# Estimating Expected Return Function Nonparametrically: Based on BART *

Yaohan Chen

School of Economics, Singapore Management University

Last version: August 20, 2019

This version: June 4, 2020

**Abstract**

This paper documents the empirical Implementation of estimating expected return function nonparametrically using Bayesian Additive Regression Tree (BART) method. Within this newly introduced nonparametric framework, general non-linearity is allowed for the specification of model when the dimension of covariates used for prediction is large and the underlying non-linear relationship is hard to detect. By applying BART, we document which firm-level characteristics should be adopted as the most influential predictors for estimating expected return and the out-of-sample performance of BART for prediction as well. I have also extended the whole framework to China stock market and global financial market for empirically comparison. Our finding suggest that (i) the performance of BART approximates the results obtained from neural-network, which is a specific machine-learning method documented with dominating out-of-sample prediction performance; (ii) Machine-learning based method (specifically BART) surely outperform the benchmark linear model, but in terms of investment strategy constructed from prediction, there is not much significant difference between machine-learning methods and linear benchmark; (iii) China stock market is relatively more predictable in comparison to the U.S. stock market in terms of out-of-sample prediction-accuracy measure.

**Keywords:** Expected Return; Prediction; Cross-sectional asset pricing; BART; Machine Learning

* All errors are my own. You may contact me at yaohan.chen.2017@phdecons.smu.edu.sg

# 1 Introduction

Prediction has always been the concern in finance research. One of the two Formidable challenges faced within community is how to increase the out-of-sample prediction accuracy and the other is the high-dimension of potential predictors to be used for prediction. To address the first issue, potential nonliearity of model has to be adopted so that to allow model mimic data as possible as it could. That is one reason it becomes more popular recently to set up nonparametric models for the purpose of making prediction and estimating the expected return. The second issue has attracted more concern given that many available firm-level characteristics have been proposed in literature. Cochrane (2017) points out this multidimensional challenge in this lecture that "Which characteristics really provide independent information about average returns and Which are subsumed by others". Harvey et al. (2016) points out the $p$-hacking issues in applying the conventional statistic-test based method to identify characteristics and apply their adjusted $p$ values to identify firm-level characteristics that potentially affect the expected returns. All these recent concerns are actually about model uncertainty or the fundamentally unknown data generating process of return in stock market. This issue has been noticed by some earlier literature such as (Cremers, 2002; Stambaugh, 1999; Stambaugh and Pastor, 2000) and some solutions from Bayesian perspective have been proposed as well. Basically, To address the high-dimension problem, some model selection techniques should be adopted. Conventionally, researchers tend to solve these two problems either separately or simultaneously in linear framework. However, with the development of modern computational power and statistical algorithms, many researches in this field begin to apply some complex models to make prediction and covariates selection simultaneously while allowing nonlinearity (see Freyberger et al., 2019). Machine learning method is one of these attempts made in community of current era.

Machine learning method usually has good performance in terms of out-of-sample fitting. It is flexible given the generic nonparametric setting, which is essentially the feature for many machine learning algorithms. Moreover, It is closely associated with the goal of selecting predictors and making return prediction, which makes it an important tool for understanding the behaviour of risk premia. Gu et al. (2019) makes a comprehensive attempt in applying and comparing many major machine learning methods to address the return predictability issues in empirical finance and reach the conclusion that machine learning method like Neural network and regression trees are able to gain much predictive advantage for predicting returns in financial market. Kozak et al. (2018) applies a Bayesian method by imposing an economically motivated prior on stochastic discount factor (SDF) to reveal how machine learning method (specifically, the penalized regression like LASSO method) can be connected with conventional factor model to provide a characteristics-sparse stochastic discount factor (SDF) which is able to summarize the explanatory power of cross-sectional stock return predictions in high-dimensional setting. Similar researches on applying machine learning method to analyse cross-sectional return include but not restricted to (Freyberger et al., 2019; Chinco et al., 2019; Han et al., 2019; Chen et al., 2019).

This paper just follows the route by applying a recently rising machine learning method based on decision tree ensembles for prediction and variable selection for high dimensional dataset. It has been documented in literature that BART usually has good performance in practice and Linero (2018) extends the pioneering work on Bayesian additive regression trees (BART) model by constructing prior on decision tree ensembles with the a sparsity-inducing Dirichlet hyper-prior for addressing variable selection issue. This newly introduced sparsity-inducing Dirichlet prior in Linero (2018) makes BART more suitable for selecting variables and is less likely to overfit. Recently, there is also ongoing work to theoretically justify BART (Ročková and van der Pas, 2019; Ročková, 2019; Ročková and Saha, 2019), which makes it a promising future that BART is theoretically grounded . Therefore, given the nice performance and properties of BART documented in statistical literature, it is meaningful to make an attempt to apply BART to a large set of documented firm-level characteristics to identify potentially useful variables for the expected return, especially for addressing the conventional concern about improving out-of-sample prediction accuracy as well as the current increasing concern of dimension reduction of covariates and selecting the most influential variables for general functional form of expected return.

The rest of this paper is structured as follows: Section 2 discusses the general background of estimating expected return and one recent proposed method in literature, which is essentially a combination of sieve-based nonparametric algorithm with the adaptive group LASSO method. Main procedure about the implementation of this proposed method and limitation of this method is discussed in this section. Section 3 describes and discusses the general implementing procedure of BART. Monte Carlo study is discussed in Section 4 for demonstrating the potential ability of BART in selecting influential variables in complex model settings and nice performance of BART in terms of making out-of-sample prediction in comparison to some existing methods as well. Section 5 applies BART to real financial data to select influential firm-level characteristics for the expected return. Out-of-sample prediction performance of BART based on out-of-sample $R$ square is discussed as well. Section 6 concludes.

## 2 Expected Return and Current Methods

### 2.1 Expected return

One major concern in empirical asset-pricing literature is to identify the firm-level characteristics of period $t-1$ which are useful for predicting returns of period $t$. Generically, it is collected in the following functional form of conditional mean

$$m_t\left(c_1,\ldots,c_S\right)=\mathbb{E}\left[R_{it}\mid C_{1,it-1}=c_1,\ldots,C_{S,it-1}=c_S\right] \tag{1}$$

where $C_{1,it-1},\ldots,C_{S,it-1}$ are the $S$ characteristics of firm $i$ in period $t-1$.

## 2.2 Current methods

The generic problem described in (1) can be approximated and estimated nonparametrically. However, if the number of available characteristics is large (which is common in finance research given that the long-time development in cross-sectional asset-pricing literature and that many potential firm-level characteristics has been proposed). How to dissect the functional form of expected return and meanwhile select the characteristics potentially useful for prediction is nascent in recent empirical finance literature. Freyberger et al. (2019) proposed a semiparametric method with LASSO method plugged in for variable selection as following

$$m_t\left(c_1,\ldots,c_S\right)=\sum_{s=1}^{S}m_{ts}\left(c_s\right) \tag{2}$$

with each $m_{ts}\left(c_s\right)$ modelled nonparametrically by approximating it using some basis functions. Specifically, Freyberger et al. (2019) apply spline function as basis functions and their method is mainly based on the following steps.

**Step 1:** Doing the following optimization

$$\tilde{\boldsymbol{\beta}}_t \;=\; \underset{\substack{b_{sk}:s=1,\cdots,S;\\ k=1,\cdots,L+2}}{\arg\min}\;\sum_{i=1}^{N}\left(R_{it}-\sum_{s=1}^{S}\sum_{k=1}^{L+2}b_{sk}p_k\left(\tilde{C}_{s,it-1}\right)\right)^2+\lambda_1\sum_{s=1}^{S}\left(\sum_{k=1}^{L+2}b_{sk}^2\right)^{\frac{1}{2}} \tag{3}$$

where $\tilde{\boldsymbol{\beta}}_t$ is a $(L+2)\times S$ matrix and $\lambda_1$ is the associated penalty parameter in LASSO. Where $p_k(c)$ is selected from the class of *quadratic spline* function such that

$$p_1(c)=1,\;\; p_2(c)=c\;, p_3(c)=c^2,\;\text{and}\;\; p_k(c)=\max\{c-t_{k-3},0\}^2,\;\;\text{for}\;\; k=4,\cdots,L+2.$$

and

$$0=t_0<t_1<\cdots<t_{L-1}<t_L=1$$

is an increasing numbers between $[0,1]$ similar to portfolios breakpoints. $\lambda$ is selected in a data-dependent way to minimize Bayesian Information Criterion (BIC), which is suggested by Yuan and Lin (2006).

**Step 2:** Doing the second optimization and first define

$$w_{ts}=\begin{cases}\left(\sum_{k=1}^{L+2}\tilde{\beta}_{sk}^2\right)^{-\frac{1}{2}} & \text{if}\quad \sum_{k=1}^{L+2}\tilde{\beta}_{sk}^2\neq 0\\ \\ \infty & \text{if}\quad \sum_{k=1}^{L+2}\tilde{\beta}_{sk}^2=0\end{cases}$$

then solve

$$\check{\boldsymbol{\beta}}_t = \underset{\substack{b_{sk}:s=1,\cdots,S;\\ k=1,\cdots,L+2}}{\arg\min} \sum_{i=1}^{N} \left( R_{it} - \sum_{s=1}^{S} \sum_{k=1}^{L+2} b_{sk} p_k \left( \tilde{C}_{s,it-1} \right) \right)^2 + \lambda_2 \sum_{s=1}^{S} \left( w_{ts} \sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}} \tag{4}$$

Note that $\lambda_2$ is selected as to minimize BIC and $\check{\boldsymbol{\beta}}_t$ is $(L+2) \times S$ matrix as well.

**Step 3:** Denote $\hat{\boldsymbol{\beta}}_{ts}$ as the $s$ column selected from $\check{\boldsymbol{\beta}}_t$, and $\hat{\beta}_{tsk}$ as the $k$' th element of this column, then the estimation for $m_{ts}(\cdot)$ is

$$\widehat{\widehat{m}}_{ts}(\tilde{c}) = \sum_{k=1}^{L+2} \hat{\beta}_{tsk} p_k(\tilde{c}) \tag{5}$$

■

Difficulties and potential problems as well for this framework is mainly due to the required optimization in **Step 1** and **Step 2** since essentially it is needed to do optimization over a large dimensional space; Moreover, the cross-validation to be used for selecting the associated penalty parameter $\lambda_1$ and $\lambda_2$ also makes this algorithm computationally heavier. Another shortcoming of this framework is the partial linear setting implying implicit exclusion of cross dependencies since

$$\frac{\partial^2 m_t\left(c_1, \ldots, c_S\right)}{\partial c_s \partial c_{s'}} = 0 \quad \forall s \neq s'.$$

This issue has to be solved by adding certain intersections as additional regressors. Therefore, a more flexible and general nonparametric setting is desired for the purpose of selecting useful predictors and making predictions by taking the underlying data nonlienarity into account under the general functional form specification. To see the why it is necessary to take nonlinearity and intersection between characteristics into account, we may briefly discuss here based on Fama and French (2015) where they set up the connections of the following four terms,

$$\begin{aligned} \frac{M}{B} &: \text{market to book ratio} \\ \mathbb{E}\left[Y\right] &: \text{firm's expected earnings} \\ \Delta B &: \text{investment} \\ r &: \text{discount rates (return in general as well)} \end{aligned}$$

as jointly summarized as

$$\frac{M_t}{B_t} = \frac{1}{B_t} \sum_{\tau=1}^{\infty} \frac{\mathbb{E}\left(Y_{t+\tau} - \Delta B_{t+\tau}\right)}{(1+r)^{\tau}}. \tag{6}$$

This connection as demonstrated in (6) actually leads to the following implications by holding everything else unchanged

(i) A lower value of $M_t$, or equivalently high book-to-market ratio, $\frac{B_t}{M_t}$, implies a higher expected return, corresponding to $r$.

(ii) Higher expected future earnings imply a high expected return.

(iii) Higher expected growth in book equity (investment) implies a lower expected return.

These implications directly imply that characteristics like book-to-market ratio, firm's expected earnings, and investment must predict future equity returns; but the unknown functional form of $\mathbb{E}[\cdot]$ and the dependence of discount rates on these characteristics lead to highly nonlinear form in general and that different characteristics of firms are interacted with each other. Thus the non-linearities and interactions between characteristics are necessarily to be taken into consideration. This is also the motivation of this project to turn to focus on nascent BART algorithm, which is naturally of nonparametric form and generally does not implicitly exclude the intersections between covariates, given the excellent performance of BART in practice.

# 3 Bayesian Additive Regression Tree

## 3.1 General additive tree model

Bayesian Additive Regression Tree (BART) as a kind of general nonparametric statistical model setting is based on the early work of ensembles of decision trees and regression trees including Breiman (1991) and Breiman (2001). It could be treated within nonparametric statistical modelling framework since essentially the generic idea of regression tree is by employing a lot of step functions to approximate unknown functional forms. Corresponding discussion about implementation algorithm under Bayesian framework can be traced back to the pioneering work as Chipman et al. (1998); Denison et al. (1998); Chipman et al. (2010). The general idea is to approximate the functional form of interest $f_0(\boldsymbol{x})$ using the following random sum of decision trees $f(\boldsymbol{x})$

$$f(\boldsymbol{x}) = \sum_{t=1}^{T} \mathcal{T}_t(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^P$$

and for a specific observation $y$, it is assumed that

$$y = f(\boldsymbol{x}) + \epsilon, \qquad \epsilon \sim \sigma \cdot \mathcal{N}(0,1). \ [1]$$

Each regression tree $\mathcal{T}_t(\boldsymbol{x})$ is constructed from a series of binary trees, the structure of which is mainly determined by splitting rules (cutoff values) collected in a vector $\mathbf{c}_t$ and the associated terminal nodes collected in a vector $\mathcal{M}_t$ such that $\mathcal{T}_t(\boldsymbol{x}) = \mu_{tl}$ if $\boldsymbol{x}$ is associated with terminal node $l$ of tree $t$. [2] To illustrate this idea clearly, the following figure is used to demonstrate the basic

[1] $|f_0(\boldsymbol{x}) - f(\boldsymbol{x})|$ is better to be treated as the approximation error while $\epsilon$ captures the randomness of sampling error in general.

[2] cutoff values $c_{tj}$ are drawn uniformly from the collection of observed $x_{1j}, \ldots, x_{nj}$ given the $j$-th covariate is selected.

structure of a single binary tree.

**[Place Figure 1 about here]**

Suppose that $\boldsymbol{x}$ is of 2 dimension ($x_1$ and $x_2$) and a single binary tree $g(\cdot)$ is in 2 depth. Splitting rule is selected as $\boldsymbol{c} = (c_1, c_2)^\top$, where in this example $c_1 = 0.5$ and $c_2 = 0.3$. Given this structure, the 2 covariates are dissected into 3 categories and tree $g(\cdot)$ just take 3 constant values on the 3 different areas. By viewing this problem from the nonparametric perspective, it is an approximation of unknown functional form using a series of step functions in general. And it is also straightforward to see the connection between the underlying partition logic of BART with that of portfolio sorting common in finance. Where for this simple example by replacing $x_1$ with "size" and $x_2$ with "value" measure respectively, it is simply a sorting with three categories.

**Remark 3.1** Here we may slightly interchange the notation to follow convention in statistical literatures corresponding to BART. $f_0(\boldsymbol{x})$ plays the same role as the analogue of the general functional form of expected return expressed in (1). And $\boldsymbol{x}$ replace the vector of characteristics of firms. Thus $\boldsymbol{x}$ here is equivalent to the cross-sectionally observed firm-level characteristics $c_1, \ldots, c_S$ as in (1). Temporarily in this section for describing BART algorithm, vector $\boldsymbol{c}$ denotes the splitting rules.

### 3.2 Complete model with regularization prior

Under Bayesian framework, let $\boldsymbol{\theta} = (\sigma, \mathcal{T}, \mathcal{M})$ collect all the parameters in tree model where $\mathcal{T}$ denotes all the tree structure and $\mathcal{M}$ denotes all the parameters collected from all the trees used. Specifically,

$$\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_T\} \qquad \mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_T\}$$

with

$$\mathcal{M}_t = \{\mu_{t1}, \ldots, \mu_{tl}, \ldots, \mu_{tb_t}\}$$

and $b_t$ denotes number of terminal nodes used in tree $\mathcal{T}_t(\boldsymbol{x})$. Regularization prior for a specific tree model is then written as

$$\begin{aligned} \pi(\boldsymbol{\theta}) &= \pi\left((\mathcal{T}_1, \mathcal{M}_1), (\mathcal{T}_2, \mathcal{M}_2, \ldots, (\mathcal{T}_T, \mathcal{M}_T)), \sigma\right) \\ &= \pi(\sigma) \prod_{t=1}^{T} \pi\left(\mathcal{T}_t\right) \pi\left(\mathcal{M}_t \mid \mathcal{T}_t\right). \end{aligned} \tag{7}$$

#### 3.2.1 Prior for $\mathcal{T}_t$

Each tree $\mathcal{T}_t(\boldsymbol{x})$ is generated from a sequence of binary tree by imposing the prior for tree structure as following

$$q(d) : \mathbb{N} \to [0, 1].$$

where $d$ denotes the tree depth ranging from $d = 0, \cdots, \infty$. Specifically, each tree is constructed starting from a single node at depth $d = 0$ and the two child nodes of depth $d + 1$ oriented from the node of depth $d$ is generated with probability $q(d)$ or the node of depth $d$ is terminal otherwise. This process iterates onwards for $d = 1, \cdots, \infty$. A common choice for $q(d)$ (Chipman et al., 1998; 2010) is

$$q(d) = \frac{\gamma}{(1+d)^{\beta}} \quad \gamma \in (0,1), \ \beta \in [0, \infty) \tag{8}$$

**Example 3.1 (Single Tree Structure)** This is an example used to demonstrate how a single tree is specified based on probability rule as in (8) with $\gamma = .95$ and $\beta = 2$. If $d = 0$ and simply there is one terminal node, it is by specification with probability $1 - .95 = .05$; If $d = 1$ and the tree is terminated at $d = 1$, then by specification it is with probability $.95 \times (1 - .95/2^2)^2 = .55$; If $d = 2$ and the tree is terminated at $d = 2$, the number of terminal nodes could be either 3 or 4 corresponding to the two cases illustrated below

**[Place Figure 2 about here]**

If tree is with 3 terminal nodes as in **(a)**, the probability for this structure is $.95 \times (1 - .95/2^2) \times \frac{.95}{2^2} \times \left(1 - \frac{.95}{3^2}\right)^2 \times 2 = 0.28$; or the tree with 4 terminal nodes: The probability assigned for the tree structures listed in **(b)-(e)** is $.95 \times (1 - .95/2^2) \times (.95/2^2) \times (0.95/3^2) \times (1 - .95/4^2)^2 \times 4 = .06$; The probability assigned for tree structure listed in **(f)** is $.95 \times (.95/2^2)^2 \times (1 - .95/3^2)^4 = .03$. Hence, the probability for tree with four terminal nodes is $0.06 + 0.03 = 0.09$. Everything else remained is tree structure with more than 5 terminal nodes and it is with probability $1 - .05 - .55 - .28 - .09 = .03$.

Once the tree depth is determined by $q(d)$, the predictors as a subset of $\boldsymbol{x}$ are chosen according to the probability vector $\omega = (\omega_1, \ldots, \omega_P)$. Dirichlet distribution is adopted such that

$$\omega \sim \mathcal{D}(\alpha/P, \ldots, \alpha/P)^{3} \tag{9}$$

and

$$\frac{\alpha}{\alpha + \rho} \sim \text{Beta}(a, b).^{4}$$

All these prior information about depth of each tree and covariates used for splitting each tree together determine the prior $\pi(\mathcal{T}_t)$.

[3] Figure 3 demonstrates how the $\alpha$ plays the role in determining the weights contributed by different variables and how different $\alpha$ specifications may lead to sparsity for shrinkage and variable selection purpose. This is consistent witht practical decsion making of investors since if all the all the investors in market are Bayesian, homogeneous, risk-neutral and hence a shrinkage interpretation reflects informative prior structure such that the model parameters cannot have arbitraryily (this is also due to the economic plausibiliy considerations) and their posterior beliefs are shrunk towards zero.

[4] $a$,$b$,$\rho$ here are hyperparameters. By default, $\rho$ is possible to be selected as $\rho = P$ but $\rho$ is possible to be less than $P$ if theres is a strong priori support that predictors used are sparse.

#### 3.2.2 Prior for $\mathcal{M}_t \mid \mathcal{T}_t$

For each tree $\mathcal{T}_t(\boldsymbol{x})$ and given $\boldsymbol{x}$, $\mathcal{T}_t(\boldsymbol{x})$ takes one value from $\{\mu_{t1}, \ldots, \mu_{tl}, \ldots, \mu_{tb_t}\}$, denoted by $\mu_t$. For a given $\boldsymbol{\theta}$,

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \mu_1 + \mu_2 + \cdots + \mu_T.$$

Assuming $\mu_t \sim \mathcal{N}(0, \tau^2)$ i.i.d. then

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) \sim \mathcal{N}\left(0, T \cdot \tau^2\right).$$

Then Jeffreys prior is adopted such that

$$\tau \propto \frac{1}{\sqrt{T}}.^{5}$$

#### 3.2.3 Prior for $\sigma$

$$\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu} \tag{10}$$

This prior specification is usually for the conjugacy consideration which would bring much convenience for implementation. [6] For practical implementation, $\nu$ is selected and $\lambda$ is chosen in this way:

**Step 1** Roughly estimate $\sigma$, denoted by $\hat{\sigma}$

$$\hat{\sigma} = \begin{cases} \text{standard deviation of residual from OLS} & \text{if } p < n \\ \text{standard deviation of } \boldsymbol{y} & \text{if } p \geq n \end{cases}$$

**Step 2** Choose $\lambda$ so that $\hat{\sigma}$ is at a chosen quantile $q$ of $\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}$ (by default, $q = .9$).

### 3.3 Sampling procedure

As usual, a sequence of Metropolis-Hastings within Gibbs sampler is adopted:

**Step 1** $\mathcal{T}_t, \mathcal{M}_t \mid \mathcal{T}_{-t}, \mathcal{M}_{-t}, \sigma^2, \boldsymbol{y}$, for $t = 1, \ldots, T$, which is done compositionally as

(a) $\mathcal{T}_t \mid \mathcal{T}_{-t}, \mathcal{M}_{-t}, \sigma^2, \boldsymbol{y}$,

(b) $\mathcal{M}_t \mid \mathcal{T}, \mathcal{M}_{-t}, \sigma^2, \boldsymbol{y}$,

**Step 2** $\sigma^2 \mid \mathcal{T}, \mathcal{M}, \boldsymbol{y}$.

---

[5] Jeffreys prior is even approved somehow from Objective Bayesian perspective given that it approximately takes the functional form of reference prior, which maximizes the expected information from data over the permissible prior space. (see Clarke, 1994; Berger et al., 2009; Chen, 2019)

[6] As pointed in (Chipman et al., 1998, p. 939), this prior specification is equivalent to the usual Inverse Gamma specification such that $\sigma^2 \sim \text{IG}(\nu/2, \nu\lambda/2)$.

One thing to remark here is that the Dirichlet prior specification as mentioned previously gives a conjugate Gibbs-sampling update for $\omega$

$$\omega \sim \mathcal{D}\left(\frac{\alpha}{P} + m_1, \ldots, \frac{\alpha}{P} + m_P\right) \tag{11}$$

where $m_j$ is number of times when $j$-th covariate is used for splitting trees. Finally it is possible to obtain an average of posterior probability for each covariate and this can be used as a measure of variable importance for the purpose of variable selection.

# 4 Monte Carlo Study

Let's consider a relatively highly nonlinear model where $f(\boldsymbol{x})$ is specified as

$$f(\boldsymbol{x}) = \sin\left(x_1^2 + x_2^2\right) + \sin\left(x_3^2 + x_4^2\right) + \frac{\left(x_1 + x_2^2\right) \cdot \left(x_{14} + x_{15}\right)^2}{3 + x_3 + x_{14}^2} \tag{12}$$

Data is generated from $f(\boldsymbol{x})$ such that for each observation $y_i$ and covariates $\boldsymbol{x}_i$,

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

where $\boldsymbol{x}_i$ is $1 \times P$ row vector generated uniformly on $[-2, 2]$ and $\sigma$ is taken as the sample standard error of $f(\boldsymbol{x}_i)$. [7] Applying the simulated observations $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ to BART, variable selection is based on the finally updated probability $\omega$ assigned to each variable. The following figure demonstrate the finally updated probability for each variable under different $n$, $p$ specifications.

**[Place Figure 4 about here]**

It is obvious from Figure 4 that potentially influential variables $x_1$, $x_2$, $x_3$, $x_4$, $x_{14}$ and $x_{15}$ as designed is possible to be identified based on the finally updated posterior probability from BART. However, as the number of variables is relatively large in comparison to the sample size (say, where $P = 100$ and $n = 500$ or even $P = 50$ and $n = 500$), some of the not influential variables may be confounded with the indeed influential ones. But with the increase of sample size, such confounding variables are possible to be excluded based on the updated posterior probability.

To demonstrate the out of sample prediction accuracy from BART in comparison to conventional methods like OLS, As inspired by the suggestions in Campbell and Thompson (2007) and Rapach et al. (2010), the out-of-sample evaluation (forecast evaluation) is via out-of-sample $R$ square defined as

$$R_{OS}^2 = 1 - \frac{\sum_{i \in \mathcal{O}} (y_i - \widehat{y}_i)^2}{\sum_{i \in \mathcal{O}} y_i^2}$$

[7] Data simulated here is not in panel setting since we want to emphasize the return predictability from cross-sectional information. This is consistent with our empirical implementation where we train the the mode within a given window for each cross-sectional data by fixing time index and take the final prediction as the time-series average of predictions made from cross-sectional prediction.

where $\mathcal{O}$ collects all the indexes for observations and corresponding prediction out of training sample. The results are listed as following

**[Place Table 2 about here]**

For relatively more fair comparison, I also constructed a forward neural network with three layers and applied it to the the simulation mechanism. I just found that even in comparison to this machine learning method which is popular in industry, the out-of-sample performance is still better especially for the scenario where the sample size is sufficiently large and the number of covariates is relatively large as well. The implementation is through TensorFlow and the result is left in appendix Table A.1. One additional finding is that the size of nodes in each layer should not be that large and the total training epochs for forward neural network should not be that large as well (about 64 nodes in each layer and training for 10 epochs should be enough in our simulation).

# 5 Empirical Application Results

## 5.1 Data construction

As in Green et al. (2017), total 102 firm-level characteristics are constructed, requiring that each characteristic be entirely calculable from CRSP, Compustat and/or I/B/E/S data. Data covers 39-year period from January 1980 to December 2018. As the argument in Green et al. (2017), the reason that the whole data sample starts from 1980 is because most firm-level characteristics only become robustly available in that year. The total 102 characteristics used are listed in the following table

**[Place Table 3 about here]**

Specifically, we do a normalization from the original firm-level characteristic $C_{s,it-1}$ to $\tilde{C}_{s,it-1}$ such that $\tilde{C}_{s,it-1}$ is supported over $[0,1]$ by taking

$$\tilde{C}_{1,it-1} = F_{s,t}\left(C_{s,it-1}\right) = \frac{\text{rank}\left(C_{s,it-1}\right)}{N_t + 1}$$

with

$$\text{rank}\left(\min_{i=1\ldots,N_t} C_{s,it-1}\right) = 1 \quad \text{rank}\left(\max_{i=1\ldots N_t} C_{s,it-1}\right) = N_t$$

where $N_t$ denotes the number of available firms such that to keep the data as balanced sheet in period $t$.

### 5.2 Out-of-sample prediction comparison

As discussed in the previous discussion, $R^2_{OS}$ here is defined as following

$$R^2_{OS} = 1 - \frac{\sum_{(i,t)\in\mathcal{C}}(r_{i,t+1} - \widehat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{C}} r^2_{i,t+1}}$$

where $r_{i,t+1}$ denotes return for $i$-th stock in period $t+1$. $(i,t) \in \mathcal{C}$ indicates that forecasting evaluation is made on testing sub-sample. If we select 1980 to 2005 as the training sample and make $(2005 - 1980 + 1) \times 12 = 312$ cross-sectional predictions for each month from January 2006 to December 2018 and take the time series sample average as the prediction, the $R^2_{OS}$ from BART is $-0.0051$. In rolling window scheme, currently it is implemented from 2013 to 2016 and for prediction made at each month within this time period, time-series average of previous 120 (10 years) months predictions (based on cross-sectional data) are used for prediction. And currently the $R^2_{OS}$ is 0.0106. [8] This is pretty good in comparison to some existing methods. Specifically, for OLS without intercept, the $R^2_{OS}$ is 0.0019. Actually, by simply calculating that

$$0.0106/0.0019 \approx 5.58$$

which implies that it is almost 6 times improvement in $R^2_{OS}$. Predicted returns versus Realized returns are possible to be compared as following, despite that the improvement is not that obvious directly from this scatter plot.

**[Place Figure 5 about here]**

Actually based on the implementation experience, some proposed nonparametric methods with LASSO plugged in is somehow computationally demanding since two large-scale optimization problems have to be solved. This may justify the application of BART from the perspective of computational efficiency as well.

**Remark 5.1** I also compared the results from rolling window. Basically, for each month from 2013 to 2018, I trained BART in the previous 10 years, which means that the time-series average of 120 predictions are used as $\widehat{r}_{i,t+1}$ for $(i,t) \in \mathcal{C}$. Two detailed comparisons about how to select training sample are made as well:

(i) exactly the data from previous 10 years are used for training BART. For example, if $t$, where the time point the prediction is made, refers to May in 2013, the training sample in this case is from January 2003 to December 2012.

(ii) exactly the data from previous 120 months are used for training BART. For example, if $t$, where the time point the prediction is made, refers to May in 2013, the training sample in this case is from May 2003 to April 2013.

[8] The reason that the results are based on testing sample from 2013 to 2016 is simply for the purpose of saving time. Basically, it is supposed to make the testing sample have longer time coverage, say, from 2010 to 2018. It could be done, just take more time on calculation, one more week needed.

For (i), the corresponding $R^2_{OS}$ is 0.0115; and the for (ii), it is exactly what discussed in the main context, thus, $R^2_{OS} = 0.0106$.

**Remark 5.2** For robustness check, I have calculated and compared the out-of sample $R$ squares, both for BART and conventional OLS, with different specifications for testing period and testing sample length iin between 2013 and 2018 as well. The basic results are listed in appendix Table A.3. And one major conclusion made from this comparison is that the U.S. market is less predictable recently after 2015.

Moreover, we can construct equally weighted portfolios based on the predicted return of BART and hold each portfolio to see the future cumulative return as following

**[Place Figure 6 about here]**

Similarly, we can construct equally weighted portfolios based on the predicted return of NN3 and hold each portfolio to see the future cumulative return as following

**[Place Figure 7 about here]**

For comparison purpose, I also construct the equally weighted portfolios based on the predicted return of OLS and hold each portfolio to see the future cumulative return as following

**[Place Figure 8 about here]**

Where specifically for all the cases demonstrated above, I construct 5 equally weighted portfolios based on the out-of-sample predictions generated from different methods. That is, the ordering of cumulative return of constructed portfolios is expected as following

$$\text{Decile } 5 > \text{Decile } 4 > \text{Decile } 3 > \text{Decile } 2 > \text{Decile } 1.$$

One thing to note by comparing the above results is that in the relatively short period (specifically here the testing sample is specified as from 2013 to 2016, 48 months), the future cumulative return performance of portfolios constructed from machine learning method (BART and NN3) is relatively more distinguishable in comparison to the future cumulative return performance of portfolios constructed from linear model (OLS): linear model in this sense performs the worst in this specified short testing sample, given the observation that the associated portfolio which would be expected to generate the best cumulative return (Decile 5) some times does not yield the best cumulative return as expected; While the Decile 5 portfolio associated with BART and NN3 always leads to the best cumulative return in this short testing sample. But there is still some "confusion" for machine learning method where the portfolios in between the "best" (Decile 5) and the "worst" (Decile 1) may interchange their expected ordering of cumulative return: for BART, Decile 3 portfolio yields the second to the best cumulative return (Decile 3 > Decile 4 > Decile 2);

while for NN3, Decile 2 portfolio finally yields better cumulative return than Decile 3 portfolio (Decile 4 > Decile 2 > Decile 3).

The longest comparison is from 2008 to 2018 and the corresponding out-of-sample $R$ square $R^2_{OS}$ are as following respectively

| | | |
|---|---|---|
| BART | : | **0.028** (%) |
| NN3 (Python TensorFlow) | : | -0.18 (%) |
| NN3 (MATLAB) | : | -0.24 (%) |
| NN3 (R Keras TensorFlow API) | : | **0.03** (%) |
| OLS | : | -0.34 (%) |

Given the results demonstrated above, it seems that Neural Network based method is still the most advanced machine learning methods in terms of prediction accuracy but it has to be tuned a lot and the result may vary a lot over different platform. Moreover, in terms of out-of-sample $R$ square, the gain from applying NN3 in comparison to BART is not that much (0.028% of BART versus 0.03% of NN3), hence the out-of-sample $R$ square is not that bad in this sense. And one more point to be noted is that the prediction results generated from BART are relatively more stable and is not that much sensitive to the platforms on which it has been implemented. Finally, the discussed ordering "confusion" vanishes as the testing sample is extended as from 2008 to 2018 (132 months), see Figure A.2 to Figure A.4 for detailed demonstration.

### 5.3 Characteristics selected

We firstly plot heat map constructed from finally updated posterior probability in each period as following

**[Place Figure 9 about here]**

To select the overall influential characteristics, we simply implement the following steps

**Step 1** For each time period, define a benchmark probability assigned to each used characteristic in that period as

$$\Pr_t^{\text{benchmark}} = \frac{1}{\#\text{ cross-sectional characteristics used in } t}.$$

**Step 2** Denote the posterior updated probability assigned to each characteristic as $\Pr_t$ (i.e., the finally updated posterior probability from BART, $\omega$), comparing $\Pr_t$ with $\Pr_t^{\text{benchmark}}$ and picking the corresponding characteristics such that $\Pr_t > \Pr_t^{\text{benchmark}}$.

**Step 3** For each characteristic, counting the number of times it has been selected in the time span based on the criteria in **Step 2**. Denote the length of whole time period span as $T$ and finally select the characteristics with the number of being selected greater then $2T/3$. [9]

Given the above described three steps, the selected influential characteristics for prediction and estimating expected return is "mve", "mom12m", "mom1m", "retvol", "baspread" and "idiovol", which is based on the data from 1980 to 2005. It could be explained as the three sources of characteristics about firms: Size ("mve"), Momentum ("mom12m" and "mom1m") and Volatility ("retvol", "baspred" and "idiovol"). Similar result holds for the time period from 2006 to 2018 where we find that Size ("mve"), Momentum ("mom12m") and Volatility ("baspread") are the most influential characteristics for expected return.

### 5.4 More Comprehensive Results from Other Markets

Previous discussed empirical studies follow the convention in literature by focusing on the U.S. financial market. In this section, I will make one step forward progress by extending the application of machine learning method to other markets: the China stock market and the main traded financial assets around the the world. The main concern of the following to be discussed is how the specific machine learning method, BART, which has been discussed and proposed to use in this paper, performs in Chins stock market and the markets. Comparison is made between the bottom benchmark (i.e. linear model like OLS) and some advanced machine learning methods (i.e. Black-box model like Neural Network as well.

#### 5.4.1 China stock market

I collect data mainly from `CSMAR` (China Stock Market & Accounting Research Database) and construct data as following (for complete construction details, see my `SAS` codes: `constructing_csmar_data.sas`)

- Market value at monthly frequency is collected directly from `csmar.csmar_t_mnth` with the corresponding acronym as `msmvttl`. And it is renamed as `mve`.
- Equity data is collected from `csmar.combas` with the corresponding acronym as `A003000000`. Different companies may report data corresponding to equity at different timing point and the values may vary for data reported at different period. To follow the convention in finance research, I keep the data reported in June as the annual fiscal equity data. In the `SAS` code, this variable is renamed as `tseq`. `tseq` is used as the proxy for book value and hence I make an attempt to construct book-to-market ratio (with acronym `bm`) as `bm` = `tseq`/`mve`.
- Return data is directly collected from `csmar_t_mnth`. I use `Mretwd` (Monthly Return with Cash Dividend Reinvested) as the proxy for monthly return. And with the available monthly return data, I can further construct firm-level characteristics corresponding to momentum. `Mretwd` is renamed as `ret`.

[9] The rule for splitting data is the similar to what implemented in (Kozak et al., 2018; Kozak, 2019).

- Moreover, I also make an attempt to construct two additional characteristics associated with daily return through `CSMAR` daily database. Specifically, `maxret` refers to the maximum of daily return within a specific month; `retvol` refers to the volatility (standard deviation) of daily return. Following code chunk extracted is for demonstration purpose:

With the description above, I summarize the characteristics used for discussion for China stock market as following

`mom1m` : Lagged monthly return

`mom12m` : 12-month momentum

`mom6m` : 6-month momentum

`chmom` : Change of 6-month momentum

`maxret` : Maximum daily return

`retvol` : Return volatility (standard deviation) of daily return

`log_mve` : Log market value

`bm` : Book to market

I follow the same forward rolling-window scheme training and testing as in the previous discussion. To document which covariate matters for return prediction, I implement the following procedure

**Step 1** For each time period, define a benchmark probability assigned to each used characteristic in that period as

$$\Pr_t^{\text{benchmark}} = \frac{1}{\#\text{ cross-sectional characteristics used in } t}.$$

**Step 2** Denote the posterior updated probability assigned to each characteristic as $\Pr_t$ (i.e., the finally updated posterior probability from BART, $\omega$), comparing $\Pr_t$ with $\Pr_t^{\text{benchmark}}$ and picking the corresponding characteristics such that $\Pr_t > \Pr_t^{\text{benchmark}}$.

**Step 3** Documenting the results for each time point within training period and counting how many times a specific variable has been selected within training period (120 months prior to each time point in the testing sample) based on **Step 1** and **Step 2**. And later by rolling the testing time point forward we can document the time-varying properties corresponding to which covariates matter and plots the associated heat map as following

**[Place Figure 10 about here]**

As demonstrated in the figure above, `log_mve`, `retvol`, `bm` and `mom1m` in general relatively matter more for return prediction in China stock market.

For out-of-sample prediction accuracy comparison, it is still firstly based on out-of-sample $R$ square. The documented out-of-sample $R$ square with testing sample from 2008 to 2018 is as following

$$\text{BART} : 0.1937\%$$

$$\text{NN3} : 0.4661\%$$

$$\text{OLS} : -0.0519\%$$

Despite it seems that BART cannot in general beats NN3 in China stock market in terms of prediction accuracy, the gains from constructed portfolios based on one-month ahead prediction is not much different, which is demonstrated in the following plot of cumulative log return.

**[Place Figure 11 about here]**

**[Place Figure 12 about here]**

Where Figure 11 refers to the corresponding results from BART; while Figure 12 refers to the corresponding results from NN3. By comparing the last row of the associated tables, the entries of which refer to the out-of-sample Sharpe ratio of different portfolios based on the one-month ahead predictions, there is no much difference between the Sharpe ratio of portfolios constructed from BART and the portfolios constructed from NN3. However, the predictions from OLS, which as a kind linear model is the bottom benchmark for comparison, are not that good by comparing the out-of-sample $R$ square with the out-of-sample $R$ square of other two machine learning methods. This is further supported through the plot of cumulative log return and Sharpe ratios of portfolios constructed from OLS predictions.

**[Place Figure 13 about here]**

Where from Figure 13, the constructed portfolios are not distinguished well with each other and the best portfolio suggested by linear model cannot lead to best cumulative return in China stock market.

### 5.4.2 Around the world

In comparison to the construction of the U.S. data, the main difficulty in constructing the Global data is from the rare data availability corresponding to prices and return at monthly frequency. Construction of many firm-level characteristics is based on price data return data. (for example, momentum and market value). To solve this problem, I simply construct the necessarily needed data at lower frequency (monthly or annually) by aggregating the daily information, which is available

at `compd.g_secd` on WRDS with the variable name for daily price as `prccd`. The major variables collected at annually (or quarterly frequency with the similar procedure) is

```
prc_yr  :  mean(prccd) as prc_yr
```

I leave the detailed description in `SAS` code chunk in appendix. The basic cross-sectional data structure is as following:

| | | |
|---|---|---|
| `date` | : | data date |
| `year` | : | year of data date |
| `month` | : | month of data date |
| `fyear` | : | fiscal year in which annually firm-level characteristics are issued |
| `gvkey` | : | global company key |
| `loc` | : | country code-headquarters |
| `sic2` | : | the first two digits of standard industry classification code |
| `ret` | : | monthly return |
| $\cdots$ | : | all the remained columns correspond to firm-level characteristics matched to monthly frequency. |

The documented out-of-sample $R$ square is as following

$$\text{BART} : 38.3612\%$$

$$\text{NN3} : 31.0721\%$$

$$\text{OLS} : 12.7569\%$$

which is much higher in comparison to the results from the U.S. stock market and China stock market. Moreover, the performance of linear model is not bad as well given the relatively high documented out-of-sample $R$ square demonstrated above. The portfolios performance is listed and demonstrated as following

**[Place Figure 14 about here]**

**[Place Figure 15 about here]**

## 6 Conclusion

This paper demonstrates how to apply nonparametric methods to study return predictability in empirical finance with specific focus on the application of the nascent Bayesian nonparametric

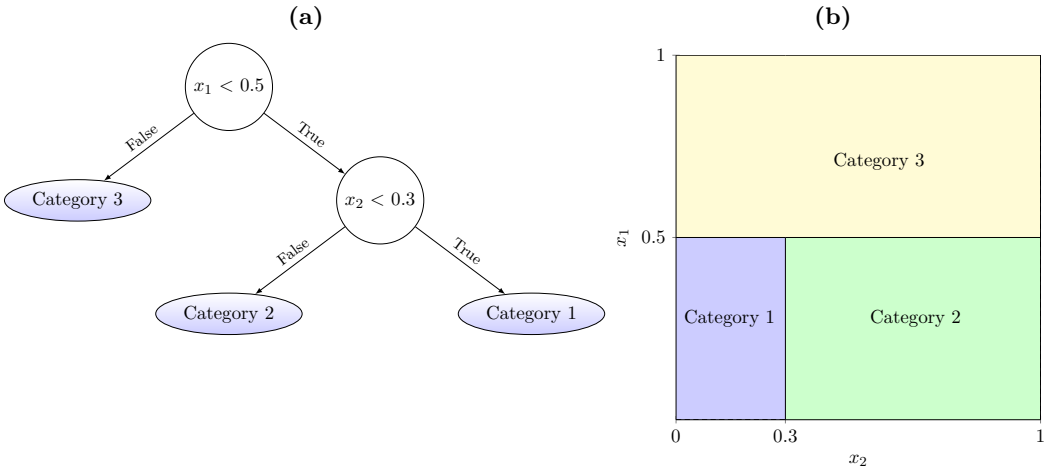
machine learning methods (BART) for variable selection and estimating expected return function in financial market. It has been discussed and approved through simulation in this paper that BART is efficient for selecting influential variables in comparison to some conventional methods, especially for the scenario where the model specification is highly nonlinear. BART is applied in this paper to select potentially influential firm-level characteristics among the major documented firm-level characteristics in literature. Based on BART, we have documented that "Size" plus momentum and volatility related characteristics contributes the most to the expected return. Prediction is further made based on the selected firm-level characteristics and the functional form of expected return function as the output from BART. The out-of-sample performance (measured in $R^2_{OS}$) is satisfactory as well in comparison to some conventional methods. (Both from simulation or empirical implementation).

Comparison is made between bottom linear benchmark model and advanced "Black-box" machine learning methods (specifically, Neural Network) but with documented good performance in prediction accuracy. Based on the results documented in this paper, prediction performance of the nascent nonparametric machine learning method BART may vary across different universe (markets): For the U.S. market, BART slight beats linear model but still cannot beats NN3 and cannot generate significantly high out-of-sample $R$ square. But for China stock market, the gain of applying machine learning method appears and NN3 beats BART in terms of out-of-sample prediction accuracy despite the performance of portfolio based on the corresponding predictions are not significantly different with each other.

Despite the nascent machine learning method BART in some scenario cannot always dominate other existing methods, its essential structure makes it a relatively more interpretable machine learning method since all the variable importance determined by BART is based on prior-posterior updating and hence is relatively more interpretable.

# Figures and Tables

**Figure 1:** Binary Regression Tree Illustration




**Note:** This figure presents the diagrams of a single binary regression of 2 depth with two covariates . (a) demonstrates how it is split and (b) demonstrates its equivalent representation in the space of two covariates ($x_1$ and $x_2$). Based on the 3 categories dissected, a single tree just take some constant value on these areas respectively.

**Figure 2:** Tree at Different Depth with Different Terminal Nodes

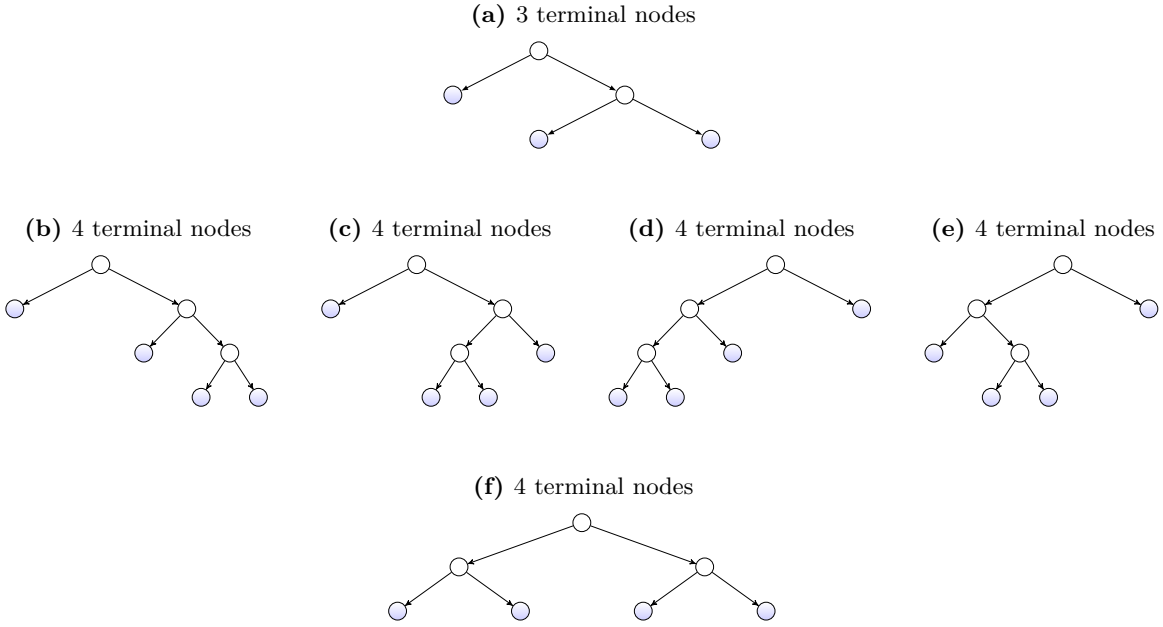
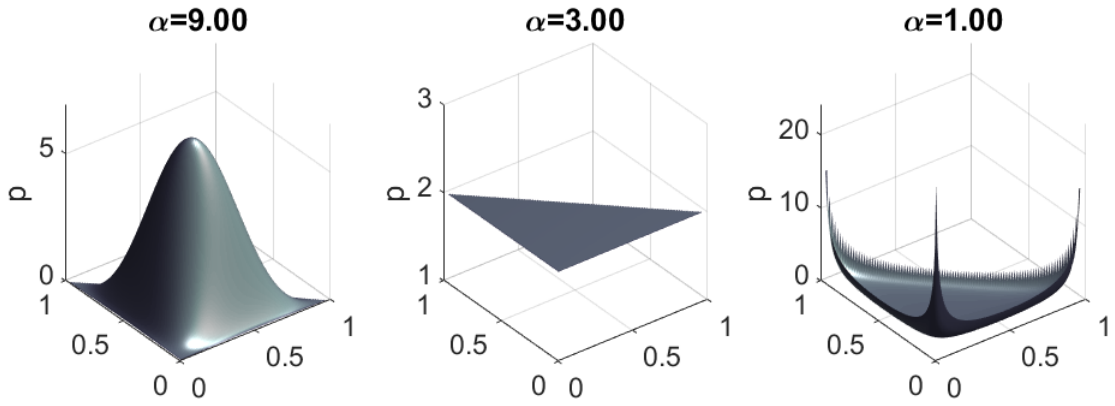



**Table 1:** Probability for Tree Structures with Different Terminal Nodes

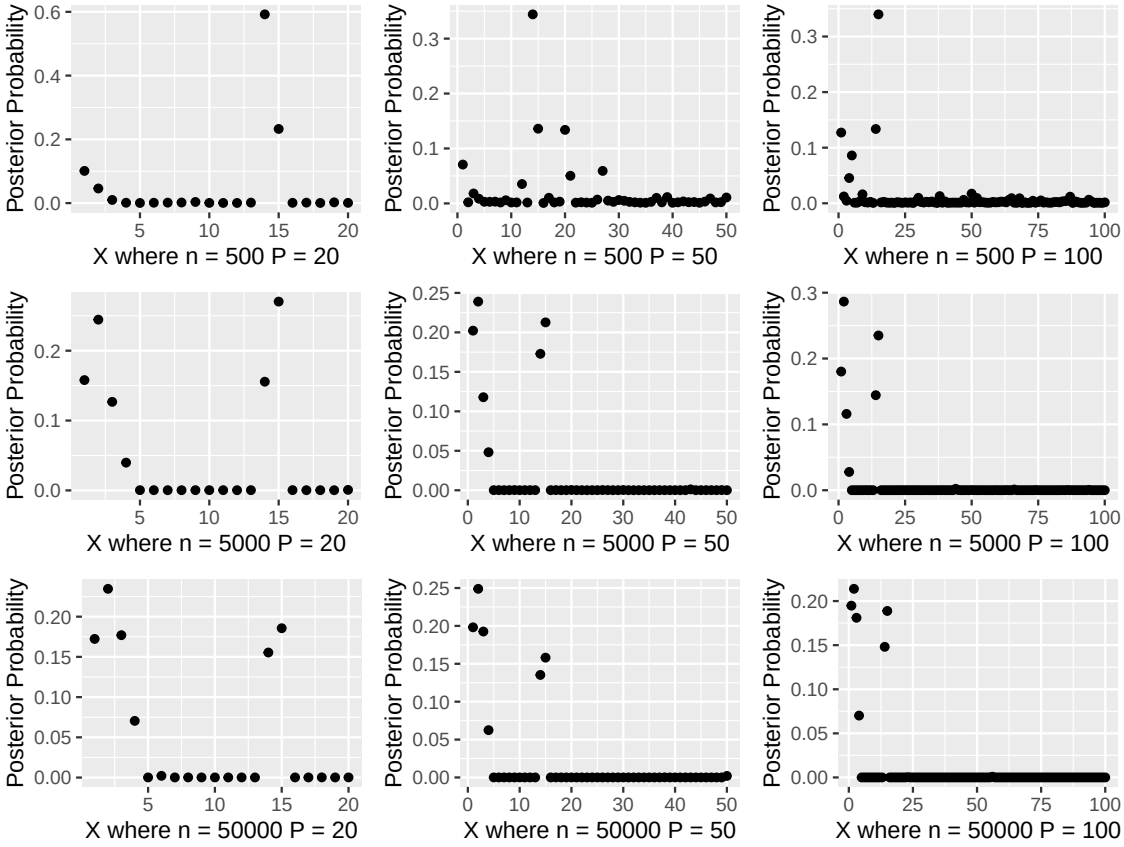
| Depth $d$ | Number of terminal nodes | Probability |
|---|---|---|
| 0 | 1 | .05 |
| 1 | 2 | .55 |
| 2 | 3 | .28 |
| | 4 | .09 |
| $\geqslant 3$ | $\geqslant 5$ | .03 |

**Figure 3:** Illustration of Dirichlet Distribution




**Note:** The above graph plots the probability density function for Dirichlet distribution $\mathcal{D}(\alpha/3, \alpha/3, \alpha/3)$ with different $\alpha$ specifications. High $\alpha$ implies that more density concentrate on the interior points over the probability simplex while relatively low $\alpha$ gets the probability density function flat. If $\alpha$ is evaluated such that $\alpha/3 < 1$, sparsity would be induced (more density would concentrate on the vertices and edges of the probability simplex, vertices of the simplex correspond to that a single variable is selected with prior probability equal to 1, edges correspond to that two variables are selected with additional variable excluded with prior probability equal to 1.)

**Figure 4:** Finally Updated Posterior Probability of Variables




**Note:** The above figure demonstrates the scatter plot of finally updated posterior probability against all the ordered number of variable index as described in the simulation example. Vertical scale corresponds to the the finally updated posterior probability and the horizontal scale corresponds to the ordered variable index. In this $3 \times 3$ grid panel, each row corresponds to the same $n$ (sample size) specification but with different $P$ (number of variables) and each column corresponds to the same $P$ (number of variables) specification but with different $n$. $n$ takes value of 500, 5000 and 50000 respectively and $P$ takes value of 20, 50 and 100 respectively.

**Table 2:** Out of Sample Prediction Comparison

| BART | | | | OLS | | | | Nonparametric-LASSO | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $p$ | $nt$ | $R^2_{OS}$ | $n$ | $p$ | $nt$ | $R^2_{OS}$ | $n$ | $p$ | $nt$ | $R^2_{OS}$ |
| 500 | 20 | 100 | -0.0178 | 500 | 20 | 100 | -0.0285 | 500 | 20 | 100 | 0.0431 |
| 500 | 50 | 100 | 0.3354 | 500 | 50 | 100 | 0.2722 | 500 | 50 | 100 | -0.2750 |
| 500 | 100 | 100 | 0.4598 | 500 | 100 | 100 | 0.1928 | 500 | 100 | 100 | -7.6196 |
| 5000 | 20 | 1000 | 0.4834 | 5000 | 20 | 1000 | 0.2428 | 5000 | 20 | 1000 | 0.3030 |
| 5000 | 50 | 1000 | 0.4607 | 5000 | 50 | 1000 | 0.2044 | 5000 | 50 | 1000 | 0.2810 |
| 5000 | 100 | 1000 | 0.4912 | 5000 | 100 | 1000 | 0.1989 | 5000 | 100 | 1000 | 0.2438 |
| 50000 | 20 | 10000 | 0.5469 | 50000 | 20 | 10000 | 0.2241 | 50000 | 20 | 10000 | 0.3239 |
| 50000 | 50 | 10000 | 0.5463 | 50000 | 50 | 10000 | 0.2247 | 50000 | 50 | 10000 | 0.2968 |
| 50000 | 100 | 10000 | 0.5542 | 50000 | 100 | 10000 | 0.2177 | 50000 | 100 | 10000 | 0.3032 |

**Note:** This table demonstrates the out-of-sample prediction comparison between BART, conventional OLS and the proposed Nonparametric-LASSO method in (Freyberger et al., 2019). For each column, $n$ denotes the training sample size (number of observations); $p$ denotes the number of covariates; $nt$ denotes the testing sample size used for calculating $R^2_{OS}$; $R^2_{OS}$ denotes the out-of-sample prediction accuracy measure discussed in the context.

**Figure 5:** Realized Returns vs Predicted Returns

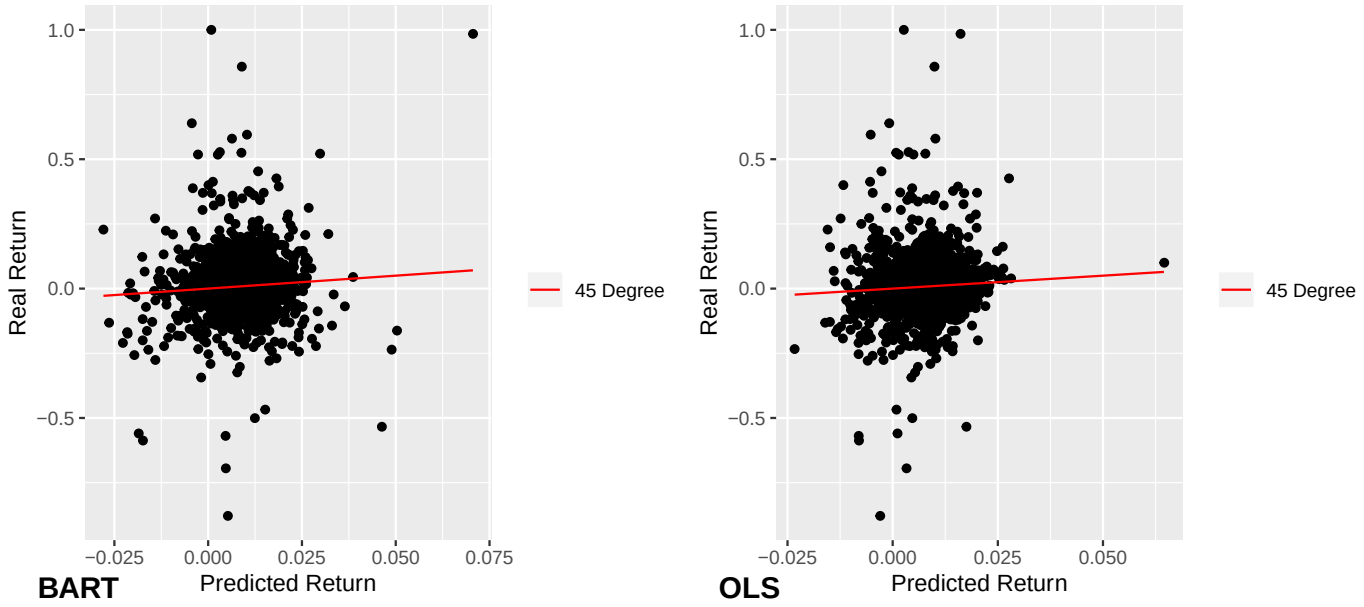
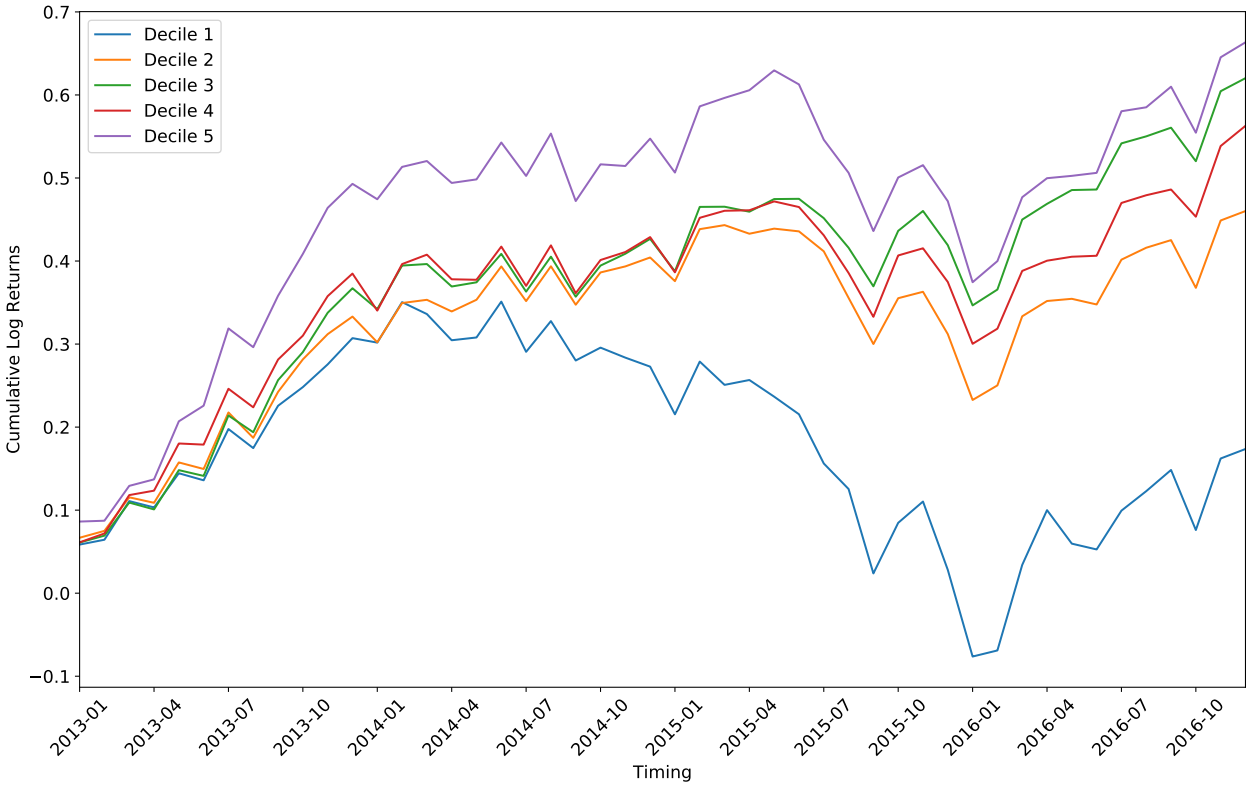
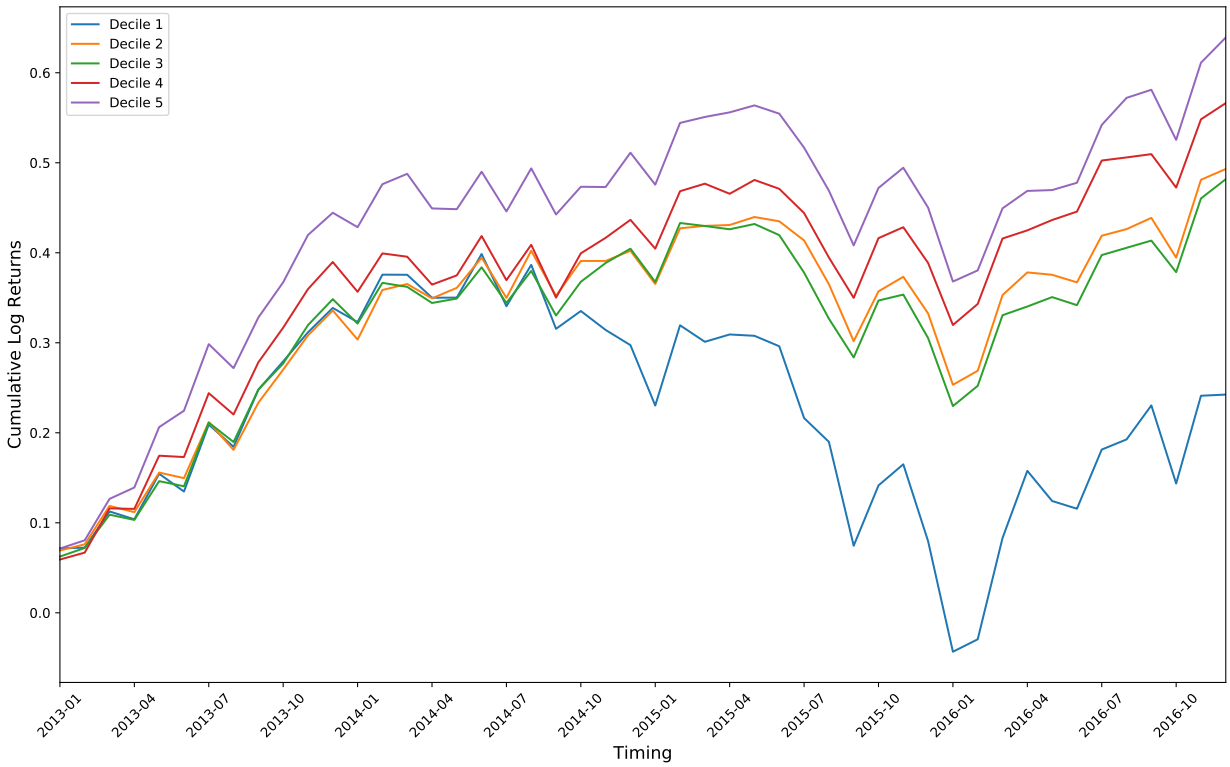


**Figure 6:** Cumulative Returns (in log transformation) of Equally-weighted Portfolios based on BART Prediction in the U.S. Stock Market from 2013 to 2016



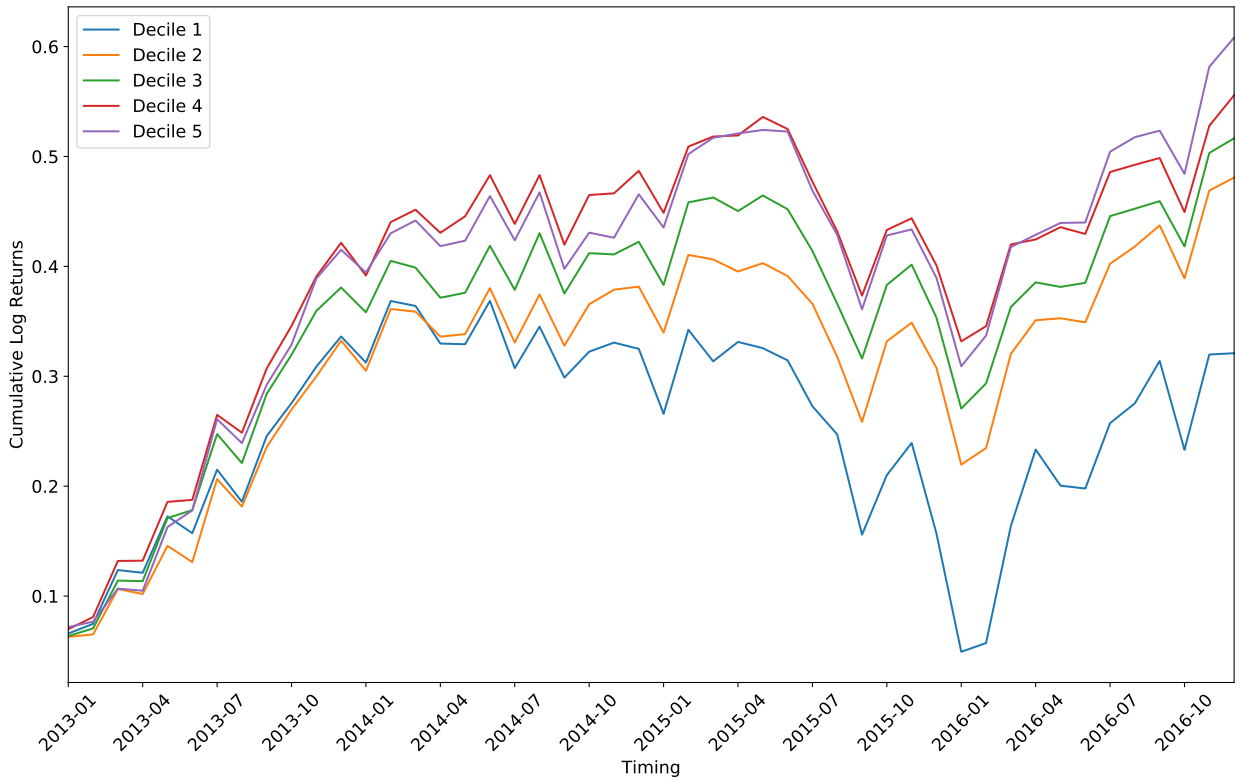

| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 48.000000 | 48.000000 | 48.000000 | 48.000000 | 48.000000 |
| mean | 0.004748 | 0.010398 | 0.013756 | 0.012594 | 0.015009 |
| std | 0.047845 | 0.039650 | 0.039412 | 0.040591 | 0.047356 |
| min | -0.099099 | -0.076256 | -0.069921 | -0.071750 | -0.092969 |
| 25% | -0.023956 | -0.011216 | -0.011156 | -0.010600 | -0.017195 |
| 50% | 0.005861 | 0.008741 | 0.014855 | 0.010184 | 0.016574 |
| 75% | 0.042555 | 0.041043 | 0.041235 | 0.047664 | 0.047009 |
| max | 0.108491 | 0.086766 | 0.087976 | 0.088855 | 0.097486 |
| Sharpe Ratio | 0.099243 | 0.262243 | 0.349023 | 0.310266 | 0.316938 |

**Figure 7:** Cumulative Returns (in log transformation) of Equally-weighted Portfolios based on NN3 Prediction in the U.S. Stock Market from 2013 to 2016



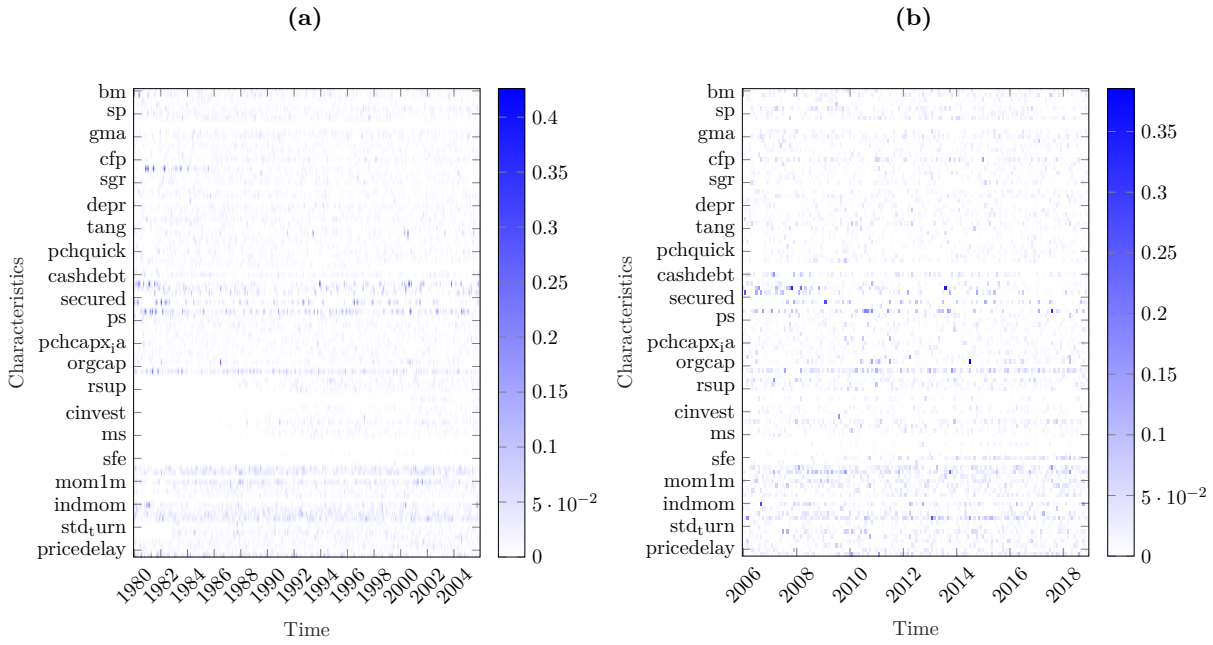

| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 48.000000 | 48.000000 | 48.000000 | 48.000000 | 48.000000 |
| mean | 0.006548 | 0.011094 | 0.010816 | 0.012613 | 0.014264 |
| std | 0.054957 | 0.039774 | 0.038753 | 0.039199 | 0.042025 |
| min | -0.115563 | -0.076199 | -0.072759 | -0.066618 | -0.078856 |
| 25% | -0.021739 | -0.010245 | -0.013661 | -0.014233 | -0.010850 |
| 50% | 0.001057 | 0.010142 | 0.008842 | 0.010946 | 0.012604 |
| 75% | 0.047694 | 0.040562 | 0.037812 | 0.043982 | 0.047539 |
| max | 0.119084 | 0.090323 | 0.085381 | 0.078791 | 0.089519 |
| Sharpe Ratio | 0.119155 | 0.278919 | 0.279093 | 0.321773 | 0.339420 |

**Figure 8:** Cumulative Returns (in log transformation) of Equally-weighted Portfolios based on OLS Prediction in the U.S. Stock Market from 2013 to 2016




| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 48.000000 | 48.000000 | 48.000000 | 48.000000 | 48.000000 |
| mean | 0.007886 | 0.010899 | 0.011632 | 0.012445 | 0.013637 |
| std | 0.049008 | 0.041406 | 0.040936 | 0.040568 | 0.042731 |
| min | -0.102083 | -0.084282 | -0.079444 | -0.066805 | -0.077382 |
| 25% | -0.026027 | -0.016627 | -0.014942 | -0.012256 | -0.008598 |
| 50% | 0.008146 | 0.009767 | 0.007097 | 0.011196 | 0.011484 |
| 75% | 0.042695 | 0.042350 | 0.043853 | 0.045999 | 0.040627 |
| max | 0.112421 | 0.089512 | 0.088791 | 0.081612 | 0.102262 |
| Sharpe Ratio | 0.160920 | 0.263232 | 0.284157 | 0.306773 | 0.319141 |

**Figure 9:** Characteristic Probability




**Note:** This figure plots the posterior updated probability assigned to each characteristic (vertical scale) in each time period (horizontal scale). (a) refers to the time span from year 1980 to year 2005 and (b) refers to the time span from year 2006 to year 2018. Different lightness of Blue indicates different values of corresponding probabilities. The color gradient ranges from light white to blue and the darker (in blue color) the corresponding area is, the larger the associated probability is.

**Figure 10:** Characteristics Heat Map for China Stock Market

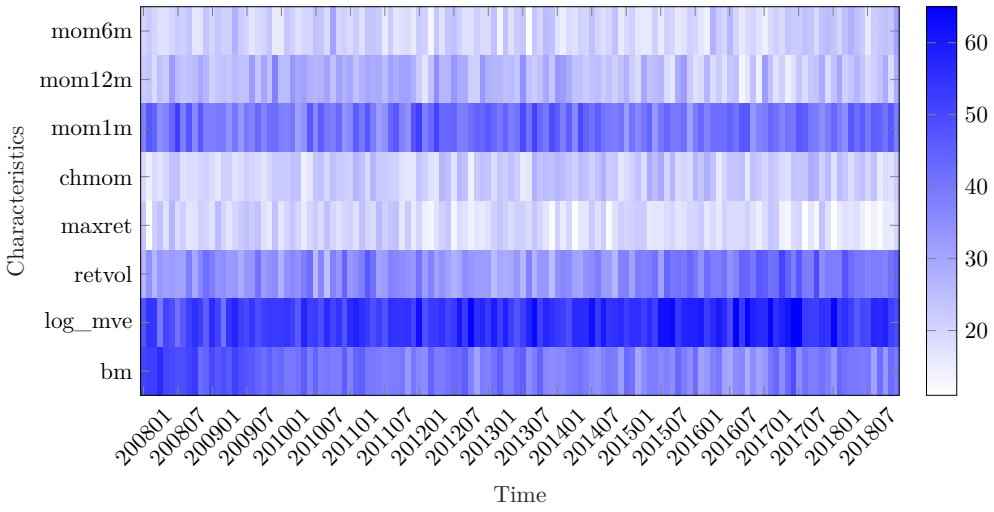
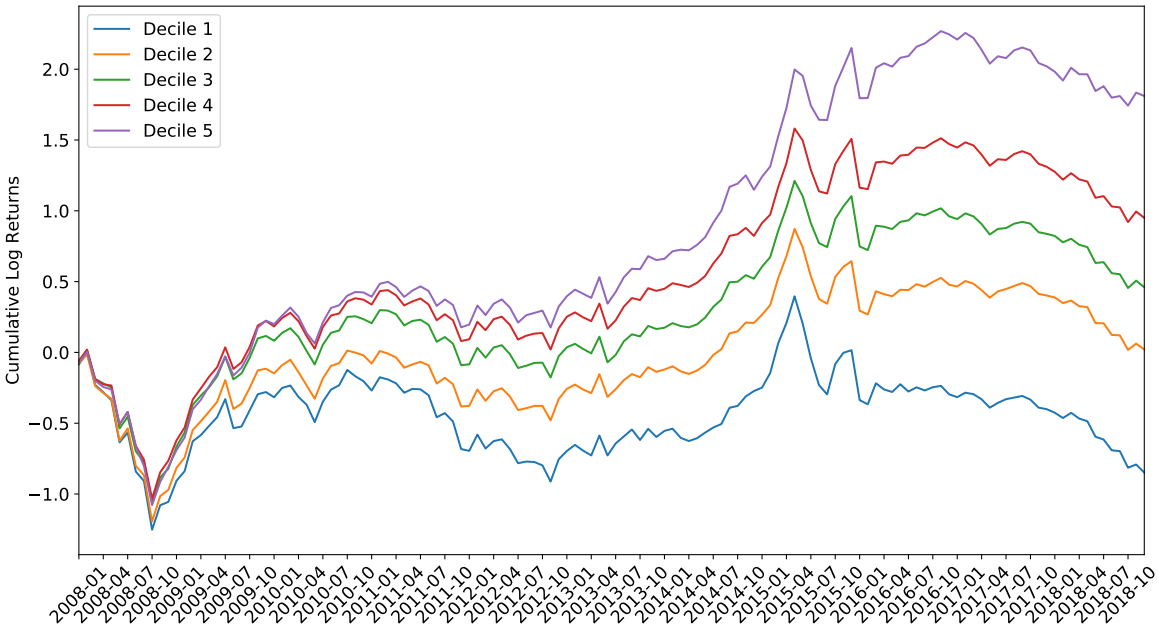
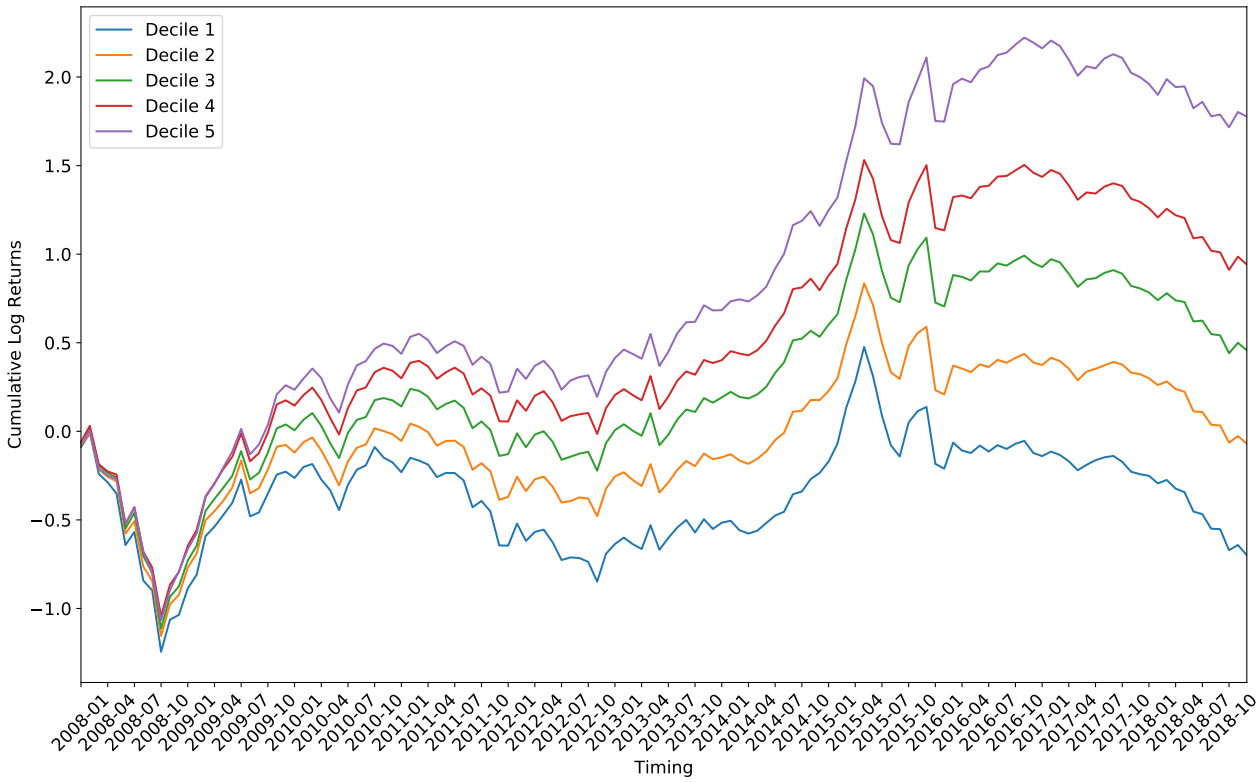


**Figure 11:** Cumulative Returns (in log transformation) of Equally-weighted portfolios based on BART prediction in China Stock Market from 2008 to 2018



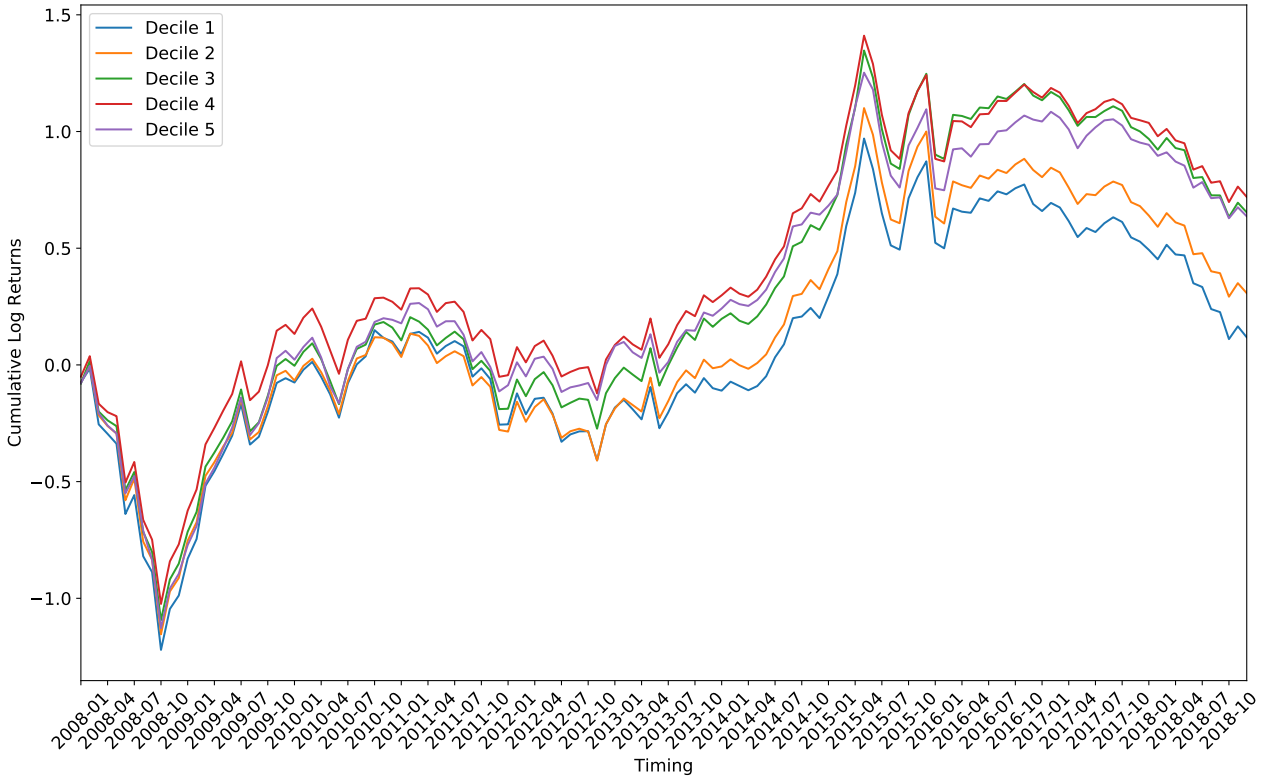

| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 132.000000 | 132.000000 | 132.000000 | 132.000000 | 132.000000 |
| mean | -0.001181 | 0.004934 | 0.008131 | 0.012135 | 0.019215 |
| std | 0.100748 | 0.096342 | 0.095362 | 0.098673 | 0.104074 |
| min | -0.296938 | -0.295408 | -0.299369 | -0.292002 | -0.298720 |
| 25% | -0.055228 | -0.040007 | -0.037058 | -0.041597 | -0.038118 |
| 50% | 0.003481 | 0.012183 | 0.007578 | 0.013284 | 0.021607 |
| 75% | 0.051726 | 0.061003 | 0.063843 | 0.078420 | 0.083340 |
| max | 0.238628 | 0.213788 | 0.219561 | 0.278204 | 0.312767 |
| Sharpe Ratio | -0.011721 | 0.051215 | 0.085265 | 0.122983 | 0.184624 |

**Figure 12:** Cumulative Returns (in log transformation) of Equally-weighted portfolios based on NN3 prediction in China Stock Market from 2008 to 2018



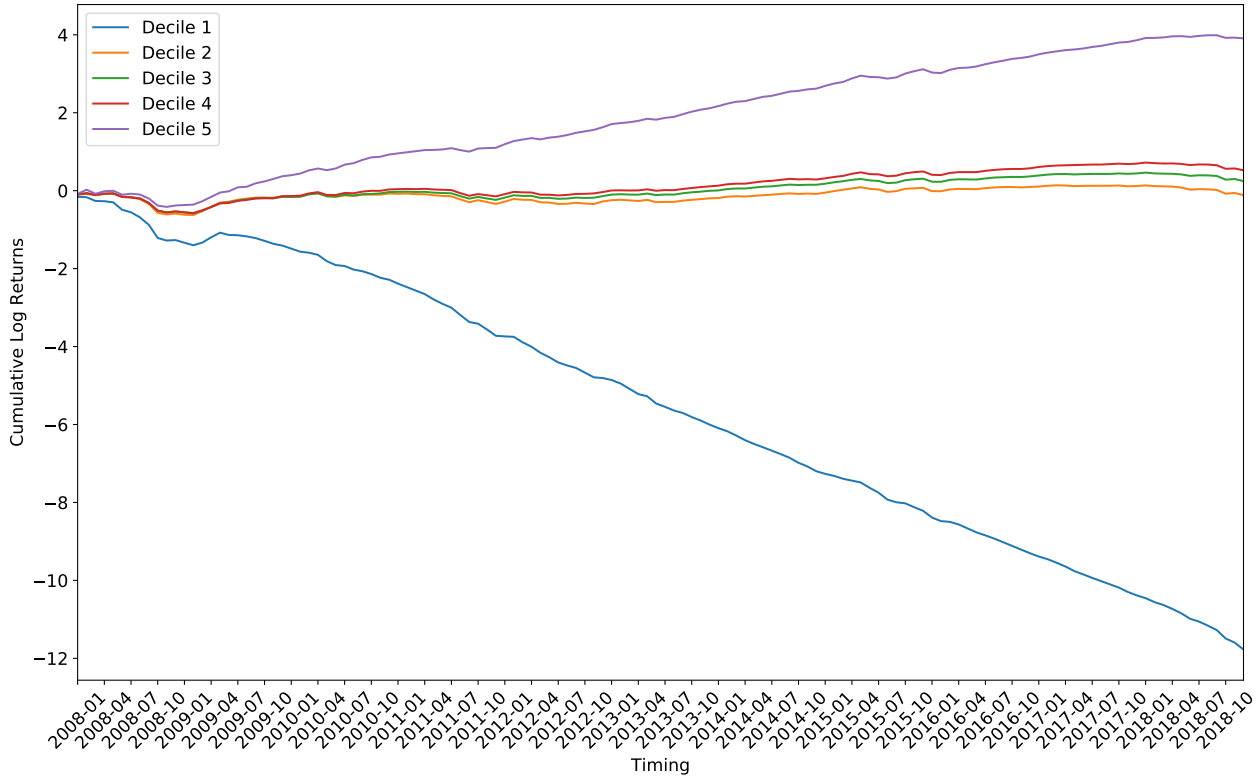

| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 132.000000 | 132.000000 | 132.000000 | 132.000000 | 132.000000 |
| mean | -0.000296 | 0.004128 | 0.008313 | 0.012185 | 0.018905 |
| std | 0.098565 | 0.095198 | 0.097386 | 0.099656 | 0.103610 |
| min | -0.293151 | -0.300826 | -0.306687 | -0.298766 | -0.301268 |
| 25% | -0.053913 | -0.041484 | -0.040400 | -0.040686 | -0.039781 |
| 50% | 0.003532 | 0.006626 | 0.011412 | 0.010805 | 0.024020 |
| 75% | 0.054781 | 0.058489 | 0.062428 | 0.075051 | 0.081953 |
| max | 0.243530 | 0.214880 | 0.230357 | 0.256440 | 0.314424 |
| Sharpe Ratio | -0.003000 | 0.043366 | 0.085360 | 0.122267 | 0.182467 |

**Figure 13:** Cumulative Returns (in log transformation) of Equally-weighted portfolios based on OLS prediction in China Stock Market from 2008 to 2018



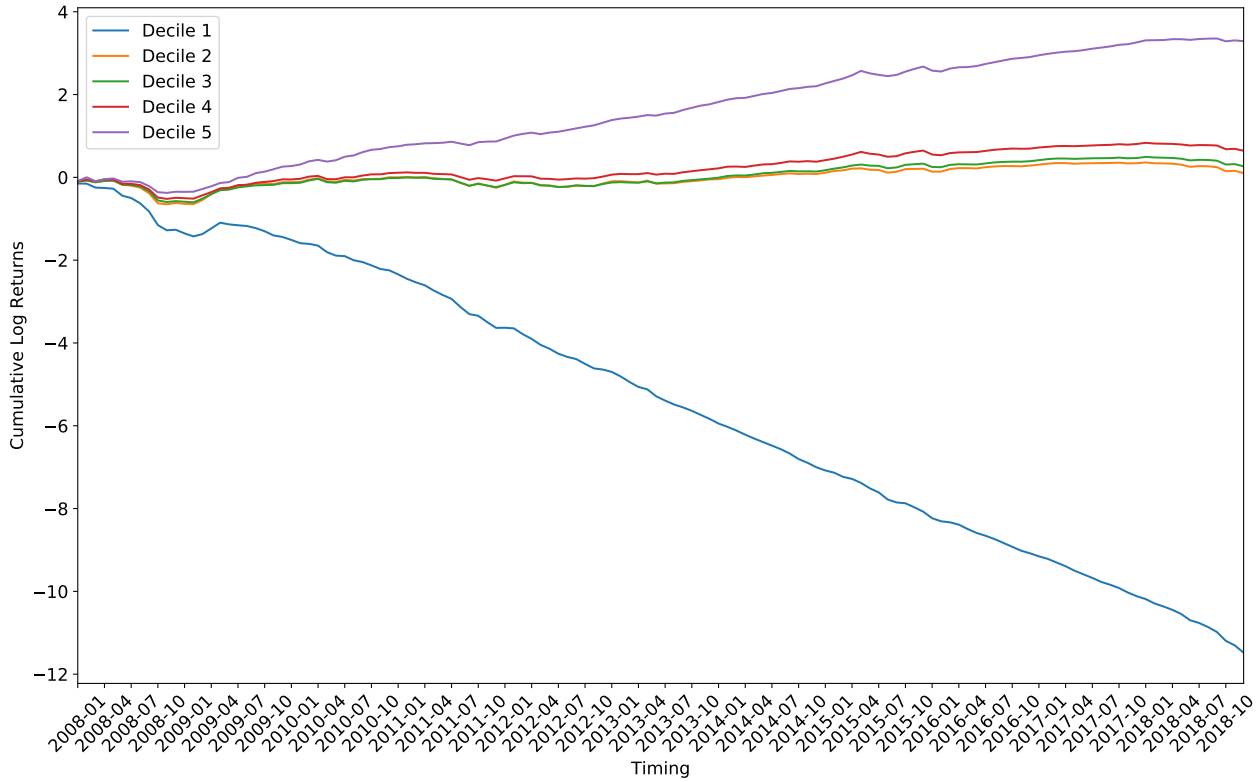

| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 132.000000 | 132.000000 | 132.000000 | 132.000000 | 132.000000 |
| mean | 0.006134 | 0.007548 | 0.010018 | 0.010302 | 0.009239 |
| std | 0.101404 | 0.101357 | 0.100018 | 0.097448 | 0.092888 |
| min | -0.294652 | -0.305872 | -0.292653 | -0.301628 | -0.287632 |
| 25% | -0.042448 | -0.041146 | -0.041078 | -0.038247 | -0.044087 |
| 50% | 0.003086 | 0.006161 | 0.014595 | 0.010521 | 0.009899 |
| 75% | 0.064353 | 0.064761 | 0.070156 | 0.070217 | 0.061040 |
| max | 0.261840 | 0.282919 | 0.278199 | 0.237953 | 0.214227 |
| Sharpe Ratio | 0.060488 | 0.074473 | 0.100160 | 0.105722 | 0.099469 |

**Figure 14:** Cumulative Returns (in log transformation) of Equally-weighted portfolios based on NN3 prediction around Global Market from 2008 to 2018




| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 132.000000 | 132.000000 | 132.000000 | 132.000000 | 132.000000 |
| mean | -0.083866 | 0.000048 | 0.002609 | 0.004745 | 0.031124 |
| std | 0.052390 | 0.042173 | 0.038282 | 0.038614 | 0.046248 |
| min | -0.286585 | -0.205280 | -0.182435 | -0.176453 | -0.166066 |
| 25% | -0.107281 | -0.015478 | -0.010673 | -0.008816 | 0.013631 |
| 50% | -0.085664 | 0.002816 | 0.006792 | 0.008566 | 0.037076 |
| 75% | -0.064104 | 0.020855 | 0.023421 | 0.027631 | 0.057363 |
| max | 0.145650 | 0.123235 | 0.106037 | 0.098307 | 0.120225 |
| Sharpe Ratio | -1.600790 | 0.001128 | 0.068157 | 0.122882 | 0.672973 |

**Figure 15:** Cumulative Returns (in log transformation) of Equally-weighted portfolios based on OLS prediction around Global Market from 2008 to 2018




| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 132.000000 | 132.000000 | 132.000000 | 132.000000 | 132.000000 |
| mean | -0.081863 | 0.001846 | 0.002885 | 0.005661 | 0.026133 |
| std | 0.052170 | 0.045781 | 0.040818 | 0.039251 | 0.042351 |
| min | -0.281405 | -0.217629 | -0.200465 | -0.178732 | -0.138601 |
| 25% | -0.103052 | -0.015345 | -0.010496 | -0.008007 | 0.011784 |
| 50% | -0.085441 | 0.004827 | 0.006917 | 0.009201 | 0.030637 |
| 75% | -0.064374 | 0.022780 | 0.025059 | 0.028414 | 0.051452 |
| max | 0.148700 | 0.154039 | 0.114332 | 0.092180 | 0.111261 |
| Sharpe Ratio | -1.569146 | 0.040315 | 0.070689 | 0.144217 | 0.617042 |

**Table 3:** Detailed descriptions of firm-level characteristics in the North American market

| No. | Acronym | Firm characteristic | Paper's author(s) | Year | Journal | Data Source | Frequency |
|---|---|---|---|---|---|---|---|
| 1 | *absacc* | Absolute accruals | Bandyopadhyay, Huang & Wirjanto | 2010 | WP | Compustat | Annual |
| 2 | *acc* | Working capital accruals | Sloan | 1996 | TAR | Compustat | Annual |
| 3 | *aeavol* | Abnormal earnings announcement volume | Lerman, Livnat & Mendenhall | 2007 | WP | Compustat +CRSP | Quarterly |
| 4 | *age* | # years since first Compustat coverage | Jiang, Lee & Zhang | 2005 | RAS | Compustat | Annual |
| 5 | *agr* | Asset growth | Cooper, Gulen & Schill | 2008 | JF | Compustat | Annual |
| 6 | *baspread* | Bid-ask spread | Ambihud & Mendelson | 1989 | JF | CRSP | Monthly |
| 7 | *beta* | Beta | Fama & MacBeth | 1973 | JPE | CRSP | Monthly |
| 8 | *betasq* | Beta squared | Fama & MacBeth | 1973 | JPE | CRSP | Monthly |
| 9 | *bm* | Book-to-market | Rosenberg, Reid & Lanstein | 1985 | JPM | Compustat +CRSP | Annual |
| 10 | *bm_ia* | Industry-adjusted book to market | Asness, Porter & Stevens | 2000 | WP | Compustat +CRSP | Annual |
| 11 | *cash* | Cash holdings | Palazzo | 2012 | JFE | Compustat | Quarterly |
| 12 | *cashdebt* | Cash flow to debt | Ou and Penman | 1989 | JAE | Compustat | Annual |
| 13 | *cashpr* | Cash productivity | Chandrashekar & Rao | 2009 | WP | Compustat | Annual |
| 14 | *cfp* | Cash-flow-to-price-ratio | Desai, Rajgopal & Venkatachalam | 2004 | TAR | Compustat | Annual |
| 15 | *cfp_ia* | Industry-adjusted cash-flow-to-price ratio | Asness, Porter & Stevens | 2000 | WP | Compustat | Annual |
| 16 | *chatoia* | Industry-adjusted change in asset turnover | Soliman | 2008 | TAR | Compustat | Annual |
| 17 | *chcsho* | Change in shares outstanding | Pontiff and Woodgate | 2008 | JF | Compustat | Annual |
| 18 | *chempia* | Industry-adjusted change in employees | Aeness, Porter & Stevens | 1994 | WP | Compustat | Annual |
| 19 | *chfeps* | Changes in forecasted EPS | Hawkins, Chamberlin & Daniel | 1984 | FAJ | Compustat | Monthly |
| 20 | *chinv* | Change in inventory | Thomas & Zhang | 2002 | JAR | Compustat | Annual |
| 21 | *chmom* | Change in 6-month momentum | Gettleman & Marks | 2006 | WP | CRSP | Monthly |
| 22 | *chnanalyst* | Change in number of analysts | Scherbina | 2008 | RF | Compustat +I/B/E/S | Monthly |
| 23 | *chpmia* | Industry-adjusted change in profit margin | Soliman | 2008 | TAR | Compustat | Annual |
| 24 | *chtx* | Change in tax expense | Thomas & Zhang | 2011 | JAR | Compustat | Quarterly |
| 25 | *cinvest* | Corporate investment | Titman, Wei & Xie | 2004 | JFQA | Compustat | Quarterly |

**Table 3:** Detailed descriptions of firm-level characteristics in the North American market (Continued)

| No. | Acronym | Firm characteristic | Paper's author(s) | Year | Journal | Data Source | Frequency |
|---|---|---|---|---|---|---|---|
| 26 | *convind* | Convertible debt indicator | Valta | 2016 | JFQA | Compustat | Annual |
| 27 | *currat* | Current ratio | Ou & Penman | 1989 | JAE | Compustat | Annual |
| 28 | *depr* | Depreciation /PP&E | Holthausen & Larcker | 1992 | JAE | Compustat | Annual |
| 29 | *disp* | Dispersion in forecasted EPS | Diether, Malloy & Scherbina | 2002 | JF | I/B/E/S | Annual |
| 30 | *divi* | Dividend initiation | Michaely, Thaler & Womack | 1995 | JF | Compustat | Annual |
| 31 | *orgcap* | Organizational capital | Eisfeldt & Papanikolaou | 2013 | JF | Compustat | Annual |
| 32 | *pchcapx_ia* | Industry adjusted % change in capital expenditures | Abarbanel & Bushee | 1998 | TAR | Compustat | Annual |
| 33 | *pchcurrat* | % change in current ratio | Ou & Penman | 1989 | JAE | Compustat | Annual |
| 34 | *pchdepr* | % change in depreciation | Holthausen & Larcher | 1992 | JAE | Compustat | Annual |
| 35 | *pchgm_pchsale* | % change in gross margin − % change in sales | Abarbanell & Bushee | 1998 | TAR | Compustat | Annual |
| 36 | *pchquick* | % change in quick ratio | Ou & Penman | 1989 | JAE | Compustat | Annual |
| 37 | *pchsale_pchinvt* | % change in sales − % change in inventory | Abarbanell & Bushee | 1998 | TAR | Compustat | Annual |
| 38 | *pchsale_pchrect* | % change in sales − % change in A/R | Abarbanell & Bushee | 1998 | TAR | Compustat | Annual |
| 39 | *pchsale_pchxsga* | % change in sales − % change in SG&A | Abarbanell & Bushee | 1998 | TAR | Compustat | Annual |
| 40 | *pchsaleinv* | % change in sales-to-inventory | Ou & Penman | 1989 | JAE | Compustat | Annual |
| 41 | *pctacc* | Percent accruals | Hafzalla, Lundholm, &Van Winke | 2011 | TAR | Compustat | Annual |
| 42 | *pricedelay* | Price delay | Hou & Moskowitz | 2005 | RFS | CRSP | Monthly |
| 43 | *ps* | Financial statements score | Piotroski | 2000 | JAR | Compustat | Annual |
| 44 | *quick* | Quick ratio | Ou & Penman | 1989 | JAE | Compustat | Annual |
| 45 | *rd* | R&D increase | Eberhart, Maxwell & Siddique | 2004 | JF | Compustat | Annual |
| 46 | *rd_mve* | R&D to market capitalization | Guo, Lev & Shi | 2006 | JBFA | Compustat | Annual |
| 47 | *rd_sale* | R&D to sales | Guo, Lev & Shi | 2006 | JBFA | Compustat | Annual |
| 48 | *realestate* | Real estate holdings | Tuzel | 2010 | RFS | Compustat | Annual |
| 49 | *retvol* | Return volatility | Ang, Hodrick, Xing & Zhang | 2006 | JF | CRSP | Monthly |
| 50 | *roaq* | Return on assets | Balakrishnan, Bartov & Faurel | 2010 | JAE | Compustat | Quarterly |

**Table 3:** Detailed descriptions of firm-level characteristics in the North American market (Continued)

| No. | Acronym | Firm characteristic | Paper's author(s) | Year | Journal | Data Source | Frequency |
|---|---|---|---|---|---|---|---|
| 51 | *roavol* | Earnings volatility | Francis, LaFond, Olsson & Schipper | 2004 | TAR | Compustat | Quarterly |
| 52 | *divo* | Dividend omission | Michaely, Thaler & Womack | 1995 | JF | Compustat | Annual |
| 53 | *dolvol* | Dollar trading volume | Chordia, Subrahmanyan & Anshuman | 2001 | JFE | CRSP | Monthly |
| 54 | *dy* | Dividend to price | Litzenberger & Ramaswamy | 1982 | JF | Compustat | Annual |
| 55 | *ear* | Earnings announcement return | Kishore, Brandt, Santa-Clara & Venkatachalam | 2008 | WP | Compustat +CRSP | Quarterly |
| 56 | *egr* | Growth in common shareholder equity | Richardson,Sloan, Soliman & Tuna | 2005 | JAE | Compstat | Annual |
| 57 | *ep* | Earnings to price | Basu | 1997 | JF | Compustat | Annual |
| 58 | *fgr5yr* | Forecasted growth in 5-year EPS | Bauman & Dowen | 1988 | FAJ | I/B/E/S | Annual |
| 59 | *gma* | Gross profitability | Norvy-Marx | 2013 | JFE | Compustat | Annual |
| 60 | *grCAPX* | Growth in capital expenditures | Anderson & Garcia-Feijoo | 2006 | JF | Compustat | Annual |
| 61 | *grltnoa* | Growth in long-term net operating assets | Fairfield, Whisenant & Yohn | 2003 | TAR | Compustat | Annual |
| 62 | *herf* | Industry sales concentration | Hou & Robinson | 2006 | JF | Compustat | Annual |
| 63 | *hire* | Employee growth rate | Bazdresch, Belo & Lin | 2014 | JPE | Compustat | Annual |
| 64 | *idiovol* | Idiosyncratic return volatility | Ali, Hwang & Trombley | 2003 | JFE | Compustat | Monthly |
| 65 | *ill* | Illiquidity | Amihud | 2002 | JFM | CRSP | Monthly |
| 66 | *indmom* | Industry momentum | Moskowitz & Grinblatt | 1999 | JF | CRSP | Monthly |
| 67 | *invest* | Capital expenditures and inventory | Chen & Zhang | 2010 | JF | Compustat | Annual |
| 68 | *IPO* | New equity issue | Loughran & Ritter | 1995 | JF | Compustat | Monthly |
| 69 | *lev* | Leverage | Bhandari | 1988 | JF | Compustat | Annual |
| 70 | *lgr* | Growth in long-term debt | Richardson, Sloan, Soliman & Tuna | 2005 | JAE | Compustat | Annual |
| 71 | *maxret* | Maximum daily return | Bali, Cakici & Whitelaw | 2011 | JFE | CRSP | Monthly |
| 72 | *mom12m* | 12-month momentum | Jegadeesh | 1990 | JF | CRSP | Monthly |
| 73 | *mom1m* | 1-month momentum | Jegadeesh & Titman | 1993 | JF | CRSP | Monthly |
| 74 | *mom36m* | 36-month momentum | Jegadeesh & Titman | 1993 | JF | CRSP | Monthly |
| 75 | *mom6m* | 6-month momentum | Jegadeesh & Titman | 1993 | JF | CRSP | Monthly |

**Table 3:** Detailed descriptions of firm-level characteristics in the North American market (Continued)

| No. | Acronym | Firm characteristic | Paper's author(s) | Year | Journal | Data Source | Frequency |
|---|---|---|---|---|---|---|---|
| 76 | *ms* | Financial statement score | Mohanram | 2005 | RAS | Compustat | Quarterly |
| 77 | *mve* | Size | Banz | 1981 | JFE | CRSP | Monthly |
| 78 | *mve_ia* | Industry-adjusted size | Asness, Porter & Stevens | 2000 | WP | Compustat | Annual |
| 79 | *nanalyst* | Number of analysts covering stock | Elgers, Lo & Pfeiffer | 2001 | TAR | Compustat +I/B/E/S | Monthly |
| 80 | *nincr* | Number of earnings increases | Barth, Elliott & Finn | 1999 | JAR | Compustat | Quarterly |
| 81 | *operprof* | Operating profitability | Fama & French | 2015 | JFE | Compustat | Annual |
| 82 | *roeq* | Return on equity | Hou, Xue & Zhang | 2015 | RFS | Compustat | Quarterly |
| 83 | *roic* | Return on invested capital | Brown & Rowe | 2007 | WP | Compustat | Annual |
| 84 | *rsup* | Revenue surprise | Kama | 2009 | JBFA | Compustat | Quarterly |
| 85 | *salecash* | Sales to cash | Ou & Penman | 1989 | JAE | Compustat | Annual |
| 86 | *saleinv* | Sales to inventory | Ou & Penman | 1989 | JAE | Compustat | Annual |
| 87 | *salerec* | Sales to receivables | Ou & Penman | 1989 | JAE | Compustat | Annual |
| 88 | *secured* | Secured debt | Valta | 2016 | JFQA | Compustat | Annual |
| 89 | *securedind* | Secured debt indicator | Valta | 2016 | JFQA | Compustat | Annual |
| 90 | *sfe* | Scaled earnings forecast | Elgers, Lo & Pfeiffer | 2001 | TAR | Compustat +I/B/E/S | Monthly |
| 91 | *sgr* | Sales growth | Lakonishok, Sheifer & Vishny | 1994 | JF | Compustat | Annual |
| 92 | *sin* | Sin stocks | Hong & Kacperczyk | 2009 | JFE | Compustat | Annual |

**Table 3:** Detailed descriptions of firm-level characteristics in the North American market (Continued)

| No. | Acronym | Firm characteristic | Paper's author(s) | Year | Journal | Data Source | Frequency |
|---|---|---|---|---|---|---|---|
| 93 | *SP* | Sales to price | Barbee, Mukherji & Raines | 1996 | FAJ | Compustat | Annual |
| 94 | *std_dolvol* | Volatility of liquidity (dollar trading volume) | Chordia, Subrahmanyam & Anshuman | 2001 | JFE | CRSP | Monthly |
| 95 | *std_turn* | Volatility of liquidity (share turnover) | Chordia, Subrahmanyam & Anshuman | 2001 | JFE | CRSP | Monthly |
| 96 | *stdacc* | Accrual volatility | Bandyopadhyay, Huang & Wirjanto | 2010 | WP | Compustat | Quarterly |
| 97 | *stdcf* | Cash flow volatility | Huang | 2009 | JEF | Compustat | Quarterly |
| 98 | *sue* | Unexpected quarterly earnings | Rendelman, Jones & Latane | 1982 | JFE | Compustat +CRSP | Quarterly |
| 99 | *tang* | Debt capacity/firm tangibility | Almeida & Campello | 2007 | RFS | Compustat | Annual |
| 100 | *tb* | Tax income to book income | Lev & Nissim | 2004 | TAR | Compustat | Annual |
| 101 | *turn* | Share turnover | Datar, Naik & Radcliffe | 1998 | JFM | CRSP | Monthly |
| 102 | *zerotrade* | Zero trading days | Liu | 2006 | JFE | CRSP | Monthly |

## References


Berger, J. O., J. M. Bernardo, and D. Sun (2009): "The Formal Definition of Reference Priors," *The Annals of Statistics*, 37, 905–938. [Cited on page 8.]

Breiman, J. H. (1991): "Multivariate Adaptive Regression Splines," *Annals of Statistics*, 19, 1–67. [Cited on page 5.]

Breiman, L. (2001): "Random forests," *Machine learning*, 45, 5–32. [Cited on page 5.]

Campbell, J. and S. P. Thompson (2007): "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average," *The Review of Financial Studies*, 21, 1509–1531. [Cited on page 9.]

Chen, L., M. Pelger, and J. Zhu (2019): "Deep Learning in Asset Pricing," Working paper. [Cited on page 1.]

Chen, Y. (2019): "Jeffreys' Prior Asymptotically and Approximately Maximizes Expected Information," Working paper. [Cited on page 8.]

Chinco, A., A. D. Clark-Joseph, and M. Ye (2019): "Sparse Signals in the Cross-Section of Returns," Manuscript, forthcoming for *Journal of Finance*. [Cited on page 1.]

Chipman, H. A., E. I. George, and R. E. McCulloch (1998): "Bayesian CART Model Search," *Journal of the American Statistical Association*, 93, 935–948. [Cited on pages 5, 7, and 8.]

——— (2010): "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298. [Cited on pages 5 and 7.]

Clarke, B. S. (1994): "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference*, 41, 37–60. [Cited on page 8.]

Cochrane, J. H. (2017): "Presidential Address: Discount rates," *Journal of Finance*, 66, 1047–1108. [Cited on page 1.]

Cremers, K. J. M. (2002): "Stock Return Predictability: A Bayesian Model Selection Perspective," *The Review of Financial Studies*, 15, 1223–1249. [Cited on page 1.]

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998): "A Bayesian CART Algorithm," *Biometrica*, 85, 363–377. [Cited on page 5.]

Fama, E. F. and K. R. French (2015): "Dissecting Anomalies with a Five-Factor Model," *The Review of Financial Studies*, 29, 69–103. [Cited on page 4.]

Freyberger, J., A. Neuhierl, and M. Weber (2019): "Dissecting Characteristics Nonparametrically," Working paper, forthcoming in *The Review of Financial Studies*. [Cited on pages 1, 3, and 23.]

Green, J., J. R. M. Hand, and X. F. Zhang (2017): "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns," *The Review of Financial Studies*, 30, 4389–4436. [Cited on page 10.]

Gu, S., B. Kelly, and D. Xiu (2019): "Empirical Asset Pricing via Maching Learning," Working paper, forthcoming in the *The Review of Financial Studies*. [Cited on pages 1, A-9, and A-10.]

Han, Y., A. He, D. E. Rapach, and G. Zhou (2019): "What Firm Characteristics Drive US Stock Returns," Working paper. [Cited on page 1.]

Harvey, C. R., Y. Liu, and H. Zhu (2016): "...and the Cross-Section of Expected Returns," *The Review of Financial Studies*, 29, 5–68. [Cited on page 1.]

James, G., D. Witten, T. Hastie, and R. Tibshirani (2017): *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. [Cited on page A-1.]

Kozak, S. (2019): "Kernel Trick for the Cross Sectioin," Working paper. [Cited on page 14.]

Kozak, S., S. Nagel, and S. Santosh (2018): "Shrinking the Cross Section," Working paper, just accepted for *Journal of Financial Economics*. [Cited on pages 1 and 14.]

Linero, A. R. (2018): "Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection," *Journal of the American Statistical Association*, 113, 626–636. [Cited on page 2.]

Rapach, D. E., J. K. Strauss, and G. Zhou (2010): "Out-of-sample Equity Premium Prediction: Combination Forecasts and links to the Real Economy," *The Review of Financial Studies*, 23, 821–862. [Cited on page 9.]

Ročková, V. (2019): "On Semi-parametric Bernstein-von Mises Theorems for BART," Working paper. [Cited on page 2.]

Ročková, V. and E. Saha (2019): "On Theory for BART," Working paper, $22^{nd}$ International Conference on Artificial Intelligence and Statistics. [Cited on page 2.]

Ročková, V. and S. van der Pas (2019): "Posterior Concentration for Bayesian Regression Trees and Forests," Manuscript, just accepted for *Annals of Statistics*. [Cited on page 2.]

Stambaugh, R. F. (1999): "Predictive Regressions," *Journal of Financial Economics*, 54, 375–421. [Cited on page 1.]

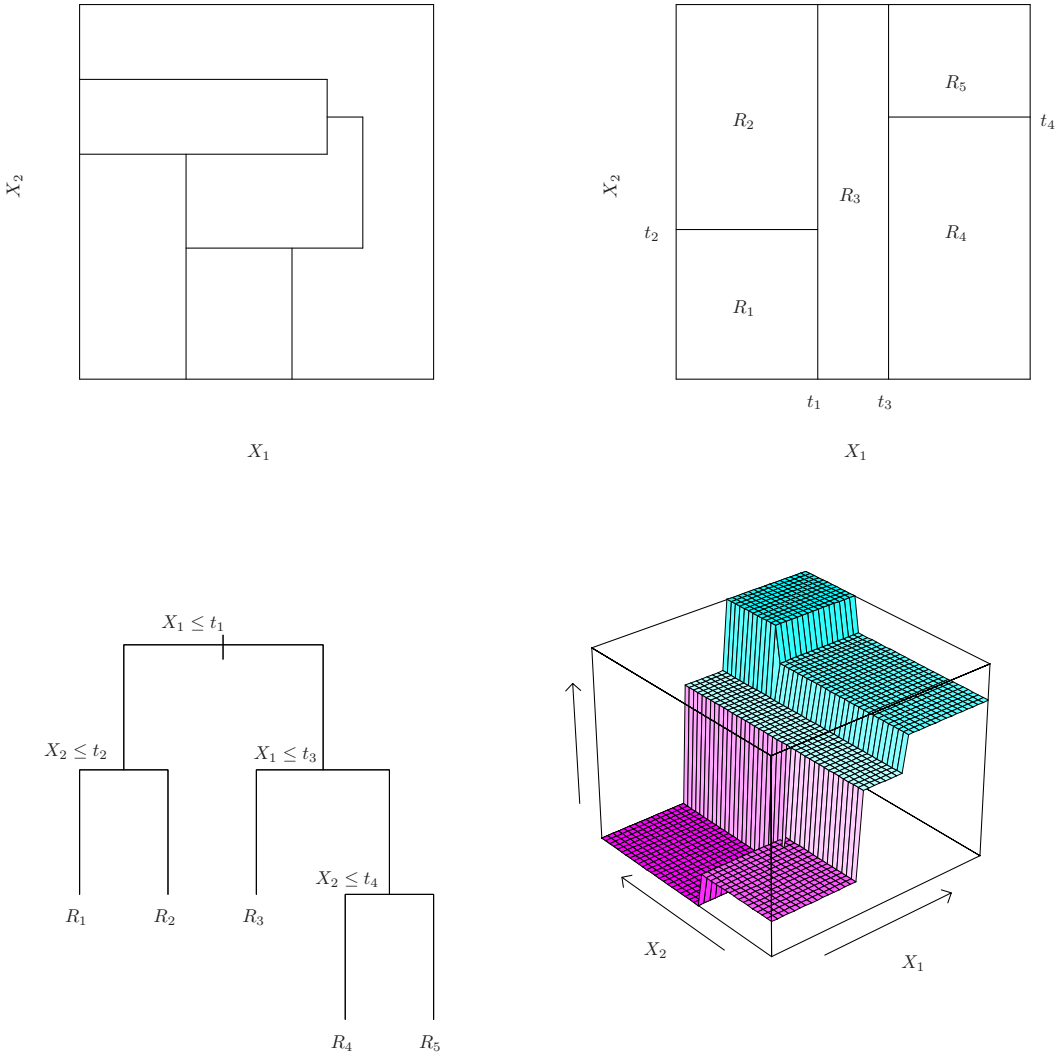
Stambaugh, R. F. and L. Pastor (2000): "Comparing Asset Pricing Models: An Investment Perspective," *Journal of Financial Economics*, 56, 335–381. [Cited on page 1.]

Yuan, M. and Y. Lin (2006): "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67. [Cited on page 3.]

# Appendix

## A More Demonstration Figures and Tables

**Figure A.1:** Additional Demonstration of Regression Tree




**Note:** This figure is for additional general demonstration of how the structure of regression tree with more partitioned areas is, where the two-dimensional covariate plane is partitioned into 5 different areas. Top Left: A partition of two-dimensional covariate plane that could not result from recursive binary splitting. Top Right: how the two-dimensional plane is partitioned into 5 different areas based on recursive binary splitting. Bottom Left: A tree structure corresponding to the partition in the top right panel. Bottom Right: based on the partition rule in the bottom left panel, how a specific step function approximation of unknown functional form is. This figure and corresponding explanation is based on ideas from James et al. (2017), with permission from authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

**Table A.1:** A Simulation Result from Network (200 nodes)

| Training Sample Size | Number of Covariates | Testing Sample Size | Out of Sample R Square |
|---|---|---|---|
| 500 | 20 | 100 | 0.156745 |
| 500 | 50 | 100 | -0.034445 |
| 500 | 100 | 100 | 0.025742 |
| 5000 | 20 | 1000 | 0.300099 |
| 5000 | 50 | 1000 | 0.082347 |
| 5000 | 100 | 1000 | -0.060066 |
| 50000 | 20 | 10000 | 0.478012 |
| 50000 | 50 | 10000 | 0.381836 |
| 50000 | 100 | 10000 | 0.266237 |

**Table A.2:** A Simulation Result from Network (64 nodes)

| Training Sample Size | Number of Covariates | Testing Sample Size | Out of Sample R Square |
|---|---|---|---|
| 500 | 20 | 100 | 0.237615 |
| 500 | 50 | 100 | 0.001637 |
| 500 | 100 | 100 | -0.039926 |
| 5000 | 20 | 1000 | 0.281587 |
| 5000 | 50 | 1000 | 0.118396 |
| 5000 | 100 | 1000 | 0.030965 |
| 50000 | 20 | 10000 | 0.493948 |
| 50000 | 50 | 10000 | 0.427575 |
| 50000 | 100 | 10000 | 0.315529 |

**Table A.3:** Additional Out-of-sample $R^2_{OS}$(%) Comparison

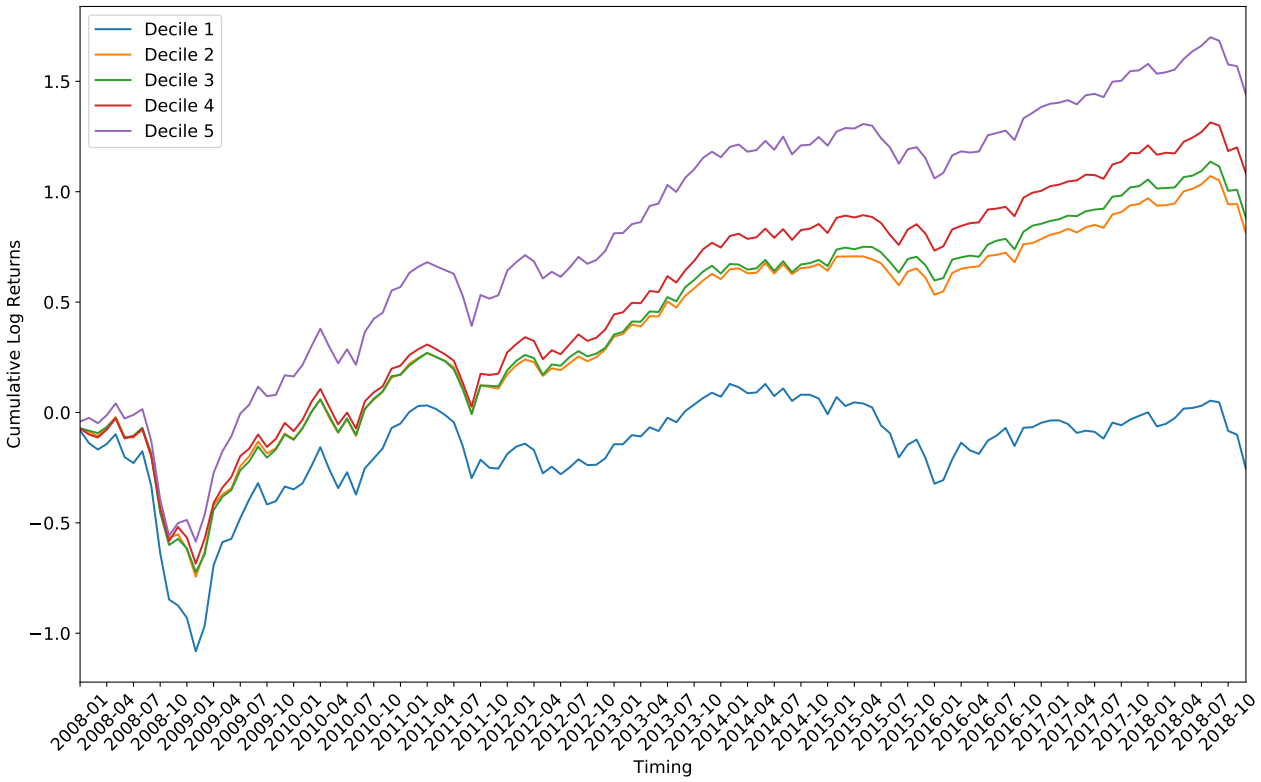
| Test Sample Length 4 (year) | | | |
|---|---|---|---|
| Test Sample Window | BART | OLS | Ratio |
| 2013-2016 | 1.06 | 0.19 | **5.58** |
| 2014-2017 | 0.09 | -0.65 | **-0.14** |
| 2015-2018 | -0.69 | -1.43 | **0.48** |
| Test Sample Length 3 (year) | | | |
| Test Sample Window | BART | OLS | Ratio |
| 2013-2015 | 1.18 | 0.16 | **7.35** |
| 2014-2016 | -0.15 | -1.11 | **-0.14** |
| 2015-2017 | 0.25 | -0.4 | **-0.63** |
| 2016-2018 | -0.60 | -1.13 | **0.53** |
| Test Sample Length 2 (year) | | | |
| Test Sample Window | BART | OLS | Ratio |
| 2013-2014 | 2.58 | 1.75 | **1.47** |
| 2014-2015 | -0.80 | -2.03 | **0.39** |
| 2015-2016 | -0.02 | -0.92 | **0.02** |
| 2016-2017 | 0.89 | 0.60 | **1.48** |
| 2017-2018 | -1.58 | -2.11 | **0.75** |
| Test Sample Length 1 (year) | | | |
| Test Sample Window | BART | OLS | Ratio |
| 2013 | 4.95 | 4.35 | **1.14** |
| 2014 | -0.58 | -1.71 | **0.34** |
| 2015 | -0.93 | -2.24 | **0.42** |
| 2016 | 0.79 | 0.25 | **3.16** |
| 2017 | 1.03 | 1.12 | 0.92 |
| 2018 | -3.79 | -4.84 | **0.78** |

**Note:** This table demonstrates the empirical out-of-sample $R^2_{OS}$ for different test sample specifications. In the table demonstrated above, the last column "Ratio" refers to the ratio calculated with the $R^2_{OS}$ of BART over the $R^2_{OS}$ of OLS. For the rows where the ratios are highlighted, BART performs better than OLS for that corresponding test sample specification indicated in the first column.

| Test Sample Length 4 (year) | | | | |
|---|---|---|---|---|
| Test Sample Window | BART | NN3 | OLS | Ratio |
| 2013-2016 | 1.06 | 0.41 | 0.19 | 5.58 |
| 2014-2017 | 0.09 | -0.57 | -0.65 | -0.14 |
| 2015-2018 | -0.69 | -1.22 | -1.43 | 0.48 |
| Test Sample Length 3 (year) | | | | |
| Test Sample Window | BART | NN3 | OLS | Ratio |
| 2013-2015 | 1.18 | 0.59 | 0.16 | 7.35 |
| 2014-2016 | -0.15 | -0.76 | -1.11 | -0.14 |
| 2015-2017 | 0.25 | -0.48 | -0.4 | -0.63 |
| 2016-2018 | -0.60 | -1.13 | -1.13 | 0.53 |
| Test Sample Length 2 (year) | | | | |
| Test Sample Window | BART | NN3 | OLS | Ratio |
| 2013-2014 | 2.58 | 1.94 | 1.75 | 1.47 |
| 2014-2015 | -0.80 | -1.26 | -2.03 | 0.39 |
| 2015-2016 | -0.02 | -0.69 | -0.92 | 0.02 |
| 2016-2017 | 0.89 | 0.05 | 0.60 | 1.48 |
| 2017-2018 | -1.58 | -1.92 | -2.11 | 0.75 |
| Test Sample Length 1 (year) | | | | |
| Test Sample Window | BART | NN3 | OLS | Ratio |
| 2013 | 4.95 | 4.14 | 4.35 | 1.14 |
| 2014 | -0.58 | -0.98 | -1.71 | 0.34 |
| 2015 | -0.93 | -1.45 | -2.24 | 0.42 |
| 2016 | 0.79 | -0.006 | 0.25 | 3.16 |
| 2017 | 1.03 | 0.14 | 1.12 | 0.92 |
| 2018 | -3.79 | -3.66 | -4.84 | 0.78 |

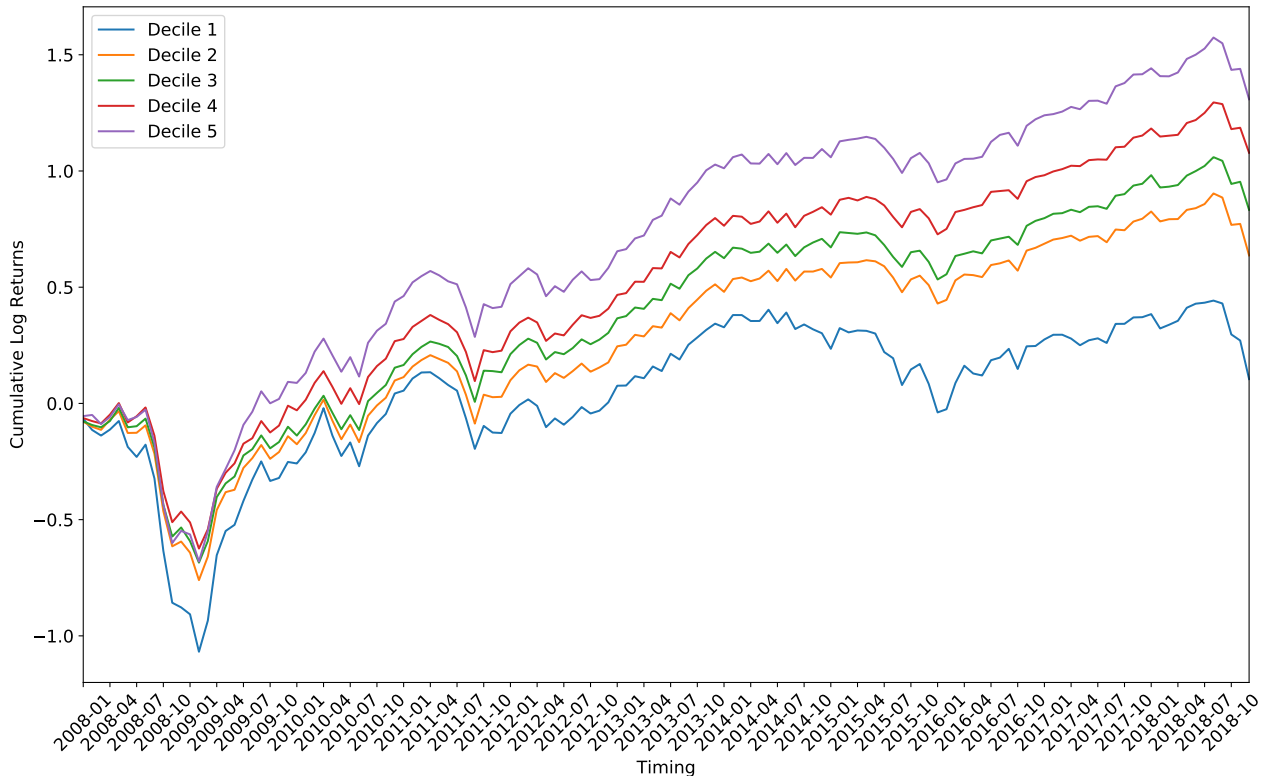
| Test Sample Length 4 (year) | | | | |
|---|---|---|---|---|
| Test Sample Window | BART | NN3 | OLS | Ratio |
| 2013-2016 | 1.06 | 0.64 | 0.19 | 5.58 |
| 2014-2017 | 0.09 | -0.28 | -0.65 | -0.14 |
| 2015-2018 | -0.69 | -0.91 | -1.43 | 0.48 |
| Test Sample Length 3 (year) | | | | |
| Test Sample Window | BART | NN3 | OLS | Ratio |
| 2013-2015 | 1.18 | 0.64 | 0.16 | 7.35 |
| 2014-2016 | -0.15 | -0.51 | -1.11 | -0.14 |
| 2015-2017 | 0.25 | -0.11 | -0.4 | -0.63 |
| 2016-2018 | -0.60 | -0.70 | -1.13 | 0.53 |
| Test Sample Length 2 (year) | | | | |
| Test Sample Window | BART | NN3 | OLS | Ratio |
| 2013-2014 | 2.58 | 2.03 | 1.75 | 1.47 |
| 2014-2015 | -0.80 | -1.28 | -2.03 | 0.39 |
| 2015-2016 | -0.02 | -0.36 | -0.92 | 0.02 |
| 2016-2017 | 0.89 | 0.63 | 0.60 | 1.48 |
| 2017-2018 | -1.58 | -1.65 | -2.11 | 0.75 |
| Test Sample Length 1 (year) | | | | |
| Test Sample Window | BART | NN3 | OLS | Ratio |
| 2013 | 4.95 | 4.30 | 4.35 | 1.14 |
| 2014 | -0.58 | -0.98 | -1.71 | 0.34 |
| 2015 | -0.93 | -1.47 | -2.24 | 0.42 |
| 2016 | 0.79 | 0.64 | 0.25 | 3.16 |
| 2017 | 1.03 | 0.61 | 1.12 | 0.92 |
| 2018 | -3.79 | -3.56 | -4.84 | 0.78 |

**Figure A.2:** Cumulative Returns of Equally-weighted Portfolios based on BART Prediction in the U.S. Stock Market from 2008 to 2018



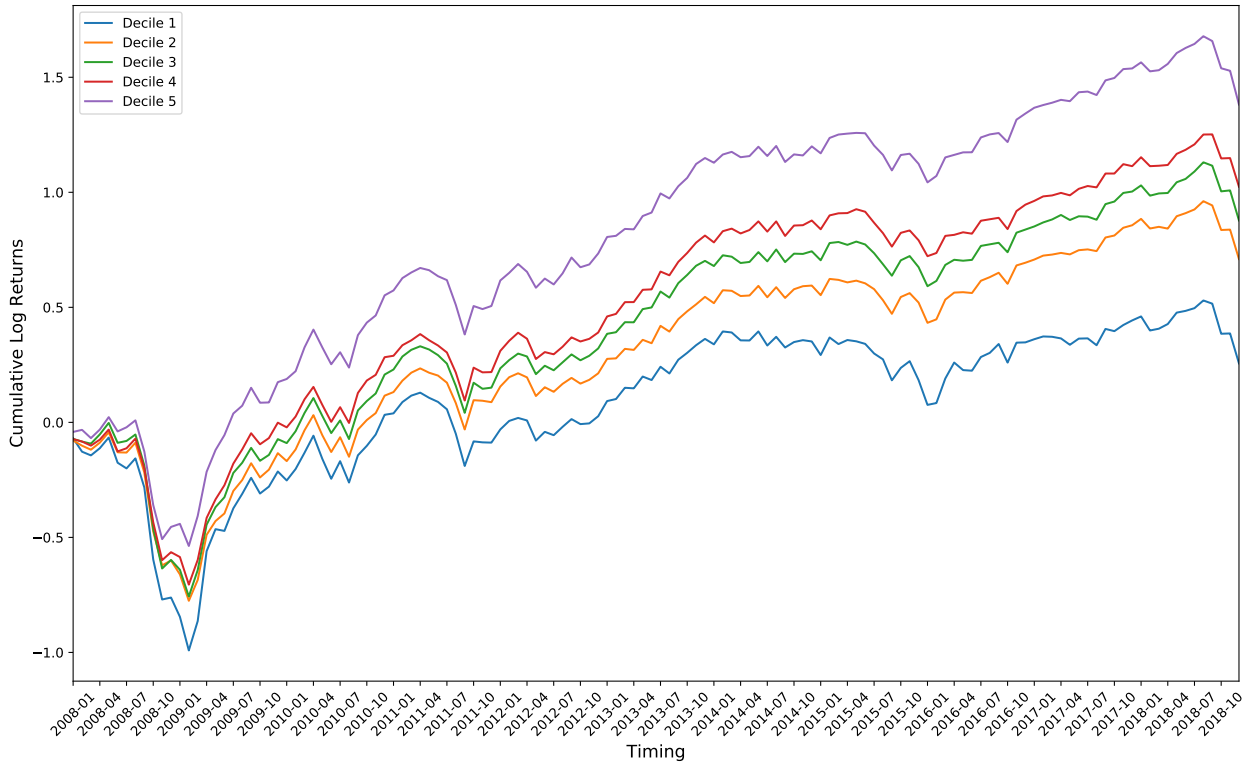

| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 132.000000 | 132.000000 | 132.000000 | 132.000000 | 132.000000 |
| mean | 0.000601 | 0.007844 | 0.008341 | 0.009953 | 0.012992 |
| std | 0.070858 | 0.057640 | 0.057130 | 0.058346 | 0.063342 |
| min | -0.262338 | -0.220831 | -0.229907 | -0.206494 | -0.231984 |
| 25% | -0.030787 | -0.018257 | -0.018778 | -0.021895 | -0.018278 |
| 50% | 0.006680 | 0.011392 | 0.010103 | 0.014230 | 0.016196 |
| 75% | 0.039924 | 0.041739 | 0.041570 | 0.045706 | 0.047745 |
| max | 0.315145 | 0.233608 | 0.223172 | 0.174141 | 0.208085 |
| Sharpe Ratio | 0.008482 | 0.136083 | 0.146004 | 0.170585 | 0.205108 |

**Figure A.3:** Cumulative Returns of Equally-weighted Portfolios based on NN3 Prediction in the U.S. Stock Market from 2008 to 2018




| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 132.000000 | 132.000000 | 132.000000 | 132.000000 | 132.000000 |
| mean | 0.003546 | 0.006563 | 0.007908 | 0.009772 | 0.011941 |
| std | 0.073964 | 0.058505 | 0.056002 | 0.055784 | 0.062619 |
| min | -0.267766 | -0.219147 | -0.218498 | -0.212000 | -0.234143 |
| 25% | -0.026869 | -0.021127 | -0.016268 | -0.018646 | -0.020009 |
| 50% | 0.008307 | 0.011557 | 0.011113 | 0.011890 | 0.017042 |
| 75% | 0.044345 | 0.040888 | 0.038291 | 0.045020 | 0.049267 |
| max | 0.324654 | 0.221566 | 0.207939 | 0.190238 | 0.209728 |
| Sharpe Ratio | 0.047940 | 0.112185 | 0.141201 | 0.175175 | 0.190688 |

**Figure A.4:** Cumulative Returns of Equally-weighted Portfolios based on OLS Prediction in the U.S. Stock Market from 2008 to 2018




| index | Port1 | Port2 | Port3 | Port4 | Port5 |
|---|---|---|---|---|---|
| count | 132.000000 | 132.000000 | 132.000000 | 132.000000 | 132.000000 |
| mean | 0.004389 | 0.007136 | 0.008529 | 0.009532 | 0.012392 |
| std | 0.070125 | 0.058769 | 0.060534 | 0.058818 | 0.061147 |
| min | -0.268633 | -0.225843 | -0.240222 | -0.209510 | -0.207347 |
| 25% | -0.027570 | -0.020325 | -0.021900 | -0.020161 | -0.020535 |
| 50% | 0.008146 | 0.013363 | 0.013634 | 0.011951 | 0.013696 |
| 75% | 0.042689 | 0.042262 | 0.042225 | 0.045521 | 0.045492 |
| max | 0.354843 | 0.216614 | 0.217992 | 0.200183 | 0.211467 |
| Sharpe Ratio | 0.062583 | 0.121427 | 0.140899 | 0.162052 | 0.202667 |

# B Additional Monte Carlo Simulations

We follow the data generating mechanism as described in Gu et al. (2019) as following, which allows the data to be generated in panel data framework.

$$r_{i,t+1} = f(\boldsymbol{x}_{i,t}) + e_{i,t+1},\ e_{i,t+1} = \beta_{i,t} v_{t+1} + \epsilon_{i,t+1},\ \boldsymbol{x}_{i,t} = (1, z_t)^\top \otimes c_{i,t},\ \beta_{i,t} = (c_{i1,t}, c_{i2,t}, c_{i3,t})$$

and $v_t$ is assumed to be normally distributed

$$v_{t+1} \sim \mathcal{N}\left(0, 0.05^2 \mathrm{I}_3\right)$$

where $\mathrm{I}_3$ denotes the identity matrix with dimension as 3. For $c_{i,t}$ and $c_t$, the structure is basically as following

$$c_{i,t} = \begin{pmatrix} c_{i1,t} \\ \vdots \\ c_{iP_c,t} \end{pmatrix}_{P_c \times 1} \qquad c_t = \begin{pmatrix} c_{1,t}^\top \\ \vdots \\ c_{N,t}^\top \end{pmatrix}_{N \times P_c}$$

In practical empirical implementation, $z_t$ is in general a $P_z \times 1$ vector as following, which collects all the macro variables common to each individual asset.

$$z_t = \begin{pmatrix} z_{1,t} \\ \vdots \\ z_{P_z,t} \end{pmatrix}_{P_z \times 1}$$

which implies that for each $\boldsymbol{x}_{i,t}$, it is a $(P_z \cdot P_c) \times 1$ vector as following

$$\boldsymbol{x}_{i,t} = z_t \otimes c_{i,t}.$$

But in simulation, $z_t$ is assumed to be scalar.

**Remark B.1** *The above discussed construction, as argued in Gu et al. (2019), is economically motivated to mimic factor structure in linear framework. To see this, suppose that*

$$\beta_{i,t} = \theta_1 c_{i,t} \qquad \lambda_t = \theta_2 z_t$$

*where $\theta_1$ is $Q \times P_c$ matrix and $\theta_2$ is $Q \times P_z$ matrix and the factor structure implies that if expected return is exposed to $Q$ factors, then*

$$\mathbb{E}\left[r_{i,t+1}\right] = \beta_{i,t}^\top \lambda_t \tag{B.1}$$

*and this is equivalent to*

$$\mathbb{E}\left[r_{i,t+1}\right] = \beta_{i,t}^{\top}\lambda_t = \left(\theta_1 c_{i,t}\right)^{\top}\left(\theta_2 z_t\right) = \underbrace{c_{i,t}^{\top}}_{1\times P_c}\underbrace{\left(\theta_1^{\top}\theta_2\right)}_{P_c\times P_z}\underbrace{z_t}_{P_z\times 1} = \left(z_t^{\top}\otimes c_{i,t}^{\top}\right)\mathrm{vec}\left(\theta_1^{\top}\theta_2\right) = \boldsymbol{x}_{i,t}^{\top}\theta$$

*where*

$$\theta = \mathrm{vec}\left(\theta_1^{\top}\theta_2\right)$$

To capture the fat-tail distribution property usually encountered in financial market, $\epsilon_{i,t+1}$ is sampled from t-distribution

$$\epsilon_{i,t+1} \sim \mathrm{t}_5\left(0, 0.05^2\right)$$

each characteristic scalar is simulated as following

$$c_{ij,t} = \frac{2}{N+1}\mathrm{rank}\left(\tilde{c}_{ij,t}\right) - 1 \quad \tilde{c}_{ij,t} = \rho_j \tilde{c}_{ij,t-1} + \varepsilon_{ij,t}$$

and

$$\rho_j \sim \mathrm{U}[0.9, 1] \text{ and } \varepsilon_{ij,t} \sim \mathcal{N}\left(0, 1\right).$$

and the univariate time series is simulated from AR(1) process with $\rho = 0.95$

$$z_t = \rho z_{t-1} + u_t$$

and

$$u_t \sim \mathcal{N}\left(0, 1-\rho^2\right).$$

Two cases for the functional form of $f(\cdot)$ is going to be considered as suggested in Gu et al. (2019)

(a) $f(\boldsymbol{x}_{i,t}) = \left(c_{i1,t}, c_{i2,t}, c_{i3,t}\times z_t\right)$ where $\theta_0 = (0.02, 0.02, 0.02)^{\top}$

(b) $f(\boldsymbol{x}_{i,t}) = \left(c_{i1,t}^2, c_{i1,t}\times c_{i2,t}, \mathrm{sgn}(c_{i3,t}\times z_t)\right)\theta_0$ where $\theta_0 = (0.04, 0.03, 0.012)^{\top}$

where (a) is basically a linear setting where 3 covariates may generate impact on $f(\cdot)$; while (b) involves non-linear setting where the interaction between $c_{i1,t}\times c_{i2,t}$, square term $c_{i1,t}^2$ and the discrete sign variables $\mathrm{sgn}(c_{i3,t}\times z_t)$ as well play the role. To keep the consistency as in our empirical implementation, we split the complete simulated data via ratio 2/3, i.e. 2/3 of the complete data is used as for training; while the remained 1/3 data is used for calculating out-of-sample $R$ square.

And we just find that in this simulation, for linear case, performance of BART in terms of out-of-sample $R$ square is not that good as simple OLS; while for nonlinear case, performance of BART in terms of out-of-sample $R$ square is better than simple OLS. Result is listed as following:

# C Data Construction for Global Market

Firstly, I need to extract daily price data, the main procedure is as following

| Model | (a) | | (b) | |
|---|---|---|---|---|
| Number of Covariates | $P_c = 50$ | $P_c = 100$ | $P_c = 50$ | $P_c = 100$ |
| $R^2_{OS}$ of BART | 1.90 | 2.24 | 6.46 | 6.28 |
| $R^2_{OS}$ of OLS | 3.84 | 3.45 | 3.91 | 3.21 |
| BART/OLS | 0.49 | 0.65 | **1.65** | **1.96** |

```
/* Collecting Price Data for Global Market from Compustat */
proc sql;
create table g_prc_data
as select gvkey,datadate,prccd,loc
from compd.g_secd
where datadate >= '01JAN1988'd;
quit;

data g_prc_data;
set g_prc_data;
year = year(datadate);
month = month(datadate);
run;

/*
proc sort data=g_prc_data;
by gvkey year month;
run;
*/

proc sort data=g_prc_data;
by gvkey datadate;
run;

/*
proc means data=g_prc_data noprint;
by gvkey year;
var prccd;
output out=g_prc_avg;
run;
*/
```

```
proc sql;
create table g_prc_avg as
select gvkey, year, mean(prccd) as prc_yr
from g_prc_data
group by gvkey, year;
quit;

proc sql;
create table funda_a as
select a.*, b.prc_yr
from funda_a a left join g_prc_avg b
on a.gvkey=b.gvkey and a.year=b.year;
quit;

data funda_a;
set funda_a;
mve_f = abs(prc_yr)*cshoi;
mve = log(mve_f);
run;
```

Hence, we can construct price of individual stock at monthly frequency so as to construct market value `mve_f`. Where `cshoi` refers to the common shares outstanding available from `compg.g_funda`. For more details, see my uploaded `SAS` code.

Simply by ignoring dividends, monthly return for global market can be constructed from daily price data as following

$$R_{t+1} = \frac{P_{t+1}}{P_t} = \frac{P_{t+1} - P_t + P_t}{P_t} = 1 + \frac{P_{t+1} - P_t}{P_t} = 1 + r_{t+1} \tag{C.1}$$

where $t$ refers to the index for daily frequency. Suppose that for each month, there are $T$ trading days with $T$ denoting the last day in that month, then by noting that

$$\frac{\cancel{P_2}}{P_1} \times \frac{P_3}{\cancel{P_2}} \times \cdots \times \frac{\cancel{P_{T-1}}}{P_{T-2}} \times \frac{P_T}{\cancel{P_{T-1}}} = \frac{P_T}{P_1}$$

which implies that return of holding a specific stock in that month can be simply calculated as

$$R_\tau = \frac{P_T}{P_1} \quad \text{and} \quad r_\tau = R_\tau - 1$$

where the subscript $\tau$ simply represents the index for monthly frequency. Where the daily price data is available through `Compustat` with the library as `compd.g_secd` and variable name as `PRCCD`.