Journal of Data Science 21 (4), 638–657
October 2023

DOI: 10.6339/22-JDS1067
Statistical Data Science

EVIboost for the Estimation of Extreme Value Index Under Heterogeneous Extremes

Jiaxi Wang[1], Yanxi Hou[1,*], Xingchi Li[2], and Tiandong Wang[3]

[1]*School of Data Science, Fudan University, China*
[2]*Department of Statistics, Texas A&M University, USA*
[3]*Shanghai Center for Mathematical Sciences, Fudan University, China*

Abstract

Modeling heterogeneity on heavy-tailed distributions under a regression framework is challenging, yet classical statistical methodologies usually place conditions on the distribution models to facilitate the learning procedure. However, these conditions will likely overlook the complex dependence structure between the heaviness of tails and the covariates. Moreover, data sparsity on tail regions makes the inference method less stable, leading to biased estimates for extreme-related quantities. This paper proposes a gradient boosting algorithm to estimate a functional extreme value index with heterogeneous extremes. Our proposed algorithm is a data-driven procedure capturing complex and dynamic structures in tail distributions. We also conduct extensive simulation studies to show the prediction accuracy of the proposed algorithm. In addition, we apply our method to a real-world data set to illustrate the state-dependent and time-varying properties of heavy-tail phenomena in the financial industry.

Keywords *gradient boosting; heterogeneous extremes; Pareto model; tail estimation; tree-based method*

1 Introduction

Estimation of the *extreme value index* is an important problem in extreme value theory. Suppose the univariate observations $\{y_i\}_{i=1}^n$ are generated from an unknown distribution F. We say that F lies in the maximum domain of attraction $\mathcal{D}_\gamma$ with an extreme value index $\gamma \in \mathbb{R}$ (write $F \in \mathbb{D}_\gamma$), if there exist sequences a_n and $b_n > 0$, and a nondegenerate distribution G_γ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\max_{1\leqslant i\leqslant n} y_i - a_n}{b_n} \leqslant y\right) = G_\gamma(y), \quad 1+\gamma y > 0, \tag{1}$$

for every continuity point y of G_γ, and $G_\gamma(y) = \exp\left(-(1+\gamma y)^{-1/\gamma}\right)$. Here the extreme value index, γ, represents the heaviness of the distributional tail, $1-F$, and is a crucial feature when learning the heavy-tail phenomena. In terms of model fitting, one can extend the one-parameter family G_γ in (1) to a three-parameter one $G_{\mu,\sigma,\gamma}$ by adding a location parameter μ and a scale parameter σ such that

$$G_{\mu,\sigma,\gamma}(y) := G_\gamma\left(\frac{y-\mu}{\sigma}\right), \quad 1+\gamma y > 0.$$

The limit can still be achieved by replacing a_n and b_n in (1) with $a_n' = a_n - \mu b_n/\sigma$ and $b_n' = b_n/\sigma$.

*Corresponding author. Email: yxhou@fudan.edu.cn.

© 2023 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.
Received June 18, 2022; Accepted September 16, 2022

The parametric limit in (1) serves as a good approximation for data in tail regions in many studies. For example, in Dekkers and De Haan (1989), high quantiles of a distribution are estimated based on the estimators of the extreme value index and intermediate quantiles, where the extrapolation technique is applied. In Gençay et al. (2003), the performance of the generalized Pareto distribution (GPD) model is compared with other well-known methods for the estimation of value-at-risk at high-risk levels, showing that the GPD model is more robust for the estimation of high quantiles. In their studies, the estimation of the *tail index*, which is defined as the reciprocal of extreme value index, is the key to fitting a GPD model.

Classical methods for univariate models are established on the assumption that univariate observations are independent and identically distributed (i.i.d). This assumption leads to a common, homogeneous extreme value index, γ, of the tail distributions for extreme values.

The value of γ can be classified into three types of tail behaviors. In general, a positive γ indicates a heavy right tail of the underlying distribution while a negative γ indicates the existence of an endpoint on the right tail; when γ equals zero, the tail behavior is even more complicated. Therefore, it is necessary to take the range of the extreme value index into account before the estimation. In the literature, one famous estimator is the Hill estimator Hill (1975) for $\gamma > 0$. Other estimators include the Pickands estimator for general $\gamma \in \mathbb{R}$, the maximum likelihood estimator for $\gamma > -1/2$; see Chapter 3 of De Haan and Ferreira (2006) for detailed discussion.

Researchers have made some great efforts to generalize the i.i.d. assumption of the underlying distributions in several ways. One is the so-called *heteroscedastic extremes*, where the underlying distributions for the observations are independent but not identical. However, a constant extreme value index is usually assumed for a benchmark distribution, and a tail equivalence condition is used between the distributions of observations and the benchmark. The tail equivalence condition implies that all the distributions share a common tail heaviness as the benchmark distribution. Under heteroscedastic extremes, Einmahl et al. (2016) develops a variant for the classical Hill estimator and shows that its asymptotic properties are similar to the traditional Hill estimator. In addition, Xu et al. (2020) further develops the regression framework for heteroscedastic extremes and applies it to the prediction of conditional expectiles with extreme levels.

Another approach is called *heterogeneous extremes*, where the underlying distributions are still independent and not identical, but the tail equivalence condition is released. Instead, each distribution of observations may have its extreme value index, which implies the tail heaviness is heterogeneous across all observations. Since the number of extreme value indices equals the sample size, the estimation problems become much more complicated than heteroscedastic extremes. However, it is possible to extend the extreme value index $\gamma(\mathbf{x})$ as a function of the predictor $\mathbf{x}$ under a regression framework. In the literature, some studies try to assume some explicit or implicit forms of $\gamma(\mathbf{x})$, and parametric or nonparametric methods can then be applied under the regression framework. For example, Wang and Tsai (2009) assumes a log-linear form of a positive tail index regression (TIR) model and applies the maximum likelihood method to estimate the coefficients. Moreover, Zhao et al. (2018) introduces the AcF model with dependent and non-identical extreme value indices, and applies the model to quantify the financial risks. In this paper, we focus on modeling positive $\gamma(\mathbf{x})$ for heavy-tailed distributions. In the statistical literature, the TIR model represents a typical parametric model for the extreme value index, which serves as a benchmark for comparison purposes in our study.

This paper aims to provide a gradient tree-boosted algorithm for the nonparametric estimation of $\gamma(\mathbf{x})$ based on heterogeneous extremes, which we refer to as EVIboost in the rest of

the paper. Gradient boosting is one of the most influential machine learning algorithms; see the influential works Friedman (2001) and Zhang and Yu (2005). Our proposed EVIboost algorithm is motivated by the nonparametric regression and classification problems in both fields of statistics and machine learning. Its advantages are two-fold. On one hand, EVIboost is a data-driven approach, and there is no assumption on the parametric form of $\gamma(\mathbf{x})$, which is more realistic for many practical problems. In our simulation studies, we focus on the comparison between the EVIboost and TIR model in Wang and Tsai (2009) to show the superior performance of the proposed method over statistical models. On the other hand, the EVIboost is a robust learning algorithm for the tail region of data distributions. It is well known that the prediction of $\gamma(\mathbf{x})$ is not reliable due to the data sparsity on the tail region, and many statistical approaches usually focus on the center region of the data based on central limit theorem. The EVIboost receives little influence of the bias on the tail region than statistical methods, making it practical to estimate a functional extreme value index.

We organize our paper as follows. Section 2 presents the proposed EVIboost algorithm, and in particular, Section 2.1 combines the regular variation condition in extreme value theory with the gradient boosting algorithm, which produces the fundamentals of our EVIboost algorithm. The rest of Section 2 further describes several issues on the selection of threshold, choice of tuning parameters as well as model interpretation. In Section 3, we conduct a series of simulation studies and compare the performance of EVIboost with the maximum likelihood method as employed in the TIR model. Section 4 is a real-world data analysis that applies our EVIboost algorithm to heavy-tailed financial data.

2 Methodology

2.1 EVIboost for Heterogeneous Extreme Value Index

Suppose the observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ are independent copies of $(Y, \mathbf{X}) \in \mathbb{R} \times \mathbb{R}^p$. Let $F_Y(y|\mathbf{x}) = \mathbb{P}(Y \leqslant y|\mathbf{X} = \mathbf{x})$ be the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$. We further assume that the conditional distribution are in the maximum domain of attraction $\mathcal{D}_{\gamma(\mathbf{x})}$ with $\gamma(\mathbf{x}) > 0$ for all $\mathbf{x}$, where $\gamma(\mathbf{x})$ is the extreme value index depending on the predictor $\mathbf{x}$. Thus, $F_Y(\cdot|\mathbf{x}) \in \mathcal{D}_{\gamma(\mathbf{x})}$ implies that there exists a slowly varying function $L(y|\mathbf{x})$ satisfying $L(y|\mathbf{x}) \to \infty$ and $L(yt|\mathbf{x})/L(y|\mathbf{x}) \to 1$ as $t \to \infty$ for any $y > 0$ and $\mathbf{x}$, such that as $y \to \infty$,

$$1 - F_Y(y|\mathbf{x}) = y^{-1/\gamma(\mathbf{x})} L(y|\mathbf{x})(1 + o(1)). \tag{2}$$

Here we follow the approach of Hall (1982) and assume that

$$L(y|\mathbf{x}) = c_0(\mathbf{x}) + c_1(\mathbf{x}) y^{-\beta(\mathbf{x})} + O(y^{-\beta(\mathbf{x})}),$$

where $c_0(\mathbf{x})$, $c_1(\mathbf{x})$ and $\beta(\mathbf{x})$ are functions of $\mathbf{x}$ with $c_0(\mathbf{x})$, $\beta(\mathbf{x}) > 0$. Furthermore, denote $L_1(y|\mathbf{x}) = \partial L(y|\mathbf{x})/\partial y$ which converges to 0 as $y \to \infty$. The tail conditional probability function can be derived as

$$f_Y(y|\mathbf{x}) = \left(\frac{1}{\gamma(\mathbf{x})} y^{-(1/\gamma(\mathbf{x})+1)} L(y|\mathbf{x}) - y^{-1/\gamma(\mathbf{x})} L_1(y|\mathbf{x}) \right) (1 + o(1)).$$

Based on the properties of $L(\cdot|\mathbf{x})$ and $L_1(\cdot|\mathbf{x})$, we have that for any given $\mathbf{x}$,

$$f_Y(y|\mathbf{x})\gamma(\mathbf{x}) y^{(1/\gamma(\mathbf{x})+1)} \to c_0(\mathbf{x}), \quad y \to \infty,$$

which implies that the tail conditional probability density function can be well approximated by $\frac{c_0(\mathbf{x})}{\gamma(\mathbf{x})} y^{-(1/\gamma(\mathbf{x})+1)}$. However, to implement statistical approaches, it is necessary to applying a sufficiently large threshold u_n to select those data with $y_i > u_n$. In this paper, we define a loss function by considering a transformed negative log-likelihood function and omitting some constants which are not related to $\gamma(\mathbf{x})$,

$$\mathcal{L}_n(\gamma(\cdot)|u_n) = \sum_{i=1}^{n} \Psi(y_i, \gamma(\mathbf{x}_i)|u_n) = \sum_{i=1}^{n} \left(\frac{\log(y_i/u_n)}{\gamma(\mathbf{x}_i)} + \log(\gamma(\mathbf{x}_i)) \right) I(y_i > u_n), \tag{3}$$

and given u_n, we intend to estimate

$$\gamma_n^*(\cdot) := \gamma_n^*(\cdot|u_n) = \operatorname{argmin}_{\gamma(\cdot)\in\mathcal{F}} \sum_{i=1}^{n} \Psi(y_i, \gamma(\mathbf{x}_i)|u_n), \tag{4}$$

where $\mathcal{F}$ is a class of functions.

For the gradient boosting, a greedy stagewise algorithm described in Friedman (2001) assumes each candidate F to be an ensemble of M base learners,

$$\gamma^{[M]}(\mathbf{x}) = \gamma^{[0]} + \sum_{m=1}^{M} \beta_m h(\mathbf{x}|\boldsymbol{\alpha}_m),$$

where $h(\cdot|\boldsymbol{\alpha}_m)$ is a base learner belonging to a class of simple functions with parameters $\boldsymbol{\alpha}_m$, and $\gamma^{[0]}$ and β_m are constants.

The initial estimate of $\gamma_n^*(\cdot)$ is the Hill estimator without including $\mathbf{x}_i$, which can be obtained by using a constant θ in Ψ:

$$\gamma^{[0]} = \operatorname{argmin}_{\theta} \sum_{i=1}^{n} \Psi(y_i, \theta|u_n) = \frac{1}{k_n} \sum_{i=1}^{n} \log(y_i/u_n) I(y_i > u_n),$$

where $k_n = \sum_{i=1}^{n} I(y_i > u_n)$ is the number of y_i above the threshold u_n. Then, by the gradient boosting algorithm, at the m-th step, the current updating is

$$\gamma^{[m]}(\mathbf{x}) = \gamma^{[m-1]}(\mathbf{x}) + \beta_m h(\mathbf{x}|\boldsymbol{\alpha}_m), \quad m = 1, 2, \ldots, M. \tag{5}$$

The current negative gradient is

$$\tilde{y}_i^{[m]} = -\frac{\partial \Psi(y_i, \gamma(\mathbf{x}_i)|u_n)}{\partial \gamma(\mathbf{x}_i)}\Big|_{\gamma(\mathbf{x}_i)=\gamma^{[m-1]}(\mathbf{x}_i)} = \left(\frac{\log(y_i/u_n) - \gamma^{[m-1]}(\mathbf{x}_i)}{\gamma^{[m-1]}(\mathbf{x}_i)^2} \right) I(y_i > u_n).$$

We use an L-terminal node regression tree as the base learner such that

$$h(\mathbf{x}|\{b_{lm}, R_{lm}\}) = \sum_{l=1}^{L} b_{lm} I(\mathbf{x} \in R_{lm}),$$

where $b_{lm} = \operatorname{ave}_{i:\mathbf{x}_i\in R_{lm}} \tilde{y}_i^{[m]}$ and $\operatorname{ave}_A(\cdot)$ denotes the sample mean on the region A. Then, we split the regions R_{lm} using the squared loss function, and (5) becomes

$$\gamma^{[m]}(\mathbf{x}) = \gamma^{[m-1]}(\mathbf{x}) + \sum_{l=1}^{L} \eta_{lm} I(\mathbf{x} \in R_{lm}), \tag{6}$$

Algorithm 1 The EVIboost algorithm for estimation of $\gamma(\cdot)$.

1: **Initialize** $\gamma^{[0]}$

$$\gamma^{[0]} = \frac{1}{k_n}\sum_{i=1}^{n}\log\left(y_i/u_n\right) I\left(y_i > u_n\right).$$

2: **For** $m = 1, 2, \ldots, M$, **do**

a. Compute the negative gradient $(\tilde{y}_1^{[m]}, \tilde{y}_2^{[m]}, \ldots, \tilde{y}_n^{[m]})^T$,

$$\tilde{y}_i^{[m]} = \left(\frac{\log(y_i/u_n) - \gamma^{[m-1]}(\mathbf{x}_i)}{\gamma^{[m-1]}(\mathbf{x}_i)^2}\right) I\left(y_i > u_n\right), \quad i = 1, 2, \ldots, n.$$

b. Fit $\{(\tilde{y}_i^{[m]}, \mathbf{x}_i)\}_{i=1}^n$ to an L-terminal node regression tree,

$$\{R_{lm}\}_1^L = L\text{-terminal node tree of } \{(\tilde{y}_i^{[m]}, \mathbf{x}_i)\}_1^n.$$

c. Approximate the optimal terminal node predictions η_{lm} of R_{lm},

$$\eta_{lm} = \frac{\sum_{i:\mathbf{x}_i\in R_{lm}} \tilde{y}_i^{[m]}}{\sum_{i:\mathbf{x}_i\in R_{lm}}\left(2\tilde{y}_i^{[m]}/\gamma^{[m]}(\mathbf{x}_i) + 1/\gamma^{[m]}(\mathbf{x}_i)^2 I\left(y_i > u_n\right)\right)}, \quad l = 1, 2, \ldots, L.$$

d. Update $\gamma^{[m]}(\mathbf{x})$ for each R_{lm}, $l = 1, 2, \ldots, L$,

$$\gamma^{[m]}(\mathbf{x}) = \gamma^{[m-1]}(\mathbf{x}) + \nu\eta_{lm} I\left(\mathbf{x} \in R_{lm}\right), \quad l = 1, 2, \ldots, L.$$

3: **End for** Return $\gamma^{[M]}(\mathbf{x})$ as the final estimate.

with the coefficient $\eta_{lm} = \beta_m b_{lm}$. Thus, the optimal coefficients are the solution

$$\eta_{lm} = \underset{\eta}{\operatorname{argmin}} \sum_{i:\mathbf{x}_i\in R_{lm}} \Psi(y_i, \gamma^{[m-1]}(\mathbf{x}_i) + \eta), \quad l = 1, 2, \ldots, L,$$

which has no explicit solution. Therefore, a single Newton-Raphson step is applied, as described in Friedman (2001). This leads to the following result

$$\eta_{lm} = \frac{\sum_{i:\mathbf{x}_i\in R_{lm}} \tilde{y}_i^{[m]}}{\sum_{i:\mathbf{x}_i\in R_{lm}}\left(2\tilde{y}_i^{[m]}/\gamma^{[m]}(\mathbf{x}_i) + 1/\gamma^{[m]}(\mathbf{x}_i)^2 I\left(y_i > u_n\right)\right)}.$$

By adding a shrinkage factor $\nu \in (0, 1]$ to (6) and update the current estimate at each region R_{lm}, we have that

$$\gamma^{[m]}(\mathbf{x}) = \gamma^{[m-1]}(\mathbf{x}) + \nu\eta_{lm} I\left(\mathbf{x} \in R_{lm}\right), \quad l = 1, 2, \ldots, L,$$

where ν is the shrinkage factor which controls the rate of learning. Thus, after M iterations, it turns out the estimator $\gamma^{[M]}(\cdot)$ of $\gamma_n^*(\cdot)$ for (4). The EVIboost algorithm is summarized in Algorithm 1.

The interval estimation for the proposed EVIboost estimator remains as an open and challenging problem due to the lack of observations in the tail region. One possible approach is to

employ bootstrap methods (see for instance DiCiccio and Efron (1996) for an introduction to bootstrap). Other possiblities may be to rely on tree-based inference methods such as Meinshausen (2006), Wager et al. (2014), and Mentch and Hooker (2016). We leave detailed analyses for future work.

2.2 Choice of Tuning Parameters

To implement Algorithm 1, we need to choose the threshold u_n as well as three critical tuning parameters in advance: (1) the learning rate, ν, (2) the number of trees, M, and (3) the number of terminal nodes, L. Here ν and M together control the length and the total number of steps in the gradient boost optimization process, and L specifies the complexity of an individual regression tree. For instance, $L = 2$ indicates only one splitting variable in the tree; thus, the tree models the main effects of predictors.

For a given threshold u_n, we first select several discrete values for ν and L, then with values of ν and L chosen, we apply a cross-validation (CV) method to tune the number of trees, M, aiming to minimize the loss function $\mathcal{L}_n(\gamma(\cdot)|u_n)$ on the validation set. Specifically, we adopt a K-fold CV approach and denote $\gamma_{(-j)}^{[M;\nu,L]}(\mathbf{x})$ as the estimator of $\gamma(\mathbf{x})$ with the j-th fold as the validation dataset and the other $K-1$ folds as training dataset. This gives the validation loss as

$$\mathrm{CV}_j(M, \nu, L) := \mathcal{L}_n\left(\gamma_{(-j)}^{[M;\nu,L]}(\mathbf{x})|u_n\right).$$

Given (ν, L), we select the optimal M by minimizing the validation loss, i.e.

$$\widehat{M}_{\nu,L} := \mathrm{argmin}_M \sum_{j=1}^{K} \mathrm{CV}_j(M, \nu, L).$$

Hence, the optimal choice of (M, ν, L) becomes,

$$(\widehat{\nu}, \widehat{L}) := \mathrm{argmin}_{\nu,L} \sum_{j=1}^{K} \mathrm{CV}_j(\widehat{M}_{\nu,L}, \nu, L), \qquad \text{and} \qquad \widehat{M} := M_{\widehat{\nu},\widehat{L}}.$$

2.3 Selection of the Threshold

Another important parameter to determine is the threshold, u_n, which may largely affect the asymptotic distribution (2).

Here we choose the threshold u_n so that it controls the essential sample size on tail region. Since (2) implies that for any given $\mathbf{x}$,

$$\lim_{t\to\infty} \frac{1 - F_Y(ty|\mathbf{x})}{1 - F_Y(t|\mathbf{x})} = y^{-1/\gamma(\mathbf{x})}, \quad y > 0,$$

we define $\tilde{U}_i := (y_i/u_n)^{-1/\gamma^{[\widehat{M};\hat{\nu},\widehat{L}]}(\mathbf{x})}$ for given values of $\mathbf{x}$, and let F_n be the empirical distribution of $\tilde{U}_i$ for which $y_i > u_n$. The optimal choice of $(\widehat{M}, \hat{\nu}, \widehat{L})$ is determined under the given u_n. Following methods in Clauset et al. (2009) and Wang and Tsai (2009), we consider three different

discrepancy measures:

$$D_1(u_n) := \frac{1}{k_n} \sum_{y_i > u_n} (\tilde{U}_i - F_n(\tilde{U}_i))^2,$$
$$D_2(u_n) := \sup_{y_i > u_n} |\tilde{U}_i - F_n(\tilde{U}_i)|,$$
$$D_3(u_n) := \frac{1}{k_n} \sum_{y_i > u_n} \frac{(\tilde{U}_i - F_n(\tilde{U}_i))^2}{\tilde{U}_i(1 - \tilde{U}_i)},$$

where $D_1(\cdot)$ is identical to the setting in Wang and Tsai (2009), and $D_2(\cdot)$, $D_3(\cdot)$ correspond to the Kolmogorov-Smirnov and Anderson-Darling distances, respectively. Then u_n can be determined by minimizing a selected discrepancy measure.

A feasible approach to select u_n is by determining the tail sample fraction $q = k_n/n$, which we can consider as a turning parameter, describing the proportion of y_i exceeding u_n used in the algorithm. Then, the selection of the threshold is equivalent to the determination of the tail fraction. Given the response $y_1, \ldots, y_n$, we consider a finite sequence of tail fractions $\{q_s\}_{s=1}^S$ equally spaced on the interval $[0, 1]$ (e.g. $\{0.01, 0.02, \ldots, 0.99\}$). For each q_s, let u_{sn} be the corresponding $(1 - q_s)$-th sample quantile of $y_1, \ldots, y_n$. Then the optimal threshold is

$$u_n^* = \underset{u_n \in \{u_{sn}\}_{s=1}^S}{\operatorname{argmin}} \; D(u_{sn}),$$

and the final estimate of $\gamma(\cdot)$ is $\gamma^{[\widehat{M};\hat{\nu},\widehat{L}]}(\cdot)$ given u_n^*.

2.4 Model Interpretation

The tree-based models are more interpretive than other nonparametric machine learning algorithms such as neural networks and support vector machines. In this subsection, we briefly discuss two tools, the (feature) importance measure and the partial dependence plot.

In many real applications, one usually wants to identify the importance of features/covariates on the predictions of the interesting objects. Here we apply an impurity-based method proposed by Breiman et al. (1984) to evaluate the importance of each individual feature $x_1, \ldots, x_p$. The definition starts within a single tree T_m. Suppose x_i is the feature of interest, then the importance of x_i on T_m is given by

$$I_m(x_i) = \sum_{j=1}^{J} \Delta\delta_j \cdot \mathbb{I}(x_i, j), \quad i = 1, 2, \ldots, p,$$

where the sum is over all J non-terminal nodes of T_m, and $\Delta\delta_j$ is the reduction of the squared error caused by node j (recall that during the construction of a tree, the algorithm greedily searches for a split that can maximize the reduction in MSE). The indicator $\mathbb{I}(x_i, j)$ equals one if the node j uses x_i to split and zero otherwise. Since the EVIboost is an ensemble of M trees, we take the average of $I_1(x_i), \ldots, I_M(x_i)$ as the importance measure of x_i, i.e.

$$I(x_i) = \frac{1}{M} \sum_{m=1}^{M} I_m(x_i).$$

Different from the log-linear form as in the TIR model Wang and Tsai (2009), the importance measure $I(x_i)$ considers not only the main effects but also interactions among variables. However, $I(x_i)$ may be biased so that a feature irrelevant to the response may still have non-zero importance if it is chosen as the splitting variable by any nodes; see White and Liu (1994) and Sandri and Zuccolotto (2008). Therefore, we follow the methods in Yang et al. (2018) to derive a modified importance measure. Let $\mathbf{x}$ be the $n \times p$ design matrix, repeat steps (1) to (3) for r from 1 to R.

1. Generate an $n \times p$ matrix $\mathbf{z}^{(r)}$ by randomly shuffling the n rows of $\mathbf{x}$, while the order of columns are unchanged. Bind $\mathbf{x}$ and $\mathbf{z}^{(r)}$ by columns, then denote the $n \times 2p$ matrix $[\mathbf{x}, \mathbf{z}^{(r)}]$ as $\mathbf{x}^{(r)}$.
2. Implement the EVIboost model using $\{y, \mathbf{x}^{(r)}\}$.
3. Compute the importance measures $I^{(r)}(x_i)$ for x_i and $I^{(r)}(z_i^{(r)})$ for $z_i^{(r)}$ respectively, where $z_i^{(r)}$ is the ith column of $\mathbf{z}^{(r)}$.

Since the pseudo-predictor $z_i^{(r)}$ is shuffled from x_i, $z_i^{(r)}$ has the same number of possible splits as x_i, and is equivalently possible of being selected by tree nodes. Hence, we take $I^{(r)}(z_i^{(r)})$ as a bias approximation for $I(x_i)$, and the modified importance measure for x_i is given by

$$I^{\star}(x_i) = \frac{1}{R}\left(\sum_{r=1}^{R} I^{(r)}(x_i) - \sum_{r=1}^{R} I^{(r)}(z_i^{(r)})\right).$$

One limitation of the importance measure is that it cannot demonstrate how the estimated function $\gamma^{[M;\nu,L]}(\cdot)$ varies along with the features. To solve this problem, we employ the partial dependence plots introduced in Friedman (2001). We divide the predictors $\mathbf{X}$ and its observation $\mathbf{x}_i$ into two non-overlapping subsets $\mathbf{X}_s$ and $\mathbf{X}_{-s}$, and $\mathbf{x}_{i,s}$ and $\mathbf{x}_{i,-s}$, where s is a nonempty index subset of $\{1, 2, \ldots, p\}$. The partial dependence of $\mathbf{X}_s$ at $\mathbf{x}_s$ is then given by

$$\bar{\gamma}_s(\mathbf{x}_s) = \frac{1}{n}\sum_{i=1}^{n} \gamma^{[M;\nu,L]}(\mathbf{X}_s = \mathbf{x}_s, \mathbf{X}_{-s} = \mathbf{x}_{i,-s}). \tag{7}$$

Note that when $\mathbf{X}_s$ are independent of $\mathbf{X}_{-s}$, $\bar{\gamma}_s(\cdot)$ will serve as an estimator of the conditional expectation $E(\hat{\gamma}(\mathbf{x}_s)|X_s = \mathbf{x}_s)$. To fully depict the marginal effects of $\mathbf{X}_s$ with respect to $\gamma^{[M]}(\cdot)$, we then plot $\bar{\gamma}_s(\cdot)$ versus domain of $\mathbf{X}$. In the real data analysis, Figure 9 illustrates the applications of the modified importance measure and the one-dimensional partial dependence plot.

3 Simulation Study

3.1 Simulated Models

In this section, we conduct simulation studies to compare the prediction accuracy of our EVIboost algorithm with the TIR model in Wang and Tsai (2009). We follow a similar simulation setup as in Wang and Tsai (2009), but we consider more models, including both log-linear and nonlinear forms. Note that TIR is restricted by a log-linear form of $\gamma(\cdot)$ whereas the proposed EVIboost algorithm possesses more model flexibility by allowing nonparametric functions. We describe the scheme of generating simulated samples and calculating performance metrics as follows.

Step 0: Determine parameters in the simulation.

We set the covariate dimension $p = 10$, the size of both the training and testing samples $n = n^* = 1000$, the number of replications $R = 100$, and the distribution parameters $m = 0.10$

and $C = 1/3$. Also, set the tail fraction $q = 0.1, 0.05, 0.025$ and let $q^\star$ be the optimal fraction given by

$$q^\star := \text{argmin}_q(D_1(u_n) + D_2(u_n) + D_3(u_n)),$$

Note that $D_i(u_n)$, $i = 1, 2, 3$, are the three discrepancy measures, and u_n is the threshold given the tail fraction q.

Step 1: Generate a training sample $(\mathbf{x}_i^{(r)}, y_i^{(r)})$ of size n.

Simulate $\mathbf{Z}_i := (Z_{i,1}, \ldots, Z_{i,p}) \sim N_p$ with zero mean and $\text{Cov}(Z_{i,j}, Z_{i,k}) = \frac{1}{2}^{|j-k|}$, $j, k = 1, \ldots, p$. Denote $x_{i,j}^{(r)} = 2\sqrt{3}(\Phi(Z_{i,j}) - \frac{1}{2})$, where Φ is the CDF of a standard normal random variable, and $x_{i,j}^{(r)}$ is the j-th coordinate of $\mathbf{x}_i^{(r)}$.

Given $\mathbf{x}_i^{(r)}$, we then simulate y_i from

$$F(y|\mathbf{x}_i^{(r)}) = 1 - \frac{(1+m)y}{y^{1/\gamma(\mathbf{x}_i^{(r)})} + my},$$

where $\gamma(\mathbf{x})$ is of one of the following forms:

(1) $\gamma_1(\mathbf{x}) = C\exp\left(-\frac{1}{2}x_1 + \frac{1}{3}x_2 - \frac{1}{3}x_3\right)$;

(2) $\gamma_2(\mathbf{x}) = C\exp\left(\frac{2}{p}\sum_{i=1}^p x_i \times (-1)^i\right)$;

(3) $\gamma_3(\mathbf{x}) = C\exp\left(-\frac{1}{2}x_1^2 + \frac{1}{3}x_2^2 - \frac{1}{3}x_3^2\right)$;

(4) $\gamma_4(\mathbf{x}) = \exp\left(-(x_1+x_2)^2 - (x_2+x_3)^4\right)$;

(5) $\gamma_5(\mathbf{x}) = \exp\left(-\sqrt{x_1 - x_2} - \frac{1}{\sqrt{x_3 - x_4}}\right)$.

Step 2: Given the tail fraction q, set the value of $u_n^{(r)}$ to be the $(1-q)$-th sample quantile of $y_1^{(r)}, y_2^{(r)}, \ldots, y_n^{(r)}$.

Step 3: Estimation of the extreme value index function $\gamma(\cdot)$.

Given the threshold $u_n^{(r)}$, we implement the EVIboost algorithm to obtain $\hat{\gamma}^{[M;\nu,L](r)}(\cdot)$ and apply the maximum likelihood estimation for TIR model in Wang and Tsai (2009) to obtain $\hat{\gamma}_{TIR}(\cdot)$. Note that the tuning parameters M, ν, L of EVIboost are obtained by following a five-fold CV approach as described in Section 2.2.

Step 4: Calculate the mean squared error of predictions on a testing sample.

Generate a testing sample, $(\mathbf{x}_i^{*(r)}, y_i^{*(r)})$ of size n^*, in the same way as in Step 1. Then use the testing sample to evaluate the two estimators in Step 3 by a mean squared error:

$$\hat{\delta}^{(r)} = \frac{1}{n^*}\sum_{i=1}^{n^*}\left(\hat{\gamma}(\mathbf{x}_i^{*(r)}) - \gamma(\mathbf{x}_i^{*(r)})\right)^2,$$

where $\hat{\gamma}(\cdot)$ is either $\hat{\gamma}^{[M;\nu,L](r)}(\cdot)$ or $\hat{\gamma}_{TIR}(\cdot)$, and $\hat{\delta}_{EVI}^{(r)}, \hat{\delta}_{TIR}^{(r)}$ denote $\hat{\delta}^{(r)}$ under the EVIboost and TIR models, respectively.

Step 5: Repeat Steps 1 to 4 for r from 1 to R independently and compare the prediction performance of $\{\hat{\delta}_{EVI}^{(r)}\}_{r=1}^R$ and $\{\hat{\delta}_{TIR}^{(r)}\}_{r=1}^R$.

We compare the prediction performance of the EVIboost and TIR models by sketching the boxplots of $\{\hat{\delta}_{EVI}^{(r)}\}_{r=1}^R$ and $\{\hat{\delta}_{TIR}^{(r)}\}_{r=1}^R$. To evaluate the advantage of our method quantitatively, we compute an efficiency ratio as $\text{med}(\{\hat{\delta}_{EVI}^{(r)}\}_{r=1}^R)/\text{med}(\{\hat{\delta}_{TIR}^{(r)}\}_{r=1}^R)$, where $\text{med}(\cdot)$ means the sample median of a sequence.

In Step 1, each marginal distribution of $\mathbf{x}_i^{(r)}$ is $U[-\sqrt{3}, \sqrt{3}]$ with unit variance, and any x_i, x_j $i \neq j$ is pairwisely correlated.

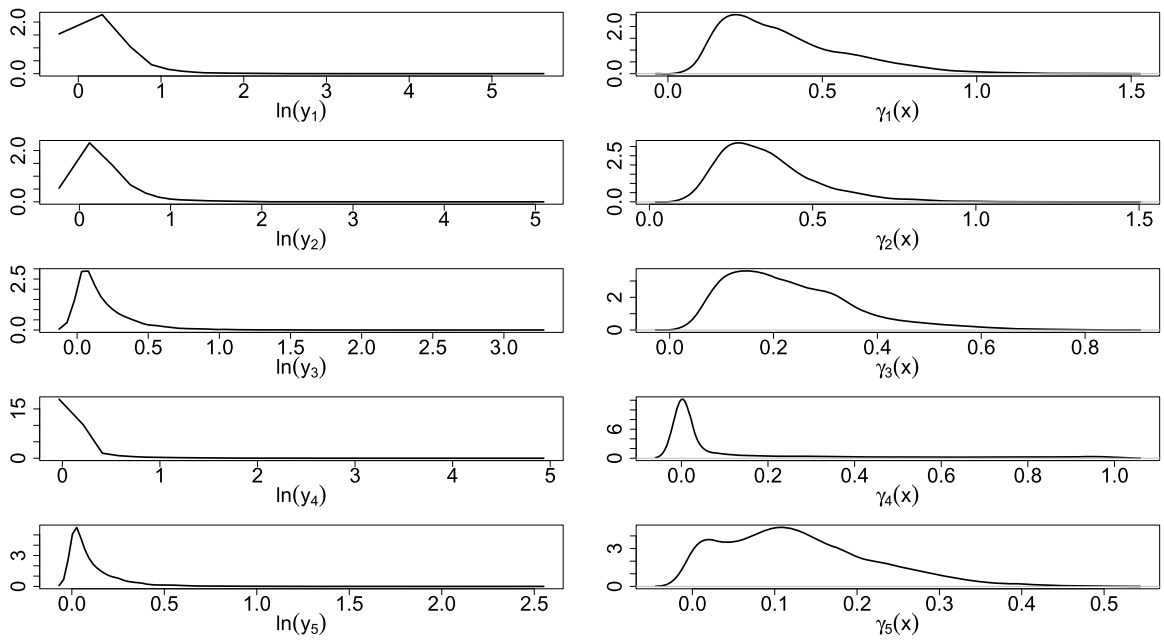


Figure 1: Density plots of simulated $\log(y_i)$ and $\gamma(\mathbf{x}_i)$.

For the choice of $\gamma(\cdot)$, we assume $\gamma(\cdot)$ to be log-linear in Cases (1) and (2), which agrees with assumptions in the TIR model. However, in Case (1), only the first three covariates have impacts on $\gamma(\cdot)$, whereas in Case (2), all covariates are equally influential. Cases (3) to (5) assume nonlinear forms with higher-order interaction terms, under which the TIR model is misspecified. Figure 1 gives the densities of simulated data in all cases. In Step 2, we intend to see the prediction performance of the EVIboost and TIR models given the upper tail fraction and the optimal one chosen by the TIR model.

3.2 Results

Now we discuss our numerical results from the simulation study. Figure 2 presents the boxplots of the mean squared error of predictions in Section 3.1 for both the EVIboost and TIR models across all five cases. When $\gamma(\mathbf{x})$ is log-linear, i.e. in Cases (1) and (2), the prediction errors of EVIboost estimators are higher than those of TIR under low thresholds ($q = 0.1$), but they are equivalent or even lower when $q = 0.05, 0.025$ (see the middle two panels of Figure 2(a)). As explained in Section 3.1, Case (1) corresponds to the simulation model in Wang and Tsai (2009), where the TIR model correctly specifies $\gamma(\cdot)$. For the optimal fraction $q^\star$ chosen by the TIR model, the TIR estimators outperform the EVIboost once $\gamma(\cdot)$ is correctly specified. However, the EVIboost performs much better than TIR when $\gamma(\mathbf{x})$ is no longer log-linear, which shows that the proposed algorithm is a data-driven method. In Cases (3) to (5), the mean squared error of the EVIboost is consistently lower for all chosen fractions q. We also summarize the efficiency ratios at all fractions in Table 1.

Overall, the accuracy of the EVIboost and TIR are both sensitive to the tail fraction q. However, EVIboost has a better performance when q is small, where in contrast, the prediction results from the TIR model show significant biases and variations. When there is no prior knowledge on the relationship between the covariates and the extreme value index, EVIboost

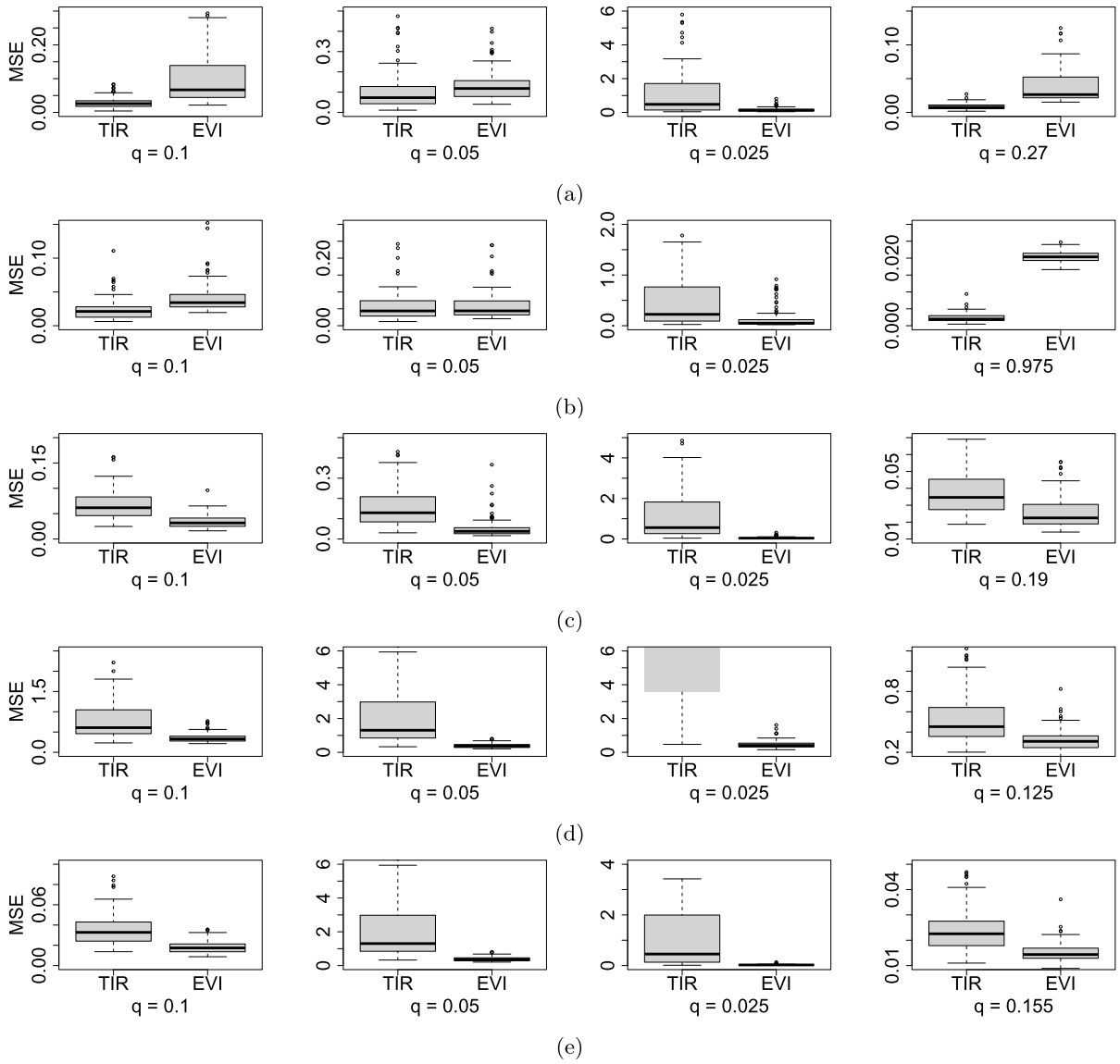


Figure 2: Mean squared error (MSE) of TIR and EVIboost estimators under $R = 100$ replications. Plots (a) to (e) correspond to Cases (1) to (5), respectively.

usually performs better and has less variation for the estimates than the TIR model, showing that EVIboost is more stable for the estimation on tail regions.

To further assess the prediction performance of the EVIboost and TIR models, we compute their mean squared errors when q is set to be uniformly spread on $[0, 1]$. In particular, we consider the results for $m = 15$, which are presented in Figure 3. The TIR model performs better than the EVIboost at most of q in Cases (1) and (2), whereas the EVIboost produces smaller MSE for all chosen q in Cases (3) to (5). These observations are consistent with what we have found in Figure 2 and Table 1.

Another interesting observation is that among the left panels of Figure 3, the MSE's of both EVIboost and TIR are decreasing as q increases. This is related to the parameter m in the setup

Table 1: The median of squared error of EVIboost and TIR.

Cases	Tail Fraction	Median of SE		Efficiency
		EVIboost	TIR	
1	0.100	0.067	0.026	0.389
	0.050	0.118	0.073	0.617
	0.025	0.130	0.479	3.673
	0.27*	0.026	0.008	0.290
2	0.100	0.034	0.021	0.618
	0.050	0.044	0.044	0.995
	0.025	0.050	0.224	4.503
	0.975*	0.020	0.002	0.096
3	0.100	0.032	0.062	1.946
	0.050	0.038	0.129	3.405
	0.025	0.042	0.561	13.501
	0.19*	0.022	0.035	1.540
4	0.100	0.329	0.607	1.847
	0.050	0.365	1.304	3.571
	0.025	0.409	15.904	38.919
	0.125*	0.308	0.453	1.470
5	0.100	0.017	0.033	1.884
	0.050	0.365	1.304	3.571
	0.025	0.025	0.455	18.146
	0.155*	0.014	0.023	1.570

NOTE: * represents the optimal tail fraction $q^\star$ selected by the TIR model.

of $F(y|\mathbf{x}_i)$ (cf. Section 3.1), which determines the rate of convergence. When m is small (e.g. $m = 0.10$), on one hand, $F(y|\mathbf{x})$ converges considerably fast and can be well approximated by its limit (see (2)) even when y is at a low level. Therefore, a high threshold u_n (or equivalently a low tail fraction q) will only lead to decreases in the accuracy of estimation since the size of the effective sample used for estimating $\gamma(\cdot)$ is small. On the other hand, for a low threshold, u_n, the asymptotic distribution in (2) may deviate from $F(y|\mathbf{x})$, thus making the loss function (3) less accurate. However, when u_n is high, the tail sample size is too small to precisely predict $\gamma(\cdot)$. The plots with $m = 15$ illustrate it, and further show that the MSE is no longer monotone along with q, and the minimum is located at the middle of the interval. For instance, the optimal values of q for the EVIboost and TIR models in Case (2) are around 0.45 and 0.15, respectively.

4 Application: Estimate the Tail Indices of Banks

4.1 Data

In this section, we conduct a real-world data analysis using the proposed EVIboost algorithm to estimate the dynamics of the extreme value index for heavy-tailed financial data. As men-

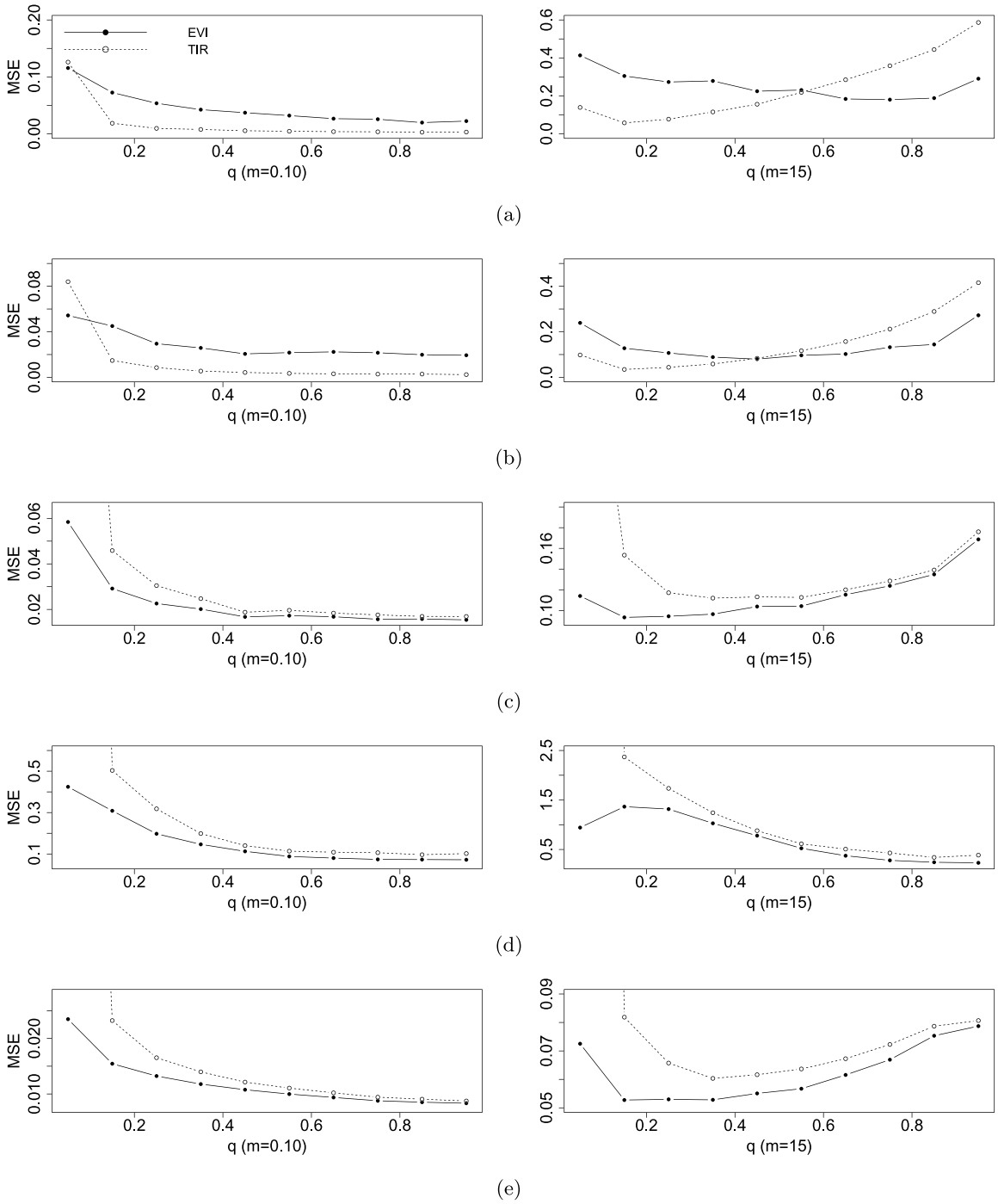


Figure 3: Mean squared error (MSE) of the EVIboost and TIR estimators in cases 1 to 5 with the fraction q ranging from 0.05 to 0.95 by step 0.10. Results are averaged over $R = 100$ replications. Apart from the original setup ($m = 0.10$), we implement a larger m here ($m = 15$).

tioned in Section 1, statistical modeling usually assumes a constant extreme value index over time. Although some studies try to make some extensions by assuming parametric functions of the extreme value index, like the TIR model, it is still hard to capture the whole dynamics of the tail heaviness given the predictor values. Therefore, our data analysis aims to measure the convolution of the distributional tail for the financial market return data along with changes in macro-economic status. Our dataset contains weekly market returns for four large commercial banks: Bank of America (BAC), Citigroup (C), JPMorgan Chase (JPM), and Wells Fargo (WFC). The entire period is from the first week of 1971 to the end of June 2013, which has been partly studied in Xu et al. (2020) and Adrian and Brunnermeier (2016). The sample sizes for the four banks are 1771, 1386, 2210, and 2210, respectively, and we consider seven macro-economic variables as predictors/covariates:

1. x_1: The weekly market return of S&P500;
2. x_2: The change in the three-month yield from the Federal Reserve Board's H.15 release;
3. x_3: Equity volatility, which is computed as the 22-day rolling standard deviation of the daily CRSP equity market return;
4. x_4: The change in the credit spread between Baa bonds (rated by Moody's) and the ten-year Treasury rate from the Federal Reserve Board's H.15 release;
5. x_5: The change in the slope of the yield curve, measured by the spread between the composite long-term bond yield and the three-month bill rate;
6. x_6: A short-term TED spread, defined as the difference between the three-month LIBOR rate and the three-month secondary market treasury bill rate. They're obtained from the British Bankers' Association and the Federal Reserve Bank of New York respectively. This term can be a measurement of short-term funding liquidity risk.
7. x_7: The weekly real estate sector return in excess of the market financial sector return (from the real estate companies with SIC code 65–66).

For data pre-processing, we first make a min-max normalization for each of the seven covariates to bound their domain by [0, 1]. That is, we transform all $x_{i,j}$ to

$$x^{\star}_{i,j} = \frac{x_{i,j} - \min_{1\leqslant i\leqslant n} x_{ij}}{\max_{1\leqslant i\leqslant n} x_{ij} - \min_{1\leqslant i\leqslant n} x_{ij}}, \quad i = 1, \ldots, n, \;\; j = 1, \ldots, p.$$

Since both extremely positive and negative returns are of interest, we take the response y to be the absolute values of weekly returns for each bank. Figure 4 is the QQ plots (with respect to the standard normal distribution) of y, from which we can spot heavy-tailed phenomenons for all four banks. Moreover, their sample kurtoses are 81.0, 81.1, 166.8, and 155.0, respectively. Figure 5 plots how the covariates vary during the whole period, from which we will interpret the influences of the seven macro-economic variables based on the extreme value indices of the four banks.

4.2 Model Fitting

We fit an independent model by the EVIboost algorithm for each bank with seven common macro-economic variables. The estimated extreme value indices are plotted in Figure 6. To demonstrate its variation more clearly, we divide the whole period into quarters (three months) and average the estimated values within each quarter. The horizontal dashed lines indicate four Hill estimators, which are time-invariant. Their values are 0.460 (BAC), 0.507 (C), 0.440 (JPM) and 0.358 (WFC), which are estimated under the same thresholds as EVIboost model (see Table 2).

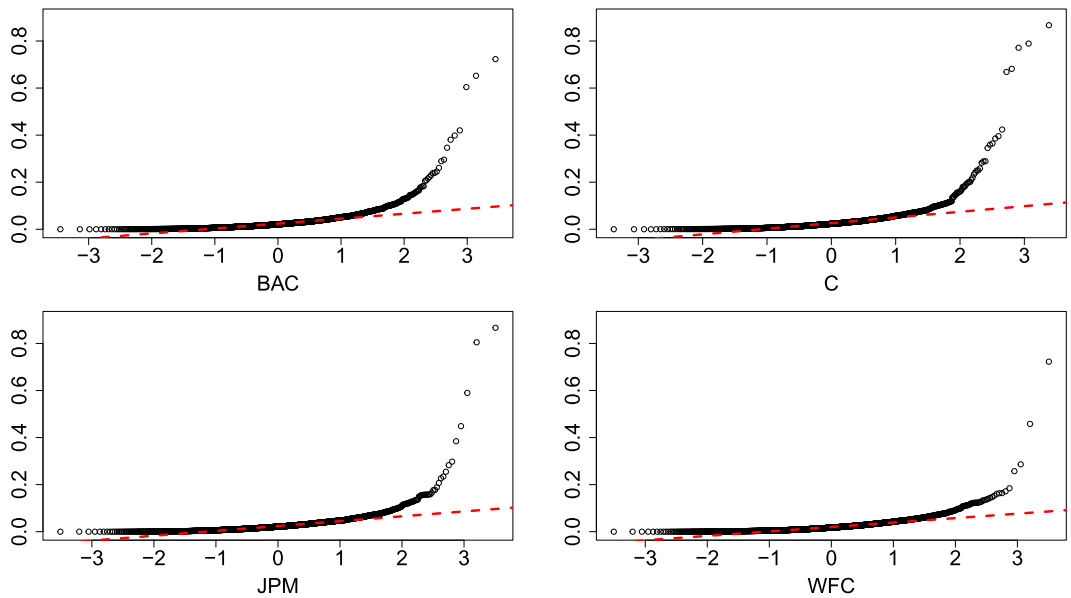


Figure 4: The QQ plots of absolute weekly returns y. The horizontal and vertical axis represent the standard normal and empirical quantiles.

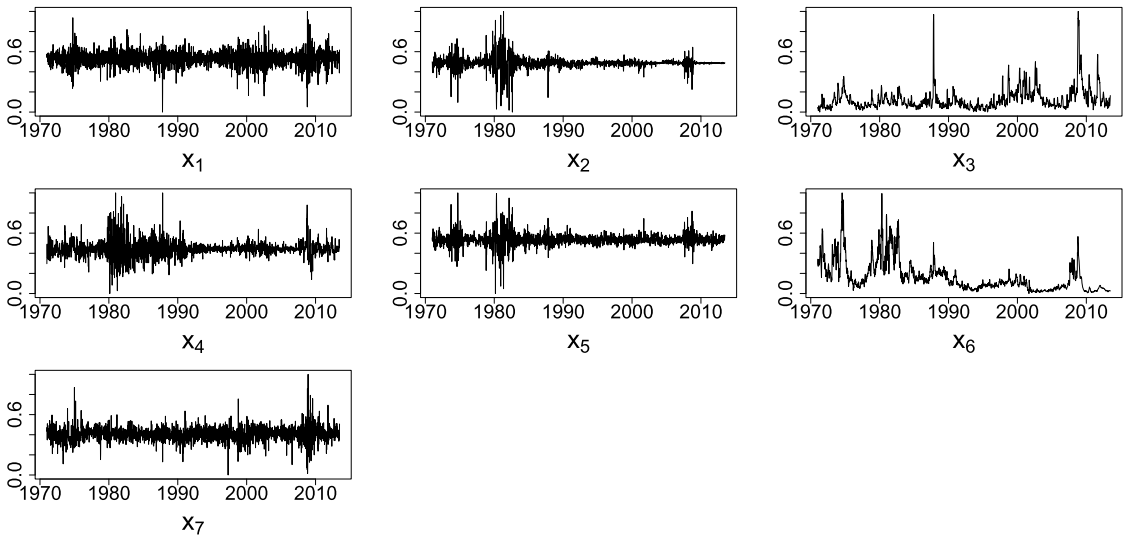


Figure 5: Time series of standardized covariates $x_1 \sim x_7$.

We then check the model fitting by considering the choice of thresholds and tuning parameters. We follow the procedures described in Sections 2.2 and 2.3 to determine the tail sample fraction, q, as well as tuning parameters, M, ν, L, for each model. Taking BAC as an example, the left panel of Figure 7(a) gives the discrepancy measures based on TIR estimators. We next compute D_1, D_2, D_3 for $q \in \{0.005, 0.010, \ldots, 0.995\}$. Overall, we see from Figure 7(a) that curves for the three discrepancy measures show similar shapes, and their optimums are almost identical: D_1 and D_3 both reach the minimum at $q = 0.075$. In contrast, D_2 reaches the minimum at $q = 0.080$. Therefore, we set $q = 0.075$ and $u_n = 0.074$, i.e., the $(1 - q)$th sample quantile of

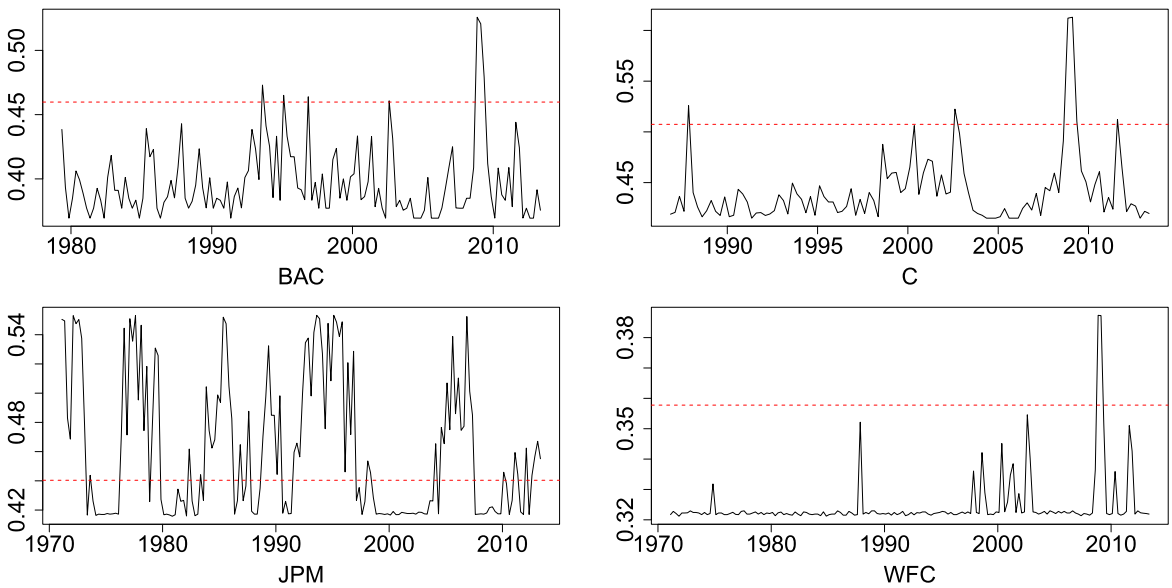


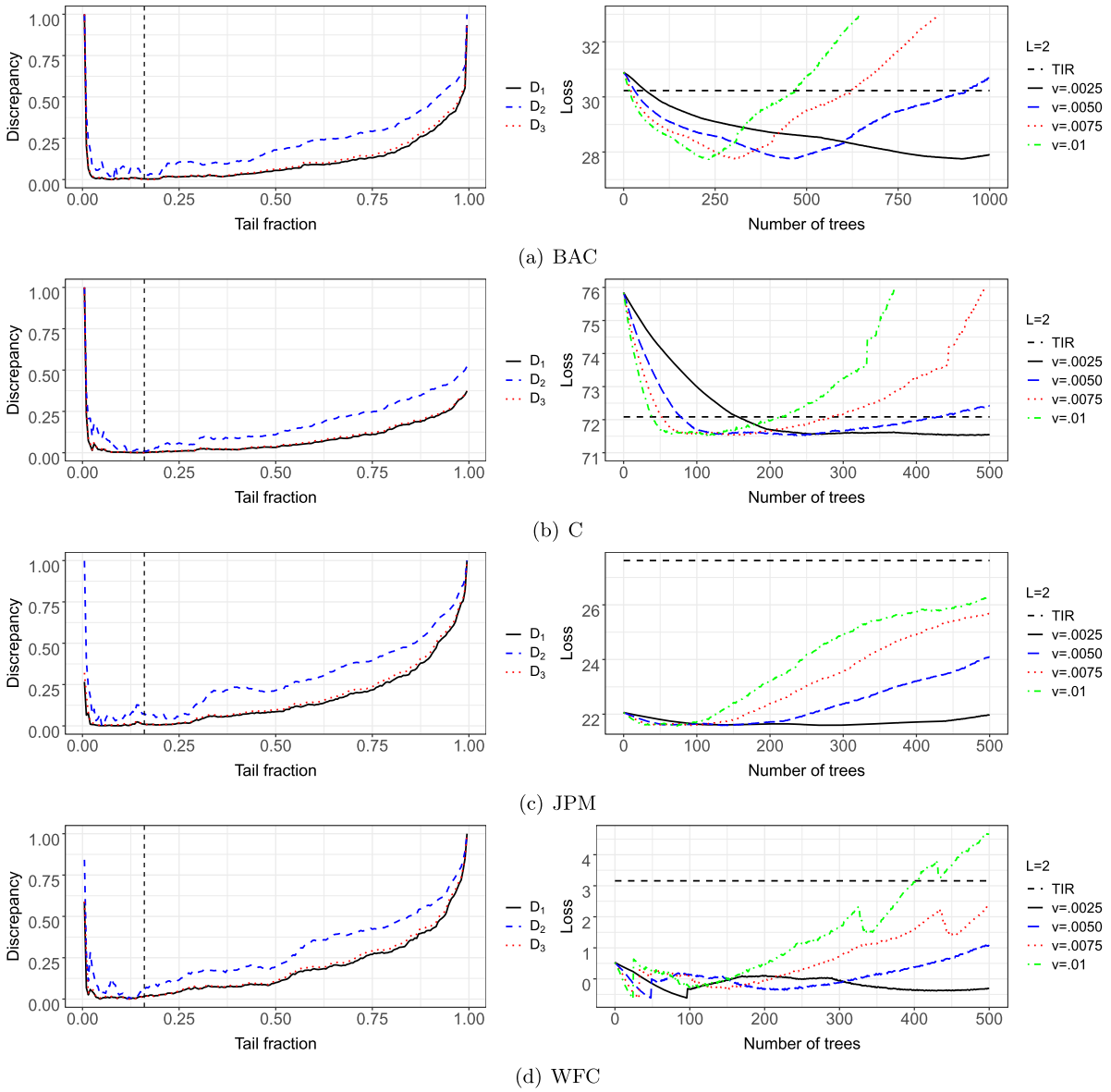
Figure 6: Estimated extreme value index of four banks (1971 to 2013).

Table 2: Tail fractions, thresholds and tuning parameters of models.

Model	Tail fraction	Threshold	Tuning parameters		
			L	ν	M
BAC	0.075	0.074	2	0.005	400
C	0.16	0.057	2	0.0075	160
JPM	0.05	0.077	2	0.0075	95
WFC	0.045	0.071	2	0.01	30

y_i for BAC. The right panel of Figure 7(a) shows the learning curve of five-fold CV on a total of 1771 observations from BAC, where the horizontal line corresponds to the loss given by TIR. The larger L we choose, the faster the loss will converge under shallow trees, i.e., trees with fewer terminal nodes (detailed simulation results are omitted here). According to methods proposed in Section 2.2, we choose $L = 2$, $\nu = 0.005$ and $M = 400$ to minimize the loss. Following the same methodology, we then determine the thresholds and tuning parameters for all other three models, and the chosen parameters are summarized in Table 2.

In addition, since the conditional distribution of $\tilde{U}_i = (y_i/u_n)^{-1/\gamma(\mathbf{x}_i)}$ given $y_i > u_n$ is close to the standard uniform distribution, we draw a QQ plot of $\tilde{U}_i$ for all $y_i > u_n$ to examine whether the EVIboost fits $\gamma(\mathbf{x})$ well (cf. Figure 8). The theoretical quantiles are those from the standard uniform distribution, and the red dashed line is the 45°-line. From Figure 8, we see that the points are closely scattered on the dashed line, indicating that the tail distributions over the thresholds fit well for all four models. Moreover, we implement the Kolmogorov-Smirnov test on $\tilde{U}_i$ for the four models, and the result shows that none of them significantly distinguishes from the standard uniform distribution under the 0.05 level (their p-values are 0.415, 0.988, 0.824 and 0.694, respectively). Therefore, we conclude that the EVIboost algorithm reasonably predicts the extreme value indices, $\gamma(\cdot)$, in all four cases.



(a) BAC

(b) C

(c) JPM

(d) WFC

Figure 7: Selection of thresholds and tuning parameters for the four models.

4.3 Model Interpretation

Figure 6 shows the dynamics of the extreme value index series estimated by the EVIboost given the seven weekly micro-economic statuses, where we use the canonical Hill estimator as the baseline. One interesting finding is that for BAC, C, and WFC, the estimation of their tail indices is below the value given by the Hill estimator (i.e., red dash lines) most of the time. Also, peaks of these three banks occur at the end of 2008, which may correspond to the 2008 US financial crisis. It suggests that the tail distributions of the three banks are not as heavy as the predictions by the Hill estimator, whose large values may be due to extreme losses during the financial crisis. On the other hand, JPM shows a different pattern, and the variation of its extreme value index is higher than the others over the study period.

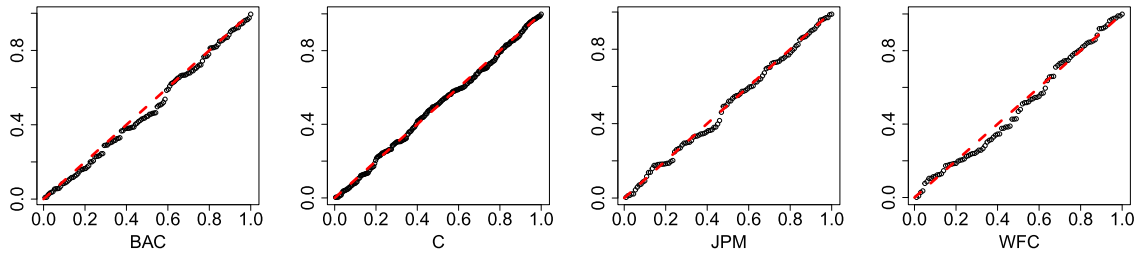


Figure 8: The QQ plots of $\tilde{U}_i$ for all four models. The horizontal and vertical axis represent the theoretical and empirical quantiles, respectively.

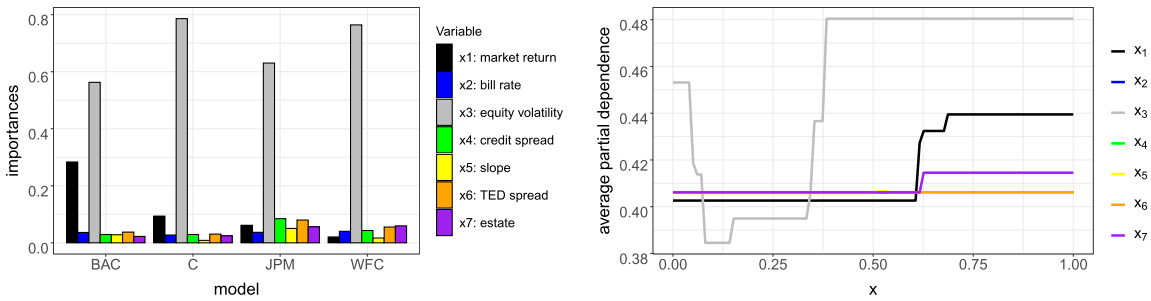


Figure 9: Importance and average partial dependence of the seven macro-economic variables.

Next, we compute the modified importance measure $I^{\star}(\cdot)$ in each model, and the results are shown in the left panel of Figure 9. Variables such as bill rate (x_2), credit spread (x_4), and slope (x_5) are of little importance in all four models, which indicates that they are rarely used in any splits of the regression trees. In contrast, the most important covariate across all four models happens to be equity volatility (x_3), especially for JPM. Market return (x_1) is another important covariate for BAC and C, but not quite for JPM and WFC. Therefore, the left panel of Figure 9 indicates the influential variables in the seven macro-economic statuses of the heterogeneous extremes for each bank.

Now we compute the partial dependence of feature x_i in the K-th model, where $K \in \{\text{BAC, C, JPM, WFC}\}$, following the definition in (7), and denote them as $\bar{\gamma}_i^{(K)}(x)$, $i = 1, \ldots, 7$. Then the average partial dependence of x_i is calculated as the average overall four values of $\bar{\gamma}_i^{(K)}(x)$. From the right panel of Figure 9, we see that the market return (x_1), volatility (x_3), and the estate return (x_7) are all positively associated with the extreme value index (the dependence curve of x_3 goes down at the beginning but there is an overall trend of ascending). Overall, the EVIboost algorithm can estimate the dynamic evolution of the tail heaviness given the macro-economic status and make model interpretations about the importance and dependence of these covariates.

5 Conclusion

This paper proposes an EVIboost algorithm to estimate the heterogeneous extremes of heavy-tailed phenomena. Modeling heterogeneous extremes is challenging in statistical methodologies, and the dynamic structures of the extreme value index are not easy to explain. However, our

EVIboost algorithm can estimate the extreme value index in nonparametric forms by borrowing ideas from gradient-boosted trees. We conduct detailed simulation studies to show that our proposed method outperforms the TIR model when the dynamic structures are unknown. Moreover, the variable importance and partial dependence analysis by boosting algorithms contribute to more substantial interpretations of the dynamic structures of the extreme value index in practice.

Supplementary Material

The following files are included in the supplementary material: (1) Programs for modeling TIR and EVIboost; (2) Code files for simulation study, along with the detailed experiment results; (3) Code and data files for financial data analysis via EVIboost; (4) Simulation results on computational time of the EVIboost model.

Funding

Yanxi Hou's research was partly supported by the National Natural Science Foundation of China Grant 72171055 and the Natural Science Foundation of Shanghai Grant 20ZR1403900.

References

Adrian T, Brunnermeier MK (2016). Covar. *The American Economic Review*, 106: 1705–1741.

Breiman L, Friedman J, Stone C, Olshen R (1984). *Classification and Regression Trees*. CRC Press, Abingdon, United Kingdom.

Clauset A, Shalizi CR, Newman MEJ (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4): 661–703.

De Haan L, Ferreira A (2006). *Extreme Value Theory: An Introduction (Vol. 21)*. Springer, New York.

Dekkers A, De Haan L (1989). On the estimation of the extreme-value index and large quantile estimation. *The Annals of Statistics*, 17(4): 1795–1832.

DiCiccio TJ, Efron B (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3): 189–228.

Einmahl J, De Haan L, Zhou C (2016). Statistics of heteroscedastic extremes. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 78(1): 31–51.

Friedman J (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(6): 1189–1232.

Gençay R, Selçuk F, Ulugülyağci A (2003). High volatility, thick tails and extreme value theory in value-at-risk estimation. *Insurance. Mathematics & Economics*, 33(2): 337–356.

Hall P (1982). On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society, Series B, Methodological*, 44(1): 37–42.

Hill B (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5): 1163–1174.

Meinshausen N (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(35): 983–999.

Mentch L, Hooker G (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26): 1–41.

Sandri M, Zuccolotto P (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, 17(3): 611–628.

Wager S, Hastie T, Hastie T (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1): 1625–1651.

Wang H, Tsai C (2009). Tail index regression. *Journal of the American Statistical Association*, 104(487): 1233–1240.

White AP, Liu WZ (1994). Technical mote: Bias in information-based measures in decision tree induction. *Machine Learning*, 15(3): 321–329.

Xu W, Hou Y, Li D (2020). Prediction of extremal expectile based on regression models with heteroscedastic extremes. *Journal of Business & Economic Statistics*, 40(2): 522–536.

Yang Y, Qian W, Zou H (2018). Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business & Economic Statistics*, 36(3): 456–470.

Zhang T, Yu B (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4): 1538–1579.

Zhao Z, Zhang Z, Chen R (2018). Modeling maxima with autoregressive conditional Fréchet model. *Journal of Econometrics*, 207(2): 325–351.