\mathbf{Z}\|_o := \max\{\|\mathbf{Z}_j\|_{op}; 1 \leq j \leq N\}$, 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. Moreover, $\|\cdot\|_2$ denotes the Euclidean norm. Denote

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

the rate of contraction of the posterior distribution of φ. The first term corresponds to the complexity of identifying s_0^{gr} non-zero groups, while the third term corresponds to the complexity of estimating s_0 non-zero elements of a parameter distributed in s_0^{gr} known groups. In absence of group-structure, the maximum between these two rates corresponds to the rate for recovery of sparse vectors over ℓ_0-balls (see *e.g.* Raskutti et al., 2011). This case can be obtained by setting either only one group (*i.e.* $N = 1$ and $g = p$) or a number of groups equal to the number of p parameters (*i.e.* $N = p$ and $g = 1$).

For two sequences a_T and b_T and two constants $c_1, c_2 > 0$ we write $a_T \asymp b_T$ if $c_1 b_T \leq a_T \leq c_2 b_T$. By using this notation, if $\log(s_0^{gr} g) \asymp \log(s_0^{gr} g/s_0)$, $\log(N) \asymp \log(N/s_0^{gr})$, and $s_0 \log(T) \leq \min\{s_0^{gr} \log(N), s_0 \log(s_0^{gr} g)\}$, then ϵ corresponds to the minimax rate for recovering φ given in Cai et al. (2022) and in Li et al. (2024). The conditions $\log(s_0^{gr} g) \asymp \log(s_0^{gr} g/s_0)$ and $\log(N) \asymp \log(N/s_0^{gr})$ guarantee a sparse setting, while the condition $s_0 \log(T) \leq \min\{s_0^{gr} \log(N), s_0 \log(s_0^{gr} g)\}$ guarantees a high-dimensional setting. If $s_0 = s_0^{gr} g$, which corresponds to the case with no sparsity within the groups, then the last condition corresponds to Assumption 4.1 *(ii)* in Mogliani and Simoni (2021).

The following assumption restricts some of the parameters of the model.

Assumption 5.1 *(i) For positive and bounded constants* $\underline{\sigma}^2$, $\overline{\sigma}^2$, $0 < \underline{\sigma}^2 \leq \sigma_0^2 \leq \overline{\sigma}^2 < \infty$*; (ii)* $\max\{N, T\} \leq e^{s_0^{gr} g}$*; (iii)* $\max_{j \in S_0^{gr}} \max_{i \in S_{0,j}} |\theta_{0,ji}| \leq \log(s_0^{gr} g)$.

Assumption 5.1 *(i)* excludes degenerate cases by restricting the model variance. Assumption 5.1 *(ii)* restricts the rate at which the number of groups can increase compared to T, s_0^{gr} and g. It is violated for instance if $N > \max\{T, \exp\{s_0^{gr} g\}\}$. Assumption 5.1 *(iii)* restricts the growth rate of the largest component of $\boldsymbol{\theta}_0$.

The next two assumptions concern the hyperparameters of the prior. Assumption 5.2 allows $\lambda_{1,j}$ (the hyperparameter of the prior for τ_j) to increase or decrease with T, g, s_0^{gr} and $|S_{0,j}|$. It rules out a $\lambda_{1,j}$ decreasing too fast to zero or increasing

too fast to infinity. In practice, any positive constant can be taken as a value for $\lambda_{1,j}$ and its choice depends on the desired tightness of the prior for τ_j.

Assumption 5.2 *Assume that* $\sqrt{T}/(\|\mathbf{Z}\|_o \min\{\log(gs_0^{gr}), \log(T)\}) < C$ *with probability* 1 *for some constant* $C > 0$. *The scale parameters* $\lambda_{1,j}$ *are allowed to change with* T *and belong to the range: for every* $j = 1, \ldots, N$,

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

for two positive constants $1 < \underline{c} < \overline{C} < \infty$, *where* $\lambda_{\max} := \max\{\lambda_{1,j};\, j \leq N\}$.

For the following assumption we introduce the function $(u)_+ := \max\{u, 0\}$.

Assumption 5.3 *There exist positive constants* κ_0 *and* κ_1 *such that the hyperparameters* c_0, d_0, c_1, d_1 *of the Beta priors for* π_0 *and* π_1 *satisfy: (i) for every* $N > 1$, $(d_0+j-1)/(c_0+N-j) \leq \kappa_0 j/[N^{u_0}(N-j+1)]$ *for every* $u_0 \in (\log(2)/\log(N), \leq s_0^{gr}]$ *and* $\forall j \in \{1, \ldots, N\} \subseteq \mathbb{N}$, *(ii) for every* $s_0^{gr} g > 1$, $(d_1 + j - 1)/(c_1 + Ng - j) \leq \kappa_1 j/[(s_0^{gr} g)^{u_1}(Ng-j+1)]$ *for every* $u_1 \in (\log(2)/\log(s_0^{gr} g), s_0]$ *and* $\forall j \in \{1, \ldots, g\} \subseteq \mathbb{N}$, *(iii) for a positive constant* C_{cd},

$$\max\{c_1, d_0 \log(N + c_0 + d_0), d_1 \log(s_0^{gr} g)\} \leq C_{cd} T \epsilon^2;$$

(iv) $c_1 \geq \lambda_{1,j}^2 g \frac{3d_1}{4} - d_1$.

Assumption 5.3 *(i)* and *(ii)* require that c_0 and c_1 increase together with N, s_0^{gr} and g and control their rate. To satisfy the assumption, if d_0 and d_1 are chosen equal to one and if κ_0 and κ_1 are chosen such that $\kappa_0 < N^{u_0}$ and $\kappa_1 < (s_0^{gr} g)^{u_1}$, then c_0 and c_1 have to be at least of the order of $1 - N + N^{u_0+1}/\kappa_0$ and $(s_0^{gr} g)^{u_1} Ng/\kappa_1 - Ng + 1$, respectively, for u_0, u_1 in the range of values given in the assumption and up to a constant. On the other hand, if d_0 and d_1 are chosen to increase with T, then c_0 and c_1 have to increase faster than $d_0 N^{u_0+1}$ and $d_1 (s_0^{gr} g)^{u_1}$. In practice, one can set

the constants κ_0 and κ_1 at very small values, as long as they are fixed and do not increase with T. Assumptions 5.3 *(i)* correspond to those in Castillo et al. (2015, Assumption 1) for the special case of a Beta prior. We provide in Appendix A.2 a deeper analysis of this assumption.

5.1 Consistency of the predictive density

We now provide a validation of our procedure for density forecasts. The next theorem states that, for every $\tau > T$, $\widehat{f}(y_\tau|\mathbf{x}_{\tau-h}, y, \mathbf{X})$ in (14) converges to the Lebesgue density $f_0(y_\tau|\mathbf{x}_{\tau-h}, \varphi_0, \sigma_0^2)$ of the true distribution $\mathcal{N}\left(\sum_{j=1}^N \varphi_{0,j}(\mathbf{x}_{j,\tau-h}), \sigma_0^2\right)$ with respect to the Hellinger distance denoted by $d_H(\cdot,\cdot)$.

Theorem 5.1 *For every $\tau > T$, let $f_0(y_\tau|\mathbf{x}_{\tau-h}, \varphi_0, \sigma_0^2)$ denote the Lebesgue density function of the true conditional distribution P_0 of y_τ given $\mathbf{x}_{\tau-h}$ evaluated at y_τ. Suppose Assumptions 2.1 and 5.1 - 5.3 hold and let $\epsilon \to 0$ as $T \to \infty$. Suppose that $|B_{0,\tau-h}(g)|^2 \lesssim \epsilon^2$, and that there exist positive constants κ_ℓ and κ_z such that for all $j \geq 1$ and some $\beta > 2$,*

$$\mathbb{E}_0\left[\Pi\left(e^{4jM_2\epsilon^2} < \frac{\sigma^2+\sigma_0^2}{2\min\{\sigma_0^2,\sigma^2\}} \leq e^{4(j+1)M_2\epsilon^2}\middle| y, \mathbf{X}\right)\right] \leq j^{-\beta},$$
$$\mathbb{E}_0\left[\Pi\left(\frac{M_3\epsilon^2}{\kappa_\ell\kappa_z}j < \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2 \leq \frac{M_3\epsilon^2}{\kappa_\ell\kappa_z}(j+1)\middle| y, \mathbf{X}\right)\right] \leq j^{-\beta}. \quad (18)$$

Then,

$$\lim_{T\to\infty} P_0\left(d_H^2(f_0(y_\tau|\mathbf{x}_{\tau-h}, \varphi_0, \sigma_0^2), \widehat{f}(y_\tau|\mathbf{x}_{\tau-h}, y, \mathbf{X})) \leq \varepsilon \middle| \mathbf{X}, \mathbf{x}_{\tau-h}\right) = 1$$

for all $\varepsilon > 0$ and all $\mathbf{x}_{\tau-h}$ such that $\sum_{j\in S^{gr}}\sum_{i\in S_j} z_{j,\tau-h,i}^2 < C$ for some constant $C > 0$, $\forall S^{gr} \subseteq \{1,\ldots,N\}$ such that $|S^{gr}| \leq s_0^{gr} + T\epsilon^2/\log(N)$, and $\forall S_j \subseteq \{1,\ldots,g\}$ such that $\sum_{j\in S_0^{gr}} |S_j| \leq s_0 + T\epsilon^2/\log(s_0^{gr} g)$.

We know from (A.6.18) and Theorem A.5.1 in the Online Appendix that the probabilities in (18) go to zero. The assumption of the theorem is then just an assumption

about the rate at which these posterior probabilities go to zero. We refer to Hoffmann et al. (2015) for primitive conditions for this assumption.

5.2 Consistency of the posterior distribution

We establish here the posterior consistency uniformly over the infinite dimensional set $\mathcal{F}(s_0, s_0^{gr}; \mathbf{Z})$ defined as:

$$\mathcal{F}(s_0, s_0^{gr}; \mathbf{Z}) := \Big\{(\varphi, \sigma^2); \|B(g)\|_2^2 \leq \frac{s_0 \sigma^2}{16}, s_{\boldsymbol{\theta}}^{gr} \leq s_0^{gr},\ s_{\boldsymbol{\theta}} \leq s_0, \\ \max_{j \in S_{\boldsymbol{\theta}}^{gr}} \max_{i \in S_{\boldsymbol{\theta},j}} |\theta_{ji}| \leq \log(s_0^{gr} g),\ \text{and } \sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2]\Big\}$$

for $s_0^{gr}, s_0 \in \mathbb{N}_+$ satisfying $s_0^{gr} \leq N$ and $s_0^{gr} \leq s_0 \leq g\, s_0^{gr}$. The next theorem establishes the in-sample consistency of our BSGS-SS procedure. It shows that the posterior contracts in a neighborhood of the true value defined by the in-sample prediction error. We use the notation $\varphi_j^{(T)}(\mathbf{X}) := (\varphi_j(\mathbf{x}_{j,-h+1}), \ldots, \varphi_j(\mathbf{x}_{j,T-h}))'$ for $j \in \{1, \ldots, N\}$ and $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_N$.

Theorem 5.2 *Suppose Assumptions 2.1 and 5.1 - 5.3 hold and let $\epsilon \to 0$ as $T \to \infty$. Then, for a sufficiently large constant $M > 0$: as $T \to \infty$,*

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

A similar result is obtained for the case with lagged values of the dependent variable among the covariates (see the Online Appendix A.7). The required sample size to achieve convergence to zero of the in-sample prediction error in Theorem 5.2 is $T > C \max\{s_0^{gr} \log(N), s_0 \log(s_0^{gr} g), s_0 \log(T)\}$ for some positive constant C. As stressed at the beginning of Section 5, if $\log(s_0^{gr} g) \asymp \log(s_0^{gr} g / s_0)$, $\log(N) \asymp \log(N / s_0^{gr})$, and $s_0 \log(T) \leq \min\{s_0^{gr} \log(N), s_0 \log(s_0^{gr} g)\}$, the contraction rate ϵ corresponds to the minimax rate for recovering φ.

In the grouped predictors example of Section 2.3.1, the norm in Theorem 5.2 is

the inner product weighted by the Gram matrix $\mathbf{X}'\mathbf{X}$, that is $\|\mathbf{X}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2$. In the mixed-frequency example of Section 2.3.2, the norm is:

$$\left\|\sum_{j=1}^{N}\left(\varphi_j^{(T)}(\mathbf{X})-\varphi_{0,j}^{(T)}(\mathbf{X})\right)\right\|_2^2=\sum_{t=1}^{T}\left(\sum_{j=1}^{N}\sum_{i=1}^{\infty}\left(\theta_{j,i}-\theta_{0,ji}\right)\Phi_i'\mathbf{x}_{j,t-h}\right)^2.$$

In addition to the entire function φ, it is worth focusing on the coefficients $\boldsymbol{\theta}$. The consistency of the marginal posterior of $\boldsymbol{\theta}$ is established in Theorem A.5.1 in the Online Appendix. The contraction rate for $\boldsymbol{\theta}$ coincides with the rate obtained in Li et al. (2024, Corollary 1) for a linear regression model. Finally, Lemma A.5.1 in the Online Appendix establishes that the posterior of the model parameter concentrates on bi-level vectors with a number of active groups and active components not smaller than the true one.

We now look at the posterior mean of $\varphi(\mathbf{x}_t)$ as a possible point estimator for $\varphi_0(\mathbf{x}_t)$. Convergence of a Bayesian point estimator towards the true φ_0 is in general not implied by the result of Theorem 5.2 if the loss function is not bounded. The next theorem analyzes the asymptotic behaviour of the posterior mean in terms of the Euclidean loss function $\ell_h(\widetilde{\varphi},\varphi):=\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{N}\left[\widetilde{\varphi}_j(\mathbf{x}_{t-h})-\varphi_j(\mathbf{x}_{t-h})\right]\right)^2$, which is not uniformly bounded if the parameter space is not compact.

Theorem 5.3 *Let us consider the posterior mean estimator* $\widehat{\varphi}(\mathbf{x}_t):=\mathbb{E}[\varphi(\mathbf{x}_t)|y,\mathbf{X}]$, *for* $t\leq T$ *and assume that as* $T\to\infty$, $\mathbb{E}_0\left[\Pi\left(\ell_h(\varphi,\varphi_0)>Mj\epsilon^2|\,y,\mathbf{X}\right)\right]\leq Cj^{-\beta}$ *for every* $j\geq 1$ *and for some* $\beta>2$ *and some constant* $C>0$. *Then under the assumptions of Theorem 5.2,*

$$\mathbb{E}_0\left[\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{N}\left[\widehat{\varphi}_j(\mathbf{x}_{t-h})-\varphi_{0,j}(\mathbf{x}_{t-h})\right]\right)^2\right]\to 0\qquad \text{as } T\to\infty.$$

The result of the theorem holds under the assumption that the posterior of the complement of a ball centered on φ_0, with radius proportional to the contraction

rate ϵ^2, becomes sufficiently small for large T. Results of this type have been established for instance in Hoffmann et al. (2015).

6 Monte Carlo experiments

We evaluate the finite sample performance of our procedure through two Monte Carlo experiments. First, we consider a DGP featuring grouped predictors, as in the example described in Section 2.3.1. Second, we consider a DGP featuring mixed-frequency data, as in the example described in Section 2.3.2. Simulations are based on 200 Monte Carlo iterations, each featuring 60'000 MCMC sweeps (with a burn-in sample of 10'000 sweeps and a chain-thinning parameter set at 5).

6.1 Experiment 1: DGP with grouped predictors

In this experiment, the DGP involves $Ng = \{100, 300\}$ predictors (sampled at the same frequency as the target variable) structured in $N = \{5, 10, 20\}$ groups. The group active set has cardinality $s_0^{gr} = \{1, 5, 10\}$ and we fix $|S_{0,j}| = 1$ for all simulations, such that $s_0 = s_0^{gr}$. The elements of S_0^{gr} and $S_{0,j}$ are fixed realizations of random draws without replacement from $\{1, \dots, N\}$ and $\{1, \dots, g\}$, respectively. The number of in-sample observations is fixed at $T = 200$, and the out-of-sample at $T_{oos} = 50$. Simulated data are obtained from the following process:

$$y_t = \alpha + \beta_y y_{t-1} + \sum_{j=1}^{N} \sum_{i=1}^{g} z_{j,t,i} \theta_{ji} + \varepsilon_t, \tag{19}$$

$$z_{j,t,i} = \rho_z z_{j,t-1,i} + \epsilon_{j,t,i}, \tag{20}$$

$$\begin{pmatrix} \varepsilon_t \\ \boldsymbol{\epsilon}_t \end{pmatrix} \sim \text{i.i.d. } \mathcal{N}\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\epsilon \end{pmatrix} \right], \tag{21}$$

where $\boldsymbol{\epsilon}_t := (\epsilon_{1,t,1}, \dots, \epsilon_{N,t,g})'$, $\boldsymbol{\Sigma}_\epsilon = \mathcal{S}_\epsilon \mathcal{R}_\epsilon \mathcal{S}_\epsilon$ is a block-diagonal matrix, $\mathcal{S}_\epsilon$ a $(Ng \times Ng)$ diagonal matrix with elements σ_ϵ, and $\mathcal{R}_\epsilon$ 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 $\rho_\epsilon^{|i-i'|}$ for all $i \neq i'$. The group structure is therefore given by the structure of this covariance matrix. We set $\sigma = 0.50$ and $\rho_\epsilon = 0.50$, and we calibrate σ_ϵ such that the noise-to-signal ratio of (19) is NSR $= 0.20$ for all the simulations. The coefficients of the active variables is set to $|\theta_{ji}| = 0.5$, for each $j \in S_0^{gr}$ and $i \in S_{0,j}$. The sign of the coefficients is a fixed realization of random draws with replacement from $\{-1, 1\}$. Finally, we set $\alpha = 0.2$, $\beta_y = 0.3$, and $\rho_z = 0.9$.

Simulation results are reported in Table 1, where we focus on both selection and density forecast performance. The former is evaluated by computing the Matthews correlation coefficient at the group level (MCC$_N$), which measures the overall quality of the groups classification, while the latter is evaluated by the means of the Continuously Ranked Probability Score (CRPS), averaged over the out-of-sample observations and expressed in relative terms with respect to an AR(1) benchmark. The results for our prior (BSGS-SS) are then compared to those obtained from the Bayesian Sparse Group Lasso (BSGL) prior of Xu and Ghosh (2015) and the Bayesian Adaptive Group Lasso with Spike-and-Slab prior (BAGL-SS) prior of Mogliani and Simoni (2021). The values reported in the table denote average outcomes across Monte Carlo iterations and their bootstrap standard errors (in parentheses).

The results for our BSGS-SS prior suggest that true group active set can be correctly selected with extremely high probability (MCC$_N$ close to one) and point to accurate density forecasts (CRPS less than one). This holds irrespective of the number of groups N and for DGPs involving a very sparse environment (*i.e.* when s_0^{gr} and s_0 are small compared to N and g). However, consistently with the contraction rates derived in Section 5, selection and predictive accuracy tend to deteriorate in less sparse DGPs. Compared to the alternative priors considered, the results also point to a substantial outperformance of our BSGS-SS prior, in particular for highly

Table 1: Monte Carlo Experiment 1 - selection and density forecast accuracy

				BSGS-SS		BSGL		BAGL-SS	
Ng	N	g	s_0^{gr}	MCC_N	CRPS	MCC_N	CRPS	MCC_N	CRPS
100	5	20	1	0.98 (0.01)	0.70 (0.01)	0.82 (0.01)	0.75 (0.01)	0.86 (0.02)	0.79 (0.01)
	10	10	1	0.99 (0.01)	0.69 (0.01)	0.88 (0.01)	0.72 (0.01)	0.84 (0.02)	0.75 (0.01)
			5	0.98 (0.01)	0.73 (0.01)	0.52 (0.01)	0.93 (0.01)	0.77 (0.01)	1.05 (0.01)
	20	5	5	1.00 (0.01)	0.72 (0.01)	0.68 (0.02)	0.87 (0.01)	0.87 (0.01)	0.90 (0.02)
			10	0.56 (0.01)	0.93 (0.02)	0.23 (0.01)	1.00 (0.01)	0.58 (0.01)	1.14 (0.02)
300	5	60	1	1.00 (0.01)	0.71 (0.01)	0.70 (0.01)	0.88 (0.01)	0.84 (0.02)	1.03 (0.02)
	10	30	1	0.99 (0.01)	0.70 (0.01)	0.76 (0.01)	0.81 (0.01)	0.91 (0.01)	0.85 (0.01)
			5	0.97 (0.01)	0.73 (0.01)	0.37 (0.01)	1.12 (0.02)	0.54 (0.02)	1.31 (0.03)
	20	15	5	0.96 (0.01)	0.74 (0.01)	0.49 (0.01)	1.04 (0.01)	0.71 (0.01)	1.15 (0.02)
			10	0.33 (0.01)	1.00 (0.01)	0.22 (0.01)	1.10 (0.01)	0.38 (0.01)	1.32 (0.03)

Notes: $T = 200$, $s_{0,j} = 1$, $s_0 = s_0^{gr}$. MCC_N and CRPS denote respectively the Matthews Correlation Coefficient at the group level and the Continuously Ranked Probability Score, the latter expressed in relative terms with respect to the AR(1) benchmark. Average values over 200 Monte Carlo simulations. Bootstrap standard errors of Monte Carlo averages in parentheses.

sparse DGPs. Focusing on density forecasts, three main findings can be remarked. First, the predictive accuracy of our model is almost unaffected by the increase in the total number of predictors, while both the BSGL and the BAGL-SS priors display a fairly strong deterioration of their forecasting performance. Second, as mentioned above, our model fails to significantly outperform the AR(1) benchmark in a less sparse environment, but it still significantly outperforms the competing models. Third, the predictive performance of single-level group sparsity priors, such as the BAGL-SS, may quickly deteriorate in presence of complex sparsity structures (such as bi-level sparsity).

Additional simulation results and robustness checks are reported in the Online Appendix A.4. In particular, we show that in a very sparse environment, the mean squared error (MSE) is extremely low (with almost zero bias) and the true active set (both at group and variables level) is correctly selected with extremely high

probability (according to both the True Positive Rate and the Matthews correlation coefficient). Further, we show that the performance of our model is robust to different correlation structures in $\mathbf{\Sigma}_\epsilon$ and alternative error distributions (*e.g.* Skew-Normal distribution). Overall, the results are very close to those reported in Table 1, irrespective of the values for Ng, N, and s_0^{gr}, although they tend to deteriorate with a substantial increase in the noise-to-signal ratio.[4]

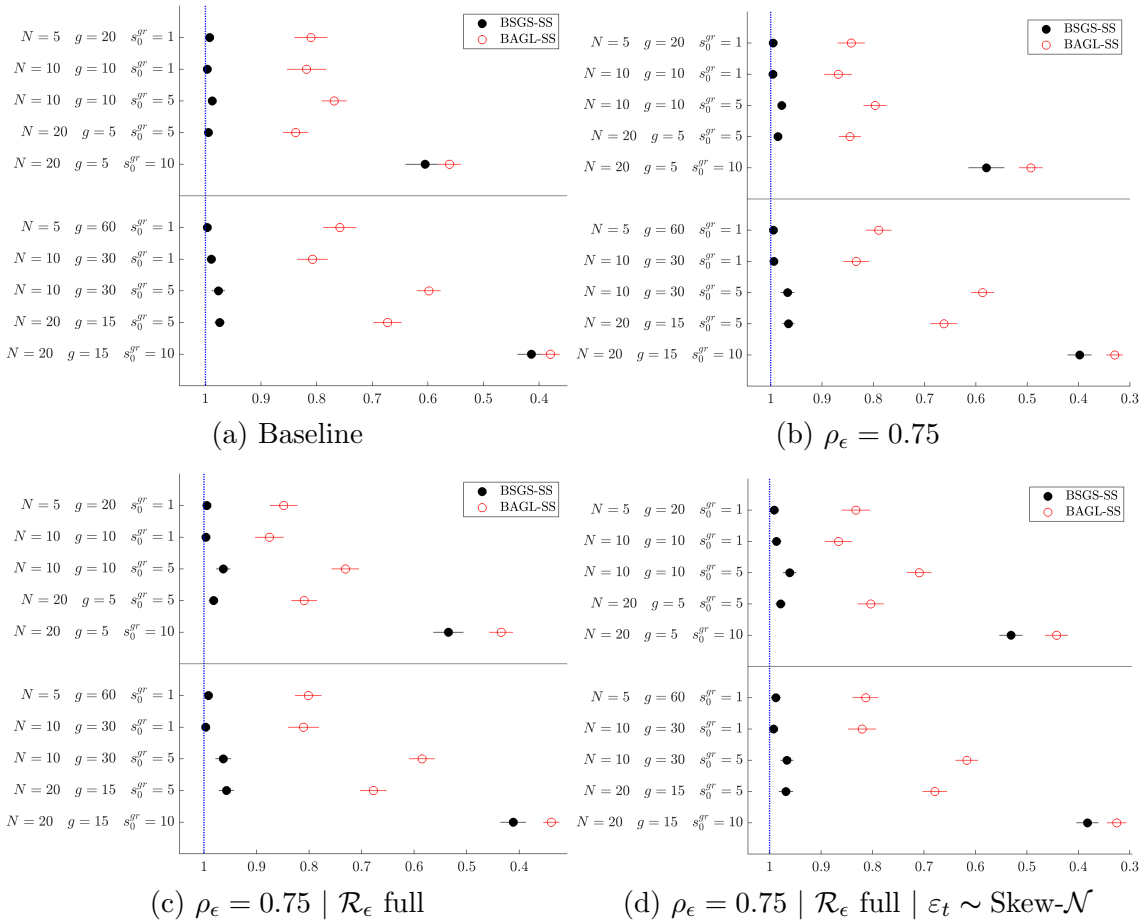
As a final exercise, we compare the variable selection accuracy of the BSGS-SS prior and the BAGL-SS prior of Mogliani and Simoni (2021). The latter relies on a single layer of sparsity (that is, sparsity at the group level only), and for this reason it is expected to be an appropriate prior only when dealing with specifications that are dense within each active group. Instead, when the true model is bi-level sparse, assuming a single-level group sparsity as in the BAGL-SS-prior, instead of bi-level group sparsity, does not ensure the posterior recovery of the correct sparsity pattern. This entails inaccurate selection and out-of-sample prediction. We investigate this issue numerically and the results are reported in Figure 1, which shows the MCC computed at the variable level (MCC_g) and across various DGPs. The simulation results confirm the theoretical results and strongly suggest a substantially lower selection accuracy for the BAGL-SS ($\text{MCC}_g < 1$) compared to our BSGS-SS prior (MCC_g close to one) when the true model is in fact bi-level sparse.

6.2 Experiment 2: DGP with mixed-frequency data

In this example, we consider a DGP with mixed-frequency data, where $N = \{50, 100\}$ predictors are sampled at a higher frequency compared to the target

[4]We perform the following exercises: *i)* we increase the within-groups correlation parameter by setting $\rho_\epsilon = 0.75$; *ii)* we allow for between-groups correlation by letting the Toeplitz correlation matrix $\mathcal{R}_\epsilon$ to be a full matrix, *i.e.* $\mathcal{R}_\epsilon = \rho_\epsilon^{|\iota - \iota'|}$ for all $\iota \neq \iota'$, with $\iota = 1, \dots, Ng$; *iii)* we allow for asymmetric shocks in the DGP by letting ε_t to follow a Skew-$\mathcal{N}(\xi, \omega^2, \alpha)$ distribution (Azzalini and Capitanio, 2014), with shape parameter $\alpha = -5$, and location and scale parameters (ξ, ω) calibrated such that ε_t has mean zero and variance σ^2; *iv)* we increase the noise-to-signal ratio to 0.5.

Figure 1: MCC_g for the BSGS-SS and the BAGL-SS priors across DGPs



(a) Baseline (b) $\rho_\epsilon = 0.75$

(c) $\rho_\epsilon = 0.75 \mid \mathcal{R}_\epsilon$ full (d) $\rho_\epsilon = 0.75 \mid \mathcal{R}_\epsilon$ full $\mid \varepsilon_t \sim$ Skew-$\mathcal{N}$

Notes: The markers indicate the average Matthews correlation coefficient at the variable level (MCC_g) across all Monte Carlo iterations. Values close to one indicate a better overall quality of the identification of the true sparsity structure. The error bars denote ± 2 bootstrap standard errors of average values obtained from 200 Monte Carlo simulations.

variable. Simulated data are obtained from the following process:

$$y_t = \alpha + \beta_y y_{t-1} + \sum_{j=1}^{N} \sum_{u=0}^{p_x} \psi_j(u) L^{u/m} x_{j,t-h}^{(m)} + \varepsilon_t, \tag{22}$$

$$x_{j,t}^{(m)} = \rho_x x_{j,t-1/m}^{(m)} + \epsilon_{j,t},$$

where $m = 3$ (which is akin to a quarterly-monthly process) and $p_x = 11$. We set the true weighting scheme $\psi_j(u)$ by relying on a three-parameters Beta function:

$$\psi_j(u) = \left(\frac{u+1}{p_x+1}\right)^{a-1} \left(1 - \frac{u+1}{p_x+1}\right)^{b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} + c.$$

We investigate four alternative weighting schemes that correspond to bell-shaped weights (DGP 1) with $(a, b, c) = (5, 15, 0)$, fast-decaying weights (DGP 2) with $(a, b, c) = (1, 10, 0)$, slow-decaying weights (DGP 3) with $(a, b, c) = (1, 4, 0)$, and flat weights (DGP 4) with $(a, b, c) = (1, 1, 0)$.[5] Note that the same weighting scheme applies to all the predictors entering the DGPs.

We define ε_t and $\boldsymbol{\epsilon}_t := (\epsilon_{1,t}, \ldots, \epsilon_{N,t})'$ as i.i.d. draws from the normal multivariate distribution in (21), where $\boldsymbol{\Sigma}_\epsilon = \mathcal{S}_\epsilon \mathcal{R}_\epsilon \mathcal{S}_\epsilon$, with $\mathcal{S}_\epsilon$ a diagonal matrix with elements σ_ϵ and $\mathcal{R}_\epsilon$ a Toeplitz correlation matrix with diagonal elements equal to one and off-diagonal elements $\rho_\epsilon^{|j-j'|}$ for all $j \neq j'$. We set $\rho_\epsilon = 0.50$, $\alpha = 0.50$, $\sigma = 0.50$, $\rho_x = 0.90$. The active set has cardinality $s_0^{gr} = \{5, 10\}$ and the indices of active variables in S_0^{gr} are fixed realizations of random draws without replacement from $\{1, \ldots, N\}$. Conditional on these parameters, we calibrate σ_ϵ^2 such that the noise-to-signal ratio of the mixed-frequency regression is NSR $= 0.20$ for all the simulations. We assume $h = 0$ (akin to a nowcasting model). The number of in- and out-of-sample observations is fixed at $T = 200$ and $T_{oos} = 50$, respectively.

In the simulations, we consider and evaluate a set of alternative approximating functions for the true weighting function $\psi_j(u)$: the U-MIDAS (Foroni et al., 2015), Almon lag polynomials (restricted and unrestricted; Mogliani and Simoni, 2021), and orthogonal lag polynomials (Legendre, Chebychev (first kind), and Bernstein orthogonal polynomials).

Selection and predictive accuracy performance of our BSGS-SS procedure are reported in Table 2, where we report the True Positive Rate at the group level (TPR_N) and the CRPS. The results point to a number of interesting features. First, among the lag polynomials considered, the best results are provided by the unrestricted, the restricted Almon and the Bernstein polynomials. Second, consistently with the theory, the results for the best-performing polynomials are overall unaltered by the increase in the total number of high-frequency predictors. However, the

[5] The weights are set to exactly zero for values of the Beta function $<$ 1e-04, and normalized to sum up to 1.

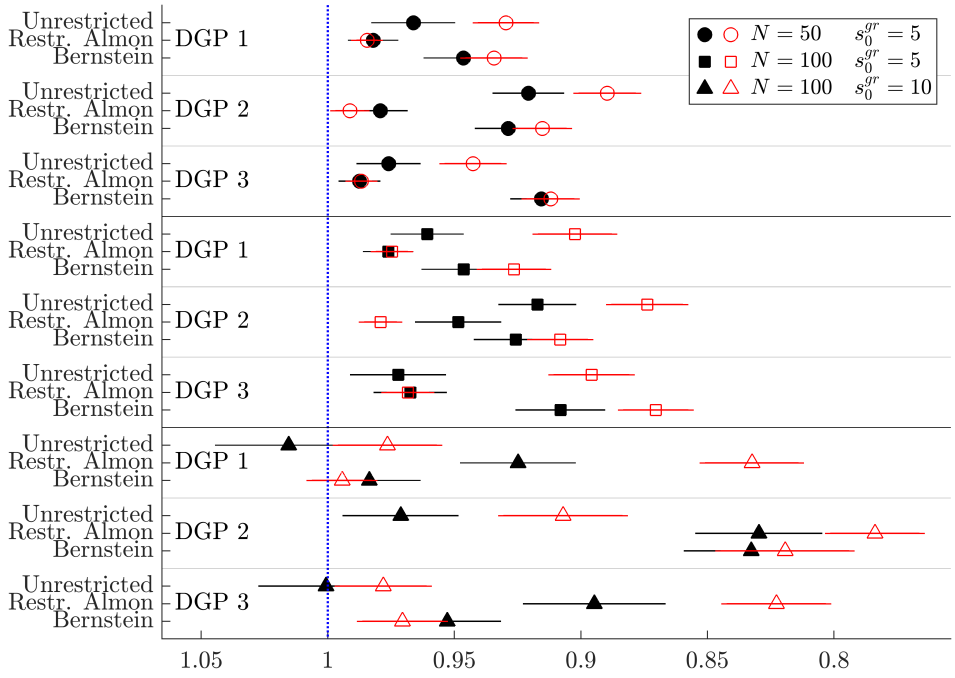
Table 2: Monte Carlo Experiment 2 - selection and density forecast accuracy

			DGP 1 bell-shaped		DGP 2 fast-decaying		DGP 3 slow-decaying		DGP 4 flat	
N	s_0^{gr}	Polynomial	TPR_N	CRPS	TPR_N	CRPS	TPR_N	CRPS	TPR_N	CRPS
50	5	Unrestricted	99.8	0.71	100.0	0.67	99.2	0.75	70.5	0.88
			(0.2)	(0.01)	(0.0)	(0.01)	(0.5)	(0.01)	(3.3)	(0.01)
		Almon	84.3	0.75	81.8	0.75	85.0	0.77	94.8	0.77
			(0.7)	(0.01)	(1.2)	(0.01)	(0.8)	(0.01)	(0.8)	(0.01)
		Restr. Almon	99.5	0.70	87.5	0.71	98.8	0.71	100.0	0.79
			(0.3)	(0.01)	(0.7)	(0.01)	(0.4)	(0.01)	(0.0)	(0.01)
		Legendre	86.7	0.77	86.2	0.77	84.8	0.78	98.3	0.75
			(0.7)	(0.01)	(0.9)	(0.01)	(1.1)	(0.01)	(0.5)	(0.01)
		Bernstein	99.3	0.71	97.2	0.67	100.0	0.70	90.8	0.84
			(0.3)	(0.01)	(0.6)	(0.01)	(0.0)	(0.01)	(2.2)	(0.01)
		Chebychev T	87.5	0.76	81.7	0.76	84.0	0.78	98.7	0.75
			(0.7)	(0.01)	(1.2)	(0.01)	(1.3)	(0.01)	(0.5)	(0.01)
100	5	Unrestricted	99.3	0.72	99.2	0.69	98.2	0.74	42.2	0.94
			(0.5)	(0.01)	(0.5)	(0.01)	(0.9)	(0.01)	(2.8)	(0.01)
		Almon	83.3	0.77	78.8	0.79	81.8	0.79	91.3	0.80
			(0.3)	(0.01)	(1.5)	(0.01)	(1.1)	(0.01)	(1.1)	(0.01)
		Restr. Almon	97.3	0.71	83.5	0.74	94.8	0.71	93.7	0.83
			(0.6)	(0.01)	(0.2)	(0.01)	(0.8)	(0.01)	(1.4)	(0.01)
		Legendre	80.2	0.79	82.7	0.81	78.3	0.81	91.0	0.78
			(1.2)	(0.01)	(1.5)	(0.01)	(1.7)	(0.01)	(2.1)	(0.01)
		Bernstein	98.7	0.72	94.5	0.69	99.3	0.70	66.8	0.89
			(0.6)	(0.01)	(0.8)	(0.01)	(0.6)	(0.01)	(3.3)	(0.01)
		Chebychev T	80.3	0.79	81.7	0.79	78.7	0.81	91.7	0.78
			(1.3)	(0.01)	(1.0)	(0.01)	(1.6)	(0.01)	(1.8)	(0.01)
100	10	Unrestricted	23.3	0.96	27.9	0.93	16.1	0.98	10.1	0.99
			(2.1)	(0.01)	(2.7)	(0.01)	(1.3)	(0.01)	(0.3)	(0.00)
		Almon	20.8	0.98	10.6	0.99	21.7	0.98	25.5	0.96
			(2.6)	(0.01)	(0.7)	(0.00)	(2.6)	(0.01)	(3.1)	(0.01)
		Restr. Almon	88.2	0.77	89.3	0.76	87.4	0.78	17.7	0.97
			(2.0)	(0.01)	(1.1)	(0.01)	(2.3)	(0.01)	(1.0)	(0.00)
		Legendre	16.5	0.98	14.2	0.98	18.6	0.97	16.2	0.98
			(1.7)	(0.01)	(0.6)	(0.00)	(2.1)	(0.01)	(2.2)	(0.01)
		Bernstein	17.1	0.98	62.0	0.80	20.4	0.95	10.5	0.99
			(0.9)	(0.01)	(3.6)	(0.02)	(2.1)	(0.01)	(0.4)	(0.00)
		Chebychev T	19.1	0.97	10.8	0.99	19.1	0.98	13.1	0.98
			(2.0)	(0.01)	(0.4)	(0.00)	(2.3)	(0.00)	(1.5)	(0.00)

Notes: TPR_N and CRPS denote respectively the True Positive Rate (at the group level) and the Continuously Ranked Probability Score, the latter in relative terms with respect to the AR(1) benchmark. Average values over 200 Monte Carlo simulations. Bootstrap standard errors of Monte Carlo averages in parentheses.

performance is substantially affected by the increase in the number of true active variables. In this case, the restricted Almon performs surprisingly better than the other polynomials, which show, for some DGPs, very low selection and predictive accuracy. Third, the ranking of the best-performing polynomials may depend, at least in part, on the shape of the underlying true weighting function. For instance, the restricted Almon seems well suited for bell-shaped and slow-decaying weights, but somewhat less for fast-decaying weights, while unrestricted and Bernstein poly-

Figure 2: Monte Carlo Experiment 2 - relative CRPS score of our BSGS-SS prior with respect to the BAGL-SS and BSGL priors



Notes: the markers denote the average relative CRPS score of our BSGS-SS prior with respect to the BAGL-SS (full markers) and BSGL (empty markers) priors. The error bars denote ± 2 bootstrap standard errors of average values obtained from 200 Monte Carlo simulations.

nomials can perform fairly well irrespective of the underlying weighting structure.[6] The performance of Legendre and Chebychev polynomials improves substantially under flat weights. This weighting structure is nevertheless less likely to describe the actual temporal aggregation for most economic data.

Finally, we compare again the predictive accuracy of our BSGS-SS prior with respect to the BSGL and BAGL-SS benchmarks. Unlike Experiment 1, here we do not control directly for the degree of sparsity (in the coefficients of the approximating functions) within the groups, but this arises indirectly from the specification of DGPs 1 to 4 considered. The results reported in Figure 2 suggest that the BSGS-SS prior provides a better predictive performance also in the mixed-frequency framework: the relative CRPS (with respect to the benchmarks) is on average less than

[6]However, we show that the U-MIDAS provides systematically less accurate parameter estimates compared to restricted Almon and Bernstein polynomials. These results are illustrated by the MSE in Figure A.2 in the Online Appendix.

one across the considered DGPs and polynomials (the best performing ones according to the results in Table 2), pointing to an average gain hovering around 5% to 10% overall.

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

We use the proposed prior for a nowcasting exercise of US GDP in the mixed-frequency framework (22) of Experiment 2, where $y_t = 400 \log(Y_t/Y_{t-1})$ is the annualized quarterly growth rate of GDP, and $\mathbf{x}_t$ is a vector of 122 macroeconomic series sampled at monthly frequency and extracted from the FRED-MD database (McCracken and Ng, 2016). The data sample starts in 1980Q1, while the pseudo out-of-sample analysis spans 2013Q1 to 2022Q4. Estimates are carried-out recursively using a rolling window of $T = 132$ quarterly observations, and h-step-ahead posterior predictive densities are generated from (14) through a direct forecast approach. We consider 3 nowcasting horizons ($h = 0, 1/3, 2/3$) and the two lag polynomials that provide overall better results in the simulation analysis reported in Section 6.2 – namely the restricted Almon and the orthonormal Bernstein polynomials. Further, as the considered data sample spans periods characterised by strong and uneven macroeconomic fluctuations (permanent and transitory), such as the Great Moderation, the Great Recession, and the Covid crisis, we also allow for time-varying volatility with heavy tails and occasional outliers in the regression errors by using the model in Section 4.1. Finally, we do not take into account real-time issues (ragged/jagged-edge data and revisions) and the data used for the analysis reflect the latest vintages available at the time of writing.[7]

We consider several modelling strategies to exploit our bi-level sparsity prior

[7]The GDP series used in the analysis was downloaded from FRED-MD database on July 2023 and we consider the 2023-06 vintage. We exclude from the analysis 5 series presenting sampling issues over the time span considered for the analysis (namely, new orders for consumer goods, non-borrowed reserves of depository institutions, 3-month AA financial commercial paper rate, 3-month commercial paper minus fed funds, consumer sentiment index).

Table 3: Empirical application: nowcasting US GDP - CRPS

	BSGS-SS			BSGL		
	$h=0$	$h=1/3$	$h=2/3$	$h=0$	$h=1/3$	$h=2/3$
Groups - Almon	0.78	0.81	0.76	0.82	0.81	0.77
	(0.02)	(0.02)	(0.00)	(0.16)	(0.02)	(0.00)
Groups - Bernstein	**0.70**	0.70	0.79	0.77	0.73	0.76
	(0.01)	(0.00)	(0.02)	(0.04)	(0.00)	(0.00)
Groups - all	0.71	**0.69**	0.77	0.77	0.74	0.78
	(0.01)	(0.00)	(0.01)	(0.04)	(0.00)	(0.01)
Whole dataset - Almon	0.71	0.81	**0.75**	0.81	0.77	0.79
	(0.01)	(0.06)	(0.00)	(0.23)	(0.01)	(0.00)
Whole dataset - Bernstein	0.71	0.75	0.83	0.85	0.89	0.86
	(0.01)	(0.01)	(0.03)	(0.27)	(0.35)	(0.07)
Whole dataset - all	0.70	0.78	0.76	0.85	0.84	0.87
	(0.01)	(0.03)	(0.00)	(0.30)	(0.05)	(0.08)

Notes: the Table displays the Continuously Ranked Probability Score (CRPS), in relative terms with respect to the AR(1) benchmark. BSGS-SS denotes the proposed bi-level sparsity prior. BSGL denotes the Bayesian Sparse Group Lasso prior (Xu and Ghosh, 2015). p-values of a test of unconditional predictive accuracy (Giacomini and White, 2006) with respect to the AR(1) benchmark in parentheses. Bold numbers denote the best outcomes across horizons and models.

approach. First, we estimate the forecasting model (22) on the whole set of 122 indicators. Then, since the simulation results in Section 6 point to a substantial deterioration of the selection and prediction performance in extreme sparse settings where $Ng \gg T$ then, we also estimate a separate model for each of the 8 groups of indicators partitioned as in McCracken and Ng (2016).[8] 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). For each of these specifications we compute density forecasts and we combine them along several dimensions (across the 8 groups, the volatility process, and the lag polynomials, see Online Appendix A.4 for more details). Density forecasts combination is carried out using the optimal prediction pool proposed by Geweke and Amisano (2011), which relies on log-scores.

The combined density forecasts are then evaluated by the means of the average CRPS. The relative scores with respect to the AR(1) benchmark are reported in

[8]The groups encompass *i)* output and income, *ii)* labour market, *iii)* housing, *iv)* consumption, orders, and inventories, *v)* money and credit, *vi)* interest and exchange rates, *vii)* prices, and *viii)* stock market. The number of indicators included in each group ranges between 5 and 31.

Table 3. Three main findings can be outlined. First, Bernstein polynomials provide accurate density forecasts for $h = 0$ and $h = 1/3$, but the restricted Almon performs best at $h = 2/3$. This outcome can be due to the way the different polynomials best describe the underlying weighting structure at each horizon. However, pooling these lag polynomials does not seem to provide here a systematic optimal strategy. Second, there is no clear-cut evidence in favour of a particular partition strategy. For $h = 0$ and $h = 2/3$, models including the whole set of indicators seem to perform better, or very closely, to those based on groups. Conversely, pooled forecasts from models based on group partitions tend to outperform for $h = 1/3$. Overall, and consistently with the simulation results reported in Section 6.2, this outcome suggests that the proposed procedure is robust to the presence of a large number of predictors. Finally, when compared to the BSGL prior (right hand side of Table 3), the results point to a considerable outperformance of our prior and substantial predictive gains for some horizons and pooling strategies.

8 Concluding remarks

This paper constructs optimal density forecasts for macroeconomic indicators in high-dimensional settings where the covariates present a known group structure. Such a structure arises, for instance, because variables in datasets are organised in groups or because of the flexible modelling of the relationship between the target variable and the set of covariates. The group-structure is important in order to reduce the dimensionality and we propose an optimal way to exploit it via the specification of a convenient prior distribution. By working under the assumption of bi-level sparsity, *i.e.* among and within groups, our procedure is able to detect the driving factors if only some regressors are relevant for forecast.

Density forecasts are constructed from the posterior predictive density, which we prove to present optimal asymptotic properties and to outperform many competitors in finite sample. We also provide parameter recovery and point forecasts, which

we prove to be minimax-optimal for our procedure under some mild conditions. Unlike alternative approaches, our procedure does not require orthogonality among covariates belonging to different groups. Finally, we show that a more general error structure, such as stochastic volatility and ARMA, can be accommodated by only slightly modifying the proposed prior.

Acknowledgements. The second author gratefully acknowledges financial support from Hi!Paris, and ANR-21-CE26-0003. The usual disclaimer applies. 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.

Online Appendix. It contains: additional elements about the prior, the Monte Carlo experiments and the application, guidelines for setting hyperparameters, the MCMC algorithm for the extended model with ARMA errors and stochastic volatility, additional theoretical results and all the proofs.

References

Adams, P.A., Adrian, T., Boyarchenko, N., Giannone, D., 2021. Forecasting macroeconomic risks. International Journal of Forecasting 37, 1173–1191.

Adrian, T., Boyarchenko, N., Giannone, D., 2019. Vulnerable growth. American Economic Review 109, 1263,Äì89.

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, 1094–1106.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives 28, 29–50.

Breheny, P., 2015. The Group Exponential LASSO for bi-level variable selection. Biometrics 71, 731–740.

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, 5975–6002.

Carriero, A., Clark, T.E., Marcellino, M., 2019. Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. Journal of Econometrics 212, 137–154.

Carriero, A., Clark, T.E., Marcellino, M., 2024a. Capturing macro-economic tail risks with Bayesian Vector Autoregressions. Journal of Money, Credit and Banking 56, 1099–1127.

Carriero, A., Clark, T.E., Marcellino, M., Mertens, E., 2024b. Addressing COVID-19 outliers in BVARs with stochastic volatility. The Review of Economics and Statistics 106, 1403–1417.

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., 2015. Bayesian linear regression with sparse priors. The Annals of Statistics 43, 1986–2018.

Chan, J.C.C., 2013. Moving average stochastic volatility models with application to inflation forecast. Journal of Econometrics 176, 162–172.

Chan, J.C.C., Hsiao, C., 2014. Estimation of stochastic volatility models with heavy tails and serial dependence, in: Jeliazkov, I., Yang, X.S. (Eds.), Bayesian Inference in the Social Sciences. John Wiley & Sons, Hoboken, New Jersey, pp. 159–180.

Chen, R.B., Chu, C.H., Yuan, S., Wu, Y.N., 2016. Bayesian sparse group selection. Journal of Computational and Graphical Statistics 25, 665–683.

Clark, T.E., 2011. Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. Journal of Business & Economic Statistics 29, 327–341.

Ferrara, L., Simoni, A., 2022. When are Google Data useful to nowcast GDP? an approach via preselection and shrinkage. Journal of Business & Economic Statistics 41, 1188–1202.

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, 57–82.

Garratt, A., Lee, K., Pesaran, M.H., Shin, Y., 2003. Forecast uncertainties in macroeconomic modeling. Journal of the American Statistical Association 98, 829–838.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. Journal of Econometrics 164, 130–141.

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

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578. doi:10.1111/j.1468-0262.2006.00718.x.

Hoffmann, M., Rousseau, J., Schmidt-Hieber, J., 2015. On adaptive posterior concentration rates. The Annals of Statistics 43, 2259 – 2295.

Huang, J., Breheny, P., Ma, S., 2012. A selective review of group selection in high-dimensional models. Statistical Science: A Review Journal of the Institute of Mathematical Statistics 27, 481–499.

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.

Lenza, M., Primiceri, G.E., 2022. How to estimate a vector autoregression after March 2020. Journal of Applied Econometrics 37, 688–699.

Li, Z., Zhang, Y., Yin, J., 2024. Estimating double sparse structures over $\ell_u(\ell_q)$-balls: Minimax rates and phase transition. IEEE Transactions on Information Theory 70, 7066–7088.

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

Mogliani, M., Simoni, A., 2021. Bayesian MIDAS penalized regressions: Estimation, selection, and prediction. Journal of Econometrics 222, 833–860.

Raskutti, G., Wainwright, M.J., Yu, B., 2011. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE Transactions on Information Theory 57, 6976–6994.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A Sparse-Group Lasso. Journal of Computational and Graphical Statistics 22, 231–245.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B: Statistical Methodology 64, 583–639.

Stock, J.H., Watson, M.W., 2016. Core inflation and trend inflation. The Review of Economics and Statistics 98, 770–784.

Xu, X., Ghosh, M., 2015. Bayesian variable selection and estimation for Group Lasso. Bayesian Analysis 10, 909–936.

Zhang, B., Chan, J.C.C., Cross, J.L., 2020. Stochastic volatility models with ARMA innovations: An application to G7 inflation forecasts. International Journal of Forecasting 36, 1318–1328.

CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

CREST
Center for Research in Economics and Statistics
UMR 9194

5 Avenue Henry Le Chatelier
TSA 96642
91764 Palaiseau Cedex
FRANCE

Phone: +33 (0)1 70 26 67 00
Email: info@crest.science
https://crest.science/

The Center for Research in Economics and Statistics (CREST) is a leading French scientific institution for advanced research on quantitative methods applied to the social sciences.

CREST is a joint interdisciplinary unit of research and faculty members of CNRS, ENSAE Paris, ENSAI and the Economics Department of Ecole Polytechnique. Its activities are located physically in the ENSAE Paris building on the Palaiseau campus of Institut Polytechnique de Paris and secondarily on the Ker-Lann campus of ENSAI Rennes.

cnrs

GENES Groupe des écoles nationales d'économie et statistique