Dynamic variable selection in high-dimensional predictive regressions*

Mauro Bernardi† Daniele Bianchi‡ Nicolas Bianco†

First draft: January 2022. This draft: May 23, 2023

Abstract

We develop methodology and theory for a general Bayesian approach towards dynamic variable selection in high-dimensional regression models with time-varying parameters. Specifically, we propose a variational inference scheme which features dynamic sparsity-inducing properties so that different subsets of "active" predictors can be identified over different time periods. We compare our modeling framework against established static and dynamic variable selection methods both in simulation and within the context of two common problems in macroeconomics and finance: inflation forecasting and equity returns predictability. The results show that our approach helps to tease out more accurately the dynamic impact of different predictors over time. This translates into significant gains in terms of out-of-sample point and density forecasting accuracy. We believe our results highlight the importance of taking a dynamic approach towards variable selection for economic modeling and forecasting.

Keywords: Economic forecasting, Bayesian methods, Dynamic variable selection, Time-varying parameters, Mean-field approximation, Variational Bayes inference.

JEL codes: C11, C32, C55, C53, G11

*We are thankful to Andrea Carriero, David Rossell (discussant), Sylvia Frühwirth-Schnatter, Andrew Harvey, and seminar participants at Universitat Pompeu Fabra, University of Padua, Queen Mary University of London, the workshop "Econometrics and Learning" at Imperial College London, and the ISBA 2022 World Conference for their helpful comments and valuable suggestions.

†Department of Statistical Sciences, University of Padova, Italy. Email: mauro.bernardi@unipd.it

‡School of Economics and Finance, Queen Mary University of London, United Kingdom. Email: d.bianchi@qmul.ac.uk Web: whitesphd.com

†Department of Economics and Business, Universitat Pompeu Fabra, Spain. Email: nicolas.bianco@phd.unipd.it Web: whitenoise8.github.io

1 Introduction

Predicting the dynamics of economic variables is a key input for policy and investment decision making processes. For example, forecasts of inflation influence both policy makers in their monetary and fiscal policy decisions as well as investors seeking to hedge against inflation risk. Similarly, forecasting asset returns is crucial for optimal portfolio allocation and represents one of the fundamental goals of empirical asset pricing. However, economic forecasting is notoriously difficult since the expected variation of economic outcomes is often buried under the noise of realised economic activity. To address this issue, market participants and researchers, alike, often leverage on large data sets and rely on high-dimensional forecasting models, of which regression analysis is often a core building block.

As the nature of the variables carrying significant explanatory power is arguably uncertain *a priori*, decision makers often consider the entire set of available predictors – mitigating the risk of omitting important information – and select *a posteriori* those which correlates the most with the targeted economic outcome. For this reason, variable selection techniques have become increasingly popular tools for economic forecasting, especially within the context of linear regression models (see, e.g., Giannone et al., 2021). However, and perhaps not surprisingly, the same variable selection method could argue in favour of different predictors for the same target outcome over different time periods. Such discrepancy stem from the fact that predictability likely changes over time, either at the intensive margin – a variable carry significant forecasting power for longer –, or at the extensive margin – more predictors carry significant explanatory power at a given point in time. In other words, the model dimension on which decision makers can act upon is potentially varying over time.[1]

In this paper, we address this issue and develop methodology and theory for a novel Bayesian dynamic variable selection method for high-dimensional predictive regressions with time-varying parameters. Specifically, we propose a dynamic Bernoulli-Gaussian (`BG` henceforth) regression specification whereby variable selection takes the form of a smooth latent stochastic process which interacts with a conventional dynamic linear model (see, e.g., West and Harrison, 2006). Posterior estimates are obtained via a novel semi-parametric variational Bayes inference approach, expanding on Rohde and Wand (2016) and Ormerod et al. (2017). We provide evidence that this approach represents a more efficient alternative to a Markov Chain Monte Carlo (MCMC) method with comparable posterior concentration properties.

[1]This is often referred in the literature as a distinction between "horizontal sparsity" and "vertical" sparsity (see, e.g., Uribe and Lopes, 2020; Ročková and McAlinn, 2021).

Our approach towards variable selection has three key features: first, the posterior estimates require only a minimal set of assumptions on hyper-parameters and initial conditions (see Section 2.2). This is particularly relevant within the context of high-dimensional regression models with time-varying parameters where tailoring individual hyper-parameters can be prohibitive. Second, our approach allows for dynamic selection of "active" predictors. That is a point mass posterior inclusion probability is placed at zero when a predictor does not carry significant explanatory power at a given time period. Third, we can sequentially reduce the model dimension by discarding those predictors which does not carry information over the entire sample, while exploring the trajectory of sparsity of the remaining time-varying regression coefficients. This improves the computational cost and the estimation efficiency, and thus the accuracy of the posterior estimates.[2]

We investigate the accuracy of both dynamic variable selection and posterior point estimates based on an extensive simulation setting in which different regression parameters display different patterns over time. As benchmarks, we first consider a variety of established static variable selection methods, such as the two-component mixture priors of George and McCulloch (1993); Ročková and George (2014) and Giannone et al. (2021), the normal-gamma prior of Griffin and Brown (2010) and the horseshoe prior of Carvalho et al. (2010). For these static prior formulations, a simple dynamic is imposed via a rolling window, a widely used non-parametric approach to approximate parameters time variation (see, e.g., Inoue et al., 2017).[3] In addition, we compare our dynamic `BG` method against two recent developments in dynamic variable selection in time-varying regression models, such as the dynamic spike-and-slab prior of Koop and Korobilis (2020) and Ročková and McAlinn (2021). Overall, the simulation results suggest that our dynamic `BG` model outperforms these competing approaches both with respect to the identification of the significant predictors over time, as well as the accuracy of the posterior point estimates.

Intuitively, the ability to identify more accurately which predictors matter over time should be of first-order importance for forecasting and decision making. To this aim, we compare our model vis-á-vis a comprehensive set of alternative regression-based forecasting strategies within the context of two common problems in macroeconomics and finance: inflation forecasting based on a large set of macroeconomic variables (see, e.g., Stock and

[2]This is akin to variance inflation when keeping irrelevant predictors in least squares estimates. Fava and Lopes (2021) showed both in simulation and empirically the effect of irrelevant predictors in the context of discrete mixture priors for variables selection.

[3]Sparsity in the posterior estimates of the global-local shrinkage priors is imposed via the signal adaptive variable selector (SAVS henceforth) of Ray and Bhattacharya (2018).

Watson, 2006; Faust and Wright, 2013) and the predictability of the equity premium (see, e.g., Welch and Goyal, 2008; Rapach et al., 2010; Dong et al., 2022).

As far as inflation forecasting is concerned, we consider more than 220 quarterly macroeconomic predictors from the FRED-QD database of McCracken and Ng (2020). The target variables consist of four measures of inflation – namely total CPI, core CPI, GDP deflator, and PCE deflator –, for four different forecasting horizons – from one to eight quarters ahead. Perhaps not surprisingly, the empirical results confirm the widespread conventional wisdom that parsimonious models, such as the unobserved component model of Stock and Watson (2007), represent rather challenging benchmarks. Nevertheless, our dynamic `BG` model outperforms all static and dynamic variable selection methods which make use of macroeconomic predictors, both in a mean-squared error sense and in density forecasts, and across different horizons.

Interestingly, a retrospective analysis of dynamic posterior inclusion probabilities show that (1) only a handful of predictors carry a meaningful explanatory power, and (2) our model provides an alternative view to some of the main theory-based inflation predictors. For instance, real consumption expenditure, which proxies demand pressure on inflation, carries predictive power on the one-quarter ahead core CPI only for a short-period during 2021, a period characterised by large fiscal stimulus. Similarly, short-term unemployment carries a significant signal to predict one-quarter ahead change in the GDP deflator from the great financial crisis towards the end of the sample for the GDP deflator. The latter could be interpreted as evidence in favour of a time-varying Phillips curve, whereby the inverse relationship between unemployment and inflation is supported by the data but only in specific periods.

For the application on financial forecasting, we build upon Jensen et al. (2022) and assess the predictive content of more than 150 characteristic-managed portfolios for the one-month ahead returns on the aggregate stock market portfolio, expanding on the original framework of Dong et al. (2022). Consistent with the latter, both the prediction from the recursively calculated sample mean and equal-weight forecasts from individual predictors represent rather challenging benchmarks. Nevertheless, the empirical results show that our dynamic sparse regression framework outperforms both static and dynamic variable selection methods. Finally, a retrospective analysis of dynamic posterior inclusion probabilities suggests that expected returns correlates with only few risk factors related to trading frictions and liquidity, such as `max1_21d` (see Bali et al., 2011) and `turnover_126d` (see Datar et al., 1998).

This paper connects to two main streams of literature. The first relates to the use of Bayesian methods for variable selection in high-dimensional regression models. Conventional approaches towards selecting predictors are mostly confined in the realm of static regression models (see, e.g., Ročková and George, 2018; Giannone et al., 2021; Fava and Lopes, 2021; Ray and Szabó, 2022 and the references therein). This is despite there is ample evidence in the economic literature on the importance of considering time-varying effects of predictors on both macroeconomic and financial forecasting (see, e.g. Primiceri, 2005; West and Harrison, 2006; Dangl and Halling, 2012; Pettenuzzo et al., 2014; Farmer et al., 2022, among others).

A notable early exception to such static approach is Nakajima and West (2013), which introduce a dynamic regression framework whereby time-varying coefficients are excluded based on a latent threshold parameter. Similarly, Kalli and Griffin (2014) proposed a normal-gamma autoregressive process to dynamically shrink towards zero unimportant coefficients. The latter approach falls into the class of dynamic shrinkage processes studied in Kowal et al. (2019). Other methods aim to perform model selection rather than shrinkage. For example, Koop and Korobilis (2020) expand on Koop and Korobilis (2012) by leveraging the flexibility of variational Bayes inference and consider a dynamic spike-and-slab prior specification for variable selection in time-varying regression models, while assuming stochastic and independent inclusion probabilities. Similarly, Uribe and Lopes (2020) and Ročková and McAlinn (2021) proposed a dynamic variable selection method that leverages on the class of mixture priors originally proposed by Mitchell and Beauchamp (1988); George and McCulloch (1997).

A second strand of literature we contribute to is related to regression-based economic forecasting. In particular, inflation forecasting represents a widely used setting to test the accuracy of large-scale predictive regression models within the context of policy making (see, e.g., Stock and Watson, 2007, 2010; Chan et al., 2012; Koop and Korobilis, 2012; Korobilis, 2013; Kalli and Griffin, 2014; Bitto and Frühwirth-Schnatter, 2019; Ročková and McAlinn, 2021, among others). The time series variation of expected inflation is particularly problematic to measure, since conventional predictors often do not seem to capture significant co-movements and cross-signals between economic activity and prices which might improve out-of-sample predictability. Similarly, forecasting the dynamics of stock returns represents a particularly challenging task for predictive regressions due to the small signal-to-noise ratio in financial returns (see, e.g., Welch and Goyal, 2008).

2 Model specification and inference

We present our approach as a dynamic linear predictive model that link a scalar response y_t at time t to a set of p known predictors $\mathbf{x}_{t-1} = (x_{1t-1}, \ldots, x_{pt-1})'$ through the relation

$$y_t = \sum_{j=1}^{p} \beta_{jt} x_{jt-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0, e^{h_t}), \qquad t = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}_t = (\beta_{1t}, \ldots, \beta_{pt})'$ is a time-varying vector of regression coefficients and $h_t = \log \sigma_t^2$ is the log of the unknown observational variance of the residual ε_t at time t (see, e.g., West and Harrison, 2006). Notably, variable selection is particular relevant for time-varying parameter regressions. This is because as the model propagates forward, the inclusion of irrelevant predictors generate a proliferation of the state space and potentially accumulates noise, with detrimental consequences for the out-of-sample predictive performance. Therefore, seeking sparsity in the set of predictors is a natural remedy against the loss of statistical efficiency and forecast ability.

In this paper, we adopt the point of view that predictors can dynamically enter or leave the regression model as time progresses. Specifically, the set of "active" predictors – meaning the variables that carry significant predictive power – can change over time according to

$$\beta_{jt} = b_{jt}\gamma_{jt}, \qquad \text{where} \qquad b_{jt} = b_{jt-1} + v_{jt} \qquad v_{jt} \sim \mathsf{N}\left(0, \eta_j^2\right), \tag{2}$$

with $b_{j0} \sim N\left(0, k_0\eta_j^2\right)$ the initial state, and $\gamma_{jt} \in \{0, 1\}$ an indicator variable which identifies if the jth predictor is included or not in the model specification. This process is reminiscent of a dynamic Bernoulli-Gaussian (`BG`) regression model (see, e.g., Soussen et al., 2011; Ormerod et al., 2017). By leveraging the first-order Markov property, the joint distribution of $\mathbf{b}_j = (b_{j0}, \ldots, b_{jn})'$ for $j = 1, \ldots, p$ can be re-written as $p(\mathbf{b}_j) = p(b_{j0})p(b_{j1}|b_{j0}) \ldots p(b_{jn}|b_{jn-1})$. This admits a Gaussian Markov random field (GMRF) representation $\mathbf{b}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \eta_j^2\mathbf{Q}^{-1})$ with $\mathbf{Q}$ a tridiagonal precision matrix with diagonal elements $q_{1,1} = 1 + 1/k_0$, $q_{n+1,n+1} = 1$, and $q_{l,l} = 2$ for $l = 2, \ldots, n$. The off-diagonal elements are $q_{l,m} = -1$ if $|l - m| = 1$ and 0 elsewhere (see Rue and Held, 2005). The same representation applies for the log-volatility process $\mathbf{h} = (h_0, \ldots, h_n)'$ with initial state $h_0 \sim \mathsf{N}\left(0, k_0\nu^2\right)$, such that $h_t = h_{t-1} + e_t$ with $e_t \sim \mathsf{N}\left(0, \nu^2\right)$ admits $\mathbf{h} \sim \mathsf{N}_{n+1}(\mathbf{0}, \nu^2\mathbf{Q}^{-1})$.

Equation (2) assumes that the time-varying process $\{b_{jt}\}_{t=1}^T$, give rise to the regression coefficients $\{\beta_{jt}\}_{t=1}^T$ only by interacting with the latent indicator $\{\gamma_{jt}\}_{t=1}^T$. This formulation

implies that the posterior inclusion probability $\mathbb{P}(\gamma_{jt} = 1)$ is a persistent latent stochastic process. This differs from Koop and Korobilis, 2020; Uribe and Lopes, 2020; Ročková and McAlinn, 2021 in which variable selection is embedded into a prior spike-and-slab structure.[4] The indicator variable γ_{jt} given the auxiliary parameters ω_{jt} is assumed to be $\gamma_{jt}|\omega_{jt} \sim \mathsf{Bern}(\text{expit}(\omega_{jt}))$ for $j = 1, \ldots, p$, where $\text{expit}(\cdot)$ is the inverse of the logit function. As a result, the persistence of the inclusion probability $\mathbb{P}(\gamma_{jt} = 1)$ is driven by $\boldsymbol{\omega}_j = (\omega_{j0}, \ldots, \omega_{jn})'$, which admits a GMRF representation of the form $\boldsymbol{\omega}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \xi_j^2 \mathbf{Q}^{-1})$. The marginal distribution for the vector $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jn})'$ is retrieved by integrating out $\boldsymbol{\omega}_j$ as,

$$p(\gamma_{j1}, \ldots, \gamma_{jn}) = \int p(\boldsymbol{\omega}_j) \prod_{t=1}^{n} p(\gamma_{jt}|\omega_{jt}) \, d\boldsymbol{\omega}_j, \tag{3}$$

so that $\gamma_{j1}, \ldots, \gamma_{jn}$ represent autocorrelated latent states for each $j = 1, \ldots, p$.

2.1 Variational Bayes inference

A variational Bayes (VB) approach to inference requires to minimize the Kullback-Leibler (Kullback and Leibler, 1951) divergence measure (KL) between an approximating density $q(\boldsymbol{\vartheta})$ and the true posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$, (see, e.g. Blei et al., 2017). The KL divergence cannot be directly minimized with respect to $\boldsymbol{\vartheta}$ because it involves the expectation with respect to the unknown true posterior distribution. Ormerod and Wand (2010) show that the problem of minimizing KL can be equivalently stated as the maximization of the variational lower bound (ELBO) denoted by $\underline{p}\,(\mathbf{y}; q)$:

$$q^*(\boldsymbol{\vartheta}) = \arg \max_{q(\boldsymbol{\vartheta}) \in \mathcal{Q}} \log \underline{p}\,(\mathbf{y}; q)\,, \quad \underline{p}\,(\mathbf{y}; q) = \int q(\boldsymbol{\vartheta}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\vartheta})}{q(\boldsymbol{\vartheta})} \right\} d\boldsymbol{\vartheta}, \tag{4}$$

where $q^*(\boldsymbol{\vartheta}) \in \mathcal{Q}$ represents the optimal variational density and $\mathcal{Q}$ is a space of functions. The choice of the family of distributions $\mathcal{Q}$ is critical and leads to different algorithmic approaches. We consider a mean-field variational Bayes (MFVB) approach which is based on a non-parametric restriction for the variational density, i.e. $q(\boldsymbol{\vartheta}) = \prod_{i=1}^{p} q_i(\boldsymbol{\vartheta}_i)$ for a partition $\{\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_p\}$ of the parameter vector $\boldsymbol{\vartheta}$. Therefore, a closed form expression for

[4] For instance, in existing dynamic spike-and-slab formulations of the evolution of γ_{jt} is assumed to be a priori independent over time (see Koop and Korobilis, 2020) or having a deterministic evolution of $\mathbb{P}(\gamma_{jt} = 1)$, given the information up to $t - 1$ (as in Ročková and McAlinn, 2021).

the optimal variational density of each component $q(\boldsymbol{\vartheta}_j)$ is defined as:

$$q^*(\boldsymbol{\vartheta}_j) \propto \exp\left\{\mathbb{E}_{q(\boldsymbol{\vartheta}\setminus\boldsymbol{\vartheta}_j)}\Big[\log p(\mathbf{y},\boldsymbol{\vartheta})\Big]\right\}, \qquad q(\boldsymbol{\vartheta}\setminus\boldsymbol{\vartheta}_j) = \prod_{\substack{i=1\\ i\neq j}}^{p} q_i(\boldsymbol{\vartheta}_i), \tag{5}$$

where the expectation is taken with respect to the joint approximating density with the j-th element of the partition removed $q(\boldsymbol{\vartheta}\setminus\boldsymbol{\vartheta}_j)$. This allows to implement a coordinate ascent variational inference (CAVI) algorithm to estimate the optimal density $q^*(\boldsymbol{\vartheta})$. Equation (5) shows that the factorization $q(\boldsymbol{\vartheta})$ plays a key role. Let $\boldsymbol{\vartheta} = (\mathbf{h}', \mathbf{b}', \boldsymbol{\gamma}', \boldsymbol{\omega}', \nu^2, \boldsymbol{\eta}^{2\prime}, \boldsymbol{\xi}^{2\prime})'$ the joint distribution of the model parameters and latent states. The key ingredient for the mean-field factorization is the joint distribution $p(\mathbf{y},\boldsymbol{\vartheta})$, which can be factorized as follows,

$$p(\mathbf{y},\boldsymbol{\vartheta}) = p(\mathbf{y}|\boldsymbol{\vartheta})p(\mathbf{h})p(\nu^2)\prod_{j=1}^{p} p(\mathbf{b}_j|\eta_j^2)p(\boldsymbol{\omega}_j|\xi_j^2)p(\eta_j^2)p(\xi_j^2)\underbrace{\prod_{t=1}^{n} p(\gamma_{jt}|\omega_{jt})}_{p(\boldsymbol{\gamma}_j|\boldsymbol{\omega}_j)}. \tag{6}$$

Notice that the full conditional distribution of $\boldsymbol{\omega}_j$ is not of a know form. Following Polson et al. (2013), we exploit a Polya-Gamma representation,

$$p(\gamma_{jt}|\omega_{jt}) = \int_0^{+\infty} p(\gamma_{jt}|z_{jt},\omega_{jt})p(z_{jt}|\omega_{jt})\,dz_{jt}, \tag{7}$$

where $p(z_{jt})$ is the probability density function of a Polya-Gamma $\mathsf{PG}(1,0)$ random variable. This allows for a computationally tractable approximation of Eq.(6). Therefore, we propose a mean-field factorization of the form,

$$q(\boldsymbol{\vartheta}) = q(\mathbf{h})q(\nu^2)\prod_{j=1}^{p} q(\mathbf{b}_j)q(\boldsymbol{\omega}_j)q(\eta_j^2)q(\xi_j^2)\prod_{t=1}^{n} q(\gamma_{jt})q(z_{jt}), \tag{8}$$

where a joint distribution for $\mathbf{h}$, $\mathbf{b}_j$, and $\boldsymbol{\omega}_j$ is required in order to preserve the time dependence and to provide a global approximation for the vector of latent states. We now discuss in turn each of the components in $q(\boldsymbol{\vartheta})$.

Optimal variational densities. We now discuss in details the main optimal variational densities for the estimation of the time-varying regression parameters $q^*(\mathbf{b}_j)$, the variable selection indicators $q^*(\gamma_{jt})$, and the stochastic log-volatility process $q^*(\mathbf{h})$. For the interested reader, the full set of analytical derivations and proofs is available in Appendix B.

Proposition 2.1. *The optimal variational density for the time-varying regression parameters* $\mathbf{b}_j = (b_{j0}, b_{j1}, \ldots, b_{jn})'$ *is a multivariate Gaussian* $q^*(\mathbf{b}_j) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(\mathbf{b}_j)}, \boldsymbol{\Sigma}_{q(\mathbf{b}_j)})$*, where:*

$$\boldsymbol{\Sigma}_{q(\mathbf{b}_j)} = (\mathbf{D}_j{}^2 + \mu_{q(1/\eta_j^2)}\mathbf{Q})^{-1}, \qquad \boldsymbol{\mu}_{q(\mathbf{b}_j)} = \boldsymbol{\Sigma}_{q(\mathbf{b}_j)}\mathbf{D}_j\boldsymbol{\mu}_{q(\varepsilon_{-j})}, \tag{9}$$

where $\mathbf{D}_j$ *and* $\mathbf{D}_j^2$ *are diagonal matrices with elements equal to* $[\mathbf{D}_j] = \mu_{q(1/\sigma_t^2)}\mu_{q(\gamma_{jt})}x_{jt-1}$ *and* $[\mathbf{D}_j]_t^2 = \mu_{q(1/\sigma_t^2)}\mu_{q(\gamma_{jt})}x_{jt-1}^2$*, respectively. Moreover,* $\boldsymbol{\mu}_{q(\varepsilon_{-j})}$ *is the vector of partial residuals with elements* $\mu_{q(\varepsilon_{-jt})} = y_t - \sum_{k=1,k\neq j}^{p} x_{kt-1}\mu_{q(\gamma_{kt})}\mu_{q(b_{kt})}$*.*

Proof. See proof B.1 in Appendix B. □

Proposition 2.1 shows that both the posterior mean and variance of a given vector of regression parameters for the variable j, depends on the posterior estimates $\mu_{q(\gamma_{jt})}$ of the entire trajectory of $\gamma_{jt}, t = 1, \ldots, n$ from the optimal variational density $q^*(\gamma_{jt})$. The latter is defined in Proposition 2.2.

Proposition 2.2. *The optimal variational density for the parameters* γ_{jt} *is a Bernoulli random variable* $q^*(\gamma_{jt}) \equiv \mathsf{Bern}(\text{expit}(\omega_{q(\gamma_{jt})}))$*, where* $\text{expit}(\cdot)$ *is the inverse of the logit function and* $\omega_{q(\gamma_{jt})} = \mu_{q(\omega_{jt})} - \frac{1}{2}\mu_{q(1/\sigma_t^2)}(x_{jt-1}^2\mathbb{E}_q[b_{jt}^2] - 2\mu_{q(b_{jt})}x_{jt-1}\mu_{q(\varepsilon_{-jt})})$*.*

Proof. See proof B.2 in Appendix B. □

The parameter $\mu_{q(1/\sigma_t^2)} \equiv \mathbb{E}_q[1/\sigma_t^2]$ and is defined as in Remark B.1 in Appendix B. In addition, $\mu_{q(\omega_{jt})}$ represents the mean of the optimal variational density for the auxiliary parameter $\boldsymbol{\omega}_j$. The latter is defined in Proposition 2.3.

Proposition 2.3. *The optimal variational density for the parameter* $\boldsymbol{\omega}_j$ *is a multivariate Gaussian* $q^*(\boldsymbol{\omega}_j) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(\omega_j)}, \boldsymbol{\Sigma}_{q(\omega_j)})$*, where:*

$$\boldsymbol{\Sigma}_{q(\omega_j)} = (\mathsf{Diag}(0, \boldsymbol{\mu}_{q(z_j)}) + \mu_{q(1/\xi_j^2)}\mathbf{Q})^{-1}, \qquad \boldsymbol{\mu}_{q(\omega_j)} = \boldsymbol{\Sigma}_{q(\omega_j)}(0, \boldsymbol{\mu}_{q(\bar{\gamma}_j)}^{\intercal})^{\intercal}, \tag{10}$$

with $\boldsymbol{\mu}_{q(\bar{\gamma}_j)} = \boldsymbol{\mu}_{q(\gamma_j)} - 1/2\boldsymbol{\iota}_n$*.*

Proof. See proof B.3 in Appendix B. □

The means $\boldsymbol{\mu}_{q(z_j)}, \mu_{q(1/\xi_j^2)}$ of the optimal variational densities for the auxiliary variable $q^*(z_{jt}) \equiv \mathsf{PG}(1, \sqrt{\mu_{q(\omega_{jt}^2)}})$ and the state variance $q^*(\xi^2)$ are defined in Appendix B in Proposition B.7 and B.9, respectively. Recall from Eq.(2) that $\beta_{jt} = b_{jt}\gamma_{jt}$. The corresponding optimal variational density is provided in Proposition 2.4.

Proposition 2.4. *Let $q^*(\mathbf{b}_j)$ and $q^*(\gamma_{jt})$ be the optimal variational densities presented in Propositions 2.1 and 2.2. Define $\boldsymbol{\beta}_j = \boldsymbol{\Gamma}_j \mathbf{b}_j$, where the matrix $\boldsymbol{\Gamma}_j = \mathsf{diag}(1, \gamma_{j1}, \ldots, \gamma_{jn})$. The optimal variational density of $\boldsymbol{\beta}_j$ is given by a mixture of multivariate Gaussian distributions:*

$$q^*(\boldsymbol{\beta}_j) = \sum_{\mathbf{s} \in \mathcal{S}} w_s \, \mathsf{N}_{n+1}(\mathbf{D}_s \boldsymbol{\mu}_{q(\mathbf{b}_j)}, \mathbf{D}_s^{1/2} \boldsymbol{\Sigma}_{q(\mathbf{b}_j)} \mathbf{D}_s^{1/2}), \tag{11}$$

where $\mathcal{S}$ is a sequence of $\{0, 1\}$ of length n with cardinality $|\mathcal{S}| = 2^n$, the diagonal matrix $\mathbf{D}_s = \mathsf{diag}(1, s_1, \ldots, s_n)$, and mixing weights:

$$w_s = \prod_{t=1}^{n} \mu_{q(\gamma_{jt})}^{s_t} (1 - \mu_{q(\gamma_{jt})})^{1 - s_t}, \tag{12}$$

where $\mathbf{s} = (s_1, \ldots, s_t, \ldots, s_n) \in \mathcal{S}$ is an element in $\mathcal{S}$. Moreover, the mean and variance can be computed analytically:

$$\boldsymbol{\mu}_{q(\beta_j)} = \boldsymbol{\mu}_{q(\Gamma_j)} \boldsymbol{\mu}_{q(\mathbf{b}_j)}, \tag{13}$$

$$\boldsymbol{\Sigma}_{q(\beta_j)} = (\boldsymbol{\mu}_{q(\gamma_j)} \boldsymbol{\mu}'_{q(\gamma_j)} + \mathbf{W}_{\mu_{q(\gamma_j)}}) \odot \boldsymbol{\Sigma}_{q(\mathbf{b}_j)} + \mathbf{W}_{\mu_{q(\gamma_j)}} \odot \boldsymbol{\mu}_{q(\mathbf{b}_j)} \boldsymbol{\mu}'_{q(\mathbf{b}_j)}, \tag{14}$$

where $\mathbf{W}_{\mu_{q(\gamma_j)}}$ is a diagonal matrix with elements $\left(1, \{\mu_{q(\gamma_{jt})}(1 - \mu_{q(\gamma_{jt})})\}_{t=1}^{n}\right)$.

Proof. See proof B.4 in Appendix B. □

For the stochastic volatility process, we adopt a parametric approach to find the optimal variational density $q^*(\mathbf{h})$. Specifically, we leverage on a GMRF representation of the vector $\mathbf{h} \sim \mathsf{N}_{n+1}(\mathbf{0}, \nu^2 \mathbf{Q}^{-1})$ and exploit the results in Rohde and Wand (2016). They provide an iterative updating scheme for the variational parameters when the approximating density is a multivariate Gaussian. Proposition 2.5 provides the optimal updating scheme.

Proposition 2.5. *Let $\boldsymbol{\varepsilon}^2 = \boldsymbol{\varepsilon} \odot \boldsymbol{\varepsilon}$ with components $[\boldsymbol{\varepsilon}^2]_t = (y_t - \boldsymbol{\beta}'_t \mathbf{x}_t)^2$. Assuming a GMRF approximation $q^*(\mathbf{h}) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Omega}_{q(h)}^{-1})$, with mean vector $\boldsymbol{\mu}_{q(h)}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{q(h)} = \boldsymbol{\Omega}_{q(h)}^{-1}$, an iterative optimization algorithm can be set as:*

$$\boldsymbol{\Sigma}_{q(h)}^{new} = \left[\nabla^2_{\boldsymbol{\mu}_{q(h)}, \boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old}) \right]^{-1} \tag{15}$$

$$\boldsymbol{\mu}_{q(h)}^{new} = \boldsymbol{\mu}_{q(h)}^{old} + \boldsymbol{\Sigma}_{q(h)}^{new} \nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}^{old}, \boldsymbol{\Sigma}_{q(h)}^{old}). \tag{16}$$

where

$$\nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}^{old}_{q(h)}, \boldsymbol{\Sigma}^{old}_{q(h)}) = -\frac{\boldsymbol{\iota}_n}{2} + \frac{1}{2}\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}^{old}_{q(h)} + \boldsymbol{\sigma}^{2\,old}_{q(h)}/2} - \mu_{q(1/\nu^2)}\mathbf{Q}\boldsymbol{\mu}^{old}_{q(h)}, \tag{17}$$

and

$$\nabla^2_{\boldsymbol{\mu}_{q(h)}, \boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}^{old}_{q(h)}, \boldsymbol{\Sigma}^{old}_{q(h)}) = -\frac{1}{2}\mathsf{Diag}(\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}^{old}_{q(h)} + \boldsymbol{\sigma}^{2\,old}_{q(h)}/2}) - \mu_{q(1/\nu^2)}\mathbf{Q}, \tag{18}$$

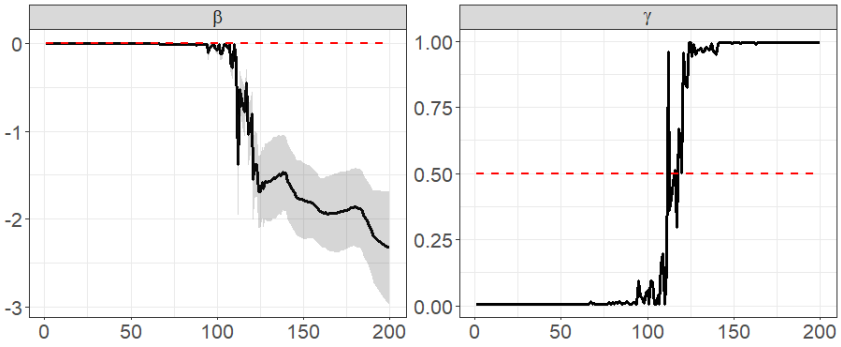
denote the first and second derivative of $S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$ *with respect to* $\boldsymbol{\mu}_{q(h)}$ *and evaluated at* $(\boldsymbol{\mu}^{old}_{q(h)}, \boldsymbol{\Sigma}^{old}_{q(h)})$*, and* $\boldsymbol{\sigma}^2_{q(h)} = \mathsf{diag}(\boldsymbol{\Sigma}_{q(h)})$.

Proof. See proof B.5 in Appendix B. □

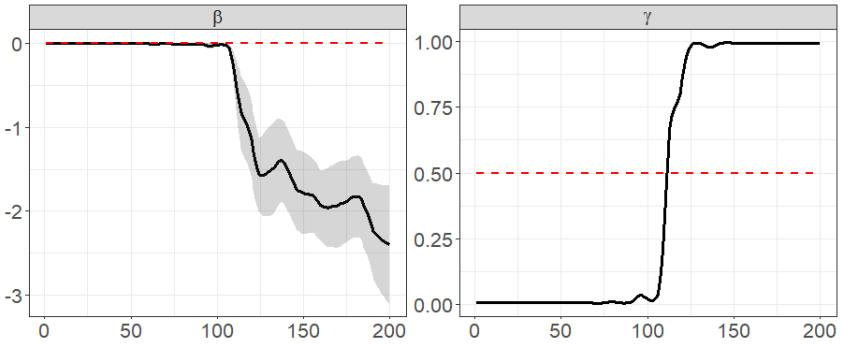
Proposition B.6 in Appendix B also reports the optimal variational density for the homoskedastic case in which the prior for the variance of the residuals in Eq.(1) is an inverse-gamma $\sigma^2 \sim \mathsf{IG}(A_\sigma, B_\sigma)$. In addition, Propositions B.8, B.9, and B.10 in Appendix B report the optimal variational densities $q^*(\eta_j^2)$, $q^*(\xi_j^2)$, and $q^*(\nu^2)$, respectively. As far as the prior distributions are concerned, we place inverse-gamma priors for the variances parameters $\nu^2 \sim \mathsf{IG}(A_\nu, B_\nu)$, $\eta_j^2 \sim \mathsf{IG}(A_\eta, B_\eta)$, and $\xi_j^2 \sim \mathsf{IG}(A_\xi, B_\xi)$, which represents a common choice in Bayesian analysis. We discuss the choice of prior hyper-parameters in Section 2.2.

Smoothing the sparsity dynamics. Proposition 2.2 shows that the variational density of $q(\boldsymbol{\gamma}_j) = \prod_{t=1}^n q(\gamma_{jt})$ is such that $\gamma_{jt} \sim \mathsf{Bern}(\mathrm{expit}(\omega_{q(\gamma_{jt})}))$. This implies that the whole time trajectory of posterior inclusion probabilities can be obtained as the mean vector $\mathbb{E}_q(\boldsymbol{\gamma}_j) = \mathrm{expit}(\boldsymbol{\omega}_{q(\gamma_j)})$. Although computationally convenient, this is an entirely data-driven approach which could produce erratic posterior inclusion probabilities, especially with noisy observations. The right panel of Figure 1(a) shows this case in point. The posterior inclusion probability $\mathbb{P}(\gamma_{jt} = 1)$ could point towards a given predictor for a very short period of time. This could be quite inconvenient in practice since $\beta_{jt} = b_{jt}\gamma_{jt}$, such that the dynamics of β_{jt} inherits the erratic behavior of the posterior inclusion probability as shown in Figure 1(a)).

To address this issue, we propose an alternative parametric approximation of the variational density of $q(\boldsymbol{\gamma}_j)$ which regularise the estimates of the time trajectory of $\mathbb{P}(\gamma_{jt} = 1)$ for $t = 1, \ldots, n$. In particular, we approximate the sequence of densities $\{q(\gamma_{jt})\}_{t=1}^n$ with the closest approximation $\{\widetilde{q}(\gamma_{jt})\}_{t=1}^n$ in terms of KL divergence; that is, $\{\widetilde{q}(\gamma_{jt})\}_{t=1}^n$ leads to a smooth sequence of posterior inclusion probabilities, whose expected values coincide with the non-smooth estimates. Proposition 2.6 explains the procedure in details.



(a) Non-smooth estimates of $\boldsymbol{\mu}_{q(\beta)}$ and $(\mu_{q(\gamma_1)}, \ldots, \mu_{q(\gamma_n)})$



(b) Smooth estimates of $\boldsymbol{\mu}_{q(\beta)}$ and $(\mu_{q(\gamma_1)}, \ldots, \mu_{q(\gamma_n)})$

Figure 1: Smoothing the time-varying parameters $\boldsymbol{\mu}_{q(\beta_j)}$ and the posterior probability of inclusion $\mathbb{P}(\gamma_{jt} = 1)$ for $t = 1, \ldots, n$.

Proposition 2.6. *A smooth estimate for the trajectory of the inclusion probabilities can be achieved assuming* $\widetilde{q}(\boldsymbol{\gamma}_j) = \prod_{t=1}^{n} \widetilde{q}(\gamma_{jt})$ *such that* $\widetilde{q}(\gamma_{jt}) \equiv \mathsf{Bern}(\text{expit}(\mathbf{w}_t'\mathbf{f}_j))$ *with constraints on the mean. Therefore, the expectation of the joint vector* $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jn})'$ *is equal to* $\mathbb{E}_{\widetilde{q}}(\boldsymbol{\gamma}_j) = \mathbf{W}\mathbf{f}_j$, *where* $\mathbf{W}$ *is a* $n \times k$ *B-spline basis matrix. The optimal value of* $\mathbf{f}_j$ *is the solution of the optimization problem* $\widehat{\mathbf{f}}_j = \arg\max_{\mathbf{f}_j \in \mathbb{R}^k} \psi(\mathbf{f}_j)$ *where* $\psi(\mathbf{f}_j) = \sum_{t=1}^{n} \left[(\omega_{q(\gamma_{jt})} - \mathbf{w}_t'\mathbf{f}_j)\text{expit}(\mathbf{w}_t'\mathbf{f}_j) + \log(1 + \exp(\mathbf{w}_t'\mathbf{f}_j))\right]$, *such that the gradient is equal to* $\nabla_{\mathbf{f}}\psi(\mathbf{f}) = \sum_{t=1}^{n} \mathbf{w}_t(\omega_{q(\gamma_{jt})} - \mathbf{w}_t'\mathbf{f}) \frac{\text{expit}(\mathbf{w}_t'\mathbf{f})}{1+\exp(\mathbf{w}_t'\mathbf{f})}$.

Proof. See proof B.11 in Appendix B. □

The right panel of Figure 1(b) shows the smoothed estimates of the original probability of inclusion (left panel) based on Proposition 2.6. As a by-product of a smoother posterior inclusion probability, the dynamics of the corresponding regression coefficient is also regu-

larised, as shown in the left panel of Figure 1(b). Notice that the matrix $\mathbf{W}$ in Proposition 2.6 does not have to be an $n \times k$ B-spline basis matrix; our formulation is general and allows for alternative forms of smoothing, such as Daubechies wavelet basis functions (see, e.g., Bianchi et al., 2022a). Appendix B.1 discusses different smoothing assumptions in detail.

2.2 Theoretical properties of the estimation algorithm

We now present some key theoretical results and numerical properties of our variational Bayes estimation algorithm. In particular, we focus on the behavior of variational densities' updates from one iteration to the next of the optimization process. The iterative optimization to perform approximate posterior inference is sketched in Algorithm 1. The theoretical properties of the optimal density updates between two consecutive iterations are cumbersome to analyze when the system of equations from Proposition 2.1 to 2.5 hold simultaneously. Instead, we will analyze the limiting properties of Algorithm 1 as the inclusion probabilities tend to zero, i.e., sparsity inducing. Proposition 2.7 extends the main result of Ormerod et al. (2017) to the dynamic variable selection with time-varying regression coefficients.

Algorithm 1: Variational Bayes for dynamic sparse regression models.

Initialize: $q(\boldsymbol{\vartheta})$, $\Delta_{\boldsymbol{\vartheta}}$, A_ν, B_ν, A_η, B_η, A_ξ, B_ξ
while $\left(\widehat{\Delta}_{\boldsymbol{\vartheta}} > \Delta_{\boldsymbol{\vartheta}}\right)$ **do**
  for $j = 1, \ldots, p$ **do**
    Update $q(\mathbf{b}_j)$ as in 2.1; and $q(\eta_j)$ as in B.8;
    Update $q(\boldsymbol{\omega}_j)$ as in 2.3 and $q(\xi_j)$ as in B.9;
    for $t = 1, \ldots, n$ **do**
      Update $q(z_{jt})$ as in B.7;
      Update $q(\gamma_{jt})$ as in 2.2 (non-smooth) or 2.6 (smooth);
    end
  end
  Update $q(\mathbf{h})$ as in 2.5 (heteroskedastic) or $q(\sigma^2)$ as in B.6 (homoskedastic);
  Update $q(\nu^2)$ as in B.10;
  Compute $\widehat{\Delta}_{\boldsymbol{\vartheta}} = q(\boldsymbol{\vartheta})^{(\text{iter})} - q(\boldsymbol{\vartheta})^{(\text{iter}-1)}$;
end

Proposition 2.7. *Assume that the maximum over time of the inclusion probabilities, for a given variable j, at the i-th iteration of the algorithm is such that $\max_{t\in\{1,\ldots,n\}} \mu^{(i)}_{q(\gamma_{jt})} = \epsilon$, and $\epsilon \ll 1$ is small enough. Moreover, let $\boldsymbol{\Sigma}^{(i)}_{q(\omega_j)} - \boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)} \geq 0$, then:*

1. $\mu^{(i+1)}_{q(\gamma_{jt})} = \text{expit}\left\{\mu^{(i+1)}_{q(\omega_{jt})} - \frac{1}{2}\mu^{(i+1)}_{q(1/\sigma^2_t)} x^2_{jt-1} \mu^{-1(i+1)}_{q(1/\eta^2_j)} q_{tt} + O(\epsilon)\right\}$, $q_{tt} = [\mathbf{Q}^{-1}]_{tt} \geq 0$;

2. $\mu^{(i+1)}_{q(\omega_{jt})} = -1/2 \sum_{k=1}^{n} s_{tk} + O(\epsilon)$, $s_{tk} = [\mathbf{\Sigma}_{q(\omega_j)}]_{tk} \geq 0$;

3. $\mu^{(i+1)}_{q(\omega_{jt})} \leq \mu^{(i)}_{q(\omega_{jt})}$ *decreases after each iteration.*

Proof. See proof C.1 in Appendix C.[5] □

Proposition 2.7 and Lemma 1[6] in Ormerod et al. (2017) leads to two key numerical results: first, for $\epsilon \approx 0$, the following approximation for the update of the inclusion probabilities holds:

$$\mu^{(i)}_{q(\gamma_{jt})} \approx \text{expit}\left\{\mu^{(i+1)}_{q(\omega_{jt})} - 1/2\mu^{(i+1)}_{q(1/\sigma^2_t)} x^2_{jt-1} \left[\mu^{(i+1)}_{q(1/\eta^2_j)}\right]^{-1} q_{tt}\right\}. \tag{19}$$

This implies that for $\mathrm{M}^{(i)} = \arg\max_{t\in\{1,\ldots,n\}} \mu^{(i)}_{q(\omega_{jt})} \ll 0$, after i iterations, the sequence $\{\mu^{(i)}_{q(\gamma_{jt})}\}^n_{t=1}$ is indistinguishable from zero. As a result, our algorithm concentrates the posterior densities to a point mass at zero for all t. Second, if $\mu^{(i)}_{q(\gamma_{jt})} \approx 0$, $\forall t$, then all successive updates $i_k \geq i$ imply $\mu^{(i_k)}_{q(\gamma_{jt})} \approx 0$ since $\mu^{(i_k)}_{q(\omega_{jt})} \leq \mu^{(i)}_{q(\omega_{jt})}$ and therefore the updates $\mathrm{M}^{(i_k)} \leq \mathrm{M}^{(i)}$.

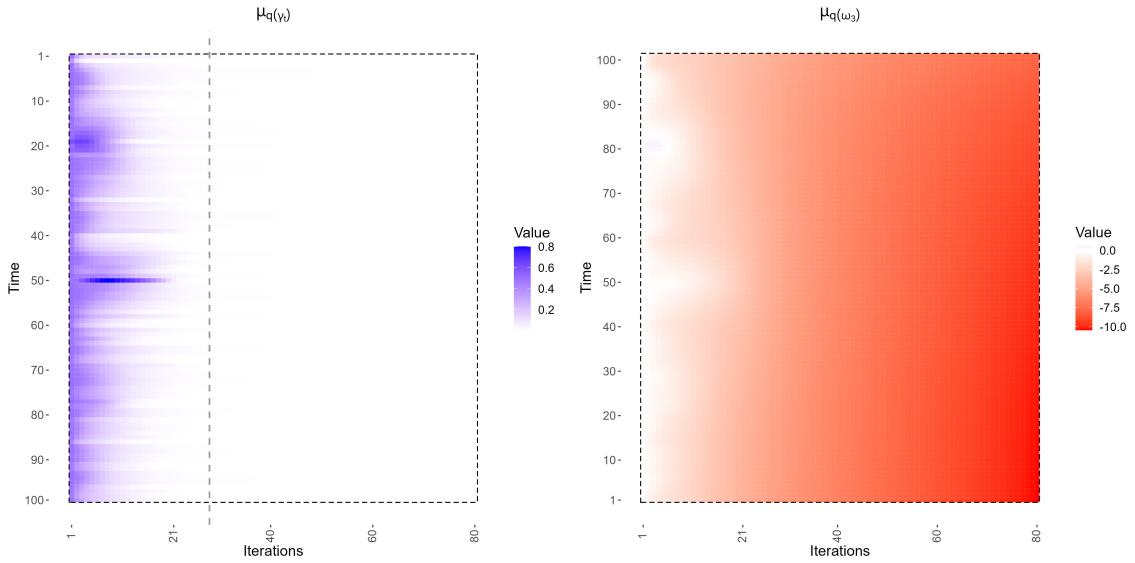


Figure 2: Left panel shows variational update over iterations (x-axis) until convergence of the vector of posterior inclusion probabilities $(\mu_{q(\gamma_{j1})}, \ldots, \mu_{q(\gamma_{jn})})$ (y-axis), for a parameter j which is always zero $\forall t$. The dashed line identifies the iteration at which the conditions of Proposition 2.7 are satisfied for $\epsilon = 0.01$. The right panel depicts the decreasing behaviour of $\mu_{q(\omega_{jt})}$, $\forall t$.

[5] For the ease of exposition, we leave the interested reader to Appendix C for some of the definitions and lemmas which are instrumental for the proof.

[6] Let $a \in \mathbb{R}^+$, then, as $a \to +\infty$, the following expansions hold: $\text{expit}(-a) = \exp(-a) + O(\exp(-2a))$ and $\text{expit}(a) = 1 - \exp(-a) + O(\exp(-2a))$.

Figure 2 provides a visual representation of Proposition 2.7 for a simple simulation set up in which a given predictor is never included in the model (see Section 3.1). The dashed line identifies the iteration at which the conditions in Proposition 2.7 are satisfied for $\epsilon = 0.01$. After few iterations $\mu_{q(\gamma_{jt})}$ remains zero $\forall t$. As a result, the corresponding j-th predictor can be deleted from the regression specification. This result provides a dimension reduction strategy which is summarised in Algorithm 2. Specifically, we can remove the j-th variable from the set of predictors during the estimation. Such automatic exclusion strategy improves the computational efficiency when p increases, but the signal $\bar{p} \leq p$ remains constant, where $\bar{p} = \mathsf{card}(\mathcal{J})$ and the set $\mathcal{J} = \{j : \sum_{t=1}^{n} \gamma_{jt} > 0\}$ collects the indexes of regression coefficients that are included in the model at least for one t.

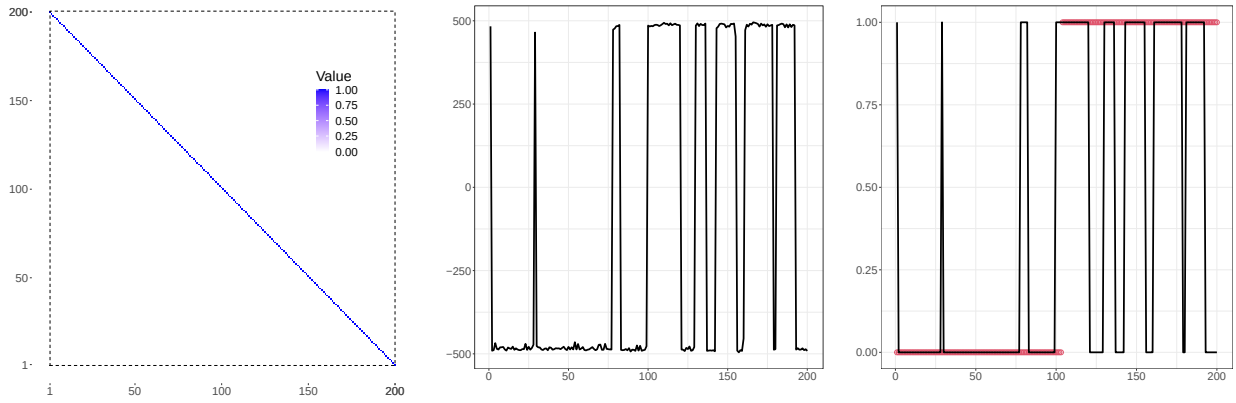
Algorithm 2: Efficient variational Bayes for dynamic sparse regression models.

Initialize: $q(\boldsymbol{\vartheta})$, $\Delta_{\boldsymbol{\vartheta}}$, A_ν, B_ν, A_η, B_η, A_ξ, B_ξ
while $\left(\widehat{\Delta}_{\boldsymbol{\vartheta}} > \Delta_{\boldsymbol{\vartheta}}\right)$ **do**
  for $j = 1, \ldots, p$ **do**
    Update $q(\mathbf{b}_j)$ as in 2.1; and $q(\eta_j)$ as in B.8;
    Update $q(\boldsymbol{\omega}_j)$ as in 2.3 and $q(\xi_j)$ as in B.9;
    for $t = 1, \ldots, n$ **do**
      Update $q(z_{jt})$ as in B.7;
      Update $q(\gamma_{jt})$ as in 2.2 (non-smooth) or 2.6 (smooth);
    end
  end
  Update $q(\boldsymbol{\sigma})$ as in B.1 (heteroskedastic) or B.6 (homoskedastic);
  Update $q(\nu^2)$ as in B.10;
  if *assumptions in 2.7 hold* **then**
    for $j = 1, \ldots, p$ **do**
      if $\max_t\{\mu_{q(\gamma_{jt})}\} < \epsilon$ **then**
        Drop the j-th variable
      end
    end
  end
  Compute $\widehat{\Delta}_{\boldsymbol{\vartheta}} = q(\boldsymbol{\vartheta})^{(\mathsf{iter})} - q(\boldsymbol{\vartheta})^{(\mathsf{iter}-1)}$;
end

Hyper-parameters and algorithm initialization. In this section we focus on the key hyper-parameters and initialization choices. As far as the inclusion probabilities are concerned, we follow Koop and Korobilis (2020) and set $\mu^{(0)}_{q(\gamma_{jt})} = 1/2$, $\forall t, j$. Next, we follow Ormerod et al. (2017) and set $A_\sigma = B_\sigma = A_\eta = B_\eta = 0.01$ to maintain non-informativeness.

Notably these are all fairly standard choices for conjugate priors. Two crucial hyper-parameters that deserve a more careful scrutiny are the couple (A_ξ, B_ξ); this is because a key property of our dynamic sparse regression model is the time-variation of $\gamma_{jt}|\omega_{jt}$, where the dynamics of the stochastic process ω_{jt} is governed by the conditional variance $\xi_j^2 \sim \mathsf{IG}(A_\xi, B_\xi)$.

In what follows we study the resulting variational mean and variance of $\{\omega_{jt}\}_{t=1}^n$, namely $\{\mu_{q(\omega_{jt})}\}_{t=1}^n$ and $\mathbf{\Sigma}_{q(\omega_j)}$, for $j = 1, \dots, p$, for three alternative limit cases. In addition, since γ_{jt} directly depends on ω_{jt}, we further show how this reflects on the posterior inclusion probability trajectory $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$, for $j = 1, \dots, p$. These comparative statics are based on Proposition 2.3 and Proposition B.9 in the Appendix, and allows us to provide a transparent strategy to select meaningful values of the couple (A_ξ, B_ξ). The first scenario considers A_ξ constant and $B_\xi \to +\infty$.

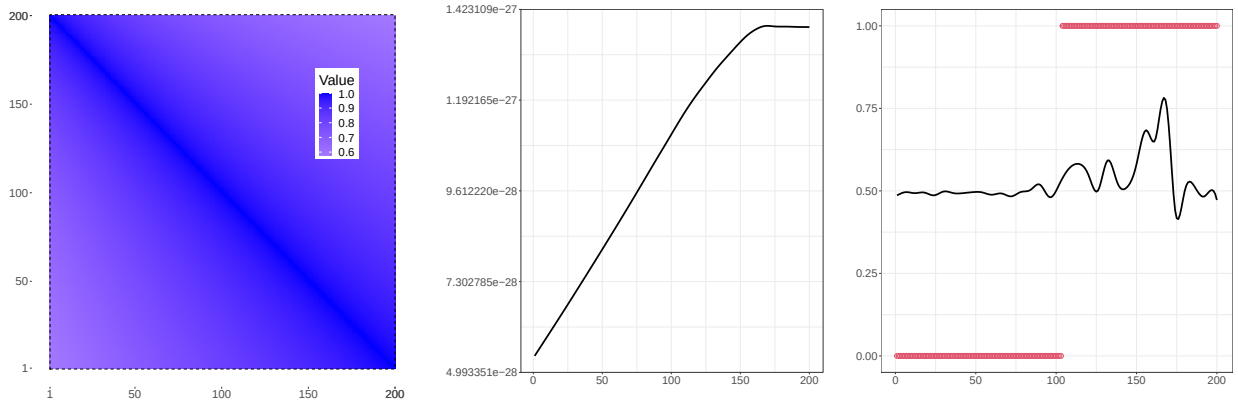


(a) (b) (c) $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$

Figure 3: Scenario A: $B_\xi \to +\infty$. (a) Depicts the variational correlation matrix for the process $\{\omega_{jt}\}_{t=1}^n$ obtained from $\mathbf{\Sigma}_{q(\omega_j)}$. (b) Plots the trajectory of $\{\mu_{q(\omega_{jt})}\}_{t=1}^n$. (c) Shows the effect on the posterior inclusion probabilities $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$ compared to the simulated (red points).

Figure 3 reports the resulting variational covariance matrix $\mathbf{\Sigma}_{q(\omega_j)}$ and the corresponding trajectory of $\{\mu_{q(\omega_{jt})}\}_{t=1}^n$ and posterior estimates of the inclusion probabilities $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$. As $B_\xi \to +\infty$, the process $\{\omega_{jt}\}_{t=1}^n$ tends to be i.i.d – $\mathbf{\Sigma}_{q(\omega_j)}$ tends to a diagonal matrix. This means that we lose the time dependence in the *a-priori* inclusion probability process, which leads to a highly erratic dynamics of ω_{jt} and, as a result, a highly irregular trajectory of $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$. The second scenario considers $A_\xi \to +\infty$ and B_ξ constant. This implies that $\mu_{q(1/\xi_j^2)} \to +\infty$ and, as a consequence, we give infinite weight to the matrix $\mathbf{Q}$ when compute $\mathbf{\Sigma}_{q(\omega_j)}$ (see Proposition 2.3). As shown by Figure 4, such strong and informative time dependence in the *a-priori* inclusion probability process, leads to posterior inclusion prob-

abilities $\{\mu_{q(\gamma_{jt})}\}_{t=1}^{n}$ with low variability around the marginal mean of the process $\{\omega_{jt}\}_{t=1}^{n}$, i.e. $\text{expit}(\mathbb{E}(\omega_{jt})) = 0$. As a result, no sparsity is captured despite being present in the underlying data generating process.



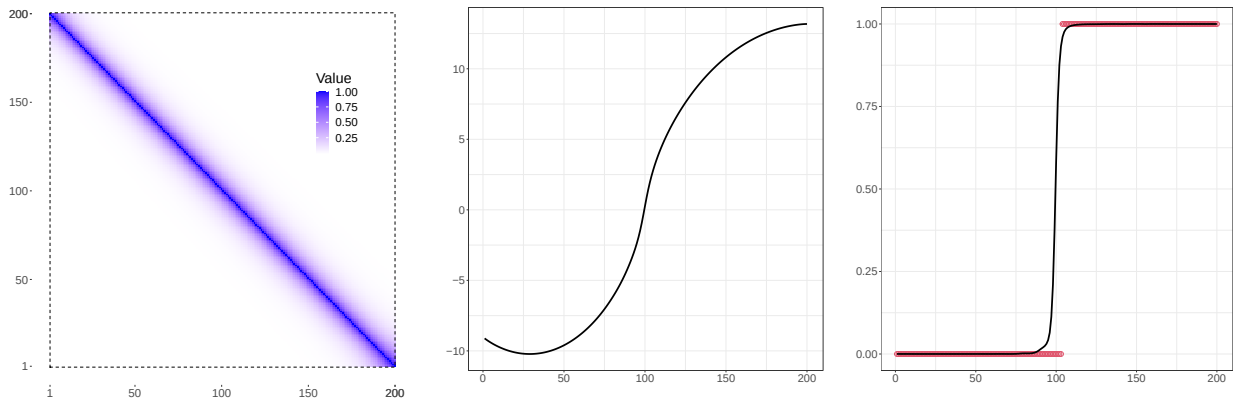
(a) (b) (c) $\{\mu_{q(\gamma_{jt})}\}_{t=1}^{n}$

Figure 4: Scenario B: $A_\xi \to +\infty$. (a) Depicts the variational correlation matrix for the process $\{\omega_{jt}\}_{t=1}^{n}$ obtained from $\boldsymbol{\Sigma}_{q(\omega_j)}$. (b) Plots the trajectory of $\{\mu_{q(\omega_{jt})}\}_{t=1}^{n}$. (c) Shows the effect on the posterior inclusion probabilities $\{\mu_{q(\gamma_{jt})}\}_{t=1}^{n}$ compared to the simulated (red points).

The last scenario considers $A_\xi/B_\xi \to c_1$, where $c_1 \in \mathbb{R}^+$ constant. This implies that $\mu_{q(1/\xi_j^2)} \to c_2$, where $c_2 \in \mathbb{R}^+$ and therefore we give moderate weight to the matrix $\mathbf{Q}$ when compute $\boldsymbol{\Sigma}_{q(\omega_j)}$ (see Proposition 2.3) – i.e, we account for a decreasing correlation as $|t_1 - t_2|$, $t_1, t_2 \in \{1, \dots, n\}$ increases. As shown by Figure 5, this translates into a moderate variability in the time dependence in the *a-priori* inclusion probability process, which leads to posterior estimates $\{\mu_{q(\gamma_{jt})}\}_{t=1}^{n}$ that accurately track the underlying dynamics of sparsity. In the following, we propose to fix $A_\xi = 2$ so that, *a-priori*, $\mathsf{Var}(\xi_j^2) = +\infty$ and B_ξ can be directly interpreted as the mean of ξ_j^2. More specifically, to estimate our dynamic sparse regression model both in the simulation study and the empirical analysis, we set $B_\xi = 5$, which satisfies $A_\xi/B_\xi \to c_1$. We also test in simulation $B_\xi = 1$ or $B_\xi = 10$ as shown in Section 3.2. The model performance are broadly consistent for $B_\xi = 5$ and $B_\xi = 1$, while slightly deteriorates for $B_\xi = 10$.

3 Simulation study

We now perform an extensive simulation study to evaluate the properties of our estimation framework in a controlled setting. We first compare our variational Bayes method against



(a) (b) (c) $\{\mu_{q(\gamma_{jt})}\}_{t=1}^{n}$

Figure 5: Scenario C: $A_\xi/B_\xi \to c_1$, $c_1 \in \mathbb{R}^+$. (a) Depicts the variational correlation matrix for the process $\{\omega_{jt}\}_{t=1}^{n}$ obtained from $\boldsymbol{\Sigma}_{q(\omega_j)}$. (b) Plots the trajectory of $\{\mu_{q(\omega_{jt})}\}_{t=1}^{n}$. (c) Shows the effect on the posterior inclusion probabilities $\{\mu_{q(\gamma_{jt})}\}_{t=1}^{n}$ compared to the simulated (red points).

an equivalent MCMC estimation algorithm. This allows to compare on an equal footing both inference approaches. Second, we compare the estimation performance of our dynamic sparse regression model against a variety of alternative static and dynamic variable selection methods. A particular emphasis will be put on the ability of each method across different model dimensions in identifying those predictors which enter and leave the model set, vis-a-vis those predictors who are either never or always in the model specification.

3.1 Variational Bayes vs MCMC

The data augmentation approach based on the Polya-Gamma representation in Eq.(7) has the main advantage to lead to a more tractable joint distribution $p(\mathbf{y}, \boldsymbol{\vartheta})$. This aspect is crucial to derive an efficient MCMC scheme for Bayesian inference. Appendix A provides a summary of the full conditionals equivalent to our variational Bayes approximation approach. This allows to make a coherent comparison between our VB and its MCMC counterpart. To this aim, we compare the posterior accuracy as proposed by Wand et al. (2011):

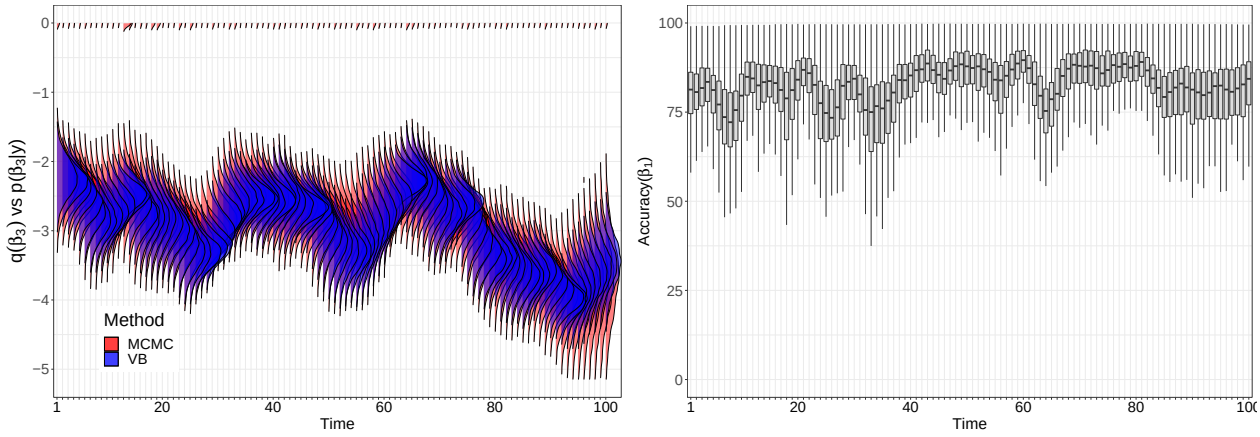
$$\mathcal{ACC}(\boldsymbol{\vartheta}) = \left\{1 - 0.5 \int |q(\boldsymbol{\vartheta}) - p(\boldsymbol{\vartheta}|\mathbf{y})|\, d\boldsymbol{\vartheta}\right\} \%, \tag{20}$$

where $\boldsymbol{\vartheta}$ is a parameter of interest, $q(\boldsymbol{\vartheta})$ is the variational density and $p(\boldsymbol{\vartheta}|\mathbf{y})$ denotes the posterior distribution sampled via MCMC. Note that the evaluation of the variational Bayes approximation compared to MCMC is fraught with difficulty. An accurate comparison is

hampered by the difficulty to determine the convergence whether the MCMC scheme converged to its stationary distribution. In addition, one can arbitrarily trade accuracy with speed in both an MCMC and in a VB context. With this in mind, we consider the same hyper-parameters for both the VB and MCMC approaches and choose an arbitrarily large number of draws, so that comparison between methods focuses on accuracy.

The simulation is set up as follows. We consider $p = 3$ and $n = 100$, and generate $\{\boldsymbol{\beta}_t\}_{t=1}^{100}$ with $\boldsymbol{\beta}_t = (\beta_{1t}, \beta_{2t}, \beta_{3t})'$ such that β_{1t} is a time-varying parameter always included in the model, β_{2t} is set equal to zero for all the time periods, while β_{3t} shows a dynamic sparsity pattern. Then, we generate $N = 100$ replicates from $y_t = \sum_{j=1}^{3} \beta_{jt} x_{jt-1} + \varepsilon_t$, with $\varepsilon_t \sim \mathsf{N}(0, 0.25)$ and x_{jt} generated from a standard normal for $t = 1, \ldots, 100$. Notice that for the purpose of comparing the accuracy of our VB versus its MCMC counterpart, the small dimension p has a limited impact on the validity of the results. A small-scale time-varying parameter regression retains the same key properties we want to investigate in terms of dynamic sparsity, with the main advantage of speeding up the MCMC computation.

For each simulated parameter we report the overlapping posterior densities for one selected replicate of β_{jt}, $j = 1, 2, 3$ obtained via VB (blue) vs MCMC (red). In addition, we report a box-chart for each time t representing the accuracy $q^*(\beta_{jt})$ with respect to the MCMC $p(\beta_{jt}|\mathbf{y})$ as per Eq.(20) across simulations. For the sake of brevity, we leave additional results for $q^*(b_{jt})$ and $q^*(\gamma_{jt})$ vs their MCMC equivalent to Appendix D.1.



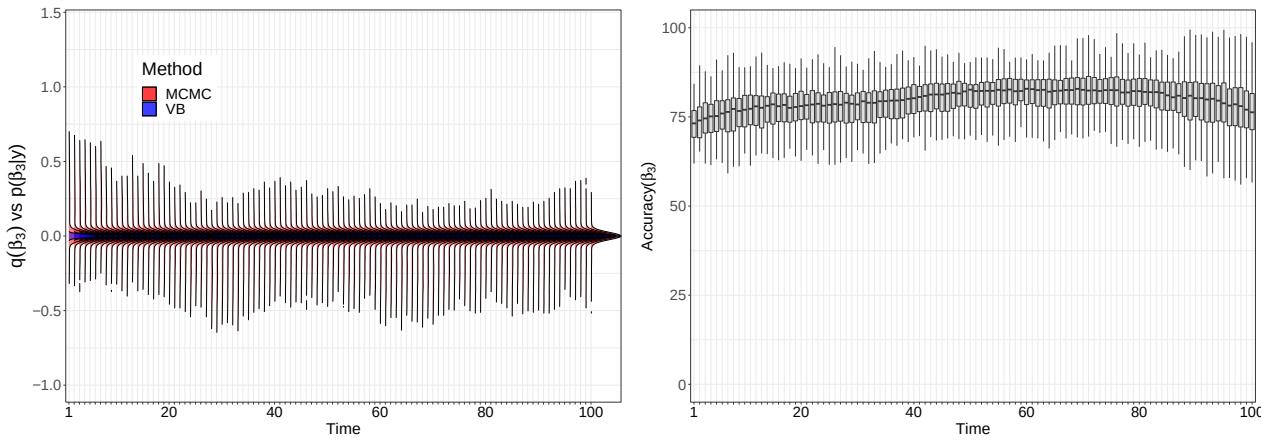
(a) β_{1t} for one simulated path

(b) Accuracy across simulations

Figure 6: Comparison with MCMC when β_{1t} is a time-varying parameter with $\gamma_{1t} = 1$, for all t. Left panel shows the overlapping posterior densities of β_{1t} obtained via VB (blue) and MCMC (red), for one selected replicate. Right panel shows the accuracy over time of $q^*(\beta_{1t})$ compared to $p(\beta_{1t}|\mathbf{y})$ across simulations.

Figure 6(a) shows that $q^*(\beta_{1t})$ is slightly more concentrated around the posterior mean compared to its MCMC equivalent. This is a by-product of the nature of the approximating density (see Proposition 2.4): the weights in the mixture defining $q^*(\boldsymbol{\beta}_1)$ are such that $w_s = 1$ if $s = (1, 1, \ldots, 1)$ and $w_s = 0$ otherwise. Thus, we only keep one component of the mixture. This is not the case for MCMC draws which still sample from a Gaussian distribution when $\gamma_{1t} = 0$. Nevertheless, the accuracy of $q^*(\beta_{1t})$ in approximating $p(\beta_{1t}|\mathbf{y})$ is as high as 80% as shown by Figure 6(b).

Next, we consider the case in which a given predictor is always excluded over time. Figure 7(a) highlights that VB provides posterior inclusion probabilities tight around zero, as highlighted in Proposition 2.7. The weights in the mixture $q^*(\boldsymbol{\beta}_2)$ are $w_s = 1$ if $s = (0, 0, \ldots, 0)$ and $w_s = 0$ otherwise. Hence, we only keep the component of the mixture that identifies a sequence of Dirac at zero $\delta_0(\beta_{2t})$. On the other hand, MCMC draws show a much lower concentration of the posterior probability mass at zero. This is reflected in a relatively lower overlapping – around 75% accuracy – of the posterior density $q^*(\beta_{2t})$ compared to $p(\beta_{2t}|\mathbf{y})$ across simulations, as shown by Figure 7(b). The right panels in Figure D.4 report a similar accuracy for both b_{2t} and γ_{2t}, respectively.



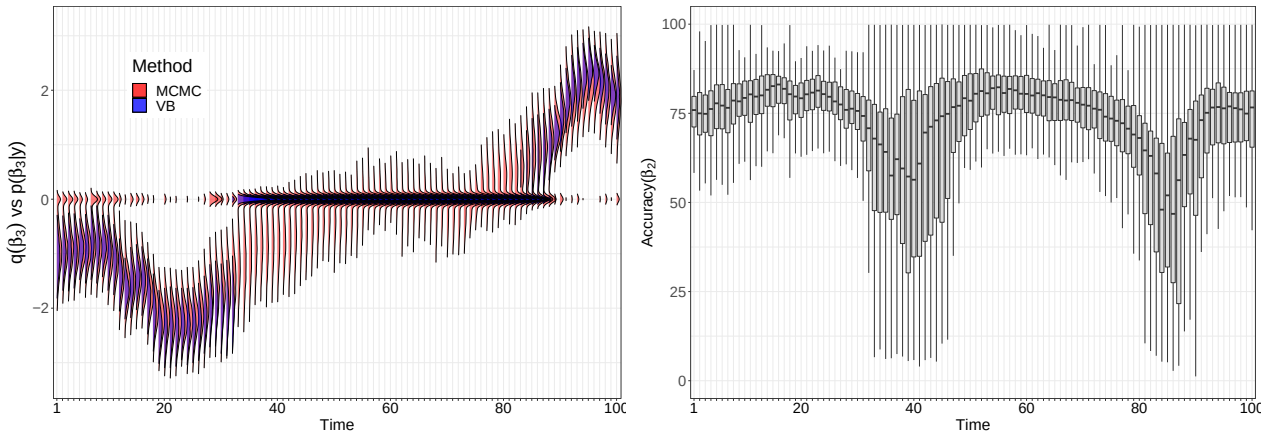
(a) β_{2t} for one simulated path

(b) Accuracy across simulations

Figure 7: Comparison with MCMC when β_{2t} is a coefficient constant at zero, i.e. $\gamma_{2t} = 0$, for all t. Left panel shows the overlapping posterior densities of β_{2t} obtained via VB (blue) and MCMC (red), for one selected replicate. Right panel shows the accuracy over time of $q^*(\beta_{2t})$ compared to $p(\beta_{2t}|\mathbf{y})$ across simulations.

Finally, we compare the accuracy of our variational Bayes inference against MCMC for the time-varying parameter β_{3t} which displays a pattern of dynamic sparsity. Figure 8(a) depicts a tight approximation to MCMC draws during periods in which the coefficient is

unambiguously included in the model (initial and final part), while the densities overlap between VB and MCMC deteriorates when $\gamma_{3t} \to 0$ as in the middle part of the sample. This confirms what highlighted in Figures 6(b) and 7(b): when the probability of inclusion approaches one, the posterior densities of our VB and the equivalent MCMC tend to overlap almost entirely. On the other hand, when there is less certainty on the inclusion of a given predictor, or outright certainty of exclusion, the MCMC posterior draws tend to be less concentrated on the actual inclusion probabilities. The middle panels in Figure D.4 show that is also applies for b_{3t} and γ_{3t}.



(a) β_{3t} for one simulated path

(b) Accuracy across simulations

Figure 8: Comparison with MCMC when β_{3t} is a coefficient that shows dynamic sparsity. Left panel shows the overlapping posterior densities of β_{3t} obtained via VB (blue) and MCMC (red), for one selected replicate. Right panel shows the accuracy over time of $q^*(\beta_{3t})$ compared to $p(\beta_{3t}|\mathbf{y})$ across simulations.

3.2 Comparison with existing variable selection methods

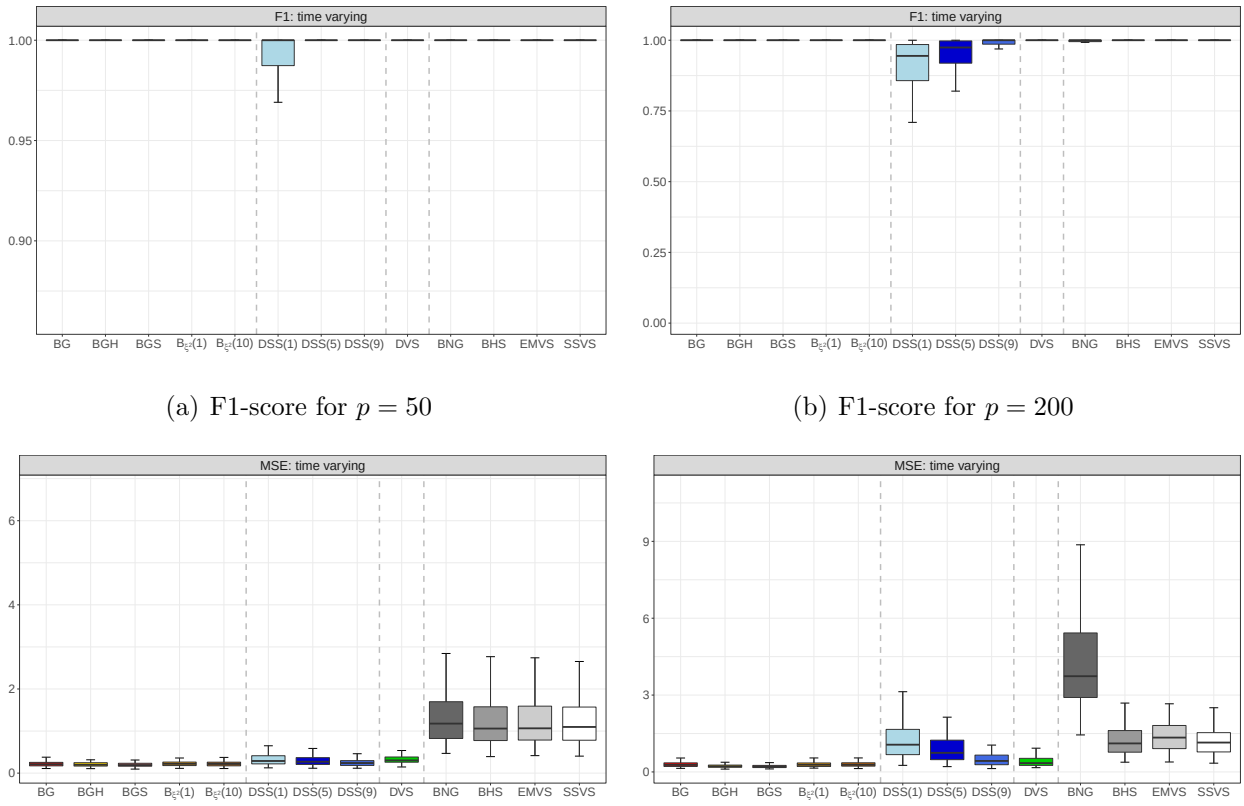
We now perform an extensive simulation study to compare our dynamic sparse regression model as outlined in Section 2 against a variety of established Bayesian static and dynamic variable selection methods. As far as the data generating process is concerned, we consider $M = 100$ replicates from the following data generating process $y_t = \sum_{j=1}^{p} \beta_{jt} x_{jt-1} + \varepsilon_t$ with $\varepsilon_t \sim \mathsf{N}(0, 0.25)$ and $\{x_{jt}\}_{j=1}^{p}$ are independently generated at each time $t = 1, \ldots, n$ from a standard normal distribution. Consistent with the empirical application we set the length of the time series $n = 200$ and $p \in \{50, 100, 200\}$. We assume that different coefficients have different dynamics; for instance, β_{1t} is a time-varying parameter which is always included in the model, i.e. $\gamma_{1t} = 1 \ \forall t$, $\beta_{2:7,t}$ show different types of dynamic sparsity – which will be

discussed later – and $\beta_{8:p,t}$ is set to zero for all t, i.e. $\gamma_{8:p,t} = 0 \; \forall t$. For the sake of brevity, in the following we focus on accuracy for regression models with $p = 50, 200$ predictors. The results for the case $p = 100$ are reported in Appendix D.2. The latter also reports examples of the simulated trajectories for each of the parameters discussed in the results.

We implement different versions of our dynamic Bernoulli-Gaussian `BG` model. Hereafter, `BGH` is the homoskedastic alternative to `BG`, and `BGS` performs smoothing on the posterior inclusion probabilities as highlighted in Section 2.1. Furthermore, we also consider our main `BG` algorithm with fixed hyper-parameters $B_\xi = 1, 10$ for the variance of the latent process ω_{jt}. We compare our dynamic sparse regression model against a series of established static sparsity inducing priors, which arguably represent the workhorse in Bayesian inference in linear regressions. We consider two continuous shrinkage priors, i.e. the normal-gamma of Griffin and Brown (2010) (`BNG`) and the horseshoe of Carvalho et al. (2010) (`BHS`), as well as the mixture of Gaussians proposed by George and McCulloch (1993) (`SSVS`) and EM spike-and-slab of Ročková and George (2014) (`EMVS`). We follow existing literature, such as Huber et al. (2021); Bianchi et al. (2022b), and use the signal adaptive variable selector (SAVS) of Ray and Bhattacharya (2018) as post-processing tool to induce sparsity in the posterior estimates from the hierarchical shrinkage priors `BNG` and `BHS`. To mimic a time-varying behavior we estimate each model based on a recursive rolling window of 100 observations. Finally, we consider two recent advancements towards dynamic variable selection in large-scale regressions, such as the dynamic spike-and-slab specification of Koop and Korobilis (2020) (`DVS`) and Ročková and McAlinn (2021) (`DSS`). The latter is estimated with three different values of the *marginal importance weight* parameter $\Theta \in \{0.1, 0.5, 0.9\}$.

We compare all models based on both point estimation accuracy and their ability to identify which predictor is significant in a given time period. Point accuracy is measured by the mean-squared error (MSE), which represents the squared distance between the true parameters $\beta_{jt}, \; t = 1, \dots, n$ observed at each simulation and its corresponding posterior estimate $\widehat{\beta}_{jt}$. As a measure of identification accuracy we quantify the type I vs type II error in variables selection from the F1-score (see, e.g., Bianchi et al., 2022b). This provides a direct assessment of the ability to correctly classify a predictor as "active" at time t.

Top panels of Figure 9 report the F1-scores for the time-varying parameter β_{1t} which follows an AR(1) process with persistence equal to 0.98 and conditional variance equal to 0.1. An example of the simulated trajectory is reported in Figure D.5 in Appendix D.2. With the partial exception of `DSS(1)`, all models provide an accurate identification of $\gamma_{1t} = 1, \; \forall t$.



(a) F1-score for $p = 50$

(b) F1-score for $p = 200$

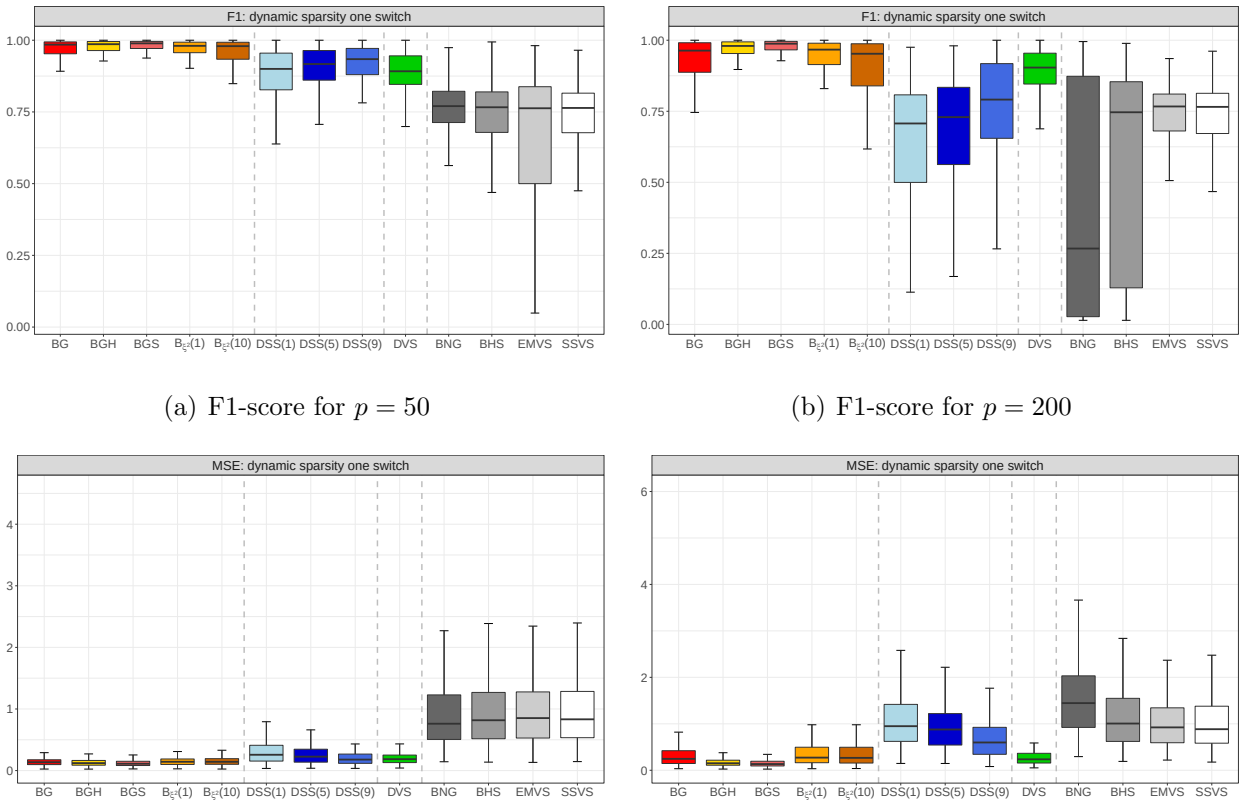
(c) MSE for $p = 50$

(d) Frobeniun norm for $p = 200$

Figure 9: Estimation accuracy for β_{1t}. Top panels report the F1-score for $p = 50$ (left) and $p = 200$ (right). Bottom panels report the MSE for $p = 50$ (left) and $p = 200$ (right).

Perhaps not surprisingly, when it comes to point estimation, a rolling window approach is less accurate in modeling the pure dynamics of the time-varying parameter. That is, the squared estimation error from all static variable selection methods is higher.

Figure 10 shows the point estimation accuracy and the F1-score for the $\beta_{2:3,t}$ which involves one switch from $\gamma_{2:3,t} = 0$ to $\gamma_{2:3,t} = 1$. Specifically, the parameter is generated by dividing the interval in sub-periods $[1, n] = [1, t_1] \cup [t_1 + 1, t_1 + t_2] \cup ... \cup [t_1 + ... + t_n + 1, n]$, where $t_k \sim \mathsf{Pois}(n/2)$, so that the expected number of sub-periods is 2, and then randomly alternate periods where $\gamma_{jt} = 0$ and $\gamma_{jt} = 1$. For the intervals where $\gamma_{jt} = 1$ we generate an AR(1) process as for β_{1t}. This represents a "structural break" type of scenario in which the estimation accuracy broadly deteriorates. Nevertheless, our `BG`, `BGS` and `BGH` approaches outperform all competing methods. This applies for both $p = 50$ and $p = 200$. On the contrary, the F1-score across the other competing methods substantially deteriorates, particularly for



(a) F1-score for $p = 50$

(b) F1-score for $p = 200$

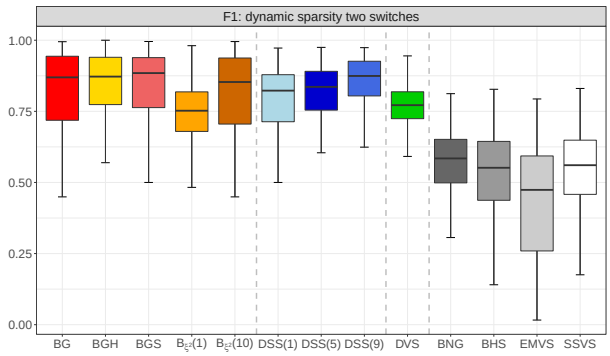
(c) MSE for $p = 50$

(d) Frobeniun norm for $p = 200$

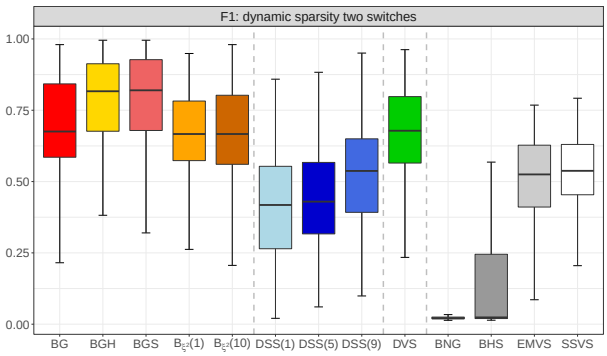
Figure 10: Estimation accuracy for $\beta_{2:3,t}$. Top panels report the F1-score for $p = 50$ (left) and $p = 200$ (right). Bottom panels report the MSE for $p = 50$ (left) and $p = 200$ (right).

a large-scale regression ($p = 200$) and static variable selection methods (see top-right panel). This translates into a visible reduction in the point estimation accuracy, as shown by the MSE in the bottom panels.

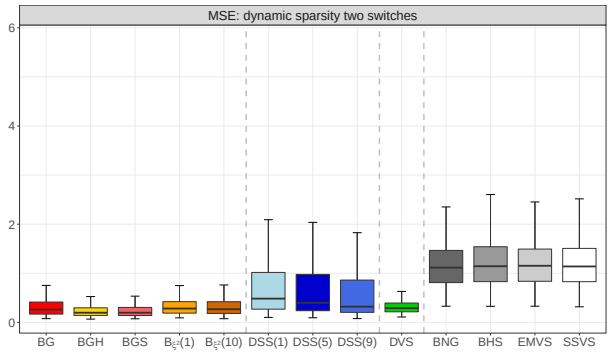
Figure 11 shows the point estimation accuracy and the F1-score for the $\beta_{4:5,t}$ which involves two switches from $\gamma_{4:5,t} = 0$ to $\gamma_{4:5,t} = 1$ and vice-versa. Specifically, the parameter is generated as follows by divide the interval in sub-periods as for $\beta_{2:3,t}$, but set $t_k \sim \mathsf{Pois}(n/4)$, so that the expected number of sub-periods is 4, and then randomly alternate periods where $\gamma_{jt} = 0$ and $\gamma_{jt} = 1$. For the intervals where $\gamma_{jt} = 1$ the process is an AR(1) as for β_{1t}. The results suggest that a more a volatile dynamics broadly poses extra challenges for parameters identification. This is particularly detrimental for static variable selection methods estimated based on a rolling window. For instance, for $p = 200$, the static and dynamic spike-and-slab methods `SSVS`, `EMWS`, `DSS(1)` and `DSS(5)` generate a rather dismal average F1-score of



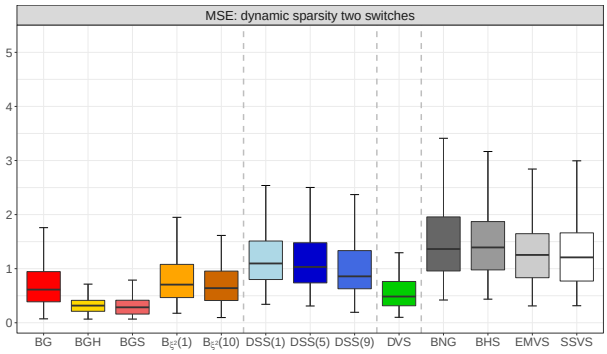
(a) F1-score for $p = 50$



(b) F1-score for $p = 200$



(c) MSE for $p = 50$



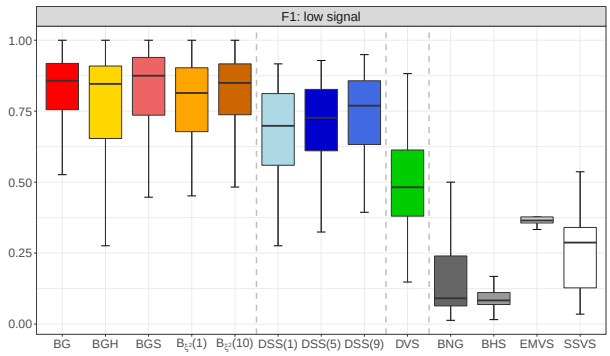
(d) Frobeniun norm for $p = 200$

Figure 11: Estimation accuracy for $\beta 4 : 5, t$. Top panels report the F1-score for $p = 50$ (left) and $p = 200$ (right). Bottom panels report the MSE for $p = 50$ (left) and $p = 200$ (right).

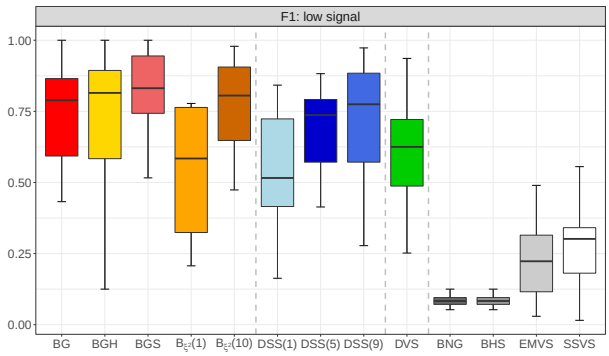
0.5. This translates in a visibly higher mean squared error in point estimates, as shown by the MSE. Nevertheless, our dynamic `BG` model achieve a visibly more accurate parameter identification and estimation accuracy, in particular for $p = 200$ and for `BGS`.

Next, we consider a short-lived signal with $\beta_{6:7,t}$. Specifically, the dynamics of the parameter is generated by sampling an interval length $\Delta_i \sim \mathsf{Pois}(n/10)$ and place it at random on the timeline such that $\gamma_{jt} = 1$ in that period, then generate a trajectory for the coefficient as for β_{1t}. This constitutes a rather extreme case in which a predictor is significant only for a very short period of time. Figure 12 broadly confirms that shorter signals are more complicated to extract, particularly for the static variable selection methods, with a median F1-score of 0.25 for both spike-and-slab priors for both $p = 50$ and $p = 200$.

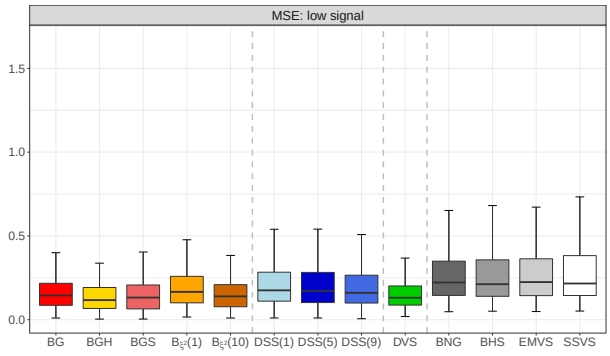
This is because rolling window procedure only tangentially capture sudden changes, and therefore strongly under-performs in terms of identification. This also holds with $p = 200$



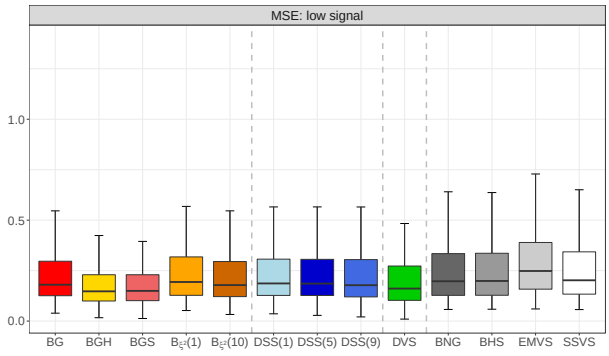
(a) F1-score for $p = 50$



(b) F1-score for $p = 200$



(c) MSE for $p = 50$

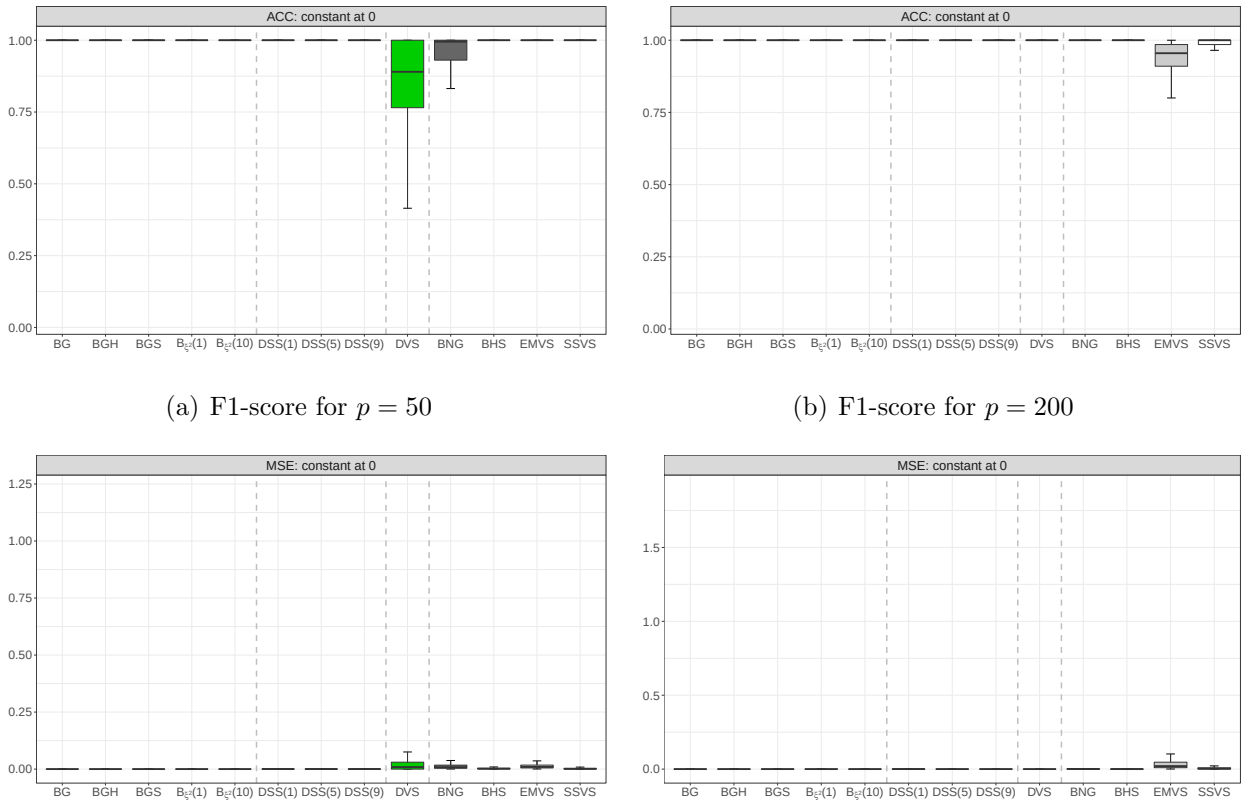


(d) Frobeniun norm for $p = 200$

Figure 12: Estimation accuracy for $\beta_{6:7,t}$. Top panels report the F1-score for $p = 50$ (left) and $p = 200$ (right). Bottom panels report the MSE for $p = 50$ (left) and $p = 200$ (right).

when we impose a very tight variance to ω_{jt} via B_ξ. Overall, our dynamic `BGS` model ranks best in terms of accuracy of signal identification. Notably, the poor performance of the classification does not affect the accuracy of the posterior estimates. This is due to the short-live nature of the $\beta_{6:7,t}$; that is, a large mis-classification for a short period is likely diluted by a good performance when the parameter is zero. This is confirmed by Figure 13, which reports the results for $\beta_{8:p,t} = 0$, $\forall t$. Within this setting, there is no signal to identify, therefore the F1-scores metric is replaced by the classification accuracy (ACC). All models provide good results in terms of identification and estimation accuracy, with the exception of `DVS` for $p = 50$ and `EMVS` for $p = 200$.

Appendix D.2 provides additional results testing two key dimensions of our dynamic variable selection method: robustness to correlated signals and computational speed. As far as the robustness to correlation is concerned, Figures 8(b)-8(a) in Section D.3 show that



(a) F1-score for $p = 50$

(b) F1-score for $p = 200$

(c) MSE for $p = 50$

(d) Frobeniun norm for $p = 200$

Figure 13: Estimation accuracy for $\beta_{8:p,t}$. Top panels report the F1-score for $p = 50$ (left) and $p = 200$ (right). Bottom panels report the MSE for $p = 50$ (left) and $p = 200$ (right).

with the exception of extremely high level of autocorrelation and cross-sectional correlation, our dynamic variable selection method is reasonably robust. The median F1-score tend to marginally deteriorate for high level of auto-correlation and/or cross-sectional correlations. This is more evident in the context of highly correlated, short-lived, predictors. However, the results suggest that our dynamic `BG` captures a great deal of sparsity in the dynamics of the regression coefficients independently on the auto- and/or cross-correlation assumptions.

In addition, Figure D.9 shows that the main advantage of our algorithmic procedure is that the computational cost of the implementation increases at a lower rate with respect to `DVS` and `DSS` as p increases and the signal $\bar{p}$ is fixed. For instance, for $p = 200$, `BG` provides posterior inference almost three times faster than `DVS` and four times faster than `DSS`, on average. Such computational efficiency is a direct consequence of the properties outlined in Section 2.2 and arguably makes our approach particularly suitable in cases in which n

is moderate and p is large. A series of unreported results, also show that sampling from the MCMC equivalent (see Section A) would be more than fifteen times more costly than sampling from our variational Bayes scheme – 71.5 secs for 20,000 draws from the MCMC vs 4.2 secs for convergence in our VB setting. This is consistent with existing evidence on the computational advantage of variational inference methods vs MCMC in the context of linear models (see, e.g., Ray and Szabó, 2022; Chan and Yu, 2022; Bianchi et al., 2022b)

4 Applications in economics and finance

We now investigate the performance of our dynamic `BG` models within the context of two common problems in macroeconomics and finance: inflation forecasting based on a large set of macroeconomic variables (see, e.g., Faust and Wright, 2013) and the predictability of the equity premium based on characteristic-managed portfolios (see, e.g., Dong et al., 2022). In both cases it is not uncommon to argue in favor of modeling parameter changes for the purpose of out-of-sample forecasting (see, e.g., Kalli and Griffin, 2014; Bitto and Frühwirth-Schnatter, 2019; Huber et al., 2021; Dangl and Halling, 2012; Farmer et al., 2022). For instance, the conventional wisdom posits that the Philips curve – the relationship between unemployment and inflation – has changed over time. If so, the regression coefficients loading on labour market variables when forecasting inflation should be time varying.[7] Similarly, it is commonly thought that the relationship between the risk premium on a given asset – that is the conditional expected excess return – and sources of systematic risk is not constant over time (see, e.g., Kelly et al., 2019).

4.1 Inflation forecasting

We retrieve the macroeconomic data from the FRED-QD database of McCracken and Ng (2020). The variables consists in quarterly data spanning the period 3rd quarter 1967 to 2nd quarter 2022, such that the sample includes oil shocks in 1973 and 1979, mild recession in 1990, the dot-com bubble and the great recession in 2007-2009, and the covid-19 pandemic since 2020. We focus on forecasting four measures of inflation, namely total CPI (CPIAUCSL), core CPI (CPILFESL), GDP deflator (GDPCTPI), and PCE deflator (PCECTPI). The name in parenthesis coincides with the variables' code in the original

[7]The necessity of capturing these dynamic trends have been discussed and explored in Stock and Watson (2007), who point out that forecasting inflation has become harder due to trend cycles and dynamic volatility processes.

database. When each of these price series P_t is used as target variable to predict h-quarters ahead we transform it according to the formula $y_{t+h} = (400/h)\ln(P_t/P_{t-1})$. The 229 predictors are transformed according to standard norms in literature (see McCracken and Ng, 2020). The set of predictors also includes the first two lags of the response variable.

In sample analysis. Before discussing the out-of-sample forecasting performance, we first report the in-sample posterior estimates of both the inclusion probabilities and the regression coefficients. This exercise is intended to demonstrate that our variational Bayes approach provides reasonable estimates of trends, volatilities and other parameters. For the sake of brevity we report the in-sample estimates for $h = 1$ quarter ahead. Figure 14 reports the time-varying posterior inclusion probabilities and posterior regression coefficients estimates $\mu_{q(\beta_{jt})}$ from our dynamic `BG` model for the CPIAUCSL inflation measure. The results show that only a handful of variables significantly predict the one-quarter ahead total CPI. This is consistent with Stock and Watson (2007); Harvey et al. (2007), whereby a great deal of time-series variation in inflation is simply captured by a time-varying mean.

Not surprisingly, past inflation plays a significant role for the one-quarter ahead growth in CPI. This is in line with Koop and Korobilis (2012). Also, the posterior estimates show an interesting intersection between demand and supply factors on inflation; for instance, on the supply side industrial production (INDPRO) is predominantly positively related to inflation until 2008/2009. On the demand side, real personal consumption expenditures (PCESVx) becomes significant since 2000 until the end of 2022. This suggests that the interplay between demand and supply pressure on inflation is potentially time varying and possibly correlates with the business cycle dynamics. Finally, the model estimates confirm that monetary policy tightening exerts a downward pressure on inflation, with the 5-year treasury interest rates (T5YFFM) negatively correlated with total CPI inflation throughout the sample.

Figure 15 reports the posterior estimates $\mu_{q(\beta_{jt})}, \mu_{q(\gamma_{jt})}$ from our dynamic `BG` model for the PCECTPI inflation measure. Interestingly, there is some overlapping in the dynamics of inflation predictability between the total CPI and the PCE deflator. For instance, lagged inflation plays a significant role for both CPIAUCSL and PCECTPI over the first part of the sample, which coincide with oil crisis in the '70 and recession in the early '80s. Similarly, some variables such as industrial production (INDPRO), 5-year treasury interest rates (T5YFFM), and producer price index (WPSFD49207) are also overlapping across both measures of inflation. This suggest that our model is able to pick up some interesting broad dynamics for the supply- and demand-side predictors for inflation. In Appendix E we

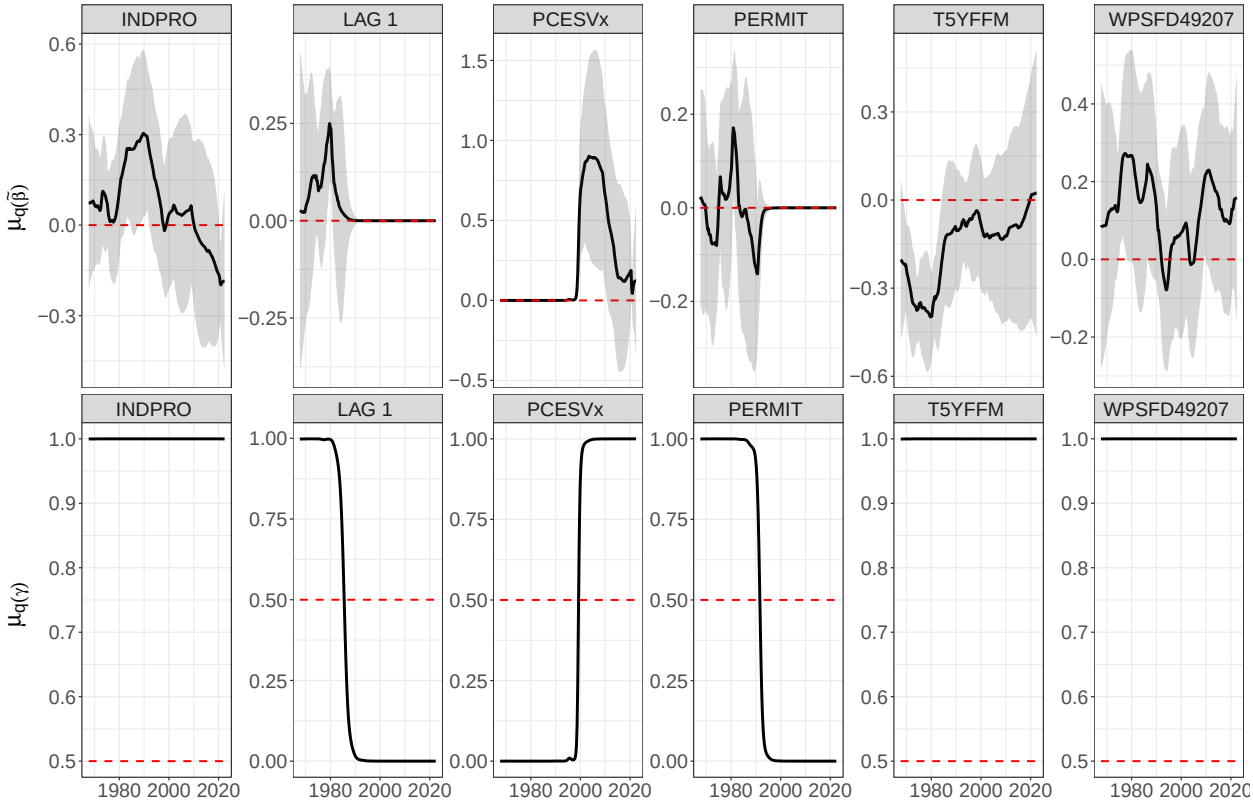


Figure 14: Time-varying coefficients estimates $\mu_{q(\beta_{jt})}$ and posterior inclusion probabilities for total CPI (CPIAUCSL).

report the in-sample estimates for both the core CPI (CPILFESL) and the GDP deflator (GDPCTPI). The results confirm two key features of our dynamic `BG` model: first, it captures predictors with some clear economic meaning. This is the case of a short-lived significance of real consumption expenditures (PCECC96) towards the end of 2020, which potentially highlights the role of stimulating demand on inflation.

Second, our model provides an alternative view to some of the main theory-based inflation predictors. For instance, short-term unemployment carries a significant signal to predict inflation as measured by the GDP deflator from the great financial crisis towards the end of the sample the GDP deflator. This evidence in favour of a time-varying Phillips curve, whereby the theoretical inverse relationship between unemployment and inflation is supported by the data but only during specific time periods. Figure E.11 in Appendix E corroborates the importance of capturing such dynamics by comparing the strength of the information available to predict inflation and idiosyncratic volatility; that is, a richer model is needed at times of

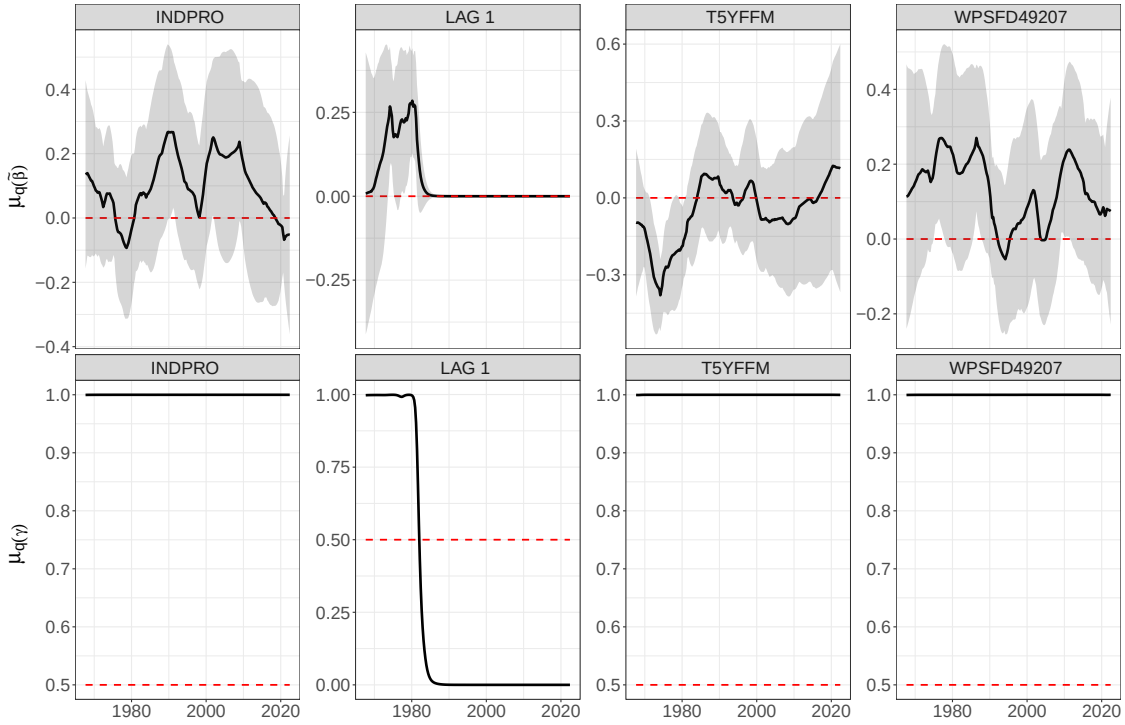


Figure 15: Time-varying coefficients estimates $\mu_{q(\beta_{jt})}$ and posterior inclusion probabilities for PCE deflator (PCECTPI).

higher uncertainty as proxied by the volatility in the residuals.

Out-of-sample forecasting. For each inflation measure we evaluate the h-quarter ahead forecasting performance based on both point forecast and density forecast accuracy. We can divide the competing methods into three groups. The first includes a series of widely used benchmarks for inflation forecasting, such as the unobserved component model of Stock and Watson (2007) (`UC`), an auto-regressive model of order two (`AR(2)`), and an auto-regressive of model of order two with time-varying parameters (`TVAR(2)`) (see, e.g., Koop and Korobilis, 2020). Notice that both `UC` and `TVAR(2)` account for stochastic volatility. In addition, we consider also a static latent factor model with five principal components (`F5`) as additional benchmark. Latent factor models also represent a rather successful approach within the context of high-dimensional regression models (see, e.g., Stock and Watson, 2006).

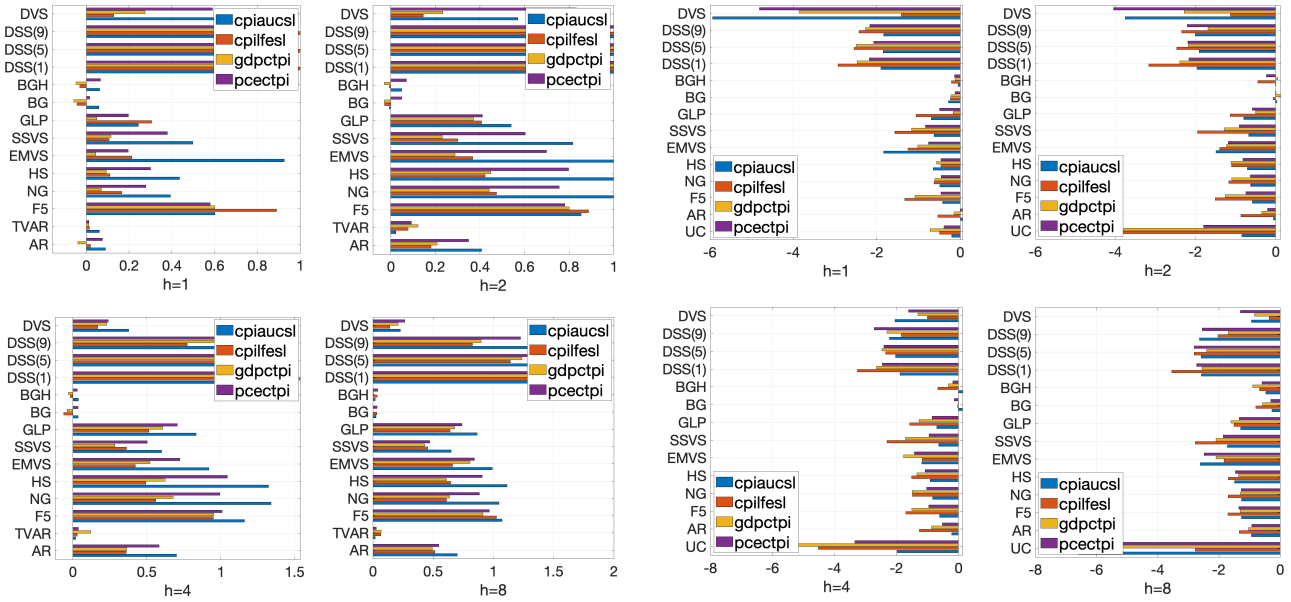
The second group of models is composed by static variable selection methods estimated using a 30-year rolling window procedure. This includes two continuous shrinkage priors, i.e. the normal-gamma prior of Griffin and Brown (2010) (`BNG`) and the horseshoe prior of

Carvalho et al. (2010) (BHS), and the variable selection methods of George and McCulloch (1993) (SSVS), Ročková and George (2014) (EMVS), and Giannone et al. (2021) (GLP). We follow existing literature, such as Huber et al. (2021); Bianchi et al. (2022b), and use the signal adaptive variable selector (SAVS) of Ray and Bhattacharya (2018) as post-processing tool to sparsify the posterior estimates from the hierarchical shrinkage priors BNG and BHS. Notice that also the AR(2) model is static in nature, and therefore estimated based on the same 30-year rolling window approach.

The third group of models considers recent advances in dynamic variable selection with time varying parameters such as Koop and Korobilis (2020) (DVS) and Ročková and McAlinn (2021) (DSS). Consistent with the simulation exercise in Section 3.2, for the DSS we consider three different values of the *marginal importance weight* parameter $\Theta \in \{0.1, 0.5, 0.9\}$. As far as our dynamic sparse regression is concerned, we test a model with (BG) and without (BGH) stochastic volatility. We consider a combination of uninformative hyper-parameters $A_\nu = 0.01, B_\nu = 0.01$, $A_\eta = 0.01, B_\eta = 0.01$, and $A_\xi = 2, B_\xi = 5$, where the choice of the latter is based on the sensitivity analysis in Section 2.2.

We first report the relative mean squared forecasting error computed as $RMSFE_i = \sum_{t=\tau}^{T} e_{i,t}^2 - \sum_{t=\tau}^{T} e_{\text{bench},t}^2$, where τ denotes the beginning of the out-of-sample period, and $e_{i,t}^2$, $e_{\text{bench},t}^2$ the forecast errors from a competing model and a benchmark specification, respectively. A value greater than zero indicates a model is under-performing the UC and vice-versa. The first prediction is generated in 1997Q3. We consider as $e_{\text{bench},t}^2$ the UC model of Stock and Watson (2007). Figure 16(a) reports the results. Not surprisingly, the UC represents a tough benchmark for point forecasts across models. The relative mean squared forecasting error with respect to UC is positive for all static variable selection methods based on rolling window estimates as well as for DVS and DSS irrespective of the choice of Θ. This holds across forecasting horizons and inflation measures, although with difference in magnitude. The gap with respect to UC tend to increase with the forecasting horizon. Nevertheless, our dynamic BG model outperforms both the benchmark and all of the competing variable selection strategies across forecasting horizons and most inflation measures. Remarkably, this also holds in comparison with both AR(2) and TVAR(2); two models which discard any information about macroeconomic factors.

As for the quality of the density forecasts, Figure 16(b) reports the average log score differential between a given model i and a benchmark, $ALS_i = \frac{1}{T-\tau-1} \sum_{t=\tau}^{T} (\log(S_{i,t}) - \log(S_{\text{bench},t}))$, where $\log(S_{i,t})$ and $\log(S_{\text{bench},t})$ represent the log-score of the ith model and the benchmark,



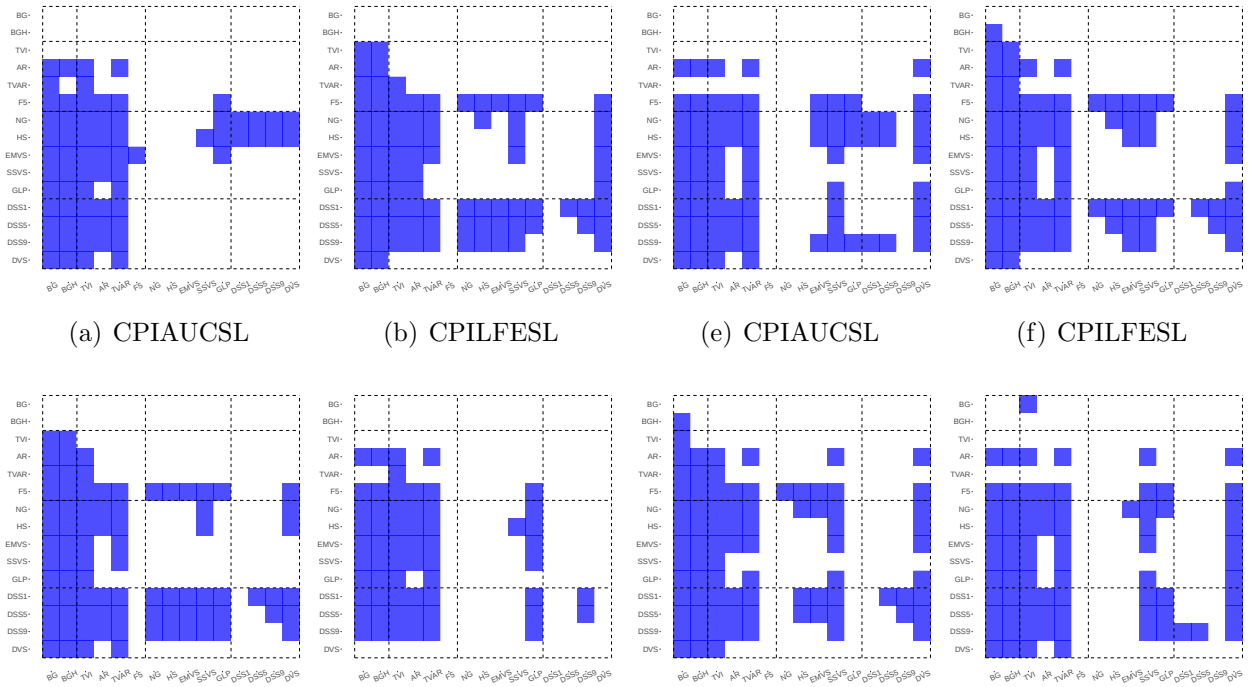
(a) Mean squared error

(b) Log-predictive score

Figure 16: Point and density forecasting. Left panel reports the relative mean squared forecasting error with respect to the unobserved component benchmark `UC`. The right panel reports the relative log-predictive socre with respect to the time-varying AR(2) model with stochastic volatility. The sample period is from 1967Q3 to 2022Q3. The first prediction is generated in 1997Q3. The scale on the x-axis is truncated to improve readability.

respectively. We consider $\log(S_{\texttt{AR(2)},t})$ as benchmark since produces the highest log-predictive score among all of the other methods. Interestingly, a simple `TVAR(2)` represents a challenging benchmark to beat when it comes to density forecasting, especially for short-term forecasts. This is in line with Koop and Korobilis (2020). However, our dynamic `BG` model represents the only competitive alternative with respect to the static and dynamic variable selection methods.

Finally, we investigate if the performance of different models are statistically different based on a series of pairwise (Diebold and Mariano, 1995) (DM) tests. Figure 17 shows the results. For a given pairwise comparison, if the null hypothesis $\mathcal{H}_0 : MSE^C \geq MSE^R$ – where MSE^C and MSE^R denote the mean-squared error of the column and row model –, is not rejected at a 10% level we report 0 (white) in the graph. If the null is rejected we report 1 (blue). For the ease of exposition, we report the results for $h = 1, 2$ forecasting horizons. The results for $h = 4, 8$ are reported in Appendix E. As far as short-term forecasts are concerned



(a) CPIAUCSL (b) CPILFESL (e) CPIAUCSL (f) CPILFESL

(c) GDPCTPI (d) PCECTPI (g) GDPCTPI (h) PCECTPI

Horizon $h = 1$ Horizon $h = 2$

Figure 17: Diebold-Mariano test for the null hypothesis $\mathcal{H}_0 : MSE^C \geq MSE^R$, where MSE^C and MSE^R denote the mean-squared error of the column and row model, respectively. If the null is not rejected at a 10% level we report 0 (white). If the null is rejected we report 1 (blue).

(first row in Figure 17), the pairwise testing results suggest that our dynamic `BG` model provides a statistically comparable performance to conventional benchmarks, such as the `UC` and the time-varying `AR(2)` model. Yet, both `BG` and `BGH` significantly outperforms all of the other static and dynamic variable selection strategies which make use of macroeconomic predictors. This holds across inflation measures. More broadly, the DM tests suggest that if the final goal is to juxtapose accurate predictions with the understanding of the key drivers of the inflation's dynamic, then our dynamic sparse regression modeling framework stands out against competing strategies.

4.2 Anomalies and the expected returns on the market

We consider the predictive content of a large set characteristic-managed portfolios, or "factors" for the one-month ahead aggregate stock market returns, expanding on the original

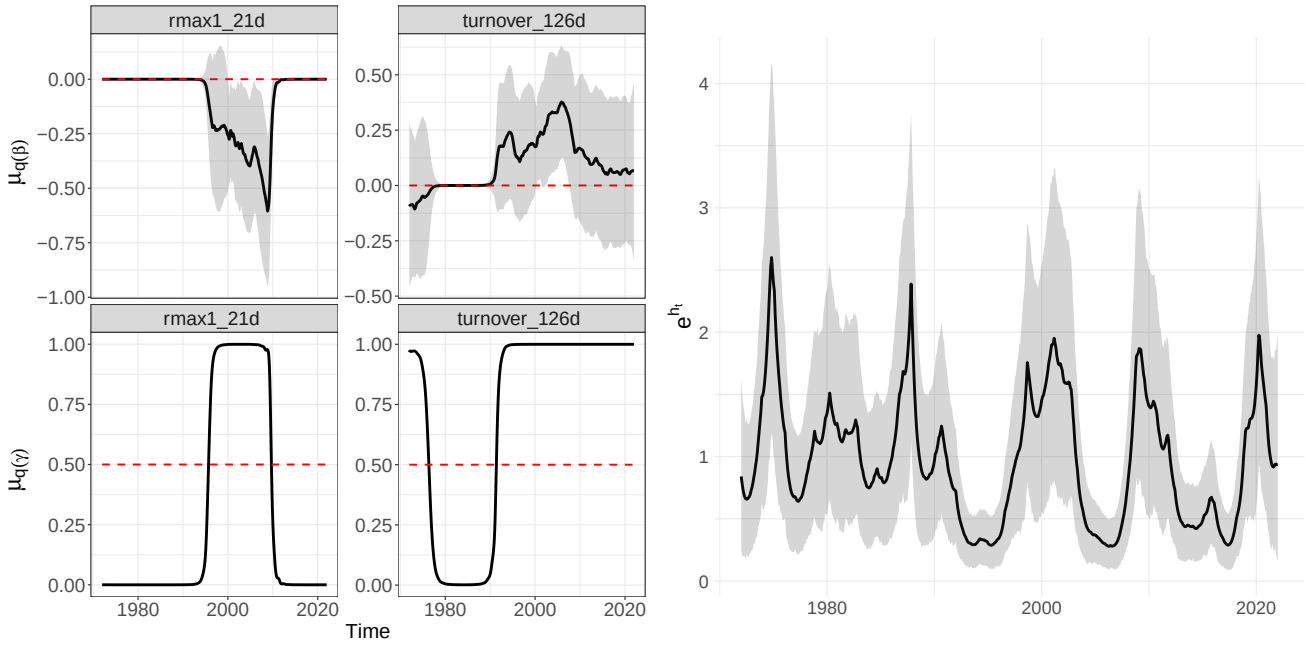
framework of Dong et al. (2022). As customary in the empirical finance literature, we restrict our analysis to value-weighted strategies that can be constructed using the Center for Research in Security Prices (CRSP) monthly and daily stock files, the Compustat Fundamental annual and quarterly files, and the Institutional Broker Estimate (IBES) database. In addition, we exclude a handful of strategies for which there are missing returns. This process identifies 149 value-weighted long-short portfolios for which we can collect monthly returns. For a more detailed description of the portfolio construction we refer to Jensen et al. (2022).[8] The portfolio returns span the period January 1971 to December 2021. The target variable is the one-month ahead returns on the value-weighted market portfolio in excess of the 30-day T-bill rate, a proxy for the equity risk premium.

In-sample analysis. We first report the in-sample posterior estimates of both the inclusion probabilities and the regression coefficients. This exercise is intended to inspect the dynamics and significance of characteristic-based portfolios for the aggregate stock market returns over time. Figure 18(a) reports the time-varying posterior inclusion probabilities and posterior regression coefficients estimates $\mu_{q(\beta_{jt})}$ from our dynamic `BG` model. We report those coefficients for which the posterior probability of inclusion is non-negligible. The results show that the model size is quite small, that is only few anomalies actually carry significant predictive power, as indicated by the regression coefficient on the `max1_21d` and the `turnover_126d` portfolios.

The `max1_21d` anomaly pertains a long-short strategy based on the maximum daily returns over the previous 21 trading trading days (see Bali et al., 2011), while `turnover_126d` pertains a long-short portfolio based on stocks the average turnover rate – number of shares traded as a fraction of the number of shares outstanding – in the previous 126 trading days as a proxy for liquidity (see Datar et al., 1998). These portfolios are primarily related to trading frictions. Figure 18(b) reports the posterior estimates of the idiosyncratic volatility $\mu_{q(\sigma_t)}$ (see Eq.B.24 in Appendix B). Idiosyncratic volatility is counter-cyclical, i.e., higher in recessions, and partly correlates with the strength of the signals as indicated by the dynamics of the regression coefficients.

Out-of-sample forecasting. As customary in the empirical asset pricing literature, we evaluate the one-month ahead forecasting performance based on both point forecast and

[8]Data on the 153 set of characteristic-based portfolios can be found at `https://jkpfactors.com`. We thank Bryan Kelly for making these data available.



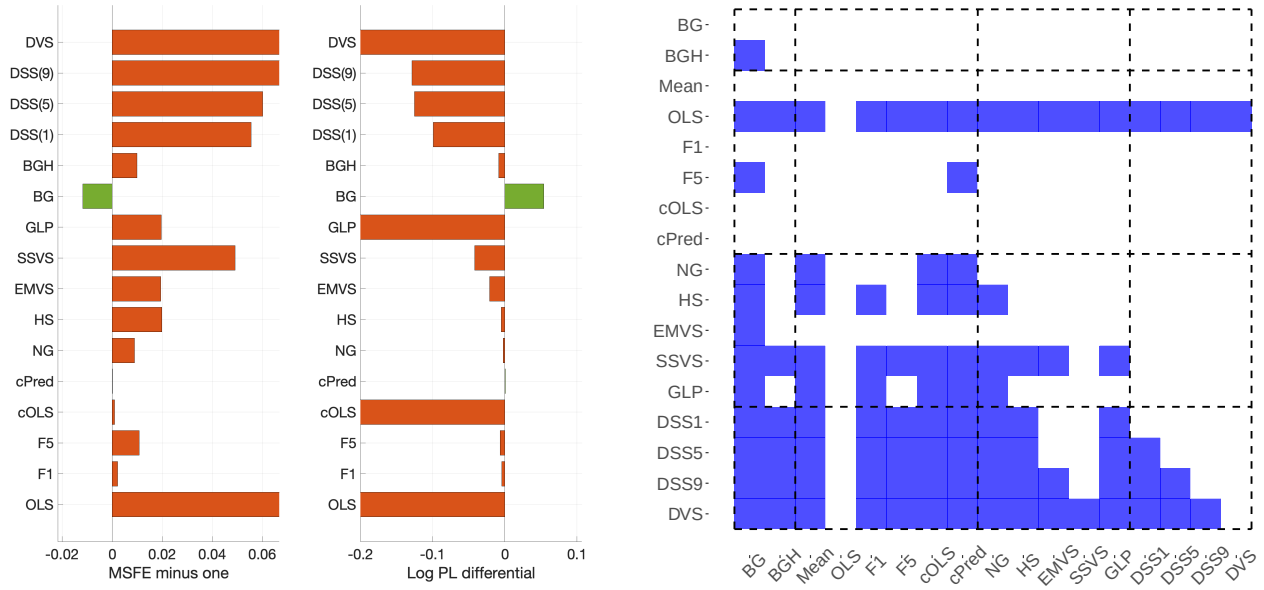
(a) Sparsity dynamics

(b) Idiosyncratic volatility dynamics

Figure 18: Dynamic sparsity and volatility. Left panel reports the time-varying coefficients estimates $\mu_{q(\beta_{jt})}$ and posterior inclusion probabilities. The right panel reports the posterior estimates of the idiosyncratic volatility $\mu_{q(\sigma_t)}$.

density forecast accuracy. In addition to our `BG` and `BGH` models, we consider the same set of static and dynamic variable selection methods as outlined in Section 4.1. However, we consider a different set of benchmark methods consistent with the existing literature. These consist of both equal-weight forecasts from individual regressions (`cOLS`) – one for each predictor — as in Rapach et al. (2010), the prediction from the recursively calculated sample mean (see, e.g., Welch and Goyal, 2008), a one-factor static principal component regression (`F1`), and a univariate regression in which the only predictor consists of the cross-sectional average of the 153 portfolio returns (`cPred`) (see, e.g., Dong et al., 2022). As for the prior hyper-parameters of our `BG` and `BGH` specifications, we consider the same combination of uninformative hyper-parameters $A_\nu = 0.01, B_\nu = 0.01$, $A_\eta = 0.01, B_\eta = 0.01$, and $A_\xi = 2, B_\xi = 5$, as for the inflation forecasting exercise (see Section 2.2). In addition, to allow for a direct comparison with Welch and Goyal (2008); Dong et al. (2022), we consider an expanding window approach in which the first 20 years of monthly returns are considered as burn-in, and forecasts are recursively generated from 1991M01 to 2021M12.

Figure 19(a) reports both the relative mean squared forecasting error and the log-score



(a) MSFE and Log score differential

(b) Diebold-Mariano tests

Figure 19: Forecasting performance. Left panel reports the mean squared error and log-predictive score differentials with respect to a simple recursive sample mean estimate. The scale on the x-axis is truncated to improve readability. Right panel reports the pairwise Diebold-Mariano tests. The sample period is from 1971M11 to 2021M12. The first prediction is generated in 1991M01.

differentials calculated as in Section 4.1, where the benchmark is the recursively calculated sample mean. The latter represents a widely used reference point to measure the extent of stock returns predictability (see, e.g., Campbell and Thompson, 2008; Fisher et al., 2020). The main message that transpire from both the point and density forecasts is that our dynamic Bernoulli-Gaussian regression strategy outperforms both static and dynamic competing variable selection strategies. Specifically, our `BG` model produces a smaller mean squared forecasting error than the no-predictability benchmark. Among the alternative benchmark considered, both the simple equal-weight average of individual forecasts `cOLS` and the `cPred` univariate regression performs on par with the recursive mean.

Figure 19(b) reports the results for a Diebold-Mariano (Diebold and Mariano, 1995) tests similar to Figure 17. Broadly speaking, the DM tests suggest that if the final goal is to leverage on the time-varying information from anomaly portfolios to the expected returns on the stock market, then our dynamic sparse regression modeling framework stands out against both static and dynamic variable selection strategies. Appendix E reports additional

results based on the mean absolute error. The results largely confirm the pattern from the relative mean squared error. Finally, as far as the quality of density forecasts is concerned, Figure 19(a) shows that our dynamic Bernoulli-Gaussian model with stochastic volatility is again the only one outperforming the no-predictability benchmark: the log-predictive score from `BG` is larger than assuming returns are generated from a normal distribution recursive sample mean and variance as sufficient statistics.

5 Concluding remarks

We are interested in modeling dynamic sparsity within the context of large-scale linear regression models with time-varying parameters. To this aim, we propose a novel variational Bayes estimation procedure which builds upon a dynamic Bernoulli-Gaussian model representation. We show both theoretically and in simulation that our inference approach concentrates the posterior estimates of time-varying regression coefficients so that different subsets of predictors can be identified over time. A comprehensive simulation study shows that our variational Bayes approach is as accurate as its MCMC counterpart, and fares favourably when compared against state-of-the-art static and dynamic variable selection methods. We evaluate empirically the performance of our model within the context of two common problems in economic forecasting, that is inflation and stock returns predictability. The empirical results suggest that a more accurate identification over time of active predictors translates into substantial out-of-sample gains compared to a variety of benchmark methods. This highlights the importance of a dynamic approach to variables selection to fully capture the extent of both inflation and stock returns predictability.

References

Bali, T. G., Cakici, N., and Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of financial economics*, 99(2):427–446.

Bianchi, D., Bernardi, M., and Bianco, N. (2022a). Smoothing volatility targeting. *Available at SSRN*.

Bianchi, D., Bernardi, M., and Bianco, N. (2022b). Variational bayes inference for large-scale multivariate predictive regressions. *Working Paper*.

Bitto, A. and Frühwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *J. Econometrics*, 210(1):75–97.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Chan, J. C., Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2012). Time varying dimension models. *Journal of Business & Economic Statistics*, 30(3):358–367.

Chan, J. C. and Yu, X. (2022). Fast and accurate variational inference for large bayesian vars with stochastic volatility. *Journal of Economic Dynamics and Control*, 143:104505.

Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106(1):157–181.

Datar, V. T., Naik, N. Y., and Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of financial markets*, 1(2):203–219.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20:134–144.

Dong, X., Li, Y., Rapach, D. E., and Zhou, G. (2022). Anomalies and the expected market return. *The Journal of Finance*, 77(1):639–681.

Farmer, L., Schmidt, L., and Timmermann, A. (2022). Pockets of predictability. *Journal of Finance, forthcoming*.

Faust, J. and Wright, J. H. (2013). Forecasting inflation. In *Handbook of economic forecasting*, volume 2, pages 2–56. Elsevier.

Fava, B. and Lopes, H. F. (2021). The illusion of the illusion of sparsity: An exercise in prior sensitivity. *Brazilian Journal of Probability and Statistics*, 35(4):699–720.

Fisher, J. D., Pettenuzzo, D., and Carvalho, C. M. (2020). Optimal asset allocation with multivariate bayesian dynamic linear models. *The annals of applied statistics*, 14(1):299–338.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, pages 339–373.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.*, 5(1):171–188.

Harvey, A. C., Trimbur, T. M., and Van Dijk, H. K. (2007). Trends and cycles in economic time series: A bayesian approach. *Journal of Econometrics*, 140(2):618–649.

Huber, F., Koop, G., and Onorante, L. (2021). Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics*, 39(3):669–683.

Inoue, A., Jin, L., and Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of econometrics*, 196(1):55–67.

Jensen, T. I., Kelly, B. T., and Pedersen, L. H. (2022). Is there a replication crisis in finance? *Journal of Finance*, (Forthcoming).

Kalli, M. and Griffin, J. (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, 178(2):779–793.

Kastner, G. (2016). Dealing with stochastic volatility in time series using the R package stochvol. *Journal of Statistical Software*, 69(5):1–30.

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

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging*. *International Economic Review*, 53(3):867–886.

Koop, G. and Korobilis, D. (2020). Bayesian dynamic variable selection in high dimensions. *International Economic Review*.

Korobilis, D. (2013). Bayesian forecasting with highly correlated predictors. *Economics Letters*, 118(1):148–150.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019). Dynamic shrinkage processes. *Journal of the Royal Statistical Society Series B*, 81(4):781–804.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

McCracken, M. and Ng, S. (2020). Fred-qd: A quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Nakajima, J. and West, M. (2013). Bayesian analysis of latent threshold dynamic models. *Journal of Business & Economic Statistics*, 31(2):151–164.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *Amer. Statist.*, 64(2):140–153.

Ormerod, J. T., You, C., and Müller, S. (2017). A variational Bayes approach to variable selection. *Electronic Journal of Statistics*, 11(2):3549 – 3594.

Pettenuzzo, D., Timmermann, A., and Valkanov, R. (2014). Forecasting stock returns under economic constraints. *Journal of Financial Economics*, 114(3):517–553.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2):821–862.

Ray, K. and Szabó, B. (2022). Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281.

Ray, P. and Bhattacharya, A. (2018). Signal adaptive variable selector for the horseshoe prior. *arXiv: Methodology*.

Rohde, D. and Wand, M. P. (2016). Semiparametric mean field variational bayes: General principles and numerical issues. *Journal of Machine Learning Research*, 17(172):1–47.

Ročková, V. and George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.*, 113(521):431–444.

Ročková, V. and McAlinn, K. (2021). Dynamic Variable Selection with Spike-and-Slab Process Priors. *Bayesian Analysis*, 16(1):233 – 269.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Soussen, C., Idier, J., Brie, D., and Duan, J. (2011). From bernoulli–gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584.

Stock, J. H. and Watson, M. W. (2006). Forecasting with many predictors. *Handbook of economic forecasting*, 1:515–554.

Stock, J. H. and Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39:3–33.

Stock, J. H. and Watson, M. W. (2010). Modeling inflation after the crisis. Technical report, National Bureau of Economic Research.

Uribe, P. W. and Lopes, H. F. (2020). Dynamic sparsity on dynamic regression models. *arXiv preprint arXiv:2009.14131*.

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011). Mean field variational bayes for elaborate distributions. *Bayesian Analysis*, 6(4):847–900.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

West, M. and Harrison, J. (2006). *Bayesian forecasting and dynamic models*. Springer Science & Business Media.

Supplementary appendix for:

Dynamic variable selection in high-dimensional predictive regressions

This appendix provide the derivation of the optimal densities used in our semi-parametric variational Bayes algorithm. The derivations concern the optimal variational densities for all model parameters. In addition, we provide formal proofs of the theoretical properties of the algorithm outlined in Section 2.2. Finally, we provide additional simulation and empirical results which have not been included in the main text for the sake of brevity.

A An equivalent MCMC sampling scheme

In this section we provide the full conditional distributions for each involved parameter. The latter enables the implementation of a Gibbs-sampling algorithm (see Algorithm 3) and constitutes the starting point to derive the variational densities in Appendix B. However, the MCMC implementation lacks of two important properties. First, it is not possible to smooth the posterior inclusion probabilities using the strategy in Proposition 2.6. Second, perhaps more important, the results described in Section 2.2 are no more valid, and therefore an efficient version of the MCMC that drops the unimportant variables on-line is not available.

Full conditional of $p(\sigma^2|\text{rest})$. Recall that the prior assumption on σ^2 is $\sigma^2 \sim \mathsf{IGa}(A_\sigma, B_\sigma)$. The full conditional distribution of σ^2 given the rest $p(\sigma^2|\text{rest}) \propto p(\mathbf{y}|\sigma^2, \mathbf{b}, \boldsymbol{\gamma})p(\sigma^2)$ is proportional to:

$$\log p(\sigma^2|\text{rest}) \propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - (A_\sigma + 1)\log\sigma^2 - \frac{B_\sigma}{\sigma^2}, \tag{A.1}$$

with $\boldsymbol{\varepsilon} = \mathbf{y} - \sum_{j=1}^{p} \mathbf{X}_j\boldsymbol{\Gamma}_j\mathbf{b}_j$, where $\mathbf{X}_k$ and $\boldsymbol{\Gamma}_k$ are diagonal matrices with elements x_{kt-1} and γ_{kt} respectively. Therefore, the full conditional distribution of variance σ^2 is an inverse-gamma $\sigma^2|\text{rest} \sim \mathsf{IG}\left(A_\sigma + \frac{n}{2}, B_\sigma + \frac{1}{2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right)$.

Sampling h. In order to get posterior samples from $p(\mathbf{h}|\mathbf{y})$, we exploit the methodology described in Kastner and Frühwirth-Schnatter (2014) where the data are transformed as $\varepsilon_t = y_t - \sum_{j=1}^{p} x_{jt-1}\gamma_{jt}b_{jt}$, for $t = 1, \ldots, n$. The latter is implemented in the stochvol package in R (Kastner, 2016).

Full conditional of $p(\mathbf{b}_j|\text{rest})$. Let the prior distribution on $\mathbf{b}_j$ be $\mathbf{b}_j \sim \mathsf{N}_{n+1}(0, \eta_j^2\mathbf{Q}^{-1})$. The full conditional distribution of $\mathbf{b}_j$ given the rest $p(\mathbf{b}_j|\text{rest}) \propto p(\mathbf{y}|\sigma^2, \mathbf{b}, \boldsymbol{\gamma})p(\mathbf{b}_j|\eta_j^2)$ is proportional to:

$$\log p(\mathbf{b}_j|\text{rest}) \propto -\frac{1}{2}\left(\mathbf{y} - \sum_{k=1}^{p}\mathbf{X}_k\boldsymbol{\Gamma}_k\mathbf{b}_k\right)' \mathbf{H}\left(\mathbf{y} - \sum_{k=1}^{p}\mathbf{X}_k\boldsymbol{\Gamma}_k\mathbf{b}_k\right) - \frac{1}{2}\mu_{q(1/\eta_j^2)}\mathbf{b}_j'\mathbf{Q}\mathbf{b}_j$$

where $\mathbf{H}$, $\mathbf{X}_k$, and $\boldsymbol{\Gamma}_k$ are diagonal matrices with elements $1/\sigma_t^2$, x_{kt-1}, and γ_{kt} for $t = 1, \ldots, n$, respectively. Define $\boldsymbol{\varepsilon}_{-j} = \mathbf{y} - \sum_{k=1,k\neq j}^{p}\mathbf{X}_k\boldsymbol{\Gamma}_k\mathbf{b}_k$, then

$$\begin{aligned}
\log p(\mathbf{b}_j|\text{rest}) &\propto -\frac{1}{2}\left(\boldsymbol{\varepsilon}_{-j} - \mathbf{X}_j\boldsymbol{\Gamma}_j\mathbf{b}_j\right)'\mathbf{H}\left(\boldsymbol{\varepsilon}_{-j} - \mathbf{X}_j\boldsymbol{\Gamma}_j\mathbf{b}_j\right) - \frac{1}{2}\mu_{q(1/\eta_j^2)}\mathbf{b}_j'\mathbf{Q}\mathbf{b}_j \\
&\propto -\frac{1}{2}\left(\mathbf{b}_j'\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{H}\mathbf{X}_j\boldsymbol{\Gamma}_j\mathbf{b}_j - 2\mathbf{b}_j'\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{H}\boldsymbol{\varepsilon}_{-j}\right) - \frac{1}{2}\mu_{q(1/\eta_j^2)}\mathbf{b}_j'\mathbf{Q}\mathbf{b}_j \qquad \text{(A.2)} \\
&\propto -\frac{1}{2}\left(\mathbf{b}_j'(\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{H}\mathbf{X}_j\boldsymbol{\Gamma}_j + 1/\eta_j^2\mathbf{Q})\mathbf{b}_j - 2\mathbf{b}_j'\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{H}\boldsymbol{\varepsilon}_{-j}\right).
\end{aligned}$$

Therefore, the full conditional distribution of $\mathbf{b}_j$ is a multivariate Gaussian distribution $\mathbf{b}_j|\text{rest} \sim \mathsf{N}_{n+1}\left(\boldsymbol{\mu}_{\mathbf{b}_j|rest}, \boldsymbol{\Sigma}_{\mathbf{b}_j|rest}\right)$, with variance-covariance $\boldsymbol{\Sigma}_{\mathbf{b}_j|rest} = (\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{H}\mathbf{X}_j\boldsymbol{\Gamma}_j + 1/\eta_j^2\mathbf{Q})^{-1}$ and mean $\boldsymbol{\mu}_{\mathbf{b}_j|rest} = (\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{H}\mathbf{X}_j\boldsymbol{\Gamma}_j + 1/\eta_j^2\mathbf{Q})^{-1}\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{H}\boldsymbol{\varepsilon}_{-j}$.

Full conditional of $p(\gamma_{jt}|\text{rest})$. Recall that the prior assumption on γ_{jt} is $\gamma_{jt} \sim \mathsf{Bern}(\text{expit}(\omega_{jt}))$. The full conditional distribution of γ_{jt}, namely $p(\gamma_{jt}|\text{rest}) \propto p(\mathbf{y}|\sigma^2, \mathbf{b}, \boldsymbol{\gamma})p(\gamma_{jt}|\omega_{jt})$ is proportional to:

$$\begin{aligned}
\log p(\gamma_{jt}|\text{rest}) &\propto -\frac{1}{2\sigma_t^2}\left(y_t - \sum_{k=1}^{p}\gamma_{kt}b_{kt}x_{kt-1}\right)^2 + \gamma_{jt}\omega_{jt} \\
&\propto -\frac{1}{2\sigma_t^2}(\gamma_{jt}^2 b_{jt}^2 x_{jt-1}^2 - 2\gamma_{jt}b_{jt}x_{jt-1}\varepsilon_{-j,t}) + \gamma_{jt}\omega_{jt} \qquad \text{(A.3)} \\
&\propto \gamma_{jt}\left\{\omega_{jt} - \frac{1}{2\sigma_t^2}(b_{jt}^2x_{jt-1}^2 - 2b_{jt}x_{jt-1}\varepsilon_{-j,t})\right\}.
\end{aligned}$$

Therefore, the full conditional distribution of the indicator variable γ_{jt} is a Bernoulli distribution $\gamma_{jt}|\text{rest} \sim \mathsf{Bern}\left(\text{expit}\left\{\omega_{jt} - \frac{1}{2\sigma_t^2}(b_{jt}^2x_{jt-1}^2 - 2b_{jt}x_{jt-1}\varepsilon_{-j,t})\right\}\right)$.

Full conditional of $p(\boldsymbol{\omega}_j|\text{rest})$. Let the prior distribution on $\boldsymbol{\omega}_j$ be $\boldsymbol{\omega}_j \sim \mathsf{N}_{n+1}(0, \xi_j^2\mathbf{Q}^{-1})$. The full conditional distribution $p(\boldsymbol{\omega}_j|\text{rest}) \propto \left[\prod_{t=1}^{n} p(\gamma_{jt}|\omega_{jt}, z_{jt})p(z_{jt}|\omega_{jt})\right] p(\boldsymbol{\omega}_j|\xi_j^2)$ is pro-

portional to:

$$\log p(\boldsymbol{\omega}_j|\text{rest}) \propto \boldsymbol{\omega}_j'(\boldsymbol{\gamma}_j - 1/2\boldsymbol{\iota}_n) - \frac{1}{2}\boldsymbol{\omega}_j'\mathsf{Diag}(\mathbf{z}_j)\boldsymbol{\omega}_j - \frac{1}{2\xi_j^2}\boldsymbol{\omega}_j'\mathbf{Q}\boldsymbol{\omega}_j$$
$$\propto -\frac{1}{2}\left(\boldsymbol{\omega}_j'(\mathsf{Diag}(\mathbf{z}_j) + 1/\xi_j^2\mathbf{Q})\boldsymbol{\omega}_j - 2\boldsymbol{\omega}_j'(\boldsymbol{\gamma}_j - 1/2\boldsymbol{\iota}_n)\right). \tag{A.4}$$

Therefore, the full conditional distribution of $\boldsymbol{\omega}_j$ is a multivariate Gaussian distribution $\boldsymbol{\omega}_j|\text{rest} \sim \mathsf{N}_{n+1}\left(\boldsymbol{\mu}_{\omega_j|rest}, \boldsymbol{\Sigma}_{\omega_j|rest}\right)$, with variance-covariance $\boldsymbol{\Sigma}_{\omega_j|rest} = (\mathsf{Diag}(\mathbf{z}_j) + 1/\xi_j^2\mathbf{Q})^{-1}$ and mean $\boldsymbol{\mu}_{\omega_j|rest} = (\mathsf{Diag}(\mathbf{z}_j) + 1/\xi_j^2\mathbf{Q})^{-1}(\boldsymbol{\gamma}_j - 1/2\boldsymbol{\iota}_n)$.

Full conditional of $p(z_{jt}|\text{rest})$. Recall the Polya-Gamma representation in Eq.(7). Then, the full conditional distribution of z_{jt}, namely $p(z_{jt}|\text{rest}) \propto p(\gamma_{jt}|z_{jt}, \omega_{jt})p(z_{jt}|\omega_{jt})$ is proportional to:

$$\log p(z_{jt}|\text{rest}) \propto -z_{jt}\omega_{jt}^2 + \log p(z_{jt}), \tag{A.5}$$

where $p(z_{jt})$ is the density function of a Polya-Gamma random variable $\mathsf{PG}(1, 0)$. Hence, $z_{jt}|\text{rest} \sim \mathsf{PG}(1, \omega_{jt}^2)$.

Full conditional of $p(\eta_j^2|\text{rest})$. Assume that a prior $\eta_j^2 \sim \mathsf{IGa}(A_\eta, B_\eta)$. Then, the full conditional distribution of η_j^2 given the rest $p(\eta_j^2|\text{rest}) \propto p(\mathbf{b}_j|\eta_j^2)p(\eta_j^2)$ is proportional to:

$$p(\eta_j^2|\text{rest}) \propto -\frac{n+1}{2}\log\eta_j^2 - \frac{1}{2\eta_j^2}\mathbf{b}_j'\mathbf{Q}\mathbf{b}_j - (A_\eta + 1)\log\eta_j^2 - \frac{B_\eta}{\eta_j^2}$$
$$\propto -(A_\eta + \frac{n+1}{2} + 1)\log\eta_j^2 - \frac{1}{\eta_j^2}\left(B_\eta + \frac{1}{2}\mathbf{b}_j'\mathbf{Q}\mathbf{b}_j\right). \tag{A.6}$$

Therefore, the full conditional distribution of the conditional variance η_j^2 is an inverse-gamma $\eta_j^2|\text{rest} \sim \mathsf{IG}\left(A_\eta + \frac{n+1}{2}, B_\eta + \frac{1}{2}\mathbf{b}_j'\mathbf{Q}\mathbf{b}_j\right)$.

Full conditional of $p(\xi_j^2|\text{rest})$. Recall that a priori $\xi_j^2 \sim \mathsf{IGa}(A_\xi, B_\xi)$. The full conditional distribution of ξ_j^2 given the rest $p(\xi_j^2|\text{rest}) \propto p(\boldsymbol{\omega}_j|\xi_j^2)p(\xi_j^2)$ is proportional to:

$$p(\xi_j^2|\text{rest}) \propto -\frac{n+1}{2}\log\xi_j^2 - \frac{1}{2\xi_j^2}\boldsymbol{\omega}_j'\mathbf{Q}\boldsymbol{\omega}_j - (A_\xi + 1)\log\xi_j^2 - \frac{B_\xi}{\xi_j^2}$$
$$\propto -(A_\xi + \frac{n+1}{2} + 1)\log\xi_j^2 - \frac{1}{\xi_j^2}\left(B_\xi + \frac{1}{2}\boldsymbol{\omega}_j'\mathbf{Q}\boldsymbol{\omega}_j\right). \tag{A.7}$$

Algorithm 3: Gibbs-sampling scheme for dynamic sparse regression models.

Initialize: $\boldsymbol{\vartheta}^{(0)}$, ndraws, A_ν, B_ν, A_η, B_η, A_ξ, B_ξ
for $r = 1, \ldots, \mathsf{ndraws}$ **do**
  for $j = 1, \ldots, p$ **do**
    Compute $\boldsymbol{\Sigma}_{\mathbf{b}_j|rest} = (\boldsymbol{\Gamma}_j^{(r-1)} \mathbf{X}_j \mathbf{H}^{(r-1)} \mathbf{X}_j \boldsymbol{\Gamma}_j^{(r-1)} + 1/\eta_j^{2(r-1)} \mathbf{Q})^{-1}$;
    Compute $\boldsymbol{\mu}_{\mathbf{b}_j|rest} = \boldsymbol{\Sigma}_{\mathbf{b}_j|rest} \boldsymbol{\Gamma}_j^{(r-1)} \mathbf{X}_j \mathbf{H}^{(r-1)} \boldsymbol{\varepsilon}_{-j}^{(r-1)}$;
    Sample $\mathbf{b}_j^{(r)} \sim \mathsf{N}_{n+1}\left(\boldsymbol{\mu}_{\mathbf{b}_j|rest}, \boldsymbol{\Sigma}_{\mathbf{b}_j|rest}\right)$;
    Sample $\eta_j^{2(r)} \sim \mathsf{IG}\left(A_\eta + \frac{n+1}{2}, B_\eta + \frac{1}{2}\mathbf{b}_j'^{(r)} \mathbf{Q} \mathbf{b}_j^{(r)}\right)$;
    Compute $\boldsymbol{\Sigma}_{\omega_j|rest} = (\mathsf{Diag}(\mathbf{z}_j^{(r-1)}) + 1/\xi_j^{2(r-1)} \mathbf{Q})^{-1}$;
    Compute $\boldsymbol{\mu}_{\omega_j|rest} = \boldsymbol{\Sigma}_{\omega_j|rest}(\boldsymbol{\gamma}_j^{(r-1)} - 1/2\boldsymbol{\iota}_n)$;
    Sample $\boldsymbol{\omega}_j^{(r)} \sim \mathsf{N}_{n+1}\left(\boldsymbol{\mu}_{\omega_j|rest}, \boldsymbol{\Sigma}_{\omega_j|rest}\right)$;
    Sample $\xi_j^{2(r)} \sim \mathsf{IG}\left(A_\xi + \frac{n+1}{2}, B_\xi + \frac{1}{2}\boldsymbol{\omega}_j'^{(r)} \mathbf{Q} \boldsymbol{\omega}_j^{(r)}\right)$;
    for $t = 1, \ldots, n$ **do**
      Sample $z_{jt}^{(r)} \sim \mathsf{PG}(1, \omega_{jt}^{2(r)})$;
      Sample $\gamma_{jt}^{(r)} \sim \mathsf{Bern}\left(\text{expit}\left\{\omega_{jt}^{(r)} - \frac{1}{2\sigma_t^{2(r-1)}}(b_{jt}^{2(r)} x_{jt}^2 - 2 b_{jt}^{(r)} x_{jt} \varepsilon_{-j,t}^{(r)})\right\}\right)$;
    end
  end
  Sample $\mathbf{h}$ with $\varepsilon_t^{(r)} = y_t - \sum_{j=1}^p x_{jt} \gamma_{jt}^{(r)} b_{jt}^{(r)}$ (heteroskedastic);
  Sample $\nu^{2(r)} \sim \mathsf{IG}\left(A_\nu + \frac{n+1}{2}, B_\nu + \frac{1}{2}\mathbf{h}_j'^{(r)} \mathbf{Q} \mathbf{h}_j^{(r)}\right)$;
  Sample $\sigma^{2(r)} \sim \mathsf{IG}\left(A_\sigma + \frac{n}{2}, B_\sigma + \frac{1}{2}\boldsymbol{\varepsilon}'^{(r)} \boldsymbol{\varepsilon}^{(r)}\right)$ (homoskedastic);
end

Hence, the full conditional distribution of the conditional variance ξ_j^2 is an inverse-gamma $\xi_j^2|\text{rest} \sim \mathsf{IG}\left(A_\xi + \frac{n+1}{2}, B_\xi + \frac{1}{2}\boldsymbol{\omega}_j' \mathbf{Q} \boldsymbol{\omega}_j\right)$.

Full conditional of $p(\nu^2|\text{rest})$. Assume that a priori $\nu^2 \sim \mathsf{IGa}(A_\nu, B_\nu)$. The full conditional distribution of ν^2 given the rest $p(\nu^2|\text{rest}) \propto p(\mathbf{h}|\nu^2) p(\nu^2)$ is proportional to:

$$\begin{aligned} p(\nu^2|\text{rest}) &\propto -\frac{n+1}{2} \log \nu^2 - \frac{1}{2\nu^2} \mathbf{h}_j' \mathbf{Q} \mathbf{h}_j - (A_\nu + 1) \log \nu^2 - \frac{B_\nu}{\nu^2} \\ &\propto -(A_\nu + \frac{n+1}{2} + 1) \log \nu^2 - \frac{1}{\nu^2}\left(B_\nu + \frac{1}{2}\mathbf{h}_j' \mathbf{Q} \mathbf{h}_j\right). \end{aligned} \tag{A.8}$$

Therefore, the full conditional distribution of the conditional variance ν^2 is an inverse-gamma $\nu^2|\text{rest} \sim \mathsf{IG}\left(A_\nu + \frac{n+1}{2}, B_\nu + \frac{1}{2}\mathbf{h}_j' \mathbf{Q} \mathbf{h}_j\right)$.

B Optimal variational densities

Proposition B.1. *The optimal variational density for the time-varying regression parameters* $\mathbf{b}_j = (b_{j0}, b_{j1}, \dots, b_{jn})'$ *is a multivariate Gaussian* $q^*(\mathbf{b}_j) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(\mathbf{b}_j)}, \boldsymbol{\Sigma}_{q(\mathbf{b}_j)})$*, where:*

$$\boldsymbol{\Sigma}_{q(\mathbf{b}_j)} = (\mathbf{D}_j{}^2 + \mu_{q(1/\eta_j^2)}\mathbf{Q})^{-1}, \qquad \boldsymbol{\mu}_{q(\mathbf{b}_j)} = \boldsymbol{\Sigma}_{q(\mathbf{b}_j)}\mathbf{D}_j\boldsymbol{\mu}_{q(\varepsilon_{-j})}, \tag{B.1}$$

where $\mathbf{D}_j$ *and* $\mathbf{D}_j^2$ *are diagonal matrices with elements* $[\mathbf{D}_j]_t = \mu_{q(1/\sigma_t^2)}\mu_{q(\gamma_{jt})}x_{jt-1}$ *and* $[\mathbf{D}_j]_t^2 = \mu_{q(1/\sigma_t^2)}\mu_{q(\gamma_{jt})}x_{jt-1}^2$*, respectively. Moreover,* $\boldsymbol{\mu}_{q(\varepsilon_{-j})}$ *is the vector of partial residuals with elements* $\mu_{q(\varepsilon_{-jt})} = y_t - \sum_{k=1,k\neq j}^{p} x_{kt-1}\mu_{q(\gamma_{kt})}\mu_{q(b_{kt})}$.

Proof. The full conditional distribution of $\mathbf{b}_j$ given the rest $p(\mathbf{b}_j|\text{rest})$ is defined in Eq.(A.2). According to Ormerod and Wand (2010), the optimal variational density is given by:

$$\begin{aligned}
\log q^*(\mathbf{b}_j) &\propto \mathbb{E}_{-\mathbf{b}_j}[\log p(\mathbf{b}_j|\text{rest})] \\
&\propto -\frac{1}{2}\left(\mathbf{b}_j'\mathbf{D}_j^2\mathbf{b}_j - 2\mathbf{b}_j'\mathbf{D}_j\boldsymbol{\mu}_{q(\varepsilon_{-j})}\right) - \frac{1}{2}\mu_{q(1/\eta_j^2)}\mathbf{b}_j'\mathbf{Q}\mathbf{b}_j \\
&\propto -\frac{1}{2}\left(\mathbf{b}_j'(\mathbf{D}_j^2 + \mu_{q(1/\eta_j^2)}\mathbf{Q})\mathbf{b}_j - 2\mathbf{b}_j'\mathbf{D}_j\boldsymbol{\mu}_{q(\varepsilon_{-j})}\right),
\end{aligned} \tag{B.2}$$

where $\mathbf{D}_j^m$ is a diagonal matrix with elements equal to

$$[\mathbf{D}_j^m]_t = \mathbb{E}_{-\mathbf{b}_j}[\gamma_{jt}x_{jt-1}^m/\sigma_t^2)] = \mu_{q(1/\sigma_t^2)}\mu_{q(\gamma_{jt})}x_{jt-1}^m,$$

and $\boldsymbol{\mu}_{q(\varepsilon_{-j})} = (0, \mu_{q(\varepsilon_{-j,1})}, \dots, \mu_{q(\varepsilon_{-j,n})})$ with

$$\mu_{q(\varepsilon_{-j,t})} = \mathbb{E}_{-\mathbf{b}_j}\left[\mathbf{y} - \sum_{k=1,k\neq j}^{p}\mathbf{X}_k\boldsymbol{\Gamma}_k\mathbf{b}_k\right] = y_t - \sum_{k=1,k\neq j}^{p} x_{kt-1}\mu_{q(\gamma_{kt})}\mu_{q(\mathbf{b}_{kt})}.$$

Equation B.2 represents the kernel of a multivariate Gaussian distribution as in B.1. □

Proposition B.2. *The optimal variational density for the parameters* γ_{jt} *is a Bernoulli random variable* $q^*(\gamma_{jt}) \equiv \mathsf{Bern}(\text{expit}(\omega_{q(\gamma_{jt})}))$*, where* $\text{expit}(\cdot)$ *is the inverse of the logit function and* $\omega_{q(\gamma_{jt})} = \mu_{q(\omega_{jt})} - \frac{1}{2}\mu_{q(1/\sigma_t^2)}(x_{jt-1}^2\mathbb{E}_q[b_{jt}^2] - 2\mu_{q(b_{jt})}x_{jt-1}\mu_{q(\varepsilon_{-jt})})$.

Proof. The full conditional distribution of $\gamma_{jt} \sim p(\gamma_{jt}|\text{rest})$ is derived in Eq.(A.3). Thus, the

optimal variational density is given by:

$$\begin{aligned}\log q^*(\gamma_{jt}) &\propto \mathbb{E}_{-\gamma_{jt}}\left[\log p(\gamma_{jt}|\text{rest})\right] \\ &\propto \gamma_{jt}\{\mu_{q(\omega_{jt})} - \frac{1}{2}\mu_{q(1/\sigma^2)}(x_{jt-1}^2\mathbb{E}_q[b_{jt}^2] - 2\mu_{q(b_{jt})}x_{jt-1}\mu_{q(\varepsilon_{-j,t})})\},\end{aligned} \tag{B.3}$$

where $\mathbb{E}_q[b_{jt}^2] = \mu_{q(b_{jt})}^2 + \sigma_{q(b_{jt})}^2$. Eq.B.3 is the kernel of a Bernoulli distribution as in B.2. □

Proposition B.3. *The optimal variational density for the parameter $\boldsymbol{\omega}_j$ is a multivariate Gaussian $q^*(\boldsymbol{\omega}_j) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(\omega_j)}, \boldsymbol{\Sigma}_{q(\omega_j)})$, where:*

$$\boldsymbol{\Sigma}_{q(\omega_j)} = (\mathsf{Diag}(0, \boldsymbol{\mu}_{q(z_j)}) + \mu_{q(1/\xi_j^2)}\mathbf{Q})^{-1}, \qquad \boldsymbol{\mu}_{q(\omega_j)} = \boldsymbol{\Sigma}_{q(\omega_j)}(0, \boldsymbol{\mu}_{q(\bar{\gamma}_j)}^{\intercal})^{\intercal}, \tag{B.4}$$

with $\boldsymbol{\mu}_{q(\bar{\gamma}_j)} = \boldsymbol{\mu}_{q(\gamma_j)} - 1/2\boldsymbol{\iota}_n$.

Proof. The full conditional distribution of $\boldsymbol{\omega}_j$ is defined in Eq.(A.4). Then, the optimal variational density is given by:

$$\begin{aligned}\log q^*(\boldsymbol{\omega}_j) &\propto \mathbb{E}_{-\omega_j}[\log p(\boldsymbol{\omega}_j|\text{rest})] \\ &\propto \boldsymbol{\omega}_j'\boldsymbol{\mu}_{q(\bar{\gamma}_j)} - \frac{1}{2}\boldsymbol{\omega}_j'\mathsf{Diag}(\boldsymbol{\mu}_{q(z_j)})\boldsymbol{\omega}_j - \frac{1}{2}\mu_{q(1/\xi_j^2)}\boldsymbol{\omega}_j'\mathbf{Q}\boldsymbol{\omega}_j \\ &\propto -\frac{1}{2}\left(\boldsymbol{\omega}_j'(\mathsf{Diag}(0, \boldsymbol{\mu}_{q(z_j)}) + \mu_{q(1/\xi_j^2)}\mathbf{Q})\boldsymbol{\omega}_j - 2\boldsymbol{\omega}_j'(0, \boldsymbol{\mu}_{q(\bar{\gamma}_j)}')'\right),\end{aligned} \tag{B.5}$$

where $\boldsymbol{\mu}_{q(\bar{\gamma}_j)} = \boldsymbol{\mu}_{q(\gamma_j)} - 1/2\boldsymbol{\iota}_n$. Equation B.5 is the kernel of a multivariate Gaussian distribution as in 2.3. □

Proposition B.4. *Let $q^*(\mathbf{b}_j)$ and $q^*(\gamma_{jt})$ be the optimal variational densities presented in Propositions 2.1 and 2.2. Define $\boldsymbol{\beta}_j = \boldsymbol{\Gamma}_j\mathbf{b}_j$, where the matrix $\boldsymbol{\Gamma}_j = \mathsf{diag}(1, \gamma_{j1}, \ldots, \gamma_{jn})$. The optimal variational density of $\boldsymbol{\beta}_j$ is given by a mixture of multivariate Gaussian distributions:*

$$q^*(\boldsymbol{\beta}_j) = \sum_{\mathbf{s}\in\mathcal{S}} w_s\, \mathsf{N}_{n+1}(\mathbf{D}_s\boldsymbol{\mu}_{q(\mathbf{b}_j)}, \mathbf{D}_s^{1/2}\boldsymbol{\Sigma}_{q(\mathbf{b}_j)}\mathbf{D}_s^{1/2}), \tag{B.6}$$

where $\mathcal{S}$ is a sequence of $\{0,1\}$ of length n with cardinality $|\mathcal{S}| = 2^n$, the diagonal matrix $\mathbf{D}_s = \mathsf{diag}(1, s_1, \ldots, s_n)$, and mixing weights:

$$w_s = \prod_{t=1}^{n} \mu_{q(\gamma_{jt})}^{s_t}(1 - \mu_{q(\gamma_{jt})})^{1-s_t}, \tag{B.7}$$

where $\mathbf{s} = (s_1, \ldots, s_t, \ldots, s_n) \in \mathcal{S}$ is an element in $\mathcal{S}$. Moreover, the mean and variance can

be computed analytically:

$$\boldsymbol{\mu}_{q(\beta_j)} = \boldsymbol{\mu}_{q(\Gamma_j)}\boldsymbol{\mu}_{q(\mathbf{b}_j)}, \tag{B.8}$$

$$\boldsymbol{\Sigma}_{q(\beta_j)} = (\boldsymbol{\mu}_{q(\gamma_j)}\boldsymbol{\mu}'_{q(\gamma_j)} + \mathbf{W}_{\mu_{q(\gamma_j)}}) \odot \boldsymbol{\Sigma}_{q(\mathbf{b}_j)} + \mathbf{W}_{\mu_{q(\gamma_j)}} \odot \boldsymbol{\mu}_{q(\mathbf{b}_j)}\boldsymbol{\mu}'_{q(\mathbf{b}_j)}, \tag{B.9}$$

where $\mathbf{W}_{\mu_{q(\gamma_j)}}$ *is a diagonal matrix with elements* $\left(1, \{\mu_{q(\gamma_{jt})}(1-\mu_{q(\gamma_{jt})})\}_{t=1}^{n}\right)$.

Proof. Recall that under the mean-field variational Bayes setting we have that

$$q(\mathbf{b}_j, \gamma_{j1}, \ldots, \gamma_{jn}) = q(\mathbf{b}_j)\prod_{t=1}^{n} q(\gamma_{jt}). \tag{B.10}$$

For the sake of simplicity, in what follows we drop the index j and define $\boldsymbol{\gamma} = \mathsf{diag}(\boldsymbol{\Gamma})$ the diagonal elements in $\boldsymbol{\Gamma}$. Consider the following transformation of random variables $(\boldsymbol{\gamma} = \boldsymbol{\gamma}, \boldsymbol{\beta} = \boldsymbol{\Gamma}\mathbf{b})$, so that $\mathbf{b} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}$. Hence it follows that:

$$\mathbf{J} = \begin{bmatrix} \nabla_\gamma(\gamma_1, \ldots, \gamma_n)' & \nabla_{\mathbf{b}}(\gamma_1, \ldots, \gamma_n)' \\ \nabla_\gamma \boldsymbol{\Gamma}^{-1}\boldsymbol{\beta} & \nabla_{\boldsymbol{\beta}}\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \nabla_\gamma \boldsymbol{\Gamma}^{-1}\boldsymbol{\beta} & \boldsymbol{\Gamma}^{-1} \end{bmatrix}, \tag{B.11}$$

and so $|\mathbf{J}| = |\boldsymbol{\Gamma}^{-1}|$. The joint distribution of $(\boldsymbol{\beta}, \gamma_1, \ldots, \gamma_n)$ can be written as:

$$q(\boldsymbol{\beta}, \gamma_1, \ldots, \gamma_n) = |\boldsymbol{\Gamma}^{-1}| q(\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}) \prod_{t=1}^{n} q(\gamma_{jt}) = f(\boldsymbol{\beta}|\gamma_1, \ldots, \gamma_n) f(\gamma_1, \ldots, \gamma_n), \tag{B.12}$$

where q are then replaced by the optimal elements q^*. For the conditional distribution in (B.12), we have that:

$$f(\boldsymbol{\beta}|\boldsymbol{\gamma}) = |\boldsymbol{\Gamma}^{-1}| \phi_{n+1}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}|\boldsymbol{\mu}_{q(\mathbf{b})}, \boldsymbol{\Sigma}_{q(\mathbf{b})}), \tag{B.13}$$

where $\phi_{n+1}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density function of a multivariate Gaussian. After some computations we have that $f(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \phi_{n+1}(\boldsymbol{\beta}|\boldsymbol{\mu}(\boldsymbol{\gamma}), \boldsymbol{\Sigma}(\boldsymbol{\gamma}))$ with mean vector $\boldsymbol{\mu}(\boldsymbol{\gamma}) = \boldsymbol{\Gamma}\boldsymbol{\mu}_{q(\mathbf{b})}$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{q(\mathbf{b})}\boldsymbol{\Gamma}^{1/2}$. The marginal for $\boldsymbol{\beta}$ can be found as:

$$q(\boldsymbol{\beta}) = \sum_{\mathbf{s}\in\mathcal{S}} \phi_{n+1}(\boldsymbol{\beta}|\boldsymbol{\mu}(\boldsymbol{\gamma} = \mathbf{s}), \boldsymbol{\Sigma}(\boldsymbol{\gamma} = \mathbf{s})) \prod_{t=1}^{n} q(\gamma_t = s_t), \tag{B.14}$$

where $\mathcal{S}$ denotes the domain of $\boldsymbol{\gamma} = (1, \gamma_1, \ldots, \gamma_n)$, and it is composed by all the possible sequences of $\{0, 1\}$ of length n, since the first element is fixed to be 1. The latter set has cardinality $|\mathcal{S}| = 2^n$. The distributional result concerning $\boldsymbol{\beta}$ is therefore proven.

Now compute the marginal mean recall that $\mathbb{E}_x(x) = \mathbb{E}_y(\mathbb{E}_x(x|y))$. Hence $\mathbb{E}_q(\boldsymbol{\beta}) = \mathbb{E}_\gamma(\boldsymbol{\Gamma}\boldsymbol{\mu}_{q(\mathbf{b})}) = \boldsymbol{\mu}_{q(\Gamma)}\boldsymbol{\mu}_{q(\mathbf{b})}$. The marginal variance-covariance matrix is then computed as $\mathsf{Var}_q(\boldsymbol{\beta}) = \mathbb{E}(\boldsymbol{\beta}\boldsymbol{\beta}') - \mathbb{E}(\boldsymbol{\beta})\mathbb{E}(\boldsymbol{\beta})'$ where

$$\begin{aligned}\mathbb{E}(\boldsymbol{\beta}\boldsymbol{\beta}') &= \mathbb{E}(\boldsymbol{\Gamma}\mathbf{b}(\boldsymbol{\Gamma}\mathbf{b})') = \mathbb{E}(\boldsymbol{\Gamma}\mathbf{b}\mathbf{b}'\boldsymbol{\Gamma}) = \mathbb{E}(\boldsymbol{\gamma}\boldsymbol{\gamma}' \odot \mathbf{b}\mathbf{b}') = \mathbb{E}(\boldsymbol{\gamma}\boldsymbol{\gamma}') \odot \mathbb{E}(\mathbf{b}\mathbf{b}') \\ &= (\boldsymbol{\mu}_{q(\gamma)}\boldsymbol{\mu}'_{q(\gamma)} + \mathbf{W}_{\mu_{q(\gamma)}}) \odot (\boldsymbol{\mu}_{q(\mathbf{b})}\boldsymbol{\mu}'_{q(\mathbf{b})} + \boldsymbol{\Sigma}_{q(\mathbf{b})}),\end{aligned} \tag{B.15}$$

where $\mathbf{W}_{\mu_{q(\gamma)}}$ is a diagonal matrix with elements $(1, \{\mu_{q(\gamma_t)}(1-\mu_{q(\gamma_t)})\}_{t=1}^n)$. Plug-in the latter in the formula for $\mathsf{Var}_q(\boldsymbol{\beta})$ and recall the analytical form of the mean $\mathbb{E}(\boldsymbol{\beta})$. After some simplification we end up with $\boldsymbol{\Sigma}_{q(\beta)} = (\boldsymbol{\mu}_{q(\gamma)}\boldsymbol{\mu}'_{q(\gamma)} + \mathbf{W}_{\mu_{q(\gamma)}}) \odot \boldsymbol{\Sigma}_{q(\mathbf{b})} + \mathbf{W}_{\mu_{q(\gamma)}} \odot \boldsymbol{\mu}_{q(\mathbf{b})}\boldsymbol{\mu}'_{q(\mathbf{b})}$, which concludes the proof. □

Proposition B.5. *Let $\boldsymbol{\varepsilon}^2 = \boldsymbol{\varepsilon} \odot \boldsymbol{\varepsilon}$ with components $[\boldsymbol{\varepsilon}^2]_t = (y_t - \boldsymbol{\beta}_t\mathbf{x}_{t-1})^2$. Assuming a Gaussian Markov Random Field (GMRF) approximation $q^*(\mathbf{h}) \equiv \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Omega}^{-1}_{q(h)})$, with mean vector $\boldsymbol{\mu}_{q(h)}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{q(h)} = \boldsymbol{\Omega}^{-1}_{q(h)}$, an iterative algorithm can be set as:*

$$\boldsymbol{\Sigma}^{new}_{q(h)} = \left[\nabla^2_{\boldsymbol{\mu}_{q(h)},\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}^{old}_{q(h)}, \boldsymbol{\Sigma}^{old}_{q(h)})\right]^{-1} \tag{B.16}$$

$$\boldsymbol{\mu}^{new}_{q(h)} = \boldsymbol{\mu}^{old}_{q(h)} + \boldsymbol{\Sigma}^{new}_{q(h)}\nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}^{old}_{q(h)}, \boldsymbol{\Sigma}^{old}_{q(h)}). \tag{B.17}$$

where

$$\nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}^{old}_{q(h)}, \boldsymbol{\Sigma}^{old}_{q(h)}) = -\frac{\boldsymbol{\iota}_n}{2} + \frac{1}{2}\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}^{old}_{q(h)} + \boldsymbol{\sigma}^{2\,old}_{q(h)}/2} - \mu_{q(1/\nu^2)}\mathbf{Q}\boldsymbol{\mu}^{old}_{q(h)}, \tag{B.18}$$

and

$$\nabla^2_{\boldsymbol{\mu}_{q(h)},\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}^{old}_{q(h)}, \boldsymbol{\Sigma}^{old}_{q(h)}) = -\frac{1}{2}\mathsf{Diag}(\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}^{old}_{q(h)} + \boldsymbol{\sigma}^{2\,old}_{q(h)}/2}) - \mu_{q(1/\nu^2)}\mathbf{Q}, \tag{B.19}$$

denote the first and second derivative of $S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$ with respect to $\boldsymbol{\mu}_{q(h)}$ and evaluated at $(\boldsymbol{\mu}^{old}_{q(h)}, \boldsymbol{\Sigma}^{old}_{q(h)})$, and $\boldsymbol{\sigma}^2_{q(h)} = \mathsf{diag}(\boldsymbol{\Sigma}_{q(h)})$.

Proof. The updating scheme follows the algorithm provided in Rohde and Wand (2016) for Gaussian variational approximations. The function S is called *non-entropy function* and it

is given by $S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = \mathbb{E}_q(\log p(\mathbf{y}, \boldsymbol{\vartheta}))$:

$$\begin{aligned} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = &-\frac{\boldsymbol{\iota}_n'}{2}\boldsymbol{\mu}_{q(h)} - \frac{1}{2}\mathbb{E}_q'(\boldsymbol{\varepsilon}^2)\mathrm{e}^{-\boldsymbol{\mu}_{q(h)}+\boldsymbol{\sigma}^2_{q(h)}/2} \\ &- \frac{1}{2}\mu_{q(1/\nu^2)}\left(\boldsymbol{\mu}_{q(h)}'\mathbf{Q}\boldsymbol{\mu}_{q(h)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(h)}\mathbf{Q}\right\}\right), \end{aligned} \tag{B.20}$$

where $\boldsymbol{\varepsilon}^2 = \boldsymbol{\varepsilon} \odot \boldsymbol{\varepsilon}$ with components $[\boldsymbol{\varepsilon}^2]_t = (y_t - \boldsymbol{\beta}_t\mathbf{x}_t)^2$, and $\boldsymbol{\sigma}^2_{q(h)} = \mathsf{diag}(\boldsymbol{\Sigma}_{q(h)})$. Then, the first derivative with respect to the variational mean vector $\boldsymbol{\mu}_{q(h)}$ is given by

$$\nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = -\frac{\boldsymbol{\iota}_n}{2} + \frac{1}{2}\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)}+\boldsymbol{\sigma}^2_{q(h)}/2} - \mu_{q(1/\nu^2)}\mathbf{Q}\boldsymbol{\mu}_{q(h)}. \tag{B.21}$$

Moreover, derive $\nabla_{\boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$ again with respect to $\boldsymbol{\mu}_{q(h)}$:

$$\nabla^2_{\boldsymbol{\mu}_{q(h)}, \boldsymbol{\mu}_{q(h)}} S(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)}) = -\frac{1}{2}\mathsf{Diag}(\mathbb{E}_q(\boldsymbol{\varepsilon}^2) \odot \mathrm{e}^{-\boldsymbol{\mu}_{q(h)}+\boldsymbol{\sigma}^2_{q(h)}/2}) - \mu_{q(1/\nu^2)}\mathbf{Q}. \tag{B.22}$$

□

Remark B.1. *Under the multivariate Gaussian approximation of $q(\mathbf{h})$ with mean vector $\boldsymbol{\mu}_{q(h)}$ and covariance matrix $\boldsymbol{\Sigma}_{q(h)}$, the optimal density of the vector $\boldsymbol{\sigma}^2 = \exp\{\mathbf{h}\}$, namely $q^*(\boldsymbol{\sigma}^2)$, is a multivariate log-normal distribution such that:*

$$\mathbb{E}_q[\sigma_t^2] = \exp\{\mu_{q(h_t)} + 1/2\sigma^2_{q(h_t)}\}, \tag{B.23}$$

$$\mathbb{E}_q[1/\sigma_t^2] = \exp\{-\mu_{q(h_t)} + 1/2\sigma^2_{q(h_t)}\}, \tag{B.24}$$

$$\mathsf{Var}_q[\sigma_t^2] = \exp\{2\mu_{q(h_t)} + \sigma^2_{q(h_t)}\}(\exp\{\sigma^2_{q(h_t)}\} - 1), \tag{B.25}$$

$$\mathsf{Cov}_q[\sigma_t^2, \sigma_{t+1}^2] = \exp\{\mu_{q(h_t)} + \mu_{q(h_{t+1})} + 1/2(\sigma^2_{q(h_t)} + \sigma^2_{q(h_{t+1})})\}(\exp\{\mathsf{Cov}_q[h_t, h_{t+1}]\} - 1).$$

Proposition B.6. *The optimal variational density for the homoskedastic variance σ^2 is an inverse-gamma $q^*(\sigma^2) \equiv \mathsf{IG}(A_{q(\sigma^2)}, B_{q(\sigma^2)})$ where:*

$$A_{q(\sigma^2)} = A_\sigma + \frac{n}{2}, B_{q(\sigma^2)} = B_\sigma + \frac{1}{2}\mathbb{E}_q\left[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right], \tag{B.26}$$

where:

$$\mathbb{E}_{-\sigma^2}\left[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right] = \mathbf{y}'\mathbf{y} - 2\left(\sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_j)}\boldsymbol{\mu}_{q(\mathbf{b}_j)}\right)'\mathbf{y} + \sum_{j=1}^{p}\mathsf{tr}\left\{\left(\boldsymbol{\mu}_{q(\mathbf{b}_j)}\boldsymbol{\mu}'_{q(\mathbf{b}_j)} + \boldsymbol{\Sigma}_{q(\mathbf{b}_j)}\right)\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_j)}\mathbf{X}_j^2\right\} + \sum_{j=1}^{p}\boldsymbol{\mu}'_{q(\mathbf{b}_j)}\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_j)}\mathbf{X}_j\sum_{k=1,k\neq j}^{p}\mathbf{X}_k\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_k)}\boldsymbol{\mu}_{q(\mathbf{b}_k)}.$$

Proof. The full conditional distribution of σ^2 given the rest $p(\sigma^2|\text{rest})$ is derived in Eq.(A.1). Thus, the optimal variational density is given by:

$$\begin{aligned}\log q^*(\sigma^2) &\propto \mathbb{E}_{-\sigma^2}[\log p(\sigma^2|\text{rest})] \\ &\propto -(A_\sigma + \frac{n}{2} + 1)\log\sigma^2 - \frac{1}{\sigma^2}\left\{B_\sigma + \frac{1}{2}\mathbb{E}_{-\sigma^2}\left[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right]\right\},\end{aligned} \tag{B.27}$$

where:

$$\begin{aligned}\mathbb{E}_{-\sigma^2}\left[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right] &= \mathbb{E}_{-\sigma^2}\left[\left(\mathbf{y} - \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\Gamma}_j\mathbf{b}_j\right)'\left(\mathbf{y} - \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\Gamma}_j\mathbf{b}_j\right)\right] = \mathbf{y}'\mathbf{y} - 2\left(\sum_{j=1}^{p}\mathbb{E}_{-\sigma^2}\left[\mathbf{X}_j\boldsymbol{\Gamma}_j\mathbf{b}_j\right]\right)'\mathbf{y} \\ &\quad + \sum_{j=1}^{p}\mathbb{E}_{-\sigma^2}\left[\mathbf{b}_j'\boldsymbol{\Gamma}_j\mathbf{X}_j\mathbf{X}_j\boldsymbol{\Gamma}_j\mathbf{b}_j + \mathbf{b}_j'\boldsymbol{\Gamma}_j\mathbf{X}_j\sum_{k=1,k\neq j}^{p}\mathbf{X}_k\boldsymbol{\Gamma}_k\mathbf{b}_k\right] \\ &= \mathbf{y}'\mathbf{y} - 2\left(\sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_j)}\boldsymbol{\mu}_{q(\mathbf{b}_j)}\right)'\mathbf{y} + \sum_{j=1}^{p}\mathsf{tr}\left\{\mathbb{E}_{\mathbf{b}_j}\left[\mathbf{b}_j\mathbf{b}_j'\right]\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_j)}\mathbf{X}_j^2\right\} \\ &\quad + \sum_{j=1}^{p}\boldsymbol{\mu}'_{q(\mathbf{b}_j)}\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_j)}\mathbf{X}_j\sum_{k=1,k\neq j}^{p}\mathbf{X}_k\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_k)}\boldsymbol{\mu}_{q(\mathbf{b}_k)} \\ &= \mathbf{y}'\mathbf{y} - 2\left(\sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_j)}\boldsymbol{\mu}_{q(\mathbf{b}_j)}\right)'\mathbf{y} + \sum_{j=1}^{p}\mathsf{tr}\left\{\left(\boldsymbol{\mu}_{q(\mathbf{b}_j)}\boldsymbol{\mu}'_{q(\mathbf{b}_j)} + \boldsymbol{\Sigma}_{q(\mathbf{b}_j)}\right)\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_j)}\mathbf{X}_j^2\right\} \\ &\quad + \sum_{j=1}^{p}\boldsymbol{\mu}'_{q(\mathbf{b}_j)}\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_j)}\mathbf{X}_j\sum_{k=1,k\neq j}^{p}\mathbf{X}_k\boldsymbol{\mu}_{q(\boldsymbol{\Gamma}_k)}\boldsymbol{\mu}_{q(\mathbf{b}_k)}.\end{aligned}$$

Equation B.27 represents the kernel of a Inverse-Gamma distribution as in B.6. □

Proposition B.7. *The optimal variational density for the z_{jt} parameters is a Polya-Gamma*

$q^*(z_{jt}) \equiv \mathsf{PG}(1, \sqrt{\mu_{q(\omega_{jt}^2)}})$ *and define*

$$\mu_{q(z_{jt})} = \mathbb{E}_q\left[z_{jt}\right] = \frac{1}{2}\frac{1}{\sqrt{\mu_{q(\omega_{jt}^2)}}}\tanh\left(\frac{\sqrt{\mu_{q(\omega_{jt}^2)}}}{2}\right) \tag{B.28}$$

Proof. The full conditional distribution of z_{jt} given the rest is proportional to Eq.(A.5). Then the optimal variational density is such that

$$\log q^*(z_{jt}) \propto -z_{jt}\mu_{q(\omega_{jt}^2)} + \log p(z_{jt}). \tag{B.29}$$

Equation B.29 represents the kernel of a Polya-Gamma distribution as in B.7. □

Proposition B.8. *The optimal variational density for the variance parameter η_j^2 is an inverse-gamma distribution $q^*(\eta_j^2) \equiv \mathsf{IG}(A_{q(\eta_j^2)}, B_{q(\eta_j^2)})$, where:*

$$A_{q(\eta_j^2)} = A_\eta + \frac{n+1}{2}, \qquad B_{q(\eta_j^2)} = B_\eta + \frac{1}{2}\left(\boldsymbol{\mu}'_{q(\mathbf{b}_j)}\mathbf{Q}\boldsymbol{\mu}_{q(\mathbf{b}_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\mathbf{b}_j)}\mathbf{Q}\right\}\right). \tag{B.30}$$

Proof. The full conditional distribution of η_j^2 given the rest $p(\eta_j^2|\text{rest})$ is described in Eq.(A.6). Then, the optimal variational density is given by:

$$\begin{aligned}
\log q^*(\eta_j^2) &\propto \mathbb{E}_{-\eta_j^2}[\log p(\eta_j^2|\text{rest})] \\
&\propto -\frac{n+1}{2}\log\eta_j^2 - \frac{1}{2\eta_j^2}\mathbb{E}_{-\eta_j^2}\left[\mathbf{b}_j'\mathbf{Q}\mathbf{b}_j\right] - (A_\eta+1)\log\eta_j^2 - \frac{B_\eta}{\eta_j^2} \\
&\propto -\left(\frac{n}{2} + A_\eta + 1\right)\log\eta_j^2 - \frac{1}{\eta_j^2}\left(B_\eta + \frac{1}{2}\left(\boldsymbol{\mu}'_{q(\mathbf{b}_j)}\mathbf{Q}\boldsymbol{\mu}_{q(\mathbf{b}_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\mathbf{b}_j)}\mathbf{Q}\right\}\right)\right).
\end{aligned} \tag{B.31}$$

Equation B.31 represents the kernel of an Inverse-Gaussian distribution as in B.8. □

Proposition B.9. *The optimal variational density for the variance parameter ξ_j^2 is an inverse-gamma distribution $q^*(\xi_j^2) \equiv \mathsf{IG}(A_{q(\xi_j^2)}, B_{q(\xi_j^2)})$, where:*

$$A_{q(\xi_j^2)} = A_\xi + \frac{n+1}{2}, \qquad B_{q(\xi_j^2)} = B_\xi + \frac{1}{2}\left(\boldsymbol{\mu}'_{q(\omega_j)}\mathbf{Q}\boldsymbol{\mu}_{q(\omega_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\omega_j)}\mathbf{Q}\right\}\right). \tag{B.32}$$

Proof. The full conditional distribution of ξ_j^2 given the rest $p(\xi_j^2|\text{rest})$ is described in Eq.(A.7).

Thus, the optimal variational density is given by:

$$\begin{aligned}\log q^*(\xi_j^2) &\propto \mathbb{E}_{-\xi_j^2}[\log p(\xi_j^2|\text{rest})] \\ &\propto -\frac{n+1}{2}\log\xi_j^2 - \frac{1}{2\xi_j^2}\mathbb{E}_{-\xi_j^2}\left[\boldsymbol{\omega}_j'\mathbf{Q}\boldsymbol{\omega}_j\right] - (A_\xi+1)\log\xi_j^2 - \frac{B_\xi}{\xi_j^2} \\ &\propto -(\frac{n}{2}+A_\xi+1)\log\xi_j^2 - \frac{1}{\xi_j^2}\left(B_\xi + \frac{1}{2}\left(\boldsymbol{\mu}_{q(\omega_j)}'\mathbf{Q}\boldsymbol{\mu}_{q(\omega_j)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(\omega_j)}\mathbf{Q}\right\}\right)\right).\end{aligned} \tag{B.33}$$

Equation B.33 represents the kernel of an Inverse-Gaussian distribution as in B.9. □

Proposition B.10. *The optimal variational density for the variance parameter ν^2 is an inverse-gamma distribution $q^*(\nu^2) \equiv \mathsf{IG}(A_{q(\nu^2)}, B_{q(\nu^2)})$, where:*

$$A_{q(\nu^2)} = A_\nu + \frac{n+1}{2}, \qquad B_{q(\nu^2)} = B_\nu + \frac{1}{2}\left(\boldsymbol{\mu}_{q(h)}'\mathbf{Q}\boldsymbol{\mu}_{q(h)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(h)}\mathbf{Q}\right\}\right). \tag{B.34}$$

Proof. The full conditional distribution of ν^2 given the rest $p(\nu^2|\text{rest})$ is described in Eq.(A.8). Thus, the optimal variational density is given by:

$$\begin{aligned}\log q^*(\nu^2) &\propto \mathbb{E}_{-\nu^2}[\log p(\nu^2|\text{rest})] \\ &\propto -\frac{n+1}{2}\log\nu^2 - \frac{1}{2\nu^2}\mathbb{E}_{-\nu^2}\left[\mathbf{h}'\mathbf{Q}\mathbf{h}\right] - (A_\nu+1)\log\nu^2 - \frac{B_\nu}{\nu^2} \\ &\propto -\left(\frac{n}{2}+A_\nu+1\right)\log\nu^2 - \frac{1}{\nu^2}\left(B_\nu + \frac{1}{2}\left(\boldsymbol{\mu}_{q(h)}'\mathbf{Q}\boldsymbol{\mu}_{q(h)} + \mathsf{tr}\left\{\boldsymbol{\Sigma}_{q(h)}\mathbf{Q}\right\}\right)\right).\end{aligned} \tag{B.35}$$

Equation B.35 represents the kernel of an Inverse-Gaussian distribution as in B.10. □

B.1 Smoothing the sparsity dynamics

Proposition B.11. *A smooth estimate for the trajectory of the inclusion probabilities can be achieved assuming $\widetilde{q}(\boldsymbol{\gamma}_j) = \prod_{t=1}^n \widetilde{q}(\gamma_{jt})$ such that $\widetilde{q}(\gamma_{jt}) \equiv \mathsf{Bern}(\text{expit}(\mathbf{w}_t'\mathbf{f}_j))$ with constraints on the mean. Therefore, the expectation of the joint vector $\boldsymbol{\gamma}_j = (\gamma_{j1},\dots,\gamma_{jn})'$ is equal to $\mathbb{E}_{\widetilde{q}}(\boldsymbol{\gamma}_j) = \mathbf{W}\mathbf{f}_j$, where $\mathbf{W}$ is a $n \times k$ B-spline basis matrix. The optimal value of $\mathbf{f}_j$ is the solution of the optimization problem $\widehat{\mathbf{f}}_j = \arg\max_{\mathbf{f}_j\in\mathbb{R}^k}\psi(\mathbf{f}_j)$ where $\psi(\mathbf{f}_j) = \sum_{t=1}^n\left[(\omega_{q(\gamma_{jt})} - \mathbf{w}_t'\mathbf{f}_j)\text{expit}(\mathbf{w}_t'\mathbf{f}_j) + \log(1+\exp(\mathbf{w}_t'\mathbf{f}_j))\right]$, such that the gradient is equal to $\nabla_{\mathbf{f}}\psi(\mathbf{f}) = \sum_{t=1}^n \mathbf{w}_t(\omega_{q(\gamma_{jt})} - \mathbf{w}_t'\mathbf{f})\frac{\text{expit}(\mathbf{w}_t'\mathbf{f})}{1+\exp(\mathbf{w}_t'\mathbf{f})}$.*

Proof. To find the best $\widetilde{q}$ that approximates q, minimize the Kullback-Leibler divergence $\mathcal{KL}(\widetilde{q} \,||\, q)$. This corresponds to maximize $\mathbb{E}_{\widetilde{q}}[\log q] - \mathbb{E}_{\widetilde{q}}[\log\widetilde{q}]$ over the parameters of the

approximating density $\widetilde{q}$. In our case we obtain:

$$\begin{aligned}\hat{\mathbf{f}} = \arg\max_{\mathbf{f}\in\mathbb{R}^k} \psi(\mathbf{f}) &= \arg\max_{\mathbf{f}\in\mathbb{R}^k} \left\{\mathbb{E}_{\widetilde{q}}[\log q(\boldsymbol{\gamma}_j)] - \mathbb{E}_{\widetilde{q}}[\log \widetilde{q}(\boldsymbol{\gamma})]\right\} \\ &= \arg\max_{\mathbf{f}\in\mathbb{R}^k} \sum_{t=1}^{n} \left\{\mathbb{E}_{\widetilde{q}}[\log q(\gamma_{jt})] - \mathbb{E}_{\widetilde{q}}[\log \widetilde{q}(\gamma_{jt})]\right\}\end{aligned}$$

and define $\psi_t(\mathbf{f}) = \mathbb{E}_{\widetilde{q}}[\log q(\gamma_{jt})] - \mathbb{E}_{\widetilde{q}}[\log \widetilde{q}(\gamma_{jt})]$. The first term is equal to:

$$\mathbb{E}_{\widetilde{q}}[\log q(\gamma_{jt})] = \mathbb{E}_{\widetilde{q}}[\gamma_{jt}\omega_{q(\gamma_{jt})}] = \omega_{q(\gamma_{jt})}\text{expit}(\mathbf{w}_t'\mathbf{f}),$$

while the second one can be written as:

$$\begin{aligned}\mathbb{E}_{\widetilde{q}}[\log \widetilde{q}(\gamma_{jt})] &= \mathbb{E}_{\widetilde{q}}[\gamma_{j,t}\mathbf{w}_t'\mathbf{f} - \log(1+\exp(\mathbf{w}_t'\mathbf{f}))] \\ &= \mathbf{w}_t'\mathbf{f}\text{expit}(\mathbf{w}_t'\mathbf{f}) - \log(1+\exp(\mathbf{w}_t'\mathbf{f})).\end{aligned}$$

Group together and obtain:

$$\psi_t(\mathbf{f}) = (\omega_{q(\gamma_{jt})} - \mathbf{w}_t'\mathbf{f})\text{expit}(\mathbf{w}_t'\mathbf{f}) + \log(1+\exp(\mathbf{w}_t'\mathbf{f})).$$

which defines the t component of the loss function in the Proposition. Now derive $\psi(\mathbf{f})$ with respect to $\mathbf{f}$:

$$\nabla_{\mathbf{f}}\psi(\mathbf{f}) = \frac{\partial}{\partial \mathbf{f}}\psi(\mathbf{f}) = \sum_{t=1}^{n}\frac{\partial}{\partial \mathbf{f}}\psi_t(\mathbf{f}).$$

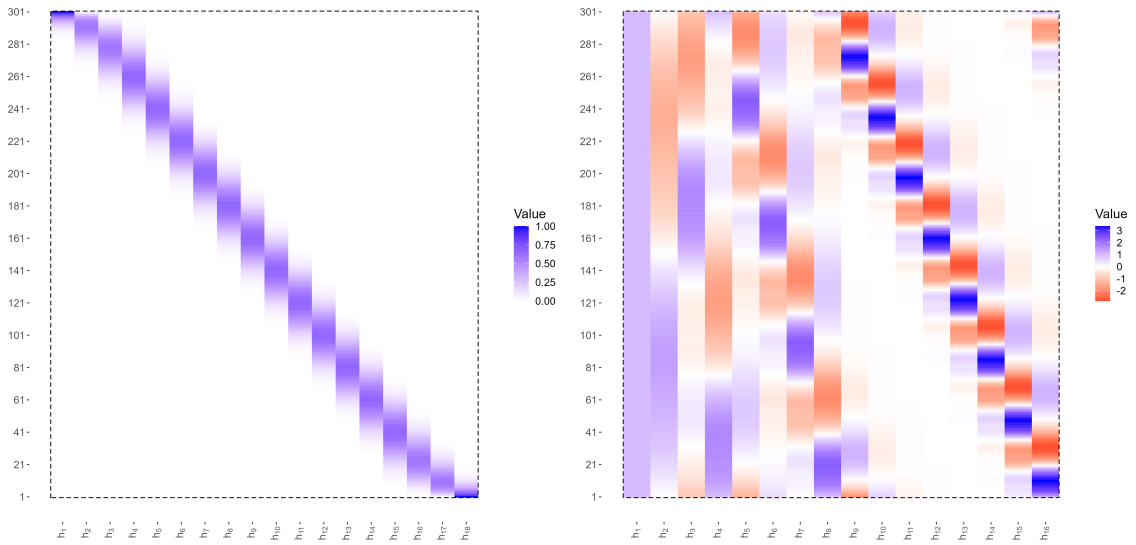
Compute the latter and get:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{f}}\psi_t(\mathbf{f}) &= -\mathbf{w}_t\text{expit}(\mathbf{w}_t'\mathbf{f}) + \mathbf{w}_t(\omega_{q(\gamma_{jt})} - \mathbf{w}_t'\mathbf{f})\frac{\text{expit}(\mathbf{w}_t'\mathbf{f})}{1+\exp(\mathbf{w}_t'\mathbf{f})} + \mathbf{w}_t\text{expit}(\mathbf{w}_t'\mathbf{f}) \\ &= \mathbf{w}_t(\omega_{q(\gamma_{jt})} - \mathbf{w}_t'\mathbf{f})\frac{\text{expit}(\mathbf{w}_t'\mathbf{f})}{1+\exp(\mathbf{w}_t'\mathbf{f})},\end{aligned}$$

which completes the proof. □

Alternative smoothing assumptions. Figure B.1 depicts the form of $\mathbf{W}$ when B-spline and Daubechies wavelets are used. The form of $\mathbf{W}$ in case of B-spline basis functions (top) and wavelet basis functions (bottom). Right panels correspond to columns of the matrix

W. The B-spline basis functions is a sequence of piecewise polynomial functions of a given degree, in this case $dg = 3$. The locations of the pieces are determined by the knots, here we assume $kn = 20$ equally spaced knots. The functions that compose the wavelet basis matrix **W** are constructed over equally spaced grids on $[0, n]$ of length R, where R is called resolution and it is equal to 2^{l-1}, where l defines the level, and as a result the degree of smoothness. The number of functions at level l is then equal to R and they are defined as dilatation and/or shift of a more general *mother* function.



(a) B-spline basis with $kn = 20$ and $dg = 3$ (b) Daubechies wavelet basis matrix

Figure B.1: Smoothing time-varying parameters. The columns of the matrix **W** in case of B-spline basis functions (left panel) or wavelet basis functions (right panel).

C Proofs of the theoretical properties

Before discussing the theoretical properties of our algorithmic procedure, we need to provide definitions and lemmas which are instrumental for the proof.

Definition C.1. **A** *is a Z-matrix if its off-diagonal elements satisfy* $a_{i,j} \leq 0$*, for* $i \neq j$*.*

Definition C.2. **A** *is a strictly diagonally dominant (SDD) matrix if, for each* i *row of* **A***,* $|a_{i,i}| > \sum_{j \neq i} |a_{i,j}|$*.*

Corollary C.1. *If a matrix* **A** *is SDD and all its diagonal elements* $a_{i,i}$ *are positive, then the real parts of its eigenvalues are positive.*

Definition C.3. *A matrix* $\mathbf{A}$ *is said to be an M-matrix if it is a strictly diagonally dominant Z-matrix and all its diagonal elements* $a_{i,i}$ *are positive.*

Corollary C.2. *If a matrix* $\mathbf{A}$ *is an M-matrix, then it belongs to inverse-positive matrices, i.e all elements of the inverse are positive* $[\mathbf{A}^{-1}]_{i,j} \geq 0$*, for all* (i, j)*.*

Lemma C.1. *The matrix* $\mathbf{Q}^{-1}$ *is a positive matrix, i.e* $[\mathbf{Q}^{-1}]_{i,j} \geq 0$*.*

Proof. Follows from the tridiagonal form of $\mathbf{Q}$ with $q_{1,1} = 1 + 1/k_0$, and $k_0 < +\infty$. □

Lemma C.2. *The matrix* $\mathbf{\Sigma}_{q(\omega_j)}$ *is a positive matrix, i.e* $[\mathbf{\Sigma}_{q(\omega_j)}]_{i,j} \geq 0$*.*

Proof. Recall that

$$\mathbf{\Sigma}_{q(\omega_j)} = \mathbf{W}^{-1} = \left(\mathsf{Diag}\left(0, \mathbb{E}_q\left[\mathbf{z}_j\right]\right) + \mu_{q(1/\xi_j^2)}\mathbf{Q}\right)^{-1}, \tag{C.1}$$

is tridiagonal, where $\mathbb{E}_q\left[z_{jt}\right] > 0$ and $\mu_{q(1/\xi_j^2)} > 0$. Notice that $\mathbf{W}$ has off-diagonal elements equal to $-\mu_{q(1/\xi_j^2)} < 0$ in the first sub/over-diagonal and 0 elsewhere and therefore it is a Z-matrix. Moreover, $w_{t,t} > 0$ for all t and:

$$w_{1,1} = (1 + k_0^{-1})\mu_{q(1/\xi_j^2)} > \mu_{q(1/\xi_j^2)} = |w_{1,2}| \tag{C.2}$$

$$w_{t,t} = 2\mu_{q(1/\xi_j^2)} + \mathbb{E}_q\left[z_{jt}\right] > 2\mu_{q(1/\xi_j^2)} = |w_{t,t-1}| + |w_{t,t+1}|, \quad t = 2, \ldots, n \tag{C.3}$$

$$w_{n+1,n+1} = \mu_{q(1/\xi_j^2)} + \mathbb{E}_q\left[z_{jn}\right] > \mu_{q(1/\xi_j^2)} = |w_{n+1,n-1}|, \tag{C.4}$$

thus $\mathbf{W}$ is SDD with positive diagonal elements. Hence, by definition C.3 is an M-matrix and corollary C.2 tells us that its inverse is a positive matrix. □

Proposition C.1. *Assume that the maximum over time of the inclusion probabilities, for a given variable* j*, at the* i*-th iteration of the algorithm is such that* $\max_{t\in\{1,\ldots,n\}} \mu^{(i)}_{q(\gamma_{jt})} = \epsilon$*, and* $\epsilon \ll 1$ *is small enough. Moreover, let* $\mathbf{\Sigma}^{(i)}_{q(\omega_j)} - \mathbf{\Sigma}^{(i-1)}_{q(\omega_j)} \geq 0$*, then:*

1. $\mu^{(i+1)}_{q(\gamma_{jt})} = \text{expit}\left\{\mu^{(i+1)}_{q(\omega_{jt})} - \frac{1}{2}\mu^{(i+1)}_{q(1/\sigma_t^2)} x^2_{jt-1}\mu^{-1(i+1)}_{q(1/\eta_j^2)} q_{tt} + O(\epsilon)\right\}$, $q_{tt} = [\mathbf{Q}^{-1}]_{tt} \geq 0$*;*
2. $\mu^{(i+1)}_{q(\omega_{jt})} = -1/2\sum_{k=1}^{n} s_{tk} + O(\epsilon)$, $s_{tk} = [\mathbf{\Sigma}_{q(\omega_j)}]_{tk} \geq 0$*;*
3. $\mu^{(i+1)}_{q(\omega_{jt})} \leq \mu^{(i)}_{q(\omega_{jt})}$ *decreases after each iteration.*

Proof. We start proving *1)*. Consider the update for $\mu^{(i+1)}_{q(\gamma_{jt})}$:

$$\mu^{(i+1)}_{q(\gamma_{jt})} = \text{expit}\left\{\mu^{(i+1)}_{q(\omega_{jt})} - 1/2\mu^{(i+1)}_{q(1/\sigma_t^2)}\left(\mathbb{E}^{(i+1)}_q[b^2_{jt}]x^2_{jt-1} - 2\mu^{(i+1)}_{q(b_{jt})}x_{jt-1}\mathbb{E}^{(i+1)}_q[\varepsilon_{jt}]\right)\right\}. \tag{C.5}$$

Notice that the vector for all times $\boldsymbol{\mu}^{(i+1)}_{q(\mathbf{b}_j)}$ has the following formula:

$$\boldsymbol{\mu}^{(i+1)}_{q(\mathbf{b}_j)} = \left(\mu^{(i+1)}_{q(1/\sigma_t^2)}(\widetilde{\mathbf{D}}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta_j^2)}\mathbf{Q}\right)^{-1} \mu^{(i+1)}_{q(1/\sigma_t^2)}\widetilde{\mathbf{D}}^{(i)}_{\gamma_j}\boldsymbol{\mu}^{(i+1)}_{q(\widetilde{\varepsilon}_{-j})}, \tag{C.6}$$

where $\widetilde{\mathbf{D}}_{\gamma_j} = \mathsf{Diag}((0, \boldsymbol{\mu}_{q(\gamma_j)}) \odot (0, \mathbf{x}_j))$ and $\widetilde{\mathbf{D}}^2_{\gamma_j} = \mathsf{Diag}((0, \boldsymbol{\mu}_{q(\gamma_j)}) \odot (0, \mathbf{x}_j \odot \mathbf{x}_j))$. Notice we can write each $\mu^{(i+1)}_{q(\gamma_{jt})} = \alpha_t\epsilon$, with $0 < \alpha_t \leq 1$. Now define $\boldsymbol{\alpha}$ the collection of the α_t, and $\mathbf{A}_{\gamma_j} = \mathsf{Diag}((0, \boldsymbol{\alpha}) \odot (0, \mathbf{x}_j))$ and $\mathbf{A}^2_{\gamma_j} = \mathsf{Diag}((0, \boldsymbol{\alpha}) \odot (0, \mathbf{x}_j \odot \mathbf{x}_j))$, such that Eq.C.6 can be written as

$$\boldsymbol{\mu}^{(i+1)}_{q(\mathbf{b}_j)} = \left(\mu^{(i+1)}_{q(1/\sigma_t^2)}(\epsilon\mathbf{A}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta_j^2)}\mathbf{Q}\right)^{-1} \mu^{(i+1)}_{q(1/\sigma_t^2)}\epsilon\mathbf{A}^{(i)}_{\gamma_j}\boldsymbol{\mu}^{(i+1)}_{q(\widetilde{\varepsilon}_{-j})}, \tag{C.7}$$

and

$$\lim_{\epsilon\to 0}\frac{\boldsymbol{\mu}^{(i+1)}_{q(\mathbf{b}_j)}}{\epsilon} < \infty \quad \Longrightarrow \quad \mu^{(i+1)}_{q(b_{jt})} = O(\epsilon). \tag{C.8}$$

Consider now the variance matrix $\boldsymbol{\Sigma}^{(i+1)}_{q(\mathbf{b}_j)}$:

$$\boldsymbol{\Sigma}^{(i+1)}_{q(\mathbf{b}_j)} = \left(\mu^{(i+1)}_{q(1/\sigma_t^2)}(\epsilon\mathbf{A}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta_j^2)}\mathbf{Q}\right)^{-1} = f(\epsilon), \tag{C.9}$$

as a scalar to matrix function f with

$$f'(\epsilon) = \left(\mu^{(i+1)}_{q(1/\sigma_t^2)}(\epsilon\mathbf{A}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta_j^2)}\mathbf{Q}\right)^{-1}\left(\mu^{(i+1)}_{q(1/\sigma_t^2)}(\mathbf{A}^{2(i)}_{\gamma_j})\right)\left(\mu^{(i+1)}_{q(1/\sigma_t^2)}(\epsilon\mathbf{A}^{2(i)}_{\gamma_j}) + \mu^{(i)}_{q(1/\eta_j^2)}\mathbf{Q}\right)^{-1}.$$

Using Taylor expansion in $\epsilon = 0$ we obtain:

$$\boldsymbol{\Sigma}^{(i+1)}_{q(\mathbf{b}_j)} = \left(\mu^{(i)}_{q(1/\eta_j^2)}\mathbf{Q}\right)^{-1} + \epsilon\left(\mu^{(i)}_{q(1/\eta_j^2)}\mathbf{Q}\right)^{-1}\left(\mu^{(i+1)}_{q(1/\sigma_t^2)}(\mathbf{A}^{2(i)}_{\gamma_j})\right)\left(\mu^{(i)}_{q(1/\eta_j^2)}\mathbf{Q}\right)^{-1} + \dots$$

and therefore each diagonal element is $\sigma^{2(i+1)}_{q(b_{jt})} = \left[\mu^{(i)}_{q(1/\eta_j^2)}\right]^{-1} q_{t,t} + O(\epsilon)$ and it follows that

$$\mathbb{E}^{(i+1)}_q[b^2_{jt}] = (\mu^{(i+1)}_{q(b_{jt})})^2 + \sigma^{2(i+1)}_{q(b_{jt})} = \left[\mu^{(i)}_{q(1/\eta_j^2)}\right]^{-1} q_{t,t} + O(\epsilon). \tag{C.10}$$

Put together (C.8) and (C.10) completes the proof. Similarly we prove *2)*. Recall the function to jointly update $\boldsymbol{\mu}^{(i+1)}_{q(\omega_j)}$:

$$\boldsymbol{\mu}^{(i+1)}_{q(\omega_j)} = \boldsymbol{\Sigma}^{(i+1)}_{q(\omega_j)}\left(0, \boldsymbol{\mu}^{(i)\prime}_{q(\gamma_j)} - 1/2\boldsymbol{\iota}'_n\right)', \tag{C.11}$$

then the update of the t-th component is:

$$\begin{aligned}\mu^{(i+1)}_{q(\omega_{jt})} &= \mathbf{s}'_t \left(0, \boldsymbol{\mu}^{(i)\prime}_{q(\gamma_j)} - 1/2\boldsymbol{\iota}'_n\right)' \\ &= -1/2\mathbf{s}'_t \left(0, \boldsymbol{\iota}'_n\right)' + \mathbf{s}'_t \left(0, \boldsymbol{\mu}^{(i)\prime}_{q(\gamma_j)}\right)' \\ &= -1/2 \sum_{k=1}^{n} s_{tk} + \sum_{k=1}^{n} s_{tk} \mu^{(i)}_{q(\gamma_{jk})}, \end{aligned} \tag{C.12}$$

where $\mathbf{s}_t$ denotes the t-th column in $\boldsymbol{\Sigma}^{(i+1)}_{q(\omega_j)}$. Notice that, since $\mu^{(i)}_{q(\gamma_{jk})} \leq \epsilon$, for all k, we can write $\mu^{(i)}_{q(\gamma_{jk})} = \alpha_k \epsilon$, where $0 < \alpha_k \leq 1$. If we plug-in the latter in Eq.C.12 we get

$$\mu^{(i+1)}_{q(\omega_{j,t})} = -1/2 \sum_{k=1}^{n} s_{tk} + \epsilon \sum_{k=1}^{n} \alpha_k s_{tk} = -1/2 \sum_{k=1}^{n} s_{tk} + O(\epsilon). \tag{C.13}$$

To prove the last statement *3)*, assume that we observe $\boldsymbol{\Sigma}^{(i)}_{q(\omega_j)} - \boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)}$ positive matrix. Then we have that, for ϵ small:

$$|\boldsymbol{\mu}^{(i)}_{q(\omega_j)}| = \frac{1}{2}\boldsymbol{\Sigma}^{(i)}_{q(\omega_j)} \left(0, \boldsymbol{\iota}'_n\right)' \geq \frac{1}{2}\boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)} \left(0, \boldsymbol{\iota}'_n\right)' = |\boldsymbol{\mu}^{(i-1)}_{q(\omega_j)}|, \tag{C.14}$$

and therefore:

$$\mathbb{E}^{(i)}_q(\boldsymbol{\omega}_j \boldsymbol{\omega}'_j) = \boldsymbol{\mu}^{(i)}_{q(\omega_j)} (\boldsymbol{\mu}^{(i)}_{q(\omega_j)})' + \boldsymbol{\Sigma}^{(i)}_{q(\omega_j)} \geq \boldsymbol{\mu}^{(i-1)}_{q(\omega_j)} (\boldsymbol{\mu}^{(i-1)}_{q(\omega_j)})' + \boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)} = \mathbb{E}^{(i-1)}_q(\boldsymbol{\omega}_j \boldsymbol{\omega}'_j), \tag{C.15}$$

which means that $\mathbb{E}^{(i)}_q(\boldsymbol{\omega}_j \boldsymbol{\omega}'_j) - \mathbb{E}^{(i-1)}_q(\boldsymbol{\omega}_j \boldsymbol{\omega}'_j)$ is a positive matrix. Consider now the update for the variable z_{jt}:

$$\mathbb{E}^{(i)}_q [z_{jt}] = \frac{1}{2} \frac{1}{\sqrt{\mathbb{E}^{(i)}_q(\omega^2_{jt})}} \tanh(\frac{\sqrt{\mathbb{E}^{(i)}_q(\omega^2_{jt})}}{2}) \leq \mathbb{E}^{(i-1)}_q [z_{jt}], \tag{C.16}$$

since it is decreasing in $\mathbb{E}^{(i)}_q(\omega^2_{jt})$, for all t. And similarly for ξ^2_j:

$$\mu^{(i)}_{q(1/\xi^2_j)} = \frac{A_\xi + \frac{n+1}{2}}{B_\xi + \frac{1}{2}\mathsf{tr}\left\{\mathbb{E}^{(i)}_q(\boldsymbol{\omega}_j \boldsymbol{\omega}'_j)\mathbf{Q}\right\}} \leq \mu^{(i-1)}_{q(1/\xi^2_j)}, \tag{C.17}$$

since it is decreasing in $\mathbb{E}^{(i)}_q(\boldsymbol{\omega}_j \boldsymbol{\omega}'_j)$ and $\mathbb{E}^{(i)}_q(\boldsymbol{\omega}_j \boldsymbol{\omega}'_j) - \mathbb{E}^{(i-1)}_q(\boldsymbol{\omega}_j \boldsymbol{\omega}'_j)$ is a positive matrix. The

next update of $\boldsymbol{\Sigma}_{q(\omega_j)}$ is equal to:

$$\boldsymbol{\Sigma}_{q(\omega_j)}^{(i+1)} = \left(\text{Diag}\left(0, \mathbb{E}_q^{(i)}\left[\mathbf{z}_j\right]\right) + \mu_{q(1/\xi_j^2)}^{(i)}\mathbf{Q}\right)^{-1}, \tag{C.18}$$

which increases as both $\mathbb{E}_q^{(i)}\left[\mathbf{z}_j\right]$ and $\mu_{q(1/\xi_j^2)}^{(i)}$ decreases. Hence also $\boldsymbol{\Sigma}_{q(\omega_j)}^{(i+1)} - \boldsymbol{\Sigma}_{q(\omega_j)}^{(i)}$ is a positive matrix and therefore, for ϵ small:

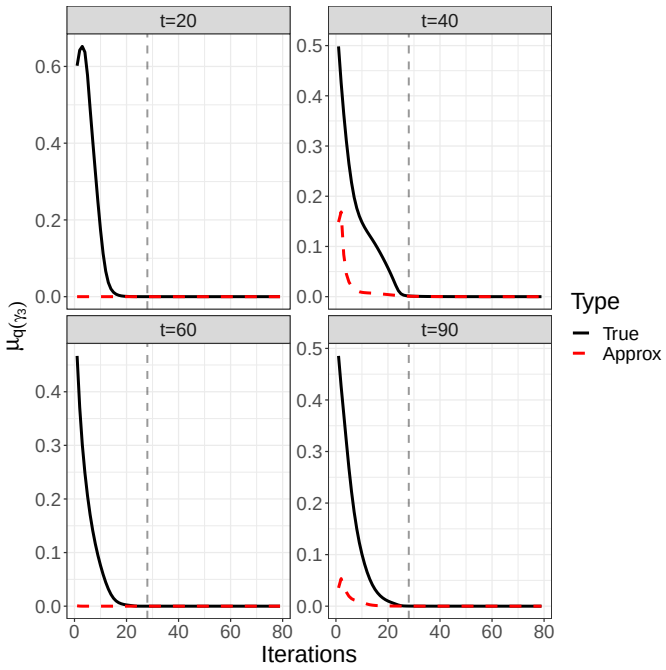
$$|\boldsymbol{\mu}_{q(\omega_j)}^{(i+1)}| \geq |\boldsymbol{\mu}_{q(\omega_j)}^{(i)}|, \tag{C.19}$$

and from statement *2)* we have that $\boldsymbol{\mu}_{q(\omega_j)}^{(i+1)} \leq \boldsymbol{\mu}_{q(\omega_j)}^{(i)}$. Set $i = i + 1$ and repeat the procedure from Eq.C.14. We can see that $\boldsymbol{\mu}_{q(\omega_j)}$ decreases after each iteration until convergence. □

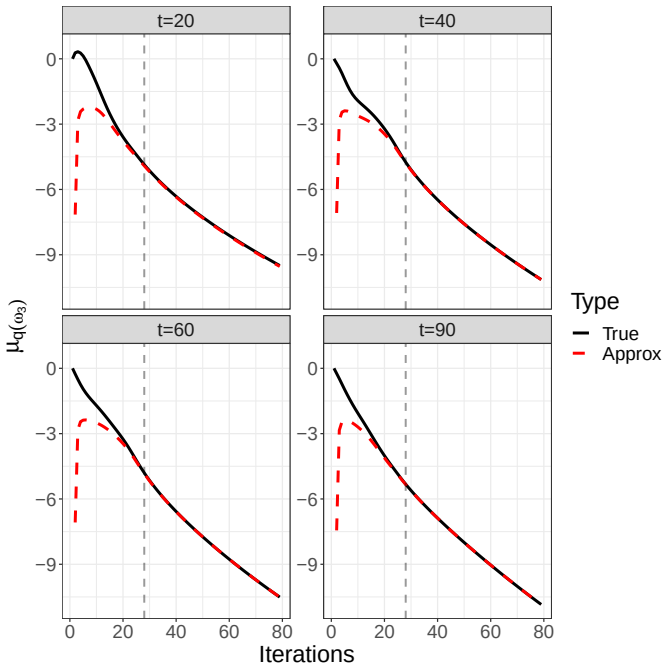
C.1 Additional convergence results

In this section we report the variational update over iterations until convergence of two key parameters to model the dynamics of sparsity, namely the auxiliary process ω_{jt} and the resulting posterior inclusion probability $\mathbb{P}(\gamma_{jt} = 1)$. Figure 2(a) reports the convergence of the algorithm updates for the posterior inclusion probability $\mu_{q(\gamma_{jt})}$, for some times t and for a parameter j which is always zero $\forall t$. This corresponds to β_{3t} in the simulation example of Section 3.1. The true update (solid black line) is compared to the approximation described in Proposition 2.7 (red-dashed line). The vertical dashed line identifies the iteration at which the conditions of Proposition 2.7 are satisfied for $\epsilon = 0.01$. Notice that the approximation is exact after the dashed line and the value of $\mu_{q(\gamma_{jt})}$ is exactly equal to zero, meaning that we induce sparsity in the posterior estimates as highlighted in Equation 19 in the main text.

Similarly, Figure 2(b) shows that the approximating variational update for $\mu_{q(\omega_{jt})}$ as from Proposition 2.7 quickly converge to the true update with convergence that is reached after less than 30 iterations for $\epsilon = 0.01$. The convergence of the updates translate in a rapid convergence of the posterior estimates to the actual sparsity dynamics. Figure C.3 reports the posterior estimates of $\mu_{q(\gamma_{jt})}$ (left panel) and $\mu_{q(\omega_{jt})}$ (right panel) across $t = 1, \ldots, n$ and for a parameter j which is significant for only part of the sample. This corresponds to β_{2t} in the simulation example of Section 3.1. The value of the update is given by the color intensity. After less than 30 iterations the posterior estimates of ω_{jt} and γ_{jt} quickly converge to their true values. This threshold corresponds to the iteration at which the conditions of Proposition 2.7 are satisfied for $\epsilon = 0.01$. In this respect, Figure C.3 complements Figure 2 in showing the oracle properties of our variational Bayes inference approach.



(a) Convergence of updates for $\mu_{q(\gamma_{jt})}$



(b) Convergence of updates for $\mu_{q(\omega_{jt})}$

Figure C.2: Variational update over iterations until convergence of the posterior inclusion probability $\mu_{q(\gamma_{jt})}$ (left panel) and the auxiliary parameter $\mu_{q(\omega_{jt})}$ (right panel), for some times t and for a parameter j which is always zero $\forall t$.

D Additional simulation results

D.1 Comparison with MCMC

In this section we report additional details on the simulation study with respect to the comparison between our variational Bayes inference approach and an MCMC equivalent. In particular, we report the accuracy of $q^*(b_{jt})$ and $q^*(\gamma_{jt})$ in approximating $p(b_{jt}|\mathbf{y})$ and $p(\gamma_{jt}|\mathbf{y})$ for $j = 1, 2, 3$, across simulations.

D.2 Comparison with existing variable selection methods

In this section we report additional details on the simulation setting implemented to compare our `BG` model vs existing variable selection methods, as well as additional simulation results for the intermediate dimension case with $p = 100$ predictors. Figure D.5 reports examples of trajectories for the time-varying intercept which is always included β_{1t}, a dynamic coefficient $\beta_{2:3,t}$ with a single switch from $\gamma_{jt} = 0$ to $\gamma_{jt} = 1$, a more complex pattern $\beta_{4:5,t}$ with two

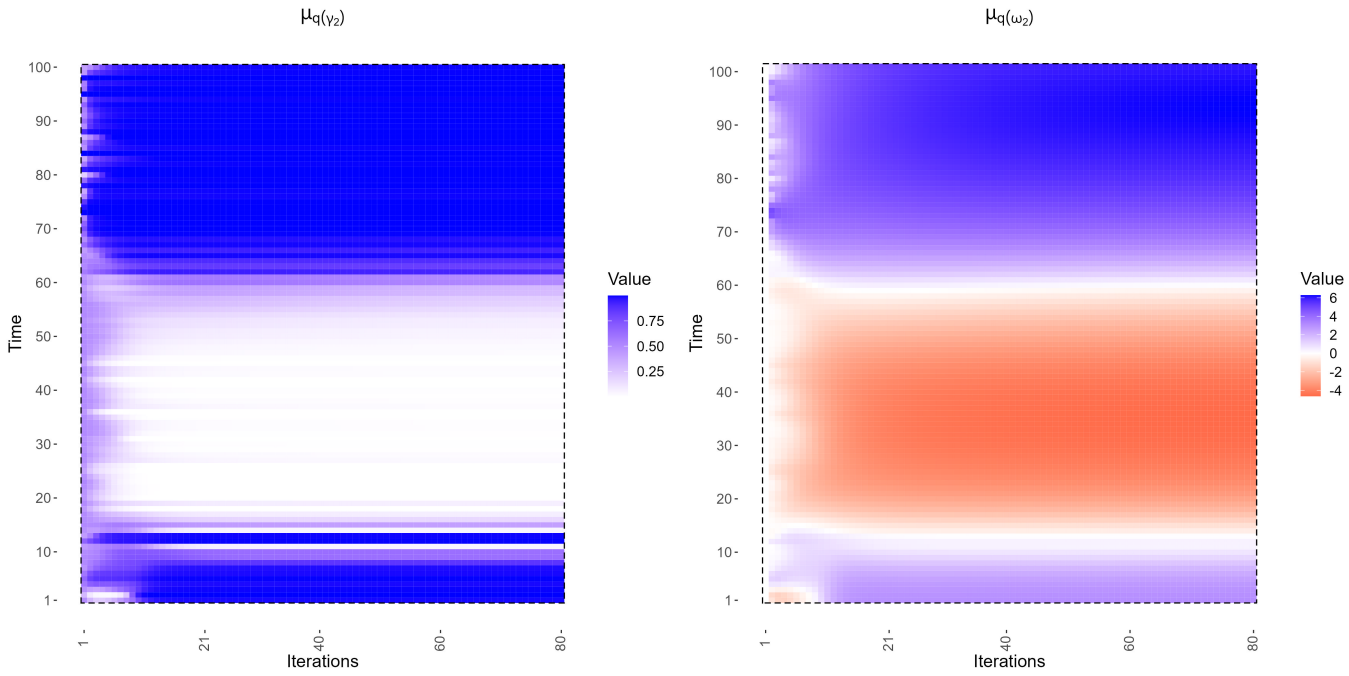
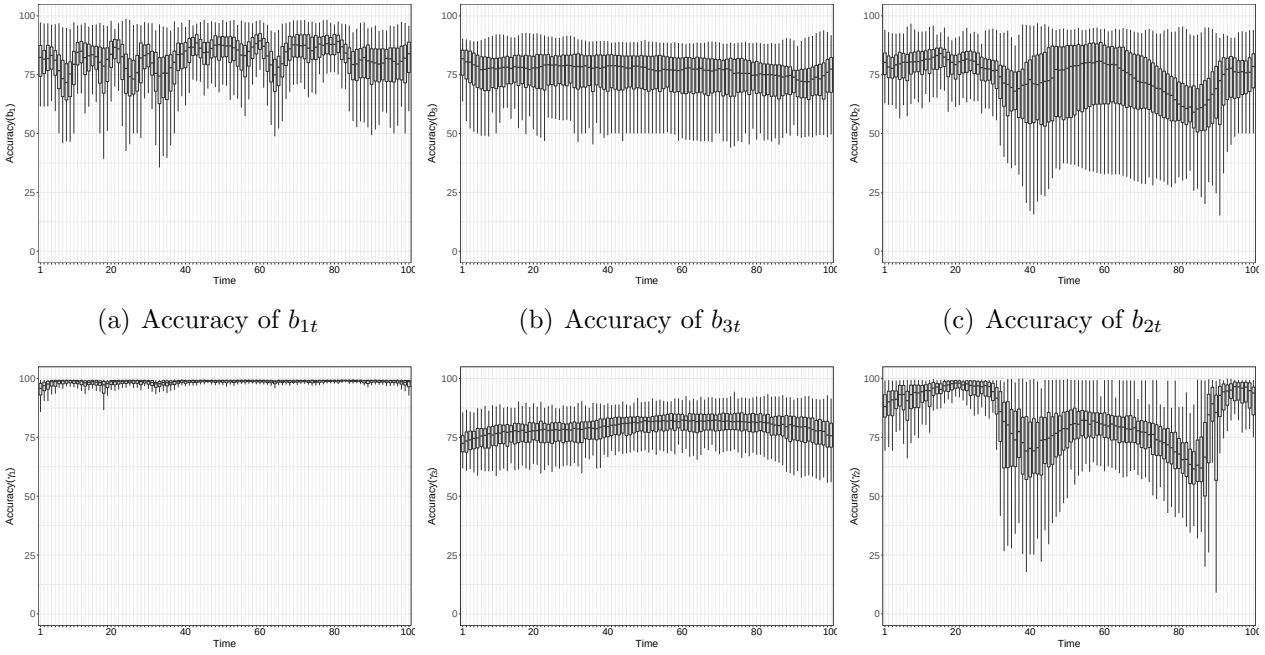


Figure C.3: Left panel shows the variational update over iterations (x-axis) until convergence of the vector of posterior inclusion probabilities $(\mu_{q(\gamma_{j1})}, \ldots, \mu_{q(\gamma_{jn})})$ (y-axis), for a parameter j which is always zero $\forall t$. The value of the update is given by the color intensity. The right panel depicts the behaviour of $\mu_{q(\omega_{jt})}$, $\forall t$.

switches from $\gamma_{jt} = 0$ to $\gamma_{jt} = 1$ and vice-versa, and a short-lived regression coefficient $\beta_{6:7,t}$ which is significant only for a short fraction of the sample.

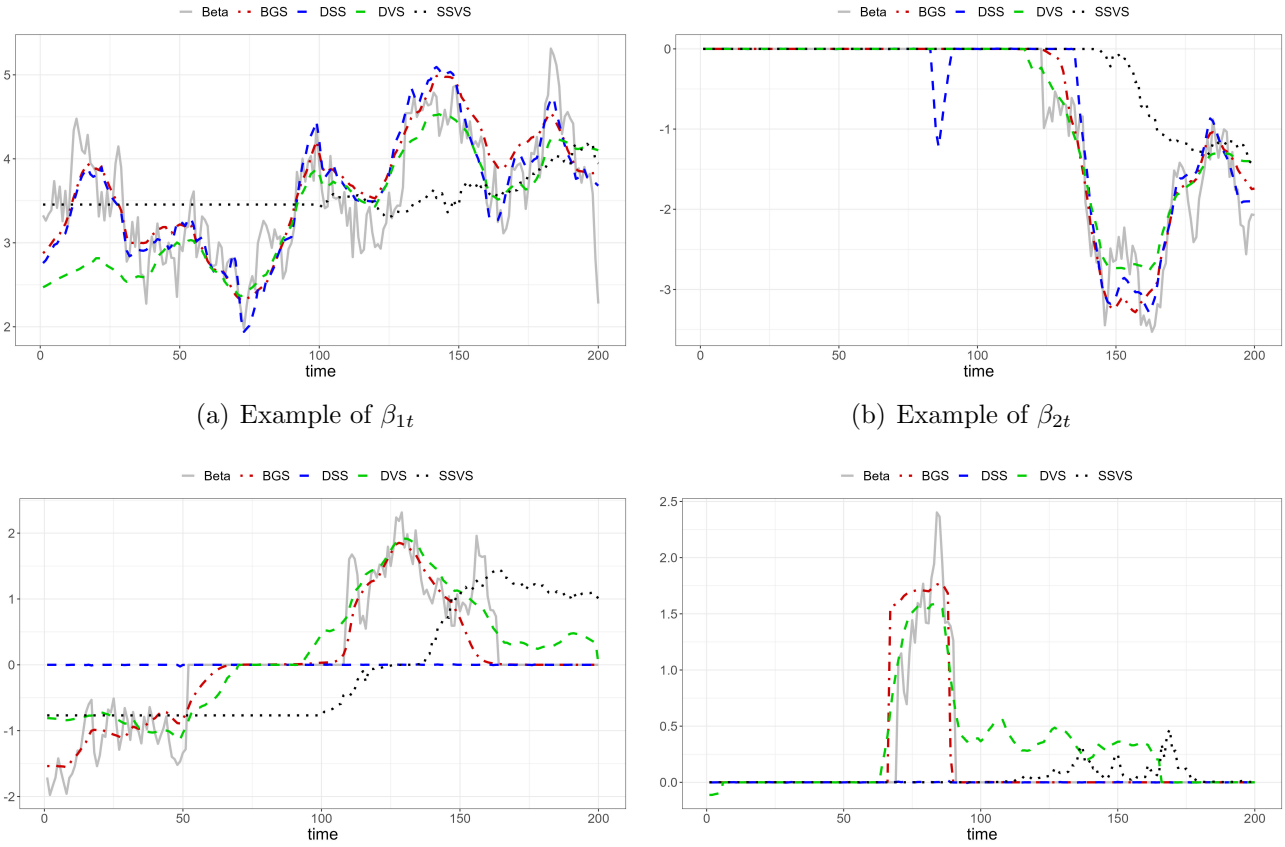
Figure D.6 reports the F1-score for β_{jt} with $j = 1, 2, 3, 4, 5, 6, 7$ and $t = 1, \ldots, 200$ when $p = 100$. A full description of the model set used for comparison against out dynamic `BG` model is in Section 3.2 in the main text. Similarly to the main simulation results, the performance of static variable selection models quickly deteriorates as the pattern of variable significance becomes more complex. For instance, for a parameter with multiple episodes of being "active" in the set of predictors the median accuracy of leading variable selection methods such as the rolling-window estimates of the spike-and-slab `SSVS` and the EM spike-and-slab `EMVS` of Ročková and George (2018) is around 50%. This compares to a more solid 75% on average for alternative dynamic variable selection methods such `DVS` from Koop and Korobilis (2020) and `DSS` from Ročková and McAlinn (2021). More importantly, the performance of our dynamic `BG` specification remains quite stable across specifications, with a clear edge in terms of signal identification of the homoschedastic version `BGH` and the `BG`



(a) Accuracy of b_{1t} (b) Accuracy of b_{3t} (c) Accuracy of b_{2t}

(d) Accuracy of γ_{1t} (e) Accuracy of γ_{3t} (f) Accuracy of γ_{2t}

Figure D.4: Comparison between the approximating variational Bayes estimate and its equivalent MCMC posterior draws. The left panels compare the accuracy of $q^*(b_{jt})$ in approximating $p(b_{jt}|\mathbf{y})$ for $j = 1, 2, 3$ across simulations. The right panels compare the accuracy of $q^*(\gamma_{jt})$ in approximating $p(\gamma_{jt}|\mathbf{y})$ for $j = 1, 2, 3$ across simulations.



(a) Example of β_{1t}

(b) Example of β_{2t}

(c) Example of β_{4t}

(d) Example of β_{6t}

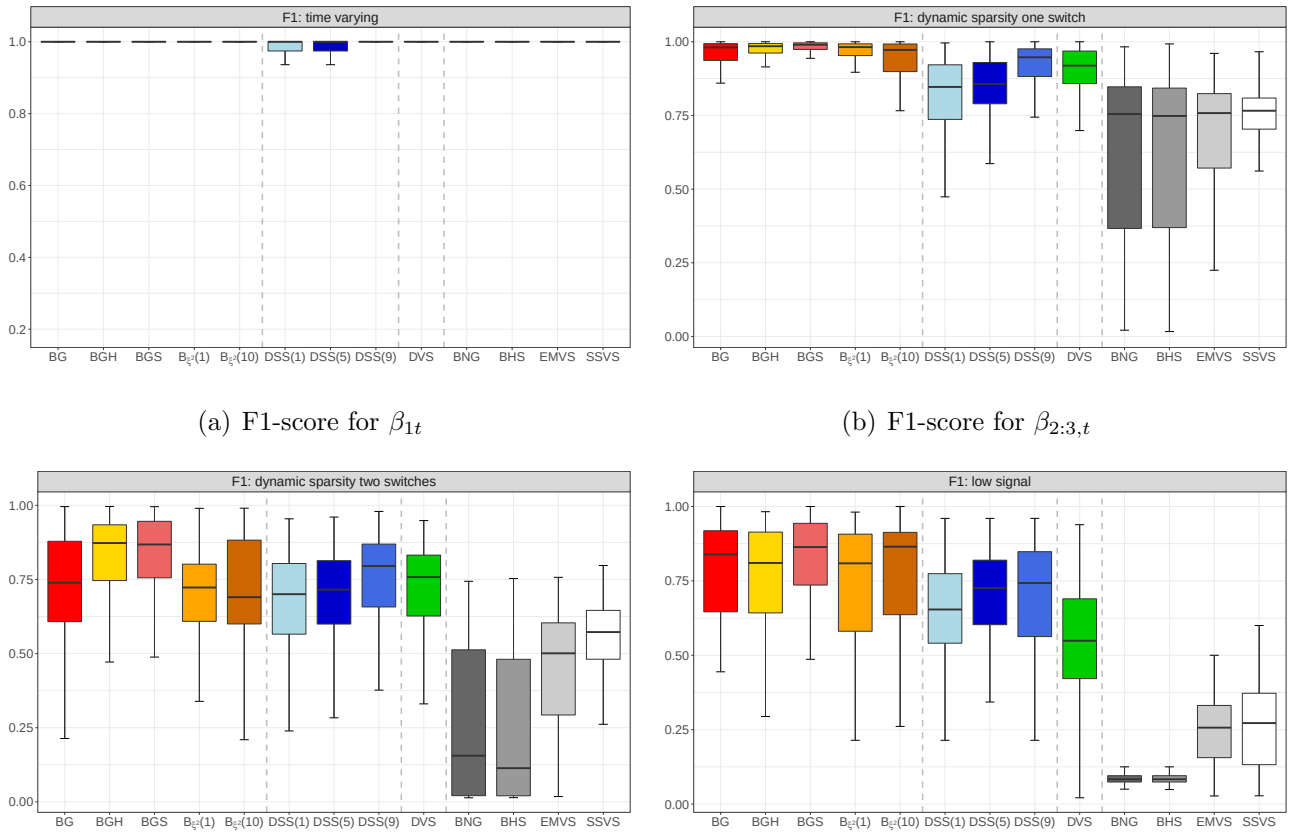
Figure D.5: Example of single simulated trajectories for β_{jt} with $j = 1, 2, 4, 6$ and $t = 1, \ldots, 200$. See description in the main text for how different coefficient dynamics are generated.

with smoothed inclusion probabilities `BGS`.

Figure D.7 reports the point accuracy as measured by the mean-squared error (MSE). The latter represents the squared distance between the true parameters β_{jt}, $t = 1, \ldots, n$ observed at each simulation and its corresponding posterior estimate $\widehat{\beta}_{jt}$. The results for $p = 100$ broadly confirms what we have observed in the main text (see Section 3.2). That is, our dynamic `BG` framework outperforms both static and dynamic sparsity inducing priors, especially within the context of complex dynamics such as $\beta_{2:3,t}$ and $\beta_{4:5,t}$.

D.3 Correlated predictors and computational speed

In this Section, we are going to investigate the robustness of our dynamic variable selection strategy with respect to different correlation assumptions on the predictors. The setting



(a) F1-score for β_{1t}

(b) F1-score for $\beta_{2:3,t}$

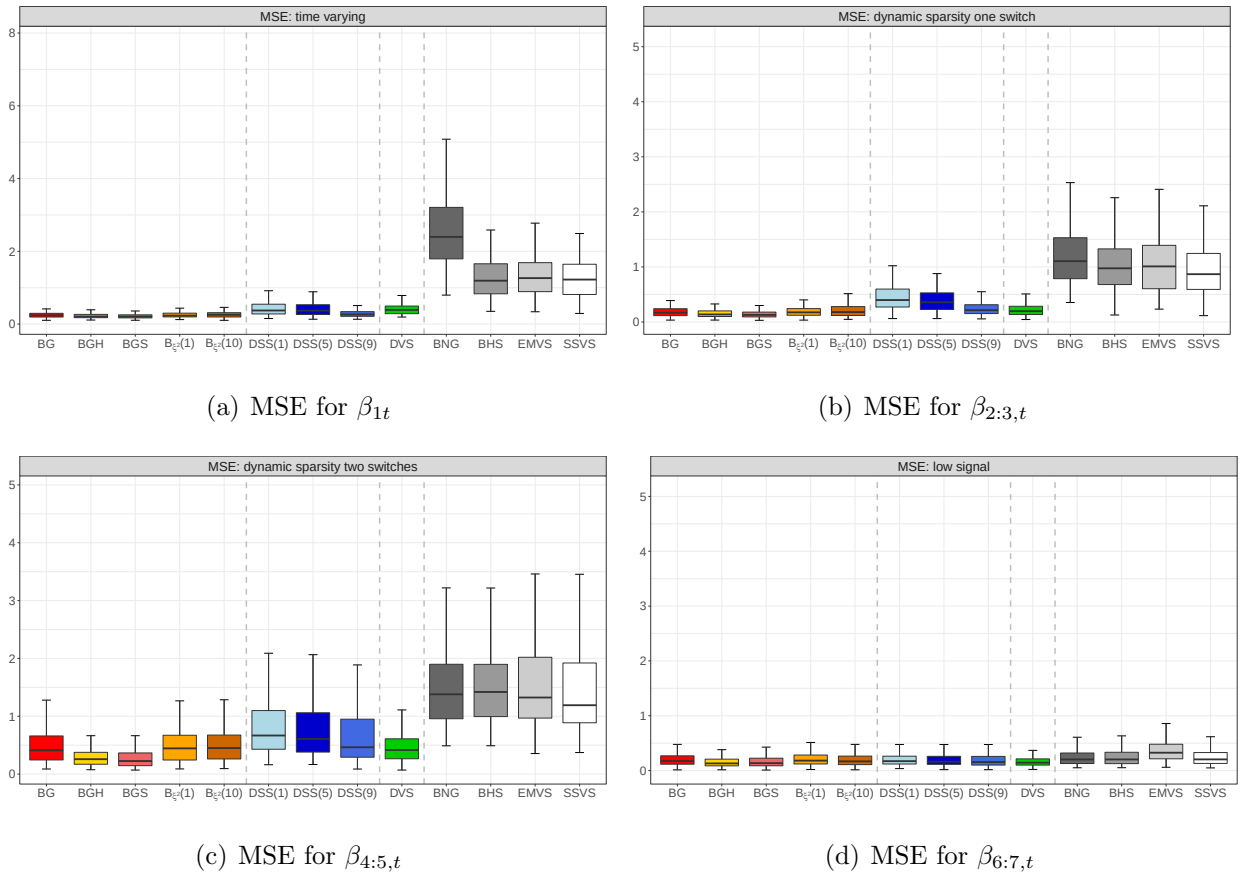
(c) F1-score for $\beta_{4:5,t}$

(d) F1-score for $\beta_{6:7,t}$

Figure D.6: F1-score for β_{jt} with $j = 1, 2, 3, 4, 5, 6, 7$ and $t = 1, \ldots, 200$ when $p = 100$.

is the same as the one outlined in Section 3.2, with the difference that we now consider for the $x_{jt},\ j = 1, \ldots, 7$ – meaning those predictors that carry a signal, either continuous or intermittent – different auto-correlation (left panel) or cross-correlation (right panel) assumptions. Specifically, we consider from 0.5 to 0.9 auto-correlation and from low to high cross-sectional correlations. The latter is based on modifying the Cholesky decomposition of a multivariate Gaussian distribution of $x_{jt},\ j = 1, \ldots, 7$.

Not surprisingly, the accuracy of the dynamic variable selection deteriorates as the cross-sectional correlation among predictors increases. For instance, Figure 8(b) shows that for the pattern in which a predictor switches from zero to be significant only once, the median F1-score goes from essentially one to 0.8/0.9, on average across methods. The deterioration in the performance in more pronounced for short-lived signals. For instance, for the pattern in which a predictor becomes significant only for a short period of time (bottom panels), the median F1-score decreases from 0.8 for the low-correlation case, on average across methods,



(a) MSE for β_{1t}

(b) MSE for $\beta_{2:3,t}$

(c) MSE for $\beta_{4:5,t}$

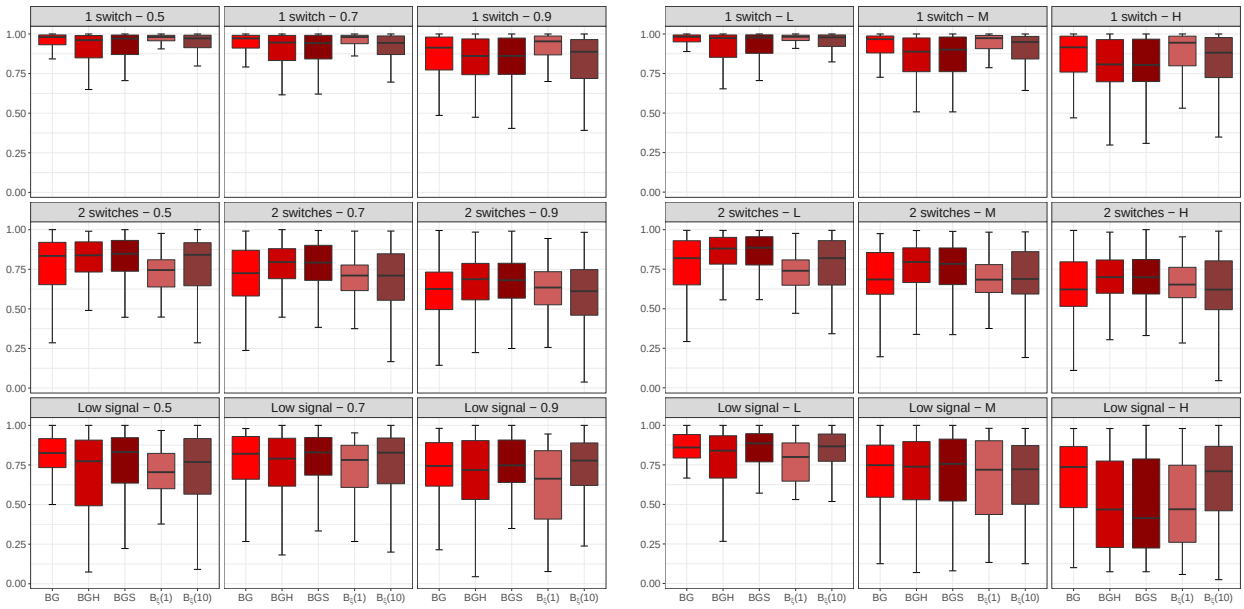
(d) MSE for $\beta_{6:7,t}$

Figure D.7: MSE for β_{jt} with $j = 1, 2, 3, 4, 5, 6, 7$ and $t = 1, \ldots, 200$ when $p = 100$.

to 0.6 for the high-correlation case.

On the other hand, Figure 8(a) shows that our dynamic variable selection is quite robust to autocorrelation in the predictors. For instance, the median F1-score remains quite intact for a 0.5 vs 0.9 autocorrelation in the one-switch pattern. This holds also for the short-lived signal (bottom panels). However, for the multiple switches from significant to non-significant and vice-versa, autocorrelation do plays a more relevant role, as highlighted in the middle panels. Nevertheless, Figures 8(b)-8(a) shows that with the exception of extremely high level of autocorrelation and cross-sectional correlation, our dynamic variable selection method is reasonably robust.

The last additional simulation result concerns a comparison in terms of computational efficiency between our variational Bayes strategy and existing dynamic variable selection methods for large-scale regressions. To evaluate computational costs across methods, we track the running time in seconds of each algorithm. Figure D.9 highlights two main con-



(a) Auto-correlation

(b) Cross-correlation

Figure D.8: F1-score for the different patterns and different assumptions on auto-correlation (left panel) and cross-correlation (right panel) for the predictors x_{jt}, $j = 1, \ldots, 7$ in the simulation exercise as outlined in Section 3.2.

clusions. First of all, the `DSS` is the slowest algorithm, regardless to the dimension of the parameter p. Instead, when $p = 50$, the `DVS` is faster that `BG`, but when we move towards higher dimensions the situation changes. Secondly, and perhaps more interesting, our algorithm scales linearly rather than exponentially as the model size p increases. This is a direct consequence of the embedded dimension reduction property as highlighted in Section 2.2.

E Additional empirical results

In this Section we are going to discuss some of the additional empirical results which have not been included in the main text for the sake of brevity. We separate the additional results between the inflation forecasting and the stock returns predictability applications.

In-sample estimates. The left panel of Figure E.10 reports the reports the time-varying posterior inclusion probabilities and posterior regression coefficients estimates $\mu_{q(\beta_{jt})}$ from our dynamic `BG` model for the core CPI (CPILFESL). Two comments are in order. First, the results show that as far as core CPI is concerned, only lagged prices and real consumption

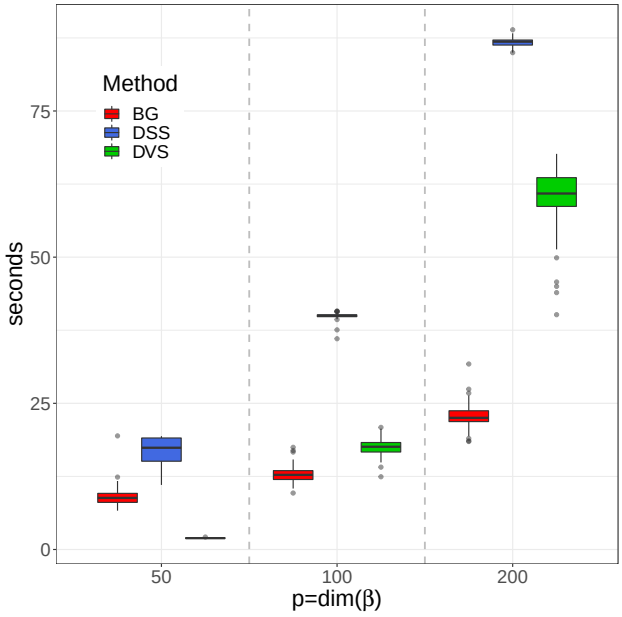
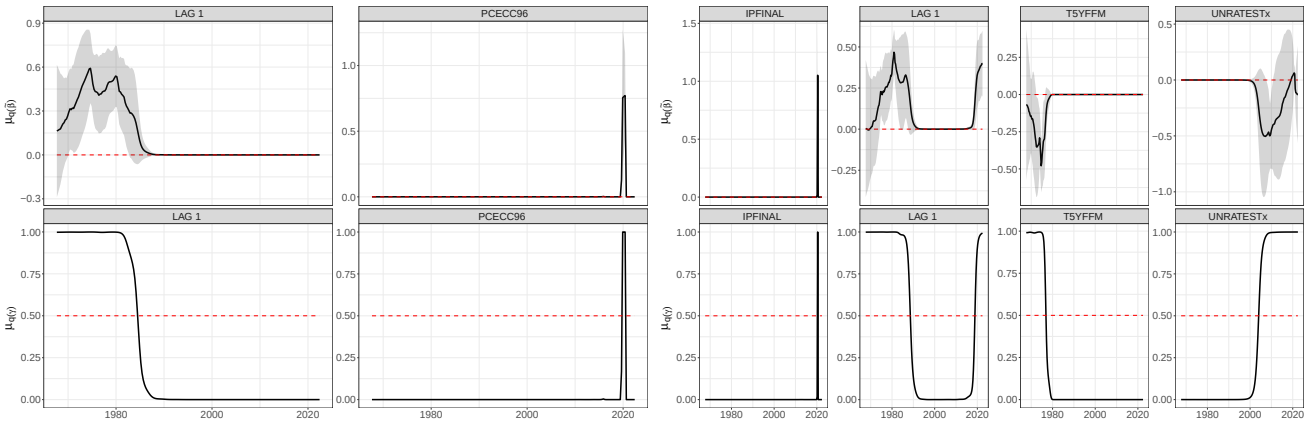


Figure D.9: Computational efficiency of the algorithms computed as running time in second, varying the dimension p.

expenditures (PCECC96) actually carry significant explanatory power above and beyond the conditional mean. Second, our dynamic `BG` model is able to capture short-lived predictors with a meaningful economic significance. This is the case of PCECC96 towards the end of 2020, which highlight the importance of demand pressure via fiscal stimulus on inflation.



(a) Core CPI (b) GDP deflator

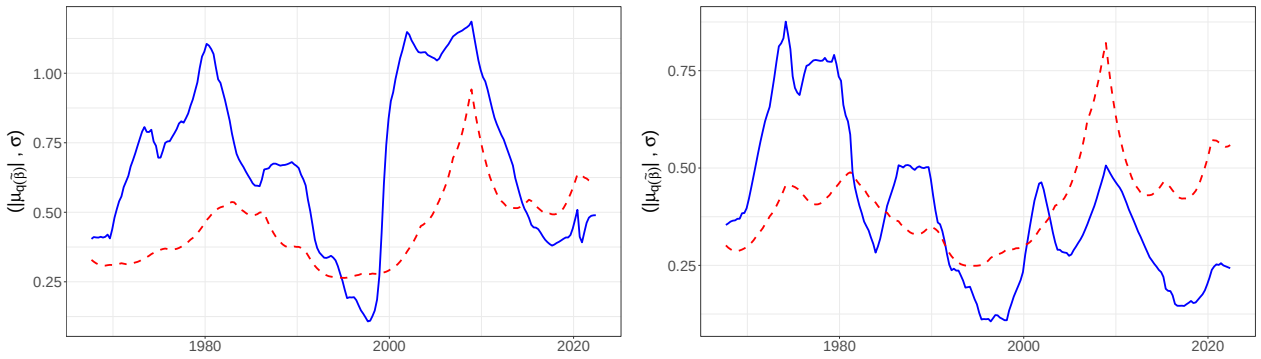
Figure E.10: Time-varying coefficients estimates $\mu_{q(\beta_{jt})}$ and posterior inclusion probabilities for core CPI (CPILFESL) and GDP deflator (GDPCTPI).

The right panel of Figure E.10 reports the dynamics of the predictors when the target variable is the one-quarter ahead GDP deflator (GDPCTPI). Again, our model is able to capture short-lived predictors which contributed to the inflationary shock during the Covid-19 crisis. For instance, increasing production of final products (IPFINAL) during late 2020 positively correlate with increasing inflation – as measured by GDPCTPI – towards the 2021. Yet, lagged prices and monetary policy, as proxied by the 5-year treasury rate, played a significant role until late 1990s/early 2000s. Interestingly, our dynamic `BG` model picks UNRATETESTx – which measures the unemployment rate for less than 27 weeks of unemployment – as significant predictor from the great financial crisis towards the end of the sample. This evidence in favour of a time-varying Phillips curve, whereby the inverse relationship between unemployment and inflation is corroborated by the data only during specific time periods.

Figure E.11 reports the sum of absolute values of the variational mean of the active regression coefficients, i.e., $\sum_{j=1}^{p} |\mu_{q(\beta_{jt})}|$, which proxies the strength of the information available to predict inflation. The information from the predictors clearly change over time, decreasing in the middle part of the sample, from the 90s to early 2000. In addition, a stronger signal from the predictors correlates with higher idiosyncratic volatility (dashed-red line); that is, a richer model is needed to predict inflation at times of higher uncertainty as proxied by the volatility in the residuals.

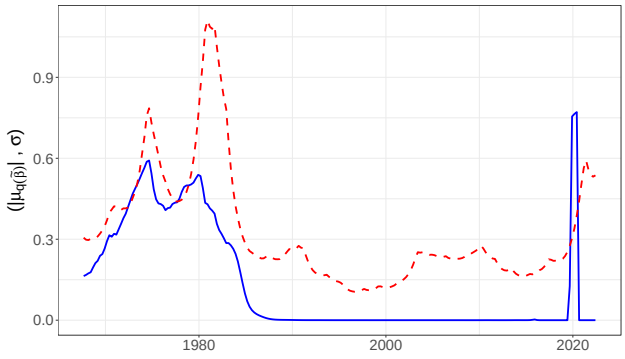
Out-of-sample forecasting. In this Section, we compare the performance of different forecasting models based on the relative mean absoluted error (RMAE) calculated as $RMAE_i = \sum_{t=\tau}^{T} |e_{i,t}| - \sum_{t=\tau}^{T} |e_{\text{bench},t}|$, where τ denotes the beginning of the out-of-sample period, and $|e_{i,t}|$, $|e_{\text{bench},t}|$ the absolute value of the forecast errors from a competing model and a benchmark specification. Figure 12(a) reports the results for inflation forecasting in which the benchmark is the `UC` model of Stock and Watson (2007). Consistent with the main empirical results in Section 4.1, our model outperforms all competing static and dynamic variable selection methods across forecasting horizons and inflation measures.

Figure 12(b) reports the results for the aggregate returns on the stock market. Consistent with the main empirical results in Section 4.2, we consider a naive forecast from the recursive sample mean as a benchmark (see, e.g., Campbell and Thompson, 2008; Welch and Goyal, 2008). The results confirm that our dynamic `BG` model outperforms most of the competing static and dynamic variable selection approaches.

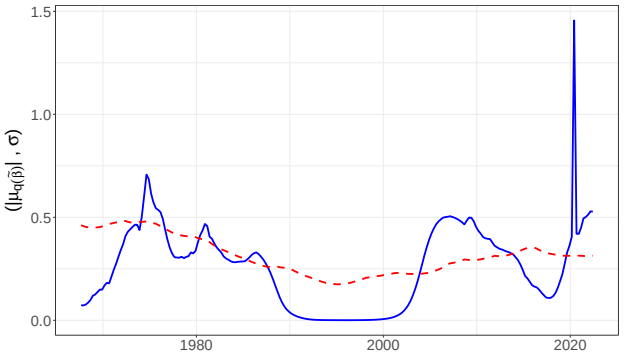


(a) Total CPI (CPIAUCSL)

(b) PCE deflator (PCECTPI)

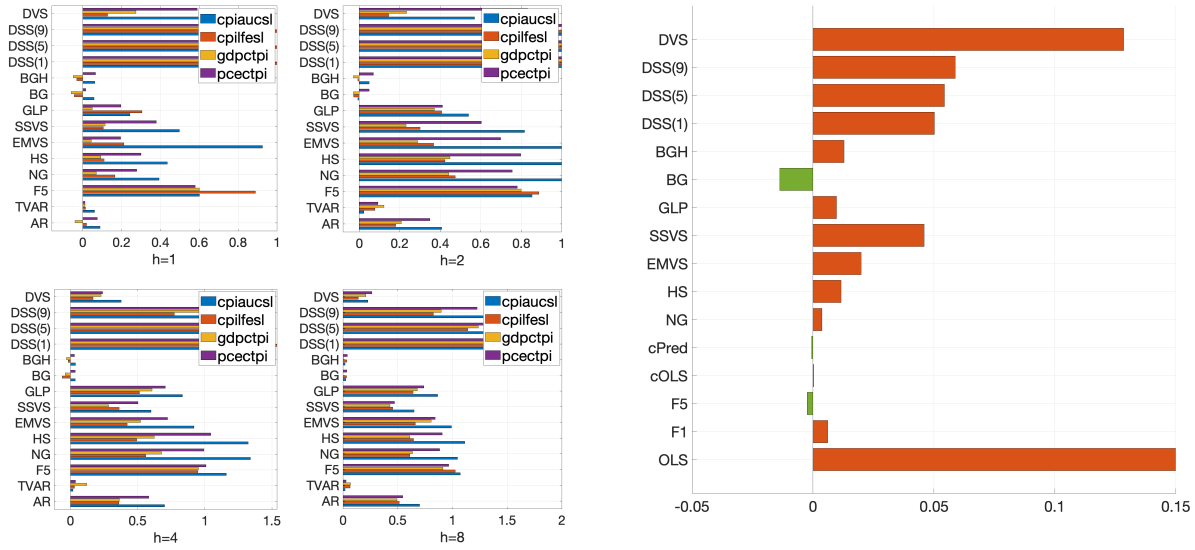


(c) Core CPI (CPILFESL)



(d) GDP deflator (GDPCTPI)

Figure E.11: Signal (blue) computed as $\sum_{j=1}^{p} |\mu_{q(\beta_{jt})}|$, for $t = 1, \ldots, n$, against the posterior estimates of stochastic volatility $\exp(h_t/2)$, for $t = 1, \ldots, n$ (dashed-red).



(a) Inflation

(b) Stock market

Figure E.12: Relative mean absolute error with respect to the benchmark. Left panel reports the results for inflation forecasting whereby the benchmark is the unobserved component benchmark `UC`. The sample period is from 1967Q3 to 2022Q3. The first prediction is generated in 1997Q3. Right panel reports the results for forecasting the excess returns on the stock market. The sample period is from 1971M11 to 2021M12. The first prediction is generated in 1991M01.