Dynamic Peer Groups of Arbitrage Characteristics[*]

Shuyi Ge [†1], Shaoran Li [‡2], and Oliver Linton [§3]

[1 2 3] *Faculty of Economics, University of Cambridge*

February 16, 2021

Abstract

We propose an asset pricing factor model constructed with semi-parametric characteristics-based mispricing and factor loading functions. We approximate the unknown functions by B-splines sieve 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, with the strengthened power approaches to one asymptotically. We also investigate the structure of mispricing components through Hierarchical K-means Clusterings. We apply our methodology to CRSP (Center for Research in Security Prices) and FRED (Federal Reserve Economic Data) 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. We also find that distinct clusters of the same characteristics lead to similar arbitrage returns, forming a "peer group" of arbitrage characteristics.

Keywords: Semiparametric; Characteristics-based; Peer Groups; 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 and Weiguang Liu for their useful comments. The editor and two referees deserve considerable credit for their help in improving this manuscript.

[†] Electronic address: `sg751@cam.ac.uk`.

[‡] Corresponding author. 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) 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 the risk factor, The APT model explains the excess returns from both cross-sectional and time-series directions. Fama and French (1993) and Fama and French (2015) 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 factor models, such as the four-factor model of Carhart (1997), the q-factor model of Hou et al. (2015), and the factor zoo by Feng et al. (2017) among others. All of above papers studied observed factors and did not assigned characteristics-based information to either alpha or beta.

Security-specific characteristics, such as capitalization and book to market ratio, are usually documented to explain asset-specific excess returns. Freyberger et al. (2017) analyzed the non-linear effects of 62 characteristics through pooling regressions. This study concluded that 13 of these characterisitcs have explanatory power on stock excess returns after selecting by adaptive group Lasso. Characterisics-based information are exploited to develop arbitrage portfolios by directly parameterizing the portfolio weights as a linear function of characteristics, as in Hjalmarsson and Manchev (2012) and Kim et al. (2019). Empirically, they showed that their portfolio outperformed other baseline competitors.

This paper's contributions are fourfold. Firstly, we build up a more flexible semi-parametric characteristics-based asset pricing factor model with a focus on mispricing component. Secondly, we extend previous estimation and testing methods, which can fit the current framework better. Especially, we extend the power enhanced test of Fan et al. (2015) in a group manner to strengthen the conventional Wald test for mispricing functions. This test can also select the characteristics that contribute to arbitrage portfolios simultaneously. Thirdly, we construct a two-layer clusterings structure of mispricing components. Finally, our methods are applied to fifty years of monthly US stock data. We detect distinct clusters of the same characteristics resulting in similar arbitrage returns, forming a "peer group" of arbitrage characteristics. This finding supplements existing portfolio management techniques by implying that the development of arbitrage portfolio through the asset weights determined by the linear mispricing function is improvable.

This class of models has a basic regression specification in Equation 1. 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 of security i at time t; f_{jt} is the j^{th} risk factor's return 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} is the mean zero idiosyncratic shock. In terms of factor loadings β_{ji}, Connor and Linton (2007) and Connor et al. (2012) 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. 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.

They restricted their beta function to be univariate and did not consider the part of factor loading function that cannot be explained by characteristics. To overcome this limitation, Fan et al. (2016) allowed β_{ji} in Equation 1 to have a component explained by observable characteristics as well as 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 to find the unobservable factors. This method has attractive properties even under large n and small T setting. However, they did not study the mispricing part (alpha), which is crucial to both asset pricing theories and portfolio management.

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 et al. (2017). 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 as $\alpha_i = h(\mathbf{X}_i) + \gamma_i$ and $\beta_{ji} = g_j(\mathbf{X}_i) + \lambda_{ij}$[1]. We then estimate $h(\mathbf{X}_i)$, $g_j(\mathbf{X}_i)$ and unobservable risk factors f_{jt}. In addition, we design a power enhanced test and Hierarchical K-mean Clusterings for the mispricing function $h(\mathbf{X}_i)$ to study the non-linear behavior of arbitrage characteristics.

Some recent papers such as Kim et al. (2019) and Kelly et al. (2019) analyzed a similar model as ours, which assume that both $h(\mathbf{X}_i)$ and $g_j(\mathbf{X}_i)$ are *linear 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 et al. (2019) determined assets weights of arbitrage portfolios using one-year rolling window estimated $\frac{1}{n}\hat{h}(\mathbf{X}_i)$. They showed that their arbitrage portfolios returns are statistically and economically significant. However, Kelly et al. (2019) applied instrumented principal component analysis (IPCA) to the entire time span from 1965 to 2014, and concluded no evidence to reject the null hypothesis $H_0 : h(\mathbf{X}_i) = \mathbf{X}_i^{\intercal}\mathbf{B} = \mathbf{0}$ through bootstrap. This dispute spurs the deveploment of a more flexible model and reliable hypothesis tests to investigate the existence and structure of $h(\mathbf{X}_i)$. The introduction of IPCA, which require both large n and T to work, was introduced at Kelly et al. (2017). This method does not fit our setting since we apply rolling window analysis with small T. Furthermore, Kelly et al. (2019) restricted the

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

function form of $h(\mathbf{X}_i)$ and $g_j(\mathbf{X}_i)$ to be time-invariant, which is not consistent with our empirical results under a semi-parametric setting. To clarfy the differences with aforementioned research, this paper proposes a semi-parametric model, which allows for both non-linearity and time-variation of $h(\mathbf{X}_i)$ and $g_j(\mathbf{X}_i)$. Furthermore, we consider a different economic question, namely, the existence and structure of mispricing functions. Our empirical study sheds light on why Kelly et al. (2019) and Kim et al. (2019) drew different conclusions: weak, time varying and nonlinear characteristics-based mispricing functions only appear in certain rolling windows, which is hard to be detected without rolling window analysis. However, given the presence of some persistent arbitrage characteristics, portfolios developed through mispricing functions can provide abitrage returns.

The unrestrictive model in this paper brings both opportunities and challenges. According to Huang et al. (2010), the number of B-spline knots must increase in the number of observations to achieve accurate approximation and good asymptotic performance. Therefore, the dimension of B-spline bases coefficients also need to grow with the sample size. Besides, mispricing functions are treated as anomalies. Under a correctly specified factor model, coefficients of these B-splines bases are very likely to be sparse. All of these circumstances make the conventional Wald tests have very low power as discussed in Fan et al. (2015). Therefore, a power enhanced 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) illustrated that if the number of coefficients diverges as the number of observations approaches infinity, the standard Wald test is power enhanceable. Fan et al. (2015) proposed a power enhanced test by introducing a screening process on all estimated coefficients one by one. They added significant components as a supplement to the standard Wald test. In this paper, we extend Fan et al. (2015) to a group manner to enhance the hypothesis test on a high dimensional additive semi-parametric function, $H_0 : h(\mathbf{X}_i) = 0$. This method allows all the significant components of $h(\mathbf{X}_i)$ to be selected and contribute to the test statistics, with the testing power approaching to one.

The careful analysis of $h(\mathbf{X}_i)$ is theoretically and practically meaningful. Firstly, the presence of $h(\mathbf{X}_i)$ 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. Secondly, as in Hjalmarsson and Manchev (2012) and Kim et al. (2019) , $h(\mathbf{X}_i)$ can be utilized to construct arbitrage portfolios through observed characteristics. However, both research was built upon the condition that $h(\mathbf{X}_i)$ is linear over characteristics. 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. In this paper, we show that some characteristics with substantially different values give rise to similar arbitrage returns. The distance of arbitrage returns between two assets i and j is $d_{ij} = |h(\mathbf{X}_i) - h(\mathbf{X}_j)|$ and the similarity of characteristics is $\|\mathbf{X}_i - \mathbf{X}_j\|_2$, where $\|\cdot\|_2$ represents L_2 distance. Inspired by Hoberg and Phillips (2016) and Vogt and Linton (2017), we employ a hierarchical K-means clustering to classify these characteristics within each mispricing return group. We present the dynamic of distinct clusters of the same characteristics leading to similar arbitrage returns, forming a "peer group" of arbitrage characteristics. Therefore, under the semiparametric setting, the asset weighting function should rely on these time-varying and non-linear peer groups.

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 investigate peer groups of arbitrage characteristics. 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 plots of peer groups 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 time window, each security has P time-invariant observed characteristics, such as market capitalization, momentum, and book-to-market ratios. Meanwhile, we may omit 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; $\mathbf{X}_i$ is a $1 \times P$ vector of P characteristics of stock i during time periods $t = 1, \dots T$. T is a small and fixed time block. In practice, most characteristics are updated annually. Thus, we assume $\mathbf{X}_i$ is time-invariant in one-year time window. $h(\mathbf{X}_i)$ 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 $E(\lambda_{ij}) = 0$. λ_{ij} is orthogonal to the $g_j(\cdot)$ function. f_{jt} is the realization of the j^{th} risk factor at time t. Finally, ϵ_{it} is homoskedastic zero-mean idiosyncratic residual of the i^{th} stock at time t. 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.

To avoid the curse of dimensionality , we impose additive forms on both $h(\cdot)$ and $g_j(\cdot)$ functions: $h(\mathbf{X}_i) = \sum_{p=1}^{P} \mu_p(X_{ip})$ 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. We rewrite the model:

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

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 et al. (2012) and Fan et al. (2016), Assumption 1 above is to standardize the model settings, including the zero mean assumption for factor loadings and mispricing functions for identification purposes. We also impose orthogonality between mispricing and factor loading parts for the identification reason. This is because the variation of 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 et al. (2012).

3 Estimation

In this section we discuss the approximation of unknown univariate functions and our estimation methods for model Equation 3. In the semi-parametric setting, we apply the Projected-PCA following Fan et al. (2016) to work on the common factors and characteristics-beta directly. Next, we project the residuals onto the characteristics-based alpha space that is orthogonal to the systematic part. The second step is similar to equality constrained OLS.

3.1 B-Splines Approximation

We use B-splines sieve to approximate unknown functions $\theta(\cdot)$ and $\mu(\cdot)$ in Equation 4. Similar to Huang et al. (2010) and Chen and Pouzo (2012), 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, $m_n = \lfloor n^v \rceil$ ($\lfloor \cdot \rceil$ gives nearest integer) is a positive integer of the number of internal knots, which is a function of security size n in period t with $0 < v < 0.5$. 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, there is a set of H_n orthogonal bases $\{\phi_{1p}(X_p), \ldots, \phi_{H_n p}(X_p)\}$. Those univariate unknown functions can be approximated as linear combinations of these bases as $\mu_p(X_p) = \sum_{q=1}^{H_n} \alpha_q \phi_{qp}(X_p) + R_p^{\mu}(X_p)$ and $\theta_p(X_p) = \sum_{q=1}^{H_n} \beta_{jq}\phi_{qp}(X_p) + R_p^{\theta}(X_p)$, 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 notational simplicity only. Therefore, the model Equation 4 can be written 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 $i = 1, 2, \ldots, n$, $p = 1, 2, \ldots, P$ and $t = 1, 2, \ldots, T$, we have:

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

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

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

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

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

where $\phi_{i,ph}(X_{ip})$ is the h^{th} basis of the p^{th} characteristic of asset i at time t. Therefore, the original model

$$\mathbf{Y} = (h(\mathbf{X}) + \mathbf{\Gamma})\mathbf{1}_\mathbf{T}^\intercal + (\mathbf{G}(\mathbf{X}) + \mathbf{\Lambda})\mathbf{F}^\intercal + \mathbf{U},$$

can be represented by B-spline sieve as:

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

$\mathbf{Y}$ is $n \times T$ matrix of y_{it}; $\mathbf{\Phi}(\mathbf{X})$ 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 $R_p^\mu(X_p) \rightarrow^p 0$ and $R_p^\theta(X_p) \rightarrow^p 0,$ as $n \rightarrow \infty$ as in Huang et al. (2010). 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}. $h(\mathbf{X})$ is a $n \times 1$ vector of characteristics-based mispricing component; $\mathbf{G}(\mathbf{X})$ is a $n \times J$ vector of characteristics-based factor loadings; $\mathbf{1_T}$ is a $T \times 1$ vector of 1. The rest are defined the same as Equation 4.

We define a 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 cluster peer groups of arbitrage characteristics.

3.2 Two Steps Projected-PCA

In this section, we combine and extend Projected-PCA by Fan et al. (2016) and equality constrained least squares similar to Kim et al. (2019) to estimate the model. 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 since they are time-invariant by $(\mathbf{\Phi}(\mathbf{X})\mathbf{A} + \mathbf{\Gamma})\mathbf{1_T}^{\intercal}\mathbf{D_T} = \mathbf{0}$. This helps us to work on the systematic part directly. Henceforth, we use $\mathbf{F}$ to represent the time-demeaned factor matrix.

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

Under Assumption 1, we have the following estimation procedures:

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 $\mathbf{P}$ with $\mathbf{P} = \Phi(\mathbf{X})(\Phi(\mathbf{X})^{\intercal}\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})^{\intercal}$. We then collect 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 $\hat{\mathbf{Y}}^{\intercal}\hat{\mathbf{Y}}$. This allows us to work directly on the sample covariance of $\mathbf{G}(\mathbf{X})\mathbf{F}^{\intercal}$, under the condition $E(g_j(\mathbf{X}_i)\epsilon_{it}) = E(g_j(\mathbf{X}_i)\lambda_{ij}) = 0$.

3 Estimating $\hat{\mathbf{F}}$ as 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 estimation accuracy and facilitates theoretical analysis even under the large n and small T . Small T is preferable in our model setting as we use one-year rolling windows analysis in both simulation and empirical studies, and large n is required for asymptotic analysis.

Factor loadings $\hat{\mathbf{G}}(\mathbf{X})$ 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 estimator of $\mathbf{A}$ is

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \text{vec}(\mathbf{Y} - \mathbf{\Phi}(\mathbf{X})\mathbf{A}\mathbf{1_T}^{\intercal} - \hat{\mathbf{G}}(\mathbf{X})\hat{\mathbf{F}}^{\intercal})^{\intercal}\text{vec}(\mathbf{Y} - \mathbf{\Phi}(\mathbf{X})\mathbf{A}\mathbf{1_T}^{\intercal} - \hat{\mathbf{G}}(\mathbf{X})\hat{\mathbf{F}}^{\intercal}),$$

subject to $\hat{\mathbf{G}}(\mathbf{X})^{\intercal}\mathbf{\Phi}(\mathbf{X})\mathbf{A} = \mathbf{0_J}$.

Let a $PH_n \times 1$ vector $\hat{\mathbf{A}}$ be a closed-form solution:

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

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

where

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

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

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

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

5. We also estimate the covariance matrix of $\hat{\mathbf{A}}$, i.e., $\boldsymbol{\Sigma}$, by extending the methods of Liew (1976). This can facilitate theoretical analysis in the next section. According to Liew (1976), $\hat{\mathbf{A}}$ is the equality constrained least-square estimator, which has the covariance matrix as (under $n \leqslant T$ and covariance shrinkage as in Ledoit et al. (2012) and Fan et al. (2013) among others.):

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

where:

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

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

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

4 Power Enhanced Tests

There are considerable discussions about the mispricing phenomenon under factor models while the existence of mispricing functions remains controversial. Namely, whether there are relevant covariates explaining remaining excess returns after subtracting co-movements components captured by risk factors. Recently, Kim et al. (2019) found the characteristics arbitrage opportunities by estimating a linear characteristic mispricing function, without providing theoretical results. However, Kelly et al. (2019) conducted a conventional Wald hypothesis test on the similar mispricing function using bootstrap, concluding that there is no evidence to reject the null hypothesis $H_0 : h(\mathbf{X}) = \mathbf{0}$. Additionally, they applied the bootstrap method to estimate the covariance matrix $\boldsymbol{\Sigma}$, which caused potential problems for theoretical analysis. Moreover, according to Fan et al. (2015), their test results may have relatively low power when the true coefficient vector of linear mispricing function $\mathbf{A}$ has a sparse structure.

Both studies adopt a parametric framework, which relies on the strong assumption of linearity. However, this assumption is not consistent with Connor et al. (2012), which showed that both characteristic-beta and mispricing functions are very likely to be non-linear. Therefore, we propose a semiparametric model to accommodate the non-linearity to a great extent.

But semi-parametric framework leads to additional challenges for inference. On the one hand, as mentioned above, the number of coefficients of mispricing B-splines diverge as $n \to \infty$, which implies that the power of standard Wald test can be quite low, (see Fan et al. (2015)). On the other hand, according to other research like Fama and French (1993) and Fama and French (2015), 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.

According to Kock and Preinerstorfer (2019), conventional hypothesis tests under these circumstances are power enhanceable. The power enhanced Wald test in this paper is an extension of Fan et al. (2015) to a group manner, namely, the hypothesis test under high-dimensional additive semi-parametric settings. The proposed test are power strengthened when the coefficients of the additive regression $\mathbf{A}$ is diverging as $n \to \infty$ without size distortion. Meanwhile, this test is robust to sparse alternatives. On top of that, 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 diverging, i.e., $P \to \infty$.

We construct a new test:

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

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:

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

where S_1 is the "original" Wald test statistics; P is the number of characteristics; PH_n is the total number of B-spline bases, and $\mathbf{A} \in \mathbb{R}^{PH_n}$. The value of 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) showed that:

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

under regularity conditions.

Potentially, sparse 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 the screening process while η_n is a data-driven threshold value to avoid potential size-distortion.

Here we discuss the choice of η_n. By construction and assumption of independent characteristics, we assume all B-Spline bases are orthogonal. Our goal is to bound the maximum of those standardized coefficients.

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

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

$$\mathbf{E}(\mathbf{Z}) = \sqrt{2\log PH_n}.$$

After grouping coefficients of bases used to approximate 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 H_n is a slowly diverging sequence, it can control the influence of the group size properly. Meanwhile, η_n also diverges slowly so that η_n is a conservative threshold value used 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.

We then define the arbitrage characteristics set, which includes the characteristics that have the strong explanation power for 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 strengthens the power of the test and drives the power to one since H_n is slowly diverging.

Therefore, our new test statistics is $S = S_0 + S_1$, and 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 $H_0 : h(\mathbf{X}) = \mathbf{0}$ is mainly enhanced without size distortions.

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

As designed, S_0 satisfies all three properties of Fan et al. (2015), 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, $\Pr(S_0 = 0 \mid H_0) \to 1$

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 listed:

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

5 Hierarchical K-Means Clustering

This section introduces a Herarchical K-means Clustering method to find peer groups of arbitrage characteristics based on their arbitrage returns. We ask whether distinct groups of the same characteristics may result in similar characteristic-based arbitrage returns in each rolling block, which is an implication for non-monotonic mispricing function, and forms a "peer group" of arbitrage characteristics. Because arbitrage portfolios rely on the linearity of characteristics-bases mispricing components to work, our clustering results can provide new evidence for the effectiveness of these arbitrage porfolios. Introduction of K-means clustering can be found in Cox (1957) and Fisher (1958).

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

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

We define a $n \times M$ matrix $\boldsymbol{M}$ of arbitrage characteristics 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. We also set characteristics-based arbitrage returns of asset i in month t as:

$$\ddot{y}_{it} = \boldsymbol{\phi}(\boldsymbol{M_i})\hat{\mathbf{A}}_{\mathrm{M}},$$

where $\boldsymbol{\phi}(\boldsymbol{M_i})$ and $\hat{\mathbf{A}}_{\mathrm{M}}$ are the corresonding parts of matrix $\boldsymbol{\Phi}(\mathbf{X_i})$ and vector $\hat{\mathbf{A}}$. For each rolling window, we classify all n assets through a 2-layer K-means clustering. At the first layer, we cluster these assets into K groups according to the similarity of their characteristics-based arbitrage returns $\ddot{y}_{it}$. At the second layer, we divide R_j subgroups within the j^{th} group from the first layer by the similarity of their arbitrage characteristics, where $j = 1, 2, \ldots, K$. Finally, the peer groups of arbitrage characteristics can be attained. We repreat this method for all rolling blocks to investigate dynamic patterns of these peer groups. These clusterings will provide illustrative evidence of linear/nonlinear and time-invariant/time-varying structure of mispricing function $h(\boldsymbol{X})$.

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

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

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

$\boldsymbol{M}_i$ represents the i^{th} row of $\boldsymbol{M}$. We set two tolerance thresholds ψ_y and ψ_x, which are used to control the biggest difference within each group of both layers separately. To accelerate the convergence of the K-means Clustering, we first apply a first difference process, which is introduced below, to obtain centroids as in Vogt and Linton (2017).

For the first layer, we have first difference process:

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 $\Delta_{i(1)} = \min\{\Delta_{i(1)} \ldots \Delta_{i(n)}\}$, and $\Delta_{i(n)} = \max\{\Delta_{i(1)} \ldots \Delta_{i(n)}\}$.

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 large jumps of "first difference", all of which are 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 $n - j_K + 1$ 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 method to further divide them into r subgroups as $\mathcal{R}_{1k}, \mathcal{R}_{2k}, \ldots, \mathcal{R}_{rk}$. Within each subgroup, 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. Step 1: 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. Step l: 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.

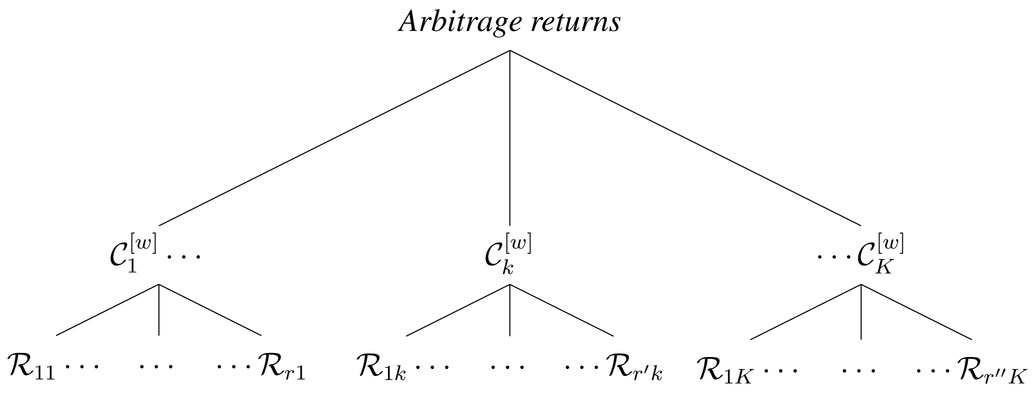
To accelerate the convergence of K-means algorithm, at the step 1, results of first difference are used. As we have already obtained our initial grouping $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, therefore starting values for the Step 1 is:

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

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

The consistency and other theoretical results of the above procedures can be found in Pollard (1981)Pollard et al. (1982), Sun et al. (2012) and Vogt and Linton (2017).

For the second layer, we repeat the procedures within each group $\mathcal{C}_k^{[w]}$ respect to Υ_{ab}, and the structure of characteristics-based arbitrage returns is:


Arbitrage returns
$\mathcal{C}_1^{[w]} \cdots$
$\mathcal{C}_k^{[w]}$
$\cdots \mathcal{C}_K^{[w]}$
$\mathcal{R}_{11} \cdots$
$\cdots$
$\cdots \mathcal{R}_{r1}$
$\mathcal{R}_{1k} \cdots$
$\cdots$
$\cdots \mathcal{R}_{r'k}$
$\mathcal{R}_{1K} \cdots$
$\cdots$
$\cdots \mathcal{R}_{r''K}$

The first layer is the structure of characteristics-based arbitrage returns, while the second layer gives peer groups of characteristics that can provide similar characteristics-based arbitrage returns.

The number of clusterings are determined by threshold values ψ_y and ψ_x directly. ψ_y and ψ_x are chosen by the tradeoff between the number of clusterings and total within sum of squares.

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 the smallest eigenvalue of matrix M, respectively. Additionally, we define C_{min} and C_{max} to be 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 to avoid non-invertibility of stock returns, characteristics, and rotation indeterminacy.

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 those in Fan et al. (2016), 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 nonvanishing explanatory power of the B-spline bases $\mathbf{\Phi}(\mathbf{X})$ on the factor loading matrix $\mathbf{G}(\mathbf{X})$.

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

Heteroskedasticity is caused by γ_i, namely, $\mathrm{var}(\gamma_i + \epsilon_{it}) = \sigma_i^2$.

6.2 Main Results

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

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 the $PH_n \times 1$ vector $\hat{\mathbf{A}}$ be the solution of constrained OLS, then*

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

where

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

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

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

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:*

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

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 approaching to 1 as:

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

7 Numerical Study

In this section, we use Compustat and Fama-French three and five factors data to simulate stocks returns and then evaluate 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 excess returns. Most of the characteristics are updated annually so we treat those variables as time-invariant during each one-year rolling block. For the characteristics that are updated every month, we substitute the mean values as their fixed values for each fiscal year. We use 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 from July of year t to June of year $t+1$. The periods we generate are the same as the empirical study, namely, 50 years from July 1967 to June 2017. For each rolling block with 12 months we have:

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

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, which is to mimic the sparse structure of the mispricing function; $g_j(\boldsymbol{X}_j)$ is the j^{th} characteristics-based factor loading, which has an additive semi-parametric structure; $\boldsymbol{X}_j$ 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 for stock i at time t, generated from $N(0, \sigma^2)$.

We generate characteristic functions:

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

$$g_j(\boldsymbol{X}_j) = X_{j1}^2 + (3X_{j2}^3 - 2X_{j2}^2) + (3X_{j3}^3 - 2X_{j3}) + X_{j4}^2,$$

X_{ji} is a randomly picked characteristic without replacement from the data in empirical study and $j = 1, 2, 3$, $i = 1, \ldots, 4$. Description of these characteristics can be found in the Appendix. Additionally, all $h(X_i)$, $g_j(\boldsymbol{X}_j)$ are rescaled to have zero mean and unit variance. As we use real data to conduct the simulation, the assumption of independent X_i may not be satisfied. Although some characterisitcs are correlated, the semi-paramtric model overcomes this problem properly when compared with the parametric model that has serious size distortion.

We do not specify $h(X_i)$ and $g_j(\boldsymbol{X}_j)$ to be orthogonal explicitly, but we draw characteristics without replacement and employ sine-waves and polynomials to approximate the orthogonality as much as possible. Orthogonality is a strong assumption in reality, which is required for Theorem 6.5. In this simulatin, our method can only estimate the component of $h(X_i)$ that is orthogonal to $g_j(\boldsymbol{X}_j)$. However, results reveal that one can still select the arbitrage characteristics even if we cannot estimate arbitrary $h(X_i)$ unrestrictively.

7.2 Model Misspecification

In this subsection, we show the necessity to consider semi-parametric analysis when the forms of factor loadings and mispricing functions are nonlinear.

Under the data generation process, we consider both semi-parametric and linear analysis to compare 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. In terms of hypothesis test, we use the same logic as in the semi-parametric settings, and we set $\eta_n = \sqrt{3 \log(P)}$.

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.

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 but also avoid size distortions. As for the linear model, it is influenced by the correlated characteristics. Therefore, during certain periods we even obtain the non-invertible characteristic matrix. The linear model can also select the relevant covariates with decent probability, but it suffers from serious size distortions. In contrast, our semi-parametric model with orthogonal bases can mitigate this problem to a great extent.

3 Because S_1 can be very small and even negative, especially when the noise σ_i is strong, 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.

Table 1: Simulation Results 1 Part1

		$\sigma^2 = 1$												$\sigma^2 = 4$											
		Linear Model						Semi-parametric Model						Linear Model						Semi-parametric Model					
Window	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					
Window	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 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. 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 models under different noise levels, $\sigma^2 = 1$ and $\sigma^2 = 4$. NA results are causes by non-invertible characteristic matrices. "Selected" means the percentage of selecting the relevant characteristic in the mispricing function in 1000 experiments. Similarly, "distortion" represents the percentage of wrongly selecting irrelevant characteristics in 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 comparatively high noise level recalling that all the components of our simulation model are rescaled to have unit variance. Another fact is that the stronger noise does deteriorate the power 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 numbers of factors considered.

We simulate according to another data generation process:

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

similarly, y_{it} is the generated stock return; $h(X_i)$ is the mispricing function consist of a non-linear characteristic function of X_i, to mimic the sparse structure of the mispricing function; $g_j(\boldsymbol{X}_j)$ is the j^{th} characteristics-based factor loading, which has an additive semi-parametric structure; 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 functions as:

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

$$g_j(\boldsymbol{X}_j) = X_{j1}^2 + (3X_{j2}^3 - 2X_{j2}^2) + (3X_{j3}^3 - 2X_{j3}) + X_{j4}^2,$$

where X_{ji} is a randomly picked characteristic without replacement from the data in empirical study with $j = 1, \ldots, 5$, $i = 1, \ldots, 4$. Furthermore, all $h(X_i)$ and $g_j(\boldsymbol{X}_j)$ are rescaled to have zero mean and unit variance.

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

The first category column is the scenario of over estimating the number of 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. For some rolling blocks, the probability of mistakenly selected irrelevant characteristics is slightly higher under over estimating the number of factors. Moreover, over estimating the number of factors can increase the model fitting marginally. Therefore, we conclude that over estimating the number of factors does not cause severe size distortion using our methods.

On the other hand, under estimating the number of factors can lead to misleading test results. We can conclude this from the last column where we estimate the number of factors to be three in a five-factor model. Compared with the correct specified model, under estimating 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-factor model rather than the three-factor 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 $\hat{J} = 3$					
Window	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 $\hat{J} = 3$					
Window	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 results under the characteristics-based beta and alpha of both Fama-French 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. Each column summarises the mean value of 1000 estimation and test results. We compare the results of both over and under estimating the number of factors, namely,$\hat{J} = 3$ and $\hat{J} = 5$. S_1 is the conventional Wald test while S_0 is the power-strengthened component. NA results are caused by non-invertible characteristic matrices. "Selected" means the percentage of selecting the relevant characteristic in mispricing functions in 1000 repetitions. Similarly, "distortion" represents the percentage of wrongly selecting irrelevant characteristics in 1000 experiments.

8 Empirical Study

8.1 Data

We use monthly stock returns from CRSP and firms' characteristics from Compustat, ranged from 1965 to 2017. We construct 33 characteristics following the methods of Freyberger et al. (2017). 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 data, we merge two data sets from CRSP and Compustat. We require all of the firms included in our analysis to have at least three years of characteristics data in Compustat.

Data is modified with regards to the following aspects:

1 Delisting is quite common for CRSP data. We use the way of Hou et al. (2015) to correct the returns of delisting stocks for all the delisted assets before 2018. Detailed methods can be found in the appendix.

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 in each window. 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. Meanwhile, we take monthly updated characteristics' mean values of 12 months as fixed characteristic values in each window. We also use rolling window method to detect peer groups of arbitrage characteristics.

3 B-splines are based on each time-invariant characteristic among n firms which are not delisted in 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 obtained by the difference between monthly stock returns and Fama-French risk-free monthly returns.

8.2 Estimation

We construct B-spline bases based on evenly distributed knots, and the degree of each basis is three. We choose $v = 0.3$, which means the number of bases for each characteristic in each window is $\lfloor n^{0.3} \rceil$, and n is the number of stocks. To get a relatively large balanced panel in each window, some characteristics with too many missing values are eliminated. Therefore, only 33 characteristics are left. Firms kept in balanced panels in our dataset range from 468 to 2928, which means that both n and $\hat{\mathbf{A}} \in \mathbb{R}^{PH_n}$ are diverging. Large n can satisfy asymptotic requirements.

These facts emphasize the necessity of introducing a power enhanced component into the hypothesis test. Before the next step, we use time-demeaning matrix $\mathbf{D_T}$ to demean excess return matrix in each window.

Next, we project the time-demeaned monthly excess return matrix $\tilde{\mathbf{Y}}$ to the B-spline space spanned by characteristics $\boldsymbol{\Phi}(\mathbf{X})$, and then we collect the fitted value $\hat{\mathbf{Y}}$. We apply Principle Component Analysis on $\frac{1}{n}\hat{\mathbf{Y}}^{\intercal}\hat{\mathbf{Y}}$, and attain the first five eigenvectors corresponding to the first five biggest eigenvalues as the estimates of unobservable factors $\mathbf{F}$. We choose the number of factors to be five according to simulation results.

Then, we estimate factor loading matrix by:

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

Moveover, we use equality-constrained OLS to estimate the mispricing function. We project excess monthly return matrix on the characteristic space $\boldsymbol{\Phi}(\mathbf{X})$ 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 the mispricing function. Therefore, our final step is to estimate covariance matrix $\boldsymbol{\Sigma}$ of $\hat{\mathbf{A}}$.

8.3 Power Enhanced Hypothesis Tests

In this section, we conduct a power enhanced test in each rolling block. 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, with $P = 33$. η_n is data-driven critical value and it diverges as the number of firms increases. We use 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 select the most relevant characteristics that can explain the variation of the mispricing function. Results of last column in Table 5 are characteristics selected in $\hat{\mathcal{M}} = \{\mathbf{X_p} \in \hat{\mathcal{M}} : \sum_{h=1}^{H} |\hat{\alpha}_{ph}|/\hat{\sigma}_{ph} \geqslant \eta_n, \quad h = 1, 2, \ldots, H_n, \quad p = 1, 2, \ldots, P\}$.

2 It contributes to the test statistics S by adding a diverging power enhanced component S_0. As $T = 12$ is small in this empirical study, we assume the homoskedasticity of ϵ_{it}. We also specify a overshrunk covariance matrix by setting off-diagonal elements to be zeros.

3 It avoids size-distortion by the conservative critical value η_n.

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

Finally, the new statistics S can be calculated as:

$$S = S_0 + S_1,$$

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

8.4 Test Results

This section presents the empirical results. Details can be found in Table 5, which list the results of 50 rolling windows from Jul.1967 to Jun.2017. Generally, the number of firms included in the 12-month rolling block is increasing period by period. The number of our characteristic B-spline bases is a function of the number of firms n in each block, which is $\lfloor n^{0.3} \rceil$. Therefore, the dimension of mispricing coefficient vector $\hat{\mathbf{A}} \in \mathcal{R}^{PH_n}$ is also diverging. This verifies the necessity of using power enhanced component S_0.

Recalling that $\mathbf{S}|\mathbf{H_0} \rightarrow^{\mathbf{d}} \mathbf{N(0,1)}$, some of the test statistics S are big enough to reject the null hypothesis. However, for some testing windows, there are no strong signals showing the existence of characteristics-based mispricing functions after subtracting systematic effects. Moreover, most S_1 values are small and even negative, which may be caused by the sparsity structure of the mispricing function or/and the low power problems due to diverging dimension of mispricing coefficients.

The power enhanced component S_0 works well in the empirical study. It selects the most important explaining characteristics and strengthens the power of S_1, mitigating the low power problem.

Apart from contributing to the power of tests, the indicator function in the power enhanced component can also screen out the most relevant explanatory characteristics, which are concluded as "Characteristics Selected" in Table 5.

Some empirical findings are worth discussing. Although short-term cumulative returns like r_{2_1} are always selected, we cannot take this as evidence of arbitrage opportunities as we construct r_{2_1} as time-invariant monthly average of the last month returns. Higher average one month lagged returns imply higher monthly returns. However, this is not the case for long-term and mid-term cumulative returns like r_{12_2}, r_{12_7} and r_{6_2}, because these average returns of these variables contain a lot of information from another rolling window.

Apart from the cumulative returns, some other characteristics contribute to the arbitrage opportunities as well. PCM (Price to Cost Margin) appears twice. From Figure 2, we find that the PCM 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 zero in Figure 3. As for Lev (ratio of long-term debt and debt in the current liabilities), it is decreasing for Lev<0 and increasing afterwards as in Figure 7. In Figure 8, IPM (pre-tax profit margin) function behaves like a "V" shape with the turning point zero during 2004-2005. DelGmSale (Difference in the percentage in gross margin and the percentage change in sales) experiences a bump at the zero during 2015-2016 in Figure 9. C2D curve behaves like "V" around the zero in 2016-2017, (see Figure 10). All characteristics in above figures are standardized as uniform distributed characteristics in the interval $[-100, 100]$. This is for presentation purpose only since most characteristics are unevenly distributed.

Table 5: 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.

Table 6: First layer 1986-1987 (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

Another finding is the persistence of some arbitrage characteristics. Arbitrage characteristics can be persistent for two years once appear, such as BEME (Ratio of the book value of equity and market value of equity) in Figure 4. Some persistent arbitrage characteristics even have similar shapes of mispricing functions in different rolling windows, such as AT (Total asset) in Figure 6 and LME (Total market capitalization of the previous month) in Figure 5.

8.5 Dynamic Peer Groups of Arbitrage Characteristics

In this section, we illustrate that there are distinguishable peer groups of the same arbitrage characteristics resulting in similar unsystematic returns. We apply the methods in section 5 and take two rolling windows, namely, Jul.1986- Jun.1987 and Jul.2004-Jun.2005 as demonstrative examples.

In the rolling window Jul.1986-Jun.1987, PCM is selected as only arbitrage characteristic that explains arbitrage returns. We reveal that similar characteristic-based arbitrage returns are determined by distinguishable groups of the characteristic PCM. We first divide arbitrage returns $\ddot{y}_{it}$ into different return groups. And then, we detect whether there are some clustering structures within groups of the highest and the lowest of characteristic-based 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.

In Table 6, group 2 has the largest positive average return while group 6 has the worst. Next, we detect the clusterings of characteristic "PCM" within each group individually, which is the second layer in section 5.

In Table 7, there are two clusterings of PCM that provide the highest positive characteristic-based arbitrage returns. Group 2.2, which has an extreme negative PCM value but a high characteristic-based arbitrage return, is

Table 7: Second layer 1986-1987 (clusterings of characteristic PCM)

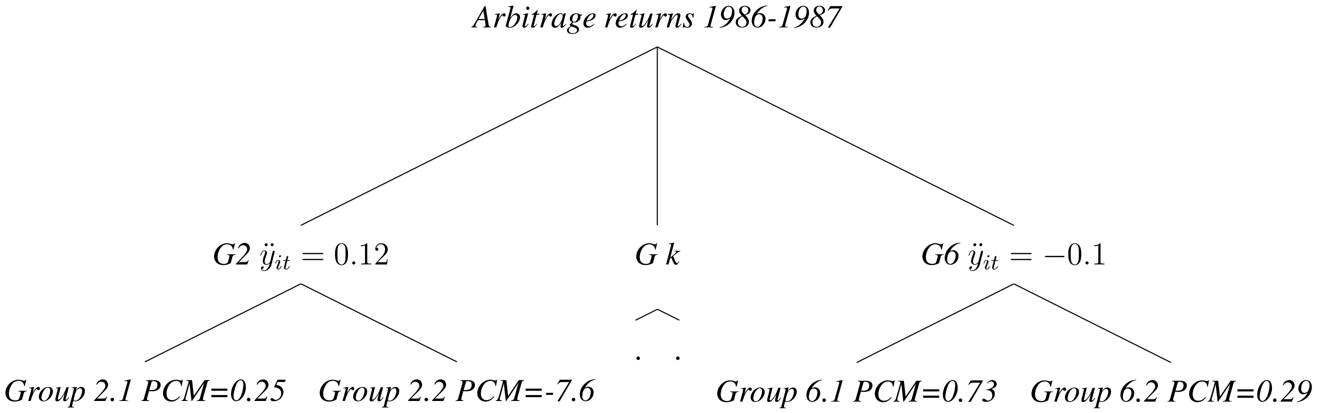
Group number	Centeroids of Arbitrage returns	Centeroids of PCM	Group size
2.1	0.1211	0.2452	25
2.2	0.1039	-7.630	1

Table 8: Second layer 1986-1987 (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

an outlier. Members in group 2.1 with excellent arbitrage performance have positive and small PCM values.

Table 8 gives groups of PCM in 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 close to that in group 2.2 with the highest arbitrage return. This is an evident illustration of the nonlinear strucure of $h(\mathbf{X})$ in this window. The structure of characteristic-based arbitrage returns during Jul.1986- Jun.1987 is:



The classification can be found at Figure 11, where assets are labeled by their "PERMNO," and both axes are rescaled.

Another example is the characteristic-based arbitrage return $\ddot{y}_{it}$ during the year 2004-2005. Power enhanced test selects characteristic "IPM" as the only explanatory variable.

We apply the Hierarchical K-means method. The results of the first layer classification can be found in Table 9. There are ten groups in total according to the similarity of charateristic-based arbitrage returns. Next, we pick two groups with the highest and the lowest returns, respectively, to check clusters of "IPM" in these two groups.

Similarly, we show classification results in Table 10 and Table 11. Positive IPM values give higher characteristic-based arbitrage returns. On the contrary, when IPM is close to zero or negative, the characteristic-based arbitrage returns fall into the lowest group (group 8).

Table 9: First layer (clusterings of $\ddot{y}_{it}$)

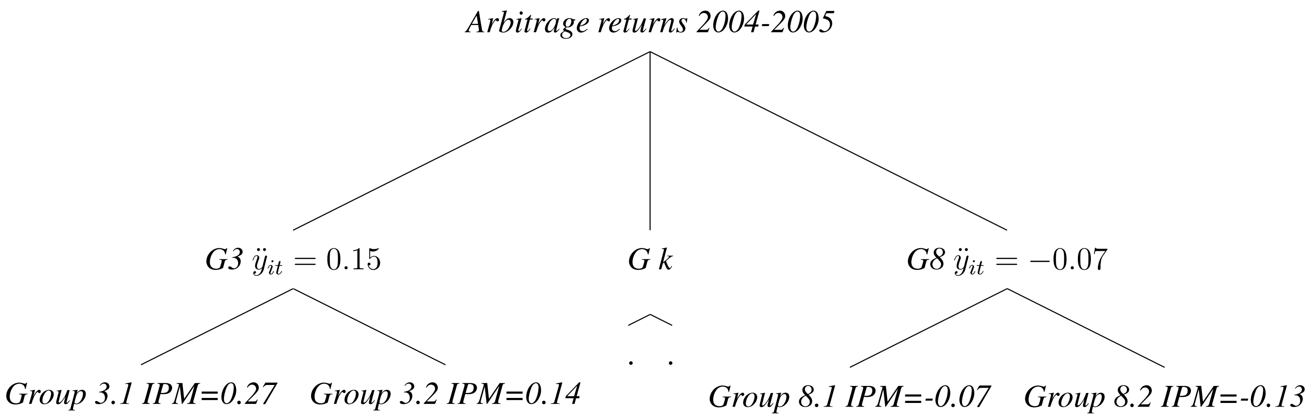
Group number	Group centeroid	Group size
1	0.0421	276
2	0.0059	459
3	0.1537	26
4	-0.024	367
5	0.0659	166
6	0.023	387
7	0.0999	120
8	-0.0758	67
9	-0.0437	244
10	-0.0082	436

Table 10: 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 11: 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


Arbitrage returns 2004-2005
G3 $\ddot{y}_{it} = 0.15$
G k
G8 $\ddot{y}_{it} = -0.07$
Group 3.1 IPM=0.27
Group 3.2 IPM=0.14
Group 8.1 IPM=-0.07
Group 8.2 IPM=-0.13

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

Finally, it is obvious that peer groups of arbitrage characteristics are dynamic in two aspects. Firstly, the selected arbitrage characteristics are time-varying. Although some of the arbitrage characteristics can show up for more than one block once appear, no arbitrage characteristic can be substantially persistent. Secondly, as in Figure 4, the same arbitrage characteristic can have different function forms in various rolling windows. However, the patterns of some characteristics show strong persistence in different time periods, such as AT in Figure 6 and LME in Figure 5. In a word, under the flexible semiparametric setting, methods for contructing arbitrage portfolio in Kim et al. (2019) may be improvable, although the characteristic-based mispricing function is significant for certain time periods. The arbitrage portfolios can perform better by considering peer groups of arbitrage characteristics.

9 Conclusion

We proposed a semi-parametric characteristics-based factor model, with a focus on the existence and structure of the mispricing function. Both unknown characteristics-based factor loadings and the mispricing component are approximated by B-spline sieve. We also develops a power enhanced test to investigate whether there are mispricing components, orthogonal to the main systematic factors. This is necessary because when the B-spline coefficients of the mispricing functions are diverging, the conventional Wald test has very low power. Our proposed methods work well for both simulations and the US stock market. Empirically, we found distinct clusters of the same characteristics resulting in similar arbitrage returns. The mispricing function and selected arbitrage characteristics are time-varying. We conclude that the traditional way of developing arbitrage portfolios can be improved by considering peer groups of arbitrage characteristics

10 Appendix

10.1 Characteristic Description

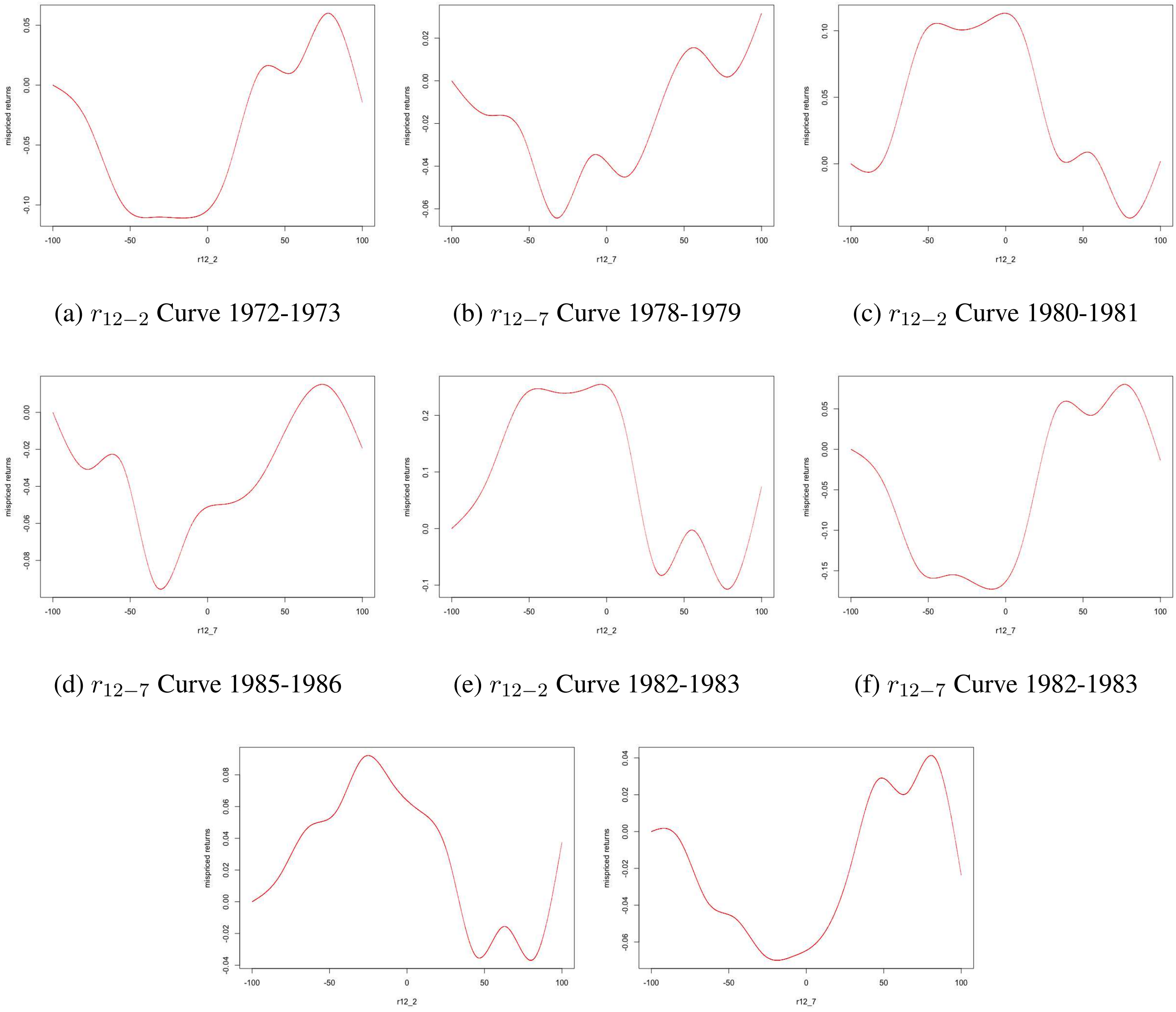
Table 12: Characteristic Details

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)
LME	Size is the total market capitalization of the previous month defined as price (PRC) times shares outstanding (SHROUT)	Fama and French (1992)
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).	



(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

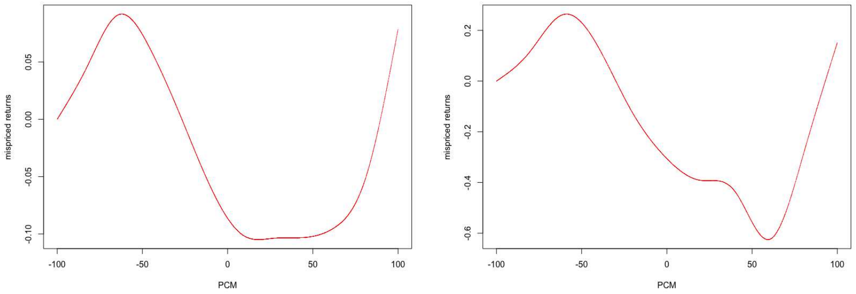
(e) r_{12-2} Curve 1982-1983

(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 Curve of standardized r_{12-2} and r_{12-7}



(a) PCM Curve 1984-1985

(b) PCM Curve 2016-2017

Figure 2: Mispricing Characteristic Curve of standardized PCM

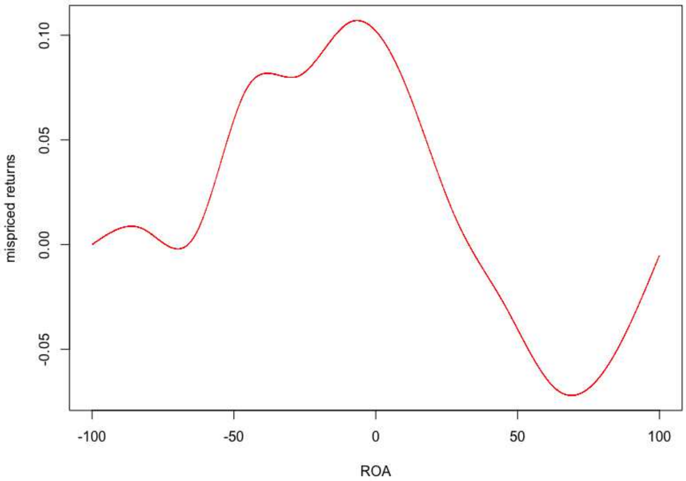
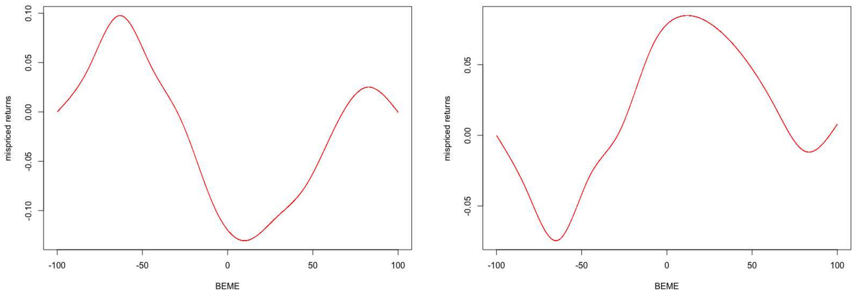


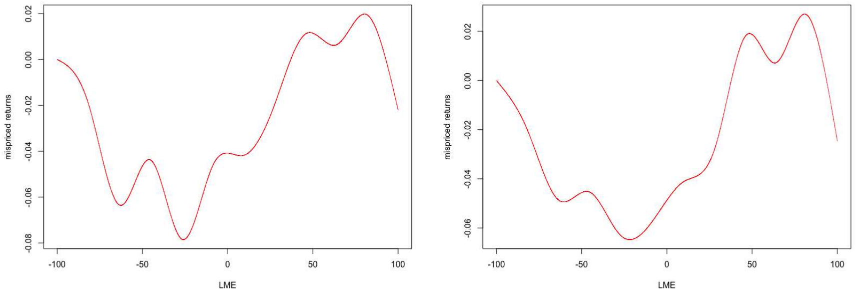
Figure 3: Mispricing Characteristic Curve of standardized ROA in 1988-1989



(a) BEME Curve 1995-1996

(b) BEME Curve 1996-1997

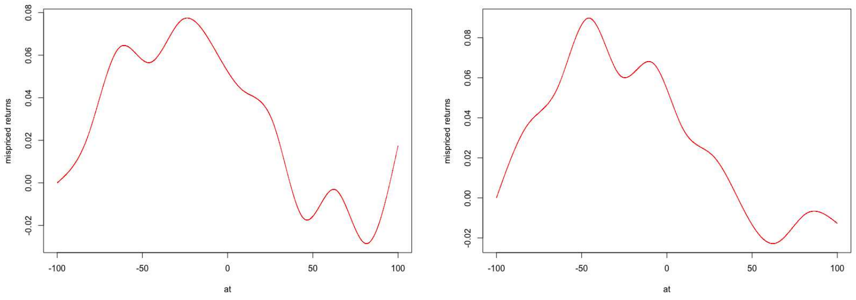
Figure 4: Mispricing Characteristic Curve of standardized BEME



(a) LME Curve 1998-1999

(b) LME Curve 2000-2001

Figure 5: Mispricing Characteristic Curve of standardized LME



(a) AT Curve 1998-1999

(b) AT Curve 2000-2001

Figure 6: Mispricing Characteristic Curve of standardized AT

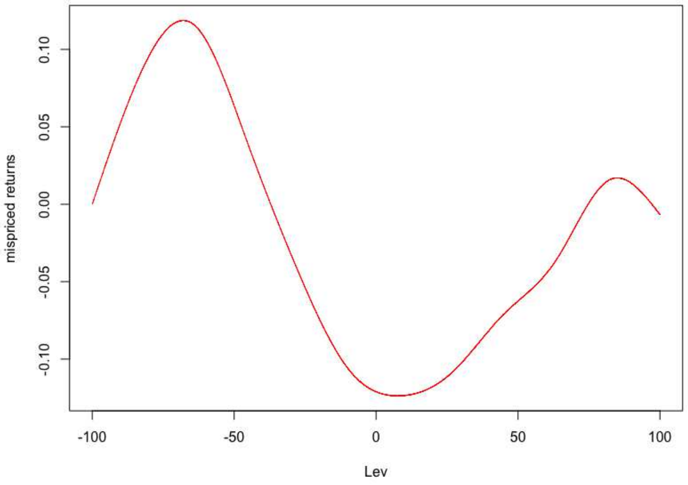


Figure 7: Mispricing Characteristic Curve of standardized LEV in 2002-2003

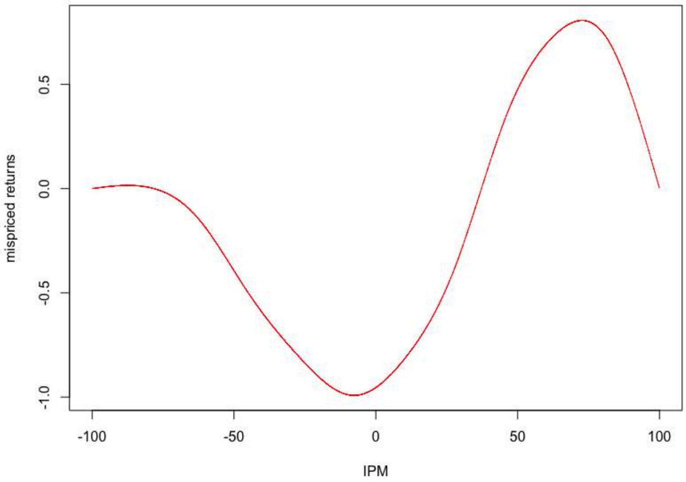


Figure 8: Mispricing Characteristic Curve of standardized IPM in 2004-2005

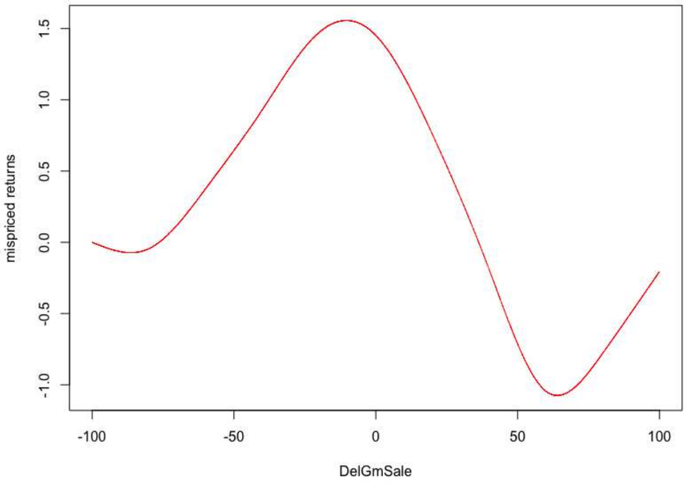


Figure 9: Mispricing Characteristic Curve of standardized DelGmSale in 2015-2016

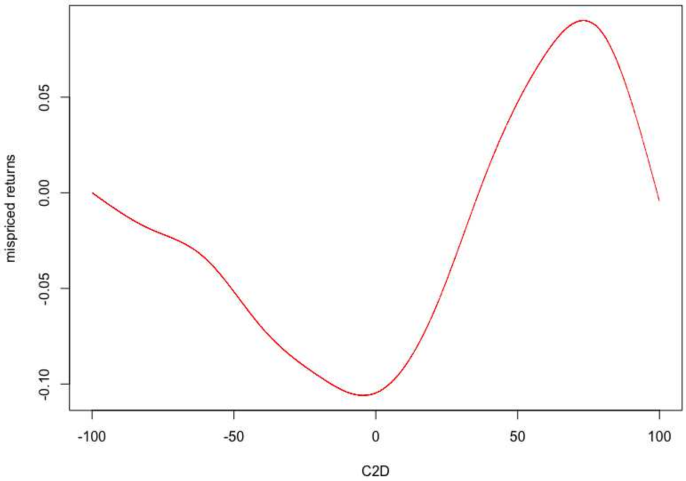
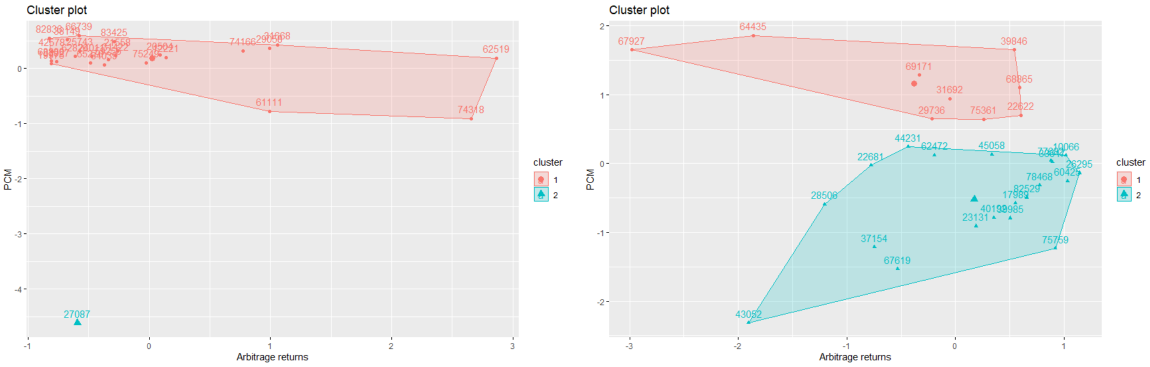


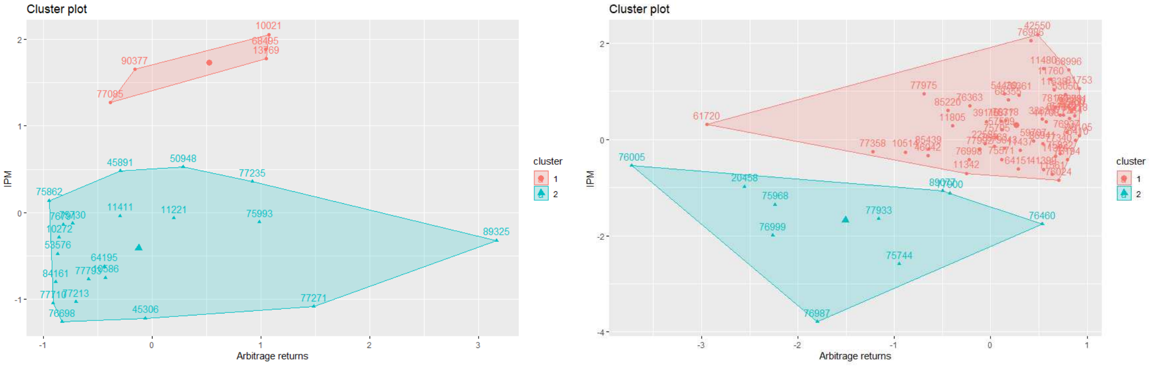
Figure 10: Mispricing Characteristic Curve of standardized C2D in 2016-2017



(a) Clustering of PCM with highest returns

(b) Clustering of PCM with lowest returns

Figure 11: Clustering of PCM 1986-1987



(a) Clustering of IPM with highest returns

(b) Clustering of IPM with lowest returns

Figure 12: Clustering of IPM 2004-2005

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

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

Multiply time-demeaned matrix $\mathbf{D_T}$ on both sides, where $\mathbf{D_T} = \mathbf{I_T} - \frac{1}{T}\mathbf{1}_{\mathbf{T}}^{\intercal}\mathbf{1_T}$. Given time-invariant mispricing components, we obtain:

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

Onwards, we define $\mathbf{Y}\mathbf{D_T} = \tilde{\mathbf{Y}}$ and $\mathbf{F}^{\intercal} = \mathbf{F}^{\intercal}\mathbf{D_T}$. Time-demeaned factors do not change their properties.

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

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

We decompose:

$$\mathbf{P}\tilde{\mathbf{Y}} = \hat{\mathbf{Y}} = \mathbf{\Phi}(\mathbf{X})\mathbf{B}\mathbf{F}^{\intercal} + \mathbf{P}\mathbf{\Lambda}\mathbf{F}^{\intercal} + \mathbf{P}\mathbf{U}\mathbf{D_T} + \mathbf{P}\mathbf{R}^{\theta}(\mathbf{X})\mathbf{F}^{\intercal} = \mathbf{e_1} + \mathbf{e_2} + \mathbf{e_3} + \boldsymbol{e_4},$$

as $n \to \infty$ and $n^{v} \to \infty$, approximation error $\mathbf{R}^{\theta}(\mathbf{X}) \to_{P} \mathbf{0}$ as in Huang et al. (2010). Thus, $\mathbf{e}_{\mathbf{4}}^{\intercal} \to^{P} \mathbf{0}$.

Under Assumption 1, we have following results:

for $\frac{1}{n}\sum_{j=1}^{3} \boldsymbol{e}_{\mathbf{2}}^{\intercal}\boldsymbol{e}_{j}$,

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

therefore,

$$\frac{1}{n}\sum_{j=1}^{3}\mathbf{e}_{\mathbf{2}}^{\intercal}\mathbf{e_j} + \frac{1}{n}\sum_{j=1}^{3}\mathbf{e}_{\mathbf{j}}^{\intercal}\mathbf{e_2} \to^{P} \mathbf{0}.$$

For $\frac{1}{n}\sum_{j=1}^{3} \boldsymbol{e}_{\mathbf{3}}^{\intercal}\boldsymbol{e}_{j}$,

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

therefore,

$$\frac{1}{n}\sum_{j=1}^{3}\mathbf{e}_{\mathbf{2}}^{\intercal}\mathbf{e_j} + \frac{1}{n}\sum_{j=1}^{3}\mathbf{e}_{\mathbf{j}}^{\intercal}\mathbf{e_2} \to^{P} \mathbf{0}.$$

And only $\frac{1}{n}\mathbf{e}_{\mathbf{1}}^{\intercal}\mathbf{e_1}$ left, namely,

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

Under Assumption 2-4 and fixed T. A much smaller $T \times T$ matrix $\frac{1}{n}\hat{\mathbf{Y}}^\intercal\hat{\mathbf{Y}}$ can be sovled by asymptotic principal component by Connor and Korajczyk (1986). $\hat{\mathbf{F}} = \frac{1}{\sqrt{T}}\{\psi_1, \psi_2, \dots, \psi_J\}$, where$\{\psi_1, \psi_2, \dots, \psi_J\}$ are eigenvectors corresponding to the first J eigenvalues of $\frac{1}{n}\hat{\mathbf{Y}}^\intercal\hat{\mathbf{Y}}$.

Thus, $\hat{\mathbf{F}} \rightarrow_P \mathbf{F}$ 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 $\hat{\mathbf{F}}^\intercal\hat{\mathbf{F}} = \mathbf{I_J}$, 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:

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

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

Furthermore,

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

Similar to the Proof of Theorem 6.1,

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

For the first term,

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

due to

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

and

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

then,

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

Theorem 6.1 and Assumption 2 give $\hat{\mathbf{F}} \rightarrow \mathbf{F}$ and $\mathbf{F}^{\mathbf{T}}\mathbf{F} = \mathbf{I}_{\mathbf{J}}$, therefore:

$$\frac{1}{n}\mathbf{q}_1^{\mathbf{T}}\mathbf{q}_1 \rightarrow^{P} \frac{\mathbf{B}^{\intercal}\mathbf{\Phi}^{\intercal}(\mathbf{X})\mathbf{\Phi}(\mathbf{X})\mathbf{B}}{\mathbf{n}},$$

Similarly,

$$\frac{1}{n}\mathbf{q}_3^{\mathbf{T}}\mathbf{q}_3 \rightarrow^{P} \frac{\mathbf{B}^{\intercal}\mathbf{\Phi}^{\intercal}(\mathbf{X})\mathbf{\Phi}(\mathbf{X})\mathbf{B}}{\mathbf{n}},$$
$$\frac{1}{n}\mathbf{q}_1^{\mathbf{T}}\mathbf{q}_3 \rightarrow^{P} -\frac{\mathbf{B}^{\intercal}\mathbf{\Phi}^{\intercal}(\mathbf{X})\mathbf{\Phi}(\mathbf{X})\mathbf{B}}{\mathbf{n}},$$
$$\frac{1}{n}\mathbf{q}_3^{\mathbf{T}}\mathbf{q}_1 \rightarrow^{P} -\frac{\mathbf{B}^{\intercal}\mathbf{\Phi}^{\intercal}(\mathbf{X})\mathbf{\Phi}(\mathbf{X})\mathbf{B}}{\mathbf{n}}.$$

Therefore,

$$\frac{1}{n}\mathbf{q}_1^{\intercal}\mathbf{q}_1 + \frac{\mathbf{1}}{\mathbf{n}}\mathbf{q}_3^{\intercal}\mathbf{q}_3 + \frac{\mathbf{1}}{\mathbf{n}}\mathbf{q}_1^{\intercal}\mathbf{q}_3 + \frac{\mathbf{1}}{\mathbf{n}}\mathbf{q}_3^{\intercal}\mathbf{q}_1 \rightarrow 0.$$

Then,

$$\frac{1}{n}((\hat{\mathbf{G}}(\mathbf{X}) - \mathbf{G}(\mathbf{X})))^{\intercal}((\hat{\mathbf{G}}(\mathbf{X}) - \mathbf{G}(\mathbf{X}))) \rightarrow^{P} 0$$

thus,

$$\hat{\mathbf{G}}(\mathbf{X}) \rightarrow^{\mathbf{P}} \mathbf{G}(\mathbf{X}).$$

Then Theorem 6.2 follows.

□

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

$$\min_{\mathbf{A}}(\dot{\mathbf{Y}} - \mathbf{\Phi}(\mathbf{X})\mathbf{A})^{\intercal}(\dot{\mathbf{Y}} - \mathbf{\Phi}(\mathbf{X})\mathbf{A}) + \lambda\hat{\mathbf{G}}^{\intercal}(\mathbf{X})\mathbf{\Phi}(\mathbf{X})\mathbf{A} \tag{8}$$

Then we take the first order condition with respect to $\mathbf{A}$ and λ separately, and we obtain:

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

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

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

Therefore, we obtain:

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

where

$$\mathbf{M} = \mathbf{I} - (\Phi(\mathbf{X})^{\intercal}\Phi(\mathbf{X}))^{-1}\mathbf{\Phi}(\mathbf{X})^{\intercal}\hat{\mathbf{G}}(\mathbf{X})(\hat{\mathbf{G}}(\mathbf{X})^{\intercal}\hat{\mathbf{G}}(\mathbf{X}))^{-1}\hat{\mathbf{G}}(\mathbf{X})^{\intercal}\mathbf{\Phi}(\mathbf{X}),$$
$$\tilde{\mathbf{A}} = \frac{\mathbf{1}}{\mathbf{T}}(\Phi(\mathbf{X})^{\intercal}\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})^{\intercal}\dot{\mathbf{Y}}\mathbf{1}_{\mathbf{T}}.$$

Furthermore, let $\Xi = \mathbf{\Phi}(\mathbf{X})\hat{\mathbf{A}} - \mathbf{h}(\mathbf{X}) = \mathbf{\Phi}(\mathbf{X})\mathbf{M}\tilde{\mathbf{A}} - \mathbf{\Phi}(\mathbf{X})\mathbf{A} - \mathbf{R}^{\mu}(\mathbf{X})$.

Under the restriction $\hat{\mathbf{G}}'(\mathbf{X})\mathbf{\Phi}(\mathbf{X})\mathbf{A} = \mathbf{0}$, we can obtain:

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

Furthermore, we have:

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

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

$$\mathbf{\Xi} = \mathbf{\Phi}(\mathbf{X})\mathbf{A} - \mathbf{\Phi}(\mathbf{X})\mathbf{A} - \mathbf{R}^{\mu}(\mathbf{X}).$$

$$\mathbf{R}^{\mu}(\mathbf{X}) \to \mathbf{0} \text{ as } n \to \infty,$$

therefore,

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

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

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

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

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

equivalently, we need to prove

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

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

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

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

Meanwhile $F_q = O(1)$, according to Fan et al. (2015) and Kock and Preinerstorfer (2019), we can show that:

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

Then the Theorem 6.5 follows. □

References

Carhart, M. M. (1997), 'On persistence in mutual fund performance', *The Journal of finance* **52**(1), 57–82.

Chen, X. and Pouzo, D. (2012), 'Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals', *Econometrica* **80**(1), 277–321.

Connor, G., Hagmann, M. and Linton, O. (2012), 'Efficient semiparametric estimation of the fama–french model and extensions', *Econometrica* **80**(2), 713–754.

Connor, G. and Korajczyk, R. A. (1986), 'Performance measurement with the arbitrage pricing theory: A new framework for analysis', *Journal of financial economics* **15**(3), 373–394.

Connor, G. and Linton, O. (2007), 'Semiparametric estimation of a characteristic-based factor model of common stock returns', *Journal of Empirical Finance* **14**(5), 694–717.

Cox, D. R. (1957), 'Note on grouping', *Journal of the American Statistical Association* **52**(280), 543–547.

Fama, E. F. and French, K. R. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of financial economics* **33**(1), 3–56.

Fama, E. F. and French, K. R. (2015), 'A five-factor asset pricing model', *Journal of financial economics* **116**(1), 1–22.

Fan, J., Liao, Y. and Mincheva, M. (2013), 'Large covariance estimation by thresholding principal orthogonal complements', *Journal of the Royal Statistical Society. Series B, Statistical methodology* **75**(4).

Fan, J., Liao, Y. and Wang, W. (2016), 'Projected principal component analysis in factor models', *Annals of statistics* **44**(1), 219.

Fan, J., Liao, Y. and Yao, J. (2015), 'Power enhancement in high-dimensional cross-sectional tests', *Econometrica* **83**(4), 1497–1541.

Feng, G., Giglio, S. and Xiu, D. (2017), 'Taming the factor zoo', *Fama-Miller Working Paper* **24070**.

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

Freyberger, J., Neuhierl, A. and Weber, M. (2017), Dissecting characteristics nonparametrically, Technical report, National Bureau of Economic Research.

Hjalmarsson, E. and Manchev, P. (2012), 'Characteristic-based mean-variance portfolio choice', *Journal of Banking & Finance* **36**(5), 1392–1401.

Hoberg, G. and Phillips, G. (2016), 'Text-based network industries and endogenous product differentiation', *Journal of Political Economy* **124**(5), 1423–1465.

Hou, K., Xue, C. and Zhang, L. (2015), 'Digesting anomalies: An investment approach', *The Review of Financial Studies* **28**(3), 650–705.

Huang, J., Horowitz, J. L. and Wei, F. (2010), 'Variable selection in nonparametric additive models', *Annals of statistics* **38**(4), 2282.

Kelly, B. T., Pruitt, S. and Su, Y. (2017), 'Instrumented principal component analysis', *Available at SSRN 2983919* .

Kelly, B. T., Pruitt, S. and Su, Y. (2019), 'Characteristics are covariances: A unified model of risk and return', *Journal of Financial Economics* .

Kim, S., Korajczyk, R. A. and Neuhierl, A. (2019), 'Arbitrage portfolios', *The Review of Financial Studies* .

Kock, A. B. and Preinerstorfer, D. (2019), 'Power in high-dimensional testing problems', *Econometrica* **87**(3), 1055–1069.

Ledoit, O., Wolf, M. et al. (2012), 'Nonlinear shrinkage estimation of large-dimensional covariance matrices', *The Annals of Statistics* **40**(2), 1024–1060.

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

Pesaran, M. H. and Yamagata, T. (2012), Testing capm with a large number of assets, *in* 'AFA 2013 San Diego Meetings Paper'.

Pollard, D. (1981), 'Strong consistency of k-means clustering', *The Annals of Statistics* pp. 135–140.

Pollard, D. et al. (1982), 'A central limit theorem for k-means clustering', *The Annals of Probability* **10**(4), 919–926.

Ross, S. (1976), 'The arbitrage theory of capital asset pricing', *Journal of Economic Theory* **13**(3), 341–360.

Sun, W., Wang, J., Fang, Y. et al. (2012), 'Regularized k-means clustering of high-dimensional data and its asymptotic consistency', *Electronic Journal of Statistics* **6**, 148–167.

Vogt, M. and Linton, O. (2017), 'Classification of non-parametric regression functions in longitudinal data models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(1), 5–27.