A Time-varying Bayesian Variable Selection for Macroeconomic Forecasting

Hyun Jae Stephen Chu

Department of Economics, Korea University, E-mail: stephenchu@korea.ac.kr

Jaeho Kim

Department of Economics, Sogang University, E-mail: jaehoecon@sogang.ac.kr

Kyu Ho Kang

Corresponding author, Department of Economics, Korea University, E-mail: kyuho@korea.ac.kr

October 2024

Abstract

Recently, the considerable attention on using big-data sets in macroeconomic forecasting has brought variable selection methods to the spotlight. This is because such methods are capable of mitigating the amplified uncertainty of parameters and inaccurate forecasts caused by the large number of predictors. While variable selection is growing in popularity, one often overlooked issue is the nonlinear relationship between variables that causes different sets of variables to be selected across time, and neglecting such relationship would bring about inaccurate forecasts. To address this, we propose a Bayesian Dirac Classification and Regression Tree (B-DART) model, which integrates the Dirac variable selection method with the Bayesian Classification and Regression Tree (B-CART) framework. We find empirical evidence that the B-DART model outperforms both the Dirac variable selection and B-CART models in forecasting certain macroeconomic variables and achieves a particularly improved forecast accuracy during relatively volatile periods, such as the post-COVID-19 era. (JEL classification: C11, C22, C52, C53).

Keywords: Variable Selection, Dirac spike-and-slab, Bayesian CART, Nonlinear Regression, Forecasting

1 Introduction

Recently, variable selection methods have gained considerable attention in macroeconomic forecasting, together with the availability of using big data sets.[1] Incorporating a large number of predictor variables into forecasting models helps mitigate the loss of forecast accuracy caused by omitting important predictors and also guards against forecast instability.[2] Despite these advantages, expanding the dimensionality of the predictor space can lead to increased parameter uncertainty and, consequently, inaccurate forecasts. As a result, variable selection methods such as Giannone et al. (2021) have become increasingly popular.

One often overlooked aspect of variable selection methods is that the set of selected predictor variables can change over time. This is because the relationship between economic variables are instable and can shift due to external shocks, such as changes in economic policies, fluctuating market conditions or technological advancements. Indeed, numerous studies have found evidence of the time-varying relationships in various macroeconomic variables, such as inflation, GDP growth, equity premium, stock market predictability, exchange rate and business cycles (Stock and Watson (1996), Stock and Watson (2003), Welch and Goyal (2008), Pesaran and Timmermann (1995), Rapach and Wohar (2006), Rossi (2013), Ng and Wright (2013)).

Forecasting accuracy may decline if the nonlinear time-varying relationships among macroeconomic variables are not fully accounted for, making it harder to detect economic shifts and leading to misaligned monetary policy or distorted policy effects. Recent studies by Jordà et al. (2022) and Bernanke and Blanchard (2023) highlight that the COVID-19 pandemic has fundamentally altered the relationships between macroeconomic variables, emphasizing that the shock was inherently different from conventional demand-side shocks, primarily driven by the supply-side inflationary pressures. Such shifts in relationships may consequently lead to instability in variable selection models as relevant variables can vary over time in response to the changing economic dynamics. Insufficient consideration of this instability may thus have led to an overly cautious monetary policy response, whereas more decisive adjustments could have been more effective in addressing such inflationary pressures.[3]

[1] Varian (2014) outlines various methods of employing big-data in econometric analysis, while Bok et al. (2018) provides examples of tracking economic conditions by applying big-data to forecasting macroeconomic variables.

[2] Rossi (2021) defines forecast instability as instability in the loss function: assuming a quadratic loss function, forecasting instability is simply time-varying forecast errors.

[3] *"So I think if we knew now—of course, if we knew now that these supply blockages, really, and the*

Therefore, the purpose of this paper is to allow for time-varying selection of predictor variables to improve forecast accuracy. This can be achieved by (1) partitioning the sample into the corresponding groups and (2) applying variable selection within the specific groups. These steps are done jointly since the source and timing of instability is assumed to be unknown. Applied to a high-dimensional predictor space framework, we argue that this combined approach offers better forecasting performance compared to implementing either (1) or (2) independently.

How, then, can the time-varying variable selection be incorporated into a forecasting model? While structural break tests are commonly used, they are not a silver-bullet; structural breaks are neither necessary nor sufficient for forecasting instability.[4] Another approach is to use nonlinear parametric models, such as Markov-switching or change point models.[5] Although these have the flexibility of handling multiple breakpoints, one crucial limitation in the high-dimensional predictor space environment is that the accuracy of the estimated breaking point decreases due to the increasing number of parameters. Therefore, we adopt the nonparametric Bayesian Classification and Regression Tree (B-CART) model, which can identify multiple and discontinuous structural break points more effectively compared to the standard methods mentioned above.[6]

Suggested by Chipman et al. (1998), the B-CART model uses a binary tree to recursively partition the predictor space into subsamples (or groups) where each y_t is heterogeneous. This allows for not only the estimation of both the subsample groups of y_t but also the data-driven functional form of the predictor variables. Specifically, the model generates a binary tree based on a prior distribution governing the structure of the tree, which consists of the tree structure prior and splitting rule prior. The tree structure prior governs the maximum depth of the tree, and once a split is determined, a threshold value of a predictor variable is randomly chosen as the splitting rule. Data within the tree is then divided into two groups accordingly.

Once the groups are specified by the B-CART model, variable selection is conducted within each group. As Giannone et al. (2021) point out, both sparse and dense models

inflation resulting from them in collision with, you know, very strong demand, if we knew that that was what was going to happen, then in hindsight, yes. It would have been appropriate to move earlier."

- Jerome Powell, March 16, 2022

[4]For detailed examples, refer to section 2.5 in Rossi (2021).

[5]See Kim and Nelson (1999) for Markov-switching and change-point models, and Teräsvirta (2006) for other nonlinear models such as the smooth transition autoregressive (STAR) model.

[6]As highlighted by Medeiros et al. (2021) and Goulet Coulombe et al. (2022), nonlinear structures imposed by machine learning models have been shown to enhance the accuracy of macroeconomic forecasts.

must be considered when dealing with a high-dimensional predictor space.[7] Thus, we implement the Bayesian variable selection model with a spike-and-slab prior, in spirit of Mitchell and Beauchamp (1988). Among several candidates for the prior distribution, we choose the Dirac spike-and-slab hierarchical prior for its advantage in deriving the conditional marginal likelihood analytically, which we later use as the criterion to select the optimal tree. Furthermore, when facing a high-dimensional predictor space, the Dirac variable selection enhances forecast accuracy by mitigating parameter uncertainty originated by the inclusion of irrelevant variables or the omission of important variables.

We denote the combination of the B-CART and Dirac variable selection method as the Bayesian Dirac Classification and Regression Tree (B-DART). Given the tree and model priors from the B-DART model, a single tree is generated. The posterior distribution of the tree is then derived by adjusting the previously generated tree, using the Metropolis-Hastings (MH) algorithm. To generate candidate trees, the proposal distribution of the MH algorithm follows four rules: grow, prune, change, and swap. The algorithm explores possible trees until it converges to the optimal tree structure. The best tree is selected by comparing the marginal likelihood, conditional on the estimated inclusion parameter of the Dirac variable selection model, denoted as δ. Since the marginal likelihood is conditional on both the tree structure and selected variables for each group, the grouping of the sample and variable selection are performed jointly.

As mentioned above, although we have implemented the flexibility of time-varying variable selection, this does not guarantee improved forecast accuracy. Therefore, it is essential to compare the out-of-sample forecasting performance of the B-DART model against other models. The baseline forecasting model is the B-DART autoregressive distributed lag (ADL) model of order (4,1), while other candidates are the Dirac ADL(4,1) model, the B-CART ADL(4,1) model and an autoregressive lag (AR) model of order 1. The ADL type models are to find evidence that jointly applying time-varying relationship and variable selection does improve forecast performance, while the AR(1) model is to examine overfitting. Each model is evaluated using the relative root-mean-squared errors (RMSE) with respect to the AR(1) case.

Following McCracken and Ng (2020), we use a big-data set comprising 221 quarterly predictor variables, classified into 14 groups, from 1967 Q1 to 2024 Q2, to forecast 8 macroeconomic variables. Specifically, we generate 1- to 8-quarter-ahead direct forecasts

[7]Sparse type models are those that assume only a handful of predictors are important but the role of each are large, while dense models suppose that many predictors are important but each has a limiting impact. For further details, see Ng (2013).

for PCE, CPI and PPI price indices, Crude Oil price, federal funds effective rate (FFE), 10-year maturity interest rate, real GDP, and the unemployment rate. In order to avoid collinearity within the high-dimensional predictor space, principal component analysis (PCA) is applied to construct representative components for each of the 14 groups.

Our empirical study provides three main results. First, the proposed B-DART model that jointly considers time-varying variable selection showed improved forecast accuracy, compared to applying either the B-CART or Dirac method independently. Moreover, the B-DART model significantly outperforms other models during the relatively volatile post-COVID-19 periods, which further reinforces this evidence. Secondly, an analysis of the selected predictor variables across each out-of-sample period reveals that the B-DART model's time-varying variable selection effectively incorporates new information, yet may also present risks of overfitting. Third, performance of variable selection and estimation of the optimal tree structure is not independent, but rather produces a synergistic effect. Summing up, model uncertainty is prevalent, and not accounting for the time-varying nature of variable selection models-where selected variables may change across time-can lead to reduced forecasting accuracy, particularly during relatively volatile periods.

The remainder of the paper proceeds as follows. Section 2 provides the general model structure along with the prior distributions of the tree and model parameters. Then Section 3 and 4 each explains the details of the Bayesian CART and Dirac variable selection model. Section 5 illustrates the details of the Metropolis-Hastings Search Algorithm and the conditional marginal likelihood when estimating the optimal tree structure. Section 6 compares the out-of-sample forecast accuracy of the proposed B-DART model to other models by applying it to an empirical big data set. Section 7 concludes.

2 General Model Structure

Let the objective be forecasting a univariate time series variable. At time t, the model consists of a univariate response variable y_t and a set of predictor variables $x_t = (x_{1t}, \cdots, x_{kt}, \cdots, x_{Kt})'$. Stacking these from $t = 1$ to T, we obtain the matrix representation $Y = (y_1, \cdots, y_t, \cdots, y_T)'$ and $X = (X_1, \cdots, X_k, \cdots, X_K)$ where $X_k = (x_{1k}, \cdots, x_{tk}, \cdots, x_{Tk})'$. Note that the X_k's can include lagged values of y_t.

When the main interest is forecasting, one must accurately predict the response variable y_t given the data of predictor variables x_t and the functional form, denoted by $\mathcal{L}(\cdot)$. The most widely used linear model assumes that the function $\mathcal{L}(\cdot)$ is linear,

such that $\mathbb{E}[y_t|x_t] = x_t'\beta$ holds.[8] The priority is to accurately estimate the coefficient parameters β. However, sometimes it is more compelling to allow the functional form to be nonlinear, where the form must be estimated to fit the data as closely as possible. In other words, one must also get an accurate estimate for a "data driven" $\widehat{\mathcal{L}}(\cdot)$ such that

$$y_t = \widehat{\mathcal{L}}(x_t, \beta, \sigma^2) + \varepsilon_t \tag{1}$$

In particular, we use the B-CART method to estimate the functional form $\mathcal{L}(\cdot)$. The functional form is determined by a binary tree, where the tree size, structure, and splitting rule of each node follow a specific prior distribution. Thus, the functional form of the model, denoted as $\mathcal{L}(\cdot)$, is

$$\mathcal{L}(x_t, \Theta|\mathcal{T}, i^x, i^q) \tag{2}$$

where Θ is the set of each $\{\beta_g, \sigma_g^2\}_{g=1}^G$ in each group, $\mathcal{T}$ is the tree structure, i^x is the set of predictors used in each terminal node, and i^q is the splitting thresholds of each terminal node. Note that $\mathcal{T}$, i^x and i^q are the tree priors where the details of these will be further explained in Section 3.

Precisely, the binary tree $\mathcal{T}$ divides the predictor space by randomly selecting a predictor variable and a threshold value of that predictor variable at each internal node. For example, **Figure 1** below depicts a regression tree model. The first node, $X_1 \leq 0.5$, is called the root node (or internal node), which implies that the initial data set is divided to the left if $X_1 \leq 0.5$ and to the right if $X_1 > 0.5$. The next two internal nodes divide their respective subsamples regarding $X_2 \leq 0.5$ and $X_2 > 0.5$. The two nodes below an internal node are referred to as child nodes, while the parent nodes are the ones above. Finally, the nodes with no splitting rules assigned are called the terminal nodes, which represent the obtained groups consisted of a group specific parameters (β_g, σ_g^2).

Once the groups are determined by the Bayesian CART method, the Dirac Variable Selection method is applied to specify the statistically significant coefficient estimators of β_g. In other words, the Dirac spike-and-slab hierarchical prior is imposed to determine the importance of each $\beta_{g,k} \in \beta_g$. Specifically, to determine the prior distribution of β_g, we consider a latent indicator variable δ_g of each group, which takes the value of

[8]Note that we define linearity following Lee et al. (1993). Consider

$$\mathbb{E}[y_t|x_t] = x_t'\beta + g(x_t)$$

where x_t includes the lagged terms of y_t. Then for y_t to be linear mean conditional on x_t, $g(x_t) = 0$ must hold.

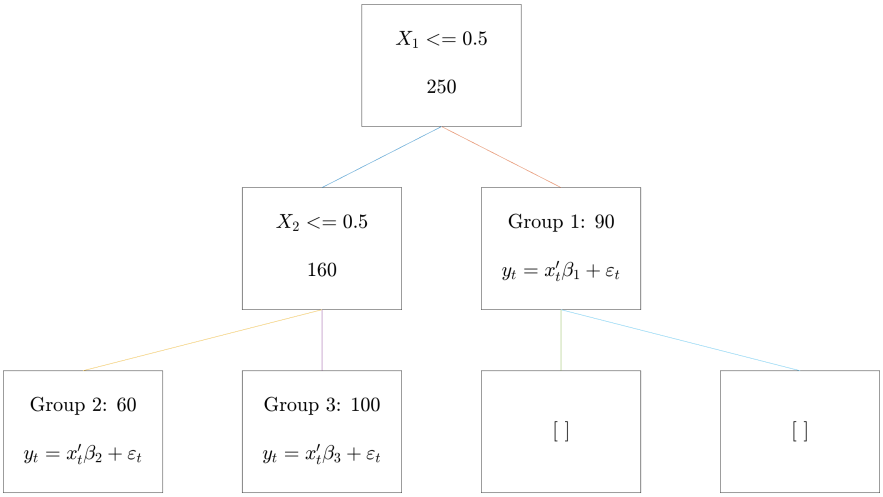


Figure 1: Example of a Regression Tree

Note. This figure plots an example of a regression tree. The tree is consisted of two internal nodes and three terminal nodes.

either 0 or 1. Then the indicator variable $\delta_{g,k} \in \delta_g$ for each predictor follows a Bernoulli distribution with probability p_g, indicating the importance of each predictor variable. If $\delta_{g,k} = 1$, the prior distribution of $\beta_{g,k}$ is considered to follow a slab density π_{slab}. The precise form of the slab density will be introduced later in Section 4. In contrast, all other $\beta_{g,k}$'s where $\delta_{g,k} = 0$ holds are assumed to be included in the spike density $\prod_{k:\delta_k=0} \pi_{spike}$, which has all its mass at zero. Finally, the probability p_g of each group that governs the Bernoulli distribution of $\delta_{g,k}$ is assumed to follow a Beta prior distribution.

Therefore, the data generating process of a Bayesian CART model with Dirac spike-and-slab prior (B-DART) is as follows. The response variable is modeled by the function $\mathcal{L}(\cdot)$ such that

$$\begin{aligned} y_t &= \left[\sum_{g=1}^{G} \mathbb{1}(d_t = g) x_t' \beta_g \right] + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0, \sigma_g^2) \\ &= \mathcal{L}(x_t, \Theta | \mathcal{T}, i^x, i^q) + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0, \sigma_g^2) \end{aligned} \tag{3}$$

where i^x is the index of the selected predictor variables, i^q is the index of splitting thresholds conditional on i^x, and g is the group index of the final terminal nodes.

Moreover, the function $\mathcal{L}(x_t | \beta, \mathcal{T}, i^x, i^q)$ is determined by the following priors of the

Bayesian CART model:

$$\mathcal{T} \sim p_{SPLIT} = \frac{1}{\kappa(1+d_\eta)^\rho}$$

where d_η is the depth of the node η (the number of splits above the ηth node), and the prior hyper parameters are κ and ρ. Also,

$$i^x, i^q \sim p_{rule}$$

where a specific p_{rule} determines i^x and i^q. Moreover, i^x and i^q each follows

$$i^x \sim Multinomial(1, K, \pi_{x_1}, \cdots, \pi_{x_K}), \text{ where } \pi_x = \frac{1}{K}$$
$$i^q \sim Uniform(\min\{i^x\}, \max\{i^x\})$$

where K denotes the number of predictor variables and $\min(i^x)$ and $\max(i^x)$ each denotes the minimum and maximum value of the selected predictor variable.

Once the tree is formed, the parameters of each group are determined by the following priors of the Dirac variable selection model. First, the coefficient vector β_g follows a Dirac spike-and-slab hierarchical prior such that

$$\beta_g|\sigma_g^2, \delta \sim \pi(\beta_g|\sigma_g^2, \delta) = \pi_{slab}(\beta_\delta) \cdot \prod_{k:\delta_k=0} \pi_{spike}(\beta_{g,k})$$

where

$$\begin{cases} \beta_{g,\delta} \sim \mathcal{N}(\beta_{0,\delta}, \sigma_g^2 \cdot B_{0,\delta}) \\ \beta_{g,-\delta} \sim \text{density function with all mass at zero} \end{cases}$$

Note that $\beta_{g,\delta}$ denotes all $\beta_{g,k}$'s where $\delta_{g,k} = 1$, while $\beta_{g,-\delta}$ denotes all $\beta_{g,k}$'s where $\delta_{g,k} = 0$. Each β_g is dependent on the latent variables δ_g, while the value of δ_g is dependent on the probability p_g, which follows the prior distribution as below.

$$\delta_{g,k} \sim Bernoulli(p)$$
$$p_g \sim Beta(a_0, c_0)$$

Finally, σ_g^2 follows an Inverse Gamma distribution such that

$$\sigma_g^2 \sim InverseGamma(\frac{\nu}{2}, \frac{\nu\lambda}{2})$$

Given the DGP above, we find the best tree among the candidate trees using the MH algorithm. Therefore, the main algorithm of our model is given as **Algorithm 1**

Algorithm 1: Main Algorithm of the B-DART Model

1. Set $j = 1$. Generate an initial tree, denoted as $\mathcal{T}_{new}$ by fixing the initial p_{move} as the Grow rule. Let $\mathcal{T}_{new} = \mathcal{T}^{(1)}$.
   - When applying the Grow rule, Algorithm 2 is used to choose the predictor variable and threshold value among the observed data of that variable.
2. If $j > 1$, i.e. if it is not the first iteration, then randomly choose a move rule. Regarding the chosen rule, generate a tree, denoted $\mathcal{T}_{new}$.
   - If the Grow rule is chosen, then apply Algorithm 2 to choose the predictor variable and threshold value.
   - If the Prune rule is chosen, eliminate the terminal nodes below the chosen internal node.
   - If the Change rule is chosen, apply Algorithm 2 again to the selected internal node.
   - If the Swap rule is chosen, swap the splitting rule of the parent and child node and update the tree regarding such new splitting rule.
3. For each terminal node of the generated tree, apply Algorithm 3 of Dirac variable selection and estimate $\hat{\delta}_g$ for each group. Let $\delta_{g,k} = 1$ if $\hat{\delta}_g > 0.5$ and $\delta_{g,k} = 0$ otherwise.
4. Compute the MH ratio of the given tree. If a randomly sampled u from the uniform distribution is smaller than the MH ratio, then accept the tree and denote it as $\mathcal{T}_{prev}$. Also, let $\mathcal{T}^{(j)} = \mathcal{T}_{prev}$ If not, reject the tree, i.e. maintain the tree to be $\mathcal{T}_{new}$ and do not save it.
5. Set $j = j + 1$ and return to Step 2 until $j \leq n$.

above, which is a detailed version of **Algorithm 4**. In sections 3, 4 and 5, the details and the precise steps of the algorithms at each blocks will be introduced.

3 Bayesian CART Model Block: Tree Prior

Note that the prior probability distribution of our model is as follows.

$$\pi(\Theta, \mathcal{T}) = \pi(\Theta|\mathcal{T}) \cdot \pi(\mathcal{T}) \tag{4}$$

In this section, we introduce the details of the tree prior, $\pi(\mathcal{T})$. The tree prior is consists of the tree structure prior and the splitting rule prior. Specifically, $\pi(\mathcal{T})$ is determined by the following recursive process stated in **Algorithm 2** below.

We first assign the size and structure of the tree and then specify the splitting rules for each internal node of the tree. Therefore, we will first explain the size and structure of the tree, followed by the splitting rule.

Algorithm 2: Tree Prior

1. Set the initial tree $\mathcal{T}$ as the simplest form with a single root node where the index of the node is denoted as $\eta = 1$.
2. Split the terminal node of index η with probability $p_{split}(\eta, \mathcal{T})$.
3. If the node is split, randomly assign the splitting rule with probability p_{rule}.
4. Assign the left and right child nodes to the split node.
5. Apply Step 2 to Step 4 to each newly created tree.

3.1 Tree Size and Structure, p_{split}

The prior for the size and structure of a binary tree is a function of η, the index of each node. Then the probability p_{split} is as follows.

$$p_{split} = \frac{1}{\kappa(1 + d_\eta)^\rho} \tag{5}$$

where d_η is the depth of the node η and the hyper-parameters of the split prior is restricted such that $\kappa < 1$ and $\rho \geq 0$.[9] κ is the global splitting hyper-parameter that determines the overall size and structure of the tree, while ρ is the local splitting hyper-parameter that governs the possibility of splitting each terminal node. One can easily observe that the p_{split} is a decreasing function of d_η, i.e. the probability of splitting branches decreases as the tree grows deeper. Also, as the local splitting hyper-parameter ρ increases, p_{split} decreases more rapidly, making deeper nodes less likely to split.[10]

By combining all p_{splits} values regarding the whole tree with η nodes, the joint prior density is given by

$$\pi(\mathcal{T}) = \prod_{\eta} p_{split}(\eta) \tag{6}$$

where $\mathcal{T}$ is the generated tree structure.

3.2 Splitting Rule, p_{rule}

To apply the splitting rule, we must select a specific type of rule. Following Chipman et al. (1998), we apply the *uniform specification* where a predictor variable X_k is chosen uniformly from all available predictors, and a splitting threshold value s is also chosen uniformly from the observed values of X_k. Such specification is natural since it assumes

[9] For example, the depth of the root node is $d_\eta = 1$. Also, the depth of the two child nodes of the root node is $d_\eta = 2$. Further, the depth of each child node of the parent nodes at $d_\eta = 2$ is $d_\eta = 3$.

[10] Refer to Figure 3 in Chipman et al. (1998) for a detailed explanation of the tree structure regarding the combination of each hyperparameter (κ and ρ).

that each predictor variable and its threshold values are equally likely to be effective.[11] Particularly, we apply uniform specification for the splitting threshold value s so that the threshold value depends on the range of the chosen predictor variable. [12] The reason of using such specification for the threshold value is to consider the fact that the range of predictor variables depends on their scale.

We first select π_X, the probability of each X_k being used as the splitting rule, with uniform probability from the uniform specification.

$$i^x \sim Multinomial(1, K, \pi_{x_1}, \cdots, \pi_{x_K}) \text{ where } \pi_{x_1} = \cdots = \pi_{x_K} = \frac{1}{K}$$

After selecting the set of predictors, the splitting threshold s is selected from the following uniform distribution based on the observed values of the chosen predictor variable.

$$s \sim Uniform(\min\{i^x\}, \max\{i^x\})$$

By this, we assume that the threshold values are equally likely to be effective, depending on the range of the observed values.

4 Dirac Variable Selection Block: Model Prior

Given each group (or terminal node) of the generated tree, the parameters β_g and σ_g for each group are estimated using the Dirac variable selection method. In other words, the Dirac spike-and-slab prior is imposed to determine the latent variable δ_g that governs the importance of each $\beta_{g,k} \in \beta_g$.

In particular, we specify the prior slab distribution to be a g-slab, suggested by Zellner (1986).[13] As suggested by Malsiner-Walli and Wagner (2011), the advantages of using such g-slab are that it simplifies the form of the marginal likelihood and significantly lowers the computational burden since there is no need to derive the determinant terms of the prior and posterior variance of β_g.

When setting the prior of $\beta_{g,\delta}$, i.e. the $\beta_{g,k}$'s where $\delta_{g,k} = 1$, to be a g-slab, the mean and variance of the slab prior becomes $\beta_{0,\delta} = 0$ and $B_{0,\delta} = g \cdot (X'_\delta X_\delta)^{-1}$. That is, the

[11] Another candidate is the *nonuniform specification* where other rules apart from the uniform specification are applied. One good example is Linero (2018) where sparsity is applied in the selection of predictor variables.

[12] Note that applying other rules, such as selecting both uniform specifications for the predictor variable and threshold value of that variable, leads to similar outcomes.

[13] Note that one can also use the f-slab which corresponds to the fractional prior suggested by O'Hagan (1995), although there is no significance difference in the results.

prior distribution is given by

$$\beta_{g,\delta} \sim \mathcal{N}(0, \sigma_g^2 \cdot g \cdot (X_\delta' X_\delta)^{-1}) \tag{7}$$

where X_δ is the design matrix consisting of columns of X_g corresponding to $\delta_{g,k} = 1$. Also, $\beta_{g,-\delta}$ is set to zero, i.e. $\beta_{g,k}$'s where $\delta_{g,k} = 0$ are set to zero. Finally, for the sake of simplicity, we assume that the data is centered, i.e. both Y_g and X_g are demeaned. Then given δ_g, the conditional marginal likelihood can be derived analytically as follows:

$$f(\mathbf{y_{g,c}}|X_g, \delta) = (\pi)^{-\frac{(T_g-1)}{2}} \cdot T_g^{-\frac{1}{2}} \cdot (\nu\lambda)^{\frac{\nu}{2}} \cdot (g+1)^{-\frac{K}{2}} \cdot \frac{\Gamma(\frac{(T_g-1)+\nu}{2})}{\Gamma(\frac{\nu}{2})} \cdot (A_\delta + \nu\lambda)^{-\frac{(T_g-1)+\nu}{2}} \tag{8}$$

where

$$A_\delta = \mathbf{y_{g,c}}'\mathbf{y_{g,c}} - \mathbf{y_{g,c}}' X_\delta B_{1,\delta} X_\delta' \mathbf{y_{g,c}}$$
$$B_{1,\delta} = \frac{g}{g+1}(X_\delta' X_\delta)^{-1}$$

Details of deriving the above expression and proofs of the equivalence between using centered and decentered data are provided in Appendix A.

With the marginalization above, it is possible to sample the posterior distribution of the parameter set $\Theta_g = \{\beta_g, \sigma_g^2, p_g, \delta_g\}$ for each group following **Algorithm 3**. It is important to stress that this sampling algorithm is based on the method of composition (MoC) and is not a standard Gibbs sampler, i.e.

$$\pi(\beta_g, \sigma_g^2, \delta_g|\mathbf{y}_{g,c}, p_g) = \pi(\beta_g|\mathbf{y}_{g,c}, p_g, \sigma_g^2, \delta_g) \cdot \pi(\sigma_g^2|\mathbf{y}_{g,c}, p_g, \delta_g) \cdot \pi(\delta_g|\mathbf{y}_{g,c}, p_g) \tag{9}$$

This is because the Gibbs-sampling algorithm is not irreducible and does not converge to the joint posterior in the presence of a Dirac specification for the spike distribution. A detailed proof of the failure of the Gibbs sampler will be shown in Appendix C.

First, set the initial values of p_g and δ_g for a given group. Since $\pi(\sigma_g^2|\mathbf{y}_{g,c}, p_g, \delta_g)$ and $\pi(\delta_g|\mathbf{y}_{g,c}, p_g)$ are not the full-conditional distributions, they must be derived analytically. Specifically, using the fact that $\delta_{g,k}$ is a discrete variable such that $\delta_{g,k} = 0$ or $\delta_{g,k} = 1$, one can sample δ_g using the ratio of the analytical conditional marginal likelihood in equation (8). If a randomly sampled u from the uniform distribution is smaller than the ratio, then $\delta_{g,k}$ is set to be 1. Otherwise, $\delta_{g,k} = 0$ holds. Now given $\delta_g^{(j)}$, one can sample $\sigma_g^{(j)}|\delta_g^{(j)}$ from $\mathcal{IG}((T_g+\nu)/2, (A_\delta+\nu\lambda)/2)$. Finally, given $\sigma_g^{(j)}$ and $\delta_g^{(j)}$, one can sample $\beta_g^{(j)}|\sigma_g^{(j)}, \delta_g^{(j)}$ from $\mathcal{N}(\beta_1, \sigma_g^2 \cdot B_1)$. It is important to note that only $\delta_g^{(j)}$ depends on its previous iteration $\delta_g^{(j-1)}$, showing that **Algorithm 3** is indeed not a Gibbs-sampler.

Algorithm 3: Dirac Variable Selection

1. Set the initial values of p_g, δ_g and let $j = 1$.
2. Sample each element of δ_g, i.e. $\delta_{g,k}$, from

$$Pr[\delta_{g,k} = 1 | \delta_{g,-k}, \mathbf{y_{g,c}}] = \frac{f(\mathbf{y}_{g,c}|\delta_k = 1, \delta_{-k}) \cdot p_g}{f(\mathbf{y}_{g,c}|\delta_{g,k} = 1, \delta_{g,-k}) \cdot p_g + f(\mathbf{y}_{g,c}|\delta_{g,k} = 0, \delta_{g,-k}) \cdot (1 - p_g)}$$

   If $u \sim Unif(0, 1)$ is smaller than the above probability, set $\delta_{g,k} = 1$. Otherwise, set $\delta_{g,k} = 0$.
3. Sample $\sigma_g^{(j)}$ from $\mathcal{IG}(\frac{\alpha_1}{2}, \frac{\delta_1}{2})$ where

$$\alpha_1 = \nu + T_g \text{ and } \delta_1 = \nu \cdot \lambda + A_\delta$$

   with

$$A_\delta = \mathbf{y_{g,c}}' \mathbf{y_{g,c}} - \mathbf{y_{g,c}}' X_\delta B_{1,\delta} X_\delta' \mathbf{y_{g,c}}$$

4. Sample $\beta_{g,\delta}^{(j)} | \sigma_g^{(j)}, \delta_g^{(j)}$ from $\mathcal{N}(\beta_{1,\delta}, \sigma_g^2 B_{1,\delta})$ where

$$B_{1,\delta} = g/(g+1)(X_\delta' X_\delta)^{-1} \text{ and } \beta_{1,\delta} = B_{1,\delta} X_\delta' \mathbf{y}_{g,c}$$

   Also set $\beta_{g,-\delta} = 0$.
5. Sample p_g from

$$p_g | \delta_g \sim Beta(a_0 + K_1, c_0 + K_0)$$

   where $K_0 = \{\#$ of $\delta_k = 0\}$ and $K_1 = \{\#$ of $\delta_k = 1\}$.
6. Set $j = j + 1$ and return to Step 2 if $j \leq N$.

The detailed derivation of the posterior distributions of each parameter and steps of the Dirac variable selection method is explained in Appendix B.

5 Metropolis-Hastings Block: Selecting Posterior

Given the two priors, i.e. the tree prior and model prior in Section 3 and 4, a single tree can be generated. Then by recursively iterating the application of the priors randomly and generating numerous trees, one can obtain the posterior of the tree $\mathcal{T}$. When exploring the posterior of $\mathcal{T}$, i.e.

$$\pi(\mathcal{T}|X, Y) \propto f(Y|X, \mathcal{T}) \times \pi(\mathcal{T}) \tag{10}$$

we use the *Metropolis-Hastings (MH) algorithm* to ensure that the Markov Chain of generated trees

$$\mathcal{T}^1, \mathcal{T}^2, \mathcal{T}^3, \cdots$$

converges to the posterior distribution $\pi(\mathcal{T}|X,Y)$. Since the MH algorithm searches through possible trees until it reaches the true tree structure, we refer to it as the MH search algorithm. During the MH algorithm, the marginal likelihood is derived conditionally on the estimated $\hat{\delta}$ of each generated tree. Here we assume that $\hat{\delta}_k = 0$ if $\hat{\delta}_k < \hat{p}_g$, while $\hat{\delta}_k = 1$ if $\hat{\delta}_k \geq \hat{p}_g$. $\hat{p}_g$ is the estimated probability governing $\hat{\delta}$ for that group.[14] Then the tree with the highest conditional marginal likelihood is selected as our final model.

Starting with an initial tree, denoted as $\mathcal{T}^1$, a candidate tree structure can be generated from a probability distribution called the *proposal distribution*. In other words, we sample $\mathcal{T}^{(j)}$ s.t.

$$\mathcal{T}^{(j)} \sim q(\mathcal{T}^{(j-1)}, \mathcal{T}^{(j)}) \tag{11}$$

Note that the proposal distribution generating $\mathcal{T}^{(j)}$ follows four rules:

- **Grow**: Randomly select a terminal node and split it into two new nodes using the splitting rule from **Algorithm 2**.
- **Prune**: Randomly select a parent of two terminal nodes and convert it into a terminal node by eliminating the nodes below it.
- **Change**: Randomly select an internal node and randomly reassign a new splitting rule from **Algorithm 2**.
- **Swap**: Randomly select a parent-child pair that are both internal nodes and swap their splitting rules.

These rules are called *moving rules*, denoted as p_{move}, where each rule is assumed to be uniform, i.e. the probability of each generating rule is $1/m$.

Once $\mathcal{T}^{(j)}$ is generated at the j-th iteration, the *MH ratio* can be computed as follows:

$$\begin{aligned}
\alpha(\mathcal{T}^{(j-1)}, \mathcal{T}^{(j)}) &= \min\left\{ \frac{P(\mathcal{T}^{(j)}|X,Y) \cdot q(\mathcal{T}^{(j-1)}, \mathcal{T}^{(j)})}{P(\mathcal{T}^{(j-1)}|X,Y) \cdot q(\mathcal{T}^{(j)}, \mathcal{T}^{(j-1)})}, 1 \right\} \\
&= \min\left\{ \frac{P(\mathcal{T}^{(j)}|X,Y) \cdot q(\mathcal{T}^{(j-1)}|\mathcal{T}^{(j)})}{P(\mathcal{T}^{(j-1)}|X,Y) \cdot q(\mathcal{T}^{(j)}|\mathcal{T}^{(j-1)})}, 1 \right\} \\
&= \min\left\{ \underbrace{\frac{P(Y|X,\mathcal{T}^{(j)})}{P(Y|X,\mathcal{T}^{(j-1)})}}_{\text{(i)}} \cdot \underbrace{\frac{P(\mathcal{T}^{(j)})}{P(\mathcal{T}^{(j-1)})}}_{\text{(ii)}} \cdot \underbrace{\frac{q(\mathcal{T}^{(j-1)}|\mathcal{T}^{(j)})}{q(\mathcal{T}^{(j)}|\mathcal{T}^{(j-1)})}}_{\text{(iii)}}, 1 \right\}
\end{aligned}$$

[14]Note that the criteria of $\hat{\delta}_k < \hat{p}_g$ is chosen so that predictor variables with an estimated average δ_g higher than the estimated inclusion probability $\hat{p}_g$ is set to one. This ensures that the selection of δ is group-specific. Alternatively, one could use a fixed threshold, e.g. 0.5, but this may occasionally suppress $\hat{\delta}_k$ to zero too frequently, leading to model mis-specification and inaccurate forecasts.

Algorithm 4: Metropolis-Hastings Search Algorithm

1. Fix the initial p_{move} as the growing rule. Grow an initial tree, denoted as $\mathcal{T}^0$ and set $j = 1$.
2. Randomly select a move rule p_{move} with probability $1/m$. Generate $\mathcal{T}^{(j)}$ from the proposal $q(\mathcal{T}^{(j-1)}, \mathcal{T}^{(j)})$ regarding the type of the randomly selected move rule.
   - For the grow or change rule, **Algorithm 2** for the tree prior is applied.
3. Once $\mathcal{T}^{(j)}$ is obtained, one can derive the MH rate as below.

$$\alpha(\mathcal{T}^{(j-1)}, \mathcal{T}^{(j)}) = \min \left\{ \frac{P(Y|X, \mathcal{T}^{(j)})}{P(Y|X, \mathcal{T}^{(j-1)})} \cdot \frac{P(\mathcal{T}^{(j)})}{P(\mathcal{T}^{(j-1)})} \cdot \frac{q(\mathcal{T}^{(j-1)}|\mathcal{T}^{(j)})}{q(\mathcal{T}^{(j)}|\mathcal{T}^{(j-1)})}, 1 \right\}$$

   Note that the likelihood ratio, prior ratio and proposal ratio differs by the selected move rule.
4. Sample $u^{(j)}$ from $Unif(0, 1)$. Then compare this with the MH rate.

$$\begin{cases} \text{accept } \mathcal{T}^{(j)} & \text{if } u^{(j)} < \alpha(\mathcal{T}^{(j-1)}, \mathcal{T}^{(j)}) \\ \text{reject } \mathcal{T}^{(j)} & \text{otherwise} \end{cases}$$

   If accepted, save $\mathcal{T}^{(j)}$ as $\mathcal{T}^{(j)}$. If rejected, save $\mathcal{T}^{(j-1)}$ as $\mathcal{T}^{(j)}$.
5. Set $j = j + 1$ and return to Step 2 if $j \leq n$.

where each term in the last equation can be interpreted as follows.

(i) The ratio of the likelihood function of the new and previous tree.
(ii) The ratio of the prior density of the new and previous tree.
(iii) The ratio between the probability of "moving" to the previous tree given the new tree and probability of "moving" to the new tree given the previous tree.

Note that while the first two terms (i) and (ii) can be easily obtained when generating a tree, the last term differs by the randomly chosen move rule p_{move}. Also, the numerator and denominator of (iii) each indicate the probability of pruning and growing a tree. Then depending on the randomly selected move rule, different equations for the MH ratio will be applied. Details of deriving the MH ratio for each move rule are explained in Appendix D.

The detailed process of applying the Metropolis-Hastings search algorithm is illustrated in **Algorithm 4** above. Note that the tree structure prior and stopping rule can be applied at this stage. For example, suppose the stopping rule is set as 30, i.e. each group must have at least 30 observations. Once any divided group has less than 30 observations, then the MH rate is set as 0 so that the tree at that iteration ($\mathcal{T}^{(j)}$) is

unequivocally rejected.

Once all trees are generated by **Algorithm 4**, following the spirit of the Bayesian model comparison method, the final tree among the generated tree by the MH algorithm is selected by comparing the conditional marginal likelihood at each splitting step. As mentioned above, the assumption of conjugate priors allows us to compute the conditional marginal likelihood function analytically as equation (8).

6 Empirical Application

Again, we underscore that considering time-varying relationship within variable selection by the proposed B-DART model does not inherently guarantee improved forecasting performance. This is because forecasting instability can arise from more than just the structural breaks in parameters. Nevertheless, we argue that, empirically in most macroeconomic variables, the application of time-varying variable selection leads to improvement of the forecasting accuracy. This is evaluated by comparing the out-of-sample forecast accuracy of the B-DART model against other candidate forecast models.

6.1 Data and Factor Construction

We use the FRED-QD data set from McCracken and Ng (2020) consisting of 221 variables of quarterly frequency, ranging from 1967 Q1 to 2024 Q2.[15] This data set is an extension of Stock and Watson (2012), where each variable is categorized into 14 group classifications and transformed using specific methods to ensure they are all stationary, i.e. $I(0)$. Using the direct forecasting method, we estimate 1- to 8-quarter-ahead forecasts for 8 macroeconomic variables: the PCE, CPI, PPI price indices, Crude Oil price, federal funds effective rate (FFE), 10-year maturity interest rate, real GDP and unemployment rate.

A crucial issue for researchers to consider is that macroeconomic variables in big datasets often exhibit similar characteristics, potentially leading to collinearity in regression models. One commonly used approach to avoid this when forecasting macroeconomic variables under high-dimensional predictor space environments are factor based methods (e.g. Stock and Watson (2006), Bai and Ng (2008) and Kim and Swanson (2018)). Particularly, in order to construct representative group principal components (PCs), we apply principal component analysis (PCA) to each of the 14 categories. As the

[15] Although the original FRED-QD data set consists of 246 variables spanning from 1959 Q1 to 2024 Q1, we only use a subsample from 1967 Q1 so that the data is balanced.

Table 1: Result of PCA for Each Category Groups

Category	Abbrev.	Variables	0.4	0.5	0.7	0.8	1st Eigenvalue	2nd Eigenvalue	3rd Eigenvalue
NIPA	NIPA	22	2	2	5	7	0.3584	0.1461	0.0923
Industrial Production	IP	16	1	2	3	4	0.4665	0.1482	0.136
Employment	EMP	49	2	2	5	9	0.3334	0.1756	0.0803
Housing	HOUSE	11	1	1	3	4	0.597	0.0939	0.0909
Inventories, Orders, and Sales	INV	6	1	2	2	3	0.4723	0.3102	0.104
Prices	PRICE	46	2	3	8	11	0.3545	0.1038	0.0653
Earnings and Productivity	EAPROD	11	2	2	3	4	0.3283	0.2919	0.1572
Interest Rates	INT	15	1	2	3	4	0.4429	0.2242	0.1203
Money and Credit	M&C	13	3	3	5	7	0.2264	0.1487	0.1299
Household Balance Sheets	HBAL	9	1	2	3	3	0.4572	0.2372	0.1232
Exchange Rates	EXCH	4	1	1	2	3	0.5249	0.2524	0.1403
Stock Market	STOCK	5	1	1	2	3	0.5848	0.2108	0.1096
Non-household Balance Sheets	NHBAL	13	2	2	4	5	0.3279	0.215	0.1064
Others (Sentiment)	SENTI	1					1 (Use original variable without PCA)		
Sum		221	21	26	49	68			

(1) 0.4, 0.5, 0.7, and 0.8 implies the total variation of each category explained by the PCs.

(2) The last category group, denoted as others, is consisted only with the Consumer Sentiment survey from the University of Michigan. Thus, we use the original variable itself without applying PCA.

estimation of PCs is sensitive to the existence of outliers, we eliminate 84 observations for the 221 predictor variables before applying PCA.[16]

By construction, the number of potential PCs is equivalent with the number of variables used. Then by comparing the eigenvalues, which represents the variation explained by the corresponding PC, one can determine the representativeness of each PC. For instance, a PC with a low eigenvalue ratio is interpreted as inadequately representing the variables. Because our objective of PCA is to identify the PCs that best represent each category group, this evaluation of variation is extremely important.

Table 1 above illustrates the list of category groups, the number of variables consisting each group, the number of PCs selected regarding the sum of eigenvalue ratios, and the eigenvalue ratio corresponding to the top 3 PCs. The criterion for selecting the number of PCs is the sum of eigenvalue ratios, representing the total variation of the predictor variables explained by the PCs. We selected 21, 26, 49, and 68 PCs based on each criteria. Note that the number of selected PCs may vary for out-of-sample forecasting, as shown in **Appendix E**.

To resolve collinearity, which is a crucial issue when applying a big dataset, we apply the PCA method to focus on obtaining the PCs that best represents each category group. In particular, we use the PCs that explain at least 40 percent of the variation within each group. This specific criteria is chosen since other permissive criteria, such as 0.5 or higher, leads to collinearity issues, making the use of linear regression infeasible as the matrix of the PCs is close to singular. Note that the category group labeled "Other" consists of a single variable, the consumer sentiment index. Thus, it is unnecessary to apply PCA to this group. Now the 20 PCs and 1 original variable are used as the predictor variables in our candidate forecast models.

6.2 Forecast Model

We examine the forecast accuracy by comparing four different models of two types: (1) B-DART ADL(4,1), (2) Dirac ADL(4,1) and (3) AR(1). These three models are defined as follows:

(1) $h-$step ahead B-DART ADL(4,1)

$$y_{t+h} = \left[\sum_{g=1}^{G} \mathbb{1}(d_t = g) x_t' \beta_{g,\delta} \right] + \varepsilon_{t+h},\ \varepsilon_{t+h} \sim \mathcal{N}(0, \sigma_g)$$

[16]The outliers are defined to be observations that deviate from the median of each variable by more than ten interquantile ranges.

where x_t is consisted of four lags of y_t and one lag of the PC f_t, while β_δ is the corresponding coefficient terms obtained by Dirac variable selection.

(2) $h-$step ahead Dirac ADL(4,1)

$$y_{t+h} = x_t'\beta_\delta + \varepsilon_{t+h}$$

where x_t is the selected terms among four lags of y_t and one lag of the PC f_t, while β_δ is the corresponding coefficient terms by Dirac variable selection.

(3) $h-$step ahead B-CART ADL(4,1)

$$y_{t+h} = \left[\sum_{g=1}^{G} \mathbb{1}(d_t = g)x_t'\beta_g\right] + \varepsilon_{t+h},\ \varepsilon_{t+h}$$

(4) $h-$step ahead AR(1) with constant

$$y_{t+h} = \mu + \rho y_t + \varepsilon_{t+h} \sim \mathcal{N}(0, \sigma_g)$$

where now β_g is simply the coefficient terms of all predictor variables of each estimated subsample.

The first three models (ADL types) consider the possibility that the predictor variables may affect the response variable and that the response variable may exhibit autoregressive tendencies. Specifically, we use the ADL(4,1) model to estimate $h = 1, \cdots, 8$ forecast horizons. The set of potential predictor variables include four lags of the response variable y_t and one lag of the estimated PCs f_{t-1}. The AR(1) model serves as a benchmark for determining whether the ADL(4,1) model is overfitted, as it only includes one lag of the response variable.

6.3 Evaluation Measures and Hyper-parameters

The out-of-sample period spans 40 periods, from 2014 Q2 to 2024 Q1. Recursive regression is performed to obtain the root-mean-squared error (RMSE) that are used as the criteria of evaluating the forecasting accuracy of each forecast model.[17]

[17] An alternative method of out-of-sample forecasting is to use the rolling-window method. However, we choose recursive regression to ensure sufficient sample sizes for constructing regression trees, which improves the accuracy of group division.

Point forecast accuracy of the forecast model is evaluated using RMSE:

$$RMSE = \left(\frac{1}{OSS} \sum_{t=1}^{OSS} (y_t - \hat{y}_t)^2 \right)^{\frac{1}{2}} \tag{12}$$

where OSS is the size of the out-of-sample (40 periods in our case), y_t is the realized value of the response variable at time t, and $\hat{y}_t$ is the predicted value of y_t at time t.

Before presenting the detailed results, the specific values of hyper-parameters used are illustrated as follows. First, the tree prior $P(\mathcal{T})$ is set by using $\kappa = 0.5$ and $\rho = 0.5$ so that the maximum depth of the tree is around 10. The variance parameter g of the g-slab prior is set as T_g^2.[18] Hyper-parameters regarding the inclusion probability of the Dirac variable selection is set as $a_0 = 5$ and $c_0 = 5$. Also, when computing the conditional marginal likelihood, we set $\hat{\delta}_k = 0$ if $\hat{\delta}_k < \hat{p}_g$ and $\hat{\delta}_k = 1$ otherwise. An additional stopping rule ensures that each group has at least 30 observations.

We tune the hyper-parameters of the prior on σ_g as the structure of the generated trees depends greatly on that value. Specifically, following Chipman et al. (2012), we calibrate σ_g^2 using the data-based estimate $\hat{\sigma}_g^2$. For that, we set $\nu = 5$ and take $\hat{\sigma}_g^2$ as $std(Y)$. Then we tune λ as a proportion of $std(Y)$. Details of the tuning process and selected hyper-parameters are reported in section 2 of the supplementary appendix.[19]

6.4 Estimation Results

In this section, the out-of-sample forecasting accuracy for the four candidate forecast models are presented. Moreover, by dividing the out-of-sample period into pre- and post-COVID-19 phases and examining the RMSEs, we show that accounting for the time-varying relationships among variables in the variable selection indeed improves forecast accuracy. To further unravel the forecasting performance, the selected variables for the B-DART and Dirac models over the out-of-sample periods are reported. Finally, we focus on the B-DART model to examine its average groups and the predictor variables used as the criteria of forming the optimal tree structure for each macroeconomic variable.

[18] Although Fernandez et al. (2001) suggest $g = T$ is enough, we use $g = T_g^2$ to guarantee the convergence of the denominator in equation (B.3). This is because when splitting the full sample with $T = 250$ into subgroups regarding the predictor variables, one can face a group with sample size at least $T_g = 30$. Then setting $g = T_g$ may lead to a biased δ_g value.

[19] Note that computation was done by a computer with an Intel i7 13700K (16-core) CPU and 64GB of DDR5 RAM, together with parallel computing that utilized 14 out of the 16 cores. The process took an average of 45.86 hours per macroeconomic variable for up to 8 forecast horizons, with 5 rounds of hyperparameter tuning and out-of-sample forecasting for 40 periods.

Table 2: Comparison of Relative RMSE for Full Out-of-samples

	PCE			CPI			PPI			Crude Oil		
H	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART
1	**0.9157**	0.9689	1.0995	1.0291	1.0474	1.1693	**0.9376**	0.9433	1.0656	1.0256	**0.9802**	1.2600
2	**0.9954**	1.0248	1.0328	**0.9707**	0.9819	1.1203	1.0188	1.0017	1.0358	**0.9264**	0.9758	1.1112
3	1.0551	1.0043	1.2695	1.0381	1.0098	1.3438	1.0702	**0.9977**	1.1281	1.0167	1.0027	1.1570
4	1.0134	1.0042	1.4407	1.0820	1.0095	1.4110	1.0480	**0.9931**	1.1719	0.9922	**0.9876**	1.1384
5	1.0083	**0.9927**	1.2733	**0.9830**	0.9998	1.1127	**0.9744**	0.9918	1.0451	0.9763	**0.9762**	1.0324
6	**0.9535**	0.9617	0.9974	0.9956	**0.9769**	1.0640	**0.9930**	1.0009	1.0517	1.0086	**0.9970**	1.0209
7	1.0219	**0.9903**	0.9908	1.0379	**0.9979**	1.2556	**0.9318**	0.9948	1.0310	1.0038	**0.9905**	1.0432
8	1.1140	1.0214	1.2222	1.0159	1.0264	1.4145	1.0005	**0.9839**	1.3730	1.0008	1.0149	1.1542
	B-DART = AR(1) ≻ Dirac ≻ B-CART			AR(1) ≻ B-DART = Dirac ≻ B-CART			B-DART ≻ Dirac ≻ AR(1) ≻ B-CART			Dirac ≻ B-DART = AR(1) ≻ B-CART		
	FFE			Long INT			Real GDP			Unemp		
H	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART
1	1.2633	1.2493	4.7426	1.0758	1.0297	1.5714	**0.9795**	1.4545	1.5894	**0.7047**	0.9131	0.7690
2	1.0071	1.0685	1.8026	1.0700	**0.9929**	1.3005	1.0741	1.0047	1.3316	**0.9133**	0.9207	1.0330
3	1.0005	1.1391	1.6068	1.2848	1.0046	1.0004	1.0431	**0.9919**	1.0396	1.0504	1.0560	1.0821
4	1.3479	1.0411	2.3198	1.0991	1.0010	1.6222	**0.8683**	0.9962	1.0927	**0.9451**	0.9874	2.0104
5	1.1737	1.0465	2.0343	1.1815	**0.9614**	1.1618	1.0251	**0.9879**	1.1566	**0.9276**	0.9944	1.0610
6	**0.9588**	1.0620	1.2073	**0.9878**	0.9883	1.4678	**0.9610**	1.0043	1.2753	**0.9787**	0.9968	1.2612
7	**0.9633**	1.0285	1.1908	1.0515	1.0058	1.1631	1.0431	1.0160	1.1384	1.0033	**0.9874**	1.2306
8	**0.8539**	0.9302	1.1468	1.5468	1.0173	1.2954	**0.9778**	0.9838	1.1457	1.0401	**0.9819**	1.0580
	AR(1) ≻ B-DART ≻ Dirac = B-CART			AR(1) ≻ Dirac ≻ B-DART ≻ B-CART			B-DART ≻ Dirac = AR(1) ≻ B-CART			B-DART ≻ Dirac ≻ AR(1) ≻ B-CART		
	Total # of B-DART 1st = 23, 35.94%						Total # of B-CART 1st = 0, 0.00%					
	Total # of Dirac 1st = 18, 28.13%						Total # of AR(1) 1st = 23, 35.94%					

(1) The numbers are relative RMSEs where the RMSE of AR(1) is normalized as 1.
(2) The bold font values denotes that the model is better than the other three.
(3) The order under each variable was determined by counting the bolded values for each forecast horizon.
(4) The numbers and percentages in the last two rows of each table represent the count of bolded values for each model and their proportion out of the total 64 cases (8 variables × 8 forecast horizons).
(5) The values depicted in the table above are all obtained under the best hyper-parameter for each forecast model. Details of the selected hyper-parameters are reported in the Appendix G.

6.4.1 Comparing Forecast Accuracy for the Full Out-of-sample Periods

Table 2 compares the out-of-sample performance of the four models using relative RMSEs, benchmarked against the AR(1) model. The full out-of-sample period spanning from 2014 Q2 to 2021 Q1 is used to forecast 8 macroeconomic variables. The bolded values in the table indicate the best-performing model for each forecast horizon, with all results obtained using the optimal σ_g^2 hyper-parameters for each model.

Overall, the B-DART model outperforms its counterparts, highlighting the importance of accounting for the time-varying relationships in variable selection to enhance forecasting accuracy. Specifically, out of 64 cases (8 variables ×8 horizons) in total, the B-DART and AR(1) models were selected as the best-performing model in 23 cases (35.94%), followed by the Dirac model in 19 cases (28.13%). The B-CART model was

Table 3: Comparison of Relative RMSE for Pre-COVID-19 Periods

	PCE			CPI			PPI			CrudeOil		
H	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART
1	1.0460	**0.9322**	0.9617	1.0493	**0.9244**	0.9727	0.8987	**0.8618**	0.8702	1.0682	**0.9868**	1.1066
2	1.0365	1.0186	1.2483	**0.9348**	1.0024	1.2090	1.1407	1.0007	1.1857	**0.9039**	0.9712	1.0687
3	1.0662	**0.9932**	1.3471	1.0489	**0.9705**	1.3126	**0.9651**	0.9975	1.2206	**0.9957**	1.0006	1.1265
4	1.0162	**0.9774**	1.2409	**0.9281**	0.9739	1.1913	1.0785	1.0234	1.1321	**0.9397**	0.9781	1.1153
5	1.0134	1.0148	1.1889	**0.9217**	1.0047	1.0286	0.9848	0.9914	**0.9666**	1.0028	**0.9842**	1.0377
6	1.0070	**0.9963**	1.0121	1.0257	**0.9917**	1.0648	**0.9824**	0.9878	1.0598	**0.9857**	0.9935	1.0219
7	**0.7976**	0.9858	0.9619	**0.9085**	0.9912	1.0070	1.0163	0.9925	**0.9662**	**0.9594**	0.9935	1.0096
8	1.1157	1.0338	1.1712	1.0483	1.0469	1.1916	1.0238	**0.9955**	1.0289	**0.9307**	1.0030	0.9907
	Dirac ≻ AR(1) ≻ B-DART ≻ B-CART			B-DART ≻ Dirac ≻ AR(1) ≻ B-CART			B-DART = Dirac = B-CART = AR(1)			B-DART ≻ Dirac ≻ AR(1) = B-CART		
	FFE			Long INT			Real GDP			Unemp		
H	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART
1	1.1900	1.2533	1.7867	1.0478	1.0237	1.5542	1.2508	1.3814	1.3883	1.1454	**0.9565**	1.2815
2	1.0362	1.0104	2.6390	1.1267	**0.9907**	1.5984	1.1756	1.1082	1.2298	**0.9449**	1.1861	1.6044
3	1.5347	1.3735	2.5768	1.3480	**0.9753**	1.1095	1.3050	1.0448	1.2706	1.3693	1.2009	1.2766
4	1.5614	**0.9956**	1.4392	1.1802	**0.9943**	1.2451	1.0605	1.0035	1.3244	1.1180	1.1484	1.3354
5	1.9528	1.2224	2.0536	1.4297	**0.9903**	1.3265	**0.9698**	0.9844	1.3148	**0.9310**	0.9551	1.0361
6	1.1247	1.1417	1.5853	1.0160	1.0008	1.5515	1.0751	**0.9995**	1.1743	1.1641	0.9597	**0.9485**
7	1.1524	**0.8643**	1.0709	1.0407	**0.9897**	1.1842	1.3917	1.0004	1.4428	1.0325	**0.9035**	1.2786
8	**0.9836**	0.9872	1.1699	1.4702	1.0063	1.4008	**0.8338**	1.0223	1.4885	1.2889	**0.8637**	1.4168
	AR(1) ≻ Dirac ≻ B-DART = B-CART			Dirac ≻ AR(1) ≻ B-DART = B-CART			AR(1) ≻ B-DART ≻ Dirac ≻ B-CART			Dirac ≻ B-DART = AR(1) ≻ B-CART		
	Total # of B-DART 1st = 18, 28.13%						Total # of B-CART 1st = 3, 4.69%					
	Total # of Dirac 1st = 22, 34.38%						Total # of AR(1) 1st = 21, 32.81%					

(1) The numbers are relative RMSEs where the RMSE of AR(1) is normalized as 1.
(2) The bold font values denotes that the model is better than the other three.
(3) The order under each variable was determined by counting the bolded values for each forecast horizon.
(4) The numbers and percentages in the last two rows of each table represent the count of bolded values for each model and their proportion out of the total 64 cases (8 variables × 8 forecast horizons).
(5) The values depicted in the table above are all obtained under the best hyper-parameter for each forecast model. Details of the selected hyper-parameters are reported in the Appendix G.

dominated by other models across all cases. While the B-DART model performed better overall, the best performing model varied across each macroeconomic variables. Therefore, it is also necessary to compare the forecast accuracy at the variable level.

Even at the variable level, the B-DART model demonstrated better performance for economic activity indicators (real GDP and unemployment rate) and certain price related variables (PCE and PPI price indices). Focusing first on economic activity indicators, the B-DART model shows sporadic superiority in forecasting 1-, 4-, 6-, 8-quarter-ahead forecasts for the real GDP, while it outperforms other candidates for the short- and medium-term (up to 6 quarters) unemployment rate forecasts. Moving on to the price indicators, the B-DART model exhibits better performance at long-term forecast horizons (5 or more quarters) for the PPI price index and short-term forecast horizons (up to 4 quarters) for the PCE price index.

6.4.2 Comparing the Forecasting Accuracy of Pre- and Post-COVID-19 Out-of-sample Periods

While **Table 2** above summarizes the forecasting performance using the full out-of-sample period (from 2014 Q2 to 2024 Q1), we further divide the out-of-sample period into pre- and post-COVID-19 sub-periods to assess whether the forecast accuracy of the models differ between stable and volatile economic regimes. The COVID-19 pandemic, particularly 2020 Q2, was chosen as the dividing point since it is the most recent period among the 40 out-of-sample periods that has the possibility of altering the data-generating process. In fact, shocks from the pandemic and subsequent geopolitical events, such as the Russia-Ukraine War, have led to the high-inflation era and the altered the relationships between macroeconomic variables. Aggregating the full out-of-sample results might mask theses dynamic changes, making a breakdown to the pre- and post-COVID-19 periods essential to gain clearer insights on how different forecast models adapt to structural shifts of this magnitude.

The forecast accuracy of candidate forecast models using the pre-COVID-19 periods is shown in **Table 3** below. In terms of the overall results, the Dirac model was selected as the best-performing model for 22 cases (34.38%), followed by the AR(1) model in 21 cases (32.81%) out of the 64 cases, the B-DART model in 18 cases (28.13%) and the B-CART model in 3 cases (4.69%). This supports the hypothesis that the Dirac and AR(1) models, which do not account for structural changes, perform better for the pre-COVID-19 period when the relationship between variables were relatively stable.

In the case of variable-specific results, the Dirac and AR(1) model outperforms their counterparts when forecasting monetary policy indicators (FFE and 10-year maturity interest rate) and economic activity indicators (real GDP and unemployment rate). These variables exhibit relatively stable dynamics prior to the COVID-19 period, making simpler models like the Dirac or AR(1) better suited for capturing trends without the need for splitting the sample. Introducing tree-based splitting in this context may unnecessarily complicate the model and lead to overfitting, which deteriorates forecast accuracy. For example, the B-DART model is dominated across all forecast horizons for the 10-year maturity interest rate case. This can be attributed to the long-term interest rate's tendency to follow a more persistent and smoother path, driven by monetary policy regimes, inflation expectations and global financial conditions. Thus, the tree-based splitting using the B-DART model introduces unnecessary complexity.

An exception is observed for price related indicators, such as the CPI, PPI and Crude Oil price, where the B-DART model outperforms other models. Specifically, it greatly

Table 4: Comparison of Relative RMSE for Post-COVID-19 Periods

	PCE			CPI			PPI			CrudeOil		
H	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART
1	**0.8183**	0.9924	1.1787	**0.9245**	1.1092	1.2641	**0.8997**	0.9793	1.1506	**0.8973**	0.9711	1.4001
2	**0.8574**	1.0317	0.9068	0.9880	**0.9716**	1.0719	**0.9171**	1.0008	0.9743	**0.9516**	0.9815	1.1537
3	1.0385	1.0068	1.2140	**0.9378**	1.0365	1.3665	1.0181	**0.9964**	1.0810	**0.9057**	1.0022	1.1842
4	1.0092	1.0199	1.5621	1.0074	1.0322	1.5270	**0.9741**	0.9829	1.1902	1.0308	**0.9960**	1.1653
5	**0.9467**	0.9817	1.3150	1.0171	**0.9971**	1.1577	**0.9722**	0.9937	1.0833	**0.9495**	0.9653	1.0223
6	**0.9018**	0.9452	0.9894	**0.9092**	0.9710	1.0620	1.0001	1.0047	1.0498	1.0171	1.0008	1.0198
7	1.0018	**0.9947**	1.0073	**0.9910**	1.0014	1.3741	**0.8545**	0.9956	1.0583	1.0090	**0.9894**	1.0730
8	**0.9793**	1.0158	1.2492	**0.9339**	1.0174	1.5190	0.9902	**0.9791**	1.5043	1.0375	1.0280	1.3097
	B-DART ≻ AR(1) ≻ Dirac ≻ B-CART			B-DART ≻ Dirac ≻ AR(1) ≻ B-CART			B-DART ≻ Dirac ≻ AR(1) ≻ B-CART			B-DART ≻ Dirac = AR(1) ≻ B-CART		
	FFE			Long INT			Real GDP			Unemp		
H	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART	B-DART	Dirac	B-CART
1	1.1376	1.2488	5.0033	1.0408	1.0352	1.5831	**0.9169**	1.4626	1.6108	**0.6926**	0.9123	0.7593
2	**0.9815**	1.0805	1.6246	1.0432	**0.9980**	1.1172	1.0038	**0.9921**	1.3418	**0.9061**	0.9135	1.0167
3	**0.8846**	1.0938	1.3804	1.1635	1.0207	**0.9321**	**0.9446**	0.9855	1.0107	1.0389	1.0518	1.0765
4	1.1172	1.0467	2.4313	**0.9988**	1.0053	1.7876	**0.8445**	0.9955	1.0643	**0.9105**	0.9812	2.0310
5	**0.9809**	1.0160	2.0322	1.0080	**0.9523**	1.0791	0.9954	**0.9880**	1.1377	**0.9276**	0.9960	1.0619
6	**0.8694**	1.0498	1.1344	**0.9401**	0.9811	1.4221	**0.9474**	1.0047	1.2861	**0.9588**	0.9986	1.2729
7	**0.9252**	1.0546	1.2070	**0.9861**	1.0177	1.1527	**0.9957**	1.0164	1.0971	1.0002	**0.9903**	1.2284
8	**0.8364**	0.9274	1.1507	**0.9070**	1.0227	1.2282	**0.9757**	0.9797	1.1025	0.9903	**0.9885**	1.0361
	B-DART ≻ AR(1) ≻ Dirac = B-CART			B-DART ≻ Dirac ≻ AR(1) = B-CART			B-DART ≻ Dirac ≻ AR(1) = B-CART			B-DART ≻ Dirac = AR(1) ≻ B-CART		
	Total # of B-DART 1st = 41, 64.06% Total # of Dirac 1st = 12, 18.75%						Total # of B-CART 1st = 1, 1.56% Total # of AR(1) 1st = 10, 15.63%					

(1) The numbers are relative RMSEs where the RMSE of AR(1) is normalized as 1.
(2) The bold font values denotes that the model is better than the other three.
(3) The order under each variable was determined by counting the bolded values for each forecast horizon.
(4) The numbers and percentages in the last two rows of each table represent the count of bolded values for each model and their proportion out of the total 64 cases (8 variables × 8 forecast horizons).
(5) The values depicted in the table above are all obtained under the best hyper-parameter for each forecast model. Details of the selected hyper-parameters are reported in the Appendix G.

outperforms its counterparts across most forecast horizons for the Crude Oil price case. This implies that commodity and price indices potentially exhibit nonlinear, regime-like behaviors, e.g. drastic shifts in oil prices. Thus, by using the B-DART model, one can capture such nonlinear relationship and guard against forecast instability.

On the other hand, the comparison of the forecast accuracy using post-COVID-19 periods is illustrated in **Table 4**. For overall results, the B-DART model greatly outperforms other models since it is selected as the leading model for 41 cases (64.06%). The Dirac, AR(1) and B-CART models show significantly lower forecast performance in comparison to the B-DART, as they are selected as the best-performing model in only 12 (18.75%), 10 (15.63%), 1 (1.56%) cases, respectively. These findings suggest that our proposed B-DART model, which jointly estimates the regimes within the sample and applies variable selection, shows better forecast accuracy in the post-COVID-19 periods.

This is because that the post-COVID-19 periods features abrupt changes or nonlinear relationships across macroeconomic variables, due to sudden shifts in consumption patterns, supply constraints and policy responses.

Across all macroeconomic variable, the B-DART model shows superior forecast power to its counterparts. For monetary policy indicators, it exhibits strong performance for long-term forecast horizons. For the case of economic activity indicators, the B-DART model shows better forecast accuracy for short- and medium-terms (up to 6 quarters) of the unemployment rate, and sporadic horizons for the real GDP. Finally, the B-DART model dominates all other models for price indicators, effectively capturing the changes in relationships between macroeconomic variables during the recent high-inflation era after the pandemic.

Notably, it is worth emphasizing that the B-DART model's strong forecast performance is not limited to volatile periods, such as the post-COVID-19 period. This is because even when considering the full out-of-sample period, the B-DART frequently outperforms the Dirac and B-CART models. This finding provides empirical evidence that it is not only important to select statistically relevant variables but also account for the changes in the relationships of variables over time, in order to achieve reliable forecasting results.

6.4.3 Selected Variables across Out-of-sample Periods

So far, we have examined which forecast models outperform using different out-of-sample periods by assessing the out-of-sample forecasting accuracy across 8 macroeconomic variables over 1- to 8-quarter-ahead forecast horizons. In this section, we delve more deeply into the reasons underlying these performance differences by analyzing which variables are selected in the Dirac and B-DART models during each out-of-sample periods. Due to spatial constraints, we limit our discussion to the variable selection results for the 1- and 8-quarter-ahead forecasts of the unemployment rate, rather than presenting all 64 outcomes. The selection is motivated by the fact that, for these specific forecasts, the B-DART model either dominates or is dominated by the Dirac model. Additional results for other variables at specific forecast horizons are summarized in the Supplementary Appendix.[20]

For forecasting the 1- and 8-quarter-ahead unemployment rate, the variables selected by the Dirac and B-DART models during each out-of-sample periods are visualized in the

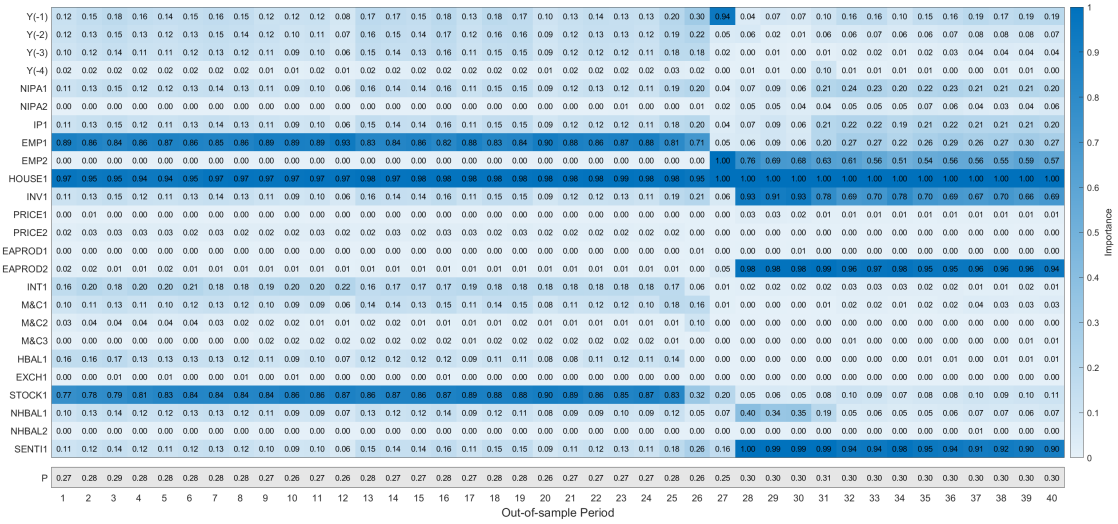
[20]Note that the variable selection results over the out-of-sample period for further cases are available upon request.

heatmaps as **Figure 2** and **Figure 3**. In these heatmaps, the x-axis represents the out-of-sample periods and the y-axis shows the employed predictor variables, which in our case include the autoregressive terms up to four lags as well as the principal components (PCs) estimated via PCA that represent each of the 14 categories. The values within the heatmap represent $\hat{\delta}_k$ from the Dirac variable selection, which indicate the importance of each predictor variable. These values, ranging from 0 to 1, are displayed with darker colors as they approach 1, signifying greater importance.[21] The final row in the heatmap, colored in gray, displays the inclusion probabilities for each out-of-sample period, which serves as the threshold for interpreting the importance of each variable. Nonetheless, in alignment with Giannone et al. (2021), we also include variables that do not meet this threshold, as their work suggests that overlooking model uncertainty in forecasting can lead to the "illusion of sparsity".

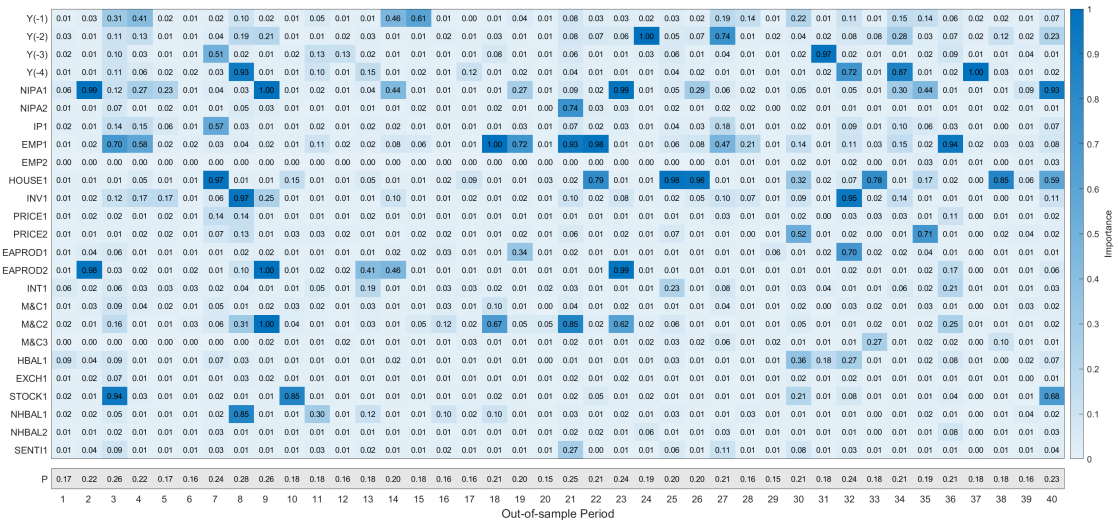
Both models exhibit changes in variable importance as information is updated over the accumulation of out-of-sample periods. However, the importance of predictors for the Dirac model tends to remain stable until new information is introduced. At some point, the importance may gradually increase or decrease, displaying a pattern of consistent adjustment. In contrast, the B-DART model groups the response variable discontinuously across periods and estimates the importance of predictor variables for each group. This leads to more abrupt changes in the importance of variables across out-of-sample periods. To determine which method proves more effective in forecasting, we also examine the forecasting outcomes for each forecast horizons of the unemployment rate in **Table 2**. Additionally, in specific out-of-sample periods, all predictor variables are not selected and the forecasting is based solely on the mean of the response variable. This observation highlights that the representative factors for each category group does not necessarily possess predictive power across all forecast horizons.

Let us now focus on the variables selected in forecasting the 1-quarter-ahead unemployment rate. With the Dirac model, we observe a marked shift in the importance of variables from the 27th out-of-sample period (2020 Q4), i.e. after the impact of COVID-19. Furthermore, certain predictors are assigned high importance, receiving substantial weights in out-of-sample forecasting. However, regarding the $H = 1$ case in **Table 2**

[21] As observed in **Figure 6** and **Table 5** in the appendix, the number of PCs representing each category varies depending on the out-of-sample period. Specifically, 19 PCs are used up to the 8th period; however an additional PC for Money and Credit is incorporated starting from the 9th period (2016 Q2), with one more added for Employment and Unemployment from the 27th period (2020 Q4) and onward, i.e. after the impact of COVID-19. For such additional PCs, the importance is written as zero in the periods where it does not exist.



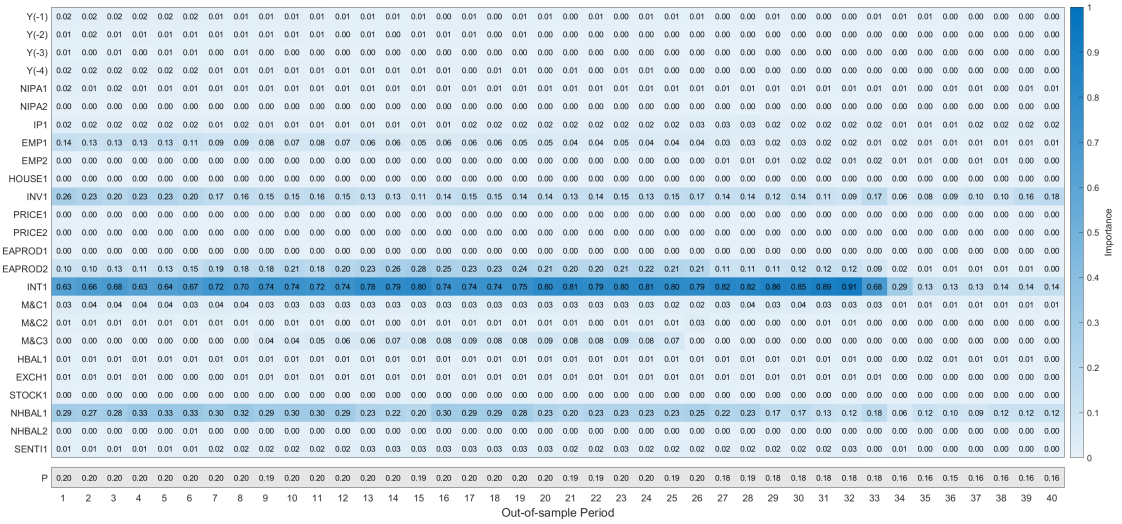
(a) Dirac Model Case



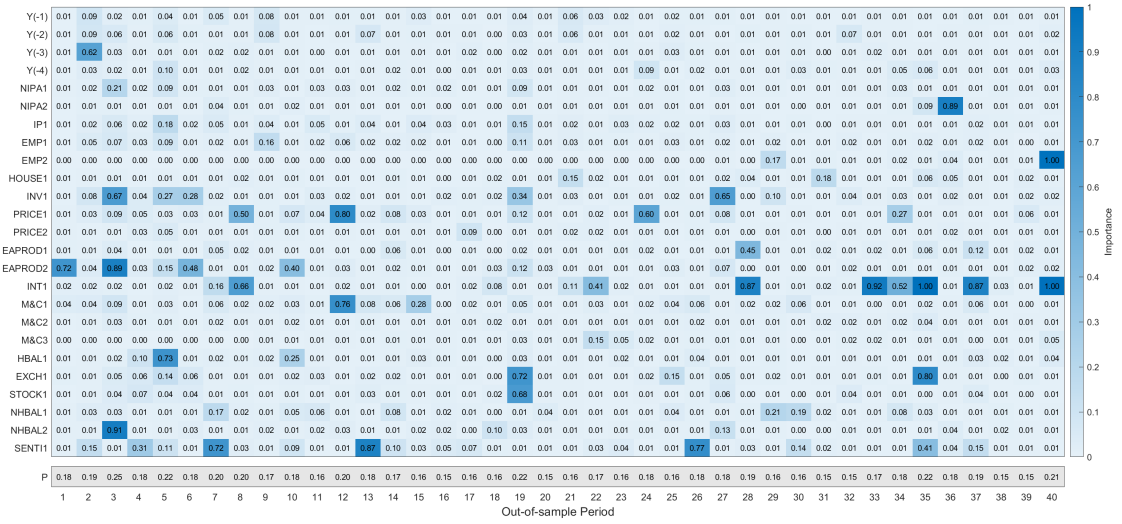
(b) B-DART Model Case

Figure 2: Importance for Forecasting 1-quarter-ahead Unemployment Rate

Note. The figure above depicts the importance of each predictor variable for 40 out-of-sample periods when forecasting the 1-quarter-ahead unemployment rate. The heatmap becomes darker as the estimated importance of the predictor variable approaches 1. The last row is the estimated inclusion probability.



(a) Dirac Model Case



(b) B-DART Model Case

Figure 3: Importance for Forecasting 8-quarter-ahead Unemployment Rate

Note. The figure above depicts the importance of each predictor variable for 40 out-of-sample periods when forecasting the 8-quarter-ahead unemployment rate. The heatmap becomes darker as the estimated importance of the predictor variable approaches 1. The last row is the estimated inclusion probability.

above, the forecast accuracy of the B-DART model outperforms the Dirac model. This suggests that when forecasting the 1-quarter-ahead unemployment rate, consistently selected variables may, in fact, exacerbate model uncertainty and impair the forecasting accuracy.

This is not to imply, however, that the flexible approach of considering time-varying variable importance, as in the B-DART model, always results in better forecasting performance. When forecasting the 8-quarter-ahead unemployment rate, we find that adjusting the importance of predictors by each out-of-sample period can actually lead to overfitting, resulting in lower forecast accuracy than that achieved by the Dirac model. Even when examining forecasts by dividing the out-of-sample period into different economic regimes, as shown in **Table 3** and **Table 4**, the Dirac model demonstrates better forecast accuracy than the B-DART model for the 8-quarter-ahead forecast. This suggests that overfitting occurs regardless of the change in the relationship between variables.

We would like to emphasize that the results above does not imply that the B-DART model outperforms other models due to its flexibility of considering time varying variable importance. Rather, these findings underscore the importance of appropriately updating new information as the out-of-sample periods extent.

6.4.4 Results of the B-DART Model

Finally, we illustrate several additional findings based on the results from applying the B-DART model to forecasting eight macroeconomic variables. The first feature is that when the B-DART model is applied to forecast macroeconomic variables, approximately five groups are estimated on average through the optimal tree structure. This grouping is crucial, as variable selection is conducted separately within each group, and forecasting is subsequently conducted using only the predictors deemed important within each group. The detailed results are provided in **Figure 4**, which depicts the sample mean of the number of groups across 40 optimal trees estimated over the out-of-sample periods, corresponding to each macroeconomic variable and forecast horizon combination. On average, around five groups were estimated for each macroeconomic variable and forecast horizon combination. In particular, apart from the FFE-whose average number of groups fluctuates between four to five groups-the average number of groups for other macroeconomic variables gradually decreases as the forecast horizon increases.

More importantly, these estimated group numbers indicate that variable selection and optimal tree structure estimation are not independent but rather combining them produces a synergistic effect. Specifically, as mentioned above, the B-DART model esti-

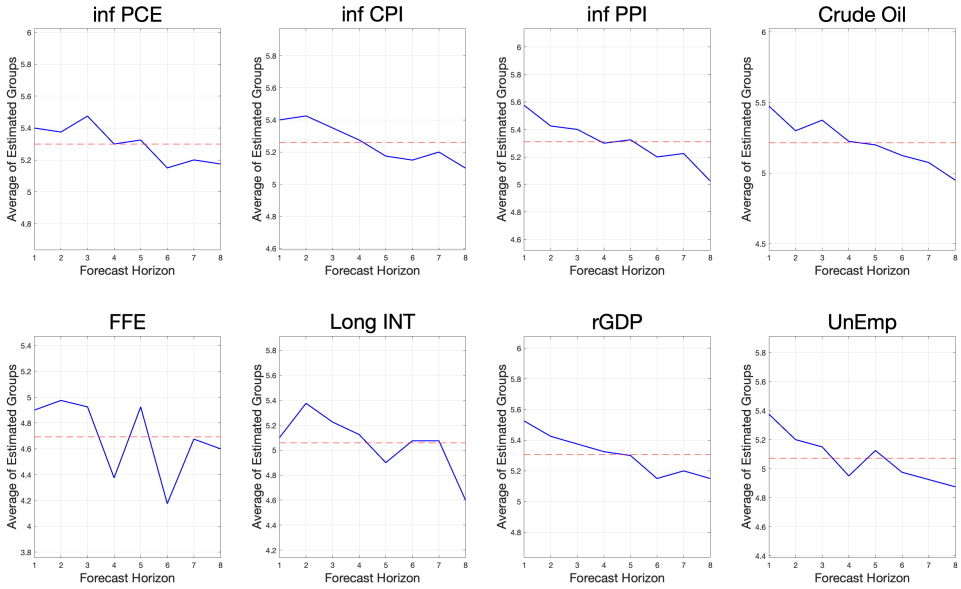


Figure 4: Average Groups of Variables across each Forecast Horizon

Note. This figure depicts the estimated average groups for macroeconomic variables of the B-DART model. The blue line is the average groups across each forecast horizon, while the red dotted line is the average groups among all 8 forecast horizon.

mated an average of five groups. In contrast, the B-CART model, which is a similar tree structure based method, exhibited not only significantly lower forecasting performance compared to its other counterparts but also failed to estimate an optimal tree structure. In fact, the average number of estimated groups from the B-CART model was merely one for most macroeconomic variables.[22] This indicates that the sample was not partitioned by the tree structure, and forecasting were performed by using the coefficients estimated from the entire sample with all available predictors. Such results highlight that the tree structure itself is a parameter, and when unnecessary parameters proliferate, the statistical significant of estimating the tree structure diminishes. Consequently, applying variable selection alongside estimating the optimal tree structure is important for accurately identifying the timing and source of instabilty.

Another intriguing finding from the B-DART model is the ability to determine which predictors were selected as the splitting criteria within the trees generated for each macroeconomic variable and forecast horizon combination. Due to spacial constraints,

[22]Refer to **Figure 7** in **Appendix F** for the detailed average groups across the forecast horizons for each of the 8 macroeconomic variables.

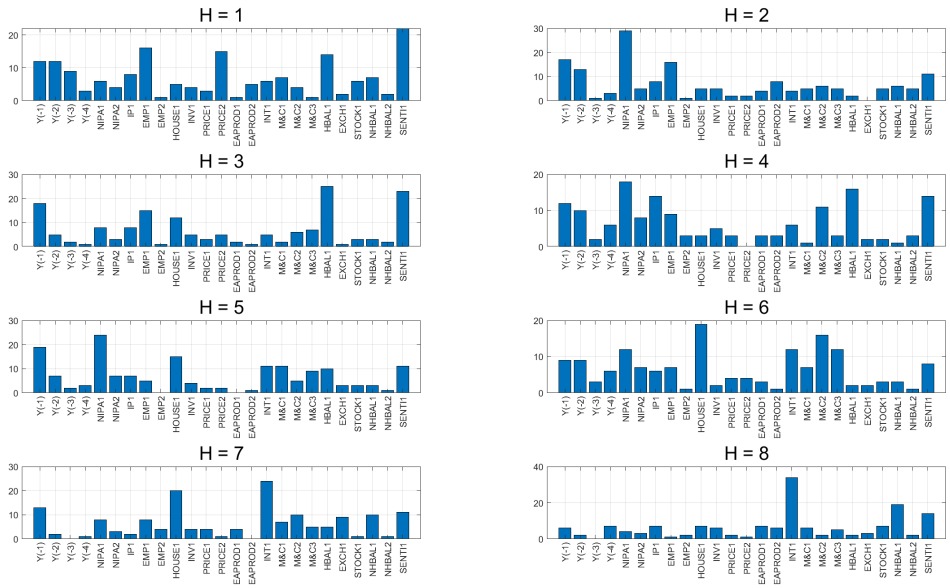


Figure 5: Frequency of Predictors used as the Splitting Criteria for Unemployment Rate

Note. This figure depicts the frequency of predictor variables that are used as the splitting criteria of the B-DART model for forecasting the unemployment rate.

we again limit our discussion to the results for the unemployment rate, with outcomes for other variables available in the Supplementary Appendix. **Figure 5** above illustrates the frequency of predictors that were used as the splitting criteria in the trees constructed for each forecast horizon of the unemployment rate.

Interestingly, certain predictors were used as the splitting criteria more frequently than others. Notably, autoregressive terms of the unemployment rate were employed as one of the main splitting criteria across all forecast horizons except that for 8-quarter-ahead forecasts. Additionally, while Economic Sentiment was consistently used as a split criterion across all horizons, its frequency of selection decreased with longer forecast horizons. For shorter horizons, the principal component representing Household Balance Sheets was one of the main split criteria from 1- to 4-quarter-ahead forecasts, but its frequency declined in later horizons. Conversely, the PC representing Housing became a more frequent split criterion in horizons 5, 6, and 7 for long-term forecasts. Finally, the PC representing Interest Rates showed an increase in frequency as a split criterion at 6-, 7-, 8-quarter-ahead forecasts.

7 Conclusion

Variable selection has become increasingly popular in macroeconomic forecasting, since dimension reduction methods are essential to handle big datasets for maximizing forecast accuracy. However, despite the inherent characteristic that selected variables may vary over time, this aspect has not been fully considered in previous literature. Incorporating time-varying variable selection into macroeconomic forecasting is particularly crucial during periods of relatively high volatility, such as the recent high-inflation era following the COVID-19 pandemic, when timely and effective decision making by economic agents and monetary policymakers becomes increasingly dependent upon it.

To this end, we propose the B-DART model which jointly considers the time-varying relationship between variables and variable selection in a big-data environment. Specifically, we apply the B-CART model to partition the sample into distinct groups, where Dirac variable selection is used within each group. The advantage of the Dirac variable selection is that the conditional marginal likelihood given the inclusion parameter δ_g can be derived analytically and serves as the criterion for selecting the optimal tree structure during the MH algorithm.

Our main empirical finding is that the B-DART model that jointly considers time-varying variable selection shows improved forecast accuracy, compared to applying the B-CART or Dirac variable selection model independently. Notably, the B-DART model demonstrates substantial improvements in forecasting accuracy for all eight macroeconomic variables during the relatively volatile post-COVID-19 out-of-sample periods. The reason of this superiority was due to the B-DART model's ability of effectively updating new information, although this flexibility also introduces a risk of overfitting. Finally, the B-DART model estimates approximately five groups on average for each of the eight macroeconomic variables, whereas the B-CART model yields only one group on average. Combined with the fact that the forecasting performance of the B-CART model is strongly dominated by the B-DART model, this highlights that the performance of variable selection and optimal tree structure estimation are not independent but rather leads to a synergistic effect.

References

Bai, J. and Ng, S. (2008), "Forecasting economic time series using targeted predictors," *Journal of Econometrics*, 146, 304–317.

Bernanke, B. and Blanchard, O. (2023), "What caused the US pandemic-era inflation?" *Peterson Institute for International Economics Working Paper*.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., and Tambalotti, A. (2018), "Macroeconomic nowcasting and forecasting with big data," *Annual Review of Economics*, 10, 615–643.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998), "Bayesian CART model search," *Journal of the American Statistical Association*, 93, pp.935–948.

— (2012), "BART: Bayesian additive regression trees," *Annals of Applied Statistics*, 6, 266–298.

Fernandez, C., Ley, E., and Steel, M. F. (2001), "Benchmark priors for Bayesian model averaging," *Journal of Econometrics*, 100, 381–427.

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

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2022), "How is machine learning useful for macroeconomic forecasting?" *Journal of Applied Econometrics*, 37, 920–964.

Jordà, Ò., Singh, S. R., and Taylor, A. M. (2022), "Longer-run economic consequences of pandemics," *Review of Economics and Statistics*, 104, 166–175.

Kim, C.-J. and Nelson, C. R. (1999), "State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications," *MIT Press Books*, 1.

Kim, H. H. and Swanson, N. R. (2018), "Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods," *International Journal of Forecasting*, 34, 339–354.

Lee, T.-H., White, H., and Granger, C. W. (1993), "Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests," *Journal of Econometrics*, 56, 269–290.

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

Malsiner-Walli, G. and Wagner, H. (2011), "Comparing Spike and Slab Priors for Bayesian Variable Selection," *Austrian Journal of Statistics*, 40, 241–264.

McCracken, M. W. and Ng, S. (2020), "FRED-QD: A Quarterly Database for Macroeconomic Research," Tech. rep., Federal Reserve Bank of St. Louis.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., and Zilberman, E. (2021), "Forecasting inflation in a data-rich environment: the benefits of machine learning methods," *Journal of Business & Economic Statistics*, 39, 98–119.

Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1032.

Ng, S. (2013), "Variable selection in predictive regressions," *Handbook of Economic Forecasting*, 2, 752–789.

Ng, S. and Wright, J. H. (2013), "Facts and challenges from the great recession for forecasting and macroeconomic modeling," *Journal of Economic Literature*, 51, 1120–1154.

O'Hagan, A. (1995), "Fractional Bayes factors for model comparison," *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 99–118.

Pesaran, M. H. and Timmermann, A. (1995), "Predictability of stock returns: Robustness and economic significance," *The Journal of Finance*, 50, 1201–1228.

Rapach, D. E. and Wohar, M. E. (2006), "Structural breaks and predictive regression models of aggregate US stock returns," *Journal of Financial Econometrics*, 4, 238–274.

Rossi, B. (2013), "Exchange rate predictability," *Journal of Economic Literature*, 51, 1063–1119.

— (2021), "Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them," *Journal of Economic Literature*, 59, 1135–1190.

Stock, J. H. and Watson, M. W. (1996), "Evidence on structural instability in macroeconomic time series relations," *Journal of Business & Economic Statistics*, 14, 11–30.

— (2003), "Forecasting output and inflation: The role of asset prices," *Journal of Economic Literature*, 41, 788–829.

— (2006), "Forecasting with many predictors," *Handbook of Economic Forecasting*, 1, 515–554.

— (2012), "Disentangling the Channels of the 2007-2009 Recession," Tech. rep., National Bureau of Economic Research.

Teräsvirta, T. (2006), "Forecasting economic variables with nonlinear models," *Handbook of Economic Forecasting*, 1, 413–457.

Varian, H. R. (2014), "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, 28, 3–28.

Welch, I. and Goyal, A. (2008), "A comprehensive look at the empirical performance of equity premium prediction," *The Review of Financial Studies*, 21, 1455–1508.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," *Bayesian Inference and Decision Techniques*.

A Deriving the Conditional Marginal Likelihood

In this section, we show the details of deriving the marginal likelihood of each terminal node when the data generating process (DGP) explicitly considers both the intercept and the coefficient of predictor variables. Then we show that this marginal likelihood is equivalent to that when using demeaned data without the intercept term. Finally, we derive the conditional marginal likelihood where the Dirac spike-and-slab prior is applied to the coefficient term.

A.1 Marginal Likelihood of Each Terminal Node (Decentered)

Suppose the Dirac spike-and-slab prior is not imposed on β_g and the DGP (model) is

$$Y = Z \cdot \alpha + \varepsilon$$

where $Z = (\mathbf{1}_T, X)'$ and $\alpha = (\mu, \beta)'$. Note that $\mathbf{1}_T$ denotes the vector of ones with $dim(\mathbf{1}_T) = T \times 1$ and $dim(Z) = T \times (K+1)$. Let the DGP of each terminal node (group) be expressed by explicitly considering both the intercept and coefficient of predictor variables as follows.

- $Y_g|\mu_g, \beta_g, \sigma_g^2 \sim \mathcal{N}(\mu_g \cdot 1_{T_g} + X_g\beta_g, \sigma_g^2 \cdot I_{T_g})$
- $\beta_g|\sigma_g^2 \sim \mathcal{N}(\beta_{0,g}, \sigma_g^2 \cdot B_{0,g})$
- $\sigma_g^2 \sim \mathcal{IG}(\frac{\nu}{2}, \frac{\nu \cdot \lambda}{2})$
- $\pi(\mu_g) \propto 1$; Improper (flat) prior[23]

Then the standard analytical simplification of the marginal likelihood can be done as follows.

$$f(Y|X, \mathcal{T}) = \iiint f(Y|X, \Theta, \mathcal{T}) \cdot \pi(\Theta|\mathcal{T}) d\Theta$$

[23]This is the case when assuming that the predictor variables are centered with the null vector as its mean, i.e. $X_g' \cdot \mathbf{1}_{T_g} = \mathbf{0}$. Precisely, let

$$Y_g = \mu_g \cdot \mathbf{1}_{T_g} + X_g \cdot \beta_g + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma_g^2 \cdot I_{T_g})$$

where $\mathbf{y_{g,c}} = Y - \overline{y_g} \cdot \mathbf{1}_{T_g}$ and $\mathbf{X_{g,c}} = X - \frac{1}{T_g}\mathbf{1}_{T_g}\mathbf{1}_{T_g}'X_g$. If $X_g' \cdot \mathbf{1}_{T_g} = 0 \iff \mathbf{1}_{T_g}' \cdot X_g = 0$, then $\mathbf{X_{g,c}} = X_g$. Thus, the model is simplified to

$$\mathbf{y_{g,c}} = X_g \cdot \beta_g + \varepsilon$$

$$= \iiint \prod_{g=1}^{G} \left(f(Y_g|\mu_g, \beta_g, \sigma_g^2) \cdot \pi(\beta_g|\sigma_g^2) \cdot \pi(\mu_g) \cdot \pi(\sigma_g^2) \right) d\{\beta_g\} d\{\mu_g\} d\{\sigma_g^2\}$$

Breaking this down to one group g for simplicity, one gets

$$\Rightarrow f(Y_g|X_g, \mathcal{T}) = \iiint f(Y_g|\mu_g, \beta_g, \sigma_g^2) \cdot \pi(\beta_g|\sigma_g^2) \cdot \pi(\mu_g) \cdot \pi(\sigma_g^2) d\beta_g d\mu_g d\sigma_g^2$$

Note that the probability density function (pdf) of each likelihood or prior distribution is as follows.

$$f(Y_g|\mu_g, \beta_g, \sigma_g^2) = (2\pi\sigma_g^2)^{\frac{T_g}{2}} \cdot \exp\left(-\frac{1}{2\sigma_g^2}(Y_g - \mu_g \mathbf{1}_{T_g} - X\beta)'(Y_g - \mu_g \mathbf{1}_{T_g} - X\beta) \right)$$

$$\pi(\beta_g|\sigma_g^2) = (2\pi\sigma_g^2)^{-\frac{K}{2}} \cdot |B_{0,g}|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2\sigma_g^2}(\beta_g - \beta_{0,g})' B_{0,g}^{-1}(\beta_g - \beta_{0,g}) \right)$$

$$\pi(\sigma_g^2) = \frac{(\frac{\nu\lambda}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \cdot (\sigma_g^2)^{-\frac{\nu}{2}-1} \cdot \exp\left(-\frac{\nu\lambda}{2\sigma_g^2} \right)$$

<**Integral Over** β_g>

Note that the joint distribution of Y_g and β_g under the given DGP is as follows.

$$\begin{aligned}
&f(Y_g, \beta_g|\mu_g, \sigma_g^2) \\
&= f(Y_g|\mu_g, \beta_g, \sigma_g^2) \cdot \pi(\beta_g|\sigma_g^2) \\
&= (2\pi\sigma_g^2)^{-\frac{T_g}{2}} \exp\left(-\frac{1}{2\sigma_g^2}(Y_g - \mu_g \mathbf{1}_{T_g} - X_g\beta_g)'(Y_g - \mu_g \mathbf{1}_{T_g} - X_g\beta_g) \right) \\
&\quad \times (2\pi\sigma_g^2)^{-\frac{k}{2}} \cdot |B_{0,g}|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2\sigma_g^2}(\beta_g - \beta_{0,g})' B_{0,g}^{-1}(\beta_g - \beta_{0,g}) \right) \\
&= \underbrace{(2\pi\sigma_g^2)^{-\frac{T_g}{2}} \cdot (2\pi\sigma_g^2)^{-\frac{k}{2}} \cdot |B_{0,g}|^{-\frac{1}{2}}}_{(*)} \\
&\quad \times \exp\left(-\frac{1}{2\sigma_g^2} \underbrace{\{(Y_g - \mu_g \mathbf{1}_{T_g} - X_g\beta_g)'(Y_g - \mu_g \mathbf{1}_{T_g} - X_g\beta_g) + (\beta_g - \beta_{0,g})' B_{0,g}^{-1}(\beta_g - \beta_{0,g})\}}_{(**)} \right)
\end{aligned} \tag{A.1}$$

Simplifying the terms in (*) yields to

$$\Rightarrow (2\pi\sigma_g^2)^{-\frac{T+K}{2}} \cdot |B_{0,g}|^{-\frac{1}{2}} \tag{A.2}$$

Now for (**), one can group the terms involving β_g as follows.

$$
\begin{aligned}
&(Y_g - \mu_g \mathbf{1}_{T_g} - X\beta_g)'(Y_g - \mu_g \mathbf{1}_{T_g} - X\beta_g) \\
&= (Y_g' - \mu_g \mathbf{1}_{T_g}' - \beta' X')(Y_g - \mu_g \mathbf{1}_{T_g} - X\beta_g) \\
&= Y_g'Y_g - 2\mu_g \mathbf{1}_{T_g}' Y_g - 2\beta_g' X_g'(Y_g - \mu_g \mathbf{1}_{T_g}) + \mu_g^2 \cdot T_g + \beta_g' X_g' X_g \beta_g \cdots \text{ since } \mathbf{1}_{T_g}' \mathbf{1}_{T_g} = T_g
\end{aligned}
$$

and

$$
\begin{aligned}
(\beta_g - \beta_{0,g}) B_{0,g}^{-1} (\beta_g - \beta_{0,g}) &= (\beta_g' - \beta_{0,g}') B_{0,g}^{-1} (\beta_g - \beta_{0,g}) \\
&= \beta_g' B_{0,g}^{-1} \beta_g - \beta_g' B_{0,g}^{-1} \beta_{0,g} - \beta_{0,g}' B_{0,g}^{-1} \beta_g + \beta_{0,g}' B_{0,g}^{-1} \beta_{0,g} \\
&= \beta_g' B_{0,g}^{1} \beta_g - 2\beta_g' B_{0,g}^{-1} \beta_{0,g} + \beta_{0,g}' B_{0,g}^{-1} \beta_{0,g}
\end{aligned}
$$

leading to

$$
\begin{aligned}
\Rightarrow & \underbrace{Y_g'Y_g - 2\mu_g \mathbf{1}_{T_g}' Y_g + \mu_g^2 T_g + \beta_{0,g}' B_{0,g}^{-1} \beta_{0,g}}_{\text{terms not involving } \beta_g} \\
& + \underbrace{\beta_g' (X_g' X_g + B_{0,g}^{-1}) \beta_g - 2\beta_g' (X_g'(Y_g - \mu_g \mathbf{1}_{T_g}) + B_{0,g}^{-1} \beta_{0,g})}_{\text{terms involving } \beta_g} \qquad \text{(A.3)}
\end{aligned}
$$

Then the terms involving β_g can be further expressed by completing the square for β_g as

$$
\begin{aligned}
&\Rightarrow \beta_g' B_{1,g}^{-1} \beta_g - 2\beta_g' B_{1,g}^{-1} \beta_{1,g} \\
&\Rightarrow (\beta_g - \beta_{1,g})' B_{1,g}^{-1} (\beta_g - \beta_{1,g}) - \beta_{1,g}' B_{1,g}^{-1} \beta_{1,g} \qquad \text{(A.4)}
\end{aligned}
$$

where $B_{1,g}^{-1} = X_g' X_g + B_{0,g}^{-1}$ and $B_{1,g}^{-1} \beta_{1,g} = X_g'(Y_g - \mu_g \mathbf{1}_{T_g}) + B_{0,g}^{-1} \beta_{0,g}$ leading to

$$
B_{1,g} = (X_g' X_g + B_{0,g}^{-1})^{-1} \text{ and } \beta_{1,g} = B_{1,g}(X_g'(Y_g - \mu_g \mathbf{1}_{T_g}) + B_{0,g}^{-1} \beta_{0,g})
$$

Therefore, by inputting equation (A.4) into (A.3), one gets

$$
\Rightarrow Y_g'Y_g - 2\mu_g \mathbf{1}_{T_g}' Y_g + \mu_g^2 T_g + \beta_{0,g}' B_{0,g}^{-1} \beta_{0,g} + (\beta_g - \beta_{1,g})' B_{1,g}^{-1} (\beta_g - \beta_{1,g}) - \beta_{1,g}' B_{1,g}^{-1} \beta_{1,g} \quad \text{(A.5)}
$$

Now by applying results (A.2) and (A.5) into (A.1),

$$
\begin{aligned}
f(Y_g, \beta_g | \mu_g, \sigma_g^2) &= (2\pi\sigma_g^2)^{-\frac{T+K}{2}} \cdot |B_{0,g}|^{-\frac{1}{2}} \\
&\quad \times \exp\left(-\frac{1}{2\sigma_g^2} (Y_g'Y_g - 2\mu_g \mathbf{1}_{T_g}' Y_g + \mu_g^2 T_g + \beta_{0,g}' B_{0,g}^{-1} \beta_{0,g} - \beta_{1,g}' B_{1,g}^{-1} \beta_{1,g}) \right)
\end{aligned}
$$

$$\times \exp\left(-\frac{1}{2\sigma_g^2}(\beta_g - \beta_{1,g})' B_{1,g}^{-1}(\beta_g - \beta_{1,g}) \right)$$

Finally, one can integrate out β_g as follows.

$$\begin{aligned}
\int f(Y_g, \beta_g | \mu_g, \sigma_g^2) d\beta_g &= (2\pi\sigma_g^2)^{-\frac{T_g+K}{2}} \cdot |B_{0,g}|^{-\frac{1}{2}} \\
&\quad \times \exp\left(-\frac{1}{2\sigma_g^2}(Y_g'Y_g + -2\mu_g \mathbf{1}_{T_g}' Y_g + \mu_g^2 T_g + \beta_{0,g}' B_{0,g}^{-1}\beta_{0,g} - \beta_{1,g}' B_{1,g}^{-1}\beta_{1,g}) \right) \\
&\quad \times (2\pi\sigma_g^2)^{\frac{K}{2}} |B_{1,g}|^{\frac{1}{2}} \\
&\quad \times \underbrace{\int (2\pi\sigma_g^2)^{-\frac{K}{2}} |B_{1,g}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_g^2}(\beta_g - \beta_{1,g})' B_{1,g}'(\beta_g - \beta_{1,g}) \right) d\beta_g}_{=1 \text{ since } \mathcal{N}(\beta_{1,g}, \sigma_g^2 B_{1,g})} \\
&= (2\pi\sigma_g^2)^{-\frac{T_g}{2}} \left(\frac{|B_{1,g}|}{|B_{0,g}|} \right)^{\frac{1}{2}} \\
&\quad \times \exp\left(-\frac{1}{2\sigma_g^2}(Y_g'Y_g - 2\mu_g \mathbf{1}_{T_g}' Y_g + \mu_g^2 T_g + \beta_{0,g}' B_{0,g}^{-1}\beta_{0,g} - \beta_{1,g}' B_{1,g}^{-1}\beta_{1,g}) \right) \\
&\equiv f(Y_g | \mu_g, \sigma_g^2)
\end{aligned}$$

<Integral Over μ_g>

The joint distribution of Y_g and μ_g under the given DGP is as follows.

$$\begin{aligned}
f(Y_g, \mu_g | \sigma_g^2) &= f(Y_g | \mu_g, \sigma_g^2) \cdot \pi(\mu_g) \\
&= (2\pi\sigma_g^2)^{-\frac{T+K}{2}} \cdot \left(\frac{|B_{1,g}|}{|B_{0,g}|} \right)^{\frac{1}{2}} \\
&\quad \times \exp\left(-\frac{1}{2\sigma_g^2} \{ \underbrace{Y_g'Y_g - 2\mu_g \mathbf{1}_{T_g}' Y_g + \mu_g^2 T_g}_{\text{terms involving } \mu_g} + \underbrace{\beta_{0,g}' B_{0,g}^{-1}\beta_{0,g} - \beta_{1,g}' B_{1,g}^{-1}\beta_{1,g}}_{\text{terms not involving } \mu_g} \} \right)
\end{aligned} \tag{A.6}$$

Then the terms involving μ_g can be further expressed by completing the square for μ_g as follows.

$$\Rightarrow Y_g'Y_g - 2\mu_g \mathbf{1}_{T_g}' Y_g + \mu_g^2 T_g = Y_g'Y_g + T_g(\mu_g - \overline{y_g})^2 - T_g \overline{y_g}^2 \tag{A.7}$$

where $\overline{y_g} = \frac{1}{T_g}\mathbf{1}'_{T_g} Y_g$. Inputting equation (B.7) into (B.6), one gets

$$f(Y_g, \mu_g|\sigma_g^2) = (2\pi\sigma_g^2)^{-\frac{T_g}{2}} \cdot \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \times \exp\left(-\frac{1}{\sigma_g^2}\{Y_g'Y_g + T_g(\mu_g - \overline{y_g})^2 - T_g\overline{y_g}^2 + \beta_{0,g}'B_{0,g}^{-1}\beta_{0,g} - \beta_{1,g}'B_{1,g}^{-1}\beta_{1,g}\}\right)$$

Finally, one can integrate out μ_g as follows.

$$\begin{aligned}
\int f(Y_g, \mu_g|\sigma_g^2) &= (2\pi\sigma_g^2)^{-\frac{T_g}{2}} \cdot \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \\
&\quad \times \exp\left(-\frac{1}{2\sigma_g^2}(Y_g'Y_g - T_g\overline{y_g}^2 + \beta_{0,g}'B_{0,g}^{-1}\beta_{0,g} - \beta_{1,g}B_{1,g}^{-1}\beta_{1,g})\right) \\
&\quad \times (2\pi\sigma_g^2)^{\frac{1}{2}} \cdot T_g^{-\frac{1}{2}} \cdot \underbrace{\int (2\pi\sigma_g^2)^{-\frac{1}{2}} \cdot T_g^{\frac{1}{2}} \exp\left(-\frac{T_g}{2\sigma_g^2}(\mu_g - \overline{y_g})^2\right) d\mu_g}_{=1 \text{ since } \mathcal{N}(\overline{y_g}, \frac{1}{T_g}\sigma_g^2)} \\
&= (2\pi\sigma_g^2)^{-\frac{(T_g-1)}{2}} \cdot T_g^{-\frac{1}{2}} \cdot \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \cdot \exp -\frac{1}{2\sigma_g^2}A \\
&\equiv f(Y_g|\sigma_g^2)
\end{aligned}$$

where $A \equiv Y_g'Y_g + \beta_{0,g}'B_{0,g}^{-1}\beta_{0,g} - \beta_{1,g}B_{1,g}^{-1}\beta_{1,g} - T_g \cdot \overline{y_g}^2$.

<Integral Over σ_g^2>

Now the joint distribution of Y_g and σ_g^2 under the given DGP is as follows.

$$\begin{aligned}
f(Y_g, \sigma_g^2) &= f(Y_g|\sigma_g^2) \cdot \pi(\sigma_g^2) \\
&= (2\pi)^{-\frac{(T_g-1)}{2}}(\sigma_g^2)^{-\frac{(T_g-1)}{2}} \cdot T_g^{-\frac{1}{2}} \cdot \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \\
&\quad \times \exp\left(-\frac{1}{2\sigma_g^2}A\right) \cdot \frac{(\frac{\nu\lambda}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\sigma_g^2)^{-\frac{\nu}{2}-1} \exp\left(-\frac{1}{2\sigma_g^2}\nu\lambda\right) \\
&= (2\pi)^{-\frac{(T_g-1)}{2}} \cdot T_g^{-\frac{1}{2}} \cdot \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \frac{(\frac{\nu\lambda}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\sigma_g^2)^{-\frac{(T_g-1)}{2}-\frac{\nu}{2}-1} \exp\left(-\frac{1}{2\sigma_g^2}(A + \nu\lambda)\right)
\end{aligned}$$

Integrating out σ_g^2,

$$
\begin{aligned}
&\int f(Y_g, \sigma_g^2) d\sigma_g^2 \\
&= (2\pi)^{-\frac{(T_g-1)}{2}} \cdot T_g^{-\frac{1}{2}} \cdot \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \frac{(\frac{\nu\lambda}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \cdot \int (\sigma_g^2)^{-\frac{(T_g-1)}{2}-\frac{\nu}{2}-1} \cdot \exp\left(-\frac{1}{2\sigma_g^2}(A+\nu\lambda)\right) d\sigma_g^2 \\
&= (2\pi)^{-\frac{(T_g-1)}{2}} \cdot T_g^{-\frac{1}{2}} \cdot \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \frac{(\frac{\nu\lambda}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \frac{\Gamma(\frac{(T_g-1)+\nu}{2})}{(\frac{A+\nu\lambda}{2})^{\frac{(T_g-1)+\nu}{2}}} \\
&\quad \times \underbrace{\int \frac{(\frac{A+\nu\lambda}{2})^{\frac{(T_g-1)+\nu}{2}}}{\Gamma(\frac{(T_g-1)+\nu}{2})} \cdot (\sigma_g^2)^{-\frac{(T_g-1)}{2}-\frac{\nu}{2}-1} \cdot \exp\left(-\frac{1}{2\sigma_g^2}(A+\nu\lambda)\right) d\sigma_g^2}_{=1 \text{ since } \mathcal{IG}(\frac{(T_g-1)+\nu}{2}, \frac{A+\nu\lambda}{2})} \\
&= (2\pi)^{-\frac{(T_g-1)}{2}} \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \left(\frac{\nu\lambda}{2}\right)^{\frac{\nu}{2}} \frac{\Gamma(\frac{(T_g-1)+\nu}{2})}{\Gamma(\frac{\nu}{2})} \cdot \left(\frac{A+\nu\lambda}{2}\right)^{-\frac{(T_g-1)+\nu}{2}} \\
&= (\pi)^{-\frac{(T_g-1)}{2}} (\nu\lambda)^{\frac{\nu}{2}} \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{(T_g-1)+\nu}{2})}{\Gamma(\frac{\nu}{2})} \cdot (A+\nu\lambda)^{-\frac{(T_g-1)+\nu}{2}}
\end{aligned}
$$

Consequently, the analytical simplification of the marginal likelihood under the given DGP is as follows.

$$
P(Y_g|X_g, \mathcal{T}) = \prod_{g=1}^{G} (\pi)^{-\frac{(T_g-1)}{2}} (\nu\lambda)^{\frac{\nu}{2}} \left(\frac{|B_{1,g}|}{|B_{0,g}|}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{(T_g-1)+\nu}{2})}{\Gamma(\frac{\nu}{2})} \cdot (A+\nu\lambda)^{-\frac{(T_g-1)+\nu}{2}} \tag{A.8}
$$

where

$$
A = Y_g' Y_g + \beta_{0,g}' B_{0,g}^{-1} \beta_{0,g} - \beta_{1,g}' B_{1,g}^{-1} \beta_{1,g} - T_g \cdot \overline{y_g}^2 \tag{A.9}
$$

$$
B_{1,g} = (X_g' X_g + B_{0,g}^{-1})^{-1} \tag{A.10}
$$

$$
\beta_{1,g} = B_{1,g}(X_g'(Y_g - \mu_g \mathbf{1}_{T_g}) + B_{0,g}^{-1}\beta_{0,g}) \tag{A.11}
$$

A.2 Equivalence of Marginal Likelihood under Centered Data

In the DGP above, there exist three parameters to be estimated, i.e. μ_g, β_g and σ_g. By assuming that the data is centered, i.e. both Y_g and X_g are demeaned, one can avoid the estimation of μ_g and consider only β_g and σ_g^2 in the estimation process. That is, suppose the DGP is as follows.

- $\mathbf{y_{g,c}}|\beta_g, \sigma_g^2 \sim \mathcal{N}(X \cdot \beta_g, \sigma_g^2 \cdot I_{T_g})$

- $\beta_g|\sigma_g^2 \sim \mathcal{N}(\beta_{0,g}, \sigma_g^2 \cdot B_{0,g})$
- $\sigma_g^2 \sim \mathcal{IG}(\frac{\nu}{2}, \frac{\nu\lambda}{2})$

where $\mathbf{y_{g,c}} = Y_g - \overline{y_g} \cdot \mathbf{1}_{T_g} = Y_g - \frac{1}{T_g}\mathbf{1}_{T_g}\mathbf{1}'_{T_g}Y_g$ since $\overline{y_g} = \frac{1}{T_g}\mathbf{1}'_{T_g}Y_g$ and $X_g = \mathbf{X_{g,c}}$.[24]

Then the standard analytical simplification of the marginal likelihood of each terminal node is now

$$f(\mathbf{y_{g,c}}|X) = \iint f(\mathbf{y_{g,c}}|\beta_g, \sigma_g^2) \cdot \pi(\beta_g|\sigma_g^2) \cdot \pi(\sigma_g^2) d\beta_g d\sigma_g^2$$

Rather than showing the detailed process of integrating out β_g and σ_g^2 as above, we show the equivalence of the marginal likelihood under the decentered data and centered data by comparing A, $B_{1,g}$ and $\beta_{1,g}$.

First, one can easily observe that $B_{1,g}$ is the same as equation (A.10) since $X_g = \mathbf{X_{g,c}}$, i.e.

$$B_{1,g} = (X_g'X_g + B_{0,g}^{-1})^{-1}$$

Next, we show that $\beta_{1,g}$ is the same as equation (A.11).

Proposition 1. *Suppose that the predictor variables are centered with the null vector as its mean, i.e.* $X_g' \cdot \mathbf{1}_{T_g} = 0$. *Then*

$$X_g'(Y_g - \mu_g \cdot \mathbf{1}_{T_g}) = X_g' \cdot \mathbf{y_{g,c}}$$

Proof. By definition, $\mathbf{y_{g,c}} = Y_g - \overline{y_g} \cdot \mathbf{1}_{T_g}$ where $\overline{y_g} = \frac{1}{T_g}\mathbf{1}'_{T_g}Y_g$. Then

$$\begin{aligned}
X_g' \cdot \mathbf{y_{g,c}} &= X_g'(Y_g - \overline{y_g} \cdot \mathbf{1}_{T_g}) \\
&= X_g'(Y_g - \frac{1}{T_g}\mathbf{1}_{T_g} \cdot \mathbf{1}'_{T_g}Y_g) \cdot \text{ since } dim(\mathbf{1}'_{T_g}Y_g) = 1 \\
&= X_g'Y_g - \frac{1}{T_g}X_g'\mathbf{1}_{T_g} \cdot \mathbf{1}'_{T_g}Y_g \\
&= X_g'Y_g
\end{aligned}$$

where the last equality is due to $X_g' \cdot \mathbf{1}_{T_g} = 0$. Thus,

$$\begin{aligned}
X_g'(Y_g - \mu_g \cdot \mathbf{1}_{T_g}) &= X_g'Y_g - \mu_g \cdot X_g'\mathbf{1}_{T_g} \\
&= X_g'Y_g \cdots \text{ since } X_g' \cdot \mathbf{1}_{T_g} \\
&= X_g' \cdot \mathbf{y_{g,c}}
\end{aligned}$$

[24] This is due to the fact that $\pi(\mu_g) \propto 1$. For details, refer to footnote 12.

which is the desired equality. □

Therefore, proposition 1 leads to the fact that $\beta_{1,g}$ is the same as before.

Finally, we show that A is the same as equation (A.9).

Proposition 2. *The following equality holds*

$$\mathbf{y_{g,c}}' \cdot \mathbf{y_{g,c}} = Y_g' Y_g - T_g \cdot \overline{y}^2$$

Proof. By definition, $\mathbf{y_{g,c}} = Y_g - \overline{y_g} \cdot \mathbf{1}_{T_g}$ where $\overline{y_g} = \frac{1}{T_g} \mathbf{1}_{T_g}' Y_g$. Then

$$\begin{aligned}
\mathbf{y_{g,c}}' \cdot \mathbf{y_{g,c}} &= (Y_g - \overline{y_g} \cdot \mathbf{1}_{T_g})'(Y_g - \overline{y_g} \cdot \mathbf{1}_{T_g}) \\
&= (Y_g' - \overline{y_g} \cdot \mathbf{1}_{T_g}')(Y_g - \overline{y_g} \cdot \mathbf{1}_{T_g}) \\
&= Y_g' Y_g - \overline{y_g} \cdot Y_g' \mathbf{1}_{T_g} - \overline{y_g} \cdot \mathbf{1}_{T_g}' Y_g + \overline{y_g}^2 \cdot \mathbf{1}_{T_g}' \mathbf{1}_{T_g} \\
&= Y_g' Y_g - \overline{y_g} \cdot T_g \cdot \overline{y_g} - \overline{y_g} \cdot T_g \cdot \overline{y_g} + \overline{y_g}^2 \cdot T_g \cdots \text{ since } \mathbf{1}_{T_g}' Y_g = Y_g' \mathbf{1}_{T_g} = T_g \cdot \overline{y_g} \\
&= Y_g' Y_g - T_g \cdot \overline{y_g}
\end{aligned}$$

□

By proposition 2, one can observe that A is the same as before.

Consequently, since A, $B_{1,g}$ and $\beta_{1,g}$ is the same as equations (A.9) to (A.11), we have shown that the marginal likelihood under the centered data and decentered data are the same. Therefore, for the simplicity of estimation, we use the centered (demeaned) data and estimate only β_g and σ_g^2 for each terminal node.

A.3 Conditional Marginal Likelihood of Each Terminal Node

Now let us consider the case where the Dirac spike-and-slab prior is applied to β_g. Then the DGP of each terminal node (group) for the centered data is as follows.

- $\mathbf{y_{g,c}} | X_g \cdot \beta_g, \sigma_g^2 \sim \mathcal{N}(X_g \beta_g, \sigma_g^2 I_{T_g})$
- $\beta_g | \sigma_g^2, \delta \sim \pi(\beta_g | \sigma_g^2, \delta) = \pi_{slab}(\beta_\delta) \cdot \prod_{k:\delta_k=0} \pi_{spike}(\beta_{g,k})$

  where
  $$\begin{cases} \beta_\delta \sim \mathcal{N}(\beta_{0,\delta}, \sigma_g^2 \cdot B_{0,\delta}) \\ \beta_{-\delta} \sim \text{density function with all mass at zero} \end{cases}$$

  - $\delta_k \sim Bernoulli(p)$
  - $p \sim Beta(a_0, c_0)$

- $\sigma_g^2 \sim InverseGamma(\frac{\nu}{2}, \frac{\nu\lambda}{2})$

Note that β_δ denotes all $\beta_{g,k}$'s where $\delta_k = 1$, while $\beta_{-\delta}$ denotes all $\beta_{g,k}$'s where $\delta_k = 0$.

Let the mean and variance of the slab prior be a g-slab as follows.

$$\beta_{0,\delta} = 0 \text{ and } B_{0,\delta} = g \cdot (X_\delta' X_\delta)^{-1} \tag{A.12}$$

where X_δ is the design matrix consisting of the predictor variables where $\beta_{g,k}$ corresponds to $\delta_k = 1$. Then given δ, the marginal likelihood given in equation (A.8) becomes

$$f(\mathbf{y_{g,c}}|X_g, \delta) = (\pi)^{-\frac{(T_g-1)}{2}} \cdot T_g^{-\frac{1}{2}} \cdot (\nu\lambda)^{-\frac{\nu}{2}} \cdot \left(\frac{|B_{1,\delta}|}{|B_{0,\delta}|}\right)^{\frac{1}{2}} \cdot \frac{\Gamma(\frac{(T_g-1)+\nu}{2})}{\Gamma(\frac{\nu}{2})} \cdot (A_\delta + \nu\lambda)^{-\frac{(T_g-1)+\nu}{2}} \tag{A.13}$$

where

$$\begin{aligned} A_\delta &= \mathbf{y_{g,c}}'\mathbf{y_{g,c}} + \beta_{0,\delta}' B_{0,\delta}^{-1} \beta_{0,\delta} - \beta_{1,\delta}' B_{1,\delta}^{-1} \beta_{1,\delta} \\ B_{1,\delta} &= (X_\delta' X_\delta + B_{0,\delta}^{-1})^{-1} \\ \beta_{1,\delta} &= B_{1,\delta}(X_\delta' \mathbf{y_{g,c}} + B_{0,\delta}^{-1}\beta_{0,\delta}) \end{aligned}$$

Now using $\beta_{0,\delta}$ and $B_{0,\delta}$ of the g-slab, one can further simplify equation (A.13) as follows. First,

$$B_{0,\delta}^{-1} = \frac{1}{g}(X_\delta' X_\delta)$$

leading to

$$B_{1,\delta} = (X_\delta' X_\delta + \frac{1}{g}(X_\delta' X_\delta))^{-1} = (\frac{g+1}{g} X_\delta' X_\delta)^{-1} = \frac{g}{g+1}(X_\delta' X_\delta)^{-1} \tag{A.14}$$

and

$$\beta_{1,\delta} = B_{1,\delta}(X_\delta' \mathbf{y_{g,c}} + \frac{1}{g}(X_\delta' X_\delta)\beta_{0,\delta}) = B_{1,\delta} X_\delta' \mathbf{y_{g,c}} \tag{A.15}$$

Further, A_δ can be expressed as follows.

$$\begin{aligned} A_\delta &= \mathbf{y_{g,c}}'\mathbf{y_{g,c}} + \beta_{0,\delta}' B_{0,\delta}^{-1} \beta_{0,\delta} - \beta_{1,\delta}' B_{1,\delta}^{-1} \beta_{1,\delta} \\ &= \mathbf{y_{g,c}}'\mathbf{y_{g,c}} - (B_{1,\delta} X_\delta' \mathbf{y_{g,c}})' B_{1,\delta}^{-1} (B_{1,\delta} X_\delta' \mathbf{y_{g,c}}) \\ &= \mathbf{y_{g,c}}'\mathbf{y_{g,c}} - \mathbf{y_{g,c}}' X_\delta B_{1,\delta}' B_{1,\delta}^{-1} B_{1,\delta} X_\delta' \mathbf{y_{g,c}} \\ &= \mathbf{y_{g,c}}'\mathbf{y_{g,c}} - \mathbf{y_{g,c}}' X_\delta B_{1,\delta} X_\delta' \mathbf{y_{g,c}} \end{aligned} \tag{A.16}$$

Finally, using the fact that $|a \cdot A| = a^K \cdot |A|$ where $dim(A) = K$, the ratio of determinants

of the prior and posterior variance of β_g can be simplified as follows.

$$\frac{|B_{1,\delta}|^{\frac{1}{2}}}{|B_{0,\delta}|^{\frac{1}{2}}} = \frac{|\frac{g}{g+1}(X_\delta' X_\delta)^{-1}|^{\frac{1}{2}}}{|g(X_\delta' X_\delta)^{-1}|^{\frac{1}{2}}} = \frac{(\frac{g}{g+1})^{\frac{K}{2}}}{g^{\frac{K}{2}}} = (g+1)^{-\frac{K}{2}} \tag{A.17}$$

This is because $B_{1,\delta}$ is a scalar multiple of $B_{0,\delta}$ as shown in equation (A.14) above. Therefore, the simplified conditional marginal likelihood given δ is

$$f(\mathbf{y_{g,c}}|X_g,\delta) = (\pi)^{-\frac{(T_g-1)}{2}} \cdot T_g^{-\frac{1}{2}} \cdot (\nu\lambda)^{\frac{\nu}{2}} \cdot (g+1)^{-\frac{K}{2}} \cdot \frac{\Gamma(\frac{(T_g-1)+\nu}{2})}{\Gamma(\frac{\nu}{2})} \cdot (A_\delta + \nu\lambda)^{-\frac{(T_g-1)+\nu}{2}} \tag{A.18}$$

where

$$A_\delta = \mathbf{y_{g,c}}' \mathbf{y_{g,c}} - \mathbf{y_{g,c}}' X_\delta B_{1,\delta} X_\delta' \mathbf{y_{g,c}}$$
$$B_{1,\delta} = \frac{g}{g+1}(X_\delta' X_\delta)^{-1}$$

B MCMC Sampling Procedure

Note that by the method of composition (MoC), one can get the posterior

$$\pi(\beta_g, \sigma_g^2, \delta_g | \mathbf{y_{g,c}}, p_g) = \pi(\beta_g | \mathbf{y_{g,c}}, p_g, \sigma_g^2, \delta_g) \cdot \pi(\sigma_g^2 | \mathbf{y_{g,c}}, p_g, \delta_g) \cdot \pi(\delta_g | \mathbf{y_{g,c}}, p_g) \tag{B.1}$$

B.1 Sampling δ_g

We first illustrate how to sample δ_g using the relation

$$\pi(\delta_g | \mathbf{y_{g,c}}, p_g) \propto f(\mathbf{y_{g,c}} | \delta_g, p_g) \cdot \pi(\delta_g | p_g)$$

Let us focus on $\delta_k \in \delta_g$. Then the posterior distribution of δ_k for each group is proportional to the likelihood and product of the prior distributions as follows.

$$\begin{aligned}\pi(\delta_k | \mathbf{y_{g,c}}, \beta_g, \sigma_g^2, \delta_{-k}, p_g) &\propto f(\mathbf{y_{g,c}} | X_g \beta_g, \sigma_g^2) \cdot \pi(\beta_g | \sigma_g^2, \delta_k, \delta_{-k}) \cdot \pi(\sigma_g^2) \\ &\quad \cdot \pi(\delta_k | p_g) \cdot \pi(\delta_k | p_g) \cdot \pi(p_g) \\ &\propto f(\mathbf{y_{g,c}} | \delta_k, \delta_{-k}) \cdot \pi(\delta_k | p_g)\end{aligned}$$

Using the property that δ_k is an indicator variable that is discrete,

$$\Pr[\delta_k = 1 | \delta_{-k}, \mathbf{y_{g,c}}] = \frac{\Pr[\delta_k = 1 | p_g] \cdot f(\mathbf{y_{g,c}} | \delta_k = 1, \delta_{-k})}{\Pr[\delta_k = 1 | p_g] \cdot f(\mathbf{y_{g,c}} | \delta_k = 1, \delta_{-k}) + \Pr[\delta_k = 0 | p_g] \cdot f(\mathbf{y_{g,c}} | \delta_k = 0, \delta_{-k})} \tag{B.2}$$

since $\Pr[\delta_k = 0 | \mathbf{y_{g,c}}, \delta_{-k}] + \Pr[\delta_k = 1 | \mathbf{y_{g,c}}, \delta_{-k}] = 1$. Now using the conditional marginal likelihood given δ_g in equation (A.18), equation (B.2) can be further simplified as follows.

$$\Rightarrow \frac{p_g \cdot (g+1)^{-\frac{K_1}{2}} \cdot (A_{\delta_1} + \nu\lambda)^{-\frac{(T_g - 1) + \nu}{2}}}{p_g \cdot (g+1)^{-\frac{K_1}{2}} \cdot (A_{\delta_1} + \nu\lambda)^{-\frac{(T_g - 1) + \nu\lambda}{2}} + (1 - p_g) \cdot (g+1)^{-\frac{K_0}{2}} \cdot (A_{\delta_0} + \nu\lambda)^{-\frac{(T_g - 1) + \nu}{2}}}$$

$$\Rightarrow \frac{p_g}{p_g + (1 - p_g) \cdot (g+1)^{\frac{1}{2}} \cdot \left(\frac{A_{\delta_0} + \nu\lambda}{A_{\delta_1} + \nu\lambda} \right)^{-\frac{(T_g - 1) + \nu}{2}}} \tag{B.3}$$

where K_0 is the number of predictor variables where $\delta_k = 0$, $K_1 = K_0 + 1$, $\delta_1 = \delta_g | (\delta_k = 1, \delta_{-k})$, $\delta_0 = \delta_g | (\delta_k = 0, \delta_{-k})$ leading to the fact that A_{δ_0} and A_{δ_1} is each A_δ derived for the $\delta_k = 1$ and $\delta_k = 0$ case.

B.2 Sampling σ_g^2

Next, σ_g^2 is sampled using the relation

$$\begin{aligned}\pi(\sigma_g^2|\mathbf{y_{g,c}}, p_g, \delta_g) &\propto f(\mathbf{y_{g,c}}|\sigma_g^2, \delta_g, p_g) \cdot \pi(\sigma_g^2) \\ &\propto f(\mathbf{y_{g,c}}|\sigma_g^2, \delta_g) \cdot \pi(\sigma_g^2) = f(\mathbf{y_{g,c}}, \sigma_g^2|\delta_g) \end{aligned} \tag{B.4}$$

Note that equation (B.4) can be further expressed from the process of integrating out σ_g^2 in Appendix A as follows.

$$\pi(\sigma_g^2|\mathbf{y_{g,c}}, \delta_g) \propto \mathcal{IG}(\frac{\alpha_1}{2}, \frac{\delta_1}{2}) \tag{B.5}$$

where $\alpha_1 = \nu + T_g$ and $\delta_1 = A_\delta + \nu\lambda$ with $A_\delta = \mathbf{y_{g,c}}'\mathbf{y_{g,c}} - \mathbf{y_{g,c}}'X_\delta B_{1,\delta} X_\delta' \mathbf{y_{g,c}}$ in equation (A.18).

B.3 Sampling β_g

One can sample β_g using the relation

$$\begin{aligned}\pi(\beta_g|\mathbf{y_{g,c}}, p_g, \sigma_g^2, \delta_g) &\propto f(\mathbf{y_{g,c}}|\beta_g, \sigma_g^2, \delta_g, p_g) \cdot \pi(\beta_g|\sigma_g, \delta_g, p_g) \\ &\propto f(\mathbf{y_{g,c}}|\beta_g, \sigma_g, \delta_g) \cdot \pi(\beta_g|\sigma_g^2, \delta_g) = f(\mathbf{y_{g,c}}, \beta_g|\sigma_g^2, \delta_g) \end{aligned} \tag{B.6}$$

Let $\beta_{g,-\delta}$, i.e. $\beta_{g,k}$'s where $\delta_{g,k} = 0$, as zero. Then the full conditional distribution of $\beta_{g,\delta}$ from equation (B.6) can be further expressed from the process of integrating out β_g in Appendix A above as follows.

$$\begin{aligned}\pi(\beta_{g,\delta}\mathbf{y_{g,c}}, \sigma_g^2) &\propto f(\mathbf{y_{g,c}}|X_g\beta_{g,\delta}) \cdot \pi(\beta_{g,\delta}) \\ &= \mathcal{N}(\mathbf{y_{g,c}}|X_\delta\beta_{g,\delta}) \cdot \mathcal{N}(\beta_{0,\delta}, \sigma_g^2 \cdot B_{0,\delta}) \\ &= \mathcal{N}(\beta_{1,\delta}, \sigma_g^2 \cdot B_{1,\delta}) \end{aligned} \tag{B.7}$$

where $\beta_{1,\delta} = B_{1,\delta}(X_\delta'\mathbf{y_{g,c}} + B_{0,\delta}^{-1}\beta_{0,\delta})$ and $B_{1,\delta} = (X_\delta'X_\delta + B_{0,\delta}^{-1})^{-1}$. Under the g-slab, $\beta_{1,\delta}$ and $B_{1,\delta}$ can be expressed as follows.

$$\beta_{1,\delta} = B_{1,\delta}X_\delta'\mathbf{y_{g,c}}$$

and

$$B_{1,\delta} = \frac{g}{g+1}(X_\delta'X_\delta)^{-1}$$

Thus,

$$\beta_{g,\delta}|\sigma_g^2, \delta_g, p_g \sim \mathcal{N}(\beta_{1,\delta}, \sigma_g^2 \cdot B_{1,\delta}) \tag{B.8}$$

B.4 Sampling p_g

Finally, the full conditional distribution of p_g for each group (or terminal node) is proportional to the likelihood and the product of prior distributions as follows.

$$\begin{aligned}
\pi(p_g|\beta_g, \sigma_g^2, \delta_g, \mathbf{y_{g,c}}) &\propto f(\mathbf{y_{g,c}}|X_g\beta_g, \sigma_g^2) \cdot \pi(\beta_g|\sigma_g^2, \delta_g) \cdot \pi(\sigma_g^2) \cdot \pi(\delta_g|p_g) \cdot \pi(p_g) \\
&\propto \pi(\delta_g|p_g) \cdot \pi(p_g) \\
&= Bernoulli(p_g) \cdot Beta(a_0, c_0) \\
&= p_g^{K_1} \cdot (1-p_g)^{K_0} \cdot \frac{\Gamma(a_0+c_0)}{\Gamma(a_0)+\Gamma(c_0)} \cdot p_g^{a_0-1} \cdot (1-p_g)^{c_0-1} \\
&= Beta(a_0+K_1, c_0+K_0)
\end{aligned}$$

where K_0 is the number predictor variables where $\delta_k = 0$ and K_1 is the number of predictor variables where $\delta_k = 1$.

C Failure of the Gibbs Sampler

Let the parameters, denoted as θ, in the Gibbs-sampler form a Markov chain with a $3K+2$ dimensional state space as follows.

$$\Theta = \mathbb{R}^{K+1} \times (0, +\infty) \times \{0,1\}^K \times (0,1)^K$$

For simplicity, let as assume that the parameters are subdivided into two blocks, each denoted as θ_1 and θ_2. That is, the state space can be expressed as a Cartesian product $\Theta = \Theta_1 \times \Theta_2$. Also, the transition probability kernel $P : \Theta \times \mathcal{B}(\Theta) \to [0,1]$ of the Gibbs-sampler is defined as

$$P(\theta, A) = \int_A \pi(t_1|Y, \theta_2) \cdot \pi(t_2|Y, t_1) dt$$

for any $\theta \in \Theta$ and Borel set $A \in \mathcal{B}(\Theta)$.

First, we can show that the posterior distribution is a stationary distribution for the transition probability kernel P as follows. For any Borel set $A \in \mathcal{B}(\Theta)$,

$$\begin{aligned}
\int_\Theta P(\theta, A) \cdot \pi(\theta|Y) d\theta &= \int_\Theta \int_A \pi(t_1|Y, \theta_2) \cdot \pi(t_2|Y, t_1) \cdot \pi(\theta_1, \theta_2|Y) dt d\theta \\
&= \int_A \int_{\Theta_2} \int_{\Theta_1} \frac{\pi(t_1, \theta_2|Y)}{\pi(\theta_2|Y)} \cdot \frac{\pi(t_1, t_2|Y)}{\pi(t_1|Y)} \cdot \pi(\theta_1, \theta_2|Y) d\theta_1 d\theta_2 dt \\
&= \int_A \pi(t|Y) \cdot \left(\int_{\Theta_2} \int_{\Theta_1} \pi(\theta_1|\theta_2, Y) \cdot \pi(\theta_2|t_1, Y) d\theta_1 d\theta_2 \right) dt \\
&= \int_A \pi(t|Y) \cdot \left(\int_{\Theta_2} \pi(\theta_2|t_1, Y) \cdot \left(\int_{\Theta_1} \pi(\theta_1|\theta_2, Y) d\theta_1 \right) d\theta_2 \right) dt \\
&= \int_A \pi(t|Y) \cdot \left(\int_{\Theta_2} \pi(\theta_2|t_1, Y) d\theta_2 \right) dt \\
&= \int_A \pi(t|Y) dt
\end{aligned}$$

Thus, the posterior density $\pi(\cdot)$ is stationary for $P(\cdot)$, by definition.

Next, we check for the convergence of the Markov chain. For the Markov chain with transition probability $P(\cdot)$ to converge to its stationary distribution $\pi(\cdot)$, it must satisfy the two sufficient conditions such that

1. Aperiodicity: The chain must not visit disjoint subsets of Θ periodically.

2. Irreducibility: The chain must visit any subset of Θ that has a positive probability under $\pi(\cdot)$ regardless of its initial point.

Under the Dirac spike-and-slab hierarchical prior, however, the irreducibility condition is not satisfied with the Gibbs-sampler. This can be shown as follows. Suppose $\delta_1 = 0$ at the initial point θ and run the Gibbs-sampler. Then the full-conditional distribution of β_1 is given as

$$\beta_1 | Y, \theta_{-\beta_1} \sim \Delta_0(\beta_1)$$

where Δ_0 is the Dirac spike with all its mass at zero. Thus, one will sample $\beta_1 = 0$. However, given $\beta_1 = 0$, the full conditional probability of $\delta_1 = 0$ is

$$\Pr[\delta_1 = 0 | Y, \theta_{-(\delta_1,\beta_1)}, \beta_1 = 0] = \frac{\Pr[\beta_1 = 0 | Y, \theta_{-(\delta_1,\beta_1)}, \delta_1 = 0] \cdot \Pr[\delta_1 = 0 | Y, \theta_{-(\delta_1,\beta_1)}]}{\Pr[\beta_1 = 0 | Y, \theta_{-(\delta_1,\beta_1)}]} \quad \text{(C.1)}$$

Specifically, the denominator of equation (C.1) above can be further expressed as follows.[25]

$$\Pr[\beta_1 = 0] = \Pr[\beta_1 = 0 | \delta_1 = 0] \cdot \Pr[\delta_1 = 0] + \Pr[\beta_1 = 0 | \delta_1 = 1] \cdot \Pr[\delta_1 = 1]$$

Since the full conditional distribution of β_1 given $\delta_1 = 1$ is absolutely continuous, we must have $\Pr[\beta_1 = 0 | \delta_1 = 1] = 0$, which implies that

$$\Pr[\beta_1 = 0] = \Pr[\beta_1 = 0 | \delta_1 = 0] \cdot \Pr[\delta_1 = 0]$$

Therefore,

$$\Pr[\delta_1 = 0 | Y, \theta_{-(\delta_1,\beta_1)}, \beta_1 = 0] = 1 \quad \text{(C.2)}$$

This implies that given $\beta_1 = 0$, δ_1 is always sampled as zero. In other words, whenever we start from the initial point where $\delta_1 = 0$, the Gibbs-sampler never enters the subset of Θ where $\beta_1 \neq 0$ or $\delta_1 = 0$. Thus, the irreducibility condition is not satisfied and the Gibbs-sampler does not converge to the joint posterior under the Dirac spike-and-slab prior framework.

[25] From now on, we omit the conditioning on Y and $\theta_{-(\delta_1,\beta_1)}$ for notational simplicity.

D MH Ratio for each Move Rule

In this section, we show that regarding the move rule in the MH algorithm one can get a different equation for the MH ratio. Recall that the MH ratio for the j-th iteration was

$$\alpha(\mathcal{T}^{(j-1)}, \mathcal{T}^{(j)}) = \min\left\{ \underbrace{\frac{P(Y|X,\mathcal{T}^{(j)})}{P(Y|X,\mathcal{T}^{(j-1)})}}_{\text{(i)}} \cdot \underbrace{\frac{P(\mathcal{T}^{(j)})}{P(\mathcal{T}^{(j-1)})}}_{\text{(ii)}} \cdot \underbrace{\frac{q(\mathcal{T}^{(j-1)}|\mathcal{T}^{(j)})}{q(\mathcal{T}^{(j)}|\mathcal{T}^{(j-1)})}}_{\text{(iii)}}, 1 \right\}$$

where (i) is the likelihood ratio, (ii) is the prior ratio and (iii) is the ratio between moving to the previous tree and moving to the new tree.

D.1 Growing Rule

Recall that if the growing rule is chosen, a randomly picked terminal node is split into two new nodes by the splitting rule p_{SPLIT}. Note that $\mathcal{T}^{(j-1)}$ is the previously generated tree while $\mathcal{T}^{(j)}$ is the newly grown tree. First, the likelihood ratio is

$$\frac{P(\mathcal{T}^{(j)}|X,Y)}{P(\mathcal{T}^{(j-1)}|X,Y)} = \frac{Lik_{new}}{Lik_{prev}} \equiv \frac{Lik_1}{Lik_2}$$

while the prior ratio is

$$\frac{P(\mathcal{T}^{(j)})}{P(\mathcal{T}^{(j-1)})} = \frac{\prod_\eta p_{split}(\eta)}{\prod_{\eta'} p_{split}(\eta')} \equiv \frac{Prior_1}{Prior_2}$$

where each prior density is obtained by equation (6).

Then the last term is consisted of the probability of "growing" a new tree given the previous tree

$$q(\mathcal{T}^{(j)}|\mathcal{T}^{(j-1)}) = \frac{1}{m} \times \frac{1}{p} \times \frac{1}{q}$$

and the probability of "pruning" the previous tree given the new tree

$$q(\mathcal{T}^{(j-1)}|\mathcal{T}^{(j)}) = \frac{1}{m} \times \frac{\text{Prune as Previous Tree}}{\text{All Possible Pruning Cases}} = \frac{1}{m} \times \frac{1}{\sum_{k=1}^{K} die_k}$$

where k is the index of each node and die_k is the nodes of the tree that can be pruned. For example, the root node cannot be pruned, which implies that $die_1 = 0$. Then $\sum_{k=1}^{K} die_k$ implies the sum of all possible pruning cases.

D.2 Pruning Rule

If the pruning rule is chosen, a randomly picked parent node is turned into a terminal node by eliminating its child nodes. Note that now $\mathcal{T}^{(j-1)}$ is the previously generated tree while $\mathcal{T}^{(j)}$ is the newly pruned tree. Then the last term is consisted of the probability of "pruning" a new tree given the previous tree

$$q(\mathcal{T}^{(j)}|\mathcal{T}^{(j-1)}) = \frac{1}{m} \times \frac{\text{Prune as Previous Tree}}{\text{All Possible Pruning Cases}} = \frac{1}{m} \times \frac{1}{\sum_{k=1}^{K} die_k}$$

and the probability of "growing" the previous tree given the new tree

$$q(\mathcal{T}^{(j-1)}|\mathcal{T}^{(j)}) = \frac{1}{m} \times \frac{1}{p} \times \frac{1}{q}$$

Note that the ratio of $q(\cdot)$ is simply the inverse of that when the growing rule is chosen. In this sense, the growing rule and pruning rule are counterparts of each other.

D.3 Changing Rule

If the changing rule is chosen, a randomly picked internal node is randomly reassigned to a new splitting rule. Now $\mathcal{T}^{(j-1)}$ is the previously generated tree while $\mathcal{T}^{(j)}$ is the newly changed tree. The last term is consisted of the probability of "changing" to a new tree given the previous tree

$$q(\mathcal{T}^{(j)}|\mathcal{T}^{(j-1)}) = \frac{1}{m} \times \frac{1}{p} \times \frac{1}{q}$$

and the probability of "re-changing" to the previous tree given the newly changed tree

$$q(\mathcal{T}^{(j-1)}|\mathcal{T}^{(j)}) = \frac{1}{m} \times \frac{1}{p} \times \frac{1}{q}$$

One can notice that if the changing rule is applied, the calculation of the $q(\cdot)$ ratio can be ignored since it is simply 1.

D.4 Swapping Rule

If the swapping rule is chosen, the splitting rule of a randomly picked parent and child node is swapped. Now $\mathcal{T}^{(j-1)}$ is the previously generated tree while $\mathcal{T}^{(j)}$ is the newly swapped tree. The last term is consisted of the probability of "swapping" to a

new tree given the previous tree

$$q(\mathcal{T}^{(j)}|\mathcal{T}^{(j-1)}) = \frac{1}{m} \times \frac{1}{p} \times \frac{1}{q}$$

and the probability of "re-swapping" to the previous tree given the newly swapped tree

$$q(\mathcal{T}^{(j-1)}|\mathcal{T}^{(j)}) = \frac{1}{m} \times \frac{1}{p} \times \frac{1}{q}$$

Again, if the swapping rule is applied, the $q(\cdot)$ ratio is 1. Note that although implicitly shown, the changing rule and swapping rule are counterparts of each other as the growing and pruning rule case.

E Number of Principal Components for each Out-of-sample

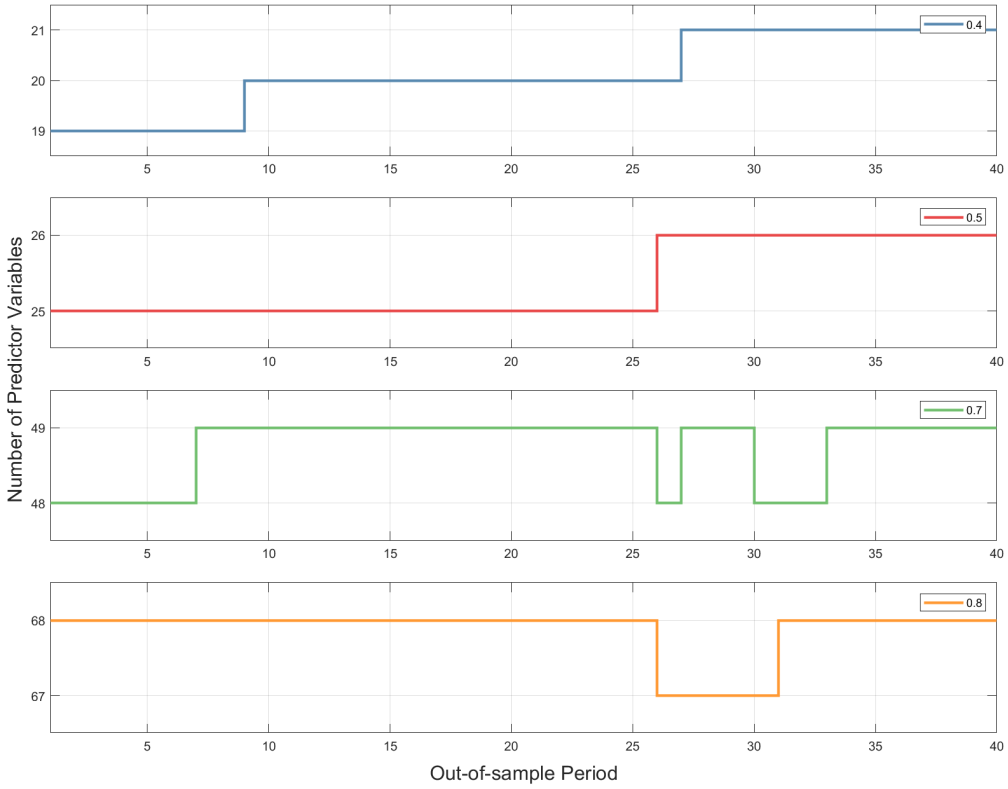


Figure 6: Number of Principal Components across Out-of-sample

Note. This figure depicts the number of principal components across 40 out-of-sample periods for each criterion of constructing the principal components. One can observe that the number of PCs varies across the out-of-sample period where the 26th period is 2020 Q3 and the 27th period is 2020 Q4.

In this section, we depict the number of principal components obtained across each out-of-sample period. First, in **Figure 7** above, we show the number of PCs across the out-of-sample period regarding the amount of total variation of each group explained by the PCs. Next, given that we select 0.4 as the criterion for constructing the PCs, the specific number of PCs obtained for each category group is depicted in **Table 5** below.

Table 5: Set of Predictor Variables across Out-of-sample Periods

	$1-8$	$9-26$	$27-40$
1	Y(-1)	Y(-1)	Y(-1)
2	Y(-2)	Y(-2)	Y(-2)
3	Y(-3)	Y(-3)	Y(-3)
4	Y(-4)	Y(-4)	Y(-4)
5	NIPA1	NIPA1	NIPA1
6	NIPA2	NIPA2	NIPA2
7	IP1	IP1	IP1
8	EM1	EM1	EM1
9			EM2
10	HOUSE1	HOUSE1	HOUSE1
11	INV1	INV1	INV1
12	PRICE1	PRICE1	PRICE1
13	PRICE2	PRICE2	PRICE2
14	EAPROD1	EAPROD1	EAPROD1
15	EAPROD2	EAPROD2	EAPROD2
16	INT1	INT1	INT1
17	M&C1	M&C1	M&C1
18	M&C2	M&C2	M&C2
19		M&C3	M&C3
20	HBAL1	HBAL1	HBAL1
21	EXCH1	EXCH1	EXCH1
22	STOCK1	STOCK1	STOCK1
23	NHBAL1	NHBAL1	NHBAL1
24	NHBAL2	NHBAL2	NHBAL2
25	SENTI1	SENTI1	SENTI1

F Selected Groups for B-CART Model

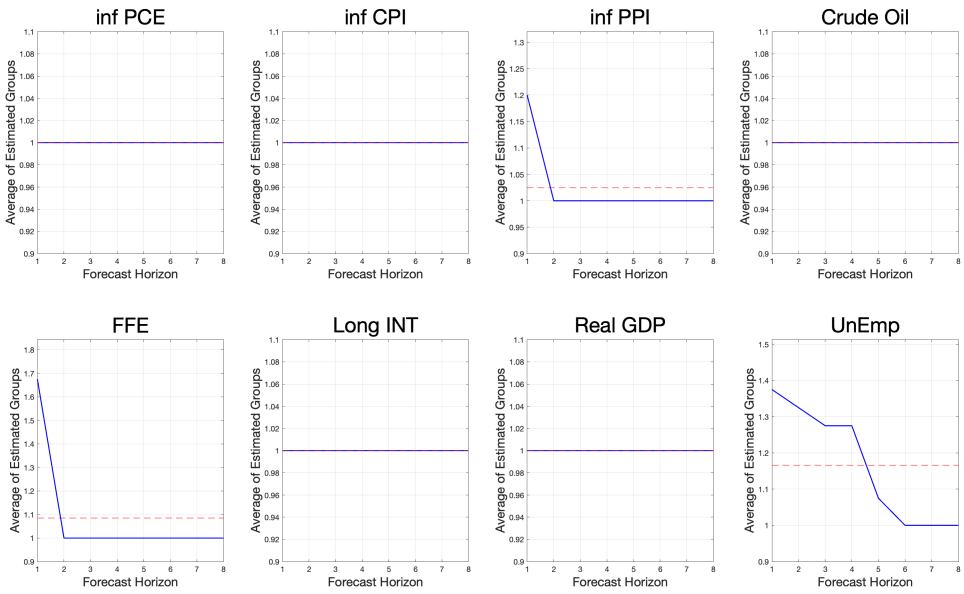


Figure 7: Average Groups of Variables across each Forecast Horizon for B-CART

Note. This figure depicts the estimated average groups for macroeconomic variables of the B-CART model. The blue line is the average groups across each forecast horizon, while the red dotted line is the average groups among all 8 forecast horizon.

The average number of groups obtained from the B-CART model across the eight macroeconomic variables and forecast horizons is depicted in **Figure 8** above. The average is derived by the sample mean of the number of groups across the optimal trees obtained for 40 out-of-sample periods. It was found that only a single group was estimated on average for macroeconomic variables, except for the PPI price index, FFE and unemployment rate. Even for these three variables, where some degree of grouping was observed, the average number of groups across the 40 out-of-sample periods quickly converged to one as the forecasting horizon increased. In other words, the overall sample period was not partitioned by the tree structure when using the B-CART model. This implies that estimating variable selection and the optimal tree structure jointly creates a compounding effect. To emphasize, incorporating variable selection is essential for accurately estimating the optimal tree structure when the timing and source of instability

among variables are unknown.