Combining shrinkage and sparsity in conjugate vector autoregressive models

Niko Hauzenberger, Florian Huber and Luca Onorante

Philipp Page

WiSo Faculty,
University of Cologne

Seminar in Econometrics, 20th January 2022

Table of Contents

Table of Contents

Introduction: Bayesian inference

Let Y and θ be a pair of random variables with joint probability distribution $p(Y, \theta)$. With θ being the parameter of interest and y the observed data, Bayes' rule is defined as follows:[1]

Bayes' rule

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1}$$

$p(y)$ with respect to θ is only a normalizing constant and can be ignored for inference such that one can write:

$$\overbrace{p(\theta|y)}^{\text{Posterior}} \propto \overbrace{p(y|\theta)}^{\text{Likelihood}} \overbrace{p(\theta)}^{\text{Prior}}$$

[1] Gelman et al., 2013, pp. 6 f.

Introduction: Shrinkage and sparsity

- Shrinkage is a form of regularization where model parameters are shrunk towards zero
  - Example: Ridge regression[2] (L2 penalization)
  - $\arg\min_\beta \left[(y - X\beta)^T(y - X\beta) + \alpha \sum_{j=1}^{p} \beta_j^2\right]$
- Sparsity refers to shrinking model parameters to exactly zero and therefore is a form of parameter selection
  - Example: Lasso regression[3] (L1 penalization)
  - $\arg\min_\beta \left[(y - X\beta)^T(y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|\right]$

Sparsity and/or shrinkage in Bayesian modeling can be realized through the choice of a prior distribution.

- Example: A regression model with a Laplace prior centered at zero is the Bayesian counterpart to a Lasso regression (proof in Appendix)

[2] Hoerl and Kennard, 2000.
[3] Tibshirani, 1996.

Introduction: Lasso example[4,5] I

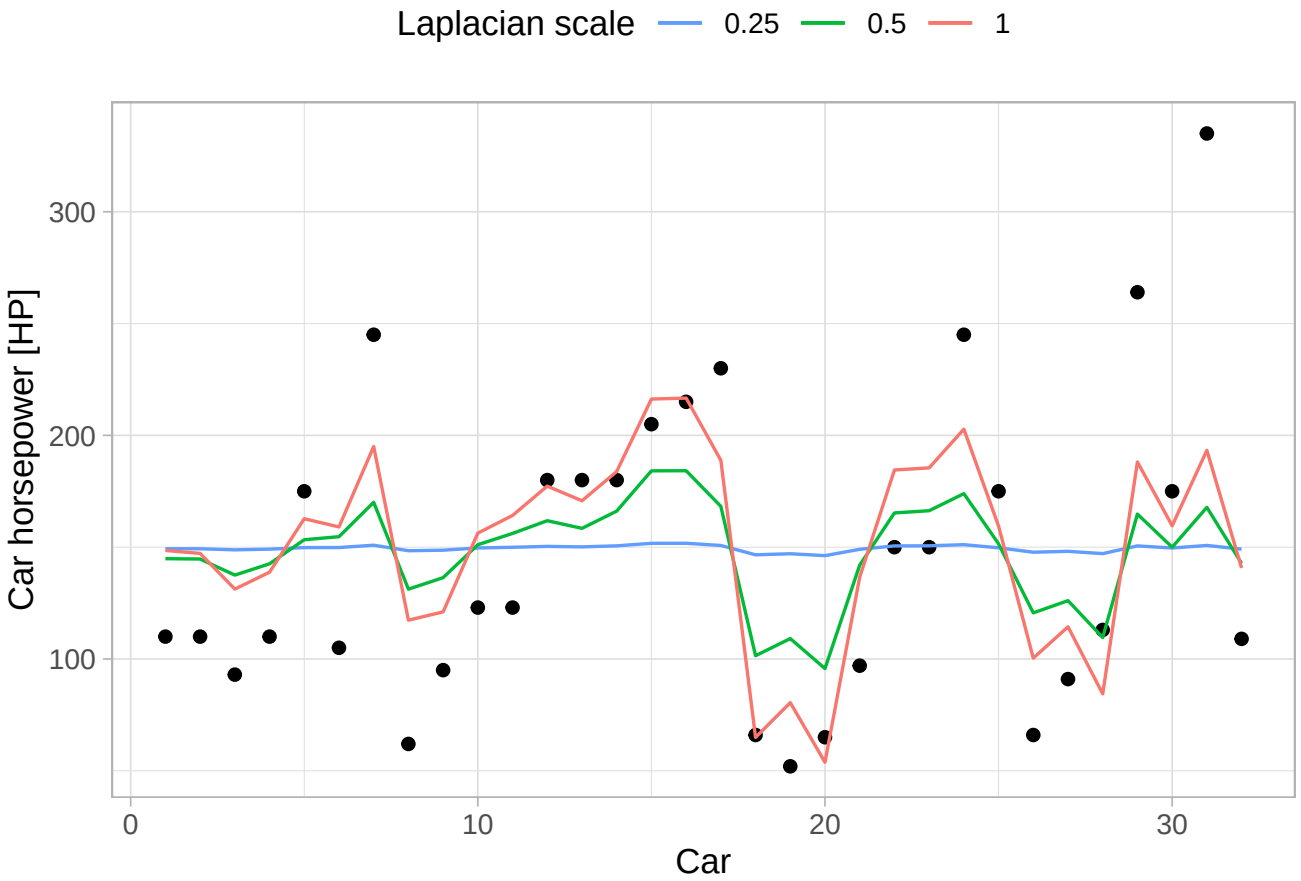

Figure 1: Bayesian Lasso model fit

[4] Source code: https://github.com/phipag/r-lasso-bayesian
[5] Data description: https://tinyurl.com/rmtcars

Introduction: Lasso example II

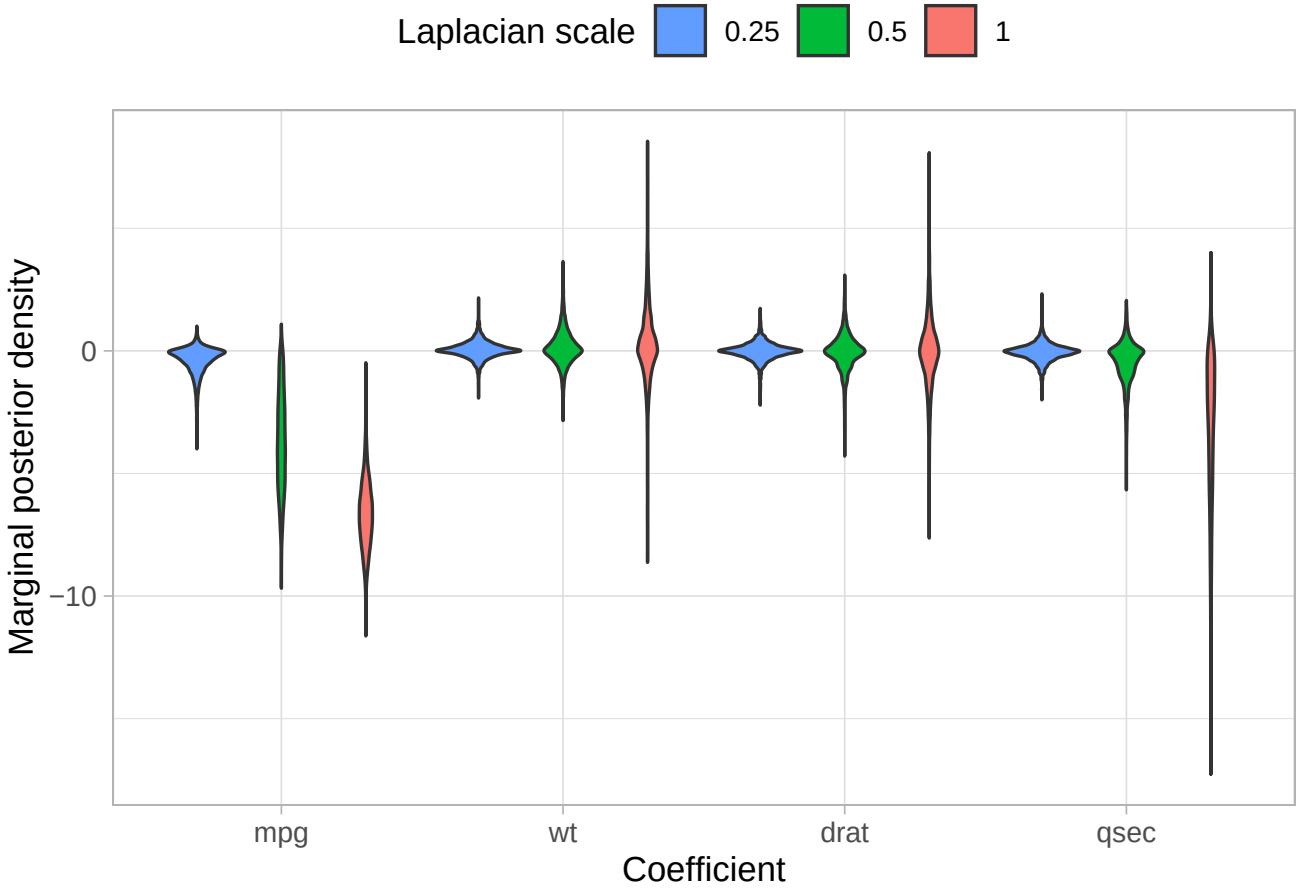

Figure 2: Marginal posterior densities for different Laplacian scale parameters (τ)

Table of Contents

VAR models

A vector auto regressive model (VAR) model with p lags ($VAR(p)$) can be written as:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + C + \epsilon_t \tag{2}$$

with $y_t = (y_{1t}, \ldots, y_{mt})'$ being the $(m \times 1)$-dimensional vector of time series data in time $t = 1, \ldots, T$; A_j $(j = 1, \ldots, p)$ the $(m \times m)$ matrix of coefficients; C the m-dimensional intercept vector and $\epsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$ a Gaussian shock vector with zero mean and $(m \times m)$-dimensional variance-covariance matrix Σ.

Example of a bivariate ($m = 2$) $VAR(1)$ model

$$y_t = A_1 y_{t-1} + C + \epsilon_t$$

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \underbrace{\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}}_{m \times m} \underbrace{\begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix}}_{m \times 1} + \underbrace{\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}}_{m \times 1} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

yields the following system of equations:

$$y_{1t} = a_{11} y_{1t-1} + a_{12} y_{2t-1} + c_1 + \epsilon_{1t}$$
$$y_{2t} = a_{21} y_{1t-1} + a_{22} y_{2t-1} + c_2 + \epsilon_{2t}$$

$m(mp + 1)$ parameters!

Model setup with natural conjugate shrinkage priors

Prior setup:

- Normal prior for coefficients and intercept: $\alpha|\Sigma \sim \mathcal{N}(\alpha_0, \Sigma \bigotimes V_0(\delta))$
- Inverted Wishart prior for variance-covariance matrix: $\Sigma \sim \mathcal{W}^{-1}(s_0, S_0)$
  - s_0 and S_0 are hyperparameters
  - Cannot shrink covariances to zero
- Minnesota prior[6] for $V_0(\delta)$
  - Traditionally shrinks parameters of first own lags towards a random walk, else towards a white noise
    - Here: Shrink every lag towards zero (including own lags)
  - $\delta = (\theta_1, \pi)$ is a hyperparameter which controls the overall tightness of the prior for the coefficients (θ_1) and the intercepts (π)

Thanks to conjugacy the posterior distributions have a closed form solution!

[6] Bańbura et al., 2010; Kadiyala and Karlsson, 1997; Koop, 2013.

The problem of dense models

The one-step-ahead predictive density follows a multivariate t-distribution with variance:

$$Var(y_{T+1}|Y,X) = \frac{1}{s_1 - 2} \underbrace{\left(1 + \sum_{i=1}^{n}\sum_{j=1}^{n}(x_{iT+1}x_{jT+1}\nu_{ij})\right)}_{\text{Parameter uncertainty}} S_1 \tag{3}$$

where s_1 and S_1 are the mean and variance of the Gaussian shocks ϵ_{T+1} and ν_{ij} the (i,j)th element in $\bar{V} = (\bar{X}'\bar{X})^{-1}$. Note that the parameter estimates depend on $\bar{V}$ as $\bar{A} = \bar{V}(X'Y + V_0(\delta)^{-1}A_0)$.

The problems:

1. ν_{ij} never becomes exactly zero with shrinkage only $\rightarrow$ Parameter uncertainty adds up in large models.
2. Macroeconomic datasets usually have many correlated variables which inflates the variance estimates S_1.

Table of Contents

Sparsity as a possible solution for predictive uncertainty I

Given the model with shrinkage prior, the authors apply ex-post sparsification algorithms on both the posterior VAR coefficients (α) and the posterior variance-covariance matrix (Σ). The high-level process can be summarized as follows:

1. Sample $\Sigma^{(rep)}$ from the marginal posterior $\Sigma|Y, X \sim \mathcal{W}^{-1}$
2. Sample $\alpha^{(rep)}$ from the marginal posterior $\alpha|\Sigma^{(rep)}, Y, X \sim \mathcal{N}$
3. Apply the sparsification algorithms on the pair of draws to obtain the sparsified point-estimates $\hat{\alpha}^{*(rep)}$ and $\hat{\Omega}^{*(rep)}$
4. Repeat (1) – (3) *rep* times

The result can be interpreted as an approximation to drawing from the sparsified joint posterior $p(\hat{\alpha}^*, \hat{\Omega}^*|Y, X)$. This is a major advantage in comparison to other literature[7] because the approximation of a sparsified posterior distribution allows statements about parameter uncertainty of the sparsified parameters.

[7] Hahn and Carvalho, 2015.

Sparsity as a possible solution for predictive uncertainty II

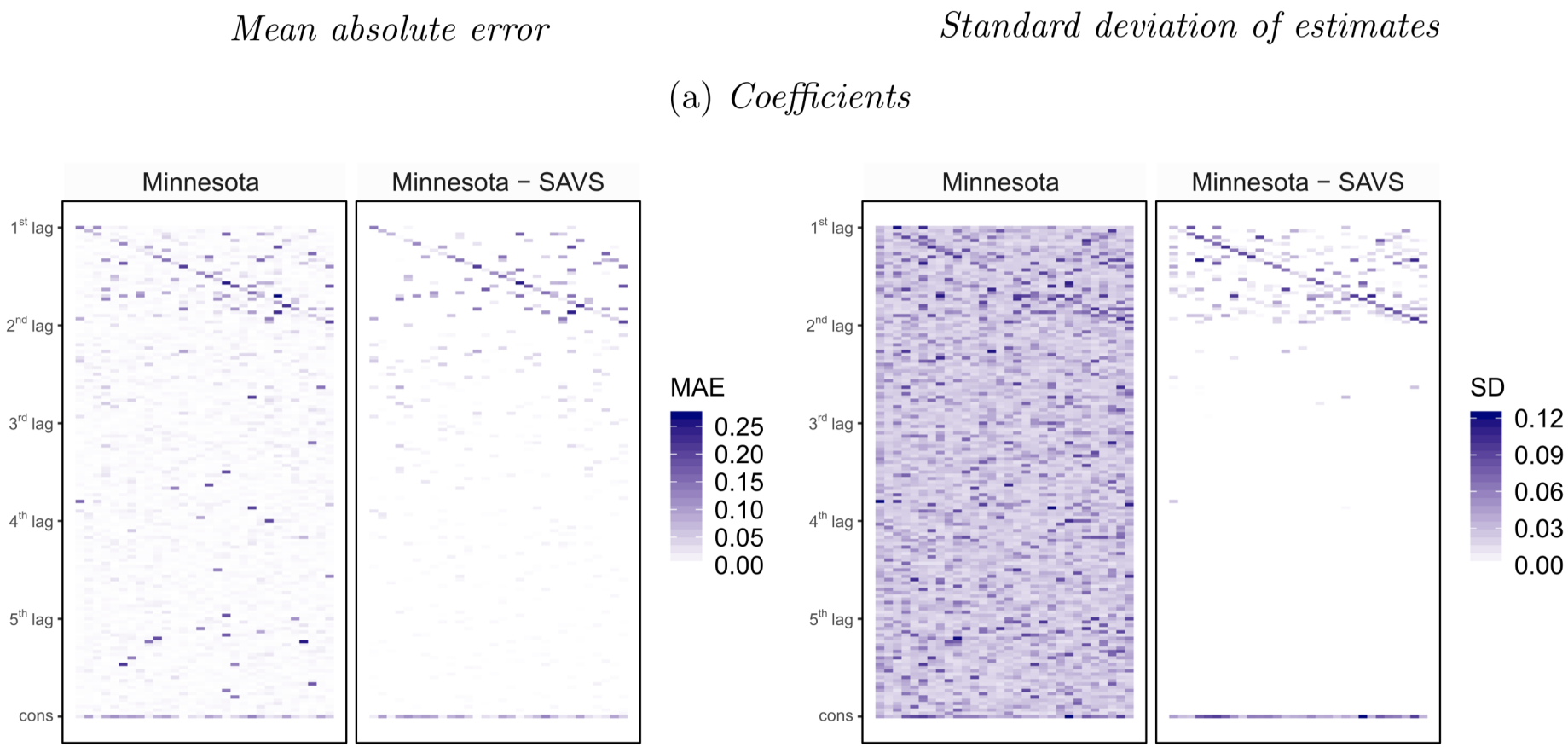

Figure 3: Sparsity reduces both the coefficients mean absolute error (MAE) and the predictive parameter uncertainty. The MAE is based on the posterior median estimates measured against the true coefficients. See Appendix for covariance estimation results.

Table of Contents

Design of the forecasting application

- Quarterly version of the McCracken and Ng, 2015 dataset
  - Spans U.S. macroeconomic data from 1959:Q1 until 2018:Q4
  - Target variables: Output (GDPC1), consumer price inflation (CPIAUCSL) and Federal Funds Rate (FEDFUNDS)
  - Large VAR (**L-VAR**): VAR(5) with $m = 165$ macroeconomic financial variables
- Quarterly recursive forecasting design:
  - Initial training period: 1959:Q1 to 1989:Q4 (first 30 years)
  - Estimation period: h-step-ahead predictive distribution for $h \in \{1, 4, 8\}$
- Evaluation:
  - Hold-out period: 1990:Q1 to 2018:Q4
  - Metrics:[8] Root-mean-squared-forecast-error (RMSE), log predictive likelihood (LPL)
- Competitive models to the **L-VAR** model:
  - **S-VAR**: Small VAR with the three target variables only (benchmark model)
  - **M-VAR**: Medium VAR with $m = 21$ variables
  - **FA-VAR**: Factor-augmented VAR adding three principal components extracted from the remaining quantities to the **S-VAR**[9]

[8] Mathematical formulation in Appendix

[9] Bańbura et al., 2010; Koop, 2013.

Average point- and density forecast performance

Overall, there is only mixed evidence that sparsification in a large-scale model improves the average forecast performance in the hold-out set in comparison to non-sparse and/or smaller models.

- $h \in \{4, 8\}$ forecasts benefit from larger information sets (**L-VAR**)
- Sparsification works well for longer run forecasts of output and interest rate
- Marginal LPL for inflation: Main driver for bad performance of large models
  - Well known result in central bank practice that small models perform better for inflation forecasts[10]

However, sparsification only very rarely hurts forecasting performance in a significant manner on average.

[10] Giannone et al., 2015.

Density forecast performance over time I

(a) *One-quarter-ahead*

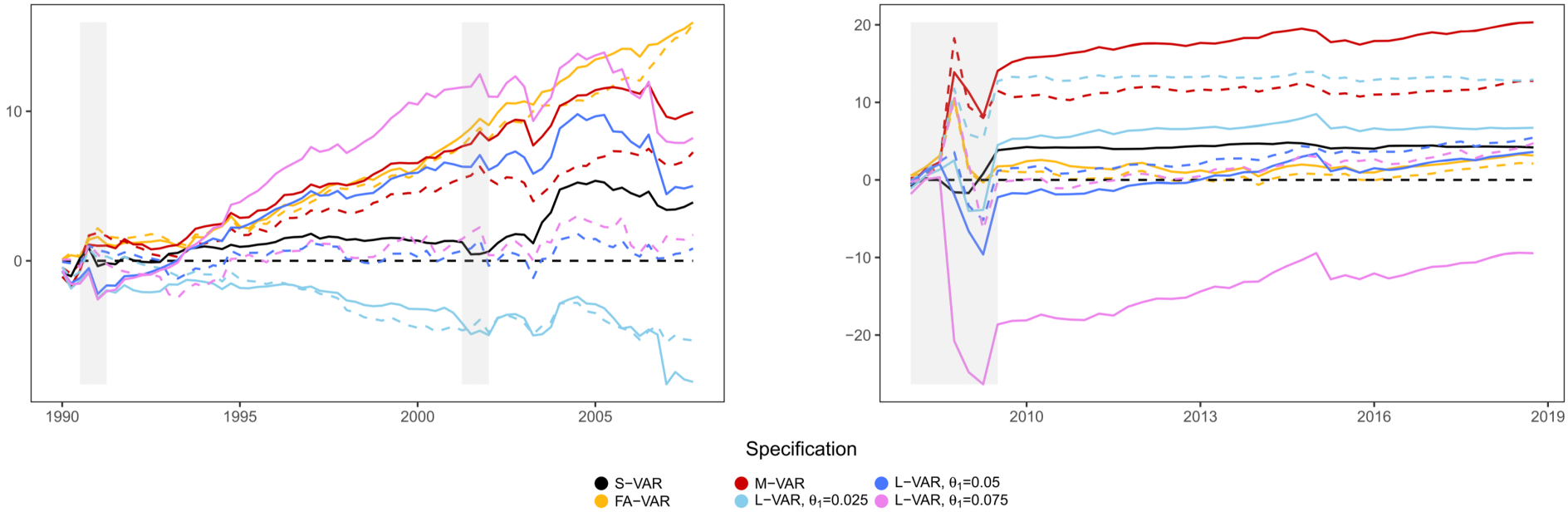

Figure 4: Before the financial crisis (left plot) the one-quarter-ahead forecasts of the sparsified large-scale models as measured by the joint cumulative LPLs are very competitive and sometimes better in comparison to the smaller and non-sparsified models. This behavior is reversed during the shock due to the financial crisis (right plot) in 2008. However, the performance increases again after the end of the financial crisis.

Dashed lines indicate non-sparse models. Solid lines depict the best sparsified model of the class. Smaller θ_1 values put more weight on shrinkage.

Density forecast performance over time II

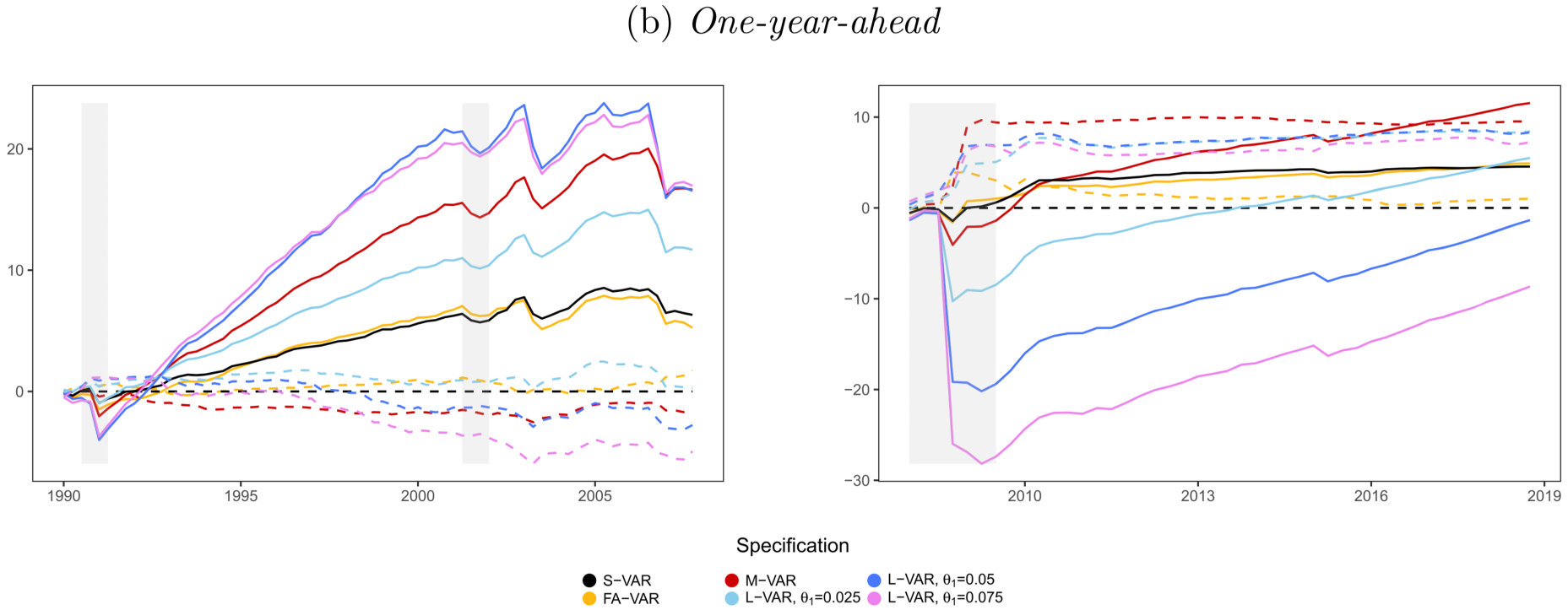

Figure 5: Before the financial crisis (left plot) the one-year-ahead forecasts of the sparsified large-scale models as measured by the joint cumulative LPLs are significantly superior to the smaller and non-sparsified models. This behavior is reversed during the shock due to the financial crisis (right plot) in 2008. However, the performance increases steeply after the end of the financial crisis.

Dashed lines indicate non-sparse models. Solid lines depict the best sparsified model of the class. Smaller θ_1 values put more weight on shrinkage.

Table of Contents

Conclusion

- Natural conjugate priors allow for direct sampling from the posterior distributions
  - This speeds up the process of approximating a sparsified posterior
- Shrinkage alone is not enough to reduce predictive uncertainty due to the curse of dimensionality in large-scale VAR models
  - Sparsification works well for many cases but has problems in catching outliers (e.g. financial crisis)
- The sparsification algorithm is heavy on CPU time even for the natural conjugate prior
  - The Paper's results could not be reproduced on a fast desktop PC

References I

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*(1), 71–92. https://doi.org/10.1002/jae.1137

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013, November 27). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b16018

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, *97*(2), 436–451. https://doi.org/10.1162/REST_a_00483

Hahn, P. R., & Carvalho, C. M. (2015). Decoupling shrinkage and selection in bayesian linear models: A posterior summary perspective [Publisher: Taylor & Francis]. *Journal of the American Statistical Association*, *110*(509), 435–448. https://doi.org/10.1080/01621459.2014.993077

Hauzenberger, N., Huber, F., & Onorante, L. (2020). Combining shrinkage and sparsity in conjugate vector autoregressive models. *arXiv:2002.08760 [econ]*. Retrieved January 15, 2022, from http://arxiv.org/abs/2002.08760

References II

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*(1), 80–86. https://doi.org/10.1080/00401706.2000.10485983

Kadiyala, K. R., & Karlsson, S. (1997). Numerical methods for estimation and inference in bayesian VAR-models [Publisher: Wiley Online Library]. *Journal of Applied Econometrics*, *12*(2), 99–132.

Koop, G. M. (2013). Forecasting with medium and large bayesian VARS. *Journal of Applied Econometrics*, *28*(2), 177–203. https://doi.org/https://doi.org/10.1002/jae.1270

McCracken, M. W., & Ng, S. (2015). *FRED-MD: A monthly database for macroeconomic research*. https://doi.org/10.20955/wp.2015.012

Park, T. H., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso [Publisher: [Royal Statistical Society, Wiley]]. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

Appendix: Bayesian Lasso I

Theorem

Under a Laplace prior and i.i.d normally distributed errors the posterior mode is the frequentist Lasso estimate.

The full Bayesian Lasso model can be found in Park and Casella, 2008.

Proof.

Consider the standard regression model $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$ with $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ and i.i.d. $\beta = \beta_1, \beta_2, \ldots, \beta_p$ and:

- Likelihood: $p(y|X, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2\right)$ with $\epsilon_i = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}$
- Laplace prior centered at zero: $p(\beta; \tau) = \prod_{j=1}^{p} \frac{1}{2\tau} \exp\left(-\frac{|\beta_j|}{\tau}\right) \propto \exp\left(-\frac{1}{\tau} \sum_{j=1}^{p} |\beta_j|\right)$

Appendix: Bayesian Lasso II

The posterior $p(\beta|y,X) \propto p(y|X,\beta)p(\beta;\tau)$ can be written as:

$$p(\beta|y,X) \propto \overbrace{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\right)}^{\text{Likelihood}} \overbrace{\exp\left(-\frac{1}{\tau}\sum_{j=1}^{p}|\beta_j|\right)}^{\text{Prior}}$$

$$= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 - \frac{1}{\tau}\sum_{j=1}^{p}|\beta_j|\right)$$

Taking the log to determine the posterior mode optimization problem:

$$\log p(\beta|y,X) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 - \frac{1}{\tau}\sum_{j=1}^{p}|\beta_j| = -\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 + \frac{1}{\tau}\sum_{j=1}^{p}|\beta_j|\right)$$

Appendix: Bayesian Lasso III

Maximizing the negative log posterior is equal to minimizing the positive log posterior:

$$\arg\min_{\beta} \left[\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 + \frac{1}{\tau} \sum_{j=1}^{p} |\beta_j| \right] = \arg\min_{\beta} \frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} \epsilon_i^2 + \frac{2\sigma^2}{\tau} \sum_{j=1}^{p} |\beta_j| \right]$$

Ignoring the normalizing constant $\frac{1}{2\sigma^2}$ and setting $\lambda = \frac{2\sigma^2}{\tau}$ yields the Lasso estimate:

$$\arg\min_{\beta} \left[\sum_{i=1}^{n} \epsilon_i^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

□

Appendix: Sparsified variance-covariance matrix estimates

(b) *Lower Cholesky factor of the variance-covariance matrix*

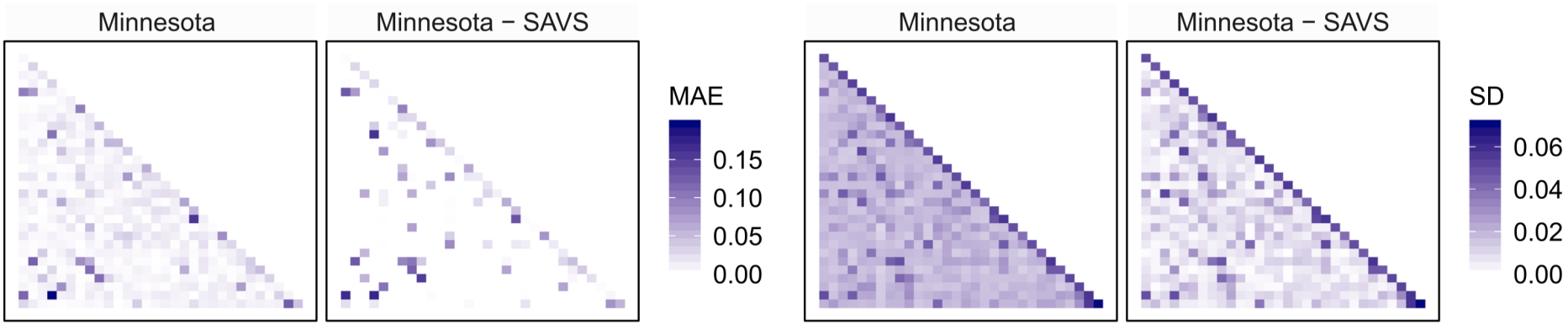

Figure 6: Similar to the results in Figure 3, sparsification also reduces the MAE as well as the predictive uncertainty of the variance-covariance matrix estimates.

Appendix: Root-mean-squared-forecast-error (RMSE)

The root-mean-squared-forecast-error (RMSE) describes the square-root of the squared average deviation of forecasts from the true values. It can be written as follows:[11]

$$\text{RMSE} = \sqrt{\frac{1}{T_1 - T_0 - H + 1} \sum_{T=T_0+H-h}^{T_1-h} (y_{i,T+h|T} - y_{i,T+h})^2}$$

where $y_{i,T+h|T}$ is the point estimate (e.g. posterior median) at time $T + h$ given the information set for the time span T and $y_{i,T+h}$ the true value at time $T + h$. T_0 and T_1 are the beginning and the end of the evaluation sample and H the longest forecast horizon.

[11] Bańbura et al., 2010.

Appendix: Predictive likelihood

The predictive likelihood is the posterior predictive density evaluated at the actual true outcome. The sum of log predictive likelihoods can be used to compare the forecasting performance of one model relative to other models. It is defined as follows:[12]

$$\sum_{T=T_0}^{T_1-h} \log[p(y_{i,T+h|T} = y_{i,T+h}|\text{Data}_T)]$$

where $y_{i,T+h|T}$ is the point estimate (e.g. posterior median) at time $T+h$ given the information set for the time span T and $y_{i,T+h}$ the true value at time $T+h$. T_0 and T_1 are the beginning and the end of the evaluation sample and h the forecast horizon. Finally, Data_T is the information set available at time T.

[12] Koop, 2013.