Large Bayesian Additive Vector Autoregressive Tree Models

Hedibert F. Lopes[1]
INSPER Institute of Education and Research

Bayesian Macroeconometric Modelling Workshop
School of Economics at The University of Queensland
Brisbain, Australia - August 2025

Access to paper and slides

The 1st draft of the paper can be found here:
https://arxiv.org/abs/2503.13759

The slides can be download from the URL address
http://hedibert.org/wp-content/uploads/2025/09/minnesota-bart-slides.pdf

or via the following QR Code

An helicopter view on VARs

- Vector autoregressive (VAR) models are the main workhorse in empirical macroeconomics: forecasting, impulse response and policy analysis.

- For m-dimensional y_t and p lags, the standard Gaussian VAR model is defined as

$$y_t = \sum_{l=1}^{p} \Phi_l y_{t-l} + \epsilon_t, \quad \epsilon_t \;\; \text{iid} \;\; N(0, \Sigma_t),$$

for $t = 1, \dots, T$.

- mp regressors per equation.

- $m^2 p$ parameters in $(\Phi_1, \dots, \Phi_p)$.

Evolution of Bayesian VAR models

- Small/medium size VAR
  - ▶ Doan, Litterman and Sims (1984/1986) - Minnesota prior
  - ▶ Kadiyala and Karlsson (1993/1997) - MC + MCMC
  - ▶ Lopes, Moreira and Schmidt (1999) - VAR + TVP via SIR
  - ▶ Primiceri (2005) - Structural VAR + TVP + SV

- Large/huge size VAR
  - ▶ Bańbura et al. (2010) - Large VAR
  - ▶ Koop and Korobilis (2013) - Large VAR + TVP
  - ▶ Carriero et al. (2019) - Large VAR + SV
  - ▶ Kastner and Huber (2020) - Huge VAR (sparsity)

- Nonparametric VAR
  - ▶ Huber and Rossini (2022) - BART
  - ▶ Clark et al. (2023) - BART
  - ▶ Huber and Koop (2024) - Dirichlet process mixture (DPM)
  - ▶ Hauzenberger et al. (2024) - Gaussian processes (GP)

Minnesota Prior

Let us focus on the 1st equation of the VAR(p) model

$$y_{t1} = \sum_{l=1}^{p} \sum_{j=1}^{m} \phi_{l,1j} y_{t-l,j} + \epsilon_{t1}$$
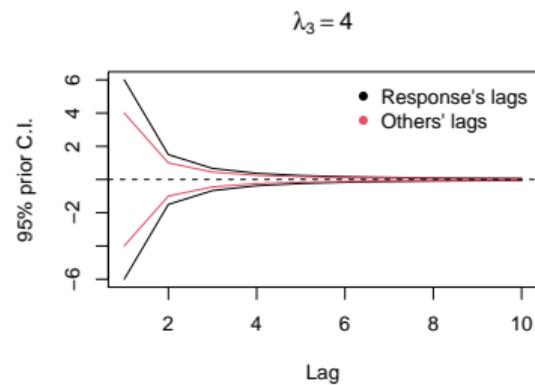
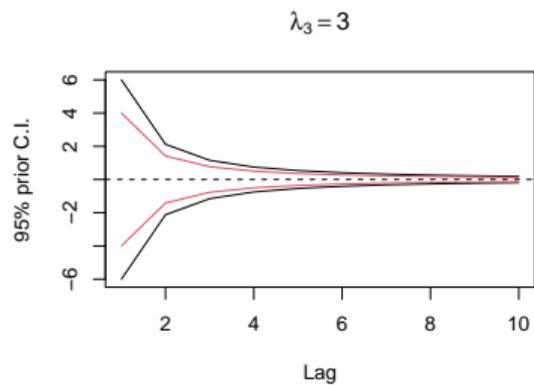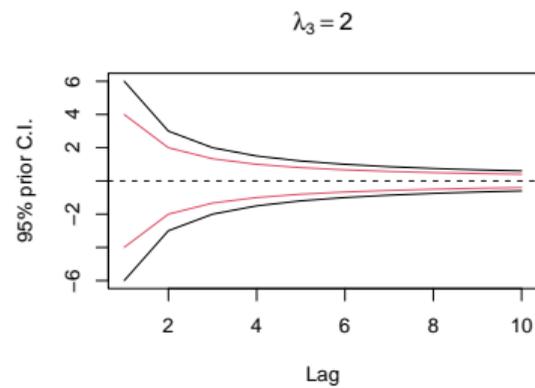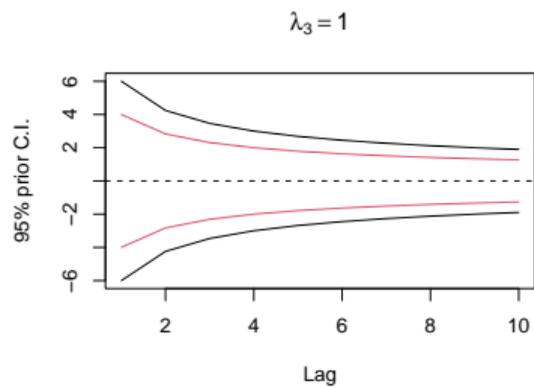The Minnesota prior induces an random walk behavior for y_{t1}:

$$E(\phi_{1,11}) = 1 \quad \text{and} \quad E(\phi_{l,1j}) = 0 \quad \forall l, j \neq 1$$

and

$$V(\phi_{l,1j}) = \begin{cases} \frac{\lambda_1}{l^{\lambda_3}} & j = 1 \\ \frac{\lambda_2}{l^{\lambda_3}} & j \neq 1 \end{cases}$$

Doan, Litterman and Sims (1984) Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1),1-100. Litterman (1986) Forecasting with Bayesian vector autoregressions - five years of experience. *JBES*, 4(1), 25-38.

$\lambda_1 = 9,\ \lambda_2 = 4$

Modeling Σ_t

Recall the VAR(p) structure

$$y_t = \sum_{l=1}^{p} \Phi_l y_{t-l} + \epsilon_t, \quad \epsilon_t \text{ iid } N(0, \Sigma_t),$$

We model Σ_t via a factor stochastic volatility (FSV) framework:

$$\Sigma_t = \Lambda \Omega_t \Lambda + H_t$$

where Λ is an $m \times r$ factor loadings matrix ($r \ll m$), with r and m components of the diagonal matrices Ω_t and H_t following univariate SV models. For further developments about FSV models, see Pitt and Shepard (1999), Aguilar and West (2000), Lopes (2000), Lopes and Carvalho (2007) and Kastner, Fruehwirth-Schnatter and Lopes (2017) and Kastner (2019), among many others.

Chan (2023) argue that for BVAR with FSV structure, SV specifications for Ω_t and H_t are crucial for producing accurate density forecasts.

Our contribution: Minnesota BART

Three-fold extension of Huber and Rossini (2022) and Clark et al. (2023):

- Allowing for high-dimensional data and variable selection via the approach by Linero (2018), and

- Introducing a Minnesota-type shrinkage specification into the BART node splitting selection.

- Factor SV specification for Σ_t.

m BART strutures

Let us introduce the following notation:

- $y_t = (y_{t1}, \ldots, y_{tm})'$.
- $x_t = (y'_{t-1}, \ldots, y'_{t-p})$.
- $G(x_t) = (g_1(x_t), \ldots, g_m(x_t))'$ is a n-dimensional vector BART mean functions.

We replace the linear AR structure by a nonlinear, nonparametric one:

$$y_t = G(x_t) + \epsilon_t, \quad \epsilon_t \sim \text{ iid } N(0, \Sigma_t)$$

The full (hierarchical) model

$$
\begin{aligned}
y_t &= G(x_t) + \Lambda f_t + \eta_t \\
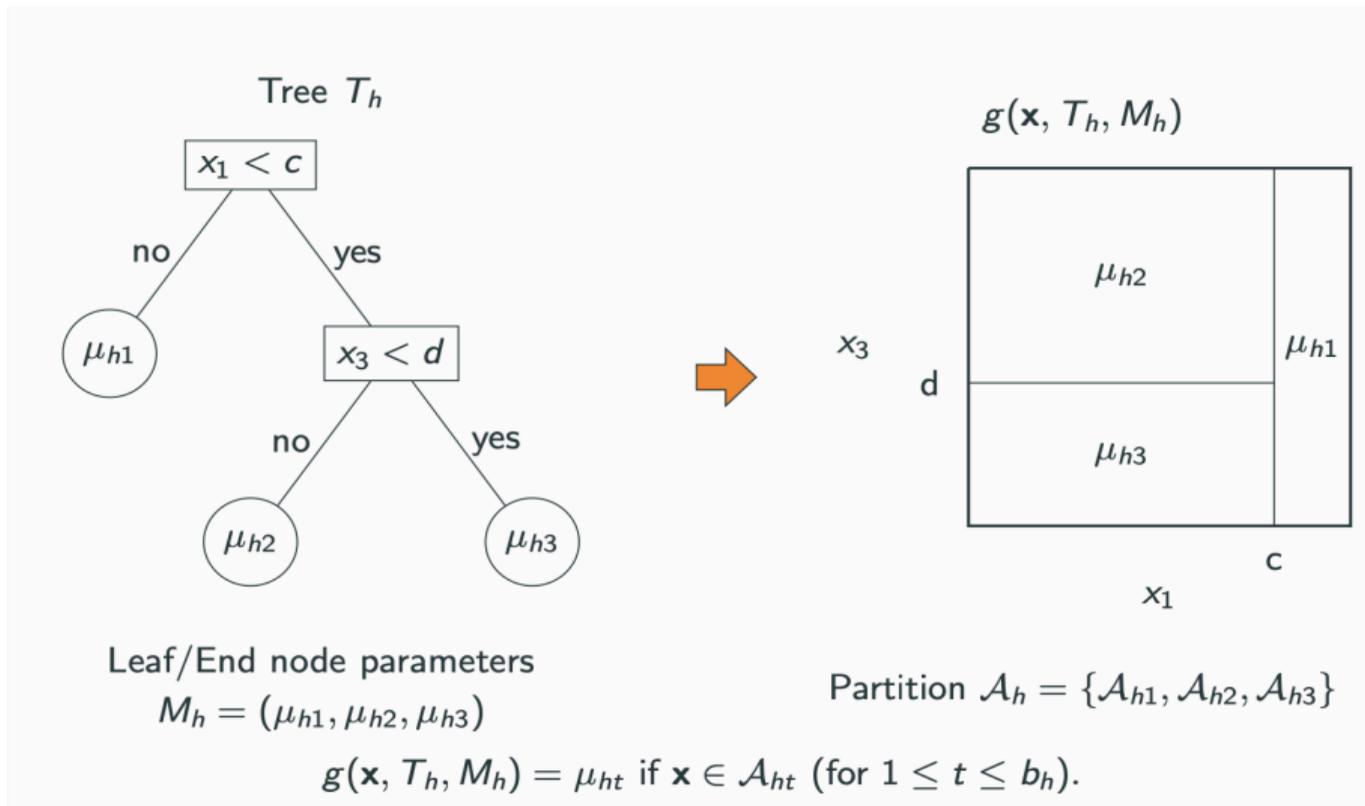f_t &\sim N(0, \Omega_t) \\
\eta_t &\sim N(0, H_t),
\end{aligned}
$$

where the components of H_t and Ω_t follow standard SV models.

Our MCMC scheme will cycle through the following blocks:

- Sampling Λ,
- sampling the components of Ω_t and H_t $\forall t$,
- Sampling the common factors f_t $\forall t$,
- Sampling $g_i(x_t)$ $\forall i = 1, \ldots, m$.

A brief introduction to a tree model

Tree T_h

$x_1 < c$

no / yes

μ_{h1}

$x_3 < d$

no / yes

μ_{h2} μ_{h3}

$g(\mathbf{x}, T_h, M_h)$

x_3

μ_{h2}

μ_{h1}

d

μ_{h3}

c

x_1

Leaf/End node parameters
$M_h = (\mu_{h1}, \mu_{h2}, \mu_{h3})$

Partition $\mathcal{A}_h = \{\mathcal{A}_{h1}, \mathcal{A}_{h2}, \mathcal{A}_{h3}\}$

$g(\mathbf{x}, T_h, M_h) = \mu_{ht}$ if $\mathbf{x} \in \mathcal{A}_{ht}$ (for $1 \le t \le b_h$).

The vector of mean functions, $G(x_t)$

Each component of $G(x_t)$ is modeled as a decision tree ensemble:

$$g(x_t) = \sum_{n=1}^{N} g_n\left(x_t; \mathcal{T}_n, \mathcal{M}_n\right),$$

where

- $\mathcal{T}_n$ denotes a *decision tree* shape,
- $\mathcal{M}_n$ denotes a collection of *leaf node parameters*, and
- $g_n(x_t; \mathcal{T}_n, \mathcal{M}_n)$ is a *regression tree function* that returns the prediction associated to x_t for the pair $(\mathcal{T}_n, \mathcal{M}_m)$.

Prior specification:
$$\pi(\mathcal{T}_m, \mathcal{M}_m) \sim \pi_{\mathcal{T}}(\mathcal{T}_m)\, \pi_{\mathcal{N}}(\mathcal{M}_m \mid \mathcal{T}_m)$$

BART prior

BART proceeds by placing a prior on the regression trees.

Prior independence, given the model hyperparameters θ:

$$\pi\left((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_M, \mathcal{M}_M) \mid \theta\right) = \prod_{m=1}^{M} \pi_{\mathcal{T}}(\mathcal{T}_m \mid \theta)\pi_{\mathcal{M}}(\mathcal{M}_m \mid \mathcal{T}_m).$$

The prior distribution for the trees $\pi_{\mathcal{T}}$ consists of three steps:

1. A prior on the shape of the tree $\mathcal{T}$;
2. A prior for the splitting rules that first selects a predictor by sampling $k_b \sim \text{Categorical}(s)$ where $s = (s_1, \ldots, s_k)^\top$ is a probability vector.
3. A prior on the splitting rules $[x_{k_b} \leq C_b]$ for each branch node of the tree, given k_b

BART splitting rule

- Select a predictor by sampling $k_b \sim \text{Categorical}(s)$, where

$$s = (1/k, \ldots, 1/k).$$

- What if $m = 100$ and $p = 5$, such that $k = mp = 500$?
  BART breaks down with larger number of irrelevant features (Linero (2018)).

- Bias will increase as k increases (again $k = mp$ for VAR's equations).

- Credible intervals will widen as well.
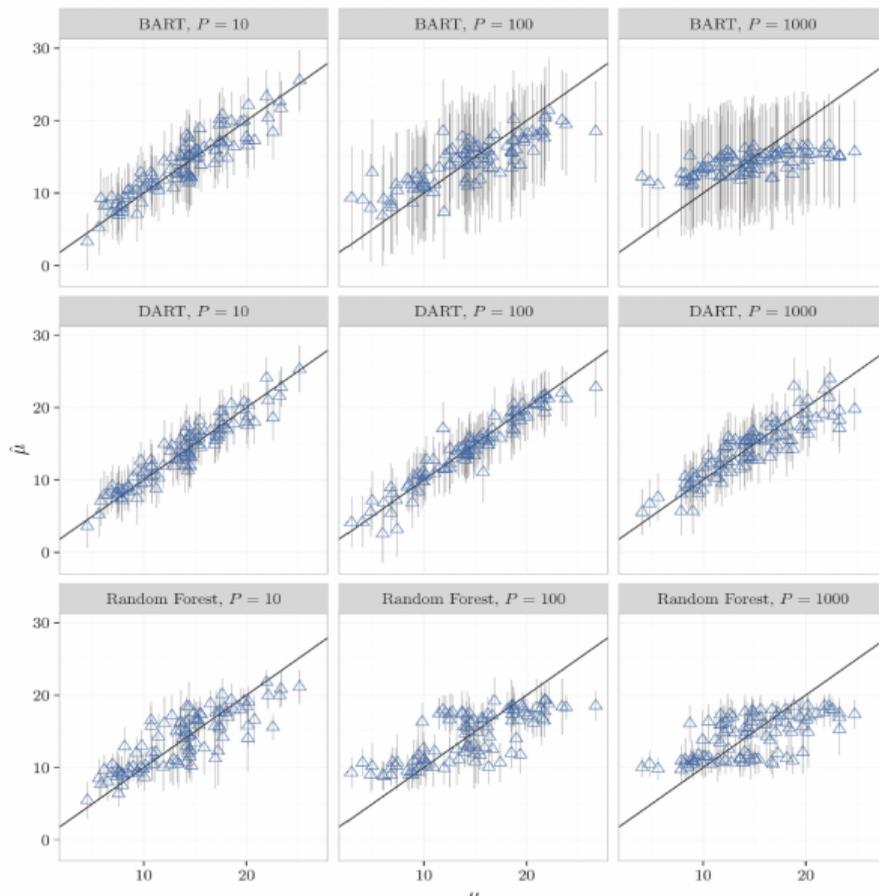
Exercise: BART in a high dimensional setting

Consider the following nonlinear regression

$$y_t = g(x_t) + \epsilon_t,$$
$$g(x_t) = 10sin(\pi x_{t1} x_{i2}) + 20(x_{t3} - 0.5)^2 + 10x_{t4} + 5x_{t5},$$

where

- $\epsilon_t \sim \mathcal{N}(0, 1)$,
- $T = 100$ observations,
- 5 relevant predictors,
- $k - 5$ irrelevant predictors,
- $k = \{10, 100, 1000\}$.

Predictions degrade as k increases, Linero (2018)

Linero's DART prior

If many predictor are potentially irrelevant, why should s_k constant over k?

Linero (2018) propose a solution when k is close or much larger than T:

$$s \sim \text{Dirichlet}(\alpha/k, \ldots, \alpha/k)$$

Full Bayesian variable selection:

$$\frac{\alpha}{\alpha + k} \sim \text{Beta}(0.5, 1).$$

UT Austin gang

Antonio & Jared
Hill, Linero, and Murray (2020) Bayesian Additive Regression Trees: A Review and Look Forward, *Annual Review of Statistics and Its Application*, Volume 7, pages 251-278 - https://doi.org/10.1146/annurev-statistics-031219-041110

Carlos, Drew, Rafael & Pedro
`stochtree` (short for "stochastic trees") - https://stochtree.ai

Boosted decision tree models (like xgboost, LightGBM, or scikit-learn's HistGradientBoostingRegressor) are great, but often require time-consuming hyperparameter tuning. stochtree can help you avoid this, by running a fast Bayesian analog of gradient boosting (called BART – Bayesian Additive Regression Trees).

Minnesota BART

Rule 1: The past values of a specific variable play a more significant role in predicting its current value compared to the past values of other variables.

Rule 2: The most recent past is considered more influential in predicting current values than events further in the past.

Therefore, for equation m, the prior for the splits probability is defined::

$$(s_{1m}, \ldots, s_{km}) \sim \text{Dirichlet}(\phi_{1n}, \ldots, \phi_{kn}) \tag{1}$$

The scale parameters of the Dirichlet distribution are defined are defined as follows:

$$\phi_{im} = \begin{cases} \frac{\lambda_1}{l^2}, & \text{for the scale on the } l\text{-th lag of variable } i, \\ \frac{\lambda_2 \cdot \rho}{l^2}, & \text{for the coefficient on the } l\text{-th lag of variable } j, \, j \neq i, \end{cases}$$

Minnesota BART

Draws from $Dirichlet\left(\lambda, \frac{\lambda}{4}, \frac{\lambda}{9}\right)$. This figure illustrates the effect of varying λ on the concentration parameters of the *Dirichlet* prior on the simplex for $\lambda = (1, 3, 10)$. The vertices of the simplex correspond to one-sparse probability vectors, the edges represent two-sparse vectors, and the interior points indicate denser probability distributions.
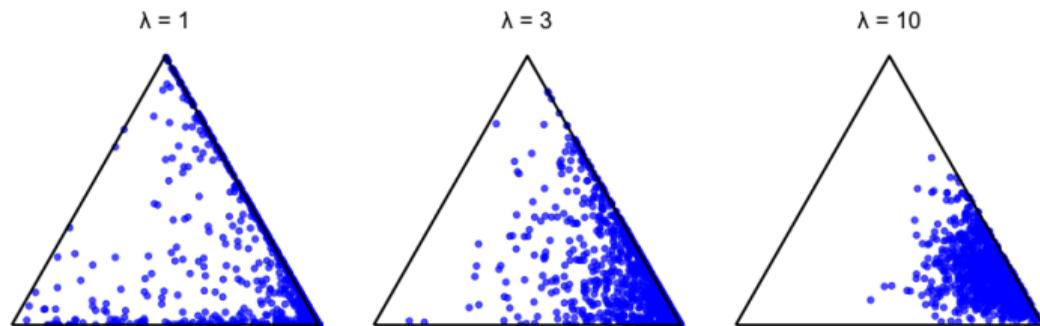
Bayesian inference

- **Prior features (in a nutshell)**
  - ▶ Choice of prior and hyperparameters from BART literature.
  - ▶ Horseshoe prior used for any linear conditional mean coefficients

- **MCMC features (in a nutshell)**
  - ▶ Standard MCMC steps from BVAR and BART.
  - ▶ Novel updating step for the split probabilities:

$$s_1, \ldots, s_k | \phi, \text{data} \sim \text{Dirichlet}(\phi_1 + n_1, \ldots, \phi_k + n_k)$$

  where n_k are the number of splits on predictor k over the ensemble.

Posterior inclusion probability

- In order to illustrate the properties of the proposed priors we conduct a simulation study where we aim to assess the efficacy of DART-VAR and Minnesota DART in recovering the sparsity pattern.

- We use *posterior inclusion probability* as metric for variable selection.

$$\text{PIP}_k = \text{Pr}(\text{predictor } k \text{ appears in the ensemble} \mid \text{data})$$

- We will report the results of the **first equation** of the estimated dynamic system.

Simple simulation exercise

The data is generated from a VAR(5) model:
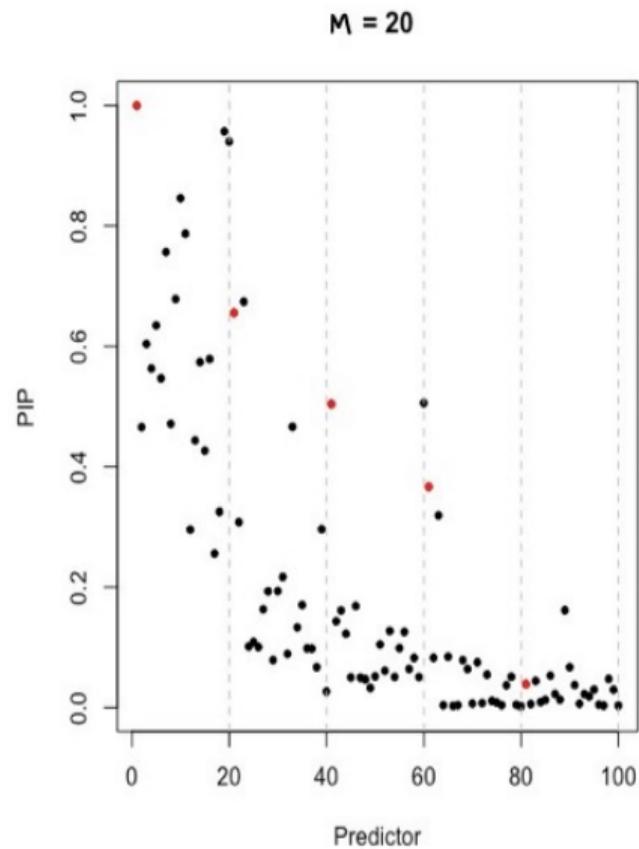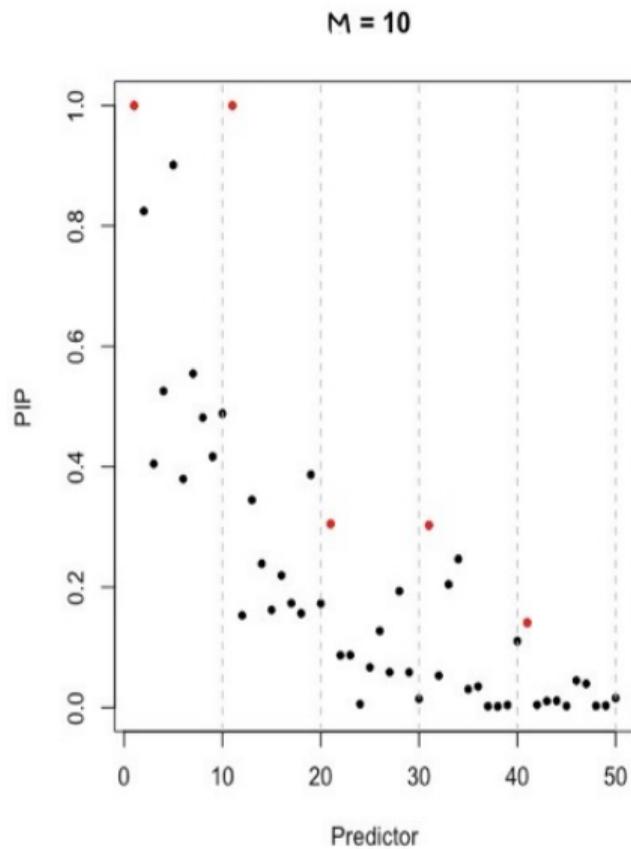
$$\Phi_1 = 0.65 I_m \tag{2}$$

and

$$\Phi_j = (-1)^{j-1}(0.4225) I_m, \quad j = 2, \ldots, 5, \tag{3}$$

for $m = 10$ or $m = 20$.

The coefficients decrease for distant lags, reflecting the conventional wisdom that recent lags hold greater importance than those further in the past.
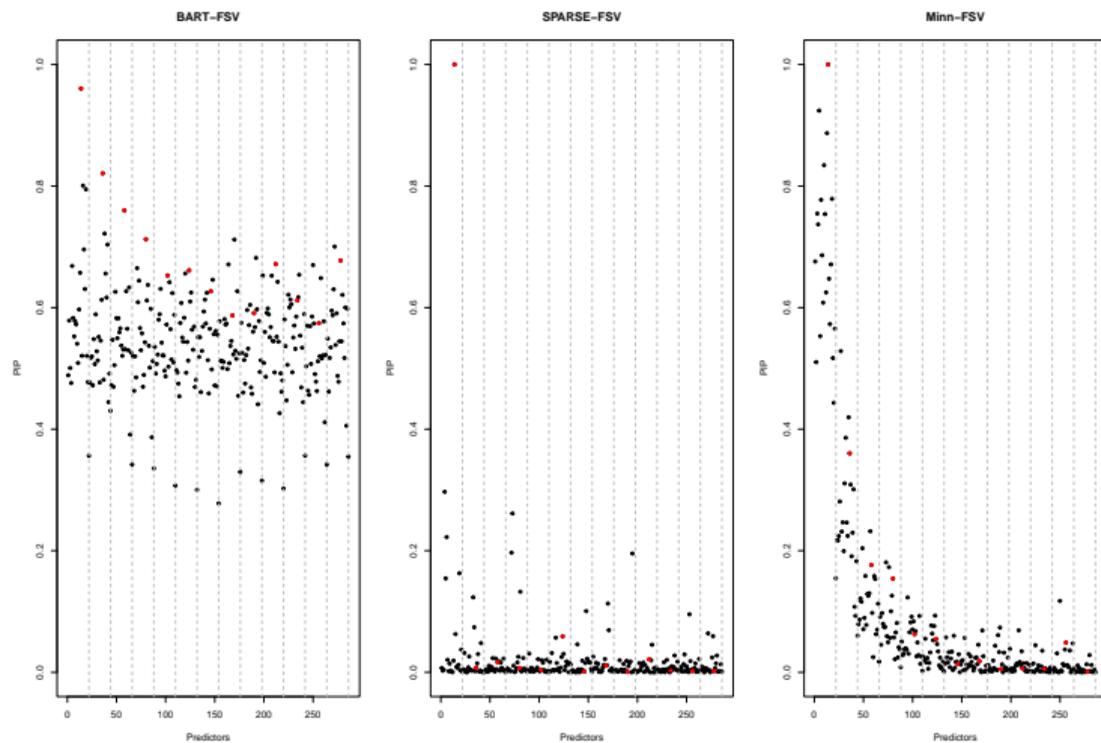
Minnesota DART prior

Real data exercise
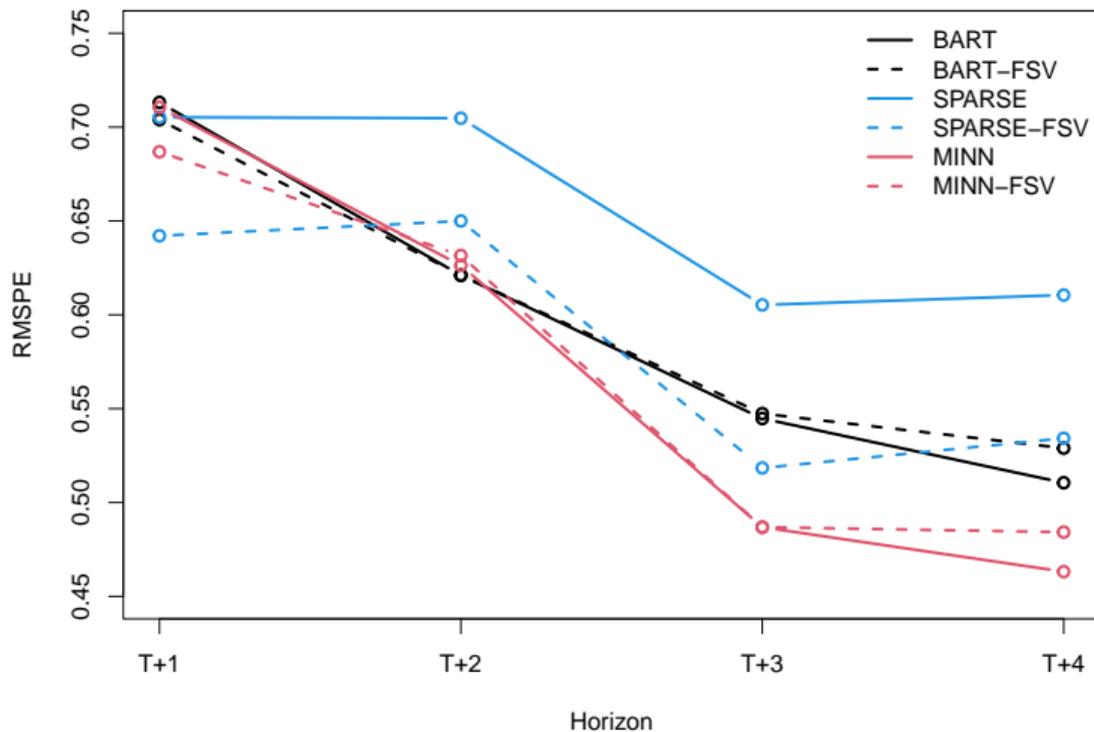
- Data: 22 series from FRED-QD, McCracken and Ng (2016).

- Time span: 1965Q1 - 2019Q4.

- Expanding window: 2005Q1 to 2019Q4.

- Horizons: $h = 1, 2, 3, 4$.

- Evaluation metric: Root mean squared predictive error (RMSPE)

- Baseline model: BVAR-FSV with Minnesota prior

Inclusion probabilities - CPI

RMSPE - CPI

BART/SPARSE/MINN = Uniform/Dirichlet/Minnesota splitting
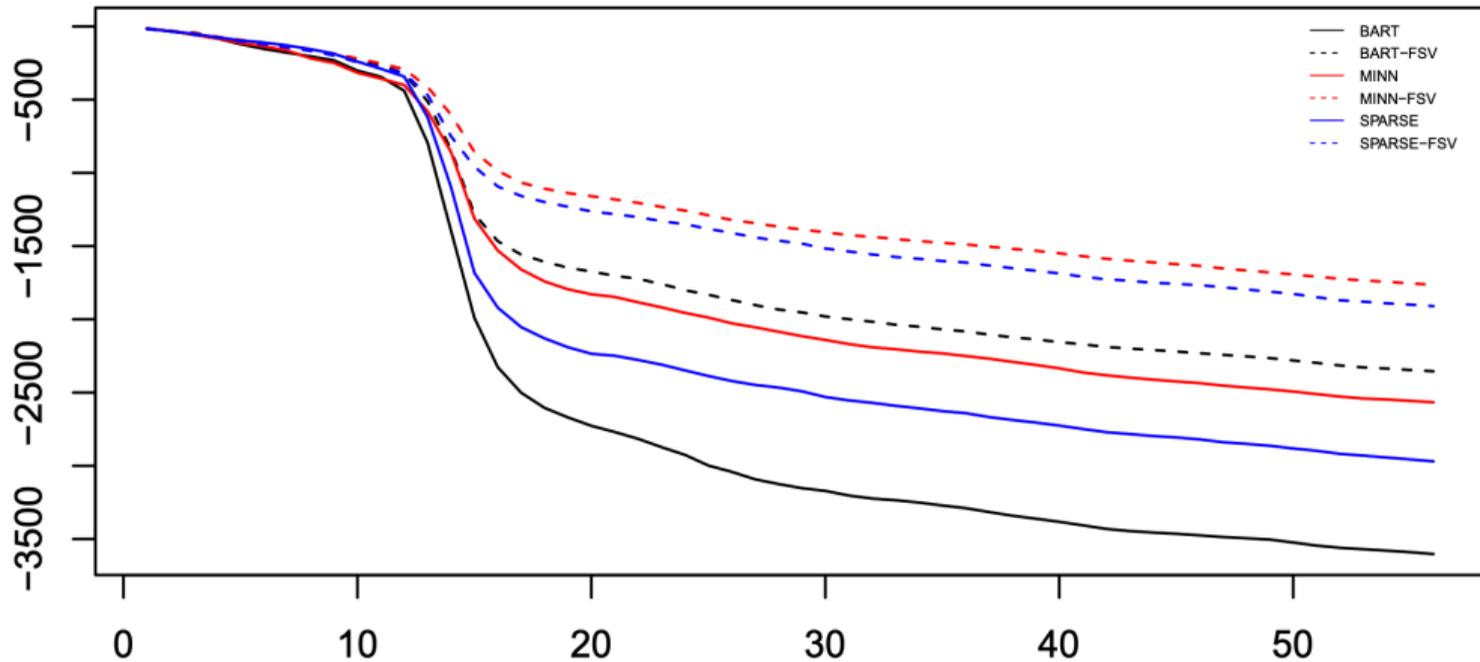
Log predictive density scores (LPDS)

- To obtain a more comprehensive evaluation, we consider a metric that account for for the models ability to predict higher-order moments of the predictive distribution - **Log Predictive Density Score**

$$\text{LPDS} = \log p(y_{t_0+1}, \ldots, y_T \mid y^{tr}) = \sum_{t=t_0+1}^{T} \log p(y_t \mid y^{t-1})$$

- The first t_0 time series observations, $y^{tr} = (y_1, \ldots, y_{t_0})$, are designated as the "training sample," while the remaining observations, $y_{t_0+1}, \ldots, y_T$, are used for evaluation based on the log predictive density.

- Each probability split prior specification for the mean function is shown under both the homoskedastic and stochastic volatility (SV) settings, where the former is represented by a continuous line and the latter by a dashed line.
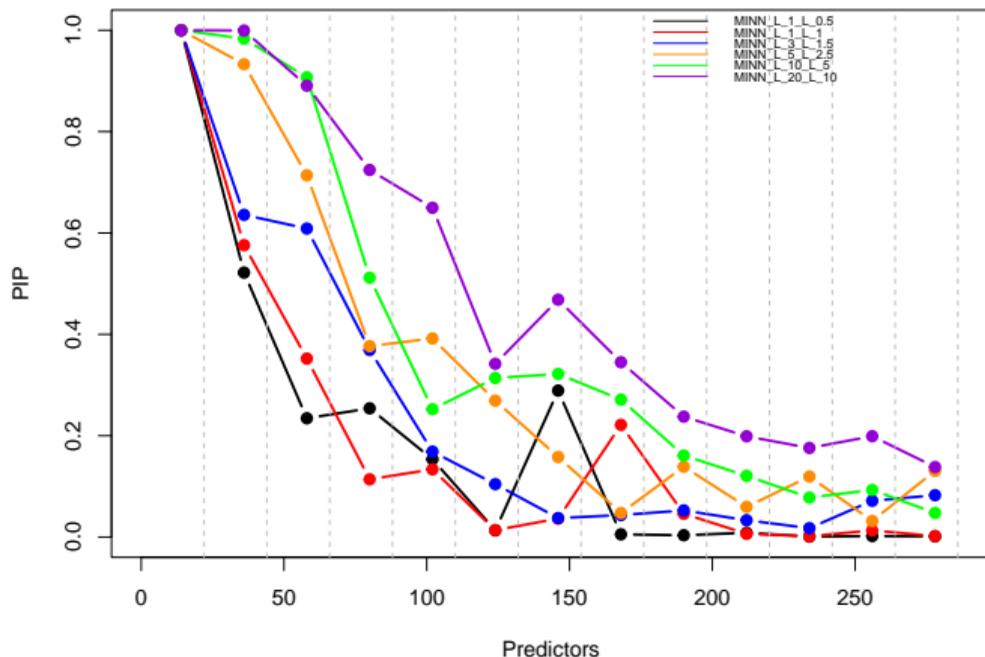
$LPDS_{t+4}$

Prior Elicitation

- The choice of λ is of critical importance, as it plays a central role in determining the expected level of shrinkage in the model.

- **Empirical Analysis:** We evaluate different levels of λ using a grid of values ($\lambda_1 = \{1, 3, 5, 10, 20\}$, $\lambda_2 = \{0.5, 1, 1.5, 2.5, 5, 10\}$) and assess their impact on the log-predictive density score relative to the standard BART prior.

- **Impact of λ on Shrinkage Forecasting:** Higher values of λ lead to a more gradual decay in posterior inclusion probabilities, preserving the influence of lags and cross-lags over a longer range. This highlights the importance of carefully selecting λ, as it directly affects variable selection, model interpretability, and forecasting accuracy.
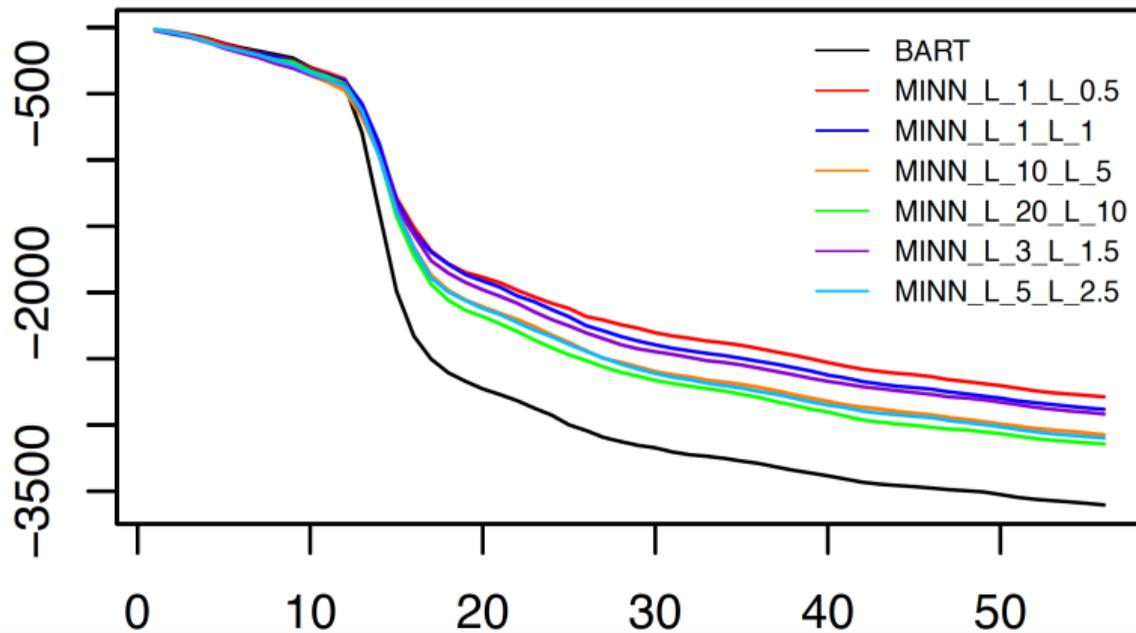
PIP for CPI

In-sample PIP for the CPI's own lag across different grid values of $\lambda_1 = \{1, 3, 5, 10, 20\}$ and $\lambda_2 = \{0.5, 1, 1.5, 2.5, 5, 10\}$.

4-steps ahead LPDS

$$LPDS_{t+4}$$

Legend:
- BART
- MINN_L_1_L_0.5
- MINN_L_1_L_1
- MINN_L_10_L_5
- MINN_L_20_L_10
- MINN_L_3_L_1.5
- MINN_L_5_L_2.5

Final Remarks

- **Advancing Multivariate BART for High-Dimensional Analysis:** We introduce a structured prior that enables shrinkage in split probabilities, addressing sparsity and time dependence limitations in high-dimensional VARs.

- **Empirical Validation & Forecasting Gains:** Our priors improve forecast accuracy, particularly for higher-order moments, with the Minnesota specification outperforming the sparse alternative.

- **Broader Applications & Future Directions:** The framework extends to structural analysis (GIRFs, LP) and can be further improved through scalable sampling methods and time-varying parameters.

Bibliography I

Bańbura, M., D. Giannone, and L. Reichlin (2010). Large bayesian vector auto regressions. *Journal of applied Econometrics 25*(1), 71–92.

Carriero, A., T. E. Clark, and M. Marcellino (2019). Large bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics 212*(1), 137–154.

Chan, J. C. (2023). Comparing stochastic volatility specifications for large bayesian vars. *Journal of Econometrics 235*(2), 1419–1446.

Clark, T. E., F. Huber, G. Koop, M. Marcellino, and M. Pfarrhofer (2023). Tail forecasting with multivariate bayesian additive regression trees. *International Economic Review 64*(3), 979–1022.

Hauzenberger, N., F. Huber, M. Marcellino, and N. Petz (2024). Gaussian process vector autoregressions and macroeconomic uncertainty. *Journal of Business & Economic Statistics*, 1–17.

Bibliography II

Huber, F. and G. Koop (2024). Fast and order-invariant inference in bayesian vars with nonparametric shocks. *Journal of Applied Econometrics*.

Huber, F. and L. Rossini (2022). Inference in bayesian additive vector autoregressive tree models. *The Annals of Applied Statistics 16*(1), 104–123.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association 113*(522), 626–636.

McCracken, M. W. and S. Ng (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Bibliography III

- Doan, Litterman and Sims (1984) Forecasting and conditional projection using realistic prior distributions, *Econometric Reviews*, 3, 1-100.
- Litterman (1986) Forecasting with BVARs: Five Years of Experience, *Journal of Business and Economic Statistics*, 4(1), 25-38.
- Kadiyala and Karlsson (1993) Forecasting with generalized BVARs. *Journal of Forecasting*, 12, 365-78.
- Kadiyala and Karlsson (1997) Numerical methods for estimation and inference in BVAR models, *Journal of Applied Econometrics*, 12, 99-132.
- Lopes, Moreira and Schmidt (1999) Hyperparameter estimation in forecasting models, *Computational Statistics and Data Analysis*, 29, 387-410.
- Primiceri (2005) Time varying structural VARs and monetary policy. *The Review of Economic Studies*, 72(3), 821-852.
- Koop and Korobilis (2013) Large time-varying parameter VARs, *Journal of Econometrics*, 177, 185-198.
- Kastner and Huber (2020) Sparse Bayesian vector autoregressions in huge dimensions, *Journal of Forecasting*, 39(7), 1142-1165.

Access to paper and slides

The 1st draft of the paper can be found here:
https://arxiv.org/abs/2503.13759

The slides can be download from the URL address
http://hedibert.org/wp-content/uploads/2025/09/minnesota-bart-slides.pdf

or via the following QR Code