Semiparametric Variational Inference for dynamic sparsity in TVP models

2022 - World Meeting of the International Society for Bayesian Analysis

Nicolas Bianco

Università degli Studi di Padova

nicolas.bianco@phd.unipd.it

whitenoise8.github.io

June 29, 2022

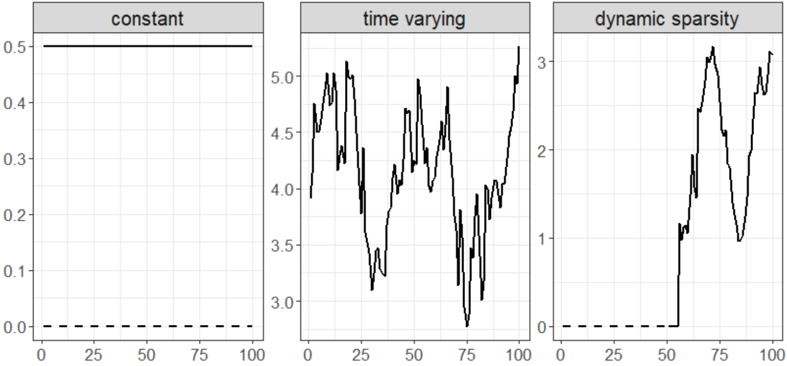
*Joint work with Mauro Bernardi (University of Padova) and Daniele Bianchi (Queen Mary University of London)

Outline

1. Introduction
2. Model and Inference
3. Properties of the model
4. Simulations
5. Application

Introduction

Different behavior of coefficients in dynamic regression:



Introduction

State of the art

Recent works on dynamic sparsity are:

- Normal-Gamma autoregressive of Kalli and Griffin (2014).
- Double-Gamma prior of Bitto and Frühwirth-Schnatter (2019).

Introduction

State of the art

Recent works on dynamic sparsity are:

- Normal-Gamma autoregressive of Kalli and Griffin (2014).
- Double-Gamma prior of Bitto and Frühwirth-Schnatter (2019).
- Dynamic variable selection of Koop and Korobilis (2020):
  - Variational Bayes Kalman Filter;
  - Spike-and-Slab type prior;
  - Stochastic and independent a priori inclusion probabilities.

Introduction

State of the art

Recent works on dynamic sparsity are:

- Normal-Gamma autoregressive of Kalli and Griffin (2014).
- Double-Gamma prior of Bitto and Frühwirth-Schnatter (2019).
- Dynamic variable selection of Koop and Korobilis (2020):
  - Variational Bayes Kalman Filter;
  - Spike-and-Slab type prior;
  - Stochastic and independent a priori inclusion probabilities.
- Dynamic Spike-and-Slab of Ročková and McAlinn (2021):
  - MCMC;
  - Spike-and-Slab type prior;
  - Deterministic evolution of inclusion probabilities, giving the past:

$$\theta_t = \frac{\Theta\psi_1(\beta_{t-1})}{\Theta\psi_1(\beta_{t-1}) + (1-\Theta)\psi_0(\beta_{t-1})},$$

    where Θ is a marginal importance weight.

Introduction

Objectives

The main feature of our method are:

- Semiparametric variational Bayes.
- Model specification without hyper-parameters tuning.
- Stochastic evolution for the inclusion probabilities.
- Fast algorithm, able to deal with regression with many predictors.

Bayesian model specification

Bernoulli-Gaussian specification (Ormerod et al., 2017) for time-varying parameter regression model:

$$y_t = \mathbf{x}_t^\intercal \boldsymbol{\Gamma}_t \boldsymbol{\beta}_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_t^2),$$

where

$$\boldsymbol{\Gamma}_t \boldsymbol{\beta}_t = \begin{bmatrix} \gamma_{1,t} & 0 & \dots & 0 \\ 0 & \gamma_{2,t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \gamma_{p,t} \end{bmatrix} \begin{bmatrix} \beta_{1,t} \\ \beta_{2,t} \\ \vdots \\ \beta_{p,t} \end{bmatrix} = \begin{bmatrix} \gamma_{1,t}\beta_{1,t} \\ \gamma_{2,t}\beta_{2,t} \\ \vdots \\ \gamma_{p,t}\beta_{p,t} \end{bmatrix},$$

and $\gamma_{j,t} \in \{0, 1\}$.

Bayesian model specification

- The random walk dynamic can be represented as a Gaussian Markov random field (GMRF) for the joint vector (see Rue and Held, 2005):

$$\boldsymbol{\beta}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \eta_j^2 \mathbf{Q}^{-1}), \quad \mathbf{h} \sim \mathsf{N}_{n+1}(\mathbf{0}, \nu^2 \mathbf{Q}^{-1}),$$

  where $h_t = \log \sigma_t^2$.

- The indicator variables are $\gamma_{j,t} | \omega_{j,t} \sim \mathsf{Bern}(p_{j,t})$ given the parameters $\omega_{j,t}$, where $\omega_{j,t} = \mathsf{logit}(p_{j,t})$.

- Dependence a priori $\boldsymbol{\omega}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \xi_j^2 \mathbf{Q}^{-1})$.

- Prior distributions for the variances parameters.
  $\nu^2 \sim \mathsf{IG}(A_\nu, B_\nu)$, $\eta_j^2 \sim \mathsf{IG}(A_\eta, B_\eta)$, and $\xi_j^2 \sim \mathsf{IG}(A_\xi, B_\xi)$.

Semiparametric Variational Inference

Goal: Find the best approximation q^* to the posterior distribution p such that $q^* = \arg\min_{q \in \mathcal{Q}} \mathsf{KL}(q||p)$.

Semiparametric Variational Inference

Goal: Find the best approximation q^* to the posterior distribution p such that $q^* = \arg\min_{q \in \mathcal{Q}} \mathsf{KL}(q||p)$.

Non-parametric: factorization of the joint variational density q:

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

closed-form updates are available and Coordinate Ascent Variational Inference (CAVI) algorithm can be implemented as in Ormerod and Wand (2010).

Semiparametric Variational Inference

Goal: Find the best approximation q^* to the posterior distribution p such that $q^* = \arg\min_{q \in \mathcal{Q}} \mathsf{KL}(q||p)$.

Non-parametric: factorization of the joint variational density q:

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

closed-form updates are available and Coordinate Ascent Variational Inference (CAVI) algorithm can be implemented as in Ormerod and Wand (2010).

Parametric: assume q to be a parametric density function.

- Gaussian approximation for $\mathbf{h}$ (see Rohde and Wand, 2016).

Properties of the model

Result (extension of Result 1 in Ormerod et al., 2017)

Assume for variable j at iteration i of the algorithm:

- $\max_t\{\mu^{(i)}_{q(\gamma_{j,t})}\} = \mu^{(i)}_{q(\gamma_{j,s_1})} = \epsilon \ll 1.$
- $\mathbf{\Sigma}^{(i)}_{q(\omega_j)} - \mathbf{\Sigma}^{(i-1)}_{q(\omega_j)}$ is a non-negative matrix.

It holds that:

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

where $q_{t,t} = [\mathbf{Q}^{-1}]_{t,t}$ and $s_{t,k} = [\mathbf{\Sigma}_{q(\omega_j)}]_{t,k}$.

Properties of the model

Remark

1) When ϵ is sufficiently small and $\mu^{(i+1)}_{q(\omega_{j,t})}$ is small enough, after i iterations:

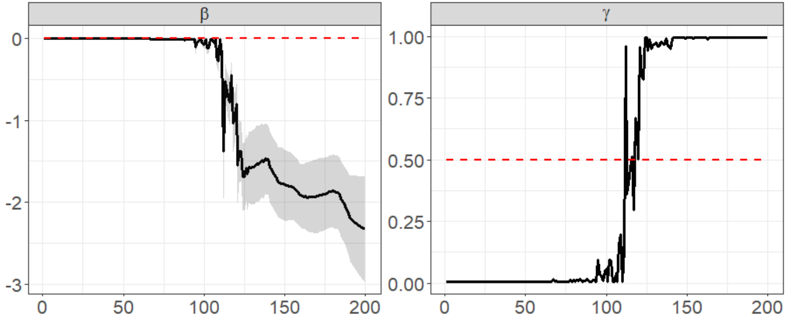
$$\mu^{(i+1)}_{q(\gamma_{j,t})} \approx \operatorname{expit}\left\{\mu^{(i+1)}_{q(\omega_{j,t})} - \frac{1}{2}\mu^{(i+1)}_{q(1/\sigma_t^2)} x_{j,t}^2 \mu^{-1(i+1)}_{q(1/\eta_j^2)} q_{t,t}\right\},$$

is represented as 0 when implemented on a computer, for all t.

2) If $\mu^{(i)}_{q(\gamma_{j,t})} \approx 0$ for all t at iteration i and the condition in Result 1 are satisfied, then the successive updates $i+k$, for $k = 1, 2, \ldots$ remains $\mu^{(i+k)}_{q(\gamma_{j,t})} \approx 0$. Thus we can remove the j-th variable from the design matrix $\mathbf{X}$ and define a reduced matrix $\mathbf{X}_\gamma$.

Smooth inclusion probabilities

The update of posterior inclusion probabilities depend on data and might be non smooth.



Smooth inclusion probabilities

- The optimal variational density $q(\boldsymbol{\gamma}_j) = \prod_{t=1}^{n} q(\gamma_{j,t})$ is a product of Bernoulli distributions.

- Approximate with $\tilde{q}(\boldsymbol{\gamma}_j) = \prod_{t=1}^{n} \tilde{q}(\gamma_{j,t})$ such that:

$$\tilde{q}(\gamma_{j,t}) \sim \mathsf{Bern}(\pi_{j,t}) \quad \text{with} \quad \mathrm{logit}(\boldsymbol{\pi}_j) = \mathbf{W}\mathbf{f}_j,$$

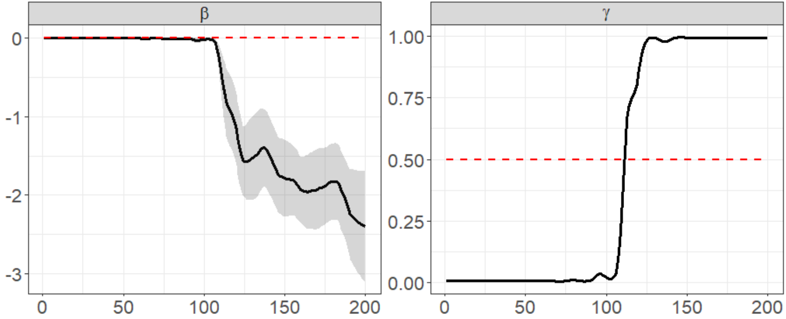
  where $\mathbf{W}$ is a b-spline basis matrix.

- The optimal value of $\mathbf{f}_j$ is obtained solving:

$$\hat{\mathbf{f}}_j = \arg\min_{\mathbf{f}_j \in \mathbb{R}^k} \mathsf{KL}\left(\tilde{q} \mid\mid q\right).$$

Smooth inclusion probabilities

The parametric variational approximation smooths the estimates.



Simulations
Setting

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

$$y_t = \mathbf{x}_t^\intercal \boldsymbol{\Gamma}_t \boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{N}(0, 0.16), \quad t = 1, \ldots, 200, \tag{1}$$

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

We compare our method (BGTVP) with:

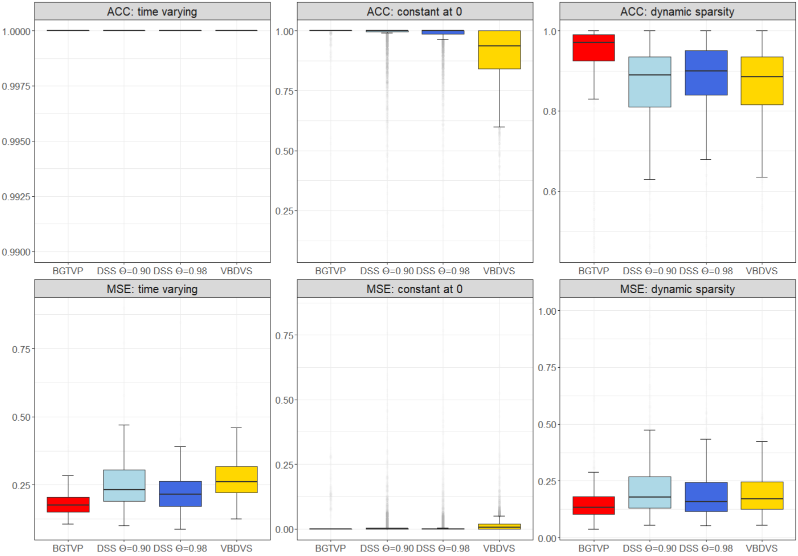
- Dynamic spike-and-slab (DSS) of Ročková and McAlinn (2021), for $\Theta = \{0.90, 0.98\}$.
- Dynamic variable selection (VBDVS) of Koop and Korobilis (2020).

We look at:

- Mean squared error (MSE).
- Classification accuracy (ACC).

Simulations

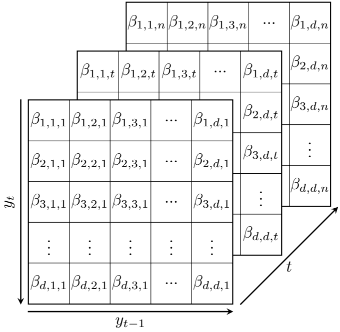
Results



Application

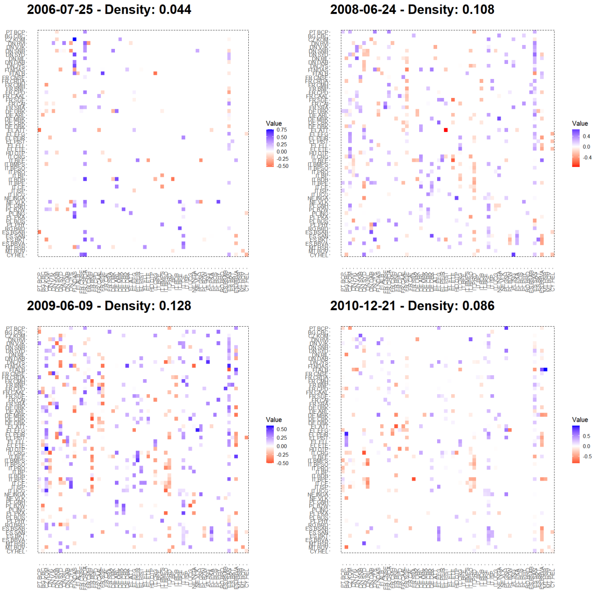
Data: returns of 60 European Banks from 2006 to 2012.
Approach: VAR(1) estimated equation-by-equation.
Output: sparse tensor of regression coefficients (see Figure below).

- Fix $t \Rightarrow$ sparse transition matrix (directed graph).
- Fix a column $\Rightarrow$ variable importance.
- Fix a row $\Rightarrow$ measure of predictability.
- Fix (row,column) $\Rightarrow$ behavior of a given lead-lag relationship across time.



Application

Dynamic directed graph



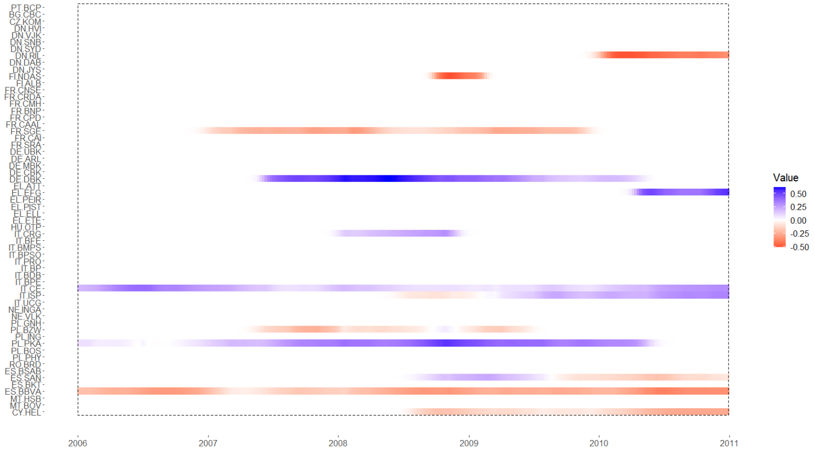
Application
Variable relevance

Relevance of ES.SAN (Santander Consumer Bank) in predict other variables.

Application

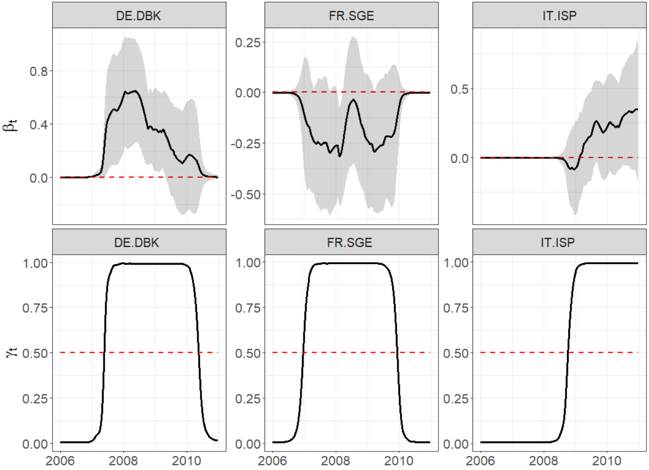
Predictability measure

Predictability of ES.SAN (Santander Consumer Bank).



Application
Time varying sparsity

Time varying dynamic and inclusion probabilities of the impact of DE.DBK (Deutsche Bank), FR.SGE (Société générale), and IT.ISP (Intesa San Paolo) on ES.SAN (Santander Consumer Bank).



Conclusion

- Extension of Bernoulli-Gaussian model for dynamic sparsity.
- Stochastic evolution of time varying inclusion probabilities.
- Fast and scalable algorithm: deal with many predictors.
- Good performances compared to state-of-the-art methods.

References I

[1] Angela Bitto and Sylvia Frühwirth-Schnatter. “Achieving shrinkage in a time-varying parameter model framework”. In: *J. Econometrics* 210.1 (2019), pp. 75–97.

[2] Maria Kalli and Jim Griffin. “Time-varying sparsity in dynamic regression models”. In: *Journal of Econometrics* 178.2 (2014), pp. 779–793.

[3] Gary Koop and Dimitris Korobilis. *Bayesian dynamic variable selection in high dimensions*. 2020.

[4] J. T. Ormerod and M. P. Wand. “Explaining variational approximations”. In: *Amer. Statist.* 64.2 (2010), pp. 140–153.

[5] J. T. Ormerod, Chong You, and Samuel Müller. “A variational Bayes approach to variable selection”. In: *Electronic Journal of Statistics* 11.2 (2017), pp. 3549 –3594.

References II

[6] Nicholas G. Polson, James G. Scott, and Jesse Windle. "Bayesian Inference for Logistic Models Using Polya Gamma Latent Variables". In: *Journal of the American Statistical Association* 108.504 (2013), pp. 1339–1349.

[7] David Rohde and M. P. Wand. "Semiparametric Mean Field Variational Bayes: General Principles and Numerical Issues". In: *Journal of Machine Learning Research* 17.172 (2016), pp. 1–47.

[8] Veronika Ročková and Kenichiro McAlinn. "Dynamic Variable Selection with Spike-and-Slab Process Priors". In: *Bayesian Analysis* 16.1 (2021), pp. 233 –269.

[9] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Vol. 104. Monographs on Statistics and Applied Probability. London: Chapman & Hall, 2005.

Thank you!

Appendix

Bayesian model specification

Computational details

Let $\boldsymbol{\vartheta} = (\mathbf{h}^\intercal, \boldsymbol{\beta}^\intercal, \boldsymbol{\gamma}^\intercal, \boldsymbol{\omega}^\intercal, \nu^2, \boldsymbol{\eta}^{2\intercal}, \boldsymbol{\xi}^{2\intercal})^\intercal$. The joint distribution of data, latent states, and parameters is $p(\mathbf{y}, \boldsymbol{\vartheta}) = p(\mathbf{y}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})$, where:

$$p(\boldsymbol{\vartheta}) = p(\mathbf{h})p(\nu^2)\prod_{j=1}^{p} p(\boldsymbol{\beta}_j|\eta_j^2)p(\boldsymbol{\gamma}_j|\boldsymbol{\omega}_j)p(\boldsymbol{\omega}_j|\xi_j^2)p(\eta_j^2)p(\xi_j^2),$$

where $p(\boldsymbol{\gamma}_j|\boldsymbol{\omega}_j) = \prod_{t=1}^{n} p(\gamma_{j,t}|\omega_{j,t})$ also factorizes over time and

$$\log p(\gamma_{j,t}|\omega_{j,t}) = \omega_{j,t}\gamma_{j,t} - \log(1 + \exp(\omega_{j,t})).$$

Problem: the second term complicates $p(\omega_{j,t}|\text{rest})$!

Solution: Polya-Gamma representation (Polson et al., 2013):

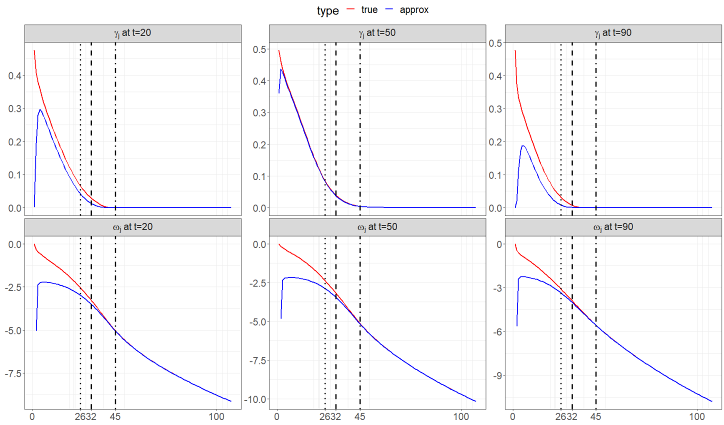
$$p(\gamma_{j,t}|\omega_{j,t}) = \int_0^{+\infty} p(\gamma_{j,t}|z_{j,t}, \omega_{j,t})p(z_{j,t})\, dz_{j,t},$$

where $p(z_{j,t})$ is the density function of a $\mathrm{PG}(1, 0)$.

Simulations

Theoretical properties

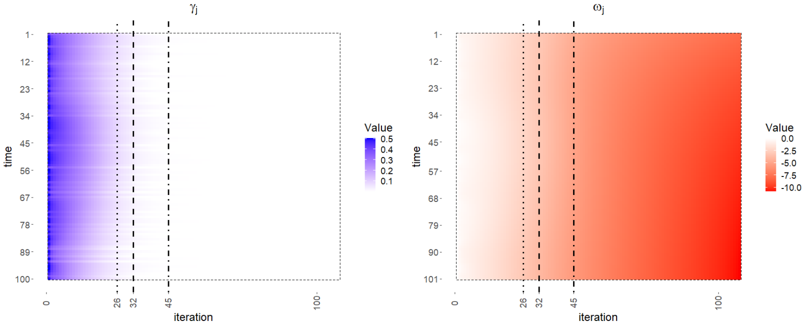
Convergence path of $\gamma_{j,t}$, $\omega_{j,t}$ and their approximations, for $t \in \{20, 50, 90\}$, $p = 50$, and $n = 100$. The conditions in Result 1 are satisfied after 26, 32 and 45 iterations when $\epsilon = 0.1, 0.05, 0.01$ (dotted, dashed, dot-dashed lines).



Simulations

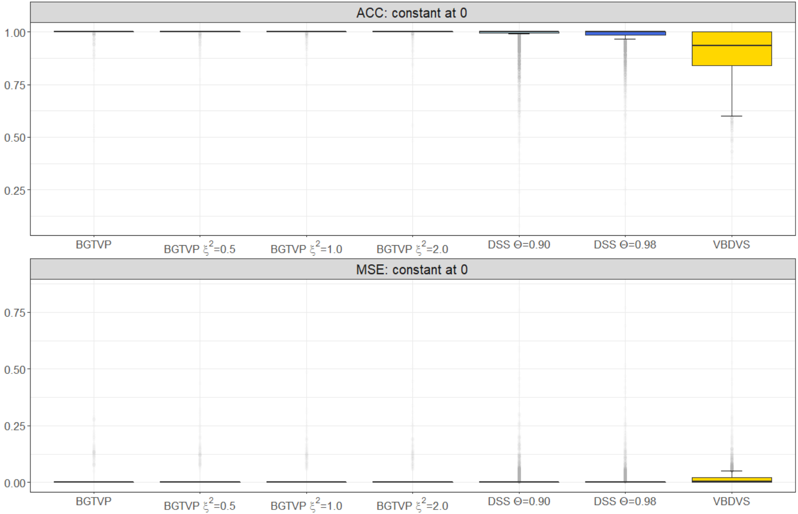
Theoretical properties

Convergence path of γ_j and ω_j, when $p = 50$ and for $n = 100$. The conditions in Result 1 are satisfied after 26, 32 and 45 iterations when $\epsilon = 0.1, 0.05, 0.01$ (dotted, dashed, dot-dashed lines).



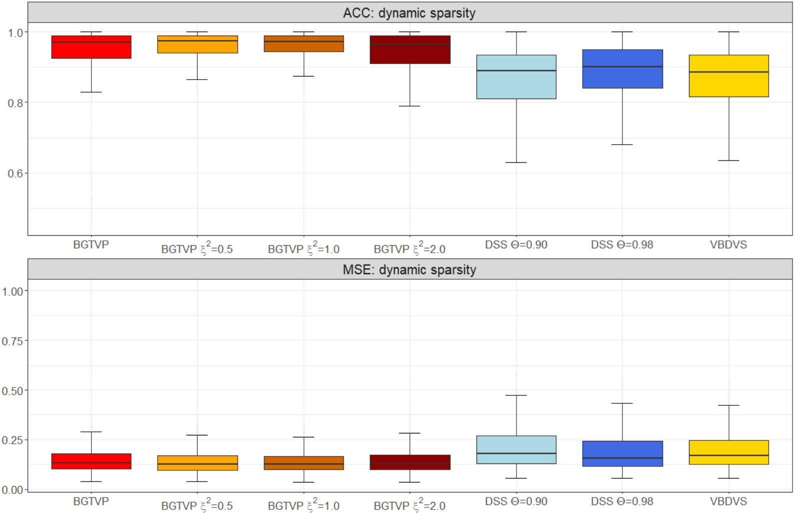
Simulations

More results



Simulations

More results



Simulations

More results