Large Bayesian Additive Vector Autoregressive Tree Models

Hedibert Freitas Lopes[1]
INSPER Institute of Education and Research

VII COBAL and XVII EBEB
UFMG, Belo Horizonte, December 2024

[1]Co-authors: Lima, Carvalho and Herren. FAPESP-2023/02538-0.

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

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

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

- Intercept $+\ mp$ regressors per equation.

- $m(1 + mp)$ parameters in $(\mu, \Phi_1, \ldots, \Phi_p)$.

Evolution of Bayesian VAR models

- Small/medium size VAR
  - Doan, Litterman and Sims (1984/1986) - Litterman's 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)

Our contribution: BAVART model

Two-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.

The BAVART model

We replace the linear autoregressive structure by a nonlinear one:

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

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

Modeling Σ_t: a factor analysis approach

Stochastic volatility specifications are crucial for producing accurate density forecasts, Chan (2023). Therefore, we model Σ_t as follows:

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

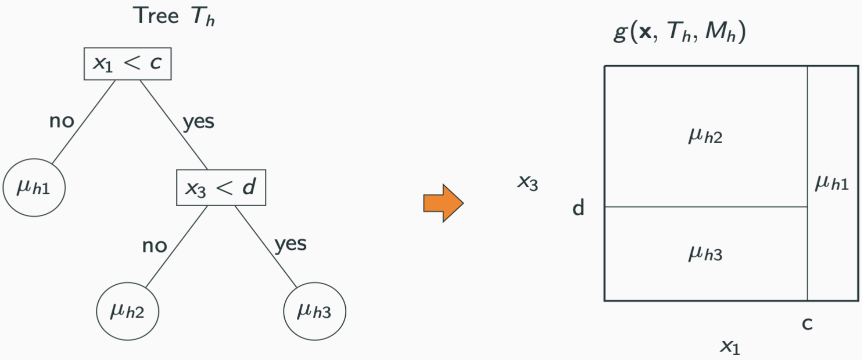
where Λ is an $m \times q$ factor loadings matrix ($q \ll m$), $H_t = \text{diag}(h_{t1}, \ldots, h_{tm})$ and $\Omega_t = \text{diag}({}_{t1}, \ldots, \omega_{tq})$.

The full (hierarchical) model is written as

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

The components of H_t and Ω_t follow standard stochastic volatility (SV) models.

A brief introduction to a tree model



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

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

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

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

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

$$g(\boldsymbol{x}_t) = \sum_{r=1}^{N} g_r\left(\boldsymbol{x}_t; \mathcal{T}_r, \mathcal{M}_r\right),$$

where

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

Prior specification:

$$\pi(\mathcal{T}_r, \mathcal{M}_r) \sim \pi_{\mathcal{T}}(\mathcal{T}_r)\, \pi_{\mathcal{M}}(\mathcal{M}_r \mid \mathcal{T}_r)$$

BART prior

BART proceeds by placing a prior on the regression trees.

Prior independence, given the model hyperparameters θ:

$$\pi\left((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_N, \mathcal{M}_N) \mid \theta\right) = \prod_{r=1}^{N} \pi_{\mathcal{T}}(\mathcal{T}_r \mid \theta) \pi_{\mathcal{M}}(\mathcal{M}_r \mid \mathcal{T}_r).$$

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 $p = 5$?
  Linero (2018): break down in the presence of larger number of potentially irrelevant features.

- Bias will increase as k increases (VAR: $k = mp$).

- Credible intervals will widen as well.

Exercise: BART in a high dimensional setting

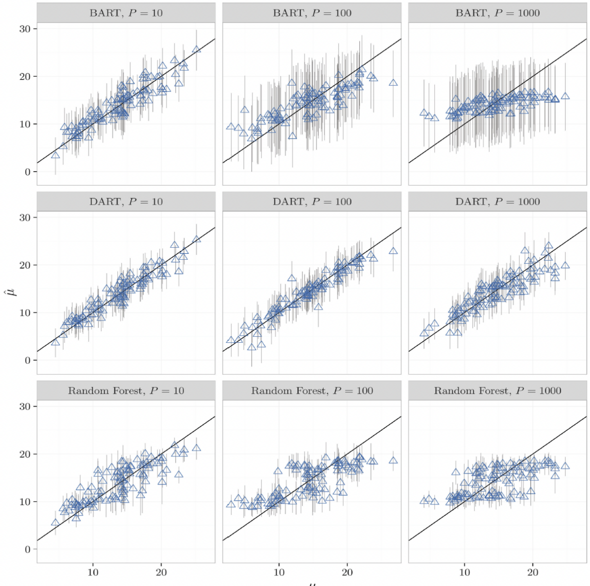
Consider the following nonlinear regression

$$
\begin{aligned}
y_i &= g(x_i) + \epsilon_t, \\
g(x_i) &= 10 sin(\pi x_{i1} x_{i2}) + 20(x_{i3} - 0.5)^2 + 10x_{i4} + 5x_{i5},
\end{aligned}
$$

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)



Solution: 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:

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

Minnesota DART

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, we model s for any node in equation m as follows:

$$s_1, \ldots, s_k \mid \phi \sim \text{Dirichlet}(\phi_1, \ldots, \phi_k)$$

with

$$\phi_k \propto \frac{\lambda_1}{l^2},$$

when k represents lag l of y_m, for $l = 1, \ldots, p$, and

$$\phi_k \propto \frac{\lambda_2}{l^2}$$

when k represents lag l of y_j, for $j \neq m$.

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.

Another simulation exercise

- 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 will be reporting the *posterior inclusion probability* as metric for variable selection.

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

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

Experiment A

The data is generated from a linear m dimensional VAR(1) model:

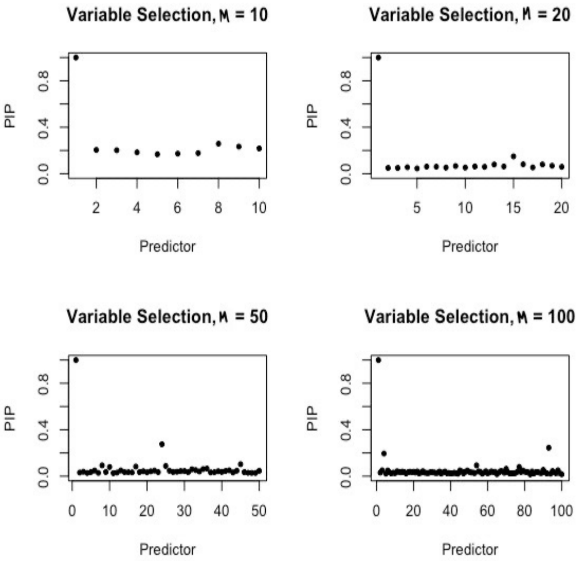
$$\Phi = 0.5I_m$$

and with $m = 10, 20, 50, 100$.

True sparsity: behavior of each variable only depends on its own past.

$m = 100$: Each equation has 99 redundant variables.

Linero's DART prior



Experiment B

The data is generated from a VAR(5) model:

$$\Phi_1 = 0.65 I_m \tag{1}$$

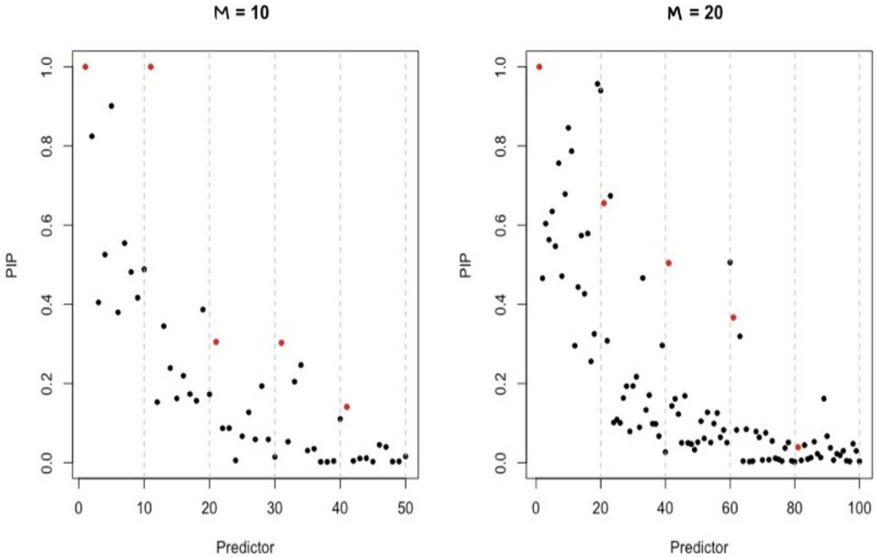
and

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

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: 15 series from FRED-QD, Jurado et al. (2015).
- Time span: 1965Q1 - 2001Q4.
- Expanding window: 2002Q1 to 2010Q4.
- Horizons: $h = 1, 2, 3, 4$.
- Evaluation metric: Root mean squared predictive error (RMSPE)
- Baseline models:
  - BART
  - BVAR-SV with Minnesota prior

Relative to RMSPE(BART) - CPI

Table: CPI Forecasting Results

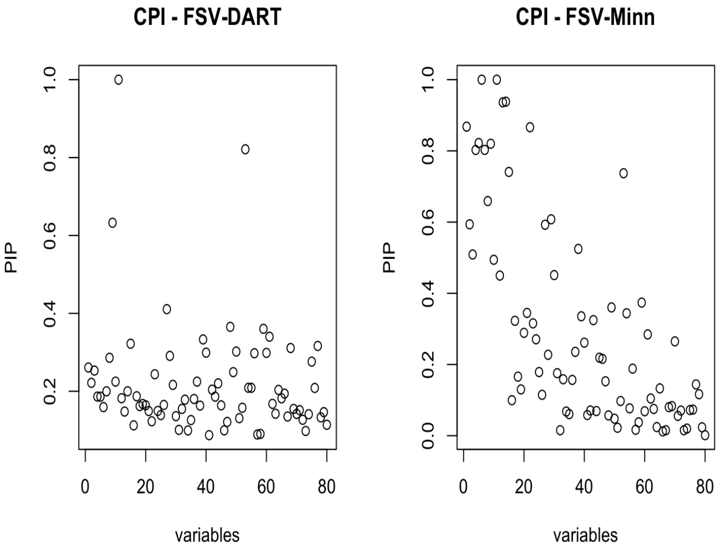
	t+1	t+2	t+3	t+4
DART	0.665	0.821	0.618	0.521
Minn	0.912	0.683	0.685	0.734
BART-FSV	0.645	0.617	0.414	0.553
DART-FSV	0.598	0.918	0.865	0.816
Minn-FSV	0.547	0.750	0.670	0.521

Relative to RMSPE(BVAR-SV) - CPI

Table: CPI Forecasting Results

	t+1	t+2	t+3	t+4
BART	1.615	1.250	1.249	1.078
DART	1.074	1.027	0.772	0.561
Minn	1.473	0.854	0.856	0.791
BART-FSV	1.042	0.772	0.517	0.596
DART-FSV	0.966	1.147	1.080	0.880
Minn-FSV	0.883	0.938	0.837	0.803

Inclusion probabilities



Relative to RMSPE(BART) - GDP

Table: GDP Forecasting

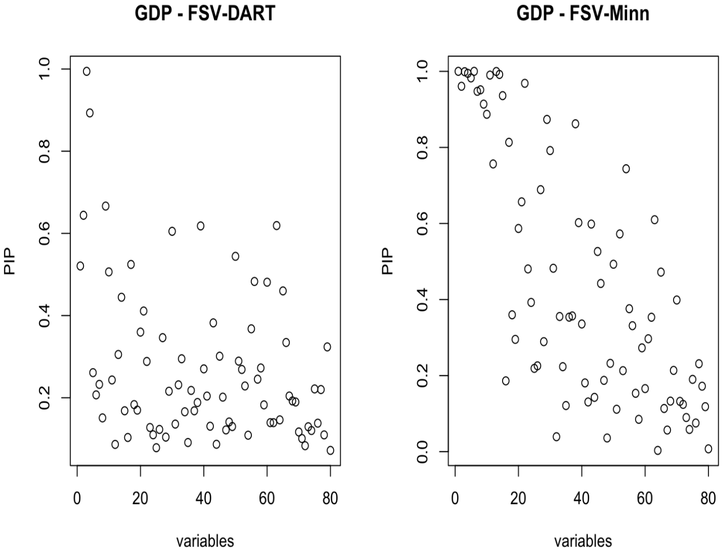
	t+1	t+2	t+3	t+4
DART	0.923	1.127	1.335	1.454
Minn	0.681	1.001	1.124	1.188
BART-FSV	0.944	0.832	0.875	0.886
DART-FSV	0.608	1.197	1.429	1.621
Minn-FSV	0.832	0.775	0.948	1.081

Relative to RMSPE(BVAR-SV) - GDP

Table: GDP Forecasting

	t+1	t+2	t+3	t+4
BART	1.547	5.418	5.302	5.557
DART	1.428	6.105	7.078	8.080
Minn	1.054	5.424	5.962	6.603
BART-FSV	1.460	4.507	4.641	4.922
DART-FSV	0.940	6.487	7.577	9.010
Minn-FSV	1.287	4.197	5.027	6.006

Inclusion probabilities



Final remarks

- Methodological contribution:
  - Extension of Huber and Rossini (2022) and Clark et al. (2023), by allowing for high-dimensional data and variable selection using Linero (2018).
  - Node splitting prior that resembles the well-known Minnesota-type shrinkage priors.
- Empirical results :
  - Better forecasting performance when compared to the baseline BART.
  - Better forecasting performance (for some horizons) when compared to a BVAR-SV.
- Current questions:
  - What is the best way to set a prior for λ?
  - Do the sparsity-inducing variable selection aligns with economic theory?
  - The sparsity/density debate: Giannone et al. (2021), Fava and Lopes (2021)

Epic slide 1: É tempo de Botafogo!

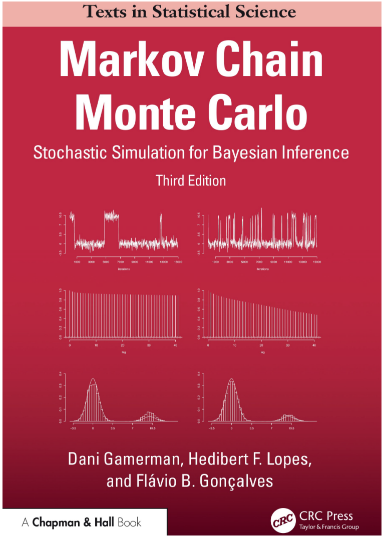
Epic slide 2: Glória Eterna

BOTAFOGO **CAMPEÃO DA LIBERTADORES 2024**

ALEJANDRO PAGNI / AFP

FOLHA DE S.PAULO

Epic slide 3: Gamerman, Lopes and Gonalves (2026)



Thanks!

Any thoughts?

plima@utexas.edu
hedibertfl@insper.edu.br

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.

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

Bibliography II

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

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

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.

Jurado, K., S. C. Ludvigson, and S. Ng (2015). Measuring uncertainty. *American Economic Review 105*(3), 1177–1216.

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

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.

You can access the slides using this QR code