Minnesota BART*

Pedro A. Lima[1], Carlos M. Carvalho[1], Hedibert F. Lopes[2], and Andrew Herren[1]

[1]The University of Texas at Austin
[2]Insper

March 18, 2025

Abstract

Vector autoregression (VAR) models are widely used for forecasting and macroeconomic analysis, yet they remain limited by their reliance on a linear parameterization. Recent research has introduced nonparametric alternatives, such as Bayesian additive regression trees (BART), which provide flexibility without strong parametric assumptions. However, existing BART-based frameworks do not account for time dependency or allow for sparse estimation in the construction of regression tree priors, leading to noisy and inefficient high-dimensional representations. This paper introduces a sparsity-inducing Dirichlet hyperprior on the regression tree's splitting probabilities, allowing for automatic variable selection and high-dimensional VARs. Additionally, we propose a structured shrinkage prior that decreases the probability of splitting on higher-order lags, aligning with the Minnesota prior's principles. Empirical results demonstrate that our approach improves predictive accuracy over the baseline BART prior and Bayesian VAR (BVAR), particularly in capturing time-dependent relationships and enhancing density forecasts. These findings highlight the potential of developing domain-specific nonparametric methods in macroeconomic forecasting.

Keywords: Bayesian non-parametrics, non-linear vector autoregressions, shrinkage prior, forecasting.

*Corresponding author: Pedro Lima. The University of Texas at Austin. *Email*: plima@utexas.edu. We would like to thank the participants of the 2024 European Seminar on Bayesian Econometrics (ESOBE), VII COBAL, and XVII EBEB for their many constructive comments and useful suggestions. Hedibert Lopes also acknowledges partial financial support from FAPESP grants 2023/02538-0 and 2024/01027-4.

1 Introduction

This paper introduces a multivariate Bayesian additive regression tree (BART) model for macroeconomic forecasting that incorporates structured priors to allow for variable selection and account for time dependency. We extend the standard BART framework to allow for a sparse vector autoregressions (VARs) estimation by introducing a sparsity-inducing Dirichlet hyperprior on the regression tree's splitting probabilities. This allows for automatic variable selection, reducing overfitting and improving computational efficiency in large-scale models.

Additionally, we propose a structured shrinkage prior that decreases the probability of splitting on higher-order lags, aligning with the Minnesota prior's principles. This addresses a fundamental limitation of existing BART-based VAR models, which fail to incorporate economic constraints on lag selection and ignore temporal dependencies. We also analyze how different levels of our shrinkage parameter affects the splitting probabilities, results demonstrate that higher values of the parameter lead to a more gradual decay in posterior inclusion probabilities, preserving the influence of lags and cross-lags for a longer range. This choice also can affect the forecasting performance of the model. By integrating these two enhancements, our approach preserves the flexibility of BART while imposing meaningful structure, leading to improved interpretability and forecasting accuracy.

We evaluate our model through an empirical application to U.S. macroeconomic forecasting. Compared to standard BART and Bayesian VAR (BVAR) models, our approach improves predictive accuracy, particularly in capturing time-dependent relationships and higher-order moments of the predictive distribution. In particular, we show that our method enhances density forecasts for key macroeconomic variables such as the Federal Funds Rate and inflation. These findings highlight the potential of structured nonparametric methods for macroeconomic forecasting.

Vector autoregression (VAR) models have been widely used for forecasting and structural analysis of macroeconomic variables (Doan et al. (1984); Litterman (1986); Bańbura et al. (2010); Koop (2013); Carriero et al. (2019); Kastner and Huber (2020)). However, as the number of time series included in the model increases, the number of parameters grows quadratically, leading to concerns about overparameterization and in-sample overfitting. To address these challenges, the Bayesian literature on VARs has developed various shrinkage prior specifications.

Despite these advancements, most Bayesian VAR models still assume a linear relationship between endogenous variables and their lags. While macroeconomic relationships are often stable over time, allowing a linear approximation to fit the data reasonably

well, this assumption can break down during shocks that alter the economy's dynamics (Huber et al. (2023)). Failing to account for such events can result in poor out-of-sample performance and misinterpretation of impulse response functions.

More recently, nonparametric approaches such as Bayesian additive regression trees (BART) have gained attention as a flexible alternative (Chipman et al. (2010)). BART uses regression trees as weak learners, allowing for complex relationships to be modeled without strong parametric assumptions. However, existing frameworks (Huber and Rossini (2022); Clark et al. (2023)) do not accommodate high-dimensional settings or account for time dependency in the construction of regression tree priors. Our paper directly addresses these gaps.

The remainder of the paper is structured as follows. Section 2 introduces the multivariate BART model and provides a necessary introduction to the BART framework. Section 3 develops the prior construction. Section 4 details the prior setup and posterior computation. Sections 5 and 6 present our empirical results: Section 5 discusses the dataset and provides in- and out-of-sample model evidence, while Section 6 focuses on macroeconomic forecasting results. The final section summarizes and concludes.

2 Tree Based Vector Autoregression model

2.1 The Model

Define $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{nt})'$ as a vector of endogenous variables of dimension $n \times 1$. Define $\boldsymbol{x}_t = \left(\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p}\right)'$ a $k(= np)$ dimensional vector of covariates, where p is the number of lags. We also define $\boldsymbol{G}(\boldsymbol{x}_t) = (g_1(\boldsymbol{x}_t), \ldots, g_n(\boldsymbol{x}_t))'$ as a n-dimensional vector of nonparametric functions :

$$\mathbf{y_t} \;=\; \boldsymbol{G}(\boldsymbol{x}_t) + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma}_t), \quad \text{for } t = 1, \ldots, T \tag{1}$$

Each function in the vector $\boldsymbol{G}(\cdot)$, where $\boldsymbol{G} : \mathbb{R}^k \to \mathbb{R}^n$, hence $g\left(\boldsymbol{x}_t\right) : \mathbb{R}^k \to \mathbb{R}$ will be approximated by a Bayesian Additive Regression Trees (BART) model, which is discussed in detail in Section 2.2.

Research on large Bayesian vector autoregressions (VARs) shows evidence that stochastic volatility specifications are well supported by the data (Carriero et al. (2016); Koop (2013); Chan (2020)) and achieving precise density forecasts. Recent work by Chan (2023) demonstrates that the factor stochastic volatility and Cholesky stochastic volatility specifications outperform the standard common stochastic volatility model.

Therefore, following Huber and Rossini (2022), the conditional covariance structure is specified as factor stochastic volatility (FSV) (Pitt and Shephard (1999); Aguilar and

West (2000); Chib et al. (2006); Lopes and Carvalho (2007); Kastner and Huber (2020)). More precisely, the error term is decomposed as:

$$\boldsymbol{\varepsilon}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\eta}_t, \tag{2}$$

where $\boldsymbol{f}_t = (f_{1,t}, \ldots, f_{r,t})$ is a $r \times 1$ vector of latent factors and $\boldsymbol{\Lambda}$ is the associated $n \times r$ factor loading matrix. This factor specification is not identified. The latent factors and the idiosyncratic errors are assumed to be independent and jointly Gaussian:

$$\begin{pmatrix} \boldsymbol{\eta}_t \\ \boldsymbol{f}_t \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Omega}_t & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_t \end{bmatrix} \right). \tag{3}$$

where $\boldsymbol{\Omega}_t = diag\left(e^{h_{1,t}}, \ldots, e^{h_{n,t}}\right)$ and $\boldsymbol{H}_t = diag\left(e^{h_{n+1,t}}, \ldots, e^{h_{n+r,t}}\right)$ are diagonal matrices. The evolution of the log-variance process for $i = 1, \ldots, n + r$ is defined as:

$$h_{i,t} = \mu_i + \phi_i(h_{i,t} - \mu_i) + \eta^h_{i,t}, \quad \eta^h_{i,t} \sim \mathcal{N}\left(0, \sigma_i^2\right) \tag{4}$$

for $t = 2, \ldots, T$. For $t = 1$, we assume a stationary distribution $h_{i,1} \sim \mathcal{N}\left(\mu_i, \frac{\sigma_i^2}{1-\phi_i^2}\right)$. The number of factors are defined using an upper bound as in Aguilar and West (2000). However, in most practical applications, a small number of factors is sufficient to capture the dynamics of the covariance structure, as seen in Bolfarine et al. (2024); Frühwirth-Schnatter et al. (2024).

To address this, we impose a shrinkage prior on the columns of the $\boldsymbol{\Lambda}$ matrix, which pushes irrelevant factors toward zero. For this purpose, we adopt the horseshoe prior proposed by Carvalho et al. (2010). Conditional on the latent factors, this non-parametric VAR becomes n unrelated regressions, estimated equation-by-equation.
Rewriting the model in terms of full-data matrices we have:

$$\boldsymbol{Y} = G(\boldsymbol{X}) + \boldsymbol{\varepsilon} \tag{5}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)'$ and $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots \boldsymbol{y}_T)'$, is a $T \times k$ and $T \times n$ matrix respectively.

2.2 The Learning Function: A BART Review

Our choice estimating $\boldsymbol{G} : \mathbb{R}^k \to \mathbb{R}^n$ is using a sequence of nonparametric decision tree ensemble such that:

$$g_j(\boldsymbol{X}) \approx \sum_{m=1}^{M} g_{j,m}\left(\boldsymbol{X} \mid \mathcal{T}_{jm}, \mathcal{M}_{jm}\right), \quad \text{for } j = 1, \ldots, n. \tag{6}$$

where m is the number of trees in the ensemble. The general idea is to aggregate individuals "weak learners" into a unified "strong learner". Each regression tree $g_{j,m}(\cdot)$ defines a piecewise constant function based on the arrangement of split rules $\mathcal{T}_{jm}$, associated with a b_{jm}-dimensional vector $\mathcal{M}_{jm} = \left(\mu_{m,1}^{(j)}, \ldots, \mu_{m,b_{jm}}^{(j)}\right)'$ of terminal nodes coefficients where b_{jm} is the number of leaves per tree m in equation j.

There is established evidence in the literature that regression trees have demostrated strong empirical performance in a wide variety of contexts, including supervised learning (Chen and Guestrin (2016); Ke et al. (2017)), casual inference (Hahn et al. (2020)), density regression (Orlandi et al. (2021)). For more discussion see Grinsztajn et al. (2022) and Hill et al. (2020).

The model prior structure follows Chipman et al. (2010). Conditioned on the model hyperparameters, BART priors are :

$$p\left((\mathcal{T}_{j,1}, \mathcal{M}_{,j1}), \ldots, (\mathcal{T}_{j,M}, \mathcal{M}_{j,M}), {\sigma_t}^2\right) = p\left(\sigma_t^2\right) \prod_{m=1}^{M} p_{\mathcal{T}}(\mathcal{T}_{jk}) p_{\mathcal{M}}(\mathcal{M}_{jk} \mid \mathcal{T}_{jk}). \tag{7}$$

The prior distribution for the trees $p_{\mathcal{T}}$ consists of two components. First a prior on the shape of the tree $\mathcal{T}$ and second a prior on the splitting rules $[x_q \leq C_q]$ for each branch node of the tree. The prior on the tree structure includes a probability that a node of depth d is "terminal" or does not split. Starting from a root, the node will split with probability probability $\gamma(1+d)^{-\beta}$. If that is not the case, the root is a terminal node. This iterates until all nodes at certain depth are terminal. We follow the convention introduced by Chipman et al. (2010) by taking $\gamma = 0.95$ and $\beta = 0.2$.

If a node is split, the split rule, defined by a variable and a cutpoint, is sampled as follows. A variable index $q \in 1, \ldots, k$ is sampled according to $q \sim Categorical\left(\boldsymbol{s}\right)$, such that $\boldsymbol{s}$ is probability vector where $s_q = \frac{1}{k}$. Subsequently, a cutpoint C_q is sampled uniformly on the observed range of values of variable $x_{q,t}$ at the current node. Finally, for each terminal node b_{jm} in the tree, we draw a mean parameter $\mu_{m,b_{jm}}^{(j)} \sim \mathcal{N}\left(0, \sigma_\mu^2/M\right)$.

3 Minnesota BART

Recent advancements in the non-parametric Vector Autoregression (VAR) literature, particularly the integration of Bayesian Additive Regression Trees (BART) to extend the traditional VAR framework, provide compelling evidence of improved forecasting performance, as demonstrated by Huber and Rossini (2022). However, these recent developments do not adequately address scenarios in which the true data-generating process (DGP) is sparse. Additionally, the original BART prior used in Huber and Rossini (2022) does not account for the temporal dependency structure present in macroeconomic data. Our contributions to the literature involve addressing these shortcomings within the vector autoregression framework.

A well-documented finding in macroeconomic forecasting—from univariate models for GDP, inflation, and interest rates to multivariate formulations is that simple random-walk or near-unit-root forecasts often perform reasonably well, particularly over short-to medium-term horizons, as initially shown by Nelson and Plosser (1982). This persistence aligns with the near-unit-root behavior frequently observed in economic time series.

Therefore, constructing a prior that embeds a random-walk assumption—where, in the absence of data, the best initial estimate is that today's value equals yesterday's value plus some small drift or shock—provides a natural way to incorporate this empirical regularity into the model.

The Minnesota prior, introduced by Litterman (1980) and Doan et al. (1984), is a shrinkage prior specifically designed to mitigate the issue of overparameterization commonly encountered in large linear VAR models. The design for this prior is based on empirical evidence on macroeconomic time-series behavior, which suggests that variables often exhibit persistent dynamics that can be effectively captured through structured shrinkage. It integrates several reasonable assumptions, including cross-variable shrinkage, where coefficients for lags of different variables are reduced more significantly than those for the variable's own lags. Additionally, it reflects the belief that higher-order lags contribute less to forecasting accuracy. For further details on the Minnesota prior, see Kadiyala and Karlsson (1997), Karlsson (2013), and Chan (2020).

While the traditional BART framework has shown promise in time series analysis, It employs a uniform prior on splitting variables when sampling split rules—an assumption that is often unrealistic in multivariate forecasting settings. To address this limitation, we leverage the DART prior framework, presented by Linero (2018), to incorporate time dependency into our trees, drawing on insights from prior specification in the linear Vector Autoregression (VAR) literature.

By integrating the principles of the Minnesota prior into this framework, our ap-

proach introduces structured shrinkage that respects the temporal dependence inherent in macroeconomic data. This allows the estimation of large dynamic systems within a multivariate BART framework while preserving interpretability and forecasting accuracy.

For equation n, the prior for the splits probability is defined:

$$(s_{1n}, \ldots, s_{kn}) \sim \text{Dirichlet}(\phi_{1n}, \ldots, \phi_{kn}) \tag{8}$$

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

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

where $\rho = \frac{\sigma_i^2}{\sigma_j^2}$ represents the variance ratio of an AR(p) process for each variable. This formulation induces a smooth shrinkage effect on the prior probabilities of splits. The choice of λ_j, where $j = 1, 2$, determines the rate at which these probabilities decay. For our analysis, we set the hyperparameters to fixed values, specifically $\lambda_1 = 1$ and $\lambda_2 = 0.5$. While λ_1 and λ_2 can be estimated directly from the data, we choose to fix them to ensure analytical tractability and simplify the estimation process. A detailed discussion on prior elicitation is provided in Section 6.

However, this configuration does not lead to a sparse solution. It is possible to elicit a prior that explicitly favors a more parsimonious model with a smaller number of covariates. Conditional on the tree topology, and for a fixed λ, substituting l^2 with k, we obtain:

$$(s_{1n}, \ldots, s_{kn}) \sim \text{Dirichlet}\left(\frac{\lambda}{k}, \ldots, \frac{\lambda}{k}\right)$$

The parameter λ governs the degree of sparsity introduced in the tree function. As demonstrated by Linero (2018), when both the number of predictors (k) and the number of branches (B) in the ensemble are large, the prior distribution of the number of relevant predictors ($Q - 1$) can be approximated by a Poisson distribution with parameter θ_B, where $\theta_B = \lambda \sum_{i=0}^{B-1} (\lambda + i)^{-1}$. The level of sparsity can be controlled by setting λ to a predetermined value. Under a fully Bayesian parameter selection framework, we follow the suggested approach of assigning λ a hyperprior, specifically $\lambda/(\lambda + k) \sim \text{Beta}(0.5, 1)$.

Figure 1 illustrates how the value of λ corresponds to different levels of sparsity. This approach may introduce some rigidity, but it aligns more closely with our objective of transparently examining the trade-off between sparsity across variables and to a variable's own lags in tree models. In Section 6, we discuss in detail how λ affect an out-of-sample

exercise.

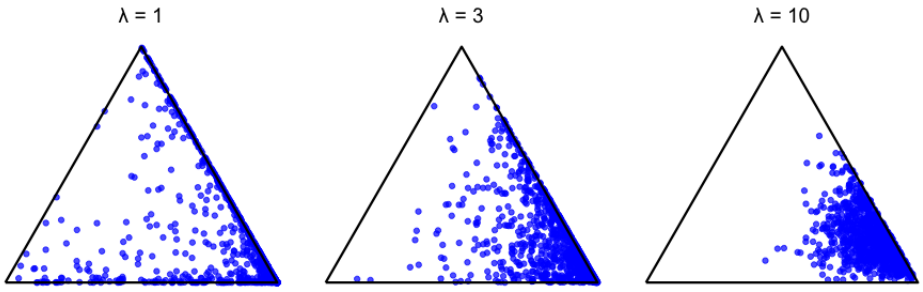


Figure 1: Draws from *Dirichlet* $\left(\lambda, \frac{\lambda}{4}, \frac{\lambda}{9}\right)$. This figure illustrates the effect of varying λ on the concentration parameters of the *Dirichlet* prior on the simplex for $\lambda = (1, 3, 10)$. The vertices of the simplex correspond to one-sparse probability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

4 Posterior Sampling Algorithm

The model is estimated using a combination of traditional Bayesian inference techniques commonly employed in the VAR and BART literature. Tree sampling updates are performed using a Metropolis-Hastings (MH) algorithm, as proposed by Chipman et al. (2010), while most of the remaining steps leverage closed-form Gibbs Update. The conditional posteriors for the factor loadings and factors follow well-known Gaussian distributions. For the stochastic volatility components, present in both the factors and idiosyncratic innovations, we employ the efficient sampler outlined in Kastner and Frühwirth-Schnatter (2014). In cases where a homoskedasticity assumption is applied, the model uses the traditional inverse-gamma prior.

As mentioned previously, conditioning on the covariance structure we can estimate the VAR equation-by-equation. The model can be rewritten the as a system of n independent equations. Let $\boldsymbol{Y}_{\bullet j}$ denote the j-th column of the matrix Y as defined in (5). Therefore we can write our dynamic system as :

$$\boldsymbol{Y}_{\bullet j} = \boldsymbol{G}_j\left(\boldsymbol{X}\right) + \boldsymbol{F}\boldsymbol{\Lambda}_{\bullet j}' + \boldsymbol{\eta}_{\bullet j} \tag{9}$$

where $\boldsymbol{\Lambda}_{\bullet j}'$ is the j-th column of the factor loading matrix. This formulation reveals that our model is a generalized additive model, where the forest component approximates the relationship of $\boldsymbol{y}_t$ with its lags, while a shared linear component across all equations captures the relationships among the variables.

4.1 Sampling the Tree Structure

The two main departures of the sampling strategy proposed by Chipman et al. (2010) is regarding the partial residuals definition, that needs to take into account the factors and loadings structure and an additional step to update the vector of split probabilities $\boldsymbol{s}$. To sample the trees using Bayesian backfitting as in the likelihood function depends of $(\mathcal{T}_{jm}, \mathcal{M}_{jm})$ through the partial residuals that should be defined for our case as:

$$\boldsymbol{R}_{jm} \equiv \boldsymbol{Y}_{\bullet j} - \boldsymbol{F}\boldsymbol{\Lambda}_{\bullet j}' - \sum_{m \neq m*}^{M} g_{jm}\left(\boldsymbol{X}|\mathcal{T}_{jm}, \mathcal{M}_{jm}\right) \tag{10}$$

Therefore we can sample tree structure marginalizing over $\mathcal{M}_{jm}$, such that:

$$p\left(\mathcal{T}_{jm}|\boldsymbol{R}_{jm}, \sigma_{j,t}\right) \propto p\left(\mathcal{T}_{jm}\right) \int p\left(\boldsymbol{R}_{jm}|\mathcal{M}_{jm}, \mathcal{T}_{jm}, \sigma_{j,t}\right) p\left(\mathcal{M}_{jm}|\mathcal{T}_{jm}, \sigma_{j,t}\right) d\mathcal{M}_{jm} \tag{11}$$

can be obtained in a closed form solution up to a constant. Allowing to carry out each draw from $(\mathcal{T}_{jm}, \mathcal{M}_{jm}|\boldsymbol{R}_{jm}, \sigma_{j,t})$ sequentially.

To draw the probability split vector $\boldsymbol{s}$, we follow Linero (2018) and leverage the conjugacy between the Dirichlet prior and multinomial sampling, enabling a full-conditional Gibbs update given by:

$$\left(s_1, \ldots, s_k\right) \sim \text{Dirichlet}\left(\phi + m_1, \ldots, \phi + m_k\right), \tag{12}$$

where ϕ represents the shape parameter of the Dirichlet and Minnesota specification, and m_k denotes the number of splitting rules for predictor k in the ensemble.

4.2 Sampling the Loadings and Factors

The factor loadings $\mathbf{\Lambda}$ are drawn from a full conditional distribution that follows a Gaussian distribution in a standard form. For each row of $\mathbf{\Lambda}$, denoted as $\mathbf{\Lambda}_i$, we sample as follows:

$$\mathbf{\Lambda}_i|\bullet \sim \mathcal{N}(\bar{\boldsymbol{L}}_i, \bar{\boldsymbol{W}}_i), \tag{13}$$

$$\bar{\boldsymbol{W}}_i = \left(\boldsymbol{F}_i'\boldsymbol{F}_i + \boldsymbol{W}_i^{-1}\right)^{-1},$$

$$\bar{\boldsymbol{L}}_i = \bar{\boldsymbol{W}}_i \left(\boldsymbol{F}_i'\tilde{y}_i\right).$$

Here, $\boldsymbol{F}_i$ is the t-th row $\boldsymbol{f}_t/e^{h_i t/2}$, and the t-th element of $\tilde{y}_i$ is given by $\left(y_{it} - f_i(x_t)\right)/e^{h_i t/2}$. The matrix $\boldsymbol{W}_i$ is a prior variance-covariance matrix of dimension $(nr) \times (nr)$. Since the number of factors are determined by an upper bound, we sample from a horseshoe prior using the auxiliary sampler proposed in Makalic and Schmidt (2015) for each column of $\mathbf{\Lambda}$. The factors are generated on a t-by-t basis using Gaussian distributions as in Aguilar and West (2000).

5 Forecasting Macroeconomic Variables

We conduct a forecasting exercise using multivariate Bayesian additive regression trees to compare our proposed sparse and Minnesota-BART priors with the baseline BART prior structure from Huber and Rossini (2022)

5.1 Data

We use a dataset consisting of 22 U.S. quarterly variables covering the period from 1965Q1 to 2019Q4. The data is sourced from the FRED-QD database at the Federal Reserve Bank of St. Louis, as described in McCracken and Ng (2016). The dataset includes a range of standard macroeconomic and financial variables, such as real GDP, industrial production, inflation rates, labor market indicators, and interest rates. These variables are transformed to achieve stationarity, typically by computing growth rates. We will include 13 lags of the endogenous variables in our model. A detailed description of the variables and their transformations can be found in Appendix A.

5.2 Predictive Distribution

We begin by presenting predictive evidence of the effectiveness of our proposed priors and their impact on predictive accuracy. To achieve this, we construct a recursive forecasting design, using 1965Q1 to 2004Q4 as the initial training period. We employ an expanding window strategy: after performing h-step-ahead forecasts, we incorporate the next observation into the dataset and re-estimate the model, obtaining a new draw from the predictive density with the updated information. This process is repeated iteratively until all available data has been utilized. The one-step-ahead predictive distribution is given by

$$p\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}^t\right) = \int p\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}^t, \boldsymbol{\vartheta}\right) p\left(\boldsymbol{\vartheta}|\boldsymbol{y}^t\right) d\boldsymbol{\vartheta}, \tag{14}$$

where $\boldsymbol{y}^t$ represents the historical time series up to time t, i.e., $\boldsymbol{y}^t = (y_1, \ldots, y_t)$. The parameter $\boldsymbol{\vartheta}$ encapsulates all unknown parameters of the model. This integral is solved using the standard Monte Carlo approach:

$$p\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}^t\right) \approx \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{y}_{t+1}|\boldsymbol{y}^t, \boldsymbol{\vartheta}^{(m)}),$$

where the one-step-ahead predictive density $p\left(\boldsymbol{y}_{t+1}|\boldsymbol{y}^t, \boldsymbol{\vartheta}\right)$ is Gaussian, conditional on knowing $\boldsymbol{\vartheta}$. Therefore, for each draw $\boldsymbol{\vartheta}^{(m)}$ from the posterior distribution $p\left(\boldsymbol{\vartheta}|\boldsymbol{y}^t\right)$, we obtain the predictive density:

$$\boldsymbol{y}_{t+1}|\boldsymbol{y}^t, \boldsymbol{\vartheta}^{(m)} \sim \mathcal{N}\left(\boldsymbol{F}^{(m)}\left(\boldsymbol{X}_{t+1}\right), \Sigma_{t+1}^{(m)}\right),$$

where $\boldsymbol{F}^{(m)}(\cdot)$ is generated by our tree sampling algorithm, and $\Sigma_{t+1}^{(m)}$ is drawn using the covariance structure specified in Eq. 2. For the stochastic volatility (SV) specification, as outlined in Eq. 4, the forecasts of Ω_{t+1} and H_{t+1} are obtained as follows. Given the posterior draws of $h_{it}^{(m)}$, we simulate $h_{it+1}^{(m)}$ from a conditional normal distribution with mean $\mu_i^{(m)} + \phi_i^{(m)}(h_{i,t}^{(m)} - \mu_i^{(m)})$ and variance $\sigma_h^{2(m)}$. Higher-order forecasts are computed recursively. In the following section, note that $\boldsymbol{y}_{t+1}$ corresponds to a one-quarter-ahead prediction, i.e., three months. We use the Gibbs sampler described in the section 4 to obtain 5,000 posterior, after a burn-in period of 30,000.

5.3 A point forecast comparison of the different priors

We use a standard small-scale Minnesota BVAR with stochastic volatility (SV) as a benchmark to evaluate our model's performance across three key variables: real GDP growth (GDPC1), inflation (CPIAUCSL), and the Federal Funds Rate (FEDFUNDS). To facilitate this comparison, we compute the (relative) root mean square prediction error (RMSPE). A value below one indicates that the model outperforms the benchmark, while a value above one suggests weaker performance. Importantly, we account for heteroskedasticity in each prior specification. A priori, we expect that the combination of our proposed priors and time variation in volatilities will enhance point forecasts. The results are summarized in Figure 2.

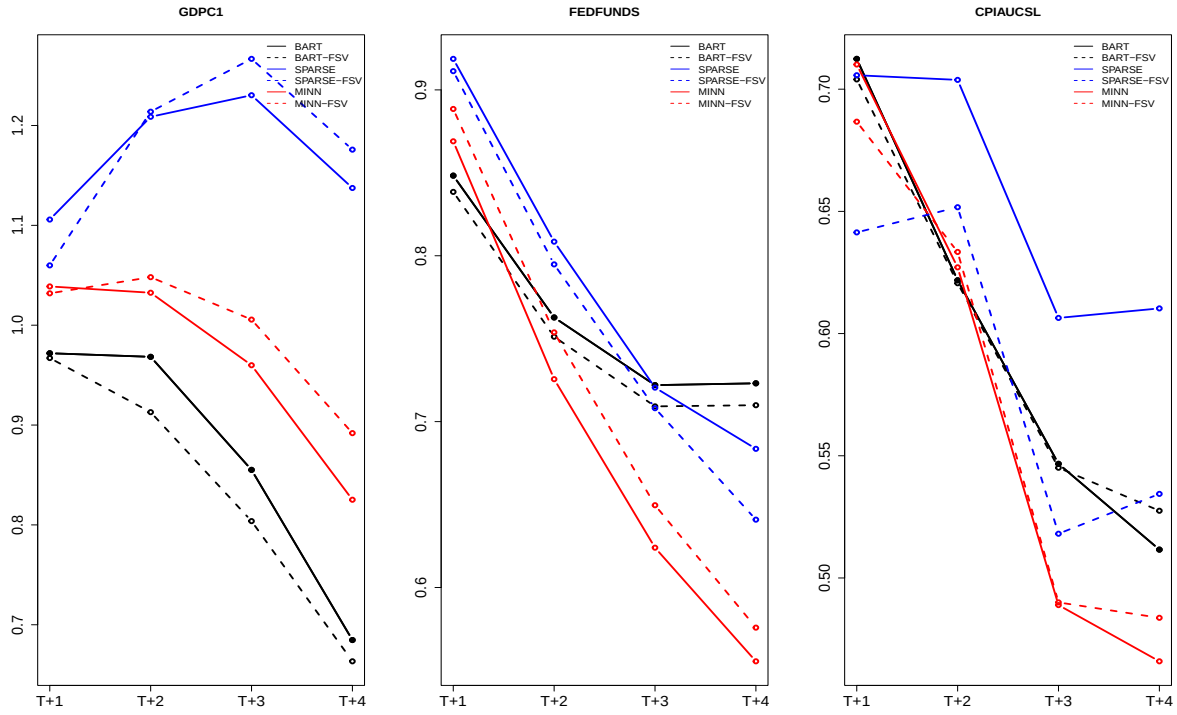


Figure 2: Point Forecast Comparison. This figure reports the Relative RMSE of the variables of interest compared to the baseline BVAR-SV, using the Minnesota-BART prior, the Sparse Prior, and the BART prior. A value below one indicates that the model outperforms the benchmark, while a value above one suggests weaker performance. 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.

Our findings reveal a consistent pattern: predictive accuracy improves as the forecasting horizon extends. Compared to the linear baseline, incorporating a richer information set and a non-linear approach enhances point forecasts. The SV variants generally improve

precision relative to their homoskedastic counterparts. However, the numerical differences are often small, particularly when comparing sparse and non-sparse versions of the prior. Importantly, for inflation forecasting, the introduction of stochastic volatility (SV) leads to a more pronounced improvement in predictive accuracy for the sparse specification compared to its non-SV counterpart. Regarding prior choice, with the exception of real GDP growth, a smoother shrinkage approach in the variable splits tends to outperform its sparse alternative for the variables of interest. This suggests that, for certain variables, a prior that imposes gradual shrinkage—leveraging more information compared to a sparse prior—is preferable.

5.4 Comparing the priors through log predictive density scores

Although point forecast exercises are important, they provide only a partial assessment of model performance. To obtain a more comprehensive evaluation, it is essential to consider additional metrics that account for the model's ability to predict higher-order moments of the predictive distribution for the variable of interest. Therefore, is necessary to utilize a metric to access the accuracy of density forecasts. As discussed by Geweke and Amisano (2010), log predictive density scores (LPDS) are often used to compare different models. In this paper, we use the LPDS as a metric to evaluate and compare the performance of various BART prior specifications. We will make slightly change of notation when presenting this calculations. The first t_0 time series observations, $\boldsymbol{y}^{tr} = (y_1, \ldots, y_{t_0})$, 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:

$$\text{LPDS} = \log p(y_{t_0+1}, \ldots, y_T \mid \boldsymbol{y}^{tr}) = \sum_{t=t_0+1}^{T} \log p(\boldsymbol{y}_t \mid \boldsymbol{y}^{t-1}) \tag{15}$$

In Equation (15), $p(\boldsymbol{y}_t \mid \boldsymbol{y}^{t-1})$ represents the one-step-ahead predictive density for time t. This density is evaluated at the observed value $\boldsymbol{y}_t$. Note that this framework not only works when evaluating the joint performance for higher order moments of the predictive distribution for the multivariate model, but also the marginal density scores, i.e, we have the LPDS for the three variables of interest as well.

Figure 3 shows that while our prior does not perform as well in point forecasting for GDP, it excels in predictive distribution forecasting. The Minnesota prior achieves the best performance, with its FSV specification closely following. Notably, this prior alternative not only outperforms the basic BART prior but also surpasses its simpler linear counterpart.

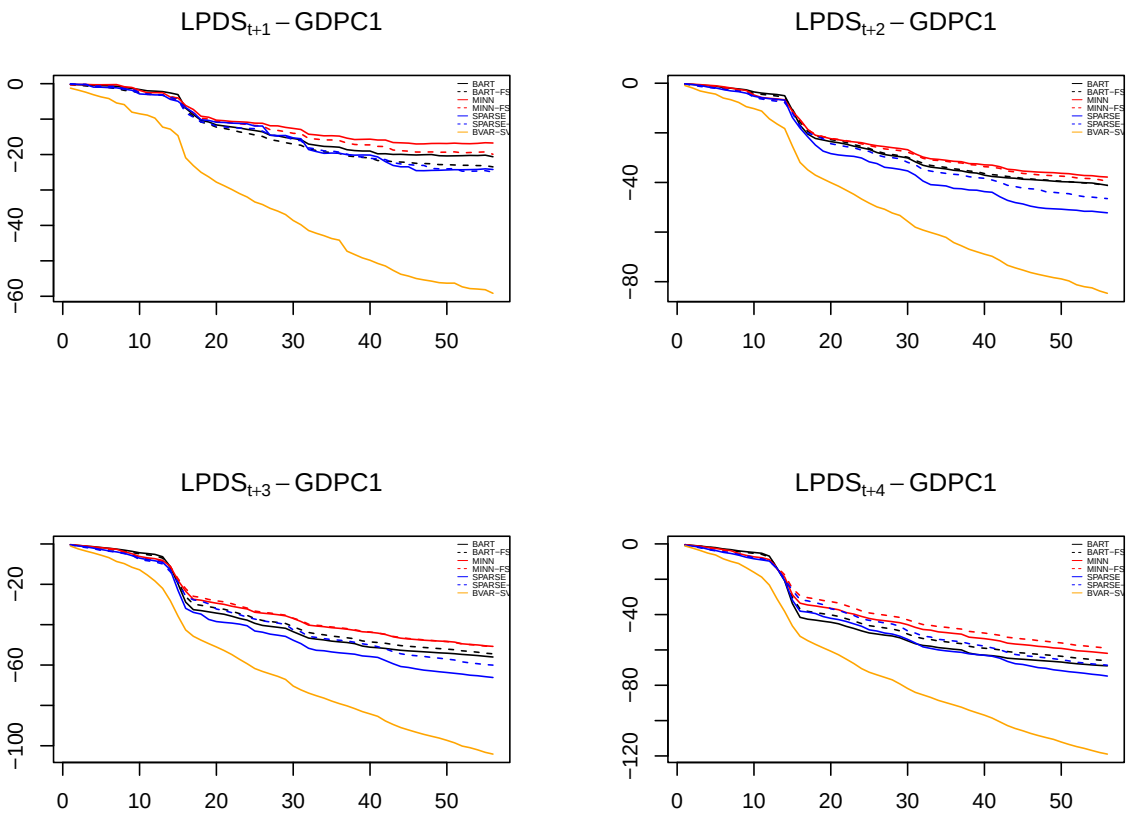


Figure 3: Marginal Log Predictive Density Score. This figure reports the Marginal Log Predictive Density Score (LPDS) for GDPC1. Cumulative Marginal log predictive scores for the last 56 time point (labeled with time index $T - t_0$, where $t_0 = 160$). Each probability split prior specification for 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.

For the FEDFUNDS variable, the results follow a different pattern. As shown in Figure 4, the more sparse specification underperforms in point forecasting when compared to the linear benchmark. However, it dominates across all horizons in predictive distribution forecasting. The smooth shrinkage prior emerges as the second-best choice.

For the CPI variable, the results favor the baseline prior with an SV correction, except for the first forecasting horizon, as shown in Figure 5. Even in this case, our smooth shrinkage option remains a close second in forecasting the predictive density distribution.

This results suggest that, for both point forecasts and marginal density, the alternative prior structures proposed in this paper are competitive and, in some cases, even outperform the baseline BART model. Figure 6 shows that when evaluating the (joint)

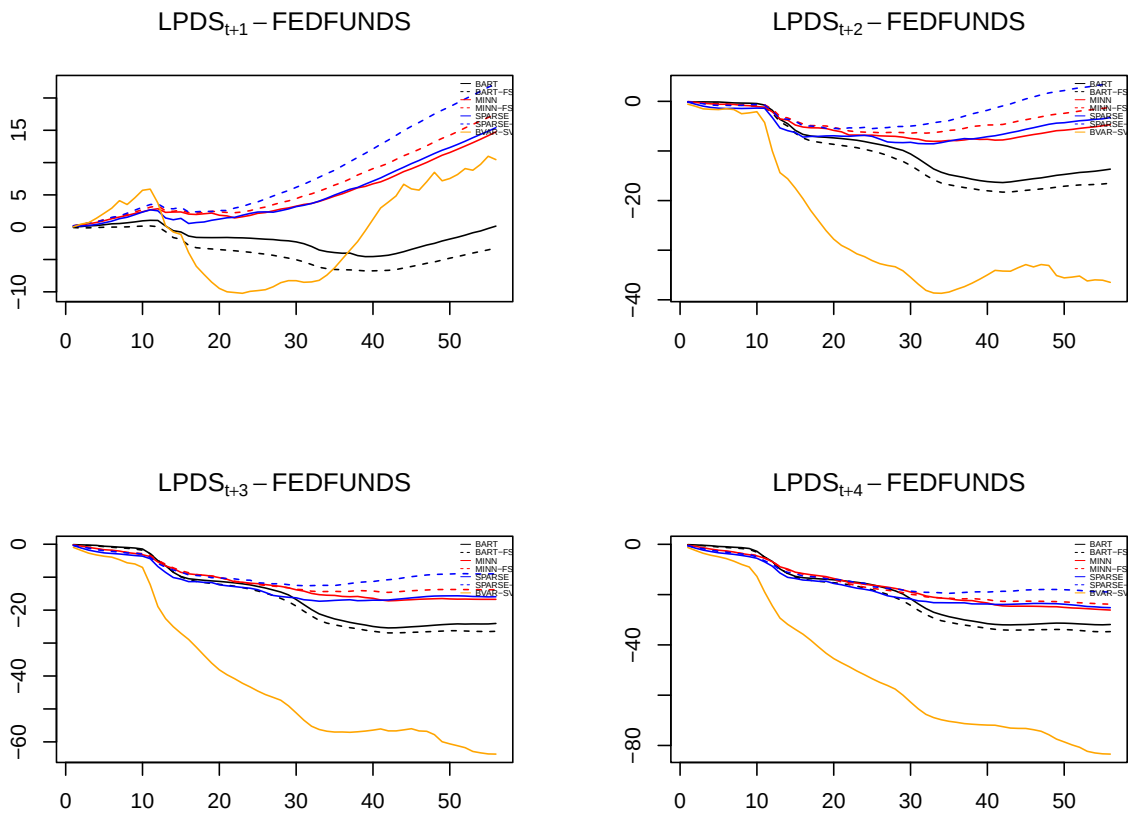


Figure 4: Marginal Log Predictive Density Score. This figure reports Marginal Log Predictive Density Score (LPDS) for FEDFUNDS. Cumulative Marginal log predictive scores for the last 56 time point (labeled with time index $T - t_0$, where $t_0 = 160$). Each probability split prior specification for the 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.

predictive density, our alternatives, which incorporate varying levels of shrinkage, consistently outperform the basic prior, particularly when heteroskedasticity is accounted for, as initially hypothesized. Notably, the Minnesota specification outperforms the sparse alternative across all time horizons.

5.5 In sample features of our model

Our prior modifications introduce key advantages to multivariate BART analysis. In the baseline BART framework, variable importance is typically assessed by counting the number of times each feature appears in a splitting rule. However, this approach has limitations. Since the traditional prior imposes a uniform split probability across all features,

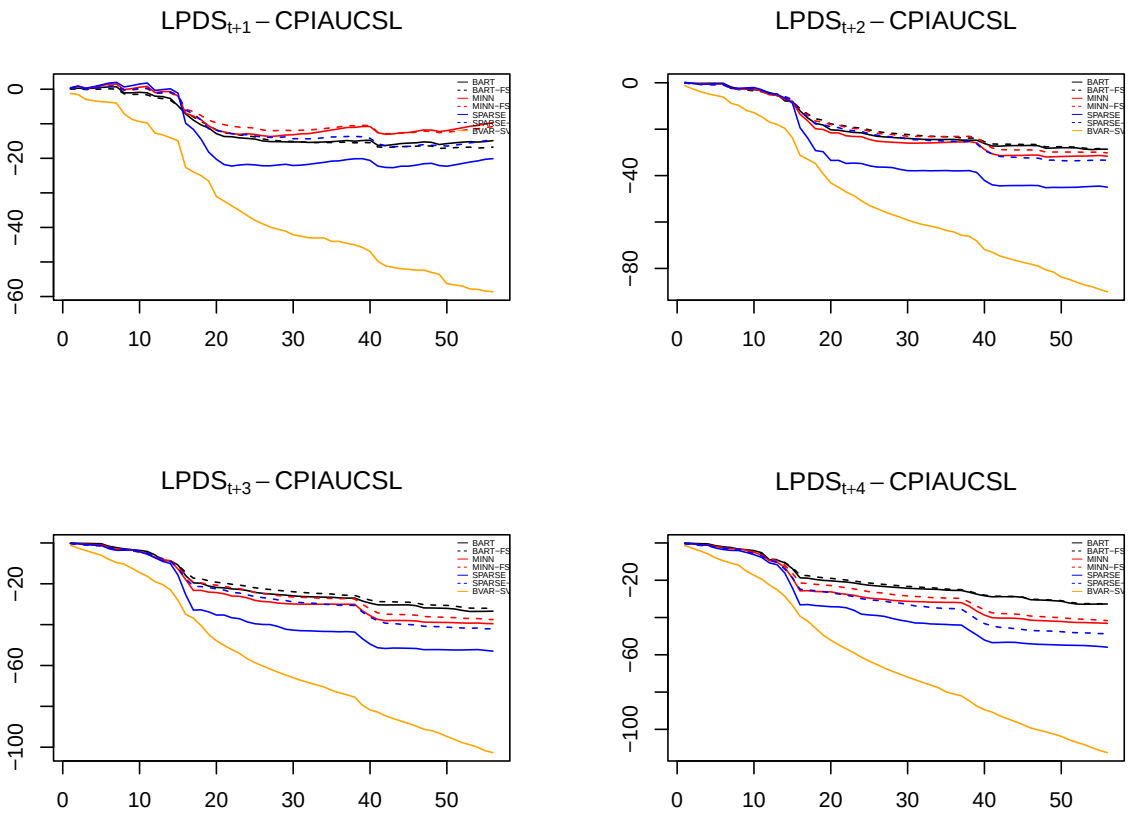


Figure 5: Marginal Log Predictive Density Score. This figure reports the Marginal Log Predictive Density Score (LPDS) for CPIAUCSL. Cumulative Marginal log predictive scores for the last 56 time point (labeled with time index $T - t_0$, where $t_0 = 160$). Each probability split prior specification for 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.

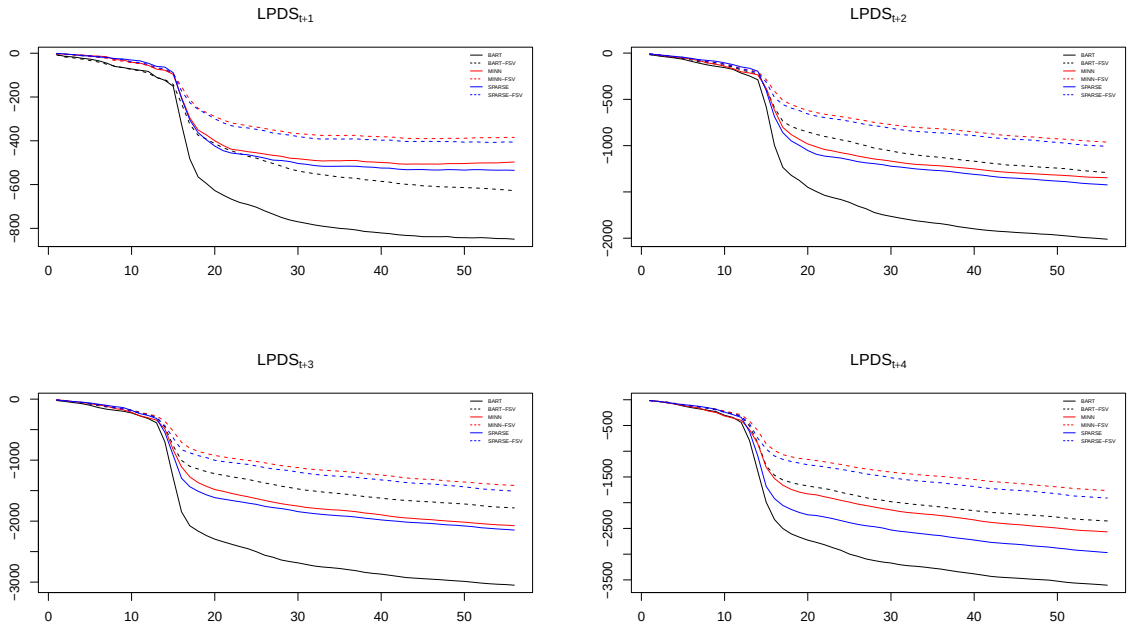


Figure 6: Cumulative Log Predictive Density Scores. Cumulative log predictive scores for the last 56 time point (labeled with time index $T - t_0$, where $t_0 = 160$). Each probability split prior specification for the mean function is shown under both the homoskedastic and stochastic volatility (SV) settings, where the former is represented by a continuous line and the latter by a dashed line.

variable importance can only be inferred if there is a trade-off between the number of trees and interpretability. In other words, improving interpretability often comes at the expense of predictive performance Chipman et al. (2010); Bleich and Kapelner (2014); Linero (2018). This limitation does not apply to our proposed prior structures. To illustrate, Figure 7 presents the Posterior Inclusion Probability (PIP) for each prior choice. The PIP is defined as:

$$\text{PIP}_j = P(\text{predictor } j \text{ appears in the ensemble}|\text{Data})$$

We present the PIP for the in-sample results of our model before applying the expanding window, focusing on the Inflation variable. As shown in Figure 7, the BART prior specification distributes splits relatively evenly across all features in the model. In contrast, the sparse prior specification shrinks the split probability of most variables to near zero while preserving the variable's own first lag as the most significant predictor. This aligns with the Bayesian VAR literature, which finds that the AR(1) term explains the largest share of variation in the variable of interest. For the Minnesota prior, the expected decay in PIP follows a structured lag hierarchy, preserving the anticipated importance ordering among the split variables.

6 Prior Elicitation

As previously discussed, the choice of λ is of critical importance, as it plays a central role in determining the expected level of shrinkage in the model. One approach we have presented is to select λ based on a shrinkage target informed by the econometric literature and subject-matter considerations. To further investigate its impact, we analyze different levels of λ, specifically considering a grid of values: $\lambda_1 = \{1, 3, 5, 10, 20\}$ and $\lambda_2 = \{0.5, 1, 1.5, 2.5, 5, 10\}$. We then examine its effects on the log-predictive density score in comparison to the standard BART prior.

First, we examine how the choice of the smooth shrinkage parameter influences the "hyperbolic" shape of the posterior inclusion probability (PIP) for the inflation variable in the in-sample analysis. As shown in Figure 8, an increase in λ values results in a slower decay of PIP, reflecting that the shrinkage of both lags and cross-lags does not reach zero as quickly as with our reference values of $\lambda_1 = 1$ and $\lambda_2 = 0.5$.

Additionally, when comparing different PIPs for CPI's own lag, we observe the influence of the hyperparameter choice on the rate at which the inclusion probability decays, as illustrated in Figure 9. To compare different prior configurations we will use the log predictive density score, calculated as how was presented in the sections before. Figure

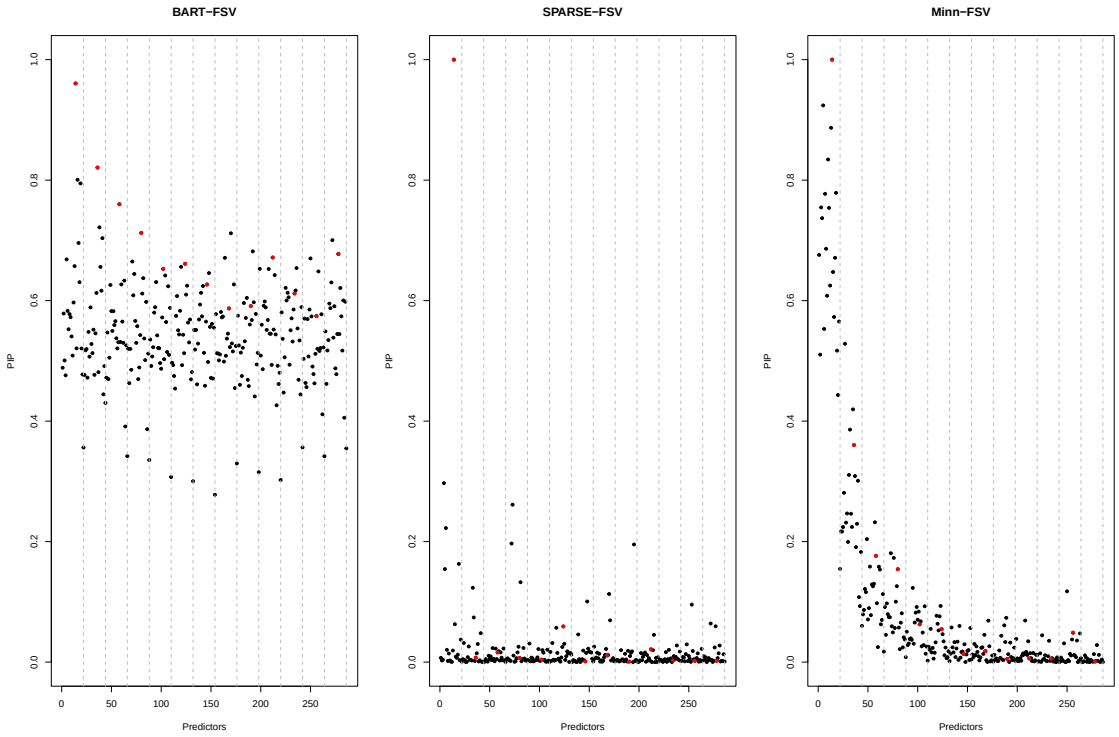


Figure 7: Posterior Inclusion Probability. In-sample Posterior Inclusion Probability (PIP) results for the CPI variable, before expanding window exercise. The dashed vertical lines indicate the bin boundaries corresponding to different lag orders. The highlighted red dots represent the variable's own lags.

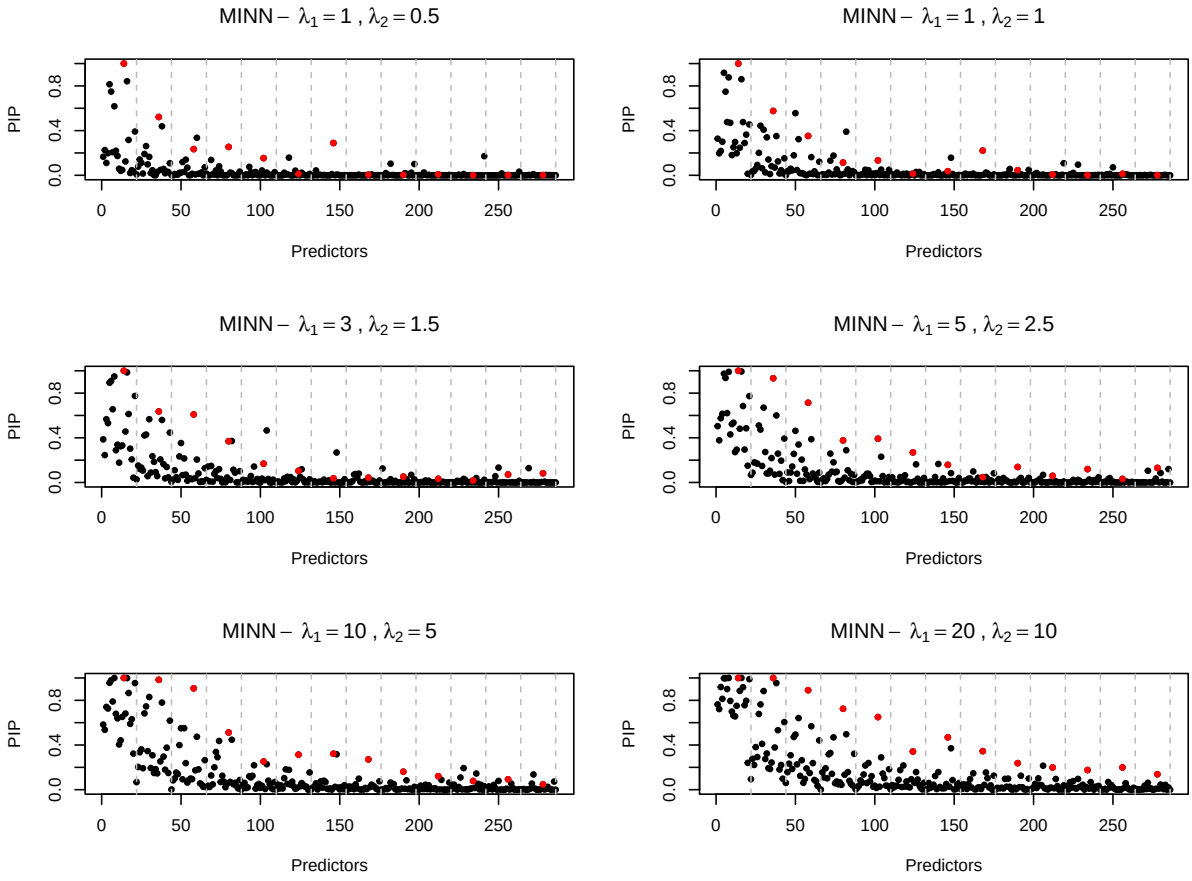


Figure 8: Posterior Inclusion Probability for different shrinkage parameters. In-sample Posterior Inclusion Probability (PIP) results for the CPI variable for each different λ_i combination. The dashed vertical lines indicate the bin boundaries corresponding to different lag orders. The highlighted red dots represent the variable's own lags.

10 shows that besides any configuration that you choose for the prior, it is clear that for accurate density forecasts, its necessary to take into account time dependency when constructing the prior.

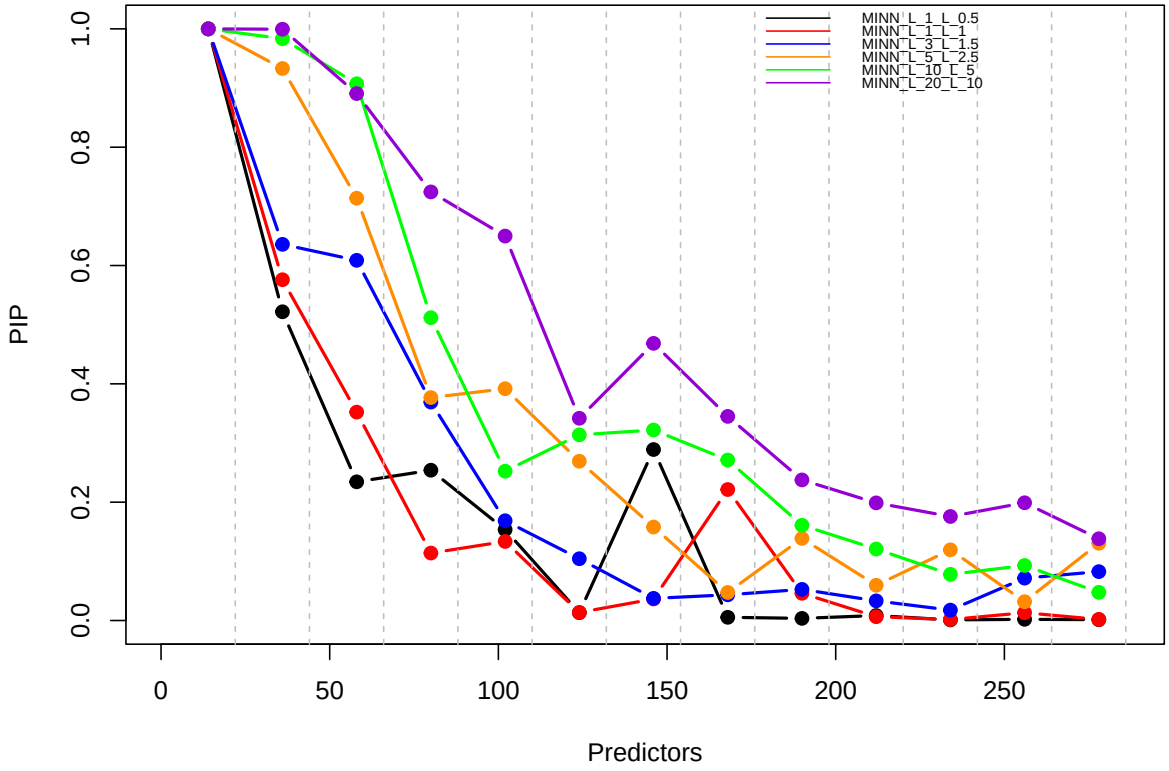


Figure 9: Own-Lag Posterior Inclusion Probability. In-sample Posterior Inclusion Probability (PIP) for the CPI's own lag across different grid values of $\lambda_1 = \{1, 3, 5, 10, 20\}$ and $\lambda_2 = \{0.5, 1, 1.5, 2.5, 5, 10\}$.

Overall, our analysis highlights the crucial role of λ in shaping the degree of shrinkage in the model and its impact on both variable selection and predictive performance. The results demonstrate that higher values of λ lead to a more gradual decay in posterior inclusion probabilities, preserving the influence of lags and cross-lags for a longer range. These findings reinforce the importance of carefully selecting the shrinkage parameter, as it directly affects model interpretability and forecasting accuracy.

7 Concluding Remarks and Future Research

The classic BART prior for split rules, which samples split variables with uniform probability, implicitly assumes that the mean function is not sparse or time-dependent in

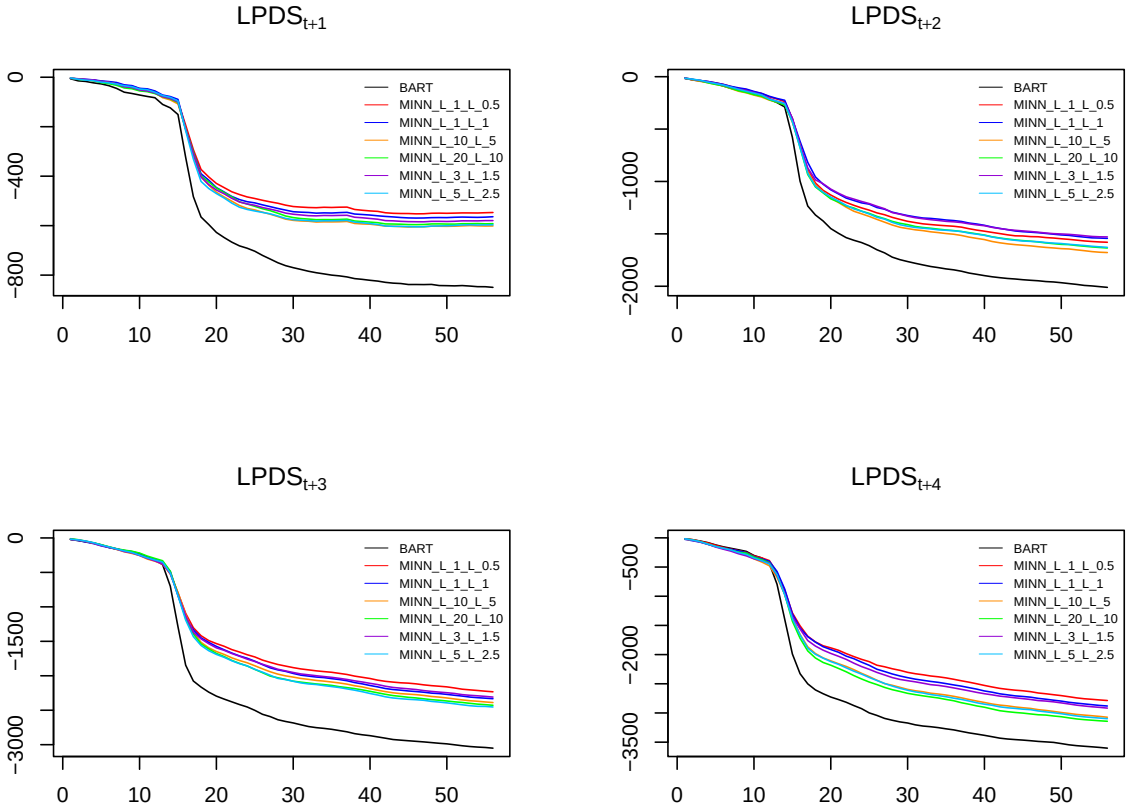


Figure 10: Log Predictive Density Score for different shrinkage values. Cumulative log predictive scores for the last 56 time points (labeled with time index $T - t_0$, where $t_0 = 160$), across different grid values of $\lambda_1 = \{1, 3, 5, 10, 20\}$ and $\lambda_2 = \{0.5, 1, 1.5, 2.5, 5, 10\}$.

its input features, limiting inference on variable importance and hindering effective high-dimensional analysis. This paper introduces a framework that integrates insights from the literature on sparse priors for BART and Bayesian VARs. The proposed model allows for shrinkage in split probabilities, enabling the estimation of large dynamic systems within a multivariate BART framework, unlike existing approaches. Additionally, we demonstrate that incorporating a prior akin to the linear Minnesota prior introduces smooth shrinkage and time dependence information during prior elicitation.

Our approach was illustrated using a large U.S. dataset, where we showed that the proposed priors yield substantial improvements in forecast accuracy, particularly for higher-order moments. In this context, the Minnesota specification appears to be especially effective in extracting forecasting gains by introducing a smoother shrinkage approach compared to the sparse alternative. While point forecast accuracy varies across the variables of interest, nonlinear models often outperform linear specifications, their gains are substantial when they do, whereas losses to linear models are relatively small. Moreover, our prior structure demonstrates predictive gains for key macroeconomic variables, such as the Federal Funds Rate and inflation, highlighting its practical relevance for economic forecasting.

Although our model focuses on a reduced-form specification for forecasting, the framework can also be adapted for structural analysis using techniques such as Generalized Impulse Response Functions (GIRFs) (Koop et al. (1996)) or Local Projection (LP) estimation (Jordà (2005)). Future work could explore alternative sampling methods or algorithmic optimizations to enhance scalability and reduce computational costs. Furthermore, expanding this approach to account for richer structural dynamics, such as incorporating time-varying parameters or state-dependent effects, could significantly enhance its applicability in macroeconomic modeling.

References

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357.

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

Bleich, J. and Kapelner, A. (2014). Bayesian additive regression trees with parametric models of heteroskedasticity. *arXiv preprint arXiv:1402.5397*.

Bolfarine, H., Carvalho, C. M., Lopes, H. F., and Murray, J. S. (2024). Decoupling shrinkage and selection in gaussian linear factor analysis. *Bayesian Analysis*, 19(1):181–203.

Carriero, A., Clark, T. E., and Marcellino, M. (2016). Common drifting volatility in large bayesian vars. *Journal of Business & Economic Statistics*, 34(3):375–390.

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

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Chan, J. C. (2020). *Large Bayesian vector autoregressions*. Springer.

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

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chib, S., Nardari, F., and Shephard, N. (2006). Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics*, 134(2):341–371.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

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

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100.

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). Sparse bayesian factor analysis when the number of factors is unknown. *Bayesian Analysis*, 1(1):1–44.

Geweke, J. and Amisano, G. (2010). Comparing and evaluating bayesian predictive distributions of asset returns. *International Journal of Forecasting*, 26(2):216–230.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7(1):251–278.

Huber, F., Koop, G., Onorante, L., Pfarrhofer, M., and Schreiner, J. (2023). Nowcasting in a pandemic using non-parametric mixed frequency vars. *Journal of Econometrics*, 232(1):52–69.

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

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American economic review*, 95(1):161–182.

Kadiyala, K. R. and Karlsson, S. (1997). Numerical methods for estimation and inference in bayesian var-models. *Journal of Applied Econometrics*, 12(2):99–132.

Karlsson, S. (2013). Forecasting with bayesian vector autoregression. *Handbook of economic forecasting*, 2:791–897.

Kastner, G. and Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, 76:408–423.

Kastner, G. and Huber, F. (2020). Sparse bayesian vector autoregressions in huge dimensions. *Journal of Forecasting*, 39(7):1142–1165.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Koop, G., Pesaran, M. H., and Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of econometrics*, 74(1):119–147.

Koop, G. M. (2013). Forecasting with medium and large bayesian vars. *Journal of Applied Econometrics*, 28(2):177–203.

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

Litterman, R. B. (1980). *Bayesian procedure for forecasting with vector autoregressions*. Massachusetts Institute of Technology.

Litterman, R. B. (1986). Forecasting with bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38.

Lopes, H. F. and Carvalho, C. M. (2007). Factor stochastic volatility with time varying loadings and markov switching regimes. *Journal of Statistical Planning and Inference*, 137(10):3082–3091.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

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

Nelson, C. R. and Plosser, C. R. (1982). Trends and random walks in macroeconmic time series: some evidence and implications. *Journal of monetary economics*, 10(2):139–162.

Orlandi, V., Murray, J., Linero, A., and Volfovsky, A. (2021). Density regression with bayesian additive regression trees. *arXiv preprint arXiv:2112.12259*.

Pitt, M. K. and Shephard, N. (1999). Time varying covariances: a factor stochastic volatility approach. *Bayesian statistics*, 6:547–570.

A Data description

Table 1: Data description.

Mnemonic	Description	Trans.	VAR-8	VAR-22
GDPC1 (RGDP)	Real Gross Domestic Product	2	x	x
CE16OV (EMP)	Civilian Employment (Thousands of Persons)	2	x	x
AWHMAN (AWH)	Average Weekly Hours of Production and Nonsupervisory Employees: Manufacturing	1	x	x
CPIAUCSL (CPI)	Consumer Price Index for All Urban Consumers: All Items	2	x	x
CES3000000008x (AHE)	Real Average Hourly Earnings of Production and Nonsupervisory Employees: Manufacturing	2	x	x
INDPRO	IP:Total index Industrial Production Index (Index 2012=100)	2	x	x
FEDFUNDS (FFR)	Effective Federal Funds Rate (Percent)	1	x	x
S.P.500 (SP500)	S&P's Common Stock Price Index: Composite	3	x	x
PCECC96	Real Personal Consumption Expenditures	2		x
FPIx	Real private fixed investment	2		x
UNRATE	Civilian Unemployment Rate (Percent)	1		x
CES0600000007	Average Weekly Hours of Production and Nonsupervisory Employees: Goods-Producing	1		x
CLAIMSx	Initial Claims	2		x
HOUST	Housing Starts: Total: New Privately Owned Housing Units Started	2		x
CES0600000008	Average Hourly Earnings of Production and Nonsupervisory Employees:	2		x
PAYEMS	Emp:Nonfarm All Employees: Total nonfarm (Thousands of Persons)	2		x
CUMFNS	Capacity Utilization: Manufacturing (SIC) (Percent of Capacity)	1		x
PERMIT	New Private Housing Units Authorized by Building Permits	2		x
BUSLOANSx	Real Commercial and Industrial Loans, All Commercial Banks	2		x
BAA10YM	Moody's Seasoned Baa Corporate Bond Yield Relative to Yield on 10-Year Treasury	1		x
GS10TB3Mx	10-Year Treasury Constant Maturity Minus 3-Month Treasury Bill, secondary market	1		x
TB3SMFFM	3-Month Treasury Constant Maturity Minus Federal Funds Rate	1		x

Notes: The data used is the quarterly version of the dataset proposed in McCracken and Ng (2016). `Trans` indicates the transformation applied to each variable with (1) implying no transformation, (2) denoting year-on-year growth rates, (3) denoting quarter-on-quarter growth rates, and (4) refers to quarter-on-quarter percentage changes.