Scientific Research Publishing

Applied Mathematics, 2024, 15, 292-312
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Measuring Causal Effect with ARDL-BART: A Macroeconomic Application

Pegah Mahdavi[1*], Mohammad Ali Ehsani[1], Daniel Felix Ahelegbey[2], Mehrnaz Mohammadpour[3]

[1]Department of Economic and Administrative Sciences, University of Mazandaran, Babolsar, Iran
[2]Department of Economics and Management Sciences, University of Pavia, Pavia, Italy
[3]Department of Mathematical Sciences, University of Mazandaran, Babolsar, Iran
Email: *pegahmahdavi74@gmail.com, m.ehsani@umz.ac.ir, danielfelix.ahelegbey@unipv.it, m.mohammadpour@umz.ac.ir

How to cite this paper: Mahdavi, P., Ehsani, M.A., Ahelegbey, D.F. and Mohammadpour, M. (2024) Measuring Causal Effect with ARDL-BART: A Macroeconomic Application. *Applied Mathematics*, **15**, 292-312.
https://doi.org/10.4236/am.2024.154018

Received: March 8, 2024
Accepted: April 27, 2024
Published: April 30, 2024

Abstract

Modeling dynamic systems with linear parametric models usually suffer limitation which affects forecasting performance and policy implications. This paper advances a non-parametric autoregressive distributed lag model that employs a Bayesian additive regression tree (BART). The performance of the BART model is compared with selection models like Lasso, Elastic Net, and Bayesian networks in simulation experiments with linear and non-linear data generating processes (DGP), and on US macroeconomic time series data. The results show that the BART model is quite competitive against the linear parametric methods when the DGP is linear, and outperforms the competing methods when the DGP is non-linear. The empirical results suggest that the BART estimators are generally more efficient than the traditional linear methods when modeling and forecasting macroeconomic time series.

Keywords

BART Model, Non Parametric Modeling, Machine Learning, Regression Trees, Bayesian Network VAR

1. Introduction

Economists have long sought to design models for identifying structural relations among endogenous macroeconomic variables, for predicting/forecasting, and for performing impulse response analysis. Some of the widely used models to achieve the above goals include vector auto-regression models, dynamic factor models, and linear projection models, among others. One of the limitations of these traditional models is the assumption of a linear parametric model. The fundamental problem of these parametric linear models is that they can be very

restrictive, in the sense that, they do not allow for nonlinear interactions among explanatory variables. This can lead to model misspecification and predictive performance of the model.

Many causal methods for observational data are conditional on the treatment and confounding covariates. Bayesian non-parametric modeling algorithm, Bayesian Additive Regression Trees (BART; [1]) with a very flexible function provides us with a strong and simpler model in estimating causal effects [2]. BART is most closely related to boosting in that it combines a large set of relatively simple decision trees to a complex high-dimensional response. Bayesian additive regression trees (BART) provide a framework for flexible non-parametric modeling of relationships of covariates to outcomes. Recently, BART models have been shown to provide excellent predictive performance, for both continuous and binary outcomes, and exceed that of its competitors. BART model has been developed in several areas of knowledge, such as medicine, biology, and genetics, mainly in classification problems (see [3] [4] [5]).

These techniques have been applied in macroeconomic and financial data sets. [6] has extended BART into the classification context by using financial statement information on solvent and insolvent firms, and therefore term the resulting classification technique as the Bayesian Additive Classification Tree (BACT). [7] empirically evaluated the performance of two machine learning models, BART, and random forest, applied to credit scoring. They compared the models? Performance to that of logistic regression and the BART and the random forest was superior to logistic regression in both the balanced sample and the unbalanced sample. [8] has evaluated the real-time forecasting performance for a set of US macroeconomic and financial indicators of the various BART models, using a variety of loss functions and a BVAR-SV model as a (strong) benchmark. BART specifications can deliver more accurate tail forecasts than BVAR-SV, in particular for unemployment.

[9] introduces a flexible local projection (LP; [10]) that generalizes the model to a non-parametric setting by using Bayesian Additive Regression Trees (BART). They apply BART-LP to US fiscal and financial shocks and show that financial shocks have non-linear effects on the economy. (VAR [11]) models assume that the lagged dependent variables in influence the contemporaneous values in a linear fashion. [12] relaxes this assumption by blending the literature on BART models and VARs. BAVART model can handle arbitrary non-linear relations between the endogenous and the exogenous variables. They apply the model to the US term structure of interest rates and show BAVART model yields precise point and density forecasts.

Leveraging BART for dynamic system modeling and forecasting in economics and other fields can lead to more accurate predictions, a better understanding of system dynamics, informed decision-making, and more effective policy development. By harnessing BART's capabilities, policymakers can make more data-driven, evidence-based decisions that contribute to the stability, growth, and resilience of economies and societies

ARDL is like VAR with just one equation and on the ARDL there is no assumption that the errors are correlated and independent. In this case, we are only looking at the explanatory variable's lag. This is like VAR but we only take them equation by equation. The covariance matrix of errors is considered in situations where we seek to identify the structural relationship. However, we are interested in forecasting each variable at the time and ARDL can help us achieve whatever the VAR and also forecasting.

We want to show that in the non-parametric framework, important variables are lower Root-mean-square Error (RMSE) compared to parametric regression coefficients. BART has the lowest RMSE in linear and non-linear data generation processes, and also the performance of BART important variables in a set of macroeconomic data has an optimal performance than other regression estimators.

2. Bayesian Additive Regression Trees

Applying Bayesian non-parametric to causal inference is rarely as simple as taking an off-the-shelf non-parametric prior and applying it in the same way one would to a prediction problem. Causal inference problems are often targeted in the sense that the final aim is to estimate a low-dimensional parameter, with non-parametric techniques used to deal with high or infinite dimensional nuisance parameters. The shrinkage induced by non-parametric models on the causal estimands introduces subtle, but serious, complications. Because of this, special care should be taken when applying Bayesian non-parametric [13].

The BART algorithm is straightforward to implement and requires the researcher only to input the outcome, treatment assignment, and confounding covariates but requires no information about how these variables are para-metrically related. Yet BART can detect interactions and non-linearities in the response surface, which (among other advantages) allows it to more readily identify heterogeneous treatment effects. Also, BART naturally produces coherent posterior intervals in contrast to methods such as propensity score matching and sub-classification [2].

Overall, BART represents a significant advancement in the field of dynamic system modeling and non-parametric regression techniques by offering a flexible, Bayesian approach that can effectively handle non-linearities, uncertainty, automatic variable selection, robustness to outliers and missing data, and modeling of temporal dynamics. Its ability to combine the strengths of tree-based methods with Bayesian inference makes it a valuable tool for a wide range of applications in various domains.

Let $\mathrm{x}=\left(x_1,\cdots,x_p\right)$ denote a p-dimensional covariate vector, or regressors. Capital letter $X=\left(\mathrm{x}_1',\cdots,\mathrm{x}_n'\right)'$ denotes the $n\times p$ predictor matrix, and $\mathrm{y}=\left(y_1,\cdots,y_n\right)$ is a vector of target values for supervised learning. Suppose, under the standard regression setting:

$$y_i=f\left(\mathrm{x}_i\right)+\varepsilon_i,\quad \varepsilon_i\sim N\left(0,\sigma^2\right) \tag{1}$$

It is assumed the residual term ε is a Gaussian noise term with mean zero and variance σ^2. The BART model [1] assumes that the unknown function $f(x)$ in the regression model (1) can be approximated by a sum of regression trees, *i.e.*

$$f(\mathrm{x}) = \sum_{l=1}^{L} g(\mathrm{x}, \mathrm{T}_l, \mu_l) \tag{2}$$

where T_l represents a tree structure, which is a set of split rules partitioning the covariate space, and μ_l is a vector of leaf parameters associated with the leaf nodes in tree T_l.

Trees are known to be prone to over-fitting due to their high flexibility. Thus, proper regularization is necessary to achieve good out-of-the-sample performance. BART assigns a regularization prior to the tree structure that strongly favors weak, or small trees. The tree prior $p(\mathrm{T}_l)$ specifies the probability for a node to split into two child nodes at depth d to be

$$\alpha(1+d)^{-\beta}, \quad \alpha \in (0,1), \beta \in [0,\infty) \tag{3}$$

which decreases exponentially as the tree grows deeper, implying strong regularization in the size of the tree. The prior of each leaf parameter $p(\mu_{lb})$ is assumed to be independent normal with variance τ, *i.e.* $\mu_{lb} \sim N(0,\tau)$. The prior of the residual variance σ^2 is set to be inverse-Gamma (a,b).

The ensemble of trees is fitted by Bayesian back fitting and MCMC sampling scheme. Let T_{-l} denotes the set of all trees except T_l, and similar define μ_{-l}. Note that the conditional posterior $p(\mathrm{T}_l, \mu_l \mid \mathrm{T}_{-l}, \mu_{-l}, \sigma^2, y)$ depends on other trees and parameters only through the residuals:

$$\mathrm{r}_l := y - \sum_{h \neq l} g(\mathrm{x}, \hat{\mathrm{T}}_h, \hat{\mu}_h) = g(\mathrm{x}, \mathrm{T}_l, \mu_l) \tag{4}$$

The original BART model [1] draws trees from the posterior using a random walk Metropolis-Hastings MCMC (MH-MCMC) algorithm. Per iteration, the algorithm randomly proposes a single growing or pruning procedure to each tree and accepts or rejects according to the MH ratios [14].

2.1. Regularization of BART

The main difference between BART and other methods is in choosing the number of trees, *i.e.* m. If BART is used to estimate $f(x)$ or predict Y, it makes sense to treat m as an unknown parameter. The best value of m is selected through cross-validation, of course, this approach is not computationally efficient, and to avoid the computational cost, [1] suggest the default value of $m = 200$. In a single-tree model (*i.e.*, $m = 1$), a tree with many terminal node may be needed to model a complicated structure. However, for a sum-of-trees model, Especially with m or number of large trees, It is essential that the regularization prior to keep the individual tree components small.

2.2. Priors and Likelihood

There are three priors for the BART model: a prior on the tree structure itself, a

prior on the leaf parameters, and a prior on the error variance σ^2.

The prior on σ^2 is independent of the other two and each tree is independent, yielding:

$$\begin{aligned} &p\left(\left(T_1,M_1\right),\cdots,\left(T_m,M_m\right),\sigma\right) \\ &=\left[\prod_j p\left(T_j,M_j\right)\right]p(\sigma)=\left[\prod_j p\left(M_j\,|\,T_j\right)p\left(T_j\right)\right]p(\sigma) \end{aligned} \tag{5}$$

where the last line follows from an additional assumption of conditional independence of the leaf parameters given the tree's structure.

To preserve the effect of each single tree, prior settings for model parameters are considered. Also, the absence of these settings causes the creation of a large number of parameters, which creates additional limitations in the calculations. According assumption of priors independence, we specify only three priors:

The first prior is on the locations of nodes within the tree. Nodes at depth d are non-terminal with probability $\alpha(1+d)^{-\beta}$ where $\alpha\in(0,1)$ and $\beta\in[0,\infty]$. This prior keeps the tree shallow, limiting the complexity of any single tree. Default values for these hyper-parameters $\alpha=0.95$ and $\beta=2$ are recommended by [1]. For non-terminal nodes, splitting rules have the following prior. First, a predictor is randomly selected to serve as the splitting variable. In the original formulation, each available predictor is equally likely to be chosen, but this is relaxed in our implementation to allow an arbitrary discrete distribution. Then, the splitting value is selected by randomly choosing a value of the selected predictor with equal probability.

The second prior is on the leaf parameters. Given a tree with a set of terminal nodes, each terminal node (or leaf) has a continuous parameter (the leaf parameter) representing the best guess of the response in this partition of predictor space.

Each leaf parameter is assigned a conjugate normal distribution $\mu_\ell \overset{iid}{\sim} \mathcal{N}\left(\mu_\mu,\sigma_\mu^2\right)$. In order to determine the parameters μ_μ and σ_μ^2, it should be noted that $E(Y\,|\,x)$ is the sum of m to μ_{ij} under the tree sum model, and since μ_{ij} have an independent prior and the same distribution, therefore, the prior of $E(Y\,|\,x)$ has a distribution of $N\left(m\mu_\mu,m\sigma_\mu^2\right)$ and most likely $E(Y\,|\,x)$ will be between $y_{\min}$ and $y_{\max}$. By choosing μ_μ and σ_μ^2 pre-selected k values, the value of $y_{\min}=m\mu_\mu+k\sqrt{m}\sigma_\mu$ and $y_{\max}=m\mu_\mu+k\sqrt{m}\sigma_\mu$ is determined. $(y_{\min}+y_{\max})/2$ is chosen as the center of range. The variance is chosen empirically so that the center of range plus or minus $k=2$, covers 95% variance of the response values provided in the training set (by default). If the values of k are between 0 and 1, the desired model will perform better and the value can be calculated from the cross-validation method. The aim of this prior is to provide model regularization by shrinking the leaf parameters towards the center of the distribution of the response.

The final prior is on the error variance and is chosen to be InvGamma $(\nu/2,\nu\lambda/2)$. We use data values to find ν and λ values. We assign a significant probability to *sigma* so as to avoid over-concentration and over-dispersion.

There are two choices for estimating $\hat{\sigma}$, a simple way is to let $\hat{\sigma}$ be the sample standard deviation of Y, or alternatively, $\hat{\sigma}$ can be expressed as the standard deviation of the residuals from a least-squares fit of a linear regression of X on the Y. considered Usually, the value of v is chosen between 3 and 10 and the value of λ is considered as the q^{th} prior quantile $\hat{\sigma}$, so that there is a prior chance $q = 90\%$ (by default) that the BART model based on the criteria RMSE is better than ordinary least squares regression. Therefore, the majority of the prior probability density is lower than the RMSE of the least squares regression.

2.3. Validity and Applicability Limitations of BART

While Bayesian Additive Regression Trees (BART) offer numerous advantages for dynamic system modeling and forecasting, it's essential to consider their limitations and assumptions, as they can affect the validity and applicability of the analysis.

First, BART assumes that the relationship between the response variable and predictors is additive. While this assumption holds for many real-world scenarios, there may be cases where interactions between predictors are significant, and an additive model may not accurately capture these relationships. Second, The performance of BART models can be sensitive to the choice of these hyperparameters, and suboptimal selections may lead to poor model performance or overfitting. Careful cross-validation and hyperparameter tuning are necessary to ensure robust and reliable results, which can add complexity to the modeling process. Third, Like many machine learning models, BART may struggle with extrapolation, particularly when making predictions outside the range of observed data. Extrapolation uncertainty can lead to unreliable forecasts, especially in dynamic systems where future conditions may differ significantly from historical observations. Fourth, BART assumes that the true relationship between predictors and the response variable can be adequately represented by a sum of regression trees. However, if the true relationship deviates substantially from this assumption (e.g., non-additive or non-tree-like relationships), BART may produce biased or misleading results.

3. Competing Regression Estimator

3.1. LASSO Estimator

The LASSO proposed by [15] is a standard technique that minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint, it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. The LASSO solves a penalized log-likelihood function given by

$$\hat{\beta} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right] \tag{6}$$

where n is the number of observations, p the number of predictors, and λ is

the penalty term, such that large values of λ shrinks a large number of the coefficients towards zero.

3.2. Elastic-Net Estimator

The Elastic-Net (EN) estimator proposed by [16] is based on a compromise between the lasso and ridge regression [17] penalties. Similar to the lasso, the elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. Simulation studies and real data examples show that the elastic net often outperforms the lasso in terms of prediction accuracy. The EN estimator solves the following

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha \left| \beta_j \right| \right] \right\} \tag{7}$$

where $0 \le \alpha \le 1$ is a penalty weight. The EN with $\alpha = 1$ is identical to the lasso, whereas it turns out to be ridge regression with $\alpha = 0$ [18]. Setting α close to 1 makes the EN behave similarly to the lasso, but eliminates problematic behavior caused by high correlations. When α increases from 0 to 1, for a given λ the sparsity of the minimization (*i.e.*, the number of coefficients equal to zero) increases monotonically from 0 to the sparsity of the lasso estimation. The elastic net can select more variables than observations.

3.3. Bayesian-Network Estimator

The Bayesian-Network estimator is based on the concept of network models as a convenient representation of the relationships among a set of variables. The networks are defined by nodes joined by a set of links, describing the statistical relationships between a pair of variables. In a regression model, the relationship between a dependent variable y and a set of p-dimensional covariate vector $\mathrm{x} = \left(x_1, \cdots, x_p \right)$ is given by:

$$y_i = \sum_{j=1}^{p} \beta_j X_{ij} + u, \quad u \sim \mathcal{N}\left(0, \sigma^2\right) \tag{8}$$

Suppose the coefficient of $\beta = \left(\beta_1, \cdots, \beta_p \right)$ has some zeros elements corresponding to sparsity (missing edges) in the underlying conditional independence structure which we refer to as a network. More specifically, if $\beta_j = 0$ then x_j has no relationship (or influence) on y. In network terms, this means a missing edge between variables y and x_j. However, if $\beta_j \neq 0$ then x_j has an impact (or influence) on y. Based on this illustration, there is a correspondence between the regression coefficients and the network structure defined by:

$$\beta = (\phi \circ g) \tag{9}$$

where operator $(\circ)$ is the element-by-element product such that $\beta_j = \phi_j g_j$ with

$$\beta_j = \begin{cases} 0 & \text{if} \quad g_j = 0 \quad \Rightarrow \quad x_j \not\to y \\ \phi_j \in \mathbb{R} & \text{if} \quad g_j = 1 \quad \Rightarrow \quad x_j \to y \end{cases} \tag{10}$$

where $x_j \not\to y$ means that x_j does not influence *y*.

4. Simulation Experiments

In this section, we demonstrate the application of BART on empirical examples and evaluate BART capabilities on simulated data used by [19]. We make this comparison in linear and non-linear Data Generated Processes (DGP). Also, the foundation of our evaluation of the Competitive models is RMSE.

The default hyperparameters generally follow the recommendations of [1] and provide a ready-to-use algorithm for many data problems. For both DGP, Our hyperparameter settings are $v=3$, $q=0.9$, $k=2$, and, $m=200$. Using the backfitting MCMC algorithm, we generated 20,000 MCMC draws from the posterior after skipping 250 burn-in iterations.

As competitors, we considered 3 estimators: lasso, elastic net, and Bayesian network. These competitors were chosen because, like BART, they are black box predictors. In this simulation, we are looking for whether the BART model performs better in selecting predictors and forecasting in linear models or non-linear models. In the following, we will test this comparison for a set of US macroeconomic data to compare the simulation results with the observed data results.

4.1. Linear Data Generation Process (DGP-L)

We next proceed to illustrate various features of BART on simulated data where we can gauge its performance against the true underlying signal. For this purpose, we consider the following function as the data generation process assuming a linear version of the [19] model. For this purpose, we constructed data by simulating values of $x=\left(x_1, x_2, \cdots, x_p\right)$ where:

$x_1, x_2, \cdots, x_p$ i.i.d. $\sim$ Uniform$(0,1)$

$$y=f(x)+\varepsilon=\pi X_1+\pi X_2+2\pi X_3+10X_4+5X_5+\varepsilon \tag{11}$$

where $\varepsilon \sim N(0,1)$. Because y only depends on $x_1, \cdots, x_5$, the predictors $x_6, \cdots, x_p$ are irrelevant. These added variables together with the interactions and non-linearities make it more challenging to find $f(x)$ by standard parametric methods.

[19] used this setup with p = 10 and to illustrate the potential of multivariate adaptive regression splines (MARS). We compare BART's performance with the same set of competitors used by increasing variables up to p = 100. We increased the number of irrelevant predictors in the data to show BART's effectiveness at detecting a low-dimensional structure in a high-dimensional setup. When p increases, the BART model is still not the best, and the LASSO model has the lowest RSME, and as a result, the BART models do not perform better than competitive estimators in linear settings.

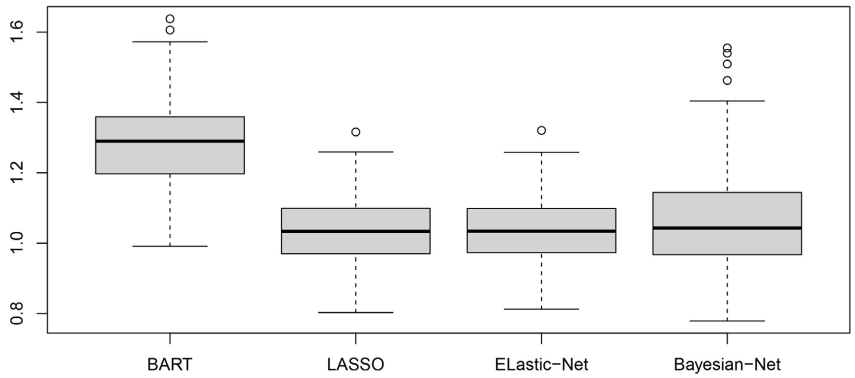
4.1.1. Comparing Model Predictive Performance

To perform comparisons across data sets, we considered RMSE and the mean and middle, and 50%, 75% RMSE quantiles are given in **Table 1**.

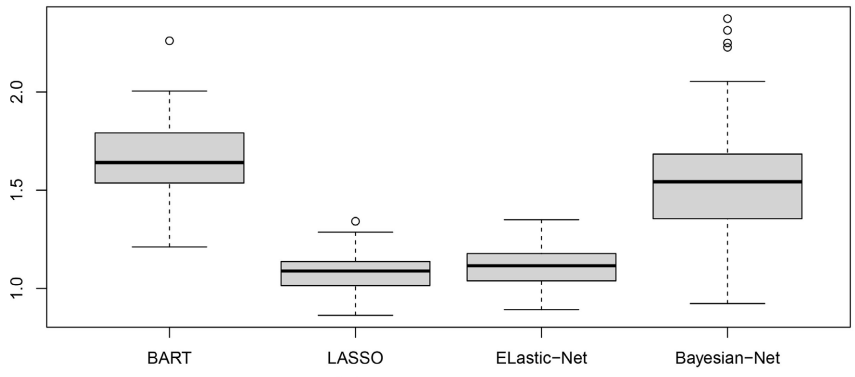
Table 1. Mean, middle and, 0.50, 0.75 quantiles of relative RMSE values for each method for DGP-L when $p = 10, 100$.

$p = 10$	mean	median	Q 0.50	Q 0.75
BART	1.292671	1.290039	1.290039	1.358976
LASSO	1.039849	1.033391	1.033391	1.098285
Elastic-Net	1.042411	1.034202	1.034202	1.098245
Bayesian-Net	1.070179	1.042845	1.042845	1.139688
$p = 100$				
BART	1.645258	1.640708	1.640708	1.790123
LASSO	1.083434	1.089314	1.089314	1.137274
Elastic-Net	1.113681	1.115589	1.115589	1.174979
Bayesian-Net	1.547082	1.543075	1.543075	1.681589

Although relative performance in **Figure 1(a)** and **Figure 1(b)** varies widely across the different problems. It is clear from the distribution of RMSE values that BART tended to more often obtain Bigger RMSE than any of its competitors. In fact, this is evidence of the lower performance of BART compared to other competitive linear models. When the p is small, the LASSO, Elastic-Net, and Bayesian-net are not so different but when the p is large, Bayesian-net is closer to BART.



(a) DGP-L, $p = 10$



(b) DGP-L, $p = 100$

Figure 1. Boxplots of the RMSE of DGP-L for the competing methods over 100 simulations when $p = 10, 100$.

4.1.2. Comparing Variable Importance

After building a predictive model, it is natural to ask the question: which variables are most important? This is assessed by examining the splitting rules in the m trees across the post-burn-in Gibbs samples which are known as "inclusion proportions" [1]. The inclusion proportion for any given predictor represents the proportion of times that variable is chosen as a splitting rule out of all splitting rules among the posterior draws of the sum-of-trees model. The segments atop the bars represent 95% confidence intervals. The predictors with inclusion proportions of zero feature identically one value (after missing data was dropped).

We do this for two linear and non-linear scenarios with $p = 10$ settings. We also repeat this process for $p = 100$ settings.

In all four models of **Figure 2**, the most important variable is X_4, and afterwards is X_3, X_5, X_2, and X_1. In the sum of tree models, the variables that are at rest show a value between 0 and 0.05, while in competitive LASSO and Elastic Net models, this value is less than 0.01, and in the Bayesian network model, it is equal to 0.

By increasing the number of variables to $p = 100$ on **Figure 3**, X_4 is still the most important variable in all four methodologies, followed by X_3 and X_5. Also, the rest of the variables are the same. But the difference is that variable X_1 is more important than X_2 and also both X_1 and X_2 have zero value in the Bayesian Net model.

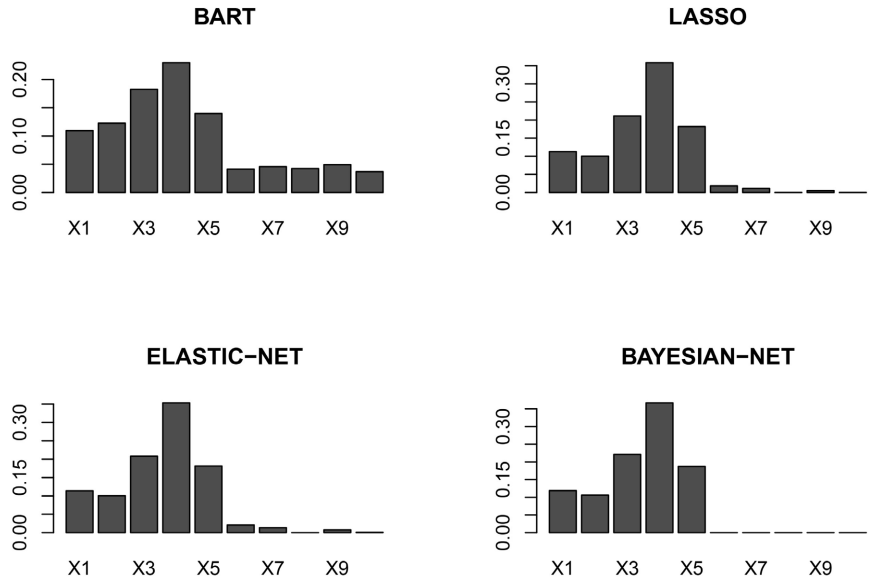


Figure 2. Variable importance in the linear scenario when $p = 10$.

4.2. Non-Linear Data Generation Process (DGP-NL)

We consider the following function as the data generation process assuming a non-linear framework:

$$y = f(x) + \varepsilon = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon \quad (12)$$

Similar to the linear DGP, we perform this simulation exercise by setting $p = \{10,100\}$. Again, we compare the BART performance with the same set of competing methods.

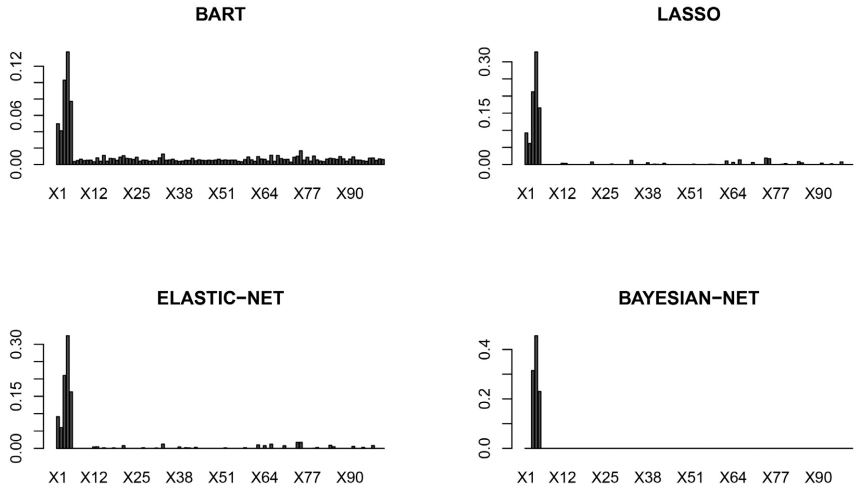


Figure 3. Variable importance in the DGP-L when $p = 100$.

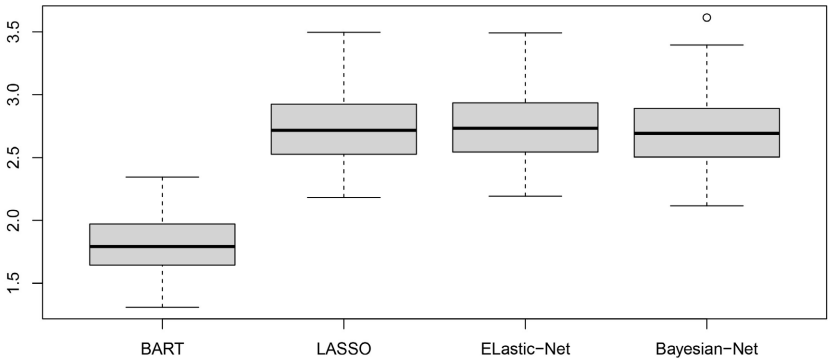
4.2.1. Comparing Model Predictive Performance

Table 2 reports the RMSE of the competing methods for $p = 10, 100$. By comparing Table 2, it can be seen that the nonlinear data generate process average RMSE is smaller than the linear models and is a significant contribution, which shows that the BART model has a better response in nonlinear functions but in the nonlinear data generation process, it has a better performance than the other competing linear estimators.

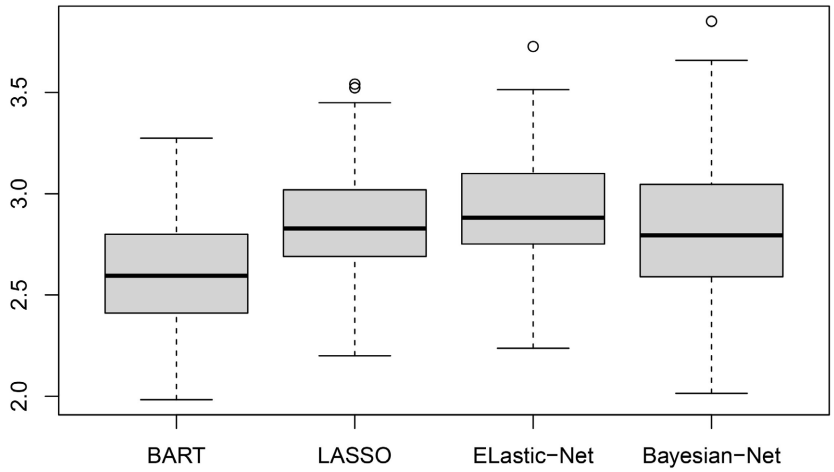
Table 2. Mean, middle and, 0.50, 0.75 quantiles of relative RMSE values for each method.

$p = 10$	mean	median	Q 0.50	Q 0.75
BART	1.806810	1.792002	1.792002	1.959392
LASSO	2.728710	2.744029	2.744029	2.933420
Elastic-Net	2.735580	2.750239	2.750239	2.949370
Bayesian-Net	2.684871	2.721498	2.721498	2.865008
$p = 100$	mean	median	Q 0.50	Q 0.75
BART	2.612356	2.593971	2.593971	2.798618
LASSO	2.848172	2.827754	2.827754	3.018718
Elastic-Net	2.910745	2.881590	2.881590	3.091359
Bayesian-Net	2.821050	2.794044	2.794044	3.042147

Figure 4(a) and Figure 4(b) reports the boxplot of the RMSE of nonlinear DGP for $p = 10, 100$, respectively. According to both plots, the lowest RMSE is for the BART model, and with the increase in the number of variables, the amount of errors has also increased. Especially in competitive models, the Elastic Net model has suffered more errors than the LASSO model.



(a) DGP-NL, $p = 10$



(b) DGP-NL, $p = 100$

Figure 4. Box plot of the RMSE of DGP-NL for the competing methods over 100 simulations when $p = 10, 100$.

4.2.2. Comparing Variable Importance

In the following, we specify the important variables for the non-linear scenario. In non-linear simulation with $p = 10$ settings, we observe that X_1 ranks first (see, **Figure 5**) then, X_2 and X_3 are our most important variables. This is even though in LASSO, Elastic Net, and Bayesian Net, X_4 is the most important variable, followed by X_1 and X_2. Also, variable X_3 is at rest.

When variables increase by a factor of 100 in the BART model, X_4 is the most important variable and this situation is the same in competitive models (see, **Figure 6**). In fact, by increasing the value of p to the settings of $p = 100$ in the MCMC algorithm, the calculation error has probably been resolved. Therefore, our suggestion is to identify the important variables of the BART model with a high number of variables.

5. Forecasting Macroeconomic Time Series

In this part, we test a set of macroeconomic data with 221 data and 8 variables in 4 lag. The data set for our empirical application consists of quarterly observations, from 1959Q1 to 2022Q3, of 8 US-macroeconomic variables which were

originally used by [20]. The macroeconomic variables include (GDP)—real gross domestic product, (INF)—consumer price index, (FF)—Federal funds rate, (M2)—money stock M2, (PC)—real personal consumption, (IP)—industrial production index, (U)—unemployment, and (INV)—real gross domestic private investment. Table 3 gives the data description and transformation code from [20] used for our application. We transferred data to the first difference and first difference of the log variable and set the lag settings to $n = 4$.

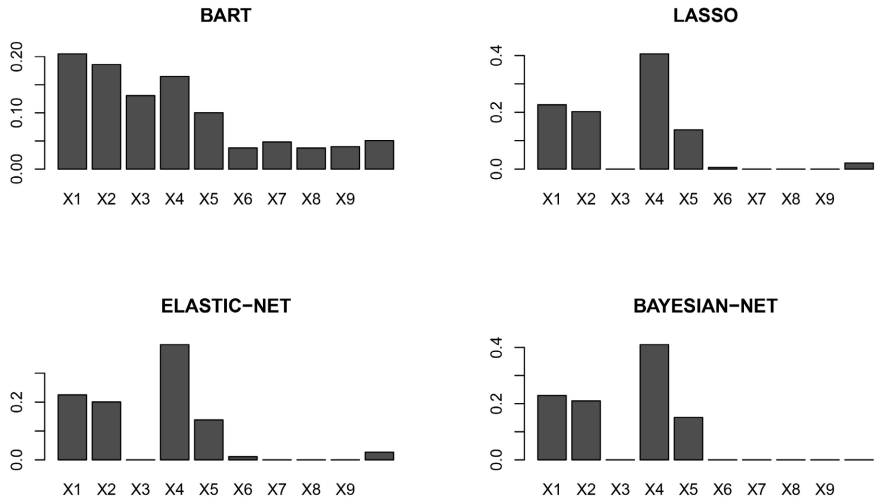


Figure 5. Variable importance in the non linear scenario when $p = 10$.

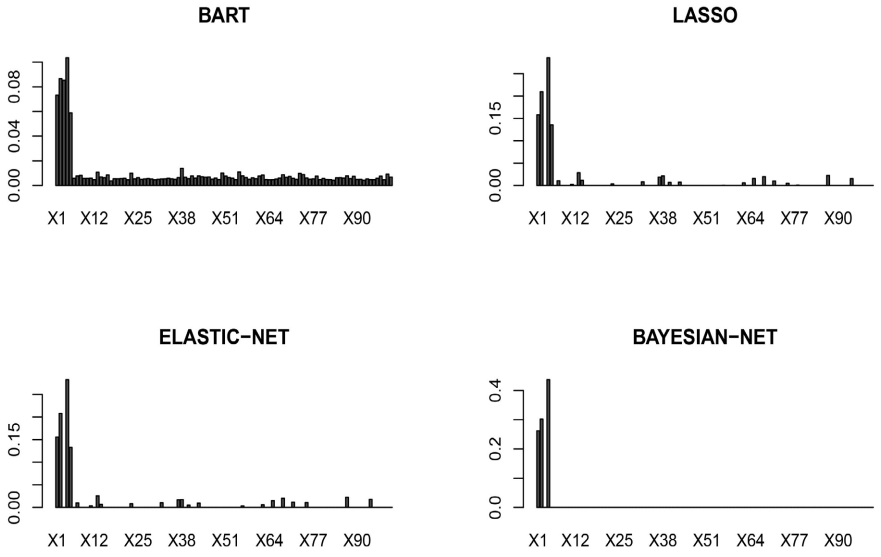


Figure 6. Variable importance in the nonlinear scenario when $p = 100$.

5.1. Comparing Model Predictive Performance

We want to show a forecast of GDP and show which model gives the best forecasting, so we rank the BART model and competitive models. For forecasting the effect of GDP shock on GDP, the BART is minimum, and INF, PC, IP, U, and INV effects on GDP have the same condition. And this shows that the BART

model has provided the best performance for forecasting most variables (see Table 4).

Table 3. Data description and transformation code to achieve stationarity. The transformation code is as follows: 1 = no transformation, 2 = first difference, 3 = second difference, 4 = log, 5 = first difference of the log variable, 6 = second difference of the log variable.

No	Short ID	Mnemonic	Code	Description
1	GDP	GDP251	5	Real GDP, Quantity Index (2000 = 100)
2	INF	CPIAUCSL	5	CPI All Items
3	FF	FEDFUNDS	2	Interest rate: Federal funds (effective) (% per annum)
4	M2	M2SL	5	Money stock: M2 (bil$)
5	PC	PCE	5	Real Personal Cons. Exp., Quantity Index
6	IP	INDPRO	5	Industrial production index: total
7	U	UNRATE	2	Unemp. rate: All workers, 16 and over (%)
8	INV	GPDIC1	5	Real gross domestic private investment

Table 4. RMSE of the competing methods by ranking.

	BART	Rank	LASSO	Rank	Elastic-Net	Rank	Bayesian-Net	Rank
GDP	1.8373	1	1.8376	2	1.9062	3	2.0475	4
INF	0.6457	1	0.9541	4	0.9104	3	0.8675	2
FF	0.5472	3	0.3593	2	0.3321	1	1.9648	4
M2	1.5440	4	1.3614	2	1.3500	1	1.3861	3
PC	3.9311	1	4.2034	4	4.0271	2	4.1056	3
IP	3.4303	1	4.0299	2	4.3242	3	4.9381	4
U	1.8548	1	1.9461	2	1.9529	4	1.9502	3
INV	4.7414	1	8.8198	4	7.8913	2	8.6739	3

5.2. Comparing Variable Importance

Next, we will check the important variables for the observed data: Figure 7 illustrates the variable importance of the competing models for predicting GDP. In the BART model, The variables M2.lag1, INF.lag3, FF.lag4, IP lag1 and INV.lag1 are key determinants.

For the LASSO and Elastic Net model, The variables M2.lag1 rank first, but in the Bayesian Net model, GDP.lag2 is highly significant. In these competitive models, important variables are selected and this is the difference between reporting important variables and BART. See Figures A1-7 for the results of the

relative importance when predicting CPI, FF, M2, PC, IP, U and INV, respectively.

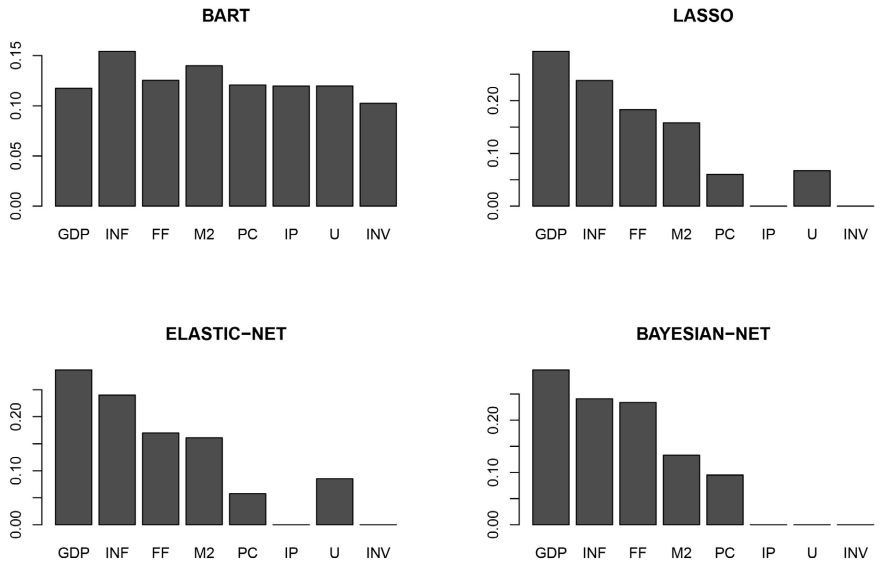


Figure 7. Variable inclusion proportions for predicting GDP according to the competing methods. The variables are arranged according to their lags as follows: (GDP, INF, FF, M2, PC, IP, U, INV).

6. Conclusions

This paper advances a non-parametric autoregressive distributed lag model that employs a Bayesian additive regression tree (BART). The performance of the BART methodology is compared with selection models like Lasso, Elastic Net, and Bayesian networks in linear and non-linear simulation scenarios, as well as application to forecasting macroeconomic data.

Our results show that in the case of non-linear relationships between the variables, the nonparametric structure like BART works better than the competing estimators, and very competitive against the linear parametric methods when the true model is linear. The result also show that when applied to modeling and forecasting macroeconomic times series, the BART non-parametric model outperforms the linear models like Lasso, Elastic Net, and Bayesian networks. This suggests that many macroeconomic variables have non-linear relationships and must therefore be modeled with non-linear models like the BART.

Since most of the relationships between macroeconomic variables have nonlinear relationships, we recommend future researchers make this comparison with the empirical application of economic theories as well as relationships between macroeconomic variables.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Chipman, H.A., George, E.I. and McCulloch, R.E. (2010) Bart: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, **4**, 266-298. https://doi.org/10.1214/09-AOAS285

[2] Hill, J.L. (2011) Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, **20**, 217-240. https://doi.org/10.1198/jcgs.2010.08162

[3] Zeldow, B.M. (2017) Bayesian Nonparametric Methods for Causal Inference and Prediction. Doctor's Thesis, University of Pennsylvania.

[4] Sparapani, R., Logan, B.R., McCulloch, R.E. and Laud, P.W. (2020) Nonparametric Competing Risks Analysis Using Bayesian Additive Regression Trees. *Statistical Methods in Medical Research*, **29**, 57-77. https://doi.org/10.1177/0962280218822140

[5] Spanbauer, C. and Pan, W. (2022) Flexible Instrumental Variable Models with Bayesian Additive Regression Trees. arXiv: 2210.01872.

[6] Zhang, J.L. and Härdle, W.K. (2010) The Bayesian Additive Classification Tree Applied to Credit Risk Modelling. *Computational Statistics & Data Analysis*, **54**, 1197-1205. https://doi.org/10.1016/j.csda.2009.11.022

[7] de Brito, D.A. and Artes, R. (2018) Application of Bayesian Additive Regression Trees in the Development of Credit Scoring Models in Brazil. *Production*, **28**, e20170110. https://doi.org/10.1590/0103-6513.20170110

[8] 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**, 979-1022. https://doi.org/10.1111/iere.12619

[9] Mumtaz, H. and Piffer, M. (2022) Impulse Response Estimation *via* Flexible Local Projections. arXiv: 2204.13150. https://doi.org/10.2139/ssrn.4088760

[10] Jordà, Ò. (2005) Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review*, **95**, 161-182. https://doi.org/10.1257/0002828053828518

[11] Sims, C.A. (1980) Macroeconomics and Reality. *Econometrica*, *Econometric Society*, **48**, 1-48. https://doi.org/10.2307/1912017

[12] Huber, F. and Rossini, L. (2022) Inference in Bayesian Additive Vector Autoregressive Tree Models. *The Annals of Applied Statistics*, **16**, 104-123. https://doi.org/10.1214/21-AOAS1488

[13] Linero, A.R. and Antonelli, J.L. (2022) The How and Why of Bayesian Nonparametric Causal Inference. *WIREs Computational Statistics*, **15**, e1583. https://doi.org/10.1002/wics.1583

[14] Wang, M., He, J. and Hahn, P.R. (2022) Local Gaussian Process Extrapolation for Bart Models with Applications to Causal Inference. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2023.2240384

[15] Tibshirani, R. (1996) Regression Shrinkage and Selection *via* the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[16] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection *via* the Elastic net. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[17] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, **12**, 69-82.

https://doi.org/10.1080/00401706.1970.10488635

[18] Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models *via* Coordinate Descent. *Journal of Statistical Software*, **33**, 1-22. https://doi.org/10.18637/jss.v033.i01

[19] Friedman, J.H. (1991) Multivariate Adaptive Regression Splines. *The Annals of Statistics*, **19**, 1-67. https://doi.org/10.1214/aos/1176347963

[20] Ahelegbey, D.F., Billio, M. and Casarin, R. (2016) Bayesian Graphical Models for Structural Vector Autoregressive Processes. *Journal of Applied Econometrics*, **31**, 357-386. https://doi.org/10.1002/jae.2443

Appendix

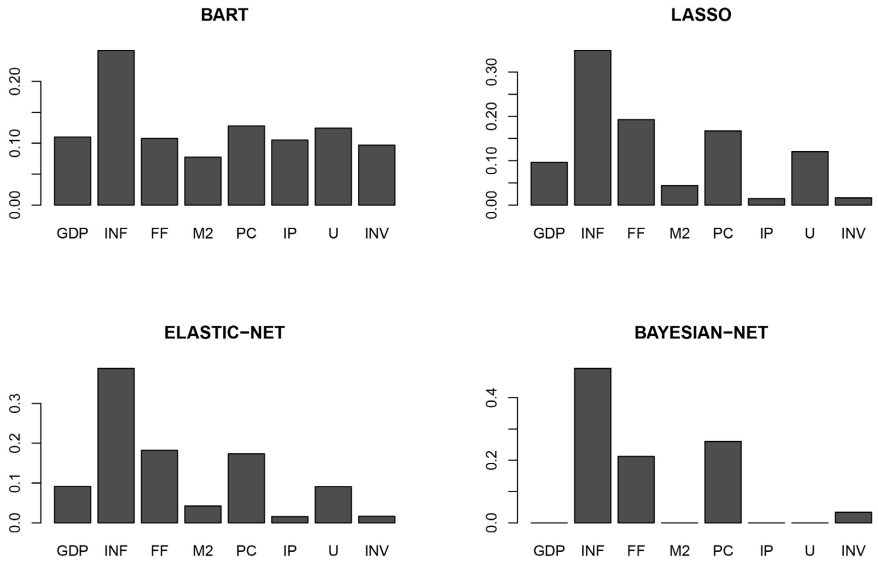


Figure A1. Variable inclusion proportions for predicting INF according to the competing methods. The variables are arranged according to their lags as follows: (GDP, INF, FF, M2, PC, IP, U, INV).

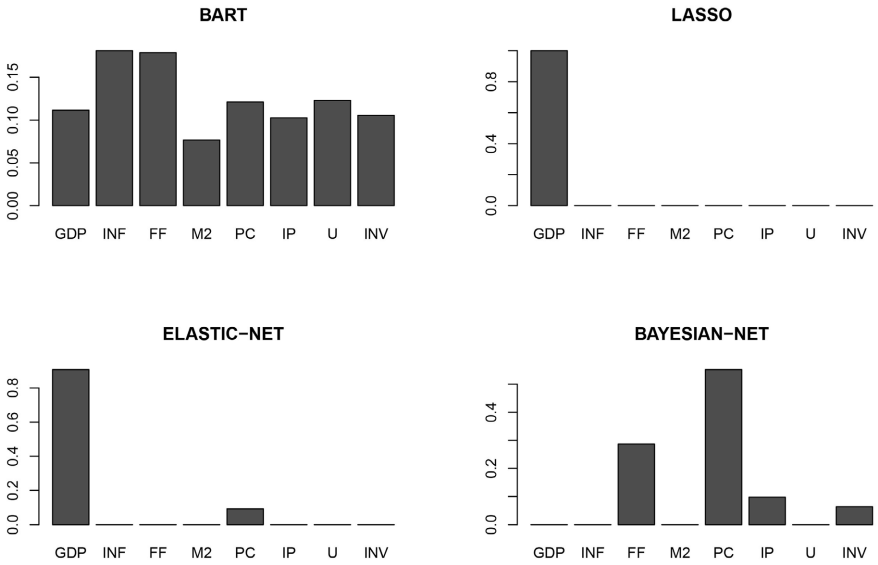


Figure A2. Variable inclusion proportions for predicting FF according to the competing methods. The variables are arranged according to their lags as follows: (GDP, INF, FF, M2, PC, IP, U, INV.

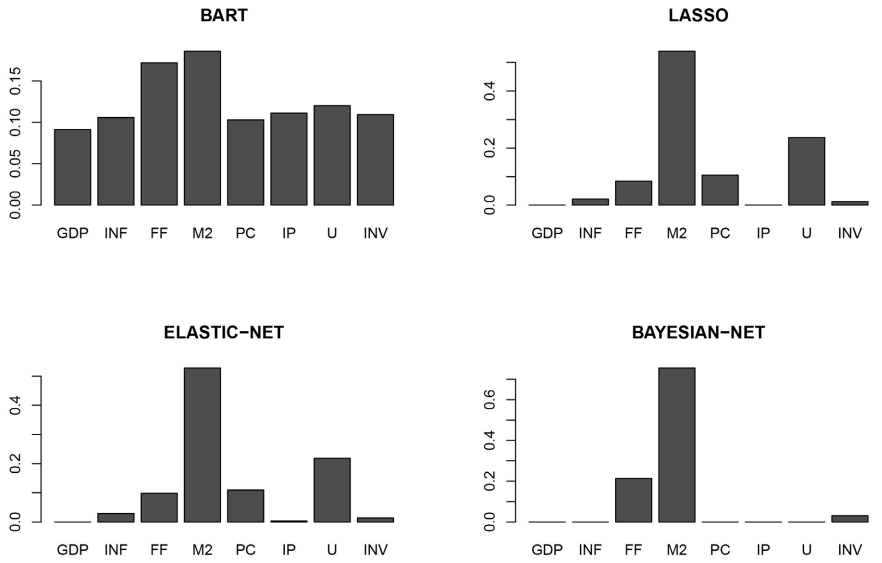


Figure A3. Variable inclusion proportions for predicting M2 according to the competing methods. The variables are arranged according to their lags as follows: (GDP, INF, FF, M2, PC, IP, U, INV.

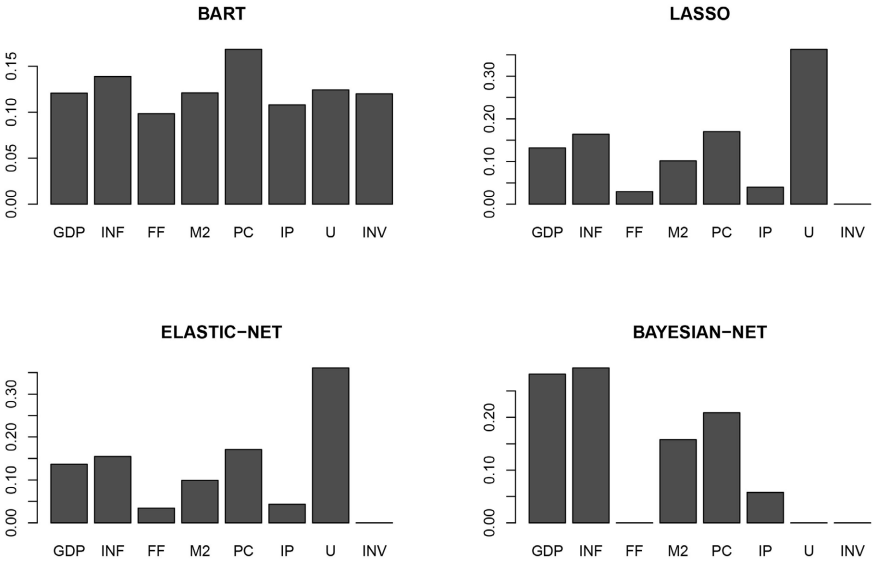


Figure A4. Variable inclusion proportions for predicting PC according to the competing methods. The variables are arranged according to their lags as follows: (GDP, INF, FF, M2, PC, IP, U, INV.

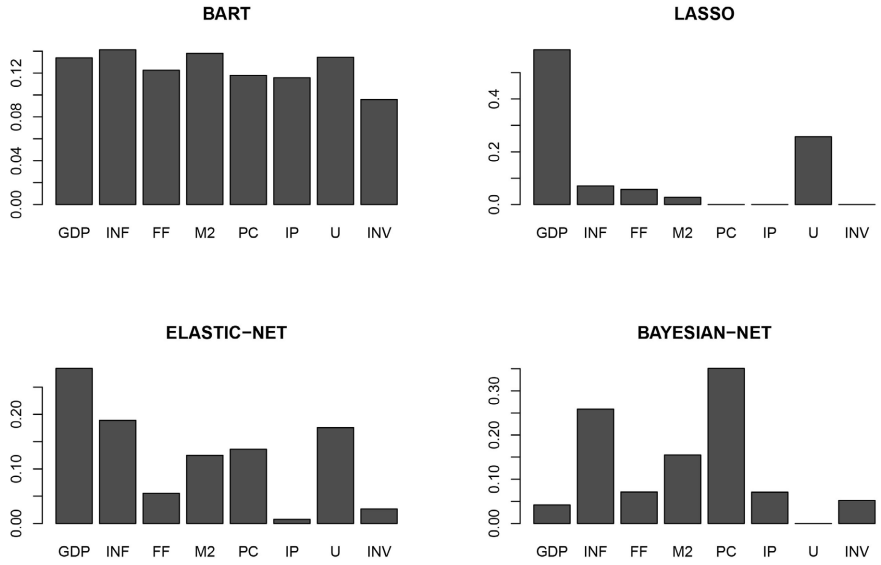


Figure A5. Variable inclusion proportions for predicting IP according to the competing methods. The variables are arranged according to their lags as follows: (GDP, INF, FF, M2, PC, IP, U, INV.

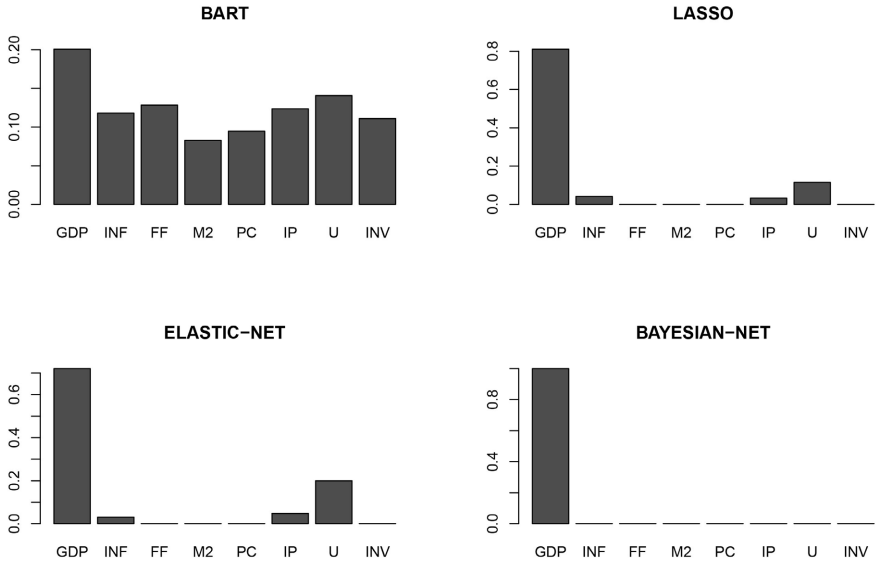


Figure A6. Variable inclusion proportions for predicting U according to the competing methods. The variables are arranged according to their lags as follows: (GDP, INF, FF, M2, PC, IP, U, INV.

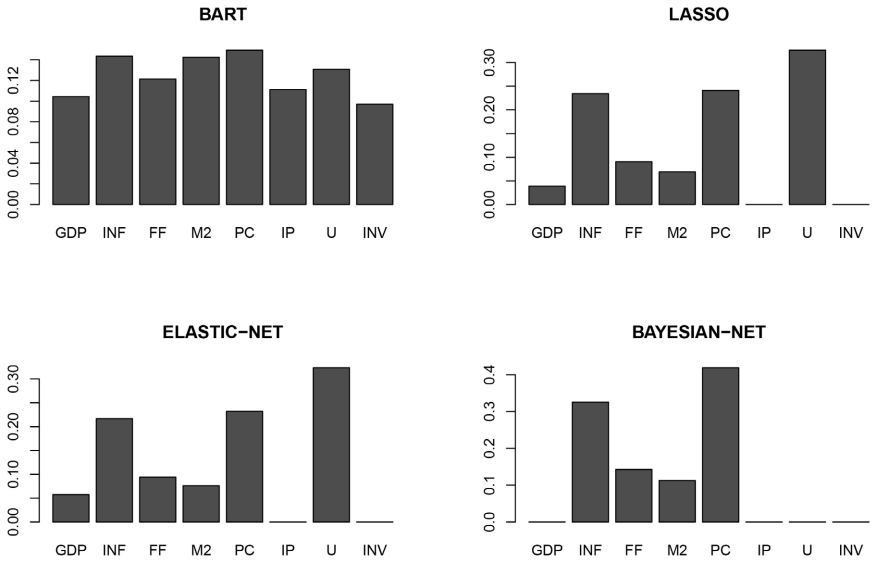


Figure A7. Variable inclusion proportions for predicting INV according to the competing methods. The variables are arranged according to their lags as follows: (GDP, INF, FF, M2, PC, IP, U, INV.