View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Apollo

UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Cambridge Working Papers in Economics

Cambridge Working Papers in Economics: 2060

A DYNAMIC NETWORK OF ARBITRAGE CHARACTERISTICS

Shaoran Li Oliver Linton

29 June 2020

We propose an asset pricing factor model constructed with semi-parametric characteristics based mispricing and factor loading functions. This model captures common movements of stock excess returns and includes a two-layer network of arbitrage returns interconnected by security-specific characteristics. We approximate the unknown functions by B-splines where the number of B-splines coefficients is diverging. We estimate this model and test the existence of the mispricing function by a power enhanced hypothesis test. The enhanced test solves the low power problem caused by diverging B-spline coefficients. Meanwhile, the strengthened power approaches to one asymptotically. And the dynamic networks are explored through Hierarchical K-Means Clusterings. We apply our methodology to CRSP monthly data for the US stock market with one-year rolling windows during 1967-2017. This empirical study shows the presence of mispricing functions in certain time blocks and a dynamic network structure of arbitrage returns through groups of some characteristics.

A Dynamic Network of Arbitrage Characteristics*

Shaoran Li [†1] and Oliver Linton [‡2]

[1 2] *Faculty of Economics, University of Cambridge*

June 28, 2020

Abstract

We propose an asset pricing factor model constructed with semi-parametric characteristics-based mispricing and factor loading functions. This model captures common movements of stock excess returns and includes a two-layer network of arbitrage returns interconnected by security-specific characteristics. We approximate the unknown functions by B-splines where the number of B-splines coefficients is diverging. We estimate this model and test the existence of the mispricing function by a power enhanced hypothesis test. The enhanced test solves the low power problem caused by diverging B-spline coefficients. Meanwhile, the strengthened power approaches to one asymptotically. And the dynamic networks are explored through Hierarchical K-Means Clusterings. We apply our methodology to CRSP monthly data for the US stock market with one-year rolling windows during 1967-2017. This empirical study shows the presence of mispricing functions in certain time blocks and a dynamic network structure of arbitrage returns through groups of some characteristics.

Keywords: Semiparametric; Characteristics-based; Network; Power-enhanced test
JEL Classification: C14; G11; G12

*The authors would like to thank Chaohua Dong, Ondrej Tobek, Yuan Liao, Anders Kock, Joachim Freyberger, Alexei Onatski, Shuyi Ge and Weiguang Liu for their useful comments.

[†]Electronic address: sl736@cam.ac.uk.

[‡]Electronic address: obl20@cam.ac.uk

1 Introduction

Stock returns have both common and firm-specific components. Ross (1976)[2] proposed Arbitrage Pricing Theory (APT) to summarize that expected returns on financial assets can be modeled as a linear combination of various factors. In such a model, each asset has a sensitivity beta to those factors-this tells us how much change in each factor affects the return on each stock. The APT model explains the excess returns from both cross-sectional and time-series directions. Moreover, Rosenberg (1974)[26] argued that stock returns behave as a linear combination of characteristic-related factors. Fama and French (1993, 2014)[7][8] approximated those factors by the returns on portfolios sorted by different characteristics, and they developed three-factor and five-factor models. After extracting the common movement parts, they treated the intercept as the mispricing *alpha,* which is asset-specific and cannot be explained by those risk factors. Many papers use a similar method to present other characteristic-based factor models, such as the four-factor model of Carhart (1997)[3], the q-factor model of Hou, Xue, and Zhang (2015)[16], and the factor zoo by Feng, Giglio and Xiu (2017)[11], see Equation 1. These models utilize risk factors to capture the co-movements while employing α_i and β_{ji} to exploit firm-specific behavior. All of the above papers studied observed factors and the beta coefficients are estimated by time series Ordinary Least Squares (OLS). However, those betas are time-invariant and not assigned characteristics-based information.

Consider the panel regression model

$$y_{it} = \alpha_i + \sum_{j=1}^{J} \beta_{ji} f_{jt} + \epsilon_{it}, \tag{1}$$

where y_{it} is the excess return to security i at time t, f_{jt} are the j^{th} risk factor's returns at time t, β_{ji} denotes the j^{th} factor loading of asset i, α_i represents the intercept (mispricing) of asset i, and ϵ_{it} are the mean zero idiosyncratic shocks. In terms of factor loadings β_{ji}, Connor and Linton [5] and Connor, Hagmann and Linton (2012)[4] studied a characteristic-beta model, which bridges the beta-coefficients and firm-specific characteristics by specifying each beta as an unknown function of one characteristic. In their model, beta functions and unobservable factors are estimated by the backfitting iteration. They concluded that those characteristic-beta functions are *significant and non-linear*. Nonetheless, they restricted their beta function to be univariate and did not consider the unexplained part within each factor loading function, see Equation 2. Their model can be summarized by

$$y_{it} = \sum_{j=1}^{J} g_j(X_{ji}) f_{jt} + \epsilon_{it}, \tag{2}$$

where X_{ji} is the j^{th} observable characteristic of firm i.

To overcome this limitation, Fan, Liao and Wang (2016)[9] allowed β_{ji} in Equation 1 to have a component which is explained by observed characteristics and an unexplained or stochastic part, written as $\beta_{ji} = g_j(X_i) + u_{ji}$, where u_{ji} is mean independent of X_{ji}. They proposed the Projected Principal Component Analysis (PPCA), which projects stocks' excess returns onto the space spanned by firm-specific characteristics and then applies Principal Component Analysis (PCA) to the projected returns in order to find the unobservable factors. This method not only facilitates theoretical analysis but also has attractive properties even under large N and small T asymptotics. However, they did not pay enough attention to the mispricing part (alpha), which is crucial to both asset pricing theories and portfolio management.

The discussion of the mispricing component has been extended from the intercepts part (alpha) of the Fama-French factor model, see Equation 1, to a portfolio management strategy such as constructing portfolios weights through the values of the characteristics-alpha function. For example, Hjalmarsson and Manchev (2012)[14] documented a method of directly parameterizing the portfolio weights as a linear function of characteristics, such as value and momentum. They showed this method outperforms other baseline methods according to empirical studies. This research also sheds light on the possibility of using characteristics to explain the remaining part of multiple factors models. Another seminal paper on firm-specific characteristics was composed by Freyberger, Neuhierl, and Weber (2017)[13], which analyzed the non-linear effects of 62 characteristics through pooling regressions. This study concluded that 13 of these characterisitcs have explanatory power on stocks' excess returns after selection by adaptive group Lasso.

The studies mentioned above demonstrate that characteristics do have explanatory power for stocks' excess returns and can be used as explanatory variables for both mispricing and factor loading functions.

In this paper, we work on a semi-parametric characteristics-based alpha and beta model, which utilizes a set of security-specific characteristics that are similar to Freyberger, Neuhierl, and Weber (2017)[13]. We use unknown multivariate characteristic functions to approximate both α_i and β_{ji} in Equation 1. Specifically, we assume α_i and β_{ji} are functions of a large set

of asset-specific characteristics, taking the form $\alpha_i = h(\mathbf{X}_i) + \gamma_i$ and $\beta_{ji} = g_j(\mathbf{X}_i) + \lambda_i$[1]. We then estimate $h(\mathbf{X}_i)$ and $g_j(\mathbf{X}_i)$ as well as the unobserved risk factors f_{jt}. In addition, we construct a power enhanced test on the mispricing function $h(\mathbf{X}_i)$ to detect possible networks of arbitrage characteristics. We will present our model formally in Equation 3.

Some recent papers such as Kim, Korajczyk and Neuhierl (2019)[20] and Kelly, Pruitt and Su (2019)[19] analyzed a similar model to ours, which assumes that both $h(\mathbf{X}_i)$ and $g_j(\mathbf{X}_i)$ are *parametric linear additive* functions. They both included around 40 characteristics in $\mathbf{X}_i$. However, they drew different conclusions on the existence of $h(\mathbf{X}_i)$. Kim, Korajczyk and Neuhierl (2019)[20] set arbitrage portfolio weights as estimated values of $h(\mathbf{X}_i)$ through one year rolling windows. And then they showed that their arbitrage portfolios are statistically and economically significant. However, Kelly, Pruitt and Su (2019) [19] applied instrumented principal component analysis (IPCA) to the entire time span from 1965 to 2014, and came to a different result, with no evidence to reject the null hypothesis $H_0 : h(\mathbf{X}_i) = \mathbf{X}_i^\intercal \mathbf{B} = \mathbf{0}$. The introduction of IPCA can be found at Kelly, Pruitt and Su (2017)[18]. They conducted their hypothesis tests by bootstrap. This dispute spurred the deveploment of a more flexible model and reliable hypothesis tests to investigate the existence and structure of $h(\mathbf{X}_i)$.

In this paper, the proposed semi-parametric model does not impose strict restrictions on functional forms. This model can exploit information contained in characteristics more efficiently and thoroughly. Meanwhile, the semi-parametric setting can also match the results of Connor, Hagmann and Linton (2012)[4] and some empirical results of Kim, Korajczyk and Neuhierl (2019)[20]. This is promising as it suggests that the more flexible approach is picking up the same relationships as in previous studies. The only assumption on the functional form in my model is *additivity of the unknown functions* $h(\mathbf{X}_i)$ and $g_j(\mathbf{X}_i)$. Each univariate component is approximated by linear combination of B-spline bases.

However, this unrestrictive model brings both opportunities and challenges. According to Huang, Horowitz and Wei (2010)[17], the number of B-spline knots must increase in the number of observations, in order to achieve a more accurate approximation and better asymptotic performance. Therefore, the dimension of B-spline bases' coefficients also grows with the sample size. Furthermore, mispricing functions are treated as anomalies, so under a correctly specified factor model, these coefficients are very likely to be sparse. All of these circumstances make the conventional Wald tests have very low power. Therefore, a power enhanced

[1] $\mathbf{X_i}$ is a vector of a large set of asset-specific characteristics of stock i.

test should be developed to strengthen the power of Wald tests and to detect the most relevant characteristics among a characteristic zoo included in $h(\mathbf{X}_i)$. Kock and Preinerstorfer (2019)[21] illustrated that if the number of coefficients diverges as the number of observations approaches infinity, the standard Wald test is power enhanceable. Meanwhile, Fan, Liao and Yao (2015)[10] proposed a power enhanced test after showing that if true coefficients have a sparse structure, the traditional Wald test has very low power. Therefore, they introduced a screening process to detect all the estimated coefficients one by one, and then select significant coefficients' statistics as a supplement to the standard Wald test. In this paper, we extend Fan, Liao and Yao (2015)[10] to a group manner. We enhance the hypothesis test on a high dimensional additive semi-parametric function $h(\mathbf{X}_i)$ and then test $H_0 : h(\mathbf{X}_i) = 0$. This method allows all the significant components of $h(\mathbf{X}_i)$ to be selected and contribute to the test statistics, with the power of the proposed test approaching to one.

The aforementioned procedures are designed to detect the dynamic structure of $h(\mathbf{X}_i)$, and a careful analysis of $h(\mathbf{X}_i)$ is theoretically and practically meaningful. Firstly, $h(\mathbf{X}_i)$ is an important component of Arbitrage Pricing Theory (APT) and can contribute to asset pricing theories, namely, linking the mispricing functions with security-related characteristics. Therefore, like Hjalmarsson and Manchev (2012)[14] and Kim, Korajczyk and Neuhierl (2019)[20] , $h(\mathbf{X}_i)$ can be utilized to construct arbitrage portfolios. Secondly, the existence of $h(\mathbf{X}_i)$ reveals a network of unsystematic returns that are interconnected through the similarity of assets' characteristics.

As noted by Aymanns, Farmer, Kleinnijenhuis and Wetzer (2018)[1], overlapping portfolios are an important factor in contagion across financial networks. Therefore, the network of mispricing returns can help diversify risks and reduce the overlap across portfolios. If the mispricing function $h(\mathbf{X_i})$ is not monotonic, simply setting portfolio weights to the estimated values of $h(\mathbf{X_i})$ can be problematic. Because some characteristics with significantly different values maybe have similar arbitrage returns. Therefore, the network structure $h(\mathbf{X_i})$ is important. Hoberg and Phillips (2016)[15] classified firms' competitors through the similarity of text-based descriptions, which is similar to the clustering of characteristics. They also compared between text-based and characteristics-based classification methods. In this paper, we analyze the characteristics-based mispricing function as a demonstration that how firms' characteristics can be used to construct a network of arbitrage returns and the distance between two assets i and j has a straightforward computation of arbitrage return distance, namely, $d_{ij} = \|h(\mathbf{X}_i) - h(\mathbf{X}_j)\|$. The similarity of characteristics is$\|\mathbf{X}_i - \mathbf{X}_j\|_2$, where $\|\cdot\|_2$

represents L_2 distance. Inspired by Vogt and Linton (2017) [28], we employ a hierarchical K-means classification to group these characteristics within each mispricing return group.

Understanding the structure of this network can not only improve the performance of arbitrage portfolios by longing assets with similar characteristics which provide high arbitrage returns, but also help us to learn the dynamics of this network through rolling windows. Although this paper mainly focuses on $h(\mathbf{X}_i)$, the same network can be computed based on $g_j(\mathbf{X}_i)$ functions, and overlapping portfolios can be avoided by choosing assets with greater sensitivity difference to a certain risk factor $\mathbf{f}_j$.

It is worth clarifying the differences between our research and others, namely, Kim, Korajczyk and Neuhierl (2019)[20], and Kelly, Pruitt and Su (2019)[19]. The model setting of this paper is *semiparametric*, which brings both flexibility and challenges. This paper considers a different economic question, which is the existence and networks of mispricing characteristics. Therefore, we develop power enhanced hypothesis tests and hierarchical K-means clusterings to investigate the dynamic network structure of $h(\mathbf{X})$ through one rolling windows. Therefore, the IPCA approach in Kelly, Pruitt and Su (2019)[19] is no longer suitable as it requires large T for consistency.

This paper's contributions are threefold. Firstly, we build up a new semi-parametric characteristics-based alpha and beta asset pricing model, which is more general and flexible. Secondly, we apply and extend previous estimation methods, which can fit the current framework better. In addition, we extend the power enhanced test of Fan, Liao, and Yao (2015) [10] in a group manner to strengthen the test on mispricing functions, which can select the characteristics that contribute to the arbitrage portfolios at the same time. Finally, we detect some remarkable networks of analogous mispricing returns, which are interconnected through the similarity of security-specific characteristics, after applying our model and methods on CRSP and Compustat data.

The rest of this paper is organized as follows. Section 2 sets out the semi-parametric model. Section 3 introduces the assumptions and estimation methods. Section 4 constructs a power enhanced test for high dimensional additive semi-parametric functions. Section 5 employs hierarchical K-Means clustering to build the network. Section 6 describes the asymptotic properties of our estimates and test statistics. Section 6 simulates data to verify the performance of our methodology. Section 7 presents an empirical study. Finally, Section 8 concludes this paper. Characteristics description tables, proofs, mispricing curves and network plots are

arranged to the Appendix.

2 Model setup

We assume that there are n securities observed over T time periods. We also assume that during a short period, each security has P time-invariant observed characteristics, such as market capitalization, momentum, and book-to-market ratios. Meanwhile, we may omit the heteroskedasticity by assuming that each characteristic shares a certain form of variation within each period for all securities. We suppose that

$$y_{it} = (h(\mathbf{X}_i) + \gamma_i) + \sum_{j=1}^{J}(g_j(\mathbf{X}_i) + \lambda_{ij})f_{jt} + \epsilon_{it}, \tag{3}$$

where y_{it} is the monthly excess return of the i^{th} stock at the month t, while $\mathbf{X}_i$ is a $1 \times P$ vector of P characteristics of stock i during time periods $t = 1, \dots T$, and where T is a small and fixed time block. In practice, most characteristics are updated annually, and thus, we assume $\mathbf{X}_i$ is time-invariant within this short time period. The $h(\mathbf{X}_i)$ is an unknown mispricing function explained by a large set of characteristics whereas γ_i is the random intercept of the mispricing part that cannot be explained by characteristics. Similarly, we have characteristics-beta function $g_j(\cdot)$ to explain the j^{th} factor loadings and the unexplained stochastic part of the loading is λ_{ij} with $E(\lambda_{ij}) = 0$, which is orthogonal to the $g_j(\cdot)$ function. The quantity f_{jt} is the realization of the j^{th} risk factor at time t. Finally, ϵ_{it} is homoskedastic zero-mean idiosyncratic residual return of the i^{th} stock at time t. The random variables γ_i and λ_{ij} are used to generalize our settings and not to be estimated. They will be treated as noise in the identification assumptions. Furthermore, γ_i can be treated as random effects under conventional panel settings.

Meanwhile, we impose additive forms for both the $h(\cdot)$ and $g_j(\cdot)$ functions to avoid the curse of dimensionality as below: $h(\mathbf{X}_i) = \sum_{p=1}^{P} \mu_p(x_{ip})$ and $g_j(\mathbf{X}_i) = \sum_{p=1}^{P} \theta_{jp}(x_{ip})$, where $\mu_p(x_{ip})$ and $\theta_{jp}(x_{ip})$ are univariate unknown functions of the p^{th} characteristic X_p for the i^{th} asset. Therefore, we compose the model below:

$$y_{it} = (\sum_{p=1}^{P} \mu_p(X_{ip}) + \gamma_i) + \sum_{j=1}^{J}(\sum_{p=1}^{P} \theta_{jp}(X_{ip}) + \lambda_{ij})f_{jt} + \epsilon_{it}, \tag{4}$$

Assumption 1. *We suppose that:*

$$E(\epsilon_{it}|\mathbf{X}, f_{jt}) = 0,$$

$$E(h(\mathbf{X}_i)) = E(g_j(\mathbf{X}_i)) = 0,$$
$$E(\gamma_i|\mathbf{X}) = E(\lambda_{ij}|\mathbf{X}) = 0,$$
$$E(h(\mathbf{X}_i)g_j(\mathbf{X}_i)) = \mathbf{0},$$

Similar to Connor, Hagmann and Linton (2012)[4] and Fan, Liao and Wang (2016)[9], the Assumption 1 above is to standardize the model settings, including the zero mean assumption of factor loadings and mispricing functions for identification purposes. We also impose the orthogonality between mispricing and factor loading parts for identification reason. This is because the variation of the risk factors can be absorbed into the mispricing part if it is not orthogonal to the factor loadings. More discussion can be found in Connor, Hagmann and Linton (2012)[4].

3 Estimation

In this section we discuss the approximation of the unknown univariate functions and our estimation methods for the model Equation 3. In our semi-parametric setting, we applied the Projected-PCA following Fan, Liao and Wang (2016)[9] to work on the common factors and characteristics-beta directly. And then, we project the residuals onto the characteristics-alpha space that is orthogonal to the beta function. The second step is similar to equality constrained OLS estimator. Finally, estimates of the characteristics-beta and alpha can be obtained correspondingly.

3.1 B-Splines Approximation

We construct B-splines to approximate unknown functions $\theta(\cdot)$ and $\mu(\cdot)$ in Equation 4. Similar to Huang, Horowitz and Wei (2010)[17], we have the following procedures. Firstly, suppose that the p^{th} covariate X_p is in the interval $[D_0, D]$, where D_0 and D are finite numbers with $D_0 < D$. Let $\mathbf{D} = \{\underbrace{D_0, D_0, \ldots, D_0}_{l+1} < d_1 < d_2 < \cdots < d_{m_n} < \underbrace{D, D, \ldots, D}_{l+1}\}$ be a simple knot sequence on the interval $[D_0, D]$. Here, $m_n = \lfloor n^v \rceil$ is a positive integer of the number of internal knots, which is a function of security size n in period T, and $0 < v < 0.5$. l is the degree of those bases. Therefore, we have $H_n = l + m_n$ bases in total, which will diverge as $n \to \infty$. Following this setting, a set of B-splines can be built for the space $\Omega_n[\mathbf{D}]$.

Secondly, for the p^{th} characteristic X_p, based on a set of H_n orthogonal bases $\{\phi_{1p}(X_p), \ldots, \phi_{H_np}(X_p)\}$, those univariate unknown functions can be approximated as linear combinations of those bases as $\mu_p(X_p) = \sum_{q=1}^{H} \alpha_q \phi_{qp}(X_p) + R_p^{\mu}(X_p)$ and $\theta_p(X_p) = \sum_{q=1}^{H} \beta_{jq}\phi_{qp}(X_p) + R_p^{\theta}(X_p)$, where $R_p^{\mu}(X_p)$ and $R_p^{\theta}(X_p)$ are approximation errors. It is not necessary to use the same bases for both unknown functions and the representation here is for notation simplicity only. Therefore, the model Equation 4 can be illustrated as:

$$y_{it} = \sum_{p=1}^{P}(\sum_{q=1}^{H_n} \alpha_{pq}\phi_{pq}(X_{ip}) + R_p^{\mu}(X_p)) + \gamma_i + \sum_{j=1}^{J}(\sum_{p=1}^{P}(\sum_{q=1}^{H_n} \beta_{jpq}\phi_{pq}(X_{ip}) + R_p^{\theta}(X_p)) + \lambda_{ij})f_{jt} + \epsilon_{it}$$

For each $i \leqslant n$, $p \leqslant P$ and $t \leqslant T$, we have:

$$\mathbf{1_T} = (1, \ldots, 1)^{\intercal} \in \mathbb{R}^T,$$

$$\beta_j = (\beta_{1,j1}, \ldots, \beta_{H_n,j1}, \ldots, \beta_{1,jP}, \ldots, \beta_{H_n,jP})^{\intercal} \in \mathbb{R}^{H_nP},$$

$$\mathbf{B} = (\beta_1, \ldots, \beta_J),$$

$$\mathbf{A} = (\alpha_{11}, \ldots, \alpha_{1H_n}, \ldots, \alpha_{P1}, \ldots, \alpha_{PH_n})^{\intercal} \in \mathbb{R}^{H_nP},$$

$$\mathbf{\Phi(X)} = \begin{bmatrix} \phi_{1,11}(X_{11}) & \cdots & \phi_{1,1H_n}(X_{11}) & \cdots & \phi_{1,P1}(X_{1P}) & \ldots & \phi_{1,PH_n}(X_{1P}) \\ \phi_{2,11}(X_{21}) & \cdots & \phi_{2,1H_n}(X_{21}) & \cdots & \phi_{2,P1}(X_{2P}) & \ldots & \phi_{2,PH_n}(X_{2P}) \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \\ \phi_{n,11}(X_{n1}) & \cdots & \phi_{n,1H_n}(X_{n1}) & \cdots & \phi_{n,P1}(X_{nP}) & \ldots & \phi_{n,PH_n}(X_{nP}) \end{bmatrix},$$

where $\phi_{i,ph}(X_{ip})$ means the h^{th} basis of the p^{th} characteristic of individual i at time t. Therefore, we have our B-spline model as:

$$\mathbf{Y} = (\mathbf{\Phi(X)A} + \mathbf{\Gamma} + \mathbf{R}^{\mu}(\mathbf{X}))\mathbf{1_T^{\intercal}} + (\mathbf{\Phi(X)B} + \mathbf{\Lambda} + \mathbf{R}^{\theta}(\mathbf{X}))\mathbf{F}^{\intercal} + \mathbf{U}, \tag{5}$$

$\mathbf{Y}$ is $n \times T$ matrix of y_{it}. $\mathbf{\Phi(X)}$ is the $n \times PH_n$ matrix of B-Spline bases. $\mathbf{A}$ is a $PH_n \times 1$ matrix of mispricing coefficients, $\mathbf{R}^{\mu}(\mathbf{X})$ is a $n \times 1$ matrix of approximation errors. $\mathbf{B}$ is a $PH_n \times J$ matrix factor loadings' coefficients. $\mathbf{R}^{\theta}(\mathbf{X})$ is a $n \times J$ matrix of approximation errors. We have $R_p^{\mu}(X_p) \to^p 0$ and $R_p^{\theta}(X_p) \to^p 0$, as $n \to \infty$, see Huang, Horowitz and Wei (2010)[17]. Therefore, we omit the approximation errors for simplicity below. $\mathbf{F}$ is the $T \times J$ matrix of f_{tj} and $\mathbf{U}$ is a $n \times T$ matrix of ϵ_{it}, the rest are defined the same as Equation 4.

Furthermore, we define the projection matrix as:

$$\mathbf{P} = \Phi(\mathbf{X})(\Phi(\mathbf{X})^{\intercal}\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})^{\intercal}.$$

The remaining goals of this paper are to estimate both $h(\boldsymbol{X})$ and $\mathbf{G}(\mathbf{X})$ consistently and conduct a power enhanced test of the hypothesis $H_0 : h(\boldsymbol{X}) = \mathbf{0}$, i.e., to check the existence of mispricing functions under semi-parametric settings. Finally, we construct a network of arbitrage characteristics, recalling that the dimension of $(\alpha_{11}, \ldots, \alpha_{1H_n}, \ldots, \alpha_{P1}, \ldots, \alpha_{PH_n}) \in \mathbb{R}^{PH_n}$ is diverging as $n \to \infty$.

3.2 Two Steps Projected-PCA

In this section, we combine and extend the method of Projected-PCA by Fan, Liao and Wang (2016)[9] and equality constrained least squares similar to Kim, Korajczyk and Neuhierl (2019) [20] to estimate the model above. To facilitate the estimation, we define a $T \times T$ time series demeaning matrix $\mathbf{D_T} = \mathbf{I_T} - \frac{1}{T}\mathbf{1_T}\mathbf{1_T^\intercal}$.[2] Next, we demean the equation above on both sides. Therefore we have

$$\mathbf{Y}\mathbf{D_T} = \tilde{\mathbf{Y}} = (\mathbf{\Phi}(\mathbf{X})\mathbf{B} + \mathbf{\Lambda})\mathbf{F}^\intercal\mathbf{D_T} + \mathbf{U}\mathbf{D_T}.$$

Mispricing terms disappear due to $(\mathbf{\Phi}(\mathbf{X})\mathbf{A} + \mathbf{\Gamma})\mathbf{1_T^\intercal}\mathbf{D_T} = \mathbf{0}$, which help us to work on the characteristics-based factor loadings and those factors only. From now on, we use $\mathbf{F}$ to represent the demeaned factor matrix.

Our procedures are designed to estimate factor loadings $\mathbf{G}(\mathbf{X})$, demeaned unobserved factors $\mathbf{F}$ and mispricing coefficients $\mathbf{A}$ in sequence.

Under the identification conditions in the above section, we have the following estimation procedures similar to Fan, Liao and Wang (2016)[9] and Kim, Korajczyk and Neuhierl (2019)[20] :

1 Projecting $\tilde{\mathbf{Y}}$ onto the spline space spanned by $\{\mathbf{X_{ip}}\}_{i\leqslant n, p\leqslant P}$ through a $n \times n$ projection matrix $\mathbf{P}$ where $\mathbf{P} = \Phi(\mathbf{X})(\Phi(\mathbf{X})^\intercal\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})^\intercal$. We then collected the projected data $\hat{\mathbf{Y}} = \Phi(\mathbf{X})(\Phi(\mathbf{X})^\intercal\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})^\intercal\tilde{\mathbf{Y}}$.

2 Applying the Principle Component Analysis to the projected data $\hat{\mathbf{Y}}^\intercal\hat{\mathbf{Y}}$, which allows us to work directly on the sample covariance of $\mathbf{G}(\mathbf{X})\mathbf{F}^\intercal$, under the condition that $E(g_j(\mathbf{X}_i)\epsilon_{it}) = E(g_j(\mathbf{X}_i)\lambda_{ij}) = 0$.

[2] $\mathbf{I_T}$ is a $T \times T$ identity matrix, and $\mathbf{1_T}$ is a $T \times 1$ matrix of 1.

3 Estimating $\hat{\mathbf{F}}$ as $\frac{1}{\sqrt{T}}$ times the eigenvectors corresponding to the first J (assumed given) eigenvalues of the $T \times T$ matrix $\frac{1}{n}\hat{\mathbf{Y}}^\intercal\hat{\mathbf{Y}}$ (covariance of projected $\hat{\mathbf{Y}}$).

The method above substantially improves the estimation accuracy and facilitates the theoretical analysis. Furthermore, it also has some good properties, even under the large n and small T circumstance. Small T is preferable in our model setting as we use one-year rolling windows analysis in both simulation and empirical studies to explore dynamic network structures of $h(\mathbf{X}_i)$. Besides, large n is required for our asymptotic analysis.

Factor loadings $\hat{\mathbf{G}}^\intercal(\mathbf{X})$ are estimated as:

$$\hat{\mathbf{G}}(\mathbf{X}) = \hat{\mathbf{Y}}\hat{\mathbf{F}}(\hat{\mathbf{F}}^\intercal\hat{\mathbf{F}})^{-1}$$

In the next step, we estimate the coefficients of the mispricing bases.

4 The estimate of $\mathbf{A}$ is

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \mathrm{vec}(\mathbf{Y} - \mathbf{\Phi}(\mathbf{X})\mathbf{A}\mathbf{1}_\mathbf{T}^\intercal - \hat{\mathbf{G}}(\mathbf{X})\hat{\mathbf{F}}^\intercal)^\intercal \mathrm{vec}(\mathbf{Y} - \mathbf{\Phi}(\mathbf{X})\mathbf{A}\mathbf{1}_\mathbf{T}^\intercal - \hat{\mathbf{G}}(\mathbf{X})\hat{\mathbf{F}}^\intercal)$$

subject to $\hat{\mathbf{G}}(\mathbf{X})^\intercal\mathbf{\Phi}(\mathbf{X})\mathbf{A} = \mathbf{0_J}$, a closed-form solution can be obtained below:

Let a $PH_n \times 1$ vector $\hat{\mathbf{A}}$ be the solution of constrained OLS above:

$$\hat{\mathbf{A}} = \mathbf{M}\tilde{\mathbf{A}},$$

where

$$\mathbf{M} = \mathbf{I} - (\Phi(\mathbf{X})^\intercal\Phi(\mathbf{X}))^{-1}\mathbf{\Phi}(\mathbf{X})^\intercal\hat{\mathbf{G}}(\mathbf{X})(\hat{\mathbf{G}}(\mathbf{X})^\intercal\hat{\mathbf{G}}(\mathbf{X}))^{-1}\hat{\mathbf{G}}(\mathbf{X})^\intercal\mathbf{\Phi}(\mathbf{X}),$$

$$\tilde{\mathbf{A}} = \frac{\mathbf{1}}{\mathbf{T}}(\Phi(\mathbf{X})^\intercal\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})^\intercal(\mathbf{Y} - \hat{\mathbf{G}}(\mathbf{X})\hat{\mathbf{F}}^\intercal)\mathbf{1_T},$$

given $\mathbf{P}\hat{\mathbf{G}}(\mathbf{X}) = \hat{\mathbf{G}}(\mathbf{X})$.

As stated in Assumption 1, the $h(\boldsymbol{X})$ is orthogonal to the characteristics-based loadings $\mathbf{G}(\mathbf{X})_\mathbf{i}$.

5 We can also estimate the covariance matrix of $\hat{\mathbf{A}}_\mathbf{i}$, i.e., $\hat{\mathbf{\Sigma}}$, extending the methods of Liew (1976)[22], which can facilitate theoretical analysis in the next section. According to Liew (1976)[22], $\hat{\mathbf{A}}$ is the equality constrained least-square estimates, which has the covariance matrix as:

$$\hat{\mathbf{\Sigma}} = \mathbf{M}\mathbf{\Sigma}_{\tilde{\mathbf{A}}}\mathbf{M}^\intercal,$$

where:

$$\hat{\mathbf{\Sigma}}_{\tilde{\mathbf{A}}} = (\Phi(\mathbf{X})^\intercal \Phi(\mathbf{X}))^{-1} \Phi(\mathbf{X})^\intercal \begin{bmatrix} \hat{\sigma}_1^2 & & \\ & \ddots & \\ & & \hat{\sigma}_n^2 \end{bmatrix} \Phi(\mathbf{X})(\Phi(\mathbf{X})^\intercal \Phi(\mathbf{X}))^{-1},$$

$$\hat{\sigma}_i^2 = \frac{\sum_1^T \hat{e}_{it}^2}{T-1},$$

where $\sum_1^T \hat{e}_{it}^2 = \sum_1^T (y_{it} - \sum_{p=1}^P \sum_{q=1}^{H_n} \hat{\alpha}_{pq}\phi_{pq}(x_{ip}) - \sum_{j=1}^J (\sum_{p=1}^P \sum_{q=1}^H \hat{\beta}_{jpq}\phi_{pq}(x_{ip}))\hat{f}_{jt})^2$.

The heteroskedasticity is caused by the random effects of γ_i.

4 Power Enhanced Tests

There are considerable discussions about the mispricing phenomenon under factor models and the existence of mispricing functions remains controversial. namely, whether there are relevant covariates after subtracting co-movements components captured by risk factors. For recent research, Kim, Korajczyk and Neuhierl (2019)[20] found the characteristics arbitrage opportunities through estimating a linear characteristic mispricing function, without providing theoretical results. However, Kelly, Pruitt and Su (2019)[19] conducted a conventional Wald hypothesis test on the similar mispricing function through bootstrap, concluding that there is no substantial evidence to reject the null hypothesis. Additionally, they applied the bootstrap method to estimate covariance matrix $\hat{\mathbf{\Sigma}}$, which caused potential problems for theoretical analysis. Moreover, according to Fan, Liao and Yao (2015)[10], their test results may have relatively low power when the true coefficient vector of linear mispricing function $\mathbf{A}$ has a sparse structure.

Meanwhile, the research from both studies is based on a parametric framework, which rely on the strong assumption of linearity. However, this assumption is not consistent with Connor, Hagmann and Linton (2012)[4], which showed that both characteristic-beta and mispricing functions are very likely to be non-linear. Therefore, we propose a semiparametric model to accommodate non-linearity to a great extent.

But semi-parametric framework leads to extra challenges for theoretical analysis and hypothesis tests. On the one hand, as mentioned above, the number of coefficients of mispricing B-splines diverge as $n \to \infty$, which implies the power of standard Wald test can be quite low, see Fan, Liao and Yao (2015)[10]. On the other hand, according to other research of

asset pricing, like Fama and French (1993,2014)[7][8], mispricing terms can be regarded as anomalies. This means that in our model setting, the true mispricing coefficient vector $\mathbf{A}$ can be high-dimensional but sparse, reducing the power of conventional Wald test further.

However, according to the results from Kock and Preinerstorfer (2019)[21], conventional hypothesis tests under these circumstances are power enhanceable. The power enhanced Wald test in this paper is an extension of Fan, Liao and Yao (2015)[10] to a group manner, which can also be generalized as hypothesis tests under high-dimensional additive semi-parametric settings. The proposed tests are power strengthened when the coefficients of the additive regression $\mathbf{A}$ is diverging as $n \to \infty$ without distorting the test's size. Meanwhile, this test is also robust under sparse alternatives. Additionally, the proposed test can select the most important components from sparse additive functions. Finally, the proposed method can also be applied when the number of characteristics is also diverging, i.e. $P \to \infty$.

Hence, we construct a new test:

$$H_0 : h(\boldsymbol{X}) = \mathbf{0}, \qquad H_1 : h(\boldsymbol{X}) \neq \mathbf{0},$$

equivalently,

$$H_0 : \mathbf{A} = \mathbf{0}, \qquad H_1 : \mathbf{A} \in \mathcal{A},$$

where $\mathcal{A} \subset \mathbb{R}^{PH_n} \backslash \mathbf{0}$.

Here, we have:

$$S_1 = \frac{\hat{\mathbf{A}}\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{A}}^{\intercal} - PH_n}{\sqrt{2PH_n}}$$

where S_1 is the "original" test statistics. P is the number of characteristics. PH_n is the total number of B-spline bases, and $\mathbf{A} \in \mathbb{R}^{PH_n}$. H_n is a function of asset number n, therefore, $H_n \to \infty$ as $n \to \infty$. Under H_0, S_1 has nondegenerate limiting distribution F as $n \to \infty$. Given the significance level q, $q \in (0, 1)$ as well as the critical value F_q:

$$S_1 | H_0 \to^d F$$

$$\lim_{N \to \infty} \Pr(S_1 > F_q | H_0) = q$$

Pesaran and Yamagata (2012)[23] illustrated that:

$$S_1 | H_0 \to^d \mathcal{N}(0, 1),$$

under regularity conditions.

Possible sparsity and diverging PH_n means that it is plausible to add a power enhanced component to S_1, which can improve the power of the hypothesis test without any size distortions.

Therefore, we can construct an extra screening component S_0 as:

$$S_0 = H_n \sum_{p=1}^{P} \mathbf{I}(\sum_{h=1}^{H_n} |\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_n),$$

where, $\hat{\sigma}_{ph}$ is the ph^{th} entry of the diagonal elements of $\hat{\mathbf{\Sigma}}$. $\mathbf{I}(\cdot)$ is an indicator for screening process while η_n is a data-driven threshold value to avoid potential size-distortion.

Here we use some space to discuss the choice of η_n. By construction and Assumption 3 below, we know all the B-Spline basis and characteristics are orthogonal. Therefore, all the elements of $\hat{\mathbf{A}}$ are asymptotically i.i.d normal distributed under $\mathbf{H_0}$. And our goal is to bound the maximum of those values.

Define $Z = \max_{\{1\leqslant p\leqslant P, 1\leqslant h\leqslant H_n\}} \{|\hat{\alpha}_{ph}|/\hat{\sigma}_{ph}\}$. Under Assumption 3 below, we have

$$\hat{\alpha}_{ph}/\hat{\sigma}_{ph}|\mathbf{H_0} \to^d N(0,1).$$

After grouping coefficients of bases that used to represent the unknown function of each characteristic, let $Q = \max(\sum_{h=1}^{H_n} |\hat{\alpha}_{1h}|/\hat{\sigma}_{1h}, \ldots, \sum_{h=1}^{H_n} |\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \ldots, \sum_{h=1}^{H_n} |\hat{\alpha}_{Ph}|/\hat{\sigma}_{Ph})$. Following this, we may set the threshold as $\eta_n = H_n\sqrt{2\log(PH_n)}$, where $H_n = l + n^v$. As the H_n is a slowly diverging sequence, it can control the influence of the group size properly. Meanwhile, the η_n also diverges slowly. η_n is a conservative threshold value to avoid potential size distortion.

Apart from strengthening the power of conventional hypothesis test, $\mathbf{I}(\cdot)$ is a screening term which can select the most relevant characteristics at the same time.

Here, we define the arbitrage characteristics set, which includes the characteristics that have the strong explanation power on the mispricing functions:

$$\mathcal{M} = \{\mathbf{X_m} \in \mathcal{M} : \sum_{h=1}^{H_n} |\alpha_{ph}|/\sigma_{ph} \geqslant \eta_n, \quad m = 1, 2, \ldots, M\}$$

$$\hat{\mathcal{M}} = \{\mathbf{X_m} \in \hat{\mathcal{M}} : \sum_{h=1}^{H_n} |\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_N, \quad m = 1, 2, \ldots, M\}$$

Therefore, we have $\mathcal{M} \cup \mathbf{0} = \mathcal{A}$ and $\mathcal{M} \cap \mathbf{0} = \emptyset$. When the set $\mathcal{M}$ is relatively small, conventional tests are likely to suffer the lower power problem. The added S_0 can strengthen the power of the tests to a great extent as H_n is a slowly diverging value .

Therefore, our new test statistics is $S = S_0 + S_1$ and conclusions of hypothesis tests are made accordingly. Asymptotic properties of S will be discussed later.

To conclude, the advantages of our new statistics $S = S_0 + S_1$ are:

1 The power of the hypothesis test on mispricing functions are mainly enhanced without size distortions.

2 We can find specific characteristics which cause the mispricing by screening mechanism.

As designed, S_0 satisfies all three properties of Fan, Liao and Yao (2015)[10], as $n \to \infty$:

1 S_0 is non-negative, $\Pr(S_0 \geqslant 0) = 1$

2 S_0 does not cause size distortion, under H_0, we have $\Pr(S_0 = 0 \mid H_0) \to 1$

3 S_0 enhances test power, under alternative H_1, S_0 diverge quickly in probability given the well chosen $\eta_{n,T}$.

Based on properties of S_0, we have three properties of S below:

1 No size distortion $\limsup_{n\to\infty} \Pr(S > F_q|H_0) = q$

2 $\Pr(S > F_q|H_1) \geqslant \Pr(S_1 > F_q|H_1)$. Hence, the power of S is at least as large as that of S_1.

3 If S_0 diverges very quickly, we have $\Pr(S > F_q|H_1) \to 1$. This happens, especially, when the true form of $\hat{\mathbf{A}}$ has a sparse structure.

5 Hierarchical K-Means Clustering

This section introduces a hierarchical K-means clustering method to construct a network of arbitrage interconnected through assets' characteristics.

After the screening process in section 4, we obtain the relevant components of mispricing function $h(\boldsymbol{X})$, which is estimated as

$$\hat{\mathcal{M}} = \{\mathbf{X_m} \in \hat{\mathcal{M}} : \sum_{h=1}^{H_n} |\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_N, \quad m = 1, 2, \ldots, M\}.$$

Therefore, we define an arbitrage characteristics $n \times M$ matrix $\boldsymbol{M}$ at time window t as :

$$\boldsymbol{M} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_M}\}, \text{ where } \boldsymbol{X_m} \in \hat{\mathcal{M}}.$$

Note that these characteristics are time-invariant within each rolling window. Here we also set arbitrage returns of asset i at rolling window t as:

$$\ddot{y}_{it} = \boldsymbol{\phi}(\boldsymbol{X_i})\hat{\boldsymbol{A}}.$$

For each rolling window, we classify all n assets through 2 layers K-means clustering. At the first layer, we group the assets into K groups according to the similarity of their arbitrage returns $\ddot{y}_{it}$. At the second layer, we divide R_j subgroups within the j^{th} clustering of the first layer groups by the similarity of their characteristics, where $j = 1, 2, \ldots, K$. Finally, the network of arbitrage portfolios can be obtained through these clusterings. K-means clustering is a popular method to classify similar groups, see Cox(1957)[6] and Fisher(1958)[12] for details.

We give the classification procedures of both layers below. We define Δ_{ij} as the difference between arbitrage returns of $\ddot{y}_{it}$ and $\ddot{y}_{jt}$, as well as Υ_{ij} as the difference between characteristics:

$$\Delta_{ij} = \ddot{y}_{it} - \ddot{y}_{jt}, \text{ where } i \neq j, j = 1, 2, \ldots, n.$$

$$\Upsilon_{ij} = \|\boldsymbol{M}_i - \boldsymbol{M}_j\|_2, \text{ where } i \neq j, i, j = 1, 2, \ldots, n,$$

M_i represents the i^{th} row of $\boldsymbol{M}$. We also set two tolerance thresholds as ψ_y and ψ_x, which are used to control the biggest difference within each group of both layers separately. Similar to Vogt and Linton (2017)[28], we apply a first difference process before the K-means clustering, as this is an efficient way to get the initial centroids for K-means to converge more quickly.

For the first layer:

1. **First difference**: We randomly pick i^{th} asset and then we calculate Δ_{ij} with other assets $j = 1, 2, \ldots, n$. Thus we obtain $\Delta_{i(1)} \ldots \Delta_{i(n)}$, with n being the total individuals for

classification. Without loss of generality, we assume $\Delta_{i(1)} = \min\{\Delta_{i(1)} \dots \Delta_{i(n)}\}$, and $\Delta_{i(n)} = \max\{\Delta_{i(1)} \dots \Delta_{i(n)}\}$.

2. **Ordering**: We rank the values obtained in Step 1 as follows:

$$\begin{aligned}\Delta_{i(1)} \leqslant \ldots \leqslant \Delta_{i(j_1-1)} &< \Delta_{i(j_1)} \leqslant \ldots \leqslant \Delta_{i(j_2-1)} \\ &< \Delta_{i(j_2)} \leqslant \ldots \leqslant \Delta_{i(j_3-1)} \\ &\vdots \\ &< \Delta_{i(j_{K-1})} \leqslant \ldots \leqslant \Delta_{i(n)}.\end{aligned}$$

We use the strict inequality mark to show the large jumps of "first difference", all of which are assumed to be larger than ψ_y , while the weak inequality means that the distance calculated is smaller than ψ_y. We identify $K-1$ jumps that are larger than ψ_y above. Thus, the initial classification is achieved and we have a total of K groups with $j_1 - 1$ members in the first group, $\mathcal{C}_1$, $j_2 - j_1$ members in the second group, $\mathcal{C}_2$, ..., and $j_n - j_K + 1$ members in the final group $\mathcal{C}_K$.

In terms of the second layer, for the assets in the k^{th} group $\mathcal{C}_k$, we use the same methods to further divide them into r subgroups as $\mathcal{R}_{1k}, \mathcal{R}_{2k}, \ldots, \mathcal{R}_{rk}$ within each subgroup until we have:

$$\Upsilon_{ab} = \|\boldsymbol{M}_a - \boldsymbol{M}_b\|_2 \leqslant \psi_x, \text{ where } a, b \in \mathcal{R}_{ik}, i = 1, 2, \ldots, r, \text{ and } k = 1, 2, \ldots, K.$$

The K-means algorithm is:

1. 1^{st} Step: Determine the starting mean values for each group $\hat{\ddot{c}}_1^{[0]}, \ldots, \hat{\ddot{c}}_K^{[0]}$ and calculate the distances $\hat{D}_k(i) = \Delta(\ddot{y}_{it}, \hat{\ddot{c}}_k^{[0]}) = |\ddot{y}_{it} - \hat{\ddot{c}}_k^{[0]}|$ for each i and k. Define the partition $\{\mathcal{C}_1^{[0]}, \ldots, \mathcal{C}_K^{[0]}\}$ by assigning the i^{th} individual to the k-th group $\mathcal{C}_k^{[0]}$ if $\hat{D}_k(i) = \min_{1 \leqslant k' \leqslant K} \hat{D}_{k'}(i)$.

2. l^{th} Step: Let $\{\mathcal{C}_1^{[l-1]}, \ldots, \mathcal{C}_K^{[l-1]}\}$ be the partition of $\{1, \ldots, n\}$ from the latest iteration step. Calculate mean functions

$$\hat{\ddot{c}}_k^{[l]} = \frac{1}{|\mathcal{C}_k^{[l-1]}|} \sum_{i \in \mathcal{C}_k^{[l-1]}} \ddot{y}_{it} \qquad \text{for } 1 \leqslant k \leqslant K$$

And then we calculate $\Delta(\ddot{y}_{it}, \hat{\ddot{c}}_k^{[l]}) = |\ddot{y}_{it} - \hat{\ddot{c}}_k^{[l]}|$ for each i and k. Define the partition $\{\mathcal{C}_1^{[l]}, \ldots, \mathcal{C}_K^{[l]}\}$ by assigning the i^{th} individual to the k-th group $\mathcal{C}_k^{[l]}$ if $\hat{D}_k(i) = \min_{1 \leqslant k' \leqslant K_0} \hat{D}_{k'}(i)$.

3. Iterate the above steps until the partition $\{\mathcal{C}_1^{[w]}, \ldots, \mathcal{C}_K^{[w]}\}$ does not change anymore.

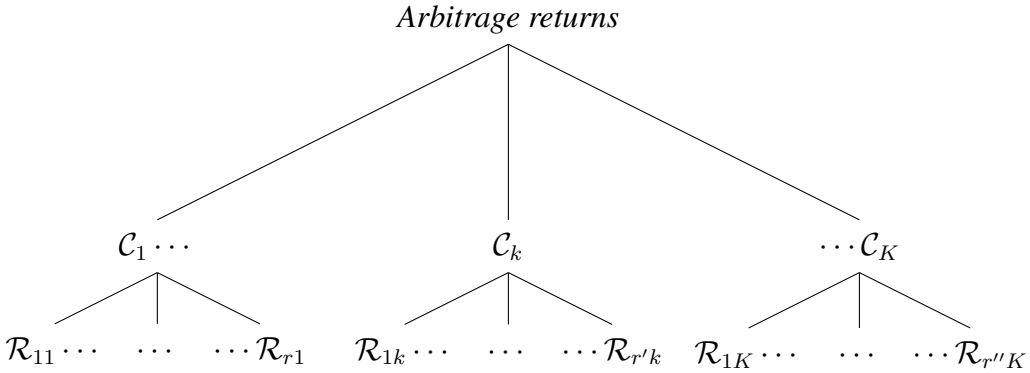
In order to accelerate the convergence of K-means algorithm, at the 1^{th} step, the results of **first difference** are used. As we have already obtained our initial grouping after double difference as $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, we therefore have our starting values for the 1^{st} Step:

$$\hat{\ddot{c}}_k^{[0]} = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \ddot{y}_{it} \qquad \text{for } 1 \leqslant k \leqslant K,$$

where $|\mathcal{C}_k|$ means the cardinality of the group $\mathcal{C}_k$.

The consistency and other theoretical results of above process can be found in Pollard(1981,1982)[24][25], Sun, Wang and Fang (2012)[27] and Vogt and Linton (2017)[28].

For the second layer, we repeat the procedures within each group $\mathcal{C}_k$ respect to Υ_{ab}, and we can obtain the network of characteristic arbitrage returns as:



The first layer is the structure of the arbitrage returns, while the second layer is the category of characteristics that can provide similar arbitrage returns.

6 Asymptotic properties

This section discusses assumptions and properties of estimates and power enhanced statistics S.

6.1 Consistency Assumptions

Assumption 2. *As $n \to \infty$, we have:*

$$\frac{1}{n}\mathbf{Y}^\intercal\mathbf{Y} \to_P \mathbf{M_Y},$$
$$\mathbf{F}^\intercal\mathbf{F} = \mathbf{I_J},$$

where $\mathbf{M_Y}$ is a positive definite matrix and $\mathbf{I_J}$ is a $J \times J$ identity matrix.

We define $\lambda_{min}(M)$ and $\lambda_{max}(M)$ as the largest and smallest eigenvalues of matrix M. Additionally, we define C_{min} and C_{max} are positive constants such that:

$$C_{min} \leqslant \lambda_{min}(\frac{1}{n}\mathbf{\Phi}^\intercal(\mathbf{X})\mathbf{\Phi}(\mathbf{X})) < \lambda_{max}(\frac{1}{n}\mathbf{\Phi}^\intercal(\mathbf{X})\mathbf{\Phi}(\mathbf{X})) \leqslant C_{max}$$

as $n \to \infty$.

We impose these restrictions above to avoid non-invertible stock returns, characteristics, and rotation indeterminacy separately.

Assumption 3.

$$\frac{1}{n}\mathbf{G}(\mathbf{X})^\intercal\mathbf{P}\mathbf{G}(\mathbf{X}) \to_P \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_{PH_n} \end{bmatrix},$$

as $n \to \infty$, where d_{PH_n} are distinct entries.

Both Assumption 2 and 3 are similar to Fan, Liao and Wang (2016)[9], which are used to separately identify risk factors and factor loadings. Given the orthogonal bases of B-splines and uncorrelated or weakly correlated characteristics, Assumption 3 is mild.

Assumption 4. *K_{min} and K_{max} are positive constants such that:*

$$K_{min} \leqslant \lambda_{min}(\frac{1}{n}\mathbf{G}(\mathbf{X})^\intercal\mathbf{P}\mathbf{G}(\mathbf{X})) < \lambda_{min}(\frac{1}{n}\mathbf{G}(\mathbf{X})^\intercal\mathbf{P}\mathbf{G}(\mathbf{X})) \leqslant K_{max}$$

as $n \to \infty$.

This assumption requires the nonvanishing explanatory power of the B-spline bases $\mathbf{\Phi}(\mathbf{X})$ on the factor loading matrix $\mathbf{G}(\mathbf{X})$. This Assumption is mild, and, as we discussed in the introduction, the explanatory power of characteristics on excess stocks' returns have been verified by a lot of previous research.

Assumption 5. *ϵ_{it} is realized i.i.d. idiosyncratic shocks with $E(\epsilon_{it}) = 0$ and $\mathrm{var}(\epsilon_{it}) = \sigma^2$.*

This assumption verifies the heteroskedasticity across different assets, which is caused by the random effect γ_i, namely, $\mathrm{var}(\gamma_i + \epsilon_{it}) = \sigma_i^2$. The off-diagonal elements are assumed to be zeros, as we have subtracted the co-movement part $\hat{\mathbf{G}}(\mathbf{X})\hat{\mathbf{F}}^\intercal$ from the $\mathbf{Y}$.

6.2 Main Results

Theorem 6.1. *Let $\hat{\mathbf{F}}$ be a $J \times T$ matrix estimate of latent risk factors, Under Assumption 1-4, as $n \to \infty$, then $\hat{\mathbf{F}} \to^P \mathbf{F}$.*

Theorem 6.2. *Define the $n \times J$ matrix $\hat{\mathbf{G}}(\mathbf{X})$ as the estimate of factor loadings $\mathbf{G}(\mathbf{X})$. Under Assumption 1-4 and Theorem 6.1 , as $n \to \infty$, then $\hat{\mathbf{G}}(\mathbf{x}) \to^P \mathbf{G}(\mathbf{X})$.*

Theorem 6.3. *Let a $PH_n \times 1$ vector $\hat{\mathbf{A}}$ be the solution of constrained OLS above,*

$$\hat{\mathbf{A}} = \mathbf{M}\tilde{\mathbf{A}},$$

where

$$\mathbf{M} = \mathbf{I} - (\Phi(\mathbf{X})^\intercal\Phi(\mathbf{X}))^{-1}\mathbf{\Phi}(\mathbf{X})^\intercal\hat{\mathbf{G}}(\mathbf{X})(\hat{\mathbf{G}}(\mathbf{X})^\intercal\mathbf{\Phi}(\mathbf{X})(\Phi(\mathbf{X})^\intercal\Phi(\mathbf{X}))^{-1}\mathbf{\Phi}(\mathbf{X})^\intercal\hat{\mathbf{G}}(\mathbf{X}))^{-1}\hat{\mathbf{G}}(\mathbf{X})^\intercal\mathbf{\Phi}(\mathbf{X}),$$

$$\tilde{\mathbf{A}} = \tfrac{\mathbf{1}}{\mathbf{T}}(\Phi(\mathbf{X})^\intercal\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})^\intercal(\mathbf{Y} - \hat{\mathbf{G}}(\mathbf{X})\hat{\mathbf{F}}^\intercal)\mathbf{1}_\mathbf{T}^\intercal.$$

Under Assumption 1-4, as $n \to \infty$, then $\mathbf{\Phi}(\mathbf{X})\hat{\mathbf{A}} \to^P h(\boldsymbol{X})$.

Theorem 6.4. *Under Assumption 3 and Assumption 5, $\mathbf{E}(\mathbf{Z}) = \sqrt{2\log PH_n}$.*

Theorem 6.5. *Define η_n as the threshold value to control the maximum noise, then:*

$$\inf_{\alpha\in\mathbf{0}} \Pr(\max_{p\leqslant P, h\leqslant H} |\hat{\alpha}_{ph} - \alpha_{ph}|/\hat{\sigma}_{ph} \leqslant \eta_n|\mathbf{A}) \to 1.$$

Under $n \to \infty$ and H_0, given the properties of S_0 and S_1, then:

$$S \to^d N(0, 1),$$

The power of S is enhanced now as:

$$\inf_{\mathbf{A}\in\mathcal{A}} \Pr(\textit{reject } H_0|\mathbf{A}) \to 1.$$

7 Numerical Study

In this section, we use Compustats and Fama-French three and five factors' data to simulate stocks' returns and then demonstrate the performance of our estimation and hypothesis test procedures.

7.1 Data Generation

Firstly, we use Fama-French three factors' monthly returns and all the characteristics that will be included in the empirical study to mimic the stocks' returns. Most of the characteristics are updated annually; therefore, we treat those variables as time-invariant during the one-year rolling block. For the characteristics that vary every month, we substitute the mean values as their fixed values per fiscal year. We match Fama-French monthly returns from July of year t to June of year $t+1$ and characteristics of fiscal year $t-1$ to generate the stock returns between July of year t to June of year $t+1$. The period we generate is the same as the empirical study, namely, the 50 years from July 1967-June 2017. Therefore, for each rolling block of 12 months we have:

$$y_{it} = h(X_i) + \sum_{j=1}^{3} g_j(X_j) f_{jt} + \epsilon_{it}, \tag{6}$$

where y_{it} is the generated stock's return. $h(X_i)$ is the mispricing function consists of a non-linear characteristic function of x_i, to mimic the sparse structure of a mispricing function. $g_j(\boldsymbol{X}_j)$ is the j^{th} characteristics-based factor loading, which has an additive semi-parametric structure, and $\boldsymbol{X}_j$ is the j^{th} subset consisting of 4 characteristics. f_{jt} is the j^{th} Fama-French factor returns at time t. ϵ_{it} is the idiosyncratic shock, generated from $N(0, \sigma^2)$.

We generate the characteristic univariate functions as:

$$h(X_i) = \sin X_i,$$

$$g_1(\boldsymbol{X}_1) = X_1^2 + (3X_2^3 - 2X_2^2) + (3X_3^3 - 2X_3) + X_4^2,$$
$$g_2(\boldsymbol{X}_2) = X_5^2 + (3X_6^3 - 2X_6^2) + (3X_7^3 - 2X_7) + X_8^2,$$
$$g_3(\boldsymbol{X}_3) = X_9^2 + (3X_{10}^3 - 2X_{10}^2) + (3X_{11}^3 - 2X_{11}) + X_{12}^2,$$

where X_i is a randomly picked characteristic and $i \neq 1, \ldots, 11, 12$. Furthermore, all the $X_1, \ldots, X_{12}$ are chosen from the characteristics of the empirical study without duplication;

description of these characteristics can be found in the Appendix. Additionlly, all $h(X_i)$, $g_1(\boldsymbol{X}_1)$, $g_2(\boldsymbol{X}_2)$ and $g_3(\boldsymbol{X}_3)$ are rescaled to be mean 0 and variance 1. As we use the real data to conduct the simulation, the assumption of independent X_i cannot be satisfied. Whilst some characterisitcs are highly correlated, we can see from the simulation that the semi-paramtric model overcome this problem properly when being compared with the serious size distortion under parametric models.

7.2 Model Misspecification

In this simulated experiment, our purpose is to show the necessity to consider semi-parametric analysis when the form of factor loading and mispricing functions are unknown.

Under the data generation process, we use both semi-parametric and linear analysis to compare the Mean Squared Error (MSE) and hypothesis test results under both specifications. We apply our estimation methodology in section 3 to estimate Equation 7.1. For semi-parametric specification, we choose the number of B-Spline bases to be $\lfloor n^{0.3} \rceil$. n is the number of assets in each balanced rolling window and $\lfloor \cdot \rceil$ means the nearest integer. We orthogonalize these bases and then use the Projected-PCA and restricted OLS to estimate model Equation 7.1. As for the hypothesis test part, we choose threshold value to be $\eta_n = H_n\sqrt{2\log(PH_n)} = \lfloor n^{0.3} \rceil \sqrt{2\log(P\lfloor n^{0.3} \rceil)}$, where P is the number of characteristics and n is the number of stocks in each rolling block. For the linear specification, each characteristic only has one basis, which is itself. And then, we repeat the procedure in section 3. In terms of hypothesis test,we use the same logic as in the semi-parametric settings. We set for $\eta_n = \sqrt{3\log(P)}$.

In all the estimation above, we assume we know the real number of factors, which is three. We will discuss the situation when the number of factors is unknown in the next subsection. Mean Squared Error (MSE) is also reported to measure the fitness of the model Equation 7.1.

As we can see from Table 1, under different noise levels, namely $\sigma^2 = 1$ and $\sigma^2 = 4$, the semi-parametric model outperforms the linear model in the following aspects:

1 The fitness of the semi-parametric model is much better than the linear model, which can be illustrated from MSE.

2 The semi-parametric model can enhance the power of S_1 by non-zero S_0, which can not only select the correct mispricing characteristics by also avoid size distortions. As for

the linear model, it is influenced by the high correlation of characteristics. Therefore, during certain periods we even obtained the non-invertible characteristic matrix. The linear model can also select the relevant covariance with decent probability, but it suffers from serious size distortions. Thus, our semi-parametric model with orthogonal bases can mitigate this problem to a great extent.

3 The additional component S_0 is necessary to strengthen the power of S_1 and select the relevant characteristics that can explain the mispricing function. Because S_1 can be very small and even negative, especially when the noise σ_i is strong.

Table 1: Simulation Results 1 Part1

		$\sigma^2 = 1$												$\sigma^2 = 4$									
		Linear Model						Semi-parametric Model						Linear Model						Semi-parametric Model					
Time	n	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%
1	468	24.9	11.5	13.4	6.4	100%	100%	-0.5	6.2	-5.7	6	81.2%	0%	14.2	10.8	3.4	8.6	100%	87.4%	-8.2	0	-8.2	8.1	0%	0%
2	894	32.8	11.6	21.2	2	100%	100%	3.4	8	-4.6	1.6	99.9%	0%	11.4	5.8	5.6	4.3	100%	2.1%	-8.5	0	-8.5	3.7	0%	0%
3	1108	34.4	5.7	28.7	11.9	100%	0%	8.6	9	-0.4	11.5	100%	0%	17.1	5.7	11.4	14.1	100%	0%	-7	0.7	-7.7	13.7	7.3%	0%
4	1199	-0.57	0	-0.57	10.2	0%	0%	9.2	9.1	0.1	9.5	96.8%	4.3%	-1.4	0	-1.4	12.5	0%	0%	-6.1		0.06	-6.2	7%	0%
5	1333	92	19.6	72.4	2.31	100%	100%	10.6	9	1.6	2	100%	0%	28.2	6.1	22	4.5	100%	6.5%	0.2	7.4	-7.2	4.1	82.8%	0%
6	1409	90	28.5	61.5	16	100%	100%	28.6	12.6	15.9	15.8	100%	28%	45.3	16.1	29.2	18.4	100%	73.4%	16.3	10.9	5.4	17.5	68.4%	35.9%
7	1466	78.4	10.6	67.8	6.4	100%	74.2%	19.5	9	10.5	6.2	100%	0%	34.8	5.7	29.1	8.6	100%	0.02%	4.3	9	-4.7	8.4	99.9%	0%
8	1560	133	16.8	116.2	3.3	100%	100%	20.3	10	10.3	3.2	100%	0%	45.2	6.1	39.1	5.5	100%	6.9%	4.2	10	-5.8	5.4	100%	0%
9	1494	117.7	13.6	104.1	3.6	100%	100%	23.1	9	14.1	3.5	100%	0%	44.1	7.6	36.5	5.8	100%	32.4%	6	9	-3	5.6	100%	0.01%
10	1292	90.7	11.5	79.2	3.7	100%	100%	16.2	9	7.2	3.6	100%	0%	39.5	9.3	30.2	5.9	100%	61.1%	3.6	8.9	-5.3	5.7	99.7%	0%
11	1393	84.7	10.6	74.1	6.1	100%	85.1%	20.7	9.1	11.6	5.8	100%	1.1%	37.1	6.5	30.6	8.3	100%	12.9%	8.9	8.9	0	7.8	98.1%	1.3%
12	1340	83.5	28	55.5	2.38	100%	100%	10.6	9	1.6	2	100%	0%	26	6.2	19.8	4.6	100%	7.1%	-1.8	5.7	-7.5	4.1	63.7	0%
13	1285	113.8	16	97.8	1.73	100%	100%	10.6	9	1.6	1.6	100%	0%	34.5	6.6	27.9	4	100%	15.3%	-2.4	5.1	-7.5	3.7	57.1%	0%
14	1181	88.5	12.8	75.7	4.7	100%	100%	15.8	9	6.8	4.5	100%	0%	31.2	5.9	25.3	6.9	100%	2.3%	3.7	9	-5.3	6.6	100%	0%
15	1110	45.7	7.5	38.1	8.9	100%	30.4%	11.5	9	2.5	8.7	100%	0%	23.9	5.8	18.1	11.1	100%	0.6%	-2	4.8	-6.8	10.8	0.54%	0%
16	1044	20.5	5.7	14.8	18.4	100%	0%	9.9	9	0.9	17.9	100%	0%	14.6	5.7	8.9	20.6	100%	0%	1.2	6.1	-4.9	20	68.1%	0.2%
17	1125	59.4	11.5	47.9	9.2	100%	100%	13.2	9	4.2	9	100%	0%	27.2	6.2	21	11.5	100%	8.4%	2.6	8.8	-6.2	11	97.9%	0%
18	2192	NA	NA	NA	NA	NA	NA	23.2	11	12.2	4.3	100%	0%	NA	NA	NA	NA	NA	NA	6.7	11	-4.3	6.4	100%	0%
19	2236	56.1	11.5	44.6	5.8	100%	100%	17.8	11	6.8	5.2	100%	0%	28.3	6.3	22	8	100%	20.3%	4.3	11	-6.7	7.4	100%	0%
20	2273	43.3	5.7	37.6	3.8	100%	0%	22.4	11	11.4	3.2	100%	0%	22.4	5.7	16.7	6.1	100%	0%	5	10.2	-5.2	5.4	92.6%	0%
21	2235	59.8	11.8	48	2.7	100%	100%	20.2	11	9.2	2	100%	0%	25	7.3	17.7	4.9	100%	28.2%	4.6	11	-6.4	4.2	100%	0%
22	2270	40.2	11.5	28.7	2.78	100%	99.5%	17.2	11.6	5.6	2.1	100%	0%	17.1	5.9	11.2	5	100%	3.5%	-6	0.1	-6.1	4.2	1.1%	0%
23	2405	41.4	8.9	32.5	4.1	100%	54.2%	16.3	11	5.3	3.3	100%	0%	18.7	5.8	12.9	6.3	100%	7.1%	-3.3	3	-6.3	5.5	27.3%	0%
24	2376	19	9.7	9.3	1.8	100%	69.9%	23.1	11	12.1	1	100%	0%	7.5	5.7	1.8	4	100%	0%	5.6	11	-5.4	3.2	100%	0%
25	2323	15.9	9.5	6.4	3.5	66.7%	98.6%	20.6	11	9.6	2.7	100%	0%	1.1	0	1.1	5.8	0%	0%	5.3	11	-5.7	4.9	100%	0%
26	2344	NA	NA	NA	NA	NA	NA	24.9	12.9	12	3.3	100%	17.1%	NA	NA	NA	NA	NA	NA	6.5	11	-4.5	5.4	100%	0%
27	2434	NA	NA	NA	NA	NA	NA	27.3	11	16.3	1.2	100%	0%	NA	NA	NA	NA	NA	NA	6.9	11	-4.1	3.4	100%	0%
28	2548	0.9	0	0.9	4.2	0%	0%	26.2	11	15.2	3.3	100%	0%	-1.3	0	-1.3	6.5	0%	0%	6.9	11	-4.1	5.5	100%	0%
29	2741	10.3	5.7	4.5	4.2	100%	0%	58.2	11.1	47.1	3.4	100%	1.3%	6.6	5.7	0.9	6.4	100%	0%	17.6	11	6.6	5.5	100%	0%
30	2928	5.6	4.6	1	7.1	80.4%	0%	59.2	11.8	47.4	6.3	100%	7.8%	-0.4	0.1	-0.5	9.3	2.5%	0%	18.8	11	7.8	8.5	100%	0.3%
31	2894	13.4	5.7	7.7	6.4	100%	0%	61	13.4	47.6	5.7	100%	21.6%	8.1	5.7	2.3	8.6	100%	0%	17.7	11	6.7	7.8	100%	0.2%
32	2905	23.1	11.5	11.6	5.9	100%	100%	33.2	11.3	21.9	5.2	100%	3%	12.9	8.5	4.4	8.1	100%	48.2%	9.8	11	-1.2	7.4	100%	0%
33	2804	9.8	5.7	4.1	9.6	100%	0%	42.7	18.5	24.2	8.9	100%	68.5%	7.3	5.7	1.6	11.9	100%	0%	9.7	11	-1.3	11.2	100%	0%
34	2570	6.9	5.7	1.2	22	99.7%	0%	37.3	12.2	25.1	21.2	100%	10.4%	2	1.9	0.1	24	34.4%	0%	12.7	11	1.7	23.3	100%	0.2%
35	2516	8.3	5.7	2.6	7.9	100%	0%	41.3	11	30.3	7.2	100%	0.4%	5.1	5.02	0.08	10.1	87.3%	0%	12.9	11	1.9	9.4	100%	0%
36	2491	10.7	5.7	4.9	2.1	100%	0%	41.3	11	30.3	1.4	100%	0.4%	0.5	0.25	0.25	4.4	4.5%	0%	12.4	11	1.4	3.6	100%	0%
37	2402	14.1	5.7	8.4	5.6	100%	0%	26.5	11.2	15.3	4.9	100%	2.2%	8.8	5.7	3.1	7.9	100%	0%	7.9	11	-3.1	7.1	100%	0%
38	2326	19.7	9.6	10.1	3	100%	66.8%	28.9	11.3	17.6	2.3	100%	2.1%	8.1	5.8	2.3	5.3	100%	0.3%	8.7	11	-2.3	4.4	99.9%	0.1%
39	2241	17	5.7	16.1	2.9	100%	0.2%	11	11	0	1.7	100%	0%	9.1	5.8	2.3	5.3	100%	0.3%	-7.5	0.1	-7.6	4	1.1%	0%
40	2178	21.8	5.7	16.1	2.9	100%	0%	9.5	11	-1.5	2.2	100%	0.3%	12.2	5.7	6.5	5.2	100%	0%	-8.1	0	-8.1	4.4	0%	0%

Table 2: Simulation Results 1 Part2

		$\sigma^2 = 1$												$\sigma^2 = 4$									
		Linear Model						Semi-parametric Model						Linear Model						Semi-parametric Model					
Time	n	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%
41	2113	24.1	6.1	18	4.7	100%	7.5%	7.8	10	-2.2	3.9	100%	0%	13.9	5.7	8.2	6.9	100%	0%	-8.1	0	-8.1	6.1	0%	0%
42	2023	18.4	5.7	12.7	6.8	100%	0%	11.3	10	1.3	6	100%	0%	10.8	5.8	5.1	9	100%	0%	-7.1	0.3	7.4	8.2	2.7%	0%
43	2007	18.8	5.7	13.1	4.9	100%	0%	9.1	10	-0.9	4.1	100%	0%	10.5	5.7	4.8	7.1	100%	0%	-8.3	0	-8.3	6.3	0%	0%
44	1924	16.6	5.8	10.8	8.18	100%	0.2%	13.6	10.8	2.8	7.5	100%	8%	11.2	5.8	5.4	10.4	100%	0.3%	-3.5	2.7	-6.2	9.7	26.3%	0.2%
45	1990	27.5	5.7	21.8	2.1	100%	0%	8.1	10	-1.9	1.4	100%	0%	13.3	5.7	7.5	4.4	100%	0%	-8	0	-8	3.6	0%	0%
46	1937	20.3	5.8	14.5	5.4	100%	0.9%	19.7	11.8	7.9	4.7	100%	18%	12.6	5.9	6.7	7.6	100%	3%	8	11.2	-3.2	6.8	100%	12.3%
47	1909	13.2	5.7	7.5	5.2	100%	0%	14.2	10.4	3.8	4.5	100%	3.5%	8.8	5.7	3.1	7.4	100%	0%	2.7	8.4	-5.7	6.7	84.9%	0%
48	1872	21.8	5.7	16.1	2.7	100%	0%	11.4	10	1.4	2	100%	0%	11.1	5.8	5.3	4.9	100%	0%	-6.8	0.6	-7.4	4.2	5.7%	0%
49	1841	16.3	5.7	10.5	2.1	100%	0%	8.7	10	-1.3	1.4	100%	0.1%	8.1	5.7	2.4	4.4	100%	0%	-8.4	0	-8.4	3.6	0%	0%
50	1826	11	5.7	5.3	4.3	100%	0%	12.6	10.6	2	3.5	100%	3.5%	6.5	5.7	0.8	6.6	99.7%	0.3%	-6.9	0	-6.9	5.7	0%	0%

This table documents the results under the characteristics-based beta and alpha of Fama-Frech 3 factors model. To mimic the empirical study, we simulated 50 12-month rolling windows, and each window is repeated for 1000 times. Therefore, each column summarises the mean value of 1000 estimations and test results. S_1 is the conventional Wald test while S_0 is the power-strengthened component. This table also compares the performance of both semi-parametric and linear model under different noise level, $\sigma^2 = 1$ and $\sigma^2 = 4$. NA results are due to non-invertible characteristic matrices. "Selected" means the percentage of selecting the relevant characteristic in the mispricing function among 1000 experiments. Similarly, "distortion" represents the percentage of wrongly selecting irrelevant characteristics among 1000 repetitions.

7.3 Robustness Under Stronger Noise

In Table 1, we set two different noise levels of random shocks, namely $\sigma^2 = 1$ and $\sigma^2 = 4$. Although $\sigma^2 = 1$ is closer to the empirical data, we conduct this comparison to show the robustness of our methods. When the noise level becomes three times bigger, the accuracy of power enhanced tests gets much lower for certain windows. However, there are no size distortions under this solid noise level, recalling that all the components of our simulation model are rescaled to be unit variance. Thus, the selection probability of relevant characteristics is affected by the higher noise level, but stronger noise will not cause size distortion under our methodology. Another fact is that the stronger noise does deteriorate the low power problem of conventional Wald tests, leading to an even smaller value of S_1, which can be mitigated through adding S_0.

Therefore, we conclude that our methods are robust to a higher noise level regarding no size distortions. However, the accuracy of selecting relevant components and the role of enhancing the power of hypothesis tests will be influenced negatively.

7.4 Number of Factors

In the empirical study, the number of factors is unknown. Therefore, in this subsection we will study whether our methodology is robust to a various number of factors estimated.

We simulate another data generation process:

$$y_{it} = h(X_i) + \sum_{j=1}^{5} g_j(X_j) f_{jt} + \epsilon_{it}, \tag{7}$$

similarly, where y_{it} is the generated stock's return. $h(X_i)$ is the mispricing function consist of a non-linear characteristic function of X_i, to mimic the sparse structure of a mispricing function. $g_j(\boldsymbol{X}_j)$ is the j^{th} characteristics-based factor loading, which has an additive semi-parametric structure, and X_j is a subset consisting of four characteristics. f_{jt} is the j Fama-French 5-factor returns at time t. ϵ_{it} is the idiosyncratic shock, generated from $N(0, \sigma^2)$. Moreover, we generate characteristic univariate functions as:

$$h(X_i) = \sin X_i,$$

$$g_1(\boldsymbol{X}_1) = X_1^2 + (3X_2^3 - 2X_2^2) + (3X_3^3 - 2X_3) + X_4^2,$$
$$g_2(\boldsymbol{X}_2) = X_5^2 + (3X_6^3 - 2X_6^2) + (3X_7^3 - 2X_7) + X_8^2,$$

$$g_3(\boldsymbol{X}_3) = X_9^2 + (3X_{10}^3 - 2X_{10}^2) + (3X_{11}^3 - 2X_{11}) + X_{12}^2,$$
$$g_4(\boldsymbol{X}_4) = X_{13}^2 + (3X_{14}^3 - 2X_{14}^2) + (3X_{15}^3 - 2X_{15}) + X_{16}^2,$$
$$g_5(\boldsymbol{X}_5) = X_{17}^2 + (3X_{18}^3 - 2X_{18}^2) + (3X_{19}^3 - 2X_{19}) + X_{20}^2,$$

where X_i is a randomly picked characteristic and $i \neq 1, \ldots, 19, 20$. All the $X_1, \ldots, X_{20}$ are chosen from the characteristics of the empirical study without duplication, details can be found in Appendix. Furthermore, all $h(X_i)$, $g_1(\boldsymbol{X}_1)$, $g_2(\boldsymbol{X}_2)$, $g_3(\boldsymbol{X}_3)$, $g_4(\boldsymbol{X}_4)$, and $g_5(\boldsymbol{X}_5)$ are rescaled to be mean 0 and variance 1.

Given the above data generation process, combining with the data generation process in Section 6.1, we test the influence of over and under-estimated number of factors. We now choose the number of estimated factors to be three and five under two different data sets and compare the results in Table 3.

The first category column is the scenario of over-estimated factors. We simulate the data generation process using the Fama-French three factors model but estimate the number of factors to be five. However, this does not cause any serious problems as we can find from the Table 3. For some rolling blocks, the probability of mistakenly selected irrelevant characteristics is slightly higher under over-estimated factor numbers. However, over-estimated factors can increase the model fitting marginally. Therefore, we conclude that overestimating the number of factors does not cause severe size distortion using our methods.

Unfortunately, underestimating the number of factors can lead to very misleading test results. We can conclude this from the last column where we estimate the number of factors to be three while the actual model contains five factors. Compared with the correct estimated results, underestimating causes not only higher MSE, but also higher distortions, which means it is more likely to select irrelevant characteristics. Therefore, in the empirical study we prefer the five factors model rather than the three factors model.

Table 3: Simulation Results 2 Part1

		Number of factors $J=3$												Number of factors $J=5$									
		Number of estimated factors $\hat{J}=5$						Number of estimated factors $\hat{J}=3$						Number of estimated factors $\hat{J}=5$						Number of estimated factors $\hat{J}=3$					
Time	n	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%
1	468	2.6	7	-4.4	5.9	99.6%	0%	-0.5	6.2	-5.7	6	81.2%	0%	4.5	6.9	-2.4	6	97.4%	1.3%	-8.5	0	-8.5	6.9	0%	0%
2	894	6	8	-2	1.5	99.9%	0%	3.4	8	-4.6	1.6	99.9%	0%	5.5	8	-2.5	2.3	100%	0%	-6.5	1	-7.5	3	12.9%	0%
3	1108	12.8	9	3.8	11.4	100%	0.1%	8.6	9	-0.4	11.5	100%	0%	14.3	9	5.3	13.6	100%	0.5%	5.3	9	-3.7	14.1	100%	0.1%
4	1199	13.5	10.3	3.2	9.4	99.8%	0%	9.2	9.1	0.1	9.5	96.8%	4.3%	16.6	11.7	4.9	9.8	99.5%	25.6%	2.3	3	-0.7	10.1	23.5%	10%
5	1333	15.4	9	6.4	1.8	100%	0.1%	10.6	9	1.6	2	100%	0%	16.7	9	7.6	2.7	100%	0%	4.5	9	-4.5	3.4	100 %	0%
6	1409	41.6	17.5	24.1	15.6	100%	51.3%	28.6	12.6	15.9	15.8	100%	28%	58.7	28.3	30.4	13.5	100%	90%	106.1	29.9	76.2	13.2	100%	100%
7	1466	26.8	9	17.8	6.1	100%	0.01%	19.5	9	10.5	6.2	100%	0%	26.3	9	17.3	9.2	100%	0.3%	3.5	9	-5.5	11.7	100%	0%
8	1560	27.6	10	17.6	3	100%	0%	20.3	10	10.3	3.2	100%	0%	30.4	10	20.4	5	100%	0.5%	26.7	24	2.7	6.7	100%	100%
9	1494	31.7	9.1	22.6	3.3	100%	0.7%	23.1	9	14.1	3.5	100%	0%	32.1	9.2	22.9	4.4	100%	1.4%	29.1	18	11.1	4.6	100%	100%
10	1292	22.5	9	13.5	3.4	100%	0.1%	16.2	9	7.2	3.6	100%	0%	26.3	10	16.3	4.3	100%	11.3%	46.7	18	28.7	4.4	100%	100%
11	1393	27.8	9.4	18.4	5.7	100%	4%	20.7	9.1	11.6	5.8	100%	1.1%	30	10.7	19.3	5.6	100%	17.2%	49	29.1	19.9	5.8	100%	100%
12	1340	15.2	9	6.2	1.8	100%	0%	10.6	9	1.6	2	100%	0%	15.2	9	6.2	1.8	100%	0%	4	9	-5	2.7	100%	0%
13	1285	15.4	9	6.4	1.4	100%	0.2%	10.6	9	1.6	1.6	100%	0%	15.1	9	6.1	1.4	100%	0.1%	3.5	9	-4.5	2.5	100%	0%
14	1181	21.9	9	12.9	4.4	100%	0.2%	15.8	9	6.8	4.5	100%	0%	21.4	9	12.4	4.7	100%	0.2%	4.5	9	-4.5	6	100%	0%
15	1110	16.4	9	7.4	8.5	100%	0%	11.5	9	2.5	8.7	100%	0%	17.1	9	8.1	9.8	100%	0.1%	5.3	9	-3.7	10.2	100%	0%
16	1044	13.3	9.1	4.3	17.8	100%	0.8 %	9.9	9	0.9	17.9	100%	0%	14.9	9.2	5.7	17.8	100%	2.1%	40	22.4	17.6	16.8	100%	100%
17	1125	18.7	9	9.7	8.8	100%	0.1%	13.2	9	4.2	9	100%	0%	24.1	9.7	14.4	10.7	100%	7.1%	101.4	27	74.4	10.3	100%	100%
18	2192	31.8	11	20.8	4.1	100%	0.2%	23.2	11	12.2	4.3	100%	0%	69.8	28.6	41.2	5.4	100%	77.6%	563.8	33	530.8	3.6	100%	100%
19	2236	24.4	11	13.4	5.1	100%	0%	17.8	11	6.8	5.2	100%	0%	25.1	11	14.1	5.4	100%	0%	10.2	11	-0.8	6.1	100%	0%
20	2273	29.4	11	18.4	3	100%	0.4%	22.4	11	11.4	3.2	100%	0%	30.3	11.1	19.2	4.2	100%	0.7%	61.1	33	28.1	5.3	100%	100%
21	2235	27.5	11	16.5	1.8	100%	0%	20.2	11	9.2	2	100%	0%	29	11	18	2.2	100%	0.3%	5.9	11	-5.1	3.3	100%	0%
22	2270	24.9	13.7	11.2	1.9	100%	23.9%	17.2	11.6	5.6	2.1	100%	0%	43.2	20.4	22.8	2.3	100%	56.7%	41.6	22.1	19.5	2.1	100%	100%
23	2405	22.5	11	11.5	3.2	100%	0.1%	16.3	11	5.3	3.3	100%	0%	21.7	11	10.7	3.3	100%	0%	10.9	11.9	-1	4.3	100%	7.8%
24	2376	30.6	11	19.6	0.8	100%	0.1%	23.1	11	12.1	1	100%	0%	30.3	11	19.3	1.2	100%	0%	20.4	21.4	-1	2.7	100%	94.8%
25	2323	27.2	11.1	16.1	2.5	100%	0.4%	20.6	11	9.6	2.7	100%	0%	26.8	11	15.8	2.8	100%	0%	8.5	11	-2.5	3.9	100%	0%
26	2344	36.4	16.7	19.7	3.1	100%	51.3%	24.9	12.9	12	3.3	100%	17.1%	36.1	17	19.1	3.2	100%	54%	47.5	23.3	24.2	4.3	100%	100%
27	2434	36.3	11.1	25.2	1	100%	0.9%	27.3	11	16.3	1.2	100%	0%	38.3	11.3	27	1.3	100%	2.6%	89.5	33	56.5	1.7	100%	100%
28	2548	34.5	11	23.5	3.2	100%	0.1%	26.2	11	15.2	3.3	100%	0%	34.8	11	23.8	3.3	0%	0.2%	50.3	22	28.3	4	100%	100%
29	2741	73	12.3	60.7	3.2	100%	10.9%	58.2	11.1	47.1	3.4	100%	1.3%	79.4	15.4	64	3.5	100%	36.8%	439.7	62.7	377	3.6	100%	100%
30	2928	73.9	13.8	60.1	6.1	24.1%	0%	59.2	11.8	47.4	6.3	100%	7.8%	84.6	18.7	65.9	7.4	100%	52.2%	94	32.6	61.4	7.2	100%	100%
31	2894	77.3	16.3	61	5.5	100%	45.4%	61	13.4	47.6	5.7	100%	21.6%	77.2	16.3	60.9	5.5	100%	45.9%	28.6	11	17.6	6.5	100%	0%
32	2905	42.4	12.9	29.5	5	100%	16%	33.2	11.3	21.9	5.2	100%	3%	41.7	12.8	28.9	6.1	100%	15.7%	8.7	11	-2.3	9.4	100%	0%
33	2804	53.8	20.5	33.3	8.8	100%	86.8%	42.7	18.5	24.2	8.9	100%	68.5%	54.1	20.4	33.6	10.1	100%	85.6%	35.5	22	13.5	12.3	100%	100%
34	2570	47.6	14.2	33.4	21.1	27.2%	0%	37.3	12.2	25.1	21.2	100%	10.4%	49.4	14.5	34.9	41.2	100%	28.9%	53.8	22	31.8	38.8	100%	100%
35	2516	50.9	11.3	39.6	7	100%	2.9%	41.3	11	30.3	7.2	100%	0.4%	38.4	11	27.4	18.4	100%	0.4%	51.2	33	18.2	20.8	100%	100%
36	2491	51.3	11.8	39.5	1.3	100%	6.8%	41.3	11	30.3	1.4	100%	0.4%	50.5	11.3	39.2	1.6	100%	3%	15.9	11	4.9	3.3	100%	0%
37	2402	34.4	12.2	22.2	4.7	100%	10.5%	26.5	11.2	15.3	4.9	100%	2.2%	37.4	14.2	23.2	5.1	100%	29.2%	68.8	22	46.8	6.1	100%	0%
38	2326	37.4	12.3	25.1	2.1	100%	10.3%	28.9	11.3	17.6	2.3	100%	2.1%	37.2	12.2	25	2.8	100%	9.4%	44.6	22	22.6	3.4	100%	0%
39	2241	14.8	11	3.8	1.6	100%	0.1%	11	11	0	1.7	100%	0%	14.9	11	3.9	1.7	100%	0%	23.4	22	1.4	2.4	100%	100%
40	2178	13.1	11.1	2	2	100%	1.1%	9.5	11	-1.5	2.2	100%	0.3%	12.9	11.2	1.8	2.2	100%	1.3%	20.4	13.1	7.3	3.4	13.3%	100%

Table 4: Simulation Results 2 Part2

		Number of factors $J=3$												Number of factors $J=5$									
		Number of estimated factors $\hat{J}=5$						Number of estimated factors $\hat{J}=3$						Number of estimated factors $\hat{J}=5$						Number of estimated factors $\hat{J}=3$					
Time	n	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%	S	S_0	S_1	MSE	Selected %	Distortion%
41	2113	11	10	1	3.8	100%	0.2%	7.8	10	-2.2	3.9	100%	0%	11.5	10	1.5	4.5	100%	0.2%	41.1	32.4	8.7	5.4	99.9%	100%
42	2023	15.2	10	5.2	5.9	100%	0%	11.3	10	1.3	6	100%	0%	15.7	10	5.7	6.4	100%	0%	-8	0	-8	9.2	0%	0%
43	2007	12.6	10	2.6	4	100%	0.5%	9.1	10	-0.9	4.1	100%	0%	13.4	10.2	3.2	4.7	100%	1.7%	-0.1	6.4	-6.5	5.6	64.4%	0%
44	1924	19.9	13.1	6.8	7.3	100%	30.7%	13.6	10.8	2.8	7.5	100%	8%	19.5	12.9	6.6	7.5	100%	28.9%	20	20	0	8.3	100%	100%
45	1990	11.4	10	1.4	1.2	100%	0.1%	8.1	10	-1.9	1.4	100%	0%	20.7	14.6	6	1.8	100%	45.2%	116	20	96	1.7	100%	100%
46	1937	27.1	14	13.1	4.5	100%	37.7%	19.7	11.8	7.9	4.7	100%	18%	28.3	14.8	13.5	5.4	100%	45.8%	24.6	20	4.6	6.2	100%	100%
47	1909	19.5	11.7	7.8	4.4	100%	16.1%	14.2	10.4	3.8	4.5	100%	3.5%	24	14	10	4.4	100%	38.1%	51.7	35.2	16.5	5.4	100%	100%
48	1872	15.2	10	5.1	1.8	100%	0.2%	11.4	10	1.4	2	100%	0%	15	10	5	2.1	100%	0.1%	5	10	-5	2.9	100%	0%
49	1841	12.3	10.1	2.2	1.2	100%	1.1%	8.7	10	-1.3	1.4	100%	0.1%	11.8	10.1	1.7	4.4	100%	0.8%	-10	0	-10	4.4	0%	0%
50	1826	18.5	12.4	6.1	3.3	100%	15%	12.6	10.6	2	3.5	100%	3.5%	20.2	13.3	6.9	3.7	100%	19.3%	-3.9	0.4	-4.3	4.4	3.9%	0%

This table presents the results under the characteristics-based beta and alpha of both Fama-Frech 3 and 5 factors model. To mimic the empirical study, we simulated 50 12-month rolling windows, and each window is repeated for 1000 times. Therefore, each column summarises the mean value of 1000 estimations and test results. We compare the different number of estimated factors, namely,$\hat{J}=3$ and $\hat{J}=5$, under both settings. Therefore, both over and under-estimated the number of factors are included. S_1 is the conventional Wald test while S_0 is the power-strengthened component. NA results are due to non-invertible characteristic matrices. "Selected" means the percentage of selecting the relevant characteristic in the mispricing functions among 1000 repetitions. Similarly, "distortion" represents the percentage of wrongly selecting irrelevant characteristics among 1000 experiments.

8 Empirical Study

8.1 Introduction

This section presents the empirical results of short-term mispricing anomalies under the semi-parametric characteristics-based model. We use monthly stocks' returns from CRSP and firms' characteristics from Compustats, from 1965 to 2017. We constructed 33 characteristics following the methods of Freyberger, Neuhierl and Weber (2017)[13]. Details of these characteristics can be found in the appendix. We use characteristics from fiscal year $t-1$ to explain stock returns between July of year t to June of year $t+1$. After adjusting the dates from the balance sheet's data, we merge the two data sets from CRSP and Computats. We require all the firms included in our analysis to have at least three years of data in Compustat.

Data is modified with regards to the following aspects:

1 Delisting is quite common for CRSP data. Therefore, we use the way of Hou, Xue and Zhang (2015)[16] to correct the returns of delisting stocks for all the delisted assets before 2018. Detailed methods can be found in the appendix of this paper.

2 Projected-PCA works well, even under small T circumstances. Thus, we choose the width of our window to be 12 months. Another reason for the short window width is that we assume mispricing functions are time-invariant. Therefore, 12 months blocks are more realistic. One of the limitations of Projected-PCA is that it can only be used for a balanced panel, which means the number of stocks will vary when we applied one-year rolling windows to obtain a short time balanced panel. At the same time, as we treat all the characteristics as time-invariant within each rolling block, we take those characteristics' mean values of 12 months as fixed characteristic values during each period.

3 B-splines are made based on each time-invariant characteristic above among all the n firms which are not delisted within each window.

4 Rolling windows are moving at a 12-month step from Jul. 1967 to Jun. 2017. The first 24 months returns are not included as they do not have corresponding characteristics.

5 Excess returns are constructed by the difference between monthly stock returns and Fama-French risk-free monthly returns, which can be found on their website.

8.2 Estimation

We first construct the characteristic B-spline bases matrices. We choose $v = 0.3$, which means the number of bases for each characteristic within a certain window is $\lfloor n^{0.3} \rceil$, n is the number of stocks in each balanced panel window. To get a considerable large balanced panel in each window, some characteristics with too many missing values are eliminated. Therefore, only 33 characteristics are left, which is also a large set of balance sheet variables compared with other similar research. We substitute each characteristic with their mean values during a window width of 12 months. Also, we construct B-spline bases based on evenly distributed knots, and the degree of each basis is three. Here we can find the dimension of characteristic bases will diverge as the number of firms in each window increases. According to the data we collected, firms that can be kept in a balanced panel vary from 1967 to 2017 and ranging from 468 to 2928, which means both n and $\hat{\mathbf{A}} \in \mathbb{R}^{PH}$ are diverging. Large n can satisfy asymptotic requirements. However, these facts also emphasize the necessity of introducing a power enhanced component into our standard hypothesis test. Furthermore, we build up monthly stock excess return $\mathbf{Y}$ by a $n \times T$ matrix, using y_{it}. Before next step, we use time-demeaned matrix $\mathbf{D_T}$ to demean excess return matrix within each window.

Secondly, we project the time-demeaned monthly excess return matrix $\tilde{\mathbf{Y}}$ to the B-spline space spanned by characteristics $\mathbf{\Phi}(\mathbf{X})$, and then we collect the fitted value $\hat{\mathbf{Y}}$. Moreover, we operate Principle Component Analysis on $\hat{\mathbf{Y}}^\intercal\hat{\mathbf{Y}}$ and obtain $\frac{1}{\sqrt{T}}$ times the first five eigenvectors corresponding to the first five biggest eigenvalues as the estimates of unobservable factors $\hat{\mathbf{F}}$. We choose the number of factors to be five because we prefer the overestimating rather than underestimating, according to simulation results.

Thirdly, we estimate factor loading matrix as:

$$\hat{\mathbf{G}}(\mathbf{X}) = \hat{\mathbf{Y}}\hat{\mathbf{F}}(\hat{\mathbf{F}}^\intercal\hat{\mathbf{F}})^{-1}.$$

Finally, we use equality-constrained OLS estimator to estimate coefficients of the mispricing function. We project excess monthly return on the characteristic space $\mathbf{\Phi}(\mathbf{X})$ that is orthogonal to factor loading matrix $\hat{\mathbf{G}}(\mathbf{X})$.

Another goal of this paper is to conduct a power enhanced test on mispricing functions. Therefore, our final step is to estimate covariance matrix $\hat{\mathbf{\Sigma}}$ of $\hat{\mathbf{A}}$. Methods can be found above.

8.3 Factor estimated

We collect all the 5 factors estimated and calculate their time series correlation with Fama-French 5 factors. We summarize the results below:

Table 5: Correlation Bewteen Estimated Factors and Fama-French 5 Factors

	MKT	SMB	HML	RMW	CMA
Fac1	0.257	-0.046	-0.01	-0.038	0.006
Fac2	0.145	-0.008	0.024	-0.004	-0.048
Fac3	-0.122	0.027	-0.043	0.106	-0.044
Fac4	-0.245	-0.054	0.00	-0.035	0.068
Fac5	-0.170	0.018	-0.072	0.097	-0.008

The table above summarises the correlation between five estimated factors from real data with Fama-French five factors. "MKT", "SMB", "HML", "RMW", and "CMA" are Fama-French five risk factors respectively, denoting "market", "size", "value", "profitability", and "investment" individually.

However, we only detect a relatively high correlation between estimated factors and the "Market" factor from Fama-French data. This may be due to the estimated factors are obtained from projected stocks' returns, which means both noise and some original information are filtered.

8.4 Power enhanced hypothesis tests

In this section, we conduct power enhanced tests on each rolling blocks. Firstly, we set threshold value for each window, $\eta_n = H_n\sqrt{2\log(PH_n)}$, where H_n is the number of bases for each characteristic whereas P is the number of total characteristics in each window, which is equal to 33 in our case. Obviously, η_n is data-driven critical value and diverge as the number of firms increase. We use this indicator function $\mathbf{I}(\sum_{h=1}^{H_n}|\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_n)$ with critical value $\eta_n = H_n\sqrt{2\log(PH_n)}$ to achieve three goals.

1 This indicator function can select the most relevant characteristics that can explain the

variation of the mispricing function, as all the test statistics that exceed the critical value will be given the value 1. Therefore, the results of last column in Table 6 are characteristics selected in $\hat{\mathcal{M}} = \{\hat{\alpha}_p \in \hat{\mathcal{M}} : \sum_{h=1}^{H} |\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_n, \quad h = 1, 2, \ldots, H, \quad p = 1, 2, \ldots, P\}$.

2 It also contributes to the test statistics S_0 by adding values from the most important covariances and let S_0 diverge. As the T is very small in the empirical study, we assume the homoskedasticity among ϵ_{it} to reduce the estimation noise.

3 It can avoid size-distortion by the conservative critical value η_n, which ignores those noises by assigning the value 0.

The diagonal elements of covariance matrix $\hat{\mathbf{\Sigma}}$ are variances of each mispricing bases. These elements can be substituted into the indicator function directly, i.e., $\mathbf{I}(|\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_n)$, where $\hat{\sigma}_{ph}$ is the ph^{th} diagonal element of covariance matrix $\hat{\mathbf{\Sigma}}$.

Finally, the new statistics S can be calculated as:

$$S = S_0 + S_1,$$

$$S_0 = H_n \sum_{p=1}^{P} \mathbf{I}(\sum_{h=1}^{H_n} |\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_n), \quad S_1 = \frac{\hat{\mathbf{A}}\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{A}}^{\intercal} - PH_n}{\sqrt{2PH_n}}.$$

8.5 Test results

This section presents the empirical results. Details can be found in the Table 6, where the table lists 50 rolling windows results from Jul.1967 to Jun.2017. Generally speaking, the number of firms that can be included in a 12 months balanced panel is increasing period by period. The number of our characteristic B-spline bases is a function of the number of firms n within each block, which is $\lfloor n^{0.3} \rceil$. Therefore, the dimension of tested mispricing coefficient vectors $\hat{\mathbf{A}} \in \mathcal{R}^{PH_n}$ is also diverging. This verifies the existence of enhanceable hypothesis tests and the necessity of using power enhanced component S_0, which are consistent with empirical results.

Recalling $\mathbf{S}|\mathbf{H_0} \rightarrow^{\mathbf{d}} \mathbf{N}(\mathbf{0}, \mathbf{1})$, some of the test statistics S are huge enough to reject the null hypothesis stating that the characteristics-based mispricing function has no explanatory power on stocks' excess monthly returns. However, for some testing windows there is no strong signal showing that characteristic mispricing functions exist after subtracting the effects of

common factors, which means all the explained variation of excess stock returns has been included in the part of the common movement, namely $\hat{\mathbf{G}}(\mathbf{X})\hat{\mathbf{F}}^{\intercal}$. Moreover, most S_1 values are minimal and even negative. These may be caused by the sparsity structure of the mispricing function or the low power problems due to diverging dimension of mispricing coefficients.

In terms of the power enhanced component S_0, it works well in our empirical study. On the one hand, it can help us to select the most important explaining characteristics. On the other hand, it strengthens the power of S_1, mitigating the low power problem caused by either sparsity or diverging coefficients. Illustrating examples can be found in Table 6, when S_1 are small or even negative, but the supplements S_0 help S to be significant. Therefore, as the $n \to \infty$, the importance of S_0 will become more obvious.

Apart from contributing to the power of tests, the indicator function in the power enhanced component, namely $S_0 = H_n \sum_{p=1}^{P} \mathbf{I}(\sum_{h=1}^{H} |\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_n)$ can also screen out the most relevant explanatory characteristics that can be used to construct arbitrage characteristic portfolios, which are concluded as "Characteristics Selected".

The last column of Table 6 summaries characteristics that contribute to the S_0 in different time windows. The reasons that those characteristics can explain the mispricings are beyond the scope of this paper. The purposes of this empirical study are mainly to conduct power enhanced hypothesis tests on mispricing functions and select potential characteristics causing the mispricing under a very flexible model.

Some phenomena of empirical findings are worth discussing. Momentum, namely cumulative past returns, is a significant category of characteristics that appears frequently. Although short-term cumulative returns like r_{2_1} are always selected, we cannot take this as evidence of arbitrage opportunities as we treat r_{2_1} as time-invariant and take the mean values of 12-month r_{2_1}. Therefore, it is very likely that higher meaned one month lagged returns associated with higher monthly returns. However, this is not the case for long-term and mid-term cumulative returns' momentum like r_{12_2}, r_{12_7} and r_{6_2}, because the 12-month mean values of these variables contain a lot of information from the past year.

Apart from the cumulative returns, some other characteristics contribute to the arbitrage opportunities as well. PCM (Price to Cost Margin) appears twice, and from **??**, we can find the mispricing curve is nonlinear and generally decreasing as the value of PCM increases. ROA (Return-on-asset) also plays a role during 1988-1989. It behaves like a parabola with fluctuations near the 0 point. Details can be found in Figure 3. As for Lev (ratio of long-term debt and

Table 6: Empirical Study Results

Time period	n	S	S_0	S_1	MSE	Characteristics Selected
Jul.1967-Jun.1968	468	-9.6	0	-9.6	0.005	NONE
Jul.1968-Jun.1969	951	-0.45	8	-8.45	0.004	r_{2_1}
Jul.1969-Jun.1970	1108	1.7	9	-7.3	0.005	r_{2_1}
Jul.1970-Jun.1971	1199	-8.7	0	-8.7	0.006	NONE
Jul.1971-Jun.1972	1333	-10	0	-10	0.004	NONE
Jul.1972-Jun.1973	1409	12.7	18	-5.3	0.005	r_{12_2},r_{6_2}
Jul.1973-Jun.1974	1466	2.1	9	-6.9	0.005	r_{2_1}
Jul.1974-Jun.1975	1560	-10.7	0	-10.7	0.01	NONE
Jul.1975-Jun.1976	1494	0.1	9	8.9	0.05	r_{2_1}
Jul.1976-Jun.1977	1292	0.1	9	-9	0.004	r_{2_1}
Jul.1977-Jun.1978	1393	-9.4	0	-9.4	0.005	NONE
Jul.1978-Jun.1979	1340	8.6	18	-9.4	0.005	r_{2_1},r_{12_7}
Jul.1979-Jun.1980	1285	1	9	-8	0.005	r_{2_1}
Jul.1980-Jun.1981	1181	9.7	18	-8.2	0.006	r_{12_7},r_{12_2}
Jul.1981-Jun.1982	1110	1.2	9	-7.8	0.01	r_{2_1}
Jul.1982-Jun.1983	1044	33.1	36	-3	0.01	r_{12_2},r_{12_7},r_{6_2},r_{2_1}
Jul.1983-Jun.1984	1125	-0.9	9	-9.9	0.006	r_{2_1}
Jul.1984-Jun.1985	2192	-0.2	11	-11.2	0.01	r_{2_1}
Jul.1985-Jun.1986	2236	13.1	22	-8.94	0.01	r_{12_7},r_{12_2}
Jul.1986-Jun.1987	2273	1.7	11	-9.3	0.01	PCM
Jul.1987-Jun.1988	2235	0.9	11	-10.1	0.01	r_{2_1}
Jul.1988-Jun.1989	2270	1.2	11	-9.8	0.01	ROA
Jul.1989-Jun.1990	2405	-0.1	11	-11.1	0.01	r_{2_1}
Jul.1990-Jun.1991	2376	1.1	11	-9.9	0.02	r_{2_1}
Jul.1991-Jun.1992	2323	2.1	11	-8.9	0.02	r_{2_1}
Jul.1992-Jun.1993	2344	12.2	22	-9.8	0.02	r_{12_7},r_{12_2}
Jul.1993-Jun.1994	2434	0.4	11	-10.6	0.01	r_{2_1}
Jul.1994-Jun.1995	2548	2.4	11	-8.6	0.01	r_{2_1}
Jul.1995-Jun.1996	2741	14.1	22	-7.9	0.02	BEME,r_{2_1}
Jul.1996-Jun.1997	2928	18.1	22	-3.9	0.01	BEME,r_{2_1}
Jul.1997-Jun.1998	2894	26.5	33	-6.5	0.02	r_{2_1},r_{12_7},r_{12_2}
Jul.1998-Jun.1999	2905	24.6	33	-8.4	0.02	AT,LME,r_{2_1}
Jul.1999-Jun.2000	2804	13.8	22	-8.2	0.03	r_{2_1},r_{12_7}
Jul.2000-Jun.2001	2570	37.7	44	-6.3	0.02	AT,LME, r_{2_1}, r_{6_2}
Jul.2001-Jun.2002	2516	1.3	11	-9.7	0.02	r_{2_1}
Jul.2002-Jun.2003	2491	15	22	-7	0.02	Lev, r_{2_1}
Jul.2003-Jun.2004	2402	3.9	11	-7.1	0.01	r_{2_1}
Jul.2004-Jun.2005	2326	1.8	11	-9.2	0.01	IPM
Jul.2005-Jun.2006	2241	2.5	11	-8.5	0.01	r_{2_1}
Jul.2006-Jun.2007	2178	1.5	11	-9.5	0.01	r_{2_1}
Jul.2007-Jun.2008	2113	12.6	20	-7.4	0.01	r_{12_2},r_{2_1}
Jul.2008-Jun.2009	2023	1.7	10	-8.3	0.02	r_{2_1}
Jul.2009-Jun.2010	2007	1	10	-9	0.01	r_{2_1}
Jul.2010-Jun.2011	1924	13.6	20	-6.4	0.01	r_{2_1}
Jul.2011-Jun.2012	1990	2.5	10	-7.5	0.01	r_{2_1}
Jul.2012-Jun.2013	1937	23.7	30	-6.3	0.01	r_{2_1},r_{12_7},r_{12_2}
Jul.2013-Jun.2014	1909	2.3	10	-7.7	0.01	r_{2_1}
Jul.2014-Jun.2015	1872	5.5	10	-4.5	0.01	r_{2_1}
Jul.2015-Jun.2016	1841	12.4	20	-7.6	0.01	DelGmSale,r_{2_1}
Jul.2016-Jun.2017	1826	26.1	30	-3.9	0.01	C2D,PCM,r_{12_7}

This table summaries the empirical results, where n represents the number of stocks in this rolling window.

debt in the current liabilities), it is decreasing when Lev<0 while increasing afterwards, see Figure 7. In Figure 8, IPM (pre-tax profit margin) behaves like a "V" shape with the turning point 0 during 2004-2005. DelGmSale (Difference in the percentage in gross margin and the percentage change in sales) experiences a bump at the 0 point during 2015-2016, see Figure 9. C2D curve behaves like "V" around the 0 point in 2016-2017, see Figure 10.

Another finding is the momentum of characteristics. Mispricing covariates can be persistent for two years once appeared, such as BEME (Ratio of the book value of equity and market value of equity) in Figure 4, AT (Total asset) in Figure 6, and LME (Total market capitalization of the previous month) in Figure 5. The long-lasting momentum of past cumulative returns is more significant. We take r_{12_2} and r_{12_7} plots as examples and details can be found in Figure 1.

8.6 A Network of Characteristic Arbitrage Returns

In this section, we illustrate the network of arbitrage returns that are interconnected through assets' characteristics. We apply the methods of section 5 to the results in Table 6. We take Jul. 1986- Jun. 1987 and Jul. 2004- Jun. 2005 as demonstrative examples.

In the rolling window Jul. 1986- Jun. 1987, "PCM" was selected as a mispricing characteristic that can help to explain the arbitrage opportunity. We are arguing that the arbitrage returns are inter-related by the characteristic "PCM." We first divide mispriced returns $\ddot{y}_{it}$ into different return groups. And then, we detect whether there are some clustering structures within groups of highest and lowest arbitrage returns, respectively. As we have 2326 assets, for the visualization purpose, we set the threshold value of the K-means method to be relatively small to have as many as ten groups.

The results are showing below:

As we can see from Table 7, group 2 is outperforming with the largest positive average return, while group 6 is the worst. Therefore, we detect the clusterings of characteristic "PCM" within each group individually, which is the second layer clustering in section 5. We apply the corresponding method and have the following results:

As we can see from Table 8, there are two clusterings of PCM, which can provide extra positive arbitrage returns. However, group 2.2 is regarded as an outlier, which has a very negative PCM value but a high arbitrage return. Members in group 2.1 with excellent arbitrage

Table 7: First layer 1984-1985 (clusterings of $\ddot{y}_{it}$)

Group number	Group centeroid	Group size
1	0.0059	435
2	0.1205	26
3	-0.0082	428
4	0.0399	189
5	0.0697	71
6	-0.1018	29
7	-0.0617	110
8	-0.0390	250
9	-0.0225	349
10	0.0208	386

Table 8: Second layer 1984-1985 (clusterings of characteristic PCM)

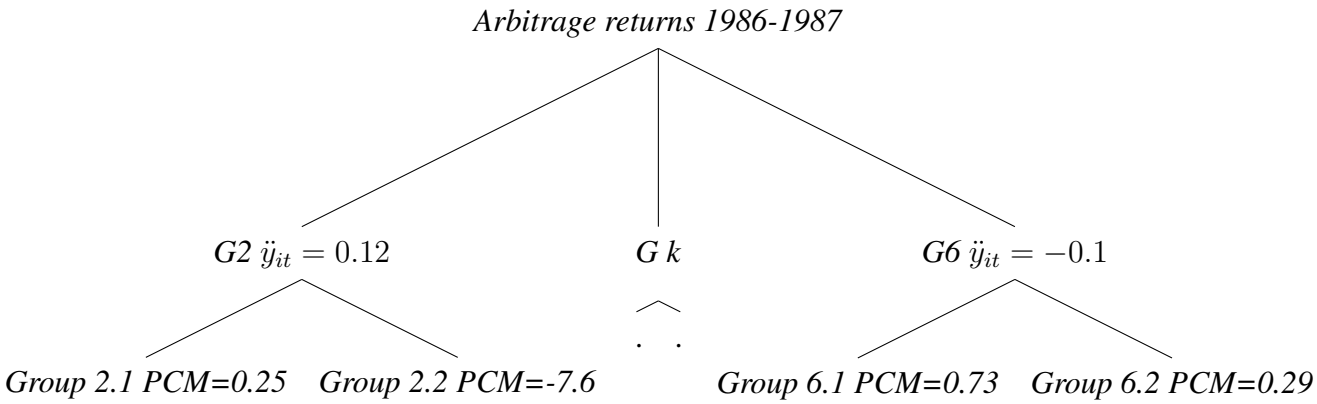
Group number	Centeroids of Arbitrage returns	Centeroids of PCM	Group size
2.1	0.1211	0.2452	25
2.2	0.1039	-7.630	1

Table 9: Second layer 1984-1985 (clusterings of characteristic PCM)

Group number	Centeroids of Arbitrage returns	Centeroids of PCM	Group size
6.1	-0.1085	0.728	9
6.2	-0.0989	0.288	20

performance have positive but small PCM values.

Table 9 illustrates clusterings of PCM within the low arbitrage group, namely group 6. Members of this group are divided into two clusterings. Group 6.1 has a relatively large PCM value, while group 6.2 has a smaller PCM, which is still bigger than that in outperforming group 2.2. The plots of the above classification can be found at Figure 11, where the assets are represented by their "PERMNO," and both axes are rescaled. Therefore, the network of arbitrage returns during Jul. 1986- Jun. 1987 is:



Another example is the arbitrage return $\ddot{y}_{it}$ during the year 2004-2005. Power enhanced screening process selects characteristic "IPM" as the only explanatory variable. Therefore, we detect the network of arbitrage returns $\ddot{y}_{it}$ that are interconnected through "IPM."

Similarly, we apply the hierarchical K-means method. The results of the first layer classification can be found in Table 10. There are ten groups in total according to the similarity of arbitrage returns. And then, we pick two groups with the highest and lowest returns, respectively, to check what is the role of "IPM" playing within these two groups.

Similarly, we have classification results in Table 11 and Table 12. Positive IPM values give higher arbitrage returns. On the contrary, when IPM is closed to zero or very negative, the arbitrage returns both reach the lowest point.

Table 10: First layer (clusterings of $\ddot{y}_{it}$)

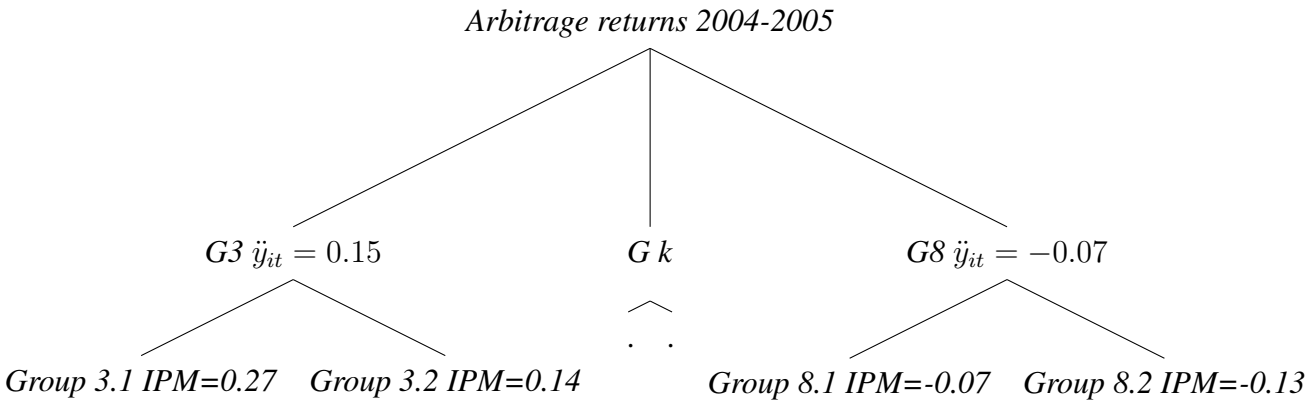
Group number	Group centeroid	Group size
1	0.0421	276
2	0.0059	459
3	0.1537	26
4	-0.024	367
5	0.0659	166
6	0.023	387
7	0.0999	120
8	-0.0758	67
9	-0.0437	244
10	-0.0082	436

Table 11: Second layer (clusterings of characteristic IPM)

Group number	Centeroids of Arbitrage returns	Centeroids of PCM	Group size
3.1	0.1681	0.266	5
3.2	0.1502	0.143	21

Table 12: Second layer (clusterings of characteristic IPM)

Group number	Centeroids of Arbitrage returns	Centeroids of PCM	Group size
8.1	-0.0713	-0.07	10
8.2	-0.1016	-0.134	57



The plots of and IPM can be found Figure 12, where the axes are rescaled and assets are represented by their "PERMNO" code with five digits.

9 Conclusion

We proposed a semi-parametric characteristics-based factor model with dynamic network structures to accommodate both common movements and asset-specific behavior of excess stock returns. We also proposed power enhanced tests to resolve challenges along with model flexibility. Our proposed methods work well in simulations and on the US stock market. Some of the rolling windows show the existence of characteristics-based mispricing functions, which provides us with theoretical evidence to construct arbitrage portfolios by parameterizing those selected security-specific characteristics. We also find the phenomenon of "Characteristics Momentum", which means the mispricing effects of some characteristics can last for several years once they have appeared. Finally, we detect a dynamic network structure among characteristics that provide arbitrage opportunities.

10 Appendix

10.1 Characteristic Description

Name	Description	Reference
A2ME	We define assets-market cap as total assets (AT) over market capitalization as of December t-1. Market capitalization is the product of shares outstanding (SHROUT) and price(PRC).	Bhandari (1988)
AT	Total assets (AT)	Gandhi and Lusting (2015)
ATO	Net sales over lagged net operating assets. Net operating assets are the difference between operating assets and operating liabilities. Operating assets are total assets (AT) minus cash and short-term investments (CHE), minus investment and other advances (IVAO). Operating liabilities are total assets (AT), minus debt in current liabilities(DLC),minus long-term debt (DLTT),minus minority interest (MIB), minus preferred stock (PSTK), minus common equity (CEQ).	Soliman(2008)
BEME	Ratio of book value of equity to market value of equity. Book equity is shareholder equity (SH) plus deferred taxes and investment tax credit (TXDITC), minus preferred stock (PS). SH is shareholder's equity (SEQ). If missing, SH is the sum of common equity (CEQ) and preferred stock (PS). If missing, SH is the difference between total assets (AT) and total liabilities (LT). Depending on availability, we use the redemption (item PSTKRV), liquidating (item PSTKL), or par value (item PSTK) for PS. The market value of equity is as of December t-1. The market value of equity is the product of shares outstanding (SHROUT) and price (PRC).	Rosenberg, Reid and Lanstein (1985) Davis, Fama, and French (2000)

C	Ration of cash and short-term investments (CHE) to total assets (AT)	Palazzo
C2D	Cash flow to price is the ratio of income and extraoridinary items (IB) and depreciation and amortization (dp) to total liabilities (LT).	
CTO	We define caoital turnover as ratio of net sales (SALE) to lagged total assets (AT).	Haugen and Baker (1996)
Debt2P	Debt to price is the radio of long-term debt (DLTT) and debt in current liabilities (DLC) to the market capitalization as of December t-1 . Market capitalization is the product of shares outstanding (SHROUT) and price (PRC).	Litzenberger and Ramaswamy (1979)
Δceq	The percentage change in the book value of equity (CEQ).	Richardson et al. (2005)
$\Delta(\Delta Gm - Sales)$	The difference in the percentage change in gross margin and the percentage change in sales (SALE). We define gross margin as the difference in sales (SALE) and costs of goods sold (COGS).	Abarbanell and Bushee (1997)
$\Delta Shrout$	The definition of the percentage change in shares outstanding (SHROUT).	Pontiff and Woodgate (2008)
$\Delta PI2A$	We define the change in property, plants ,and equipment as changes in property,plants,and equipment (PPEGT) and inventory (INVT) over lagged total assets (TA).	Lyandres , Sun, and Zhang (2008)
DTO	We define turnover as ratio of daily volume (VOL) to shares outstanding (SHROUT) minus the daily market turnover and de-trend it by its 180 trading day median. We scale down the volume of NASDAQ securities by 38% after 1997 and by 50% before that to address the issue of double-counting of volume for NASDAQ securities.	Garfinkel (2009); Anderson and Dyl (2005)

E2P	We define earnings to price as the ratio of income before extraordinary items (IB) to the market capitalization as December t-1 Market capitalization is the product of share outstanding (SHROUT) and price (PRC).	Basu (1983)
EPS	We define earnings per share as the ratio of income before extraordinary items (IB) to share outstanding (SHROUT) as of December t-1	Basu (1997)
Investment	We define investment as the percentage year-on-year growth rate in total assets (AT).	Cooper, Gulen, and Schill(2008)
IPM	We define pre-tax profit margin as ratio of pre-tax income (PI) to sales (SALE).	
Lev	leverage is the ratio of long-term debt (DLTT) and debt in the current liabilities (DLC) to the sum of long-term debt, debt in current liabilities, and stockholders' equity (SEQ)	Lewenllen (2015)
Turnover	Turnover is last month's volume (VOL) over shares outstanding (SHROUT).	Datar, Naik, and Radcliffe (1998)
OL	Operating leverage is the sum of cost of goods sold (COGS) and selling, general, and administrative expenses (XSGA) over total assets.	Novy-Marx (2011)
PCM	The price-to-cost margin is the difference between net sales (SALE) and costs of goods sold (COGS) divided by net sales (SALE).	Gorodnichenko and Weber (2016) and D'Acunto, Liu, Pflucger, and Wcber (2017)
PM	The profit margin is operating income after depreciation (OIADP) over sales (SALE)	Soliman (2008)
Q	Tobin's Q is total assets (AT), the market value of equity (SHROUT times PRC) minus cash and short-term investments (CEQ) minus deferred taxes (TXDB) scaled by total assets (AT).	
ROA	Return-on-assets is income before extraordinary items (IB) to lagged total assets (AT).	Balakrishnan, Bartov, and Faurel (2010)

ROC	ROC is the ratio of market value of equity (ME) plus long-term debt (DLTT)minus total assets to Cash and Short-Term Investments (CHE).	Chandrashekar and Rao (2009)
ROE	Return-on-equity is income before extraordinary items (IB) to lagged book-value of equity.	in Haugen and Baker (1996)
r_{12-2}	We define momentum as cumulative return from 12 months before the return prediction to two months before.	Fama and French (1996)
r_{12-7}	We define intermediate momentum as cumulative return from 12 months before the return prediction to seven months before.	Novy-Marx (2012)
r_{6-2}	We definer_{6-2} as cumulative return from 6 months before the return prediction to two months before.	Jegadeesh and Titman (1993)
r_{2-1}	We define short-term reversal as lagged one-month return.	Jegadeesh(1990)
S2C	Sales-to-cash is the ratio of net sales (SALE) to Cash and Short-Term Investments (CHE).	following Ou and Penman (1989)
Sales-G	Sales growth is the percentage growth rate in annual sales (SALE).	Lakonishok, Shleifer , and Vishmy (1994)
SAT	We define asset turnover as the ratio of sales (SALE) to total assets (AT).	Soliman (2008)
SGA2S	SGA to sales is the ratio of selling ,general and administrative expenses (XSGA) to net sales (SALE).	

Table 13: Characteristic Details

10.2 Proofs

Through out the proofs, we have the number of observations $n \to \infty$ and time T is fixed.

Proof of Theorem 6.1 : In equation 5, we have

$$\mathbf{Y} = (\mathbf{\Phi(X)A} + \mathbf{\Gamma} + \mathbf{R}^{\mu}(\mathbf{X}))\mathbf{1}_{\mathbf{T}}^{\intercal} + (\mathbf{\Phi(X)B} + \mathbf{\Lambda} + \mathbf{R}^{\theta}(\mathbf{X}))\mathbf{F}^{\intercal} + \mathbf{U},$$

After multiplying time demeaned matrix $\mathbf{D_T}$, where $\mathbf{D_T} = \mathbf{I_T} - \frac{1}{T}\mathbf{1}_{\mathbf{T}}'\mathbf{1_T}$, as the mispricing components are time-invariant, therefore we obtain:

$$\mathbf{YD_T} = (\mathbf{\Phi(X)B} + \mathbf{\Lambda} + \mathbf{R}^{\theta}(\mathbf{X}))\mathbf{F}^{\intercal}\mathbf{D_T} + \mathbf{UD_T},$$

Onwards, we define $\mathbf{YD_T} = \tilde{\mathbf{Y}}$ and $\mathbf{F}^{\intercal} = \mathbf{F}'\mathbf{D_T}$. Demeaned time factors do not change their properties.

And then multiple both sides by $\mathbf{P} = \Phi(\mathbf{X})(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})^{\intercal}$,

$$\hat{\mathbf{Y}} = (\mathbf{\Phi(X)B} + \mathbf{P\Lambda} + \mathbf{PR}^{\theta}(\mathbf{X}))\mathbf{F}^{\intercal} + \mathbf{PUD_T}.$$

And then, we decompose :

$$\hat{\mathbf{Y}} = \mathbf{\Phi(X)BF}^{\intercal} + \mathbf{P\Lambda F}^{\intercal} + \mathbf{PUD_T} + \boldsymbol{PR^{\theta}(X)F}^{\intercal} = \boldsymbol{e_1} + \boldsymbol{e_2} + \boldsymbol{e_3} + \boldsymbol{e_4},$$

as $n \to \infty,$ therefore $n^v \to \infty$, approximation error $\mathbf{R}^{\theta}(\mathbf{X}) \to_P \mathbf{0}$, see Huang, Horowitz and Wei (2010)[17], thus, under orthogonal bases, $\mathbf{e}_{\mathbf{4}}^{\intercal} \to^P \mathbf{0}$. Therefore,

$$\frac{1}{n}\hat{\mathbf{Y}}^{\intercal}\hat{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{3}\boldsymbol{e_i^{\intercal}e_j}.$$

Under Assumption 1, we have following results:

for $\frac{1}{n}\sum_{j=1}^{3}\boldsymbol{e_2^{\intercal}e_j}$,

$$\frac{1}{n}\boldsymbol{P\Lambda} \to^P \mathbf{0},$$

therefore,

$$\frac{1}{n}\sum_{j=1}^{3}\boldsymbol{e_2^{\intercal}e_j} + \frac{1}{n}\sum_{j=1}^{3}\boldsymbol{e_j^{\intercal}e_2} \to^P \mathbf{0}.$$

for $\frac{1}{n}\sum_{j=1}^{3} \boldsymbol{e_3^\intercal e_j}$,

$$\frac{1}{n}\boldsymbol{PU} \to^P \mathbf{0},$$

therefore,

$$\frac{1}{n}\sum_{j=1}^{3} \boldsymbol{e_2^\intercal e_j} + \frac{1}{n}\sum_{j=1}^{3} \boldsymbol{e_j^\intercal e_2} \to^P \mathbf{0}.$$

And the only $\frac{1}{n}\boldsymbol{e_1^\intercal e_1}$ left, namely,

$$\frac{1}{n}\boldsymbol{e_1^\intercal e_1} = \boldsymbol{F}\frac{\boldsymbol{B^\intercal \Phi^\intercal(X)\Phi(X)B}}{\boldsymbol{n}}\boldsymbol{F^\intercal}.$$

Therefore, under Assumption 2-4, and fixed T. A much smaller $T \times T$ matrix $\frac{1}{n}\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ can be sovled by asymtotic principal component by Connor and Korajczyk (1986). Estimating $\hat{\mathbf{F}} = \frac{1}{\sqrt{T}}\{\psi_1, \psi_2, \ldots, \psi_J\}$, where$\{\psi_1, \psi_2, \ldots, \psi_J\}$ are eigenvectors corresponding with the first J eigenvalues of $\frac{1}{n}\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$.

Thus, $\hat{\mathbf{F}} \to_P \mathbf{F}$ follows. □

Proof of Theorem 6.2 : Given $\hat{\mathbf{F}}$, we have:

$$\hat{\mathbf{G}}(\mathbf{X}) = \hat{\mathbf{Y}}\hat{\mathbf{F}}(\hat{\mathbf{F}}'\hat{\mathbf{F}})^{-1},$$

as $\hat{\mathbf{F}}'\hat{\mathbf{F}} = \mathbf{I_J}$, therefore,

$$\hat{\mathbf{G}}(\mathbf{X}) = \tilde{\mathbf{Y}}\hat{\mathbf{F}}.$$

Then we need to show:

$$E((\hat{\mathbf{G}}(\mathbf{X_i}) - \mathbf{G}(\mathbf{X_i}))^{\mathbf{2}}) = 0.$$

Take the sample analogue,

$$\frac{1}{n}((\hat{\mathbf{G}}(\mathbf{X}) - \mathbf{G}(\mathbf{X})))^\intercal((\hat{\mathbf{G}}(\mathbf{X}) - \mathbf{G}(\mathbf{X}))).$$

Given:

$$\mathbf{G}(\mathbf{X}) = \mathbf{\Phi}(\mathbf{X})\mathbf{B} + \mathbf{R}^\theta(\mathbf{X}).$$

$$\hat{\mathbf{G}}(\mathbf{X}) = (\mathbf{\Phi}(\mathbf{X})\mathbf{B} + \mathbf{P\Lambda} + \mathbf{PR}^\theta(\mathbf{X}))\mathbf{F}^\intercal\hat{\mathbf{F}} + \mathbf{PUD_T}\hat{\mathbf{F}}$$

Therefore,

$$\mathbf{G(X)} - \hat{\mathbf{G}}\mathbf{(X)} = (\mathbf{\Phi(X)B} + \mathbf{P\Lambda} + \mathbf{PR}^{\theta}\mathbf{(X)})\mathbf{F}^{\intercal}\hat{\mathbf{F}} + \mathbf{PUD_T}\hat{\mathbf{F}} - \mathbf{\Phi(X)B} - \mathbf{R}^{\theta}\mathbf{(X)} = \mathbf{q_1} + \mathbf{q_2} + \mathbf{q_3} + \mathbf{q_4}.$$

As shown in section 10.2,

$$\frac{1}{n}((\hat{\mathbf{G}}\mathbf{(X)} - \mathbf{G(X)}))^{\intercal}((\hat{\mathbf{G}}\mathbf{(X)} - \mathbf{G(X)})) \to^{P} \frac{1}{n}\mathbf{q_1^{\intercal}q_1} + \frac{\mathbf{1}}{\mathbf{n}}\mathbf{q_3^{\intercal}q_3} + \frac{\mathbf{1}}{\mathbf{n}}\mathbf{q_1^{\intercal}q_3} + \frac{\mathbf{1}}{\mathbf{n}}\mathbf{q_3^{\intercal}q_1}.$$

Therefore,

$$\frac{1}{n}\mathbf{q_1^{\intercal}q_1} = \hat{\mathbf{F}}^{\intercal}\mathbf{F}(\mathbf{\Phi(X)B} + \mathbf{P\Lambda} + \mathbf{PR}^{\theta}\mathbf{(X)})^{\intercal}(\mathbf{\Phi(X)B} + \mathbf{P\Lambda} + \mathbf{PR}^{\theta}\mathbf{(X)})\mathbf{F}^{\intercal}\hat{\mathbf{F}},$$

due to

$$\frac{1}{n}\sum_{j=1}^{3}\boldsymbol{e_2^{\intercal}e_j} + \frac{1}{n}\sum_{j=1}^{3}\boldsymbol{e_j^{\intercal}e_2} \to^{P} \mathbf{0},$$

and

$$\frac{1}{n}\mathbf{e_1^T e_1} \to^{P} \boldsymbol{F}\frac{\boldsymbol{B^{\intercal}\Phi^{\intercal}(X)\Phi(X)B}}{\boldsymbol{n}}\boldsymbol{F^{\intercal}}$$

then,

$$\frac{1}{n}\mathbf{q_1^T q_1} \to^{P} \hat{\boldsymbol{F}}^{\intercal}\boldsymbol{F}\frac{\boldsymbol{B^{\intercal}\Phi^{\intercal}(X)\Phi(X)B}}{\boldsymbol{n}}\boldsymbol{F^{\intercal}}\hat{\boldsymbol{F}}.$$

Under Theorem 6.1 and Assumption 2, which gives $\hat{\mathbf{F}} \to \mathbf{F}$ and $\mathbf{F^T F} = \mathbf{I_J}$:

$$\frac{1}{n}\mathbf{q_1^T q_1} \to^{P} \frac{\boldsymbol{B^{\intercal}\Phi^{\intercal}(X)\Phi(X)B}}{\boldsymbol{n}},$$

Similarly,

$$\frac{1}{n}\mathbf{q_3^T q_3} \to^{P} \frac{\boldsymbol{B^{\intercal}\Phi^{\intercal}(X)\Phi(X)B}}{\boldsymbol{n}},$$
$$\frac{1}{n}\mathbf{q_1^T q_3} \to^{P} -\frac{\boldsymbol{B^{\intercal}\Phi^{\intercal}(X)\Phi(X)B}}{\boldsymbol{n}},$$
$$\frac{1}{n}\mathbf{q_3^T q_1} \to^{P} -\frac{\boldsymbol{B^{\intercal}\Phi^{\intercal}(X)\Phi(X)B}}{\boldsymbol{n}}.$$

Therefore,

$$\frac{1}{n}\mathbf{q_1^{\intercal}q_1} + \frac{\mathbf{1}}{\mathbf{n}}\mathbf{q_3^{\intercal}q_3} + \frac{\mathbf{1}}{\mathbf{n}}\mathbf{q_1^{\intercal}q_3} + \frac{\mathbf{1}}{\mathbf{n}}\mathbf{q_3^{\intercal}q_1} \to 0,$$

thus,

$$\hat{\mathbf{G}}\mathbf{(X)} - \mathbf{G(X)} \to^{P} \mathbf{0}$$

□

Proof of Theorem 6.3 : Let $\dot{\boldsymbol{Y}} = \frac{1}{\boldsymbol{T}}(\boldsymbol{Y} - \hat{\boldsymbol{G}}(\boldsymbol{X})\hat{\boldsymbol{F}})\mathbf{1}_{\boldsymbol{T}}$. By substituting the restriction, we have the Lagrangian equation:

$$\min_{\boldsymbol{A}}(\dot{\boldsymbol{Y}} - \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A})'(\dot{\boldsymbol{Y}} - \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A}) + \boldsymbol{\lambda}\hat{\boldsymbol{G}}'(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A} \tag{8}$$

Then we take the first order condition with repect to $\boldsymbol{A}$ and $\boldsymbol{\lambda}$ separately, the we obtain:

$$\begin{pmatrix} 2\boldsymbol{\Phi}'(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X}) & \boldsymbol{\Phi}'(\boldsymbol{X})\hat{\boldsymbol{G}}(\boldsymbol{X}) \\ \hat{\boldsymbol{G}}(\boldsymbol{X})'\boldsymbol{\Phi}'(\boldsymbol{X}) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{A}} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} 2\boldsymbol{\Phi}'(\boldsymbol{X})\dot{\boldsymbol{Y}} \\ \mathbf{0} \end{pmatrix}. \tag{9}$$

Under Assumption 2, the above matrice are invertible, which can be written as:

$$\begin{pmatrix} \hat{\boldsymbol{A}} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} 2\boldsymbol{\Phi}'(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X}) & \boldsymbol{\Phi}'(\boldsymbol{X})\hat{\boldsymbol{G}}(\boldsymbol{X}) \\ \hat{\boldsymbol{G}}(\boldsymbol{X})'\boldsymbol{\Phi}'(\boldsymbol{X}) & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} 2\boldsymbol{\Phi}'(\boldsymbol{X})\dot{\boldsymbol{Y}} \\ \mathbf{0} \end{pmatrix}. \tag{10}$$

Therefore, we obtain:

$$\hat{\mathbf{A}} = \mathrm{M}\tilde{\mathbf{A}},$$

where

$$\mathrm{M} = \mathbf{I} - (\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})'\hat{\mathbf{G}}(\mathbf{X})(\hat{\mathbf{G}}(\mathbf{X})'\hat{\mathbf{G}}(\mathbf{X}))^{-1}\hat{\mathbf{G}}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}),$$

$$\tilde{\mathbf{A}} = \frac{\mathbf{1}}{\mathbf{T}}(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})'\dot{\boldsymbol{Y}}\mathbf{1}_{\mathbf{T}}.$$

Furthermore, let $\boldsymbol{\Xi} = \boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{A}} - h(\boldsymbol{X}) = \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{M}\tilde{\boldsymbol{A}} - \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A} - \boldsymbol{R}^{\mu}(\boldsymbol{X})$, and under the restriction $\hat{\boldsymbol{G}}'(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A} = \mathbf{0}$, we can obtain:

$$\boldsymbol{\Xi} = \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{M}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})'\frac{\mathbf{1}}{\mathbf{T}}(\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A}+\boldsymbol{R}^{\mu}(\boldsymbol{X})+\boldsymbol{\Gamma}+(\boldsymbol{\Lambda}+\boldsymbol{R}^{\theta}(\boldsymbol{X}))\boldsymbol{F}')\mathbf{1}_{\mathbf{T}}-\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A}-\boldsymbol{R}^{\mu}(\boldsymbol{X}). \tag{11}$$

Furthermore, we have:

$$\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{M}(\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\Phi}(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})' = (\mathbf{I}-(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1}\boldsymbol{\Phi}(\mathbf{X})'\hat{\mathbf{G}}(\mathbf{X})(\hat{\mathbf{G}}(\mathbf{X})'\hat{\mathbf{G}}(\mathbf{X}))^{-1}\hat{\mathbf{G}}(\mathbf{X})')\mathbf{P}. \tag{12}$$

And then, substituteEquation 12 into Equation 11 and under Assumption 1, Theorem 6.2:

$$\boldsymbol{\Xi} = \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A} - \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A} - \boldsymbol{R}^{\mu}(\boldsymbol{X}),$$

therefore,

$$\frac{1}{n}\boldsymbol{\Xi}^{\intercal}\boldsymbol{\Xi} \to \mathbf{0}.$$

And the Theorem 6.3 follows. □

Proof of Theorem 6.4 : Define $Z = \max_{\{1\leqslant p\leqslant P, 1\leqslant h\leqslant H_n\}} \{|\hat{\alpha}_{ph}|/\hat{\sigma}_{ph}\}$. Under Assumption 3.2.3, we have

$$\hat{\alpha}_{ph}/\hat{\sigma}_{ph}|\mathbf{H_0} \rightarrow^d N(0,1).$$

Therefore, under the $\mathbf{H_0}$, we have:

$$\begin{aligned} e^{tE(Z)} &\leqslant E[e^{tZ}] \\ &= E[max\{t|\hat{\alpha}_{ph}|/\hat{\sigma}_{ph}\}] \\ &\leqslant \sum_{p=1,h=1}^{p=P,h=H_n} E[e^{t|\hat{\alpha}_{ph}|/\hat{\sigma}_{ph}}] \\ &= ne^{t^2/2}. \end{aligned}$$

Then take the logarithm of both sides we can obtain:

$$E[Z] \leqslant \frac{\log n}{t} + \frac{t}{2}.$$

If we set $t = \sqrt{2\log n}$ to minimise $\frac{\log n}{t} + \frac{t}{2}$, then we have:

$$E[Z] \leqslant \sqrt{2\log n}.$$

Therefore, we can bound the $|\hat{\alpha}_{ph}|/\hat{\sigma}_{ph}$ by $\sqrt{2\log n}$. □

Proof of Theorem 6.5 : To proof

$$\inf_{\mathbf{A}\in\mathcal{A}} P(\text{reject } H_0|\mathbf{A}) \rightarrow 1,$$

equivalently, we need to prove

$$\inf_{\mathbf{A}\in\mathcal{A}} P(S_0 + S_1 > F_q|\mathbf{A}) \rightarrow 1.$$

$S_0 = H_n \sum_{p=1}^{P} \mathbf{I}(\sum_{h=1}^{H_n} |\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_n)$, as $H_n = n_v \rightarrow \infty$ as $n \rightarrow \infty$.

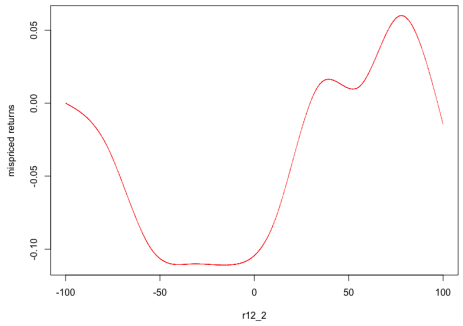
Under Theorem 6.4 and $n \rightarrow \infty$, we have:

$$E(S_0|\mathbf{A}) \rightarrow \infty.$$

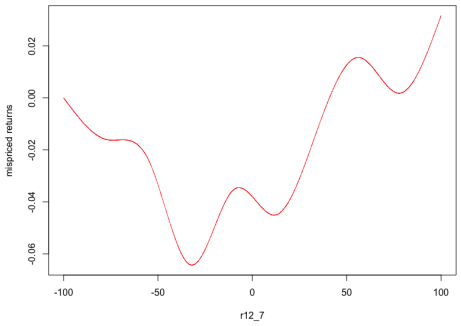
Meanwhile $F_q = O(1)$, we can show that:

$$\inf_{\mathbf{A}\in\mathcal{A}} P(S_0 + S_1 > F_q|\mathbf{A}) \rightarrow 1.$$

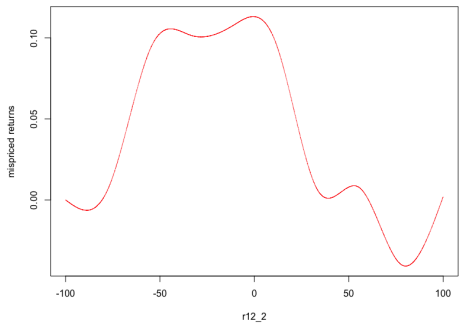
Then the Theorem 6.5 follows. □



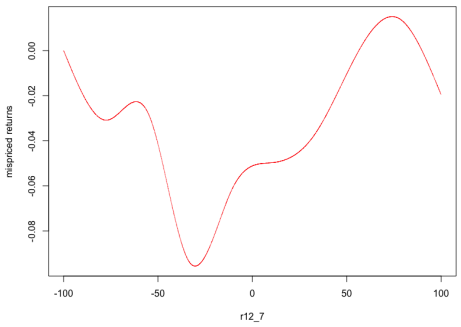
(a) r_{12-2} Curve 1972-1973



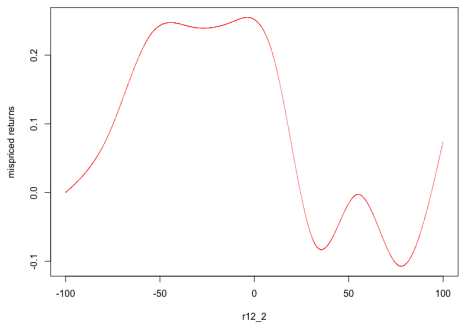
(b) r_{12-7} Curve 1978-1979



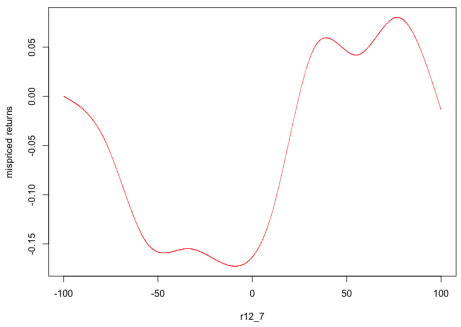
(c) r_{12-2} Curve 1980-1981



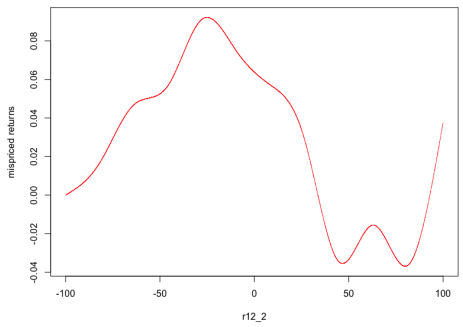
(d) r_{12-7} Curve 1985-1986



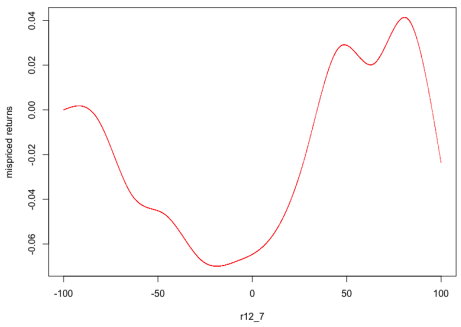
(e) r_{12-2} Curve 1982-1983



(f) r_{12-7} Curve 1982-1983

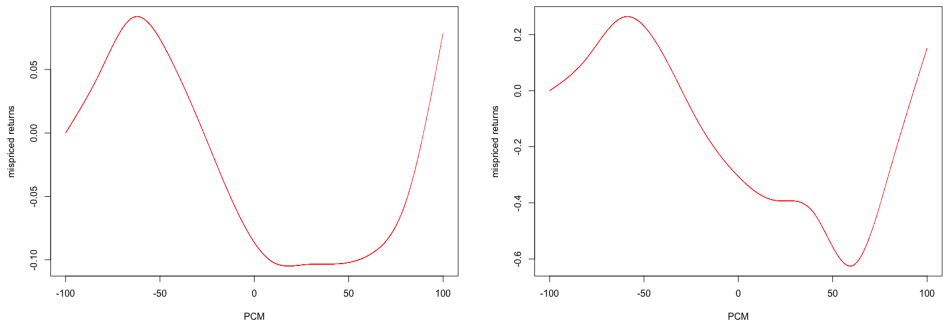


(g) r_{12-2} Curve 1985-1986



(h) r_{12-2} Curve 1985-1986

Figure 1: Mispricing Characteristic r_{12-2} and r_{12-7}



(a) PCM Curve 1984-1985 (b) PCM Curve 2016-2017

Figure 2: Mispricing Characteristic PCM

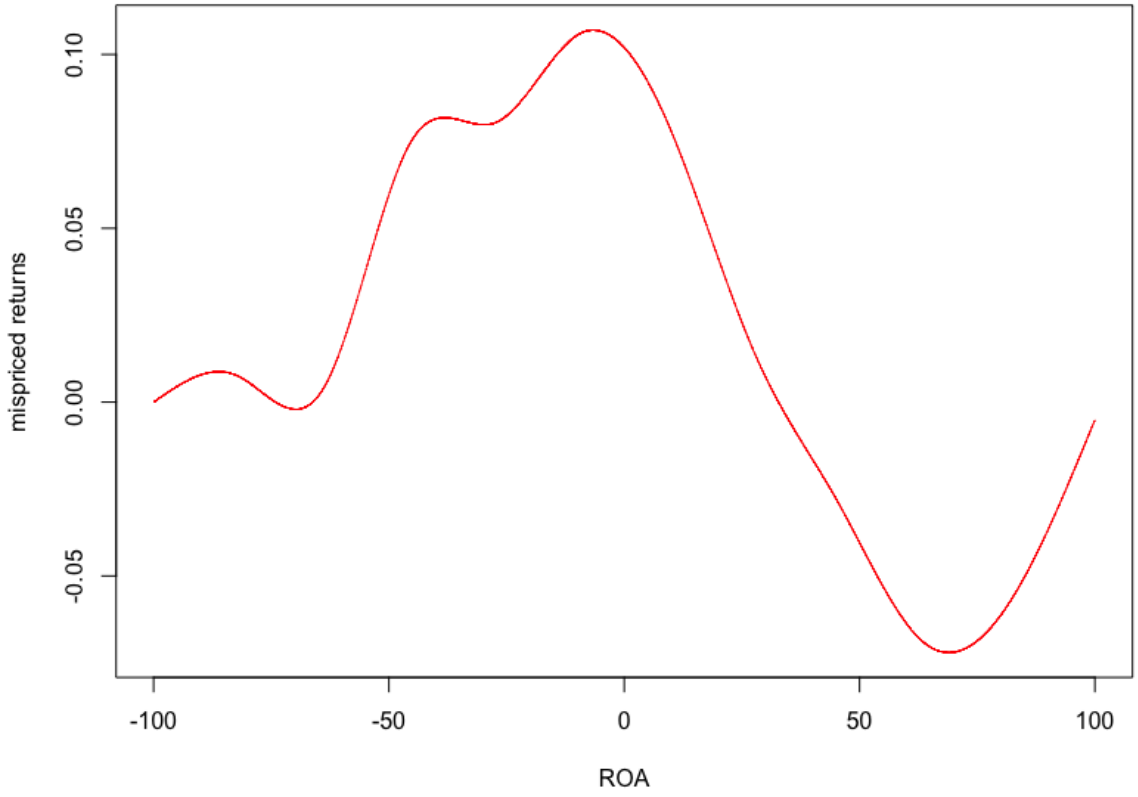
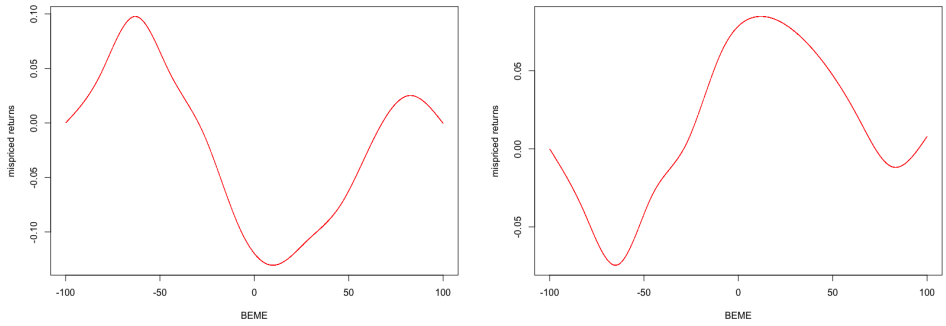
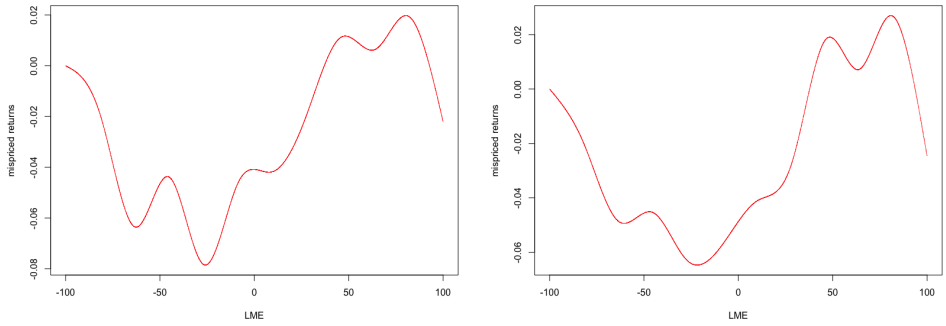


Figure 3: ROA Curve in 1988-1989



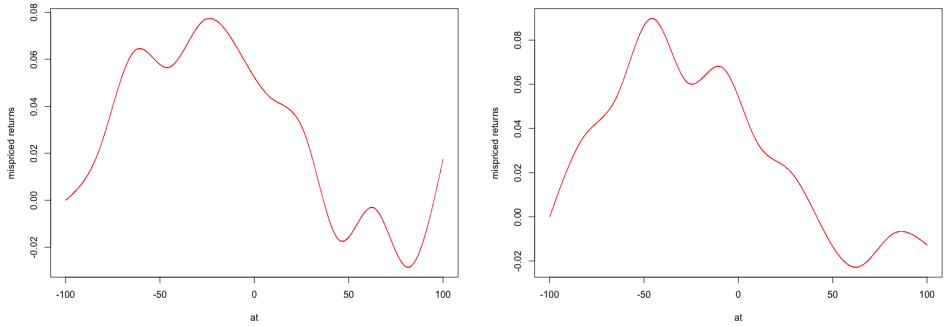
(a) BEME Curve 1995-1996 (b) BEME Curve 1996-1997

Figure 4: Mispricing Characteristic BEME



(a) LME Curve 1998-1999 (b) LME Curve 2000-2001

Figure 5: Mispricing Characteristic BEME



(a) AT Curve 1998-1999 (b) AT Curve 2000-2001

Figure 6: Mispricing Characteristic AT

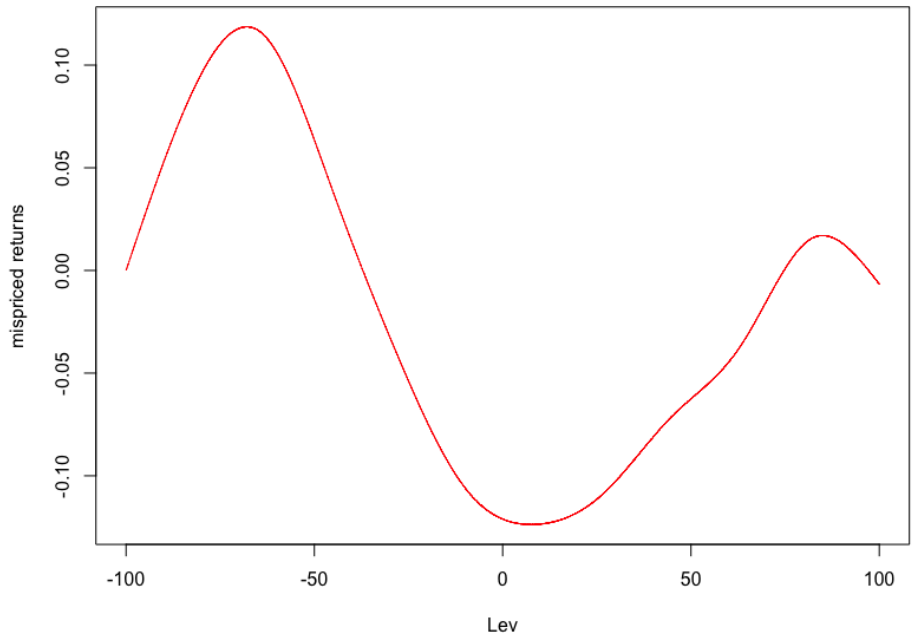


Figure 7: LEV Curve in 2002-2003

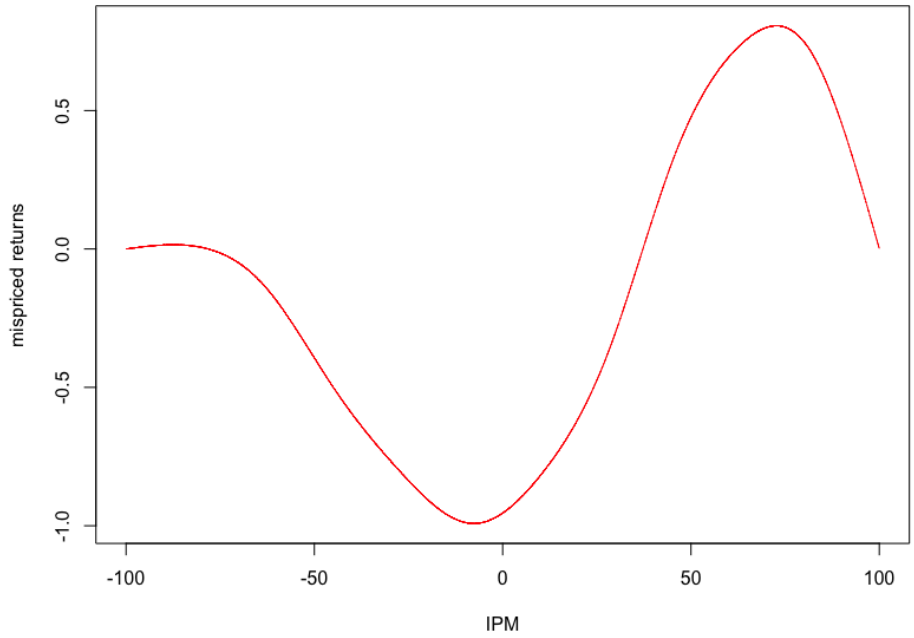


Figure 8: IPM Curve in 2004-2005

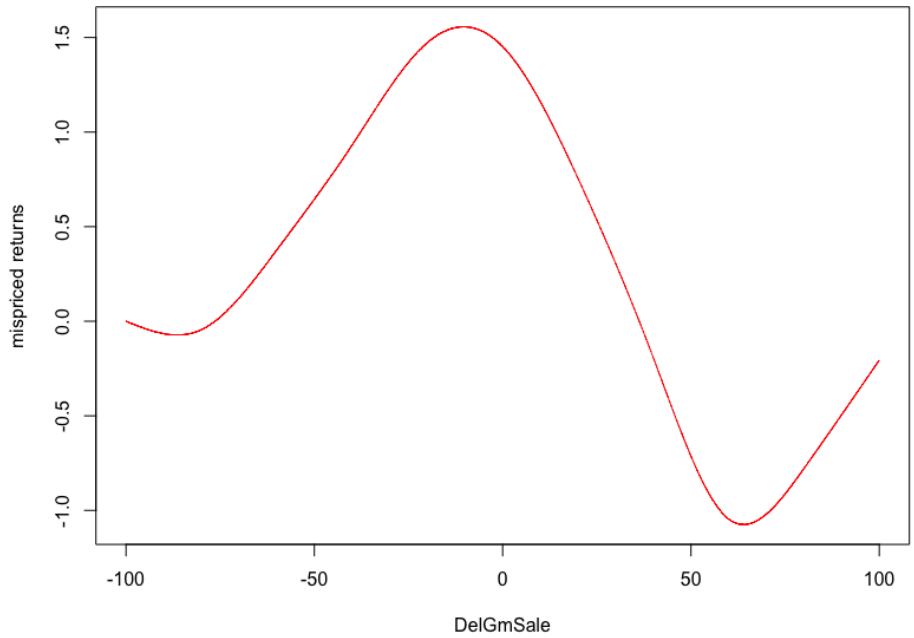


Figure 9: DelGmSale Curve in 2015-2016

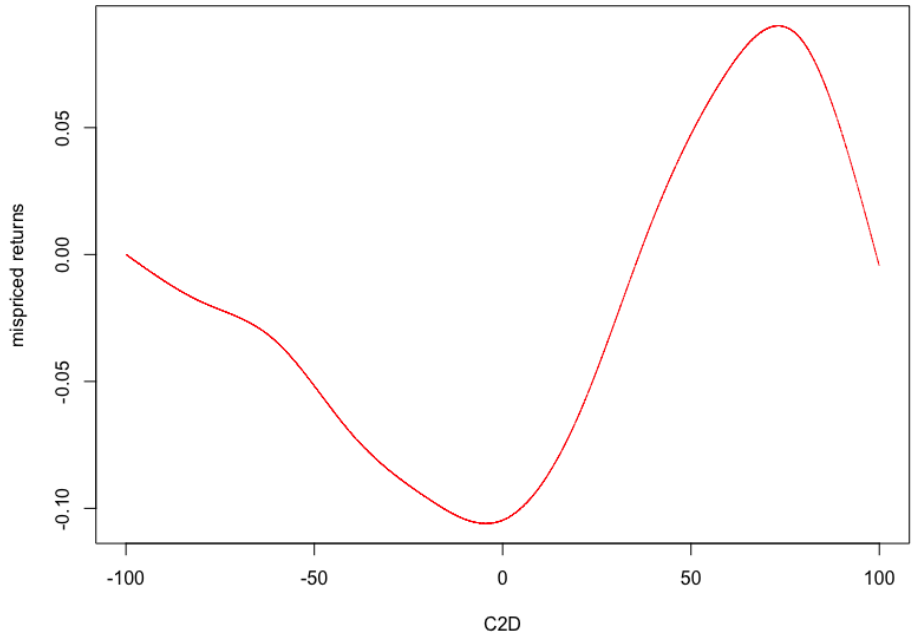
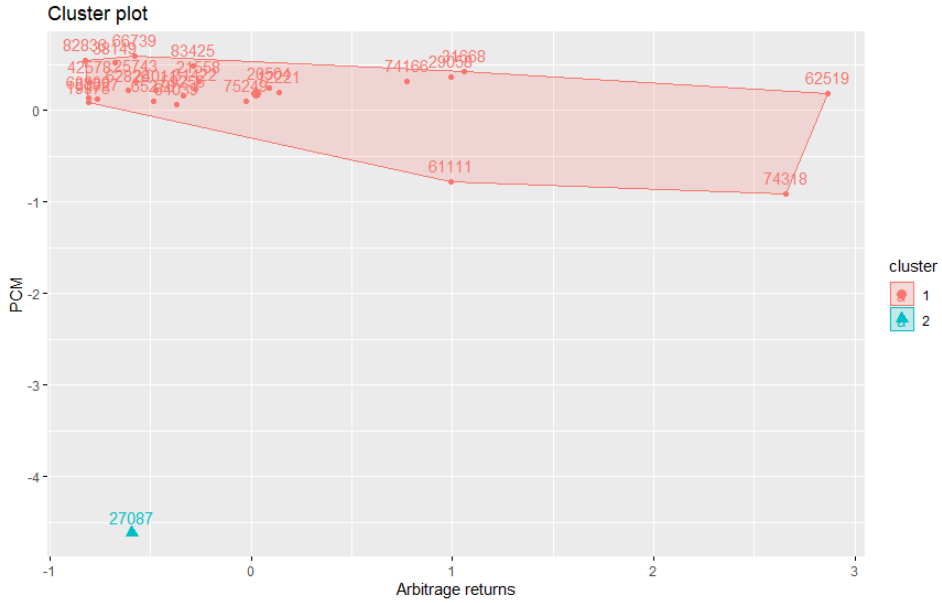
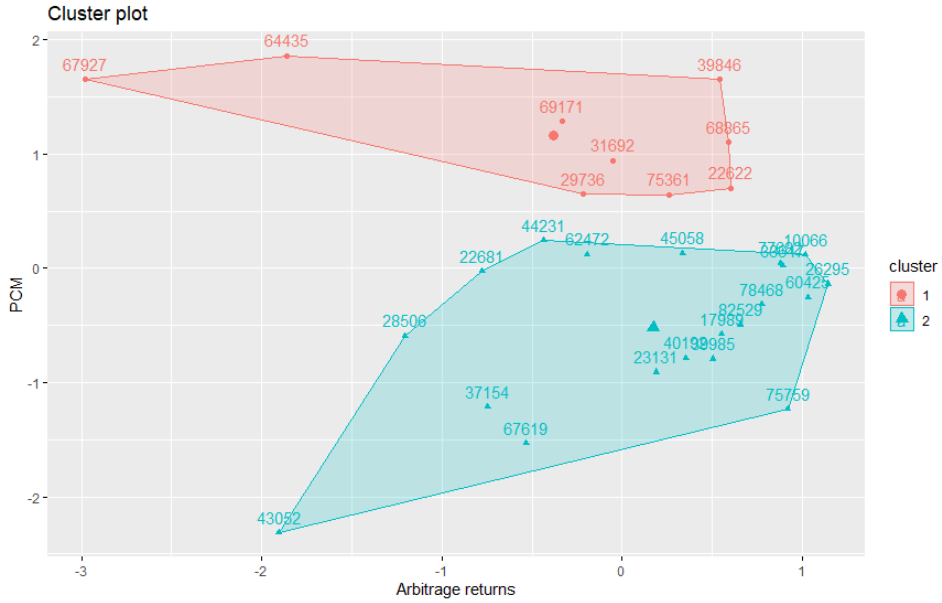


Figure 10: C2D Curve in 2016-2017

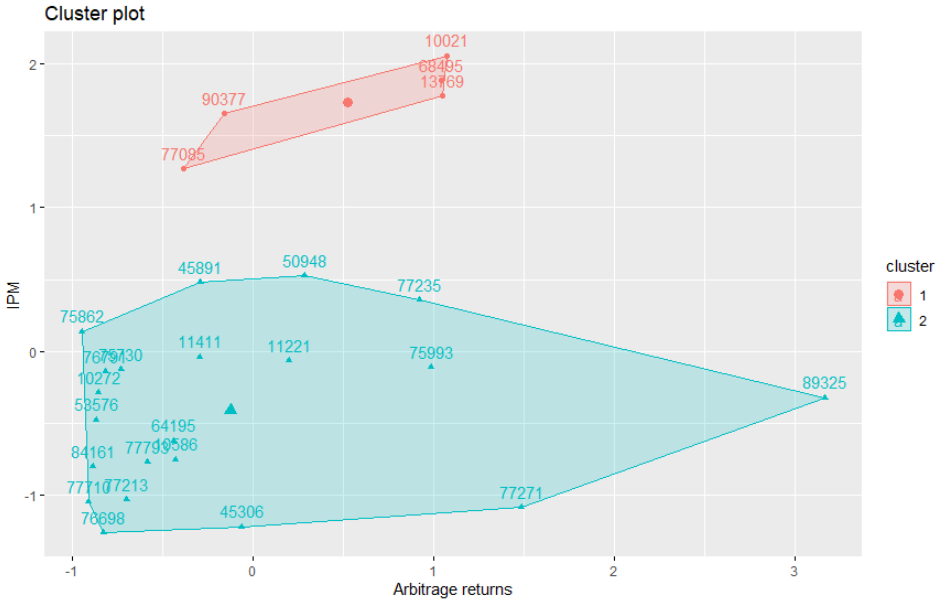


(a) Clustering of PCM with highest returns

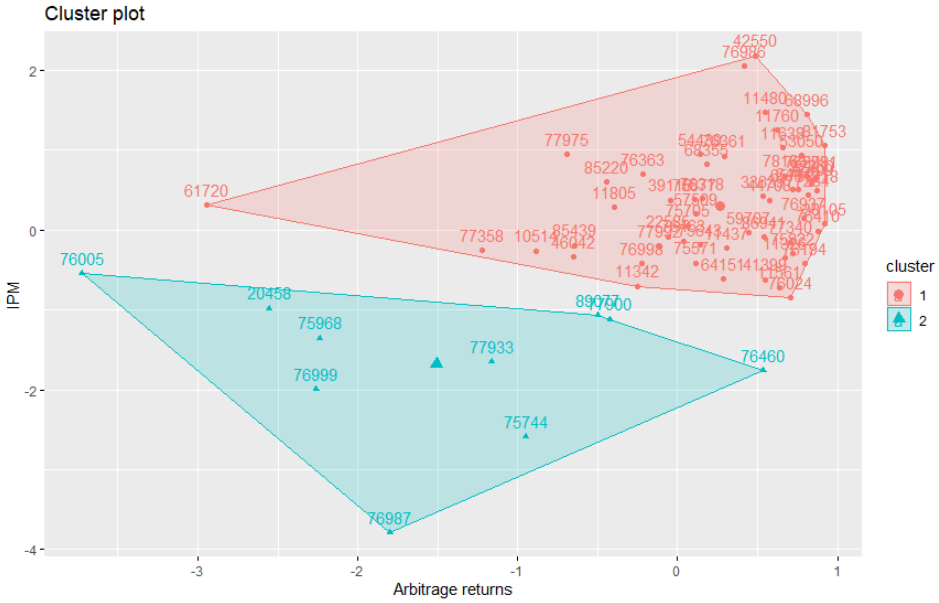


(b) Clustering of PCM with highest returns

Figure 11: Clustering of PCM 1986-1987



(a) Clustering of IPM with highest returns



(b) Clustering of IPM with highest returns

Figure 12: Clustering of IPM 2004-2005

References

[1] AYMANNS, C., FARMER, J. D., KLEINNIJENHUIS, A. M., AND WETZER, T. Models of financial stability and their application in stress tests. In *Handbook of Computational Economics*, vol. 4. Elsevier, 2018, pp. 329–391.

[2] BOLLERSLEV, T., ENGLE, R. F., AND WOOLDRIDGE, J. M. A capital asset pricing model with time-varying covariances. *Journal of political Economy 96*, 1 (1988), 116–131.

[3] CARHART, M. M. On persistence in mutual fund performance. *The Journal of finance 52*, 1 (1997), 57–82.

[4] CONNOR, G., HAGMANN, M., AND LINTON, O. Efficient semiparametric estimation of the fama–french model and extensions. *Econometrica 80*, 2 (2012), 713–754.

[5] CONNOR, G., AND LINTON, O. Semiparametric estimation of a characteristic-based factor model of common stock returns. *Journal of Empirical Finance 14*, 5 (2007), 694–717.

[6] COX, D. R. Note on grouping. *Journal of the American Statistical Association 52*, 280 (1957), 543–547.

[7] FAMA, E. F., AND FRENCH, K. R. Common risk factors in the returns on stocks and bonds. *Journal of financial economics 33*, 1 (1993), 3–56.

[8] FAMA, E. F., AND FRENCH, K. R. A five-factor asset pricing model. *Journal of financial economics 116*, 1 (2015), 1–22.

[9] FAN, J., LIAO, Y., AND WANG, W. Projected principal component analysis in factor models. *Annals of statistics 44*, 1 (2016), 219.

[10] FAN, J., LIAO, Y., AND YAO, J. Power enhancement in high-dimensional cross-sectional tests. *Econometrica 83*, 4 (2015), 1497–1541.

[11] FENG, G., GIGLIO, S., AND XIU, D. Taming the factor zoo. *Fama-Miller Working Paper 24070* (2017).

[12] FISHER, W. D. On grouping for maximum homogeneity. *Journal of the American statistical Association 53*, 284 (1958), 789–798.

[13] FREYBERGER, J., NEUHIERL, A., AND WEBER, M. Dissecting characteristics nonparametrically. Tech. rep., National Bureau of Economic Research, 2017.

[14] HJALMARSSON, E., AND MANCHEV, P. Characteristic-based mean-variance portfolio choice. *Journal of Banking & Finance 36*, 5 (2012), 1392–1401.

[15] HOBERG, G., AND PHILLIPS, G. Text-based network industries and endogenous product differentiation. *Journal of Political Economy 124*, 5 (2016), 1423–1465.

[16] HOU, K., XUE, C., AND ZHANG, L. Digesting anomalies: An investment approach. *The Review of Financial Studies 28*, 3 (2015), 650–705.

[17] HUANG, J., HOROWITZ, J. L., AND WEI, F. Variable selection in nonparametric additive models. *Annals of statistics 38*, 4 (2010), 2282.

[18] KELLY, B. T., PRUITT, S., AND SU, Y. Instrumented principal component analysis. *Available at SSRN 2983919* (2017).

[19] KELLY, B. T., PRUITT, S., AND SU, Y. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics* (2019).

[20] KIM, S., KORAJCZYK, R. A., AND NEUHIERL, A. Arbitrage portfolios.

[21] KOCK, A. B., AND PREINERSTORFER, D. Power in high-dimensional testing problems. *Econometrica 87*, 3 (2019), 1055–1069.

[22] LIEW, C. K. Inequality constrained least-squares estimation. *Journal of the American Statistical Association 71*, 355 (1976), 746–751.

[23] PESARAN, M. H., AND YAMAGATA, T. Testing capm with a large number of assets. In *AFA 2013 San Diego Meetings Paper* (2012).

[24] POLLARD, D. Strong consistency of k-means clustering. *The Annals of Statistics* (1981), 135–140.

[25] POLLARD, D., ET AL. A central limit theorem for k-means clustering. *The Annals of Probability 10*, 4 (1982), 919–926.

[26] ROSENBERG, B. Extra-market components of covariance in security returns. *Journal of Financial and Quantitative Analysis 9*, 2 (1974), 263–274.

[27] SUN, W., WANG, J., FANG, Y., ET AL. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics 6*.

[28] VOGT, M., AND LINTON, O. Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*, 1 (2017), 5–27.