Dynamic variable selection in high-dimensional predictive regressions

Forecasting @ Risk, ECB

Daniele Bianchi[1] & Mauro Bernardi[2] & Nicolas Bianco

[1] School of Economics and Finance, Queen Mary, University of London
[2] Department of Statistical Sciences, University of Padua, Italy
[3] Department of Economics and Business, Universitat Pompeu Fabra, Spain.

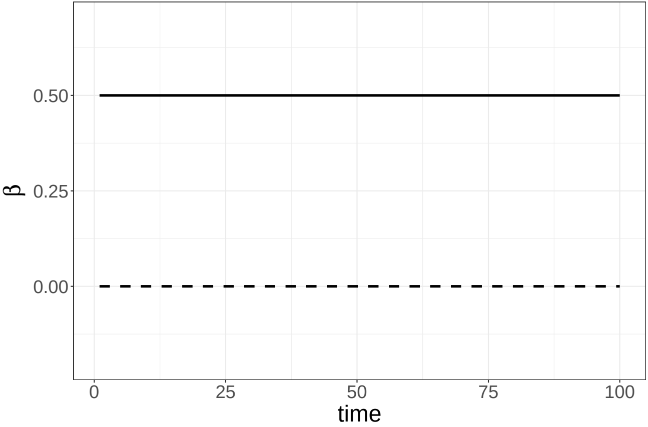
Set the stage

Set the stage

$$y_t = \beta \; x_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_t^2)$$

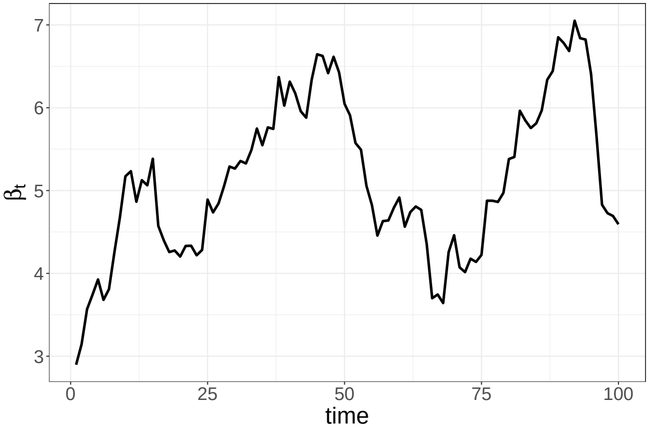
Set the stage

$$y_t = \beta\ x_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2)$$



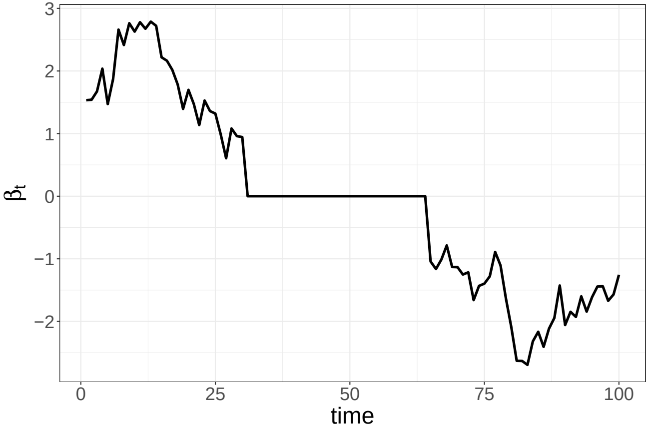
Set the stage

$$y_t = \beta_t x_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2)$$



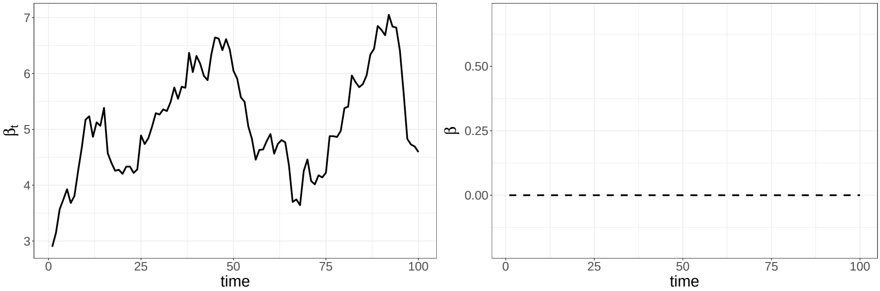
Set the stage

$$y_t = \beta_t x_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2)$$



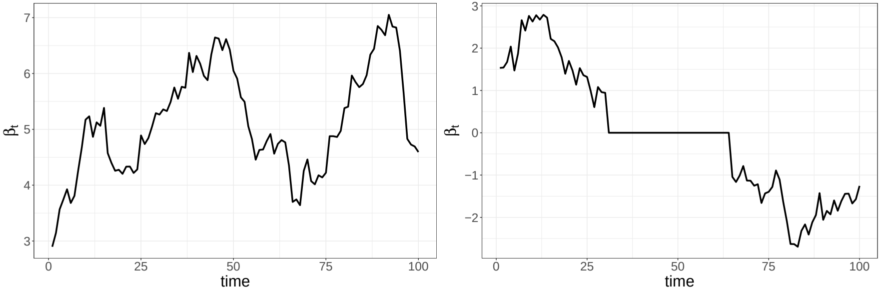
Set the stage

$$y_t = \sum_{j=1}^{p} \color{red}{\beta_{jt}} x_{jt-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_t^2)$$



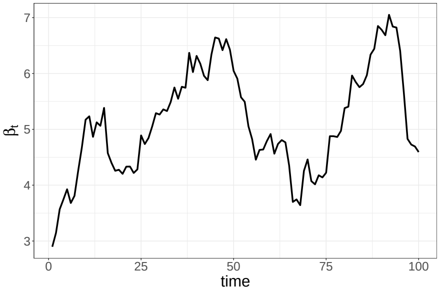
Set the stage

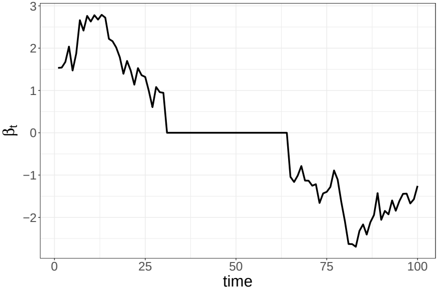
$$y_t = \sum_{j=1}^{p} \beta_{jt} x_{jt-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2)$$



Set the stage

$$y_t = \sum_{j=1}^{p} \beta_{jt} x_{jt-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2)$$





Key issues:

- ↪ The set p could be large ($p \to T$).
- ↪ Which predictor matters and when is unknown a priori...
- ↪ ...this matters even more in dynamic settings.

This paper

Bayesian method for dynamic variable selection in high-dimensional time-varying predictive regressions.

This paper

Bayesian method for dynamic variable selection in high-dimensional time-varying predictive regressions.

$$\Downarrow$$

A single predictor can be "active"
over multiple time periods.
$\hookrightarrow$ *Intensive margin*.

This paper

Bayesian method for dynamic variable selection in high-dimensional time-varying predictive regressions.

$\Downarrow$ $\Downarrow$

A single predictor can be "active" over multiple time periods.
$\hookrightarrow$ *Intensive margin*.

Multiple predictors can be "active" in a single time period.
$\hookrightarrow$ *Extensive margin*.

This paper

Bayesian method for dynamic variable selection in high-dimensional time-varying predictive regressions.

$\Downarrow$ $\Downarrow$

A single predictor can be "active" over multiple time periods.
$\hookrightarrow$ *Intensive margin*.

Multiple predictors can be "active" in a single time period.
$\hookrightarrow$ *Extensive margin*.

Method $\Longrightarrow$ Novel variational Bayes inference approach:

$\hookrightarrow$ Minimal hyper-parameters tuning.
$\hookrightarrow$ Posterior concentration properties comparable to MCMC.
$\hookrightarrow$ On-line dimension reduction (efficiency).

This paper

Bayesian method for dynamic variable selection in high-dimensional time-varying predictive regressions.

$\Downarrow$ $\Downarrow$

A single predictor can be "active" over multiple time periods.
$\hookrightarrow$ *Intensive margin*.

Multiple predictors can be "active" in a single time period.
$\hookrightarrow$ *Extensive margin*.

Empirical exercise(s) $\Longrightarrow$ Economic forecasting (point and density):

$\hookrightarrow$ Inflation predictability over several quarterly horizons based on $+230$ macroeconomic predictors (FRED-QD data).

$\hookrightarrow$ Equity risk premium predictability one-month ahead based on $+150$ anomaly/characteristic-based portfolios.

Some reference literature

A non-exhaustive list of references:

↪ Bayesian methods for variable/model selection:
(e.g., Mitchell & Beauchamp (1988), George & McCulloch (1997), Nakajima & West (2013), Kalli & Griffin (2014), Kowal et al. (2019), Bitto & Frühwirth-Schnatter (2019), Koop & Korobilis (2020), Ročková & McAlinn (2021), Giannone et al. (2021), etc.)

↪ Economic forecasting in large-dimensional models:
(e.g., Stock & Watson (2007), Stock & Watson (2010), Faust & Wright (2013), Huber et al. (2021), Dong et al. (2022), etc.)

↪ Variational Bayes inference methods:
(e.g., Ormerod & Wand (2010), Ormerod et al. (2017), Gefang et al. (2019), Koop & Korobilis (2020), Chan & Yu (2022), etc.)

Model specification

Bayesian model specification

Dynamic Bernoulli-Gaussian (BG) regression specification:

$$y_t = \sum_{j=1}^{p} \beta_{jt} x_{jt-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_t^2),$$

where

$$\beta_{jt} = \gamma_{jt} b_{jt}, \qquad \text{and} \qquad \gamma_{jt} \in \{0, 1\},$$

Bayesian model specification

Dynamic Bernoulli-Gaussian (BG) regression specification:

$$y_t = \sum_{j=1}^{p} \beta_{jt} x_{jt-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_t^2),$$

where

$$\beta_{jt} = \gamma_{jt} b_{jt}, \qquad \text{and} \qquad \gamma_{jt} \in \{0, 1\},$$

Dynamics based on two latent processes:

$\hookrightarrow$ Time-varying coefficients b_{jt}, $j = 1, \ldots, p$.

Bayesian model specification

Dynamic Bernoulli-Gaussian (BG) regression specification:

$$y_t = \sum_{j=1}^{p} \beta_{jt} x_{jt-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_t^2),$$

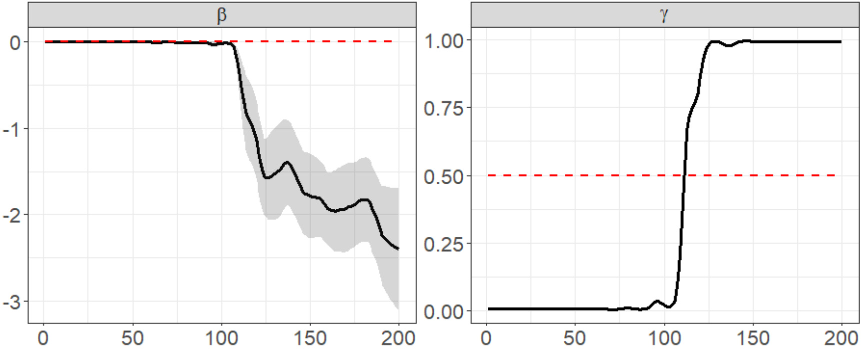
where

$$\beta_{jt} = \gamma_{jt} b_{jt}, \qquad \text{and} \qquad \gamma_{jt} \in \{0, 1\},$$

Dynamics based on two latent processes:

- $\hookrightarrow$ Time-varying coefficients b_{jt}, $j = 1, \ldots, p$.
- $\hookrightarrow$ Dynamic variable indicator γ_{jt}, $j = 1, \ldots, p$.

In other words...



Bayesian model specification

For the latent process b_{jt} we assume:

$$b_{jt} = b_{jt-1} + u_{jt}, \qquad u_{jt} \sim \mathsf{N}(0, \eta_j^2),$$

Bayesian model specification

For the latent process b_{jt} we assume:

$$b_{jt} = b_{jt-1} + u_{jt}, \qquad u_{jt} \sim \mathsf{N}(0, \eta_j^2), \quad \Longrightarrow \quad \boldsymbol{b}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \eta_j^2 \mathbf{Q}^{-1})$$

with $$\mathbf{Q} = \begin{bmatrix} 1 + k_0^{-1} & -1 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

GAUSSIAN MARKOV RANDOM FIELD (GMRF) REPRESENTATION

Bayesian model specification

For the latent process b_{jt} we assume:

$$b_{jt} = b_{jt-1} + u_{jt}, \qquad u_{jt} \sim \mathsf{N}(0, \eta_j^2), \quad \Longrightarrow \quad \boldsymbol{b}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \eta_j^2 \mathbf{Q}^{-1})$$

with $$\mathbf{Q} = \begin{bmatrix} 1 + k_0^{-1} & -1 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

GAUSSIAN MARKOV RANDOM FIELD (GMRF) REPRESENTATION

Similarly, for $h_t = \log \sigma_t^2$ we assume:

$$h_t = h_{t-1} + e_t, \qquad e_t \sim \mathsf{N}\left(0, \nu^2\right),$$

Bayesian model specification

For the latent process b_{jt} we assume:

$$b_{jt} = b_{jt-1} + u_{jt}, \qquad u_{jt} \sim \mathsf{N}(0, \eta_j^2), \quad \Longrightarrow \quad \boldsymbol{b}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \eta_j^2 \mathbf{Q}^{-1})$$

with $$\mathbf{Q} = \begin{bmatrix} 1 + k_0^{-1} & -1 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

GAUSSIAN MARKOV RANDOM FIELD (GMRF) REPRESENTATION

Similarly, for $h_t = \log \sigma_t^2$ we assume:

$$h_t = h_{t-1} + e_t, \qquad e_t \sim \mathsf{N}\left(0, \nu^2\right), \quad \Longrightarrow \quad \underbrace{\mathbf{h} \sim \mathsf{N}_{n+1}(\mathbf{0}, \nu^2 \mathbf{Q}^{-1})}_{\text{GMRF}}$$

Bayesian model specification

For the dynamics of sparsity we assume:

$$\gamma_{jt}|\omega_{jt} \sim \text{Bern}\left(\text{expit}\left(\omega_{jt}\right)\right)$$

Bayesian model specification

For the dynamics of sparsity we assume:

$$\gamma_{jt}|\omega_{jt} \sim \mathsf{Bern}\left(\mathsf{expit}\left(\omega_{jt}\right)\right) \qquad \Longrightarrow \qquad \underbrace{\boldsymbol{\omega}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \xi_j^2 \mathbf{Q}^{-1})}_{\mathsf{GMRF}}$$

Bayesian model specification

For the dynamics of sparsity we assume:

$$\gamma_{jt}|\omega_{jt} \sim \mathsf{Bern}\left(\mathsf{expit}\left(\omega_{jt}\right)\right) \qquad \Longrightarrow \qquad \underbrace{\boldsymbol{\omega}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \xi_j^2 \mathbf{Q}^{-1})}_{\text{GMRF}}$$

Such that the marginal for $\gamma_{j1}, \ldots, \gamma_{jn}$

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

has correlated components.

$\hookrightarrow$ i.e., time series dependence driven by $\boldsymbol{\omega}_j$.

Bayesian model specification

For the dynamics of sparsity we assume:

$$\gamma_{jt}|\omega_{jt} \sim \text{Bern}\left(\text{expit}\left(\omega_{jt}\right)\right) \qquad \Longrightarrow \qquad \underbrace{\boldsymbol{\omega}_j \sim \text{N}_{n+1}(\mathbf{0}, \xi_j^2 \mathbf{Q}^{-1})}_{\text{GMRF}}$$

Such that the marginal for $\gamma_{j1}, \ldots, \gamma_{jn}$

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

has correlated components.

$\hookrightarrow$ i.e., time series dependence driven by $\boldsymbol{\omega}_j$.

Prior distributions for the variances parameters.

$\hookrightarrow$ $\nu^2 \sim \text{IG}(\underbrace{A_\nu, B_\nu}_{=0.01})$, $\eta_j^2 \sim \text{IG}(\underbrace{A_\eta, B_\eta}_{=0.01})$, and $\xi_j^2 \sim \text{IG}(A_\xi, B_\xi)$.

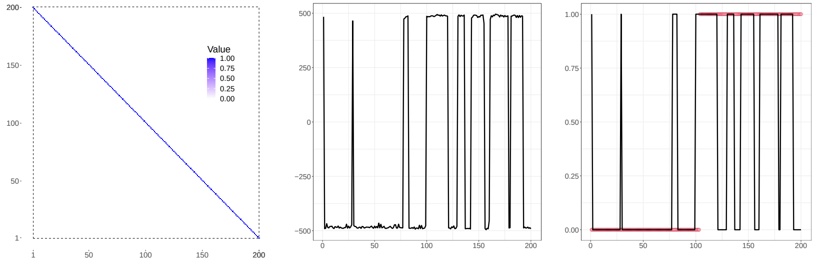
Digression: Some comparative statics on the priors

How posterior estimates are affected by the choice of (A_ξ, B_ξ)?

Digression: Some comparative statics on the priors

How posterior estimates are affected by the choice of (A_ξ, B_ξ)?

$\hookrightarrow$ Scenario A $\Rightarrow A_\xi$ constant and $B_\xi \to +\infty$.



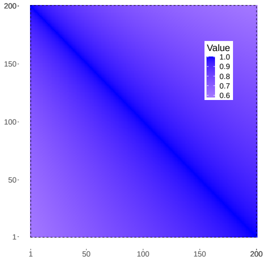
(d) (e) (f) $\{\mu_{q(\gamma_{jt})}\}_{t=1}^{n}$

Scenario A: $B_\xi \to +\infty$. (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 estimated $\{\mu_{q(\gamma_{jt})}\}_{t=1}^{n}$ compared to the simulated (red points).

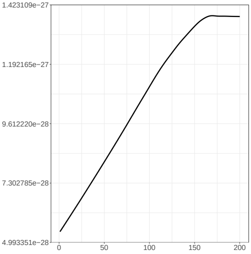
Digression: Some comparative statics on the priors

How posterior estimates are affected by the choice of (A_ξ, B_ξ)?

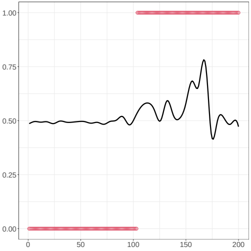
$\hookrightarrow$ Scenario B $\Rightarrow B_\xi$ constant and $A_\xi \to +\infty$.



(g)



(h)



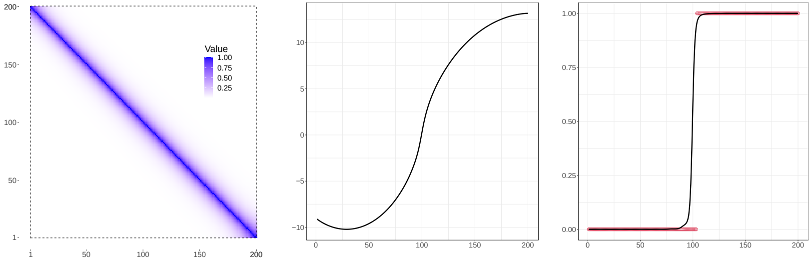
(i) $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$

Scenario B: $A_\xi \to +\infty$. (a) Depicts the variational correlation matrix for the process $\{\omega_{j,t}\}_{t=1}^n$ 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 estimated $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$ compared to the simulated (red points).

Digression: Some comparative statics on the priors

How posterior estimates are affected by the choice of (A_ξ, B_ξ)?

$\hookrightarrow$ Scenario C $\Rightarrow A_\xi/B_\xi \quad \Longrightarrow \quad c_1,\ c_1 \in R^+$.



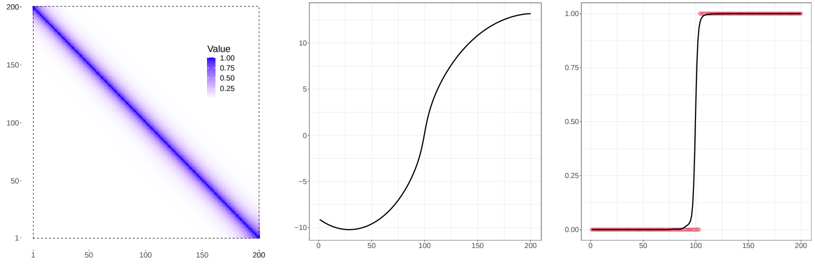
(j) (k) (l) $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$

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 estimated $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$ compared to the simulated (red points).

Digression: Some comparative statics on the priors

How posterior estimates are affected by the choice of (A_ξ, B_ξ)?

$\hookrightarrow$ Scenario C $\Rightarrow A_\xi / B_\xi \quad \Longrightarrow \quad \xi_j^2 \sim \text{IG}(A_\xi = 2, B_\xi = 5)$



(m) (n) (o) $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$

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 $\mathbf{\Sigma}_{q(\omega_j)}$. (b) Plots the trajectory of $\{\mu_{q(\omega_{jt})}\}_{t=1}^n$. (c) Shows the effect on the posterior inclusion probabilities estimated $\{\mu_{q(\gamma_{jt})}\}_{t=1}^n$ compared to the simulated (red points).

Variational Bayes inference

A re-cap on Variational Bayes (VB) inference

Minimize the Kullback-Leibler (KL) divergence between a variational density $q(\boldsymbol{\vartheta})$ and the true posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$.

A re-cap on Variational Bayes (VB) inference

Minimize the Kullback-Leibler (KL) divergence between a variational density $q(\boldsymbol{\vartheta})$ and the true posterior density $p(\boldsymbol{\vartheta}|\mathbf{y})$.

This corresponds to find $q^*(\boldsymbol{\vartheta})$ such that:

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

with (see Ormerod & Wand 2010),

$$\underline{p}\left(\mathbf{y}; q\right) = \int q(\boldsymbol{\vartheta}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\vartheta})}{q(\boldsymbol{\vartheta})} \right\} d\boldsymbol{\vartheta},$$

the variational, or “effective”, lower bound (ELBO).

$\hookrightarrow$ N.B., both $q(\boldsymbol{\vartheta})$ and $p(\mathbf{y}, \boldsymbol{\vartheta})$ are known.

A re-cap on Variational Bayes (VB) inference (cont'd)

The choice of $\mathcal{Q}$ leads to different approaches.

A re-cap on Variational Bayes (VB) inference (cont'd)

The choice of $\mathcal{Q}$ leads to different approaches.

Mean–field variational Bayes (non-parametric):

$$\mathcal{Q} = \{q(\boldsymbol{\vartheta}) : \prod_{i=1}^{M} q(\boldsymbol{\vartheta}_i), \text{for a partition } (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_M)\}.$$

$\hookrightarrow$ e.g., in a typical linear regression $q\left(\boldsymbol{\beta}, \sigma^2\right) = q\left(\boldsymbol{\beta}\right) q\left(\sigma^2\right)$.

$\hookrightarrow$ Closed-form updates based on coordinate ascent algorithm.

A re-cap on Variational Bayes (VB) inference (cont'd)

The choice of $\mathcal{Q}$ leads to different approaches.

Mean–field variational Bayes (non-parametric):

$$\mathcal{Q} = \{q(\boldsymbol{\vartheta}) : \prod_{i=1}^{M} q(\boldsymbol{\vartheta}_i), \text{for a partition } (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_M)\}.$$

↪ e.g., in a typical linear regression $q\left(\boldsymbol{\beta}, \sigma^2\right) = q\left(\boldsymbol{\beta}\right) q\left(\sigma^2\right)$.

↪ Closed-form updates based on coordinate ascent algorithm.

Parametric variational Bayes:

$$\mathcal{Q} = \{q(\boldsymbol{\vartheta}) : q(\boldsymbol{\vartheta}|\boldsymbol{\lambda}_{q(\vartheta)}) = f(\boldsymbol{\vartheta}; \boldsymbol{\lambda}_{q(\vartheta)})\},$$

↪ e.g., $f(\cdot)$ Gaussian s.t., $\boldsymbol{\lambda}_{q(\vartheta)} = (\boldsymbol{\mu}_{q(\vartheta)}, \boldsymbol{\Sigma}_{q(\vartheta)})$.

Semi-parametric variational Bayes

We propose an hybrid approach which merge parametric and non-parametric VB to estimate $\boldsymbol{\vartheta}$.

Non-parametric $\Rightarrow$ mean–field factorization of $q(\boldsymbol{\vartheta})$:

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

Semi-parametric variational Bayes

We propose an hybrid approach which merge parametric and non-parametric VB to estimate $\boldsymbol{\vartheta}$.

Non-parametric $\Rightarrow$ mean–field factorization of $q(\boldsymbol{\vartheta})$:

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

Parametric $\Rightarrow$ the red-q belong to a parametric family of distributions.

$\hookrightarrow$ Multivariate Gaussian for $q(\mathbf{h})$ (recall the GMRF representation).

$\hookrightarrow$ Polya-Gamma representation $\gamma_{jt}|\omega_{jt}$ s.t., $p(\gamma_{jt}|\omega_{jt}) \approx q\left(\gamma_{jt}\right) q\left(z_{jt}\right)$ where $p(z_{jt}) \sim \mathsf{PG}(1,0)$.

Variational density of $\beta = \gamma b$

Proposition (see paper)

Define $\boldsymbol{\beta}_j = \boldsymbol{\Gamma}_j \boldsymbol{b}_j$, where $\boldsymbol{b}_j = (b_{j0}, b_{j1}, \ldots, b_{jn})'$ and $\boldsymbol{\Gamma}_j = \mathrm{diag}(1, \gamma_{j1}, \ldots, \gamma_{jn})$. The optimal variational density of $\boldsymbol{\beta}_j$ is given by a mixture of multivariate Gaussian distributions:

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

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

$$w_s = \prod_{t=1}^{n} \mu_{q(\gamma_{jt})}^{s_t} (1 - \mu_{q(\gamma_{jt})})^{1-s_t}, \tag{2}$$

where $\mathbf{s} = (s_1, \ldots, s_t, \ldots, s_n) \in \mathcal{S}$. Moreover, the mean $\boldsymbol{\mu}_{q(\beta_j)}$ and variance $\boldsymbol{\Sigma}_{q(\beta_j)}$ can be computed analytically... (see paper)

Properties of the variational updates

Proposition (see paper)

Assume for variable j at iteration i of the algorithm:

- $\hookrightarrow$ $\max_t\{\mu^{(i)}_{q(\gamma_{jt})}\} = \epsilon \ll 1.$
- $\hookrightarrow$ $\boldsymbol{\Sigma}^{(i)}_{q(\omega_j)} - \boldsymbol{\Sigma}^{(i-1)}_{q(\omega_j)}$ is a non-negative matrix.

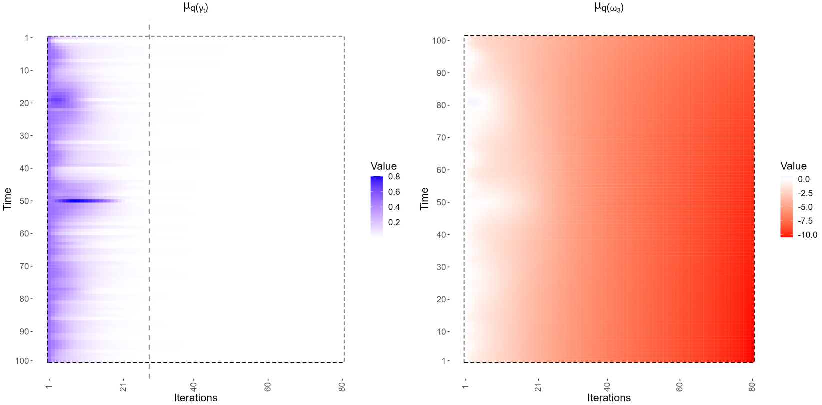
It holds that:

1. $\mu^{(i+1)}_{q(\gamma_{jt})} = \text{expit}\left\{\mu^{(i+1)}_{q(\omega_{jt})} - \frac{1}{2}\mu^{(i+1)}_{q(1/\sigma_t^2)} x_{jt}^2 \mu^{-1(i+1)}_{q(1/\eta_j^2)} q_{t,t} + O(\epsilon)\right\}$,
2. $\mu^{(i+1)}_{q(\omega_{jt})} = -\frac{1}{2}\sum_{k=1}^{n} s_{t,k} + O(\epsilon)$,
3. $\mu^{(i+1)}_{q(\omega_{jt})} \leq \mu^{(i)}_{q(\omega_{jt})}$ decreases after each iteration,

where $\text{expit} = \text{logit}^{-1}$, $q_{tt} = [\mathbf{Q}^{-1}]_{t,t}$ and $s_{tk} = [\boldsymbol{\Sigma}_{q(\omega_j)}]_{tk}$.

Properties of the variational updates

Dimension reduction



Variational update over iterations (x-axis) until convergence of the vector of posterior inclusion probabilities $(\mu_{q(\gamma_{j1})}, \ldots, \mu_{q(\gamma_{jn})})$ (left panel) and $(\mu_{q(\omega_{j1})}, \ldots, \mu_{q(\omega_{jn})})$ (right panel), for a parameter j which is always zero $\forall t$. The value of the update is given by the blue intensity. The dashed line identifies the iteration at which convergence is reached for $\epsilon = 0.01$.

Efficient variational Bayes inference scheme

Algorithm 1: Efficient variational Bayes for dynamic sparse regressions.

```
q(ϑ), Δ_ϑ, A_ν, B_ν, A_η, B_η, A_ξ, B_ξ while (Δ̂_ϑ > Δ_ϑ) do
    for j = 1, ..., p do
        Update q(b_j) as in 2.1; and q(η_j) as in A.8;
        Update q(ω_j) as in 2.3 and q(ξ_j) as in A.9;
        for t = 1, ..., n do
            Update q(z_jt) as in A.7;
            Update q(γ_jt) as in 2.2 (non-smooth) or 2.6 (smooth);
        end
    end
    Update q(σ) as in A.1 (heteroskedastic) or A.2 (homoskedastic);
    Update q(ν^2) as in A.10;
    if assumption in the previous Proposition holds then
        for j = 1, ..., p do
            if max_t{μ_q(γ_jt)} < ϵ then
                Drop the j-th variable
            end
        end
    end
    Compute Δ̂_ϑ = q(ϑ)^(iter) − q(ϑ)^(iter−1) ;
end
```

Simulation study

Comparison with MCMC

Simulation setting:

- ↪ Generate 3 processes $\{\beta_{1t}, \beta_{2t}, \beta_{3t}\}_{t=1}^{100}$ such that $\beta_{1t} \neq 0, \forall t$, $\beta_{2t} = 0, \forall t$, and β_{3t} shows dynamic sparsity.
- ↪ Generate $N = 100$ replicates from $y_t = x_{1t}\beta_{1t} + x_{2t}\beta_{2t} + x_{3t}\beta_{3t} + \varepsilon_t$, with $\varepsilon_t \sim \mathsf{N}(0, 0.25)$.
- ↪ Estimate the model with both VB and MCMC.

Comparison with MCMC

Simulation setting:

- ↪ Generate 3 processes $\{\beta_{1t}, \beta_{2t}, \beta_{3t}\}_{t=1}^{100}$ such that $\beta_{1t} \neq 0, \forall t$, $\beta_{2t} = 0, \forall t$, and β_{3t} shows dynamic sparsity.
- ↪ Generate $N = 100$ replicates from $y_t = x_{1t}\beta_{1t} + x_{2t}\beta_{2t} + x_{3t}\beta_{3t} + \varepsilon_t$, with $\varepsilon_t \sim \mathsf{N}(0, 0.25)$.
- ↪ Estimate the model with both VB and MCMC.

The accuracy of the approximation is quantified as in Wand et al. (2011):

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

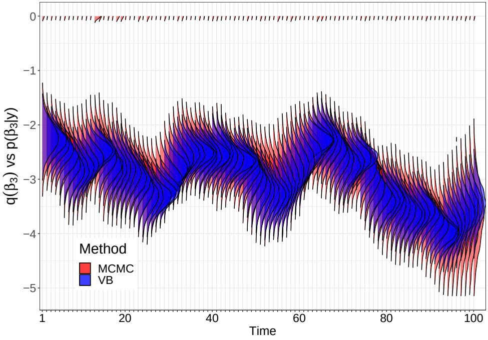
where $q(\boldsymbol{\beta})$ is the variational approximation and $p(\boldsymbol{\beta}|\mathbf{y})$ is the posterior from an equivalent MCMC with a large number of draws.

Comparison with MCMC

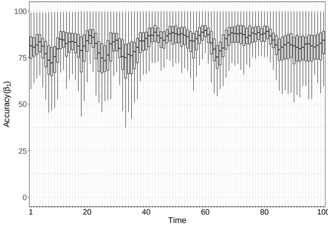
Time-varying, $\beta_{1t} \neq 0, \forall t$

Posterior densities for β_1

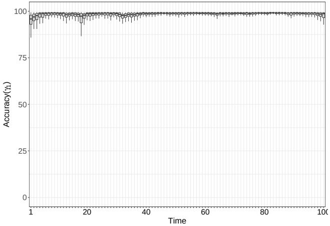
VB (blue) against MCMC (red)



Accuracy for β_1



Accuracy for γ_1

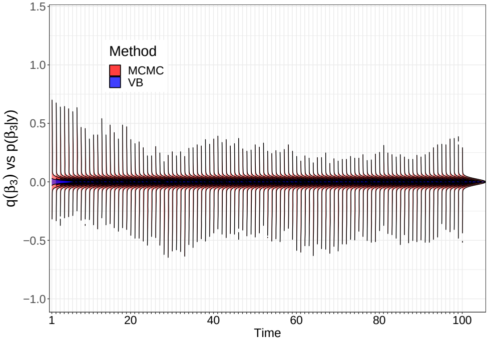


Comparison with MCMC

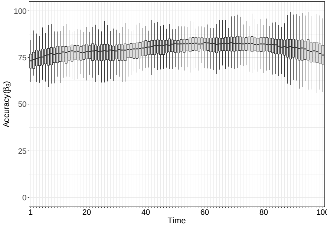
Constant at zero, i.e., $\beta_{2t} = 0, \forall t$

Posterior densities for β_2

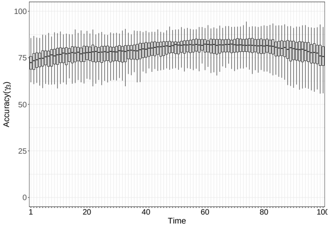
VB (blue) against MCMC (red)



Accuracy for β_2



Accuracy for γ_2

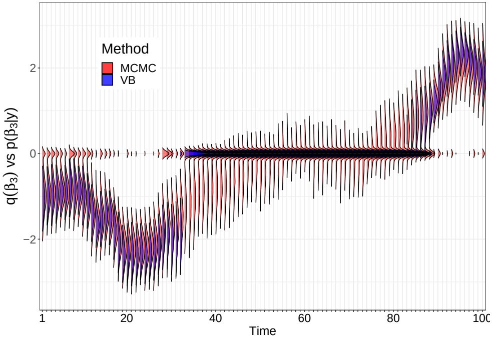


Comparison with MCMC

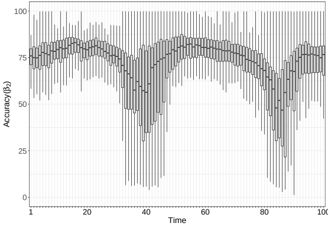
Dynamic sparsity

Posterior densities for β_3

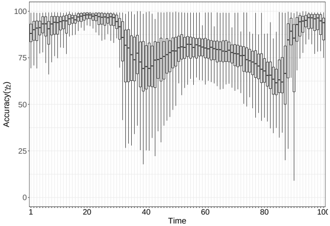
VB (blue) against MCMC (red)



Accuracy for β_3



Accuracy for γ_3



Comparison with Bayesian variable selection methods

Simulation setting

$N = 100$ replicates from the following data generating process:

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{N}(0, 0.25), \quad t = 1, \ldots, 200,$$

The dimension of the parameter $\boldsymbol{\beta}_t$ is equal to $p = 50, 100, 200$.

- $\hookrightarrow$ β_{1t} is always included, i.e. $\gamma_{1t} = 1$, $\forall t$;
- $\hookrightarrow$ $\beta_{2:7,t}$ dynamic sparsity, i.e. $\gamma_{2:7,t}$ vary over time;
- $\hookrightarrow$ $\beta_{8:p,t}$ is always excluded, i.e. $\gamma_{8:p,t} = 0$, $\forall t$.

Comparison with Bayesian variable selection methods

Simulation setting

$N = 100$ replicates from the following data generating process:

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{N}(0, 0.25), \quad t = 1, \dots, 200,$$

The dimension of the parameter $\boldsymbol{\beta}_t$ is equal to $p = 50, 100, 200$.

- ↪ β_{1t} is always included, i.e. $\gamma_{1t} = 1$, $\forall t$;
- ↪ $\beta_{2:7,t}$ dynamic sparsity, i.e. $\gamma_{2:7,t}$ vary over time;
- ↪ $\beta_{8:p,t}$ is always excluded, i.e. $\gamma_{8:p,t} = 0$, $\forall t$.

We consider different versions of our estimation algorithm:

- ↪ BG $\Longrightarrow$ basic model, with no smoothing on $\mathbb{P}(\gamma_{jt} = 1)$;
- ↪ BGH $\Longrightarrow$ as BG, but homoschedastic assumption;
- ↪ BGS $\Longrightarrow$ model with smoothing on $\mathbb{P}(\gamma_{jt} = 1)$;
- ↪ BG with fixed latent process variances ξ_j^2.

Comparison across methods

Benchmarks and metrics

Benchmark methods:

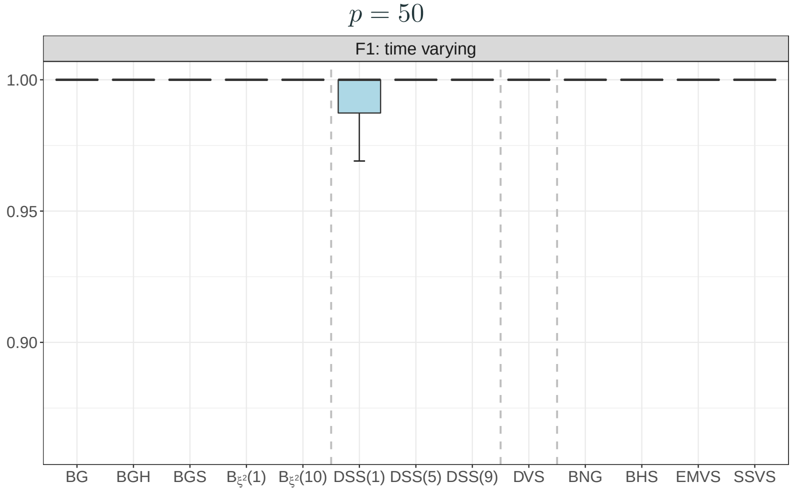
- ↪ Static models with rolling windows estimate: normal-gamma (NG), horseshoe (HS) and spike-and-slab methods (SSVS, EMVS);
- ↪ Dynamic spike-and-slab (DSS) of Ročková & McAlinn (2021), for $\Theta = \{0.1, 0.5, 0.9\}$;
- ↪ VB Dynamic variable selection (DVS) of Koop & Korobilis (2020);

Performance metrics:

- ↪ Signal/variable identification/selection (F1-score);
- ↪ Computational efficiency (running time in seconds).

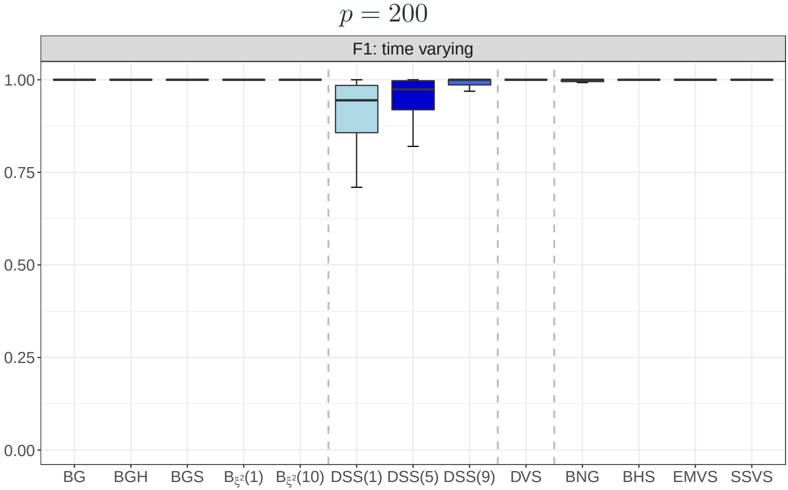
Simulation study

Scenario 1 $\implies$ $\beta_{jt} \neq 0 \; \forall t$



Simulation study

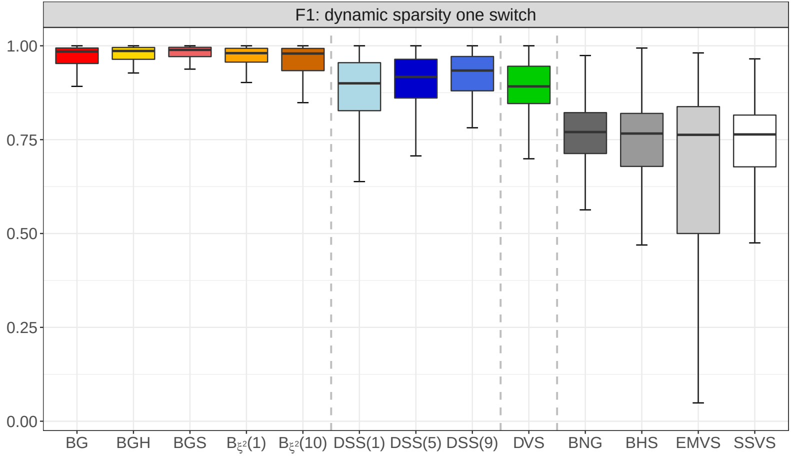
Scenario 1 $\implies$ $\beta_{jt} \neq 0 \ \forall t$



Simulation study

Scenario 2 $\implies$ single switch from $\beta_{jt} = 0$ to $\beta_{jt} \neq 0$

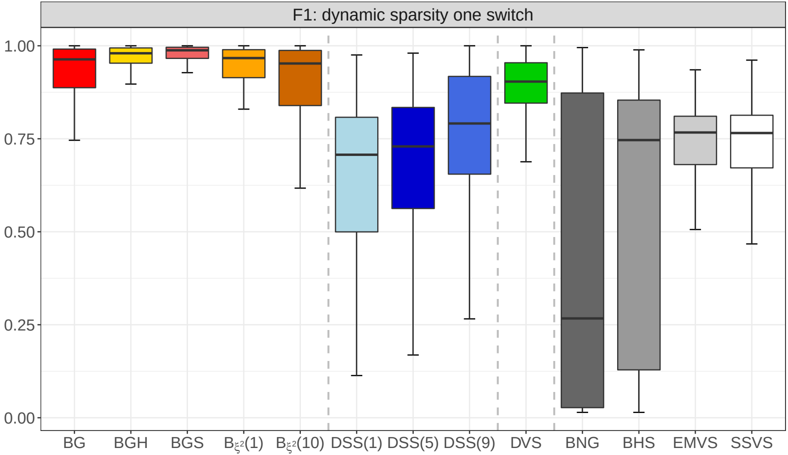
$p = 50$



Simulation study

Scenario 1 $\implies$ single switch from $\beta_{jt} = 0$ to $\beta_{jt} \neq 0$

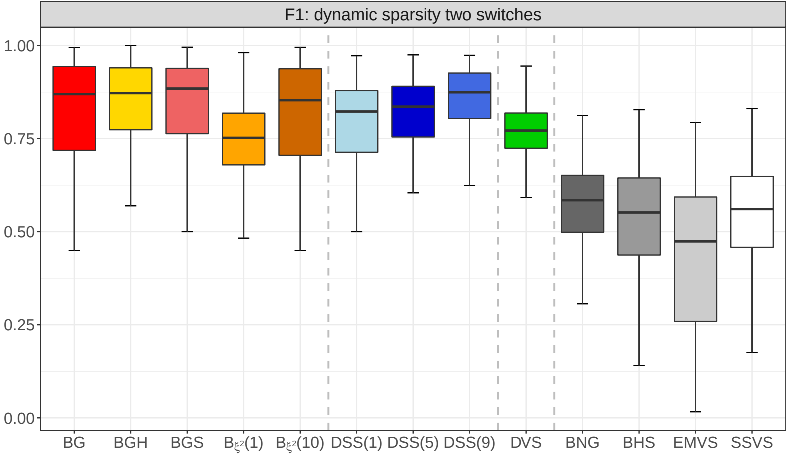
$$p = 200$$



Simulation study

Scenario 3 – two switches from $\beta_{jt} = 0$ to $\beta_{jt} \neq 0$

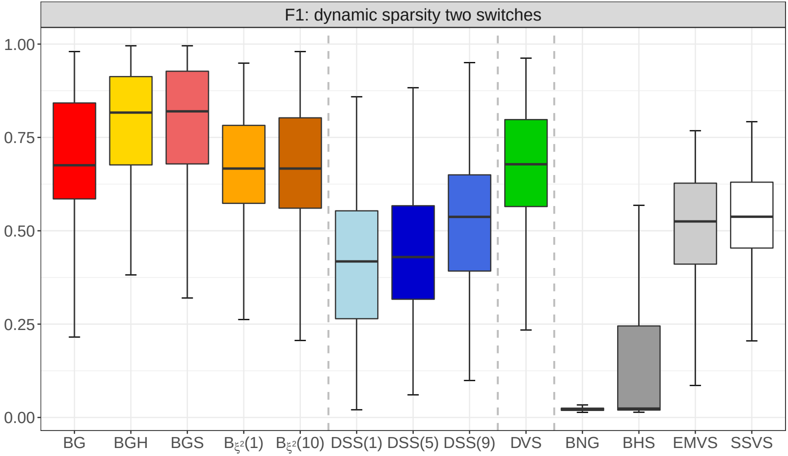
$p = 50$



Simulation study

Scenario 3 – two switches from $\beta_{jt} = 0$ to $\beta_{jt} \neq 0$

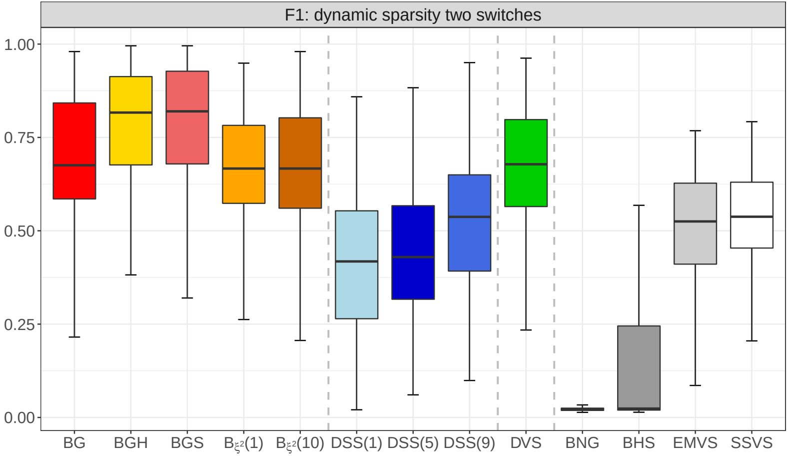
$p = 200$



Simulation study

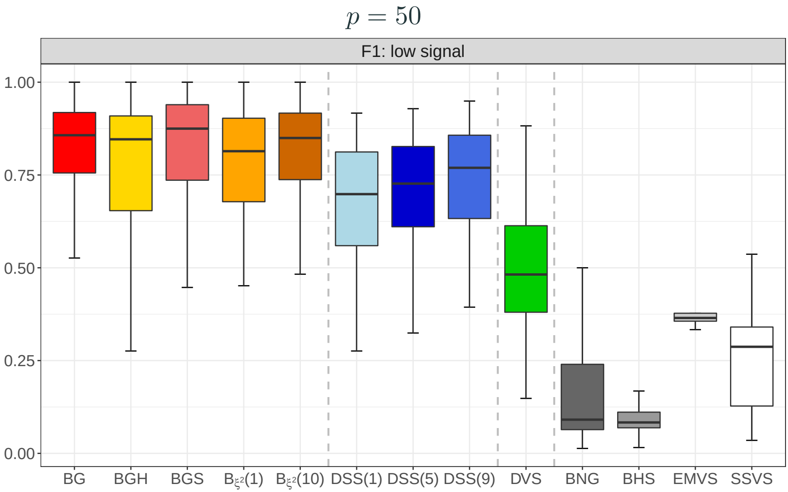
Scenario 3 – two switches from $\beta_{jt} = 0$ to $\beta_{jt} \neq 0$

$p = 200$



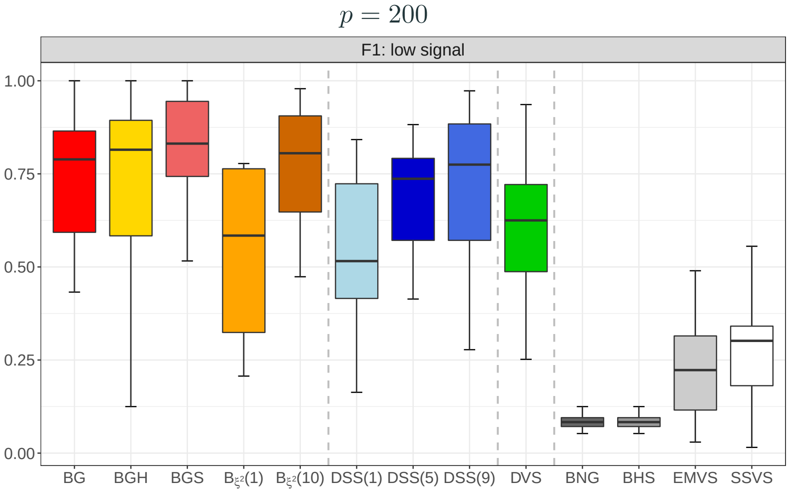
Simulation study

Scenario 4 – one short-lived switch from $\beta_{jt} = 0$ to $\beta_{jt} \neq 0$



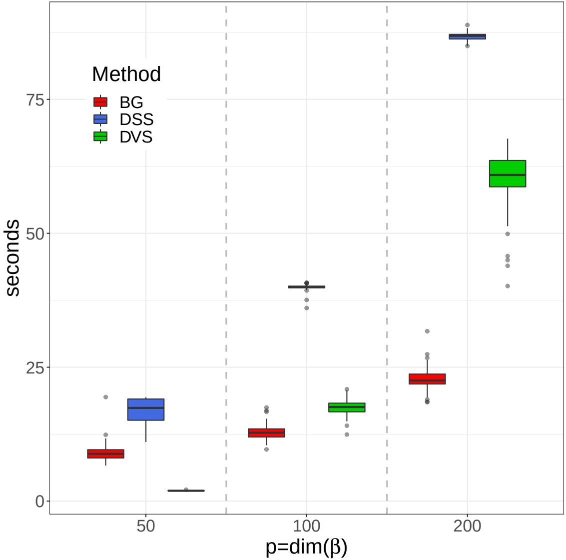
Simulation study

Scenario 4 – one short-lived switch from $\beta_{jt} = 0$ to $\beta_{jt} \neq 0$



Simulation study

Results | Running time (secs)



Inflation forecasting

Forecasting inflation based on macroeconomic variables

Empirical setting

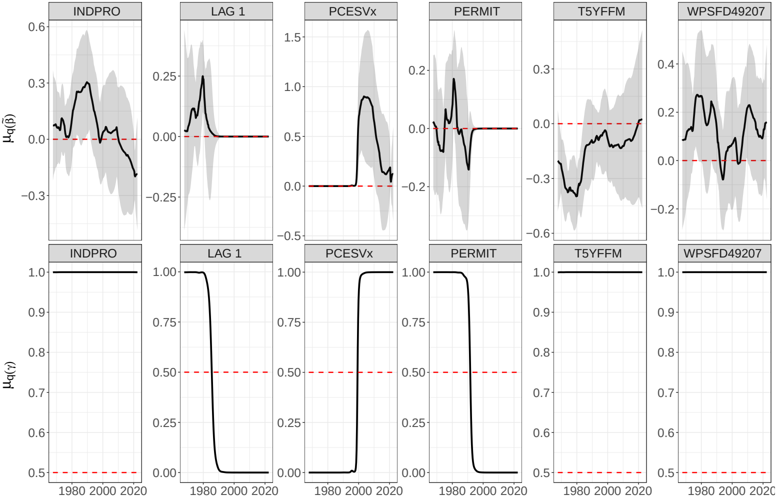
Building upon Koop & Korobilis (2020), Ročková & McAlinn (2021)

- $\hookrightarrow$ Target $\Rightarrow$ $h = 1, 2, 4, 8$ quarter-ahead inflation. Four measures of inflation: total CPI (`CPIAUCSL`), core CPI (`CPILFESL`), GDP deflator (`GDPCTPI`), PCE deflator (`PCECTPI`).
- $\hookrightarrow$ Predictors $\Rightarrow$ 229 macroeconomic variables from FRED-QD (see McCracken & Ng 2020 + 2 lags of the response (quarterly change).
- $\hookrightarrow$ Sample period $\Rightarrow$ Quarterly data 1967Q3-2022Q2
- $\hookrightarrow$ Forecasting benchmarks $\Rightarrow$ Unobserved component model (see Stock & Watson 2007) and TVP AR(1) (see Koop & Korobilis 2020).
- $\hookrightarrow$ Recursive forecasts $\Rightarrow$ 10 years "burn-in", then recursive forecasts based on an expanding window.

Forecasting inflation based on macroeconomic variables

In-sample analysis: Total CPI

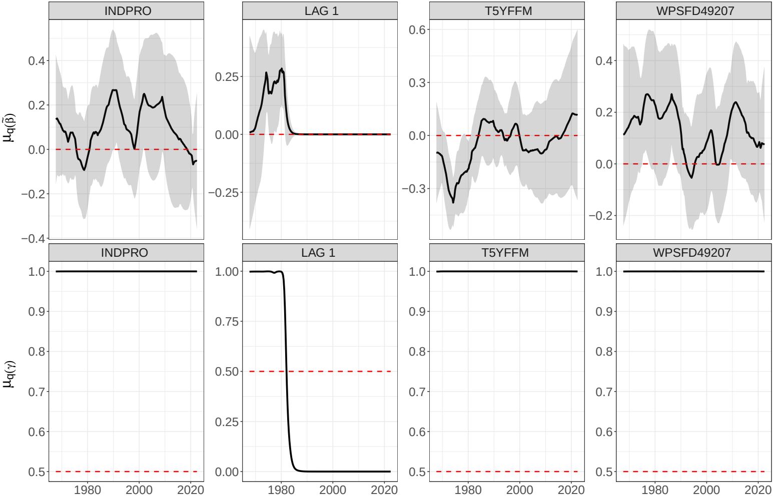
Dynamics of the selected regression coefficients $\gamma\beta$



Forecasting inflation based on macroeconomic variables

In-sample analysis: PCE deflator

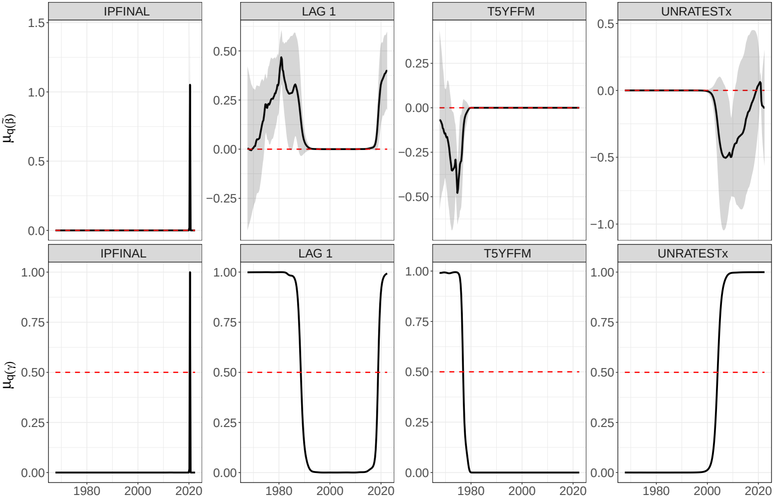
Dynamics of the selected regression coefficients $\gamma\beta$



Forecasting inflation based on macroeconomic variables

In-sample analysis: GDP deflator

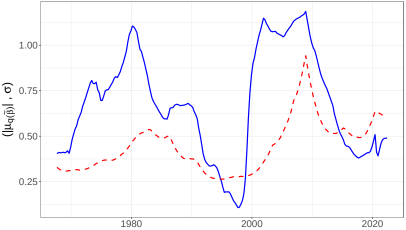
Dynamics of the selected regression coefficients $\gamma\beta$

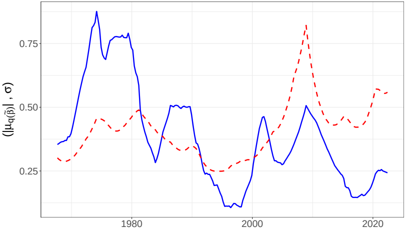


Forecasting inflation based on macroeconomic variables

In-sample analysis: Total CPI and PCE deflator

Signal $\sum_{j=1}^{p} \left|\mu_q\left(\beta_{jt}\right)\right|$ vs $\widehat{\sigma}_t$

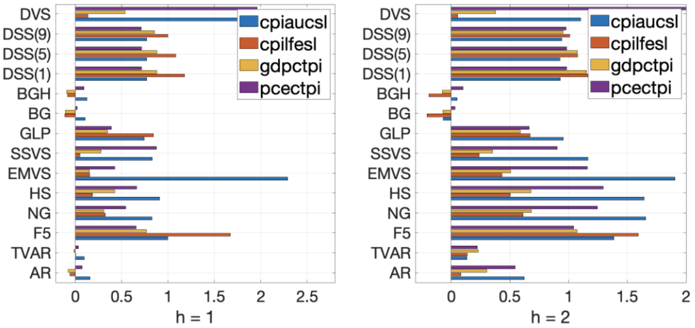




Forecasting inflation based on macroeconomic variables

Relative mean Squared Error (benchmark unobserved component model).

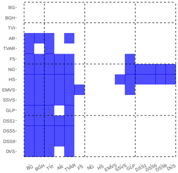
$$\sum_{t=\tau}^{T} e_{i,t}^2 - \sum_{t=\tau}^{T} e_{\text{bench},t}^2$$



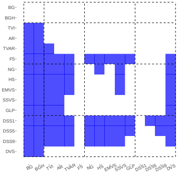
(p) Relative mean squared error

Forecasting inflation based on macroeconomic variables

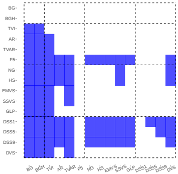
Diebold-Mariano tests



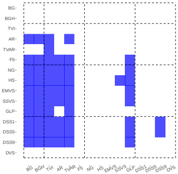
(q) CPIAUCSL



(r) CPILFESL

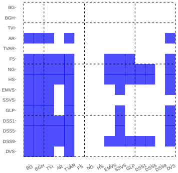


(s) GDPCTPI

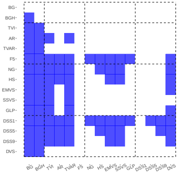


(t) PCECTPI

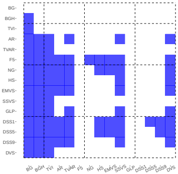
Horizon $h = 1$



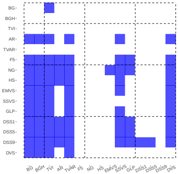
(u) CPIAUCSL



(v) CPILFESL



(w) GDPCTPI



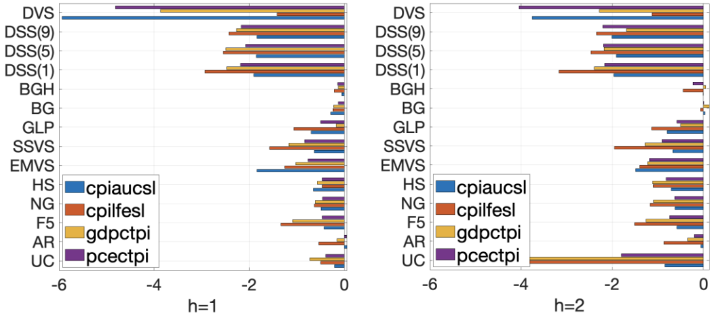
(x) PCECTPI

Horizon $h = 2$

Forecasting inflation based on macroeconomic variables

Relative log predictive score (benchmark TV-AR(2)).

$$\frac{1}{T-\tau-1}\sum_{t=\tau}^{T}\left(\log(S_{i,t})-\log(S_{\text{bench},t})\right)$$



(y) Relative log predictive score

Conclusion & what's next

This paper:

- ↪ Dynamic variable selection in large-scale time-varying predictive regressions.
- ↪ Fast and scalable semi-parametric variational Bayes algorithm.
- ↪ Competitive compared to existing variable selection methods and MCMC.

Future research:

- ↪ Extension to Generalised Linear Models.
- ↪ Dynamic group variable selection.
- ↪ Change the dependence:
    - ↪ irregular time points, spatial data, data over networks.