Improved inference for fund alphas using high-dimensional cross-sectional tests[☆]

Abstract

The traditional fund-by-fund alpha inference suffers from various econometric problems (e.g., cross-sectional independence assumption, lack of power, time-invariant coefficient assumption, multiple-hypothesis-testing). Recognizing the panel nature of fund industries, we tailor four high-dimensional cross-sectional tests to shed light into both the zero-alpha hypothesis and ratio of non-zero alphas. Particularly, we augment Gagliardini et al. (2016) with a time-varying alpha estimator. Our results reject the zero-alpha joint hypothesis as the statistical significance of alphas is too high to be explained by luck. After controlling for luck, our empirical studies show that the power enhancement helps to identify a large portion of significant fund alphas, which cannot be achieved using the usual Wald test. Meanwhile, the time-varying approach shows that fund alphas diverge during the late 2000s Global Financial Crisis, which cannot be observed using the time-invariant model. Overall, relative to the literature, we draw a more accurate and complete picture, and provide several powerful tools for future research.

Keywords: Alpha, Mutual fund, Hedge fund, Crisis, Arbitrage

JEL Classification: C15, G11, G12, G23

'A pervasive problem with performance evaluation is the presence of similar strategies among funds, which produces correlated residuals from commonly used models and, therefore, reduces the power of such models to separate skilled from unskilled fund managers.'

Hunter et al. (2014)

1. Introduction

The zero-alpha hypothesis and ratio of non-zero alphas in fund industries are of essential importance in Asset Pricing.[1] Since the seminal work of Jensen (1968), numerous studies regress fund returns on a linear combination of pricing factors via Ordinary Least Squares (OLS) and take the unexplained constant (alpha) as evidence of skill.[2] However, this approach may suffer from various distortions in statistical size and power, and cause misleading results. The statistical size of a test is the probability of falsely rejecting the null hypothesis ($H_0 : \alpha = 0$) or the probability of making a Type I error (false discovery rate). The statistical power of a test is the probability that the test correctly rejects the null hypothesis when the alternative hypothesis (H_1) is true (i.e., one minus probability of Type II error). Importantly, the statistical size and power of a test interact with each other, and it is difficult to improve one aspect while keeping the other aspect intact. The recent literature has presented several challenges to the estimation of fund alphas from both perspectives of statistical size and power.

---

[1] See, e.g., Kosowski et al. (2006); Barras et al. (2010); Fama and French (2010); Pástor and Stambaugh (2012); Pástor et al. (2015); Ferson and Chen (2020); Harvey and Liu (2018, 2019*a*); Barras et al. (2018); Andrikogiannopoulou and Papakonstantinou (2019).

[2] Although Berk and van Binsbergen (2015) argue that fund skills should be measured by the value-added (i.e., gross alpha multiplying fund size) instead of fund alphas, researchers still need to estimate alpha before computing the value-added measure. We follow the literature and focus on the net alphas in our benchmark analyses but gross alphas and value-added measure in our robustness section.

First of all, the traditional approach ignores the valuable inter-fund information in the cross-section such as Cross-Sectional Dependence (CSD) in residuals,[3] although a cross-section of funds is a panel by nature (e.g., Blake et al. (2014)). On the one hand, Barras et al. (2010), Ferson and Chen (2020), and Andrikogiannopoulou and Papakonstantinou (2019) demonstrate the importance of increasing statistical power by injecting non-zero alphas,[4] while Harvey and Liu (2018) use a power-enhanced procedure and find a much larger ratio of outperforming (i.e., 10%) mutual funds than the extant literature.[5] On the other hand, Cheng and Yan (2017) demonstrate that the size of the popular bootstrap performance evaluation approach can be distorted without properly controlling for CSD.[6] As funds have a limited length of overlapping period, the bootstrap approach only has limited capability to control for CSD.

Moreover, size distortion may also arise from the common practice of treating fund alpha as a constant, although fund alphas (and betas) are affected by many factors and unlikely to be a constant. This time-invariant coefficient assumption holds at neither asset nor portfolio level (see, e.g., Ang and Kristensen (2012) and the references therein). Mamaysky et al. (2007), Mamaysky et al. (2008) and Cai et al. (2018) have justified time-varying against time-invariant

[3]For instance, Petersen (2009) highlights the problems induced by residual dependence in finance panel data when calculating standard errors.

[4]Using the same False Discovery Rate (FDR) framework, Barras et al. (2010), Ferson and Chen (2020), and Andrikogiannopoulou and Papakonstantinou (2019) assume different degrees of statistical power. Their results differ from each other, and echo recent research in economics highlighting the importance of examining test power (e.g., Ioannidis et al. (2017) find under-powered tests in many research areas with exaggerated results.)

[5]Harvey and Liu (2018) argue that the result from traditional models *'is likely biased given the limited power of the test of an individual fund's alpha. The fund-by-fund approach ignores valuable information in the cross-section that can potentially improve the inference on the cross-sectional distribution of fund alphas.'*

[6]While Kosowski et al. (2006) find skilled funds outperforming the market, Fama and French (2010) overturn the conclusion by refining the bootstrap approach in Kosowski et al. (2006) via controlling for CSD.

alphas and betas with both theoretical and empirical evidence. For instance, Mamaysky et al. (2008) document that the traditional OLS models *'produce false positives (nonzero alphas) at too high a rate with either daily or monthly data'*.

In addition, the primary reason for size distortion is the multiple-hypothesis-testing problem.[7] However, since there are thousands of funds in the fund industry, even by luck some funds are prone to be associated with significant alphas from the traditional fund-by-fund OLS approach. The idea to mitigate this multiplicity problem typically involves choosing a test statistic (which may be based on the values or the t-statistic of alphas) to strike a balance between Type I and Type II errors: either via bootstrap methods (e.g., Kosowski et al. (2006, 2007); Fama and French (2010); Blake et al. (2013, 2014, 2017)),[8] or controlling for false discoveries (e.g., Barras et al. (2010); Ferson and Chen (2020)),[9] or applying a 'haircut' to the t-statistic of individual fund alphas (Harvey et al., 2016, 2020).

To shed light into both the zero-alpha hypothesis and the ratio of non-zero alphas for both equity-oriented mutual funds and hedge funds in the U.S., we tackle the problems above altogether based on high-dimensional cross-sectional tests from Pesaran and Yamagata (2018), Fan et al. (2015), Gagliardini et al. (2016) and Ma et al. (2020). We account for the panel nature of funds as well as cross-sectional inter-fund information, which was typically neglected

---

[7]The multiple-hypothesis-testing problem applies not only to fund performance evaluation, but also to asset pricing in general (e.g., Harvey et al. (2016); Harvey (2017); Harvey and Liu (2019*b*, 2018, 2020); Chordia et al. (2020); Yan and Zheng (2017); Barras (2019); Giglio et al. (2019)) It also applies to other subjects in both natural science and social science. For instance, it is called 'looking elsewhere' in physics, 'multiple comparisons' in medical science especially in genetic association studies, and 'data mining, data snooping or multiple testing' in Economics and Finance (Harvey (2017); Harvey and Liu (2019*b*)).

[8]The bootstrap method is not always reliable (see, e.g., Cheng and Yan (2017), Zhang and Yan (2018)).

[9]The False Discovery Rate (FDR) method hinges on the correct assumption of the number of fund subgroups, which is empirically challenging (Yan and Cheng, 2019).

in the extant literature. To capture the time-varying fund alphas, we not only implement the tests from Pesaran and Yamagata (2018) and Fan et al. (2015) in a rolling fashion, but also augment the procedure in Gagliardini et al. (2016) with a time-varying alpha estimator. To tackle the multiple-hypothesis-testing problem, we rely on simulations in Appendix A and apply a 'haircut' to the t-statistic of individual fund alphas by using a threshold of t-statistic 3.

To control for CSD, Pesaran and Yamagata (2018) propose a statistic based on other existing methods in the literature (Gibbons et al., 1989; Ledoit and Wolf, 2004; Fan et al., 2013, 2015; Gagliardini et al., 2016; Gungor and Luger, 2016), especially when N is larger than T and/or when there is strong cross-sectional dependence among residuals. The pivotal Wald-type statistic from Pesaran and Yamagata (2018) can be used to test whether all alphas are jointly zero, and we use their proposed statistic as our starting point to control for CSD.

To boost the statistical power of testing a high-dimensional vector ($H_0 : \alpha = 0$) against sparse alternatives, Fan et al. (2015) propose a screening technique to construct a power enhancement component, which is zero under the null with a large probability, but diverges to non-zeros quickly under the sparse alternatives. It does not strike a balance between Type I error and Type II error, but enhance the statistical power without sacrificing the statistical size asymptotically.

To infer time-varying risk premia from a large unbalanced panel of stock returns, Gagliardini et al. (2016) develop two-pass conditional linear asset pricing models. They derive empirical testable no-arbitrage pricing restrictions (Ross, 1976) in a multi-period economy, and provide a test of the restrictions for conditional factor models. They account for the unbalanced panel nature of individual assets, and allow for time-varying betas. While the original purpose of their model is to estimate time-varying equity risk premia, we extend their time-varying beta estimator to time-varying alpha estimator and further derive the associated asymptotic distribution, which constitutes our methodological contribution.

To assess the efficiency of stock markets, Ma et al. (2020) propose a high-dimensional test for long-run alphas (i.e., simple average of time-varying alphas) while allowing for time-varying factor loadings. Different from the structural framework of Gagliardini et al. (2016), Ma et al. (2020) consider a time-varying factor model in which the factor loadings are assumed to be unknown smooth functions of time t. To ensure the robustness of our results, we use this 'smoothing approach' as an alternative approach to dealing with time-varying alphas.

Applying these methods to both U.S. equity-orientated mutual funds and hedge funds, our results strongly reject the zero-alpha joint hypothesis (Jensen, 1968; Carhart, 1997; Fama and French, 2010; Blake et al., 2014, 2017). Interestingly, we find some evidence supporting the existence of some arbitrage opportunities among mutual funds, but little evidence supporting the existence of arbitrage opportunities among hedge funds, which suggests that it is profitable for Fund-of-Fund (FoF) managers to build up their portfolios using mutual funds rather than hedge funds. The performance of both mutual funds and hedge funds diverges during the late 2000s Global Financial Crisis (GFC).

After controlling for luck, our empirical studies show that the power enhancement helps to identify a large portion of significant fund alphas, which cannot be achieved using the usual Wald test. Meanwhile, the time-varying approach shows that fund alphas diverge during the late 2000s Global Financial Crisis, which cannot be observed using the time-invariant model. Specifically, for mutual funds we find about 5% with significant alphas, and more negative than positive alphas. On the one hand, the number of significant and positive mutual fund alphas increases during the GFC, which suggests that financial crises are not necessarily a bad thing even for mutual fund managers. On the other hand, the substantially increased significant and negative mutual fund alphas with larger magnitudes provide direct evidence for the fire sales of mutual funds during the GFC and hence transmitting crises (see, e.g., Jotikasthira et al. (2012); Raddatz and Schmukler (2012); Hau and Lai (2017)).

Hedge funds perform better than mutual funds on average, and both mutual funds and hedge funds suffer from the GFC given the number and portion of negative alphas increases for both. After controlling for luck, for hedge funds we find a larger fraction (about 10%) with significant alphas, and more positive than negative alphas. The average alpha of our selected hedge funds increases during the GFC, which persists during the post-crisis period even when both the number and portion of positive alphas decrease. This finding is new and providing fresh evidence supporting the conjecture of hedge funds as arbitrageurs and some of them reap huge profits during crises (e.g., Ben-David et al. (2012); Jiao (2012); Akbas et al. (2015); Kokkonen and Suominen (2015); Gao et al. (2018); Aragon et al. (2019)).

We mainly contribute to the extant literature from several aspects. First of all, we are the first to tailor high-dimensional cross-sectional tests for fund performance evaluation, which tackle several prominent econometric problems in the literature and hence provide researchers with more reliable inferences. Moreover, rather than the aggregated/average performance of each fund industry, our methodologies allow us to tease out diverged fund performances during the GFC. Put differently, we provide four new tools to investigate the cross-sectional variation in fund returns, and hence connect to the latest literature about the extreme performances of banks/funds during the GFC (Beltratti and Stulz, 2012; Aragon et al., 2019). In addition, we develop a time-varying alpha estimator and further derive the associated asymptotic distributions, as well as propose a high-dimensional cross-sectional test of the non-arbitrage condition for fund alphas, which is of essential importance for both academia and the practical use of FoF managers. More importantly, we find that the 'smoothing approach' cannot capture the presence of a substantially increased portion of extreme alphas during financial crises but revealed by our model based on Gagliardini et al. (2016), as the time-varying α_{it} in the setting of Gagliardini et al. (2016) can capture potential structural breaks which are ruled out by the 'smoothing approach' as it requires α_{it} change smoothly over time. Finally, we provide an overarching framework to reconcile and interpret many results in the literature on performance evaluation.

The remainder of the paper proceeds as follows. Section 2 presents the methodology. Section 3 and 4 discuss the data and empirical results, respectively. Section 5 briefly discusses the robustness checks we have done, while detailed results are delegated to online appendices. Section 6 concludes.

2. Methodology

This section gives a brief review of the four high-dimensional approaches to tackling the aforementioned potential distortions in the size and power of traditional fund performance evaluation approach: the high-dimensional test for alphas in linear factor pricing models based on Pesaran and Yamagata (2018), the power enhancement method in high-dimensional test based on Fan et al. (2015), the high-dimensional test for time-varying risk premium based on Gagliardini et al. (2016) and the high-dimensional test for long-run alphas based on Ma et al. (2020). The original versions of these four tests are variants of the Wald statistic that tests the joint hypothesis that all stock alphas are not significantly different from zero, but we propose using them to test the zero-alpha joint hypothesis for funds (Jensen, 1968; Carhart, 1997; Fama and French, 2010; Blake et al., 2014, 2017). It is perhaps noteworthy that we propose a new method based on Gagliardini et al. (2016) in Section 2.3, which makes this subsection necessarily longer as we extend the time-varying beta estimator in Gagliardini et al. (2016) to time-varying alpha estimator and further derive the associated asymptotic distribution.

2.1. High-dimensional test for alphas in linear factor pricing models

Pesaran and Yamagata (2018) proposed a J test for testing alphas in linear factor pricing models with a large number of securities. They propose a test procedure for the multivariate tests of $H_0 : \alpha^\top = (\alpha_1, \cdots, \alpha_N) = 0$, when $N > T$, that allows for non-Gaussianity and general forms of weakly cross-correlated errors. Specifically, the individual fund return regressions can be written in the form of

$$r_{it} = \alpha_i + \beta_i^\top f_t + \epsilon_{it}, \quad i = 1, 2, \cdots, N, \; t = 1, 2, \cdots, T, \tag{1}$$

where r_{it} denotes the fund returns; β_i and f_t respectively denote the $m \times 1$ vectors of factor loadings and common factors; α_i measures the abnormal performance for the ith fund. Stacking by time-series observations, we have $r_i = \alpha_i \tau_T + F\beta_i + u_i$, where $r_i = (r_{i1}, \cdots, r_{iT})^\top$, $\tau_T = (1, 1, \cdots, 1)^\top$, $F^\top = (f_1, \cdots, f_T)$ and $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{iT})^\top$.

Define $M_F = I_T - F(F^\top F)^{-1}F^\top$ and $\widehat{W}_d = \left(\tau_T^\top M_F \tau_T\right)\widehat{\alpha}^\top \widehat{D}_v^{-1}\widehat{\alpha}$, where $\widehat{D}_v = (v^{-1}T)\,\mathrm{diag}(\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_N^2)$, $\widehat{\sigma}_i^2 = \widehat{\epsilon}_i^\top \widehat{\epsilon}_i / T$, and the degree of freedom $v = T - m - 1$ is introduced to correct for small sample bias of the test. Pesaran and Yamagata (2018) first present a standardized version of $\widehat{W}_d$, denoted as J_{wald}

$$J_{wald} = \frac{N^{-1/2}[\widehat{W}_d - E(\widehat{W}_d)]}{\sqrt{Var(\widehat{W}_d)}}, \tag{2}$$

in which $E(\widehat{W}_d) = \frac{vN}{v-2}$ and $Var(\widehat{W}_d) = \frac{2N(v-1)}{v-4}\left(\frac{v}{v-2}\right)^2 + 2\sum_{i=2}^{N}\sum_{j=1}^{i} Cov(t_i^2, t_j^2)$, where t_i denotes the standard t-ratio of α_i in the OLS regression of r_{it} on an intercept and f_t (i.e. model (1)). We can see that this test statistic allows for the effects of non-zero off-diagonal elements of the underlying error covariance matrix.

By applying the thresholding approach to $\widehat{\epsilon}_{it}$ and imposing sparsity conditions on error covariance matrix, then Pesaran and Yamagata (2018) propose a feasible version of J_{wald}, which can be computed by

$$\widehat{J}_{wald} = \frac{N^{-1/2}\sum_{i=1}^{N}(t_i^2 - \frac{v}{v-2})}{\frac{v}{v-2}\sqrt{\frac{2(v-1)}{(v-4)}\left[1 + (N-1)\widetilde{\rho}_{N,T}^2\right]}}, \tag{3}$$

where $\widetilde{\rho}_{N,T}^2 = \frac{2}{N(N-1)}\sum_{i=2}^{N}\sum_{j=1}^{i-1}\widetilde{\rho}_{ij}^2$, in which $\widetilde{\rho}_{ij} = \widehat{\rho}_{ij} I\left[|\sqrt{v}\widehat{\rho}_{ij}| > c_p(N)\right]$, $v = T - m - 1$, m

is the number of factors in the regression, $c_p(N) = \Phi^{-1}\left(1 - \frac{p}{2f(N)}\right)$, p is the nominal p-value ($0 < p < 1$), and $f(N) = N^{\delta}$, $T = c_d N^d$, with c_d, δ and d being finite positive constants, $\widehat{\rho}_{ij} = \widehat{\sigma}_{ij}/\widehat{\sigma}_{ii}^{1/2}\widehat{\sigma}_{jj}^{1/2}$, where $\widehat{\sigma}_{ij} = T^{-1}\sum_{t=1}^{T}\widehat{\epsilon}_{it}\widehat{\epsilon}_{jt}$.

We can see that the statistic $\widehat{J}_{wald}$ does not require an estimation of an invertible error covariance matrix. It is much faster to implement, and is valid even if N is much larger than T. This test can also be viewed as a robust version of a standardized Wald test, where the off-diagonal elements of the error covariance matrix become relatively less important as $N \to \infty$. Moreover, this statistic employs the degrees of freedom adjustment for the standardization of t_i^2, which is shown to provide a more accurate normal approximation, especially when N is much larger than T.

It is also worth mentioning that J_{wald} is a way to achieve good statistical size, but it may suffer from low power and is not robust to time-varyingness. Therefore, in the following subsections, we also consider the power enhancement test proposed by Fan et al. (2015), high-dimensional test for time-varying risk premium based on Gagliardini et al. (2016) and the high-dimensional test for long-run alphas based on Ma et al. (2020).

2.2. Power enhancement method in high-dimensional test

The test statistic based on Fan et al. (2015) combines a power enhancement component with an asymptotically pivotal statistic (say, the Wald test statistic), and strengthens the power under sparse alternatives. The construction of the power enhancement component is based on a screening technique. The null distribution of the new statistic is completely determined by that of the pivotal statistic. The power enhancement component takes the value of zero under the null hypothesis with high probability, but diverges quickly under sparse alternatives. With this property, it can greatly increase the power of the following test without distorting the size.

To test a high-dimensional vector $H_0 : \alpha^{\top} = (\alpha_1, \cdots, \alpha_N) = 0$ against a sparse alternative where the null hypothesis is violated by only a few components, most extant tests are based on a quadratic form $W = \widehat{\alpha}^\top D_v^{-1} \widehat{\alpha}$, where $\widehat{\alpha}$ is a consistent estimator of α, and D_v^{-1} is often the inverse of the asymptotic covariance matrix of $\widehat{\alpha}$. However, in the case of sparse alternative, the statistic W suffers from low power due to the accumulation of errors in estimating high-dimensional parameters.

To address this issue, Fan et al. (2015) add a power enhancement component J_0 to a traditional test statistic J_{wald} given by (2). Then the newly constructed test statistic is

$$J^* = J_0 + J_{wald},$$

where the power enhancement component J_0 is asymptotically zero under the null, but diverges and dominants J_{wald} under sparse alternatives. To compute the power enhancement component J_0, the screening set in Fan et al. (2015) is defined as $\widehat{S} = \{j : |\widehat{\alpha}_j| > \widehat{\sigma}_j \delta_{N,T},\ j = 1, \cdots, N\}$, in which $\delta_{N,T} = \log(\log T)\sqrt{\log N}$ according to the Proposition 4.1 in Fan et al. (2015), $\widehat{\sigma}_j$ is estimated based on the residuals of OLS estimator. The corresponding power enhancement component is given by

$$J_0 = \sqrt{N} \sum_{j \in \widehat{S}} \widehat{\alpha}_j^2 \widehat{\sigma}_j^{-2}. \tag{4}$$

Combining equation (2) and (4), the new statistic $J^* = J_0 + J_{wald}$ is used to test the null hypothesis $H_0 : \alpha^\top = (\alpha_1, \cdots, \alpha_N) = 0$.

2.3. High-dimensional test for time-varying risk premium

Gagliardini et al. (2016) developed another econometric methodology to infer the path of risk premia from a large unbalanced panel of asset returns and derived the no-arbitrage pricing restrictions including a test for $H_0 : \alpha^\top = (\alpha_1, \cdots, \alpha_N) = 0$. They estimate the time-varying alphas, betas and risk premia implied by conditional linear factor models under cross-sectional dependence. Different with model (1), their individual fund return regression is in the form of

$$r_{it} = \alpha_{it} + \beta_{it}^{\top} f_t + \epsilon_{it}, \quad i = 1, 2, \cdots, N, \ t = 1, 2, \cdots, T, \tag{5}$$

in which α_{it} and β_{it} denote the time-varying alpha and the $m \times 1$ vectors of time-varying factor loadings, respectively. Define Ω_t be the information flow available to investors. Under some mild assumptions,[10] Gagliardini et al. (2016) proved that there exists an unique random vector $v_t \in R^K$ such that v_t is Ω_{t-1} measurable and $\alpha_{it} = \beta_{it}^{\top} v_t$ for almost all i. Then the asset pricing restriction becomes

$$E[r_{it} | \Omega_{t-1}] = \beta_{it}^{\top} \lambda_t, \tag{6}$$

where $\lambda_t = v_t + E[f_t | \Omega_{t-1}]$ is the vector of the conditional risk premia.

To have a workable version of equations (5) and (6), Gagliardini et al. (2016) specify the conditioning information and how the model coefficients depend on it via a functional specification. To be specific, the conditioning information includes instruments $Z_t \in R^p$ and $Z_{it} \in R^q$, in which the instruments Z_t are common to all funds such as macroeconomic variables; the instruments Z_{it} are specific to fund i such as fund specific characteristics. The factor loadings are such that $\beta_{it} = B_i Z_{t-1} + C_i Z_{i,t-1}$, where $B_i \in R^{m \times p}$ and $C_i \in R^{m \times q}$, for $i = 1, 2, \cdots, N, \ t = 1, 2, \cdots, T$. Furthermore, the risk premia vector is such that $\lambda_t = \Lambda Z_{t-1}$, where $\Lambda \in R^{m \times p}$ and $E[f_t | \Omega_{t-1}] = D Z_{t-1}$, where $D \in R^{m \times p}$, for any t. Under the above setting, model (5) can be rewritten as

$$r_{it} = b_i^{\top} x_{i,t} + \epsilon_{it}, \qquad \alpha_{it} = x_{1,i,t}^{\top} b_{1,i}, \tag{7}$$

[10]For example, the assumptions include an approximate factor structure for the conditional distribution of the error terms given Ω_{t-1} and no asymptotic arbitrage opportunities in the economy.

where $x_{i,t} = (x_{1,i,t}^\top, x_{2,i,t}^\top)^\top$ are transformed new regressors based on factors f_t, instruments Z_t and Z_{it}; $b_i = (b_{1,i}^\top, b_{2,i}^\top)^\top$ is a vector of unknown parameters which is a function of Λ, D, B_i and C_i. Please refer to Appendix B for definitions of $x_{i,t}$ and b_i. From equation (7), we can see that the key to addressing the time-varyingness in this approach is to formulate α_{it} as a function of some observable variables $x_{1,i,t}$ and unknown parameters $b_{1,i}$, where $x_{1,i,t}$ may vary across time and $b_{1,i}$ is time-invariant and can be estimated well. Hence, the time-varying components of α_{it} is fully captured by observable variables. The unobservable components of α_{it}, which is $b_{1,i}$, is time-invariant. Therefore, we can simply estimate α_{it} by

$$\widehat{\alpha}_{it} = x_{1,i,t}^\top \widehat{b}_{1,i}. \tag{8}$$

2.3.1. Asymptotic properties of time-varying alphas

The technical details of the estimators of the high-dimensional test with time-varying alphas are presented in Appendix B. Here, we derive the asymptotic properties of estimated time-varying alphas $\widehat{\alpha}_{it}$, which is lacking in Gagliardini et al. (2016).

By assuming the stationarity of regressors, the OLS estimator $\widehat{b}_i$ is then asymptotically normal with $\sqrt{T}\left(\widehat{b}_i - b_i\right) \to_d N\left(0, \tau_i Q_{x,i}^{-1} S_{ii} Q_{x,i}^{-1}\right)$, where $\tau_i = E(I_{i,t})$, $I_{i,t}$ is the observability indicator, $Q_{x,i} = E(I_{i,t} x_{i,t} x_{i,t}^\top)$ and $S_{ii} = E(I_{i,t} \epsilon_{it}^2 x_{i,t} x_{i,t}^\top)$. In order to construct confidence intervals for α_{it}, we deduce that

$$\frac{\widehat{\alpha}_{it} - \alpha_{it}}{B_T} \to_d N(0,1), \tag{9}$$

where $B_T^2 = \frac{1}{T_i} x_{1,i,t}^\top E_1^\top \widehat{Q}_{x,i}^{-1} \widehat{S}_{ii} \widehat{Q}_{x,i}^{-1} E_1 x_{1,i,t}$, $\widehat{Q}_{x,i}^{-1} = \frac{1}{T_i} \sum_t I_{i,t} x_{i,t} x_{i,t}^\top$, $\widehat{S}_{ii} = \frac{1}{T_i} \sum_t I_{i,t} \widehat{\epsilon}_{it}^2 x_{i,t} x_{i,t}^\top$, $\widehat{\epsilon}_{it} = r_{it} - \widehat{b}_i^\top x_{it}$, T_i is the number of observations for fund i and E_1 is the selection matrix.

2.3.2. High-dimensional test for time-varying alphas

Here, we consider a test for $H_0 : \alpha_{it} = 0\ \forall i,\ \forall t$, which is equivalent to $H_0 : b_{1,i} = 0\ \forall i$ against the alternative hypothesis $H_1 : E[b_{1,i}^\top b_{1,i}] > 0$. To see that, let $\alpha_i = (\alpha_{i1}, ..., \alpha_{iT})^\top$ and $x_{1,i} = (x_{1,i,1}, ..., x_{1,i,T})^\top$, we have $\alpha_i = x_{1,i} b_{1,i}$ and $(x_{1,i}^\top x_{1,i})^{-1} x_{1,i}^\top \alpha_i = b_{1,i}$, which follows that $\alpha_i = 0$ if and only if $b_{1,i} = 0$. Hence, $H_0 : \alpha_{it} = 0\ \forall i,\ \forall t$ is equivalent to $H_0 : b_{1,i} = 0\ \forall i$. This fact shows that under the framework of Gagliardini et al. (2016) one can transform a complicated joint test for an unbalanced panel data over all time periods to a relatively simpler test for time–invariant parameters.[11] Let $\widehat{\xi}_{wald}$ denote the traditional Wald statistic. It can be shown that after some normalization $\widehat{\xi}_{wald}$ converges to a standard normal distribution. For clearer presentation, we move the technical details to Appendix B.3. Similar to Section 2.2, one can combine the original statistic $\widehat{\xi}_{wald}$ with a power enhancement term.[12] This is done by adding a screening statistic denoted by $\widehat{\xi}_0$ and then our constructed power enhancement test statistic is given by $\widehat{\xi} = \widehat{\xi}_{wald} + \widehat{\xi}_0$. The procedure of constructing $\widehat{\xi}_0$ and a skeleton of its theoretical justification are given as follows.

Define a screen set $\widehat{S} = \{i : |\widehat{b}_{1,i}| > \widehat{v}_i^{1/2} \delta_{NT}, i = 1, ..., N\}$ and the screening statistic $\widehat{\xi}_0 = \sqrt{N} \sum_{j \in \widehat{S}} \widehat{b}_{1,i}^2 \widehat{v}_i^{-1}$, where $\widehat{v}_i$ is the estimated asymptotic variance of $\widehat{b}_{1,i}$. For strictly stationary regressors that satisfy strong mixing conditions, following from the large deviation theory (typically using exponential inequality to prove uniform convergence results), we have $\max_i |\widehat{b}_{1,i} - b_{1,i}| / \widehat{v}_i^{1/2} = O_P(\log N)$. Choose $\delta_{NT} = \log(\log T)\sqrt{\log N}$, and then under the null hypothesis $H_0 : b_{1,i} = 0$ for $\forall i$, $\Pr(\widehat{S} = \emptyset | H_0) \to 1$. Therefore, the asymptotic null distribution of $\widehat{\xi}$ is determined by that of $\widehat{\xi}_{wald}$ and the size distortion due to adding $\widehat{\xi}_0$ is negligible. In addition, as shown in Theorem 3.1 of Fan et al. (2015), $\Pr(\widehat{\xi}_0 > \sqrt{N} | H_1) \to 1$ and thus the power of $\widehat{\xi}_{wald}$ is enhanced after adding $\widehat{\xi}_0$.

[11] We thank an anonymous referee for providing this insight.

[12] We thank an anonymous referee for providing this idea.

2.4. High-dimensional test for long-run alphas

In the setting of Subsection 2.3, all the the time-varying components of α_{it} are driven by some observed variables $x_{1,i,t}$, a transformation of Z_t and Z_{it}. This indicated that it requires the prior knowledge of Z_t and Z_{it}, which may be misspecified or some components may be omitted. As an alternative method, Ang and Kristensen (2012) and Ma et al. (2020) consider a nonparametric time-varying factor model, in which $\alpha_{it} = \alpha_i(t/T)$ and $\beta_{it} = \beta_i(t/T)$ are unknown smooth functions of re-scaled time t/T. For ease of exposition, we call this approach as 'smooth approach'. The advantage of such modeling strategy is that it doesn't require researchers to know the true functional forms of the time-varying alphas. In addition, Ang and Kristensen (2012) and Ma et al. (2020) aim to test the null hypothesis of whether the long-run alphas (i.e, averages of time-varying alphas) are equal to zero under the fixed N and the large N frameworks, respectively. Since the number of funds are much larger than the length of time period, here we consider the high-dimensional alpha test proposed in Ma et al. (2020).

Define the long-run alpha as $\delta_i = \frac{1}{T}\sum_{t=1}^{T}\alpha_{it}$ for $i = 1, ..., N$, and then we can rewritten model (5) as

$$r_{it} = \delta_i + \delta_i(t/T) + \beta_i^\top(t/T) f_t + \epsilon_{it},$$

where $\delta_i = \alpha_i(t/T) - T^{-1}\sum_{t=1}^{T}\alpha_i(t/T)$. Hence, the null and alternative hypothesis for testing the long-run alphas across the N assets are formulated as $H_0 : \delta_i = 0$ for all $i = 1, ..., N$ versus $H_1 : \delta_i = 0$ for some $i = 1, ..., N$. Here, we outline the testing procedures for high-dimensional long-run alphas. For the details of nonparametric estimation, see Ma et al. (2020). Under the null hypothesis, the resulting residuals are $\widehat{\epsilon}_{it} = r_{it} - \widehat{\delta}_i(t/T) - \widehat{\beta}_i^\top(t/T) f_t$, where $\widehat{\delta}_i(t/T)$ and $\widehat{\beta}_i(t/T)$ are nonparametric estimates of $\delta_i(t/T)$ and $\beta_i(t/T)$, respectively. Define the following statistic:

$$J_{NT} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\widehat{\epsilon}_{it}\right)^2. \tag{10}$$

Under H_0, Ma et al. (2020) show that $\widehat{Z}_{NT} = \widehat{\sigma}_{NT}^{-1}(J_{NT} - \widehat{\mu}_{NT})$ converges to a standard normal variable, where $\widehat{\mu}_{NT}$ and $\widehat{\sigma}_{NT}$ are consistent estimates of the mean and standard deviation of J_{NT}, respectively.

Comparing the two settings of time-varying alphas presented in subsections 2.3 and 2.4, we can see that in subsection 2.3, $x_{1,i,t}$ may be weakly dependent over time, so allowing the formulation $\alpha_{it} = g(x_{1,i,t}, b_i)$ to have a large degree of varyingness over time, and potentially, structural breaks. In contrast, the formulation $\alpha_{it} = \alpha_i(t/T)$ requires that α_{it} change smoothly over time, ruling out structural breaks. So we expect that during periods where there are financial crises or market instability, the method outlined in the previous subsection would produce a much better estimation than the smoothing approach, because α_{it} may change dramatically. We will demonstrate this point in our empirical study below. Of course, as we mentioned earlier, there are also some drawbacks of the approach in subsection 2.3 that it requires prior knowledge of $x_{1,i,t}$ which may be misspecified.

3. Data and benchmark factor models for mutual funds and hedge funds

In this section, we first discuss our data sets about the fund returns net of all management expenses and 12b-fees for U.S. mutual funds and hedge funds, respectively.[13] We introduce our benchmark factor models for these two funds afterward.

3.1. Data and descriptive statistics

Our cross-section sample of mutual funds is similar to Ferson and Chen (2020) and Harvey and Liu (2018). We obtain active U.S. equity mutual funds data from the Center for Research

[13]Although our methodology is flexible enough for the gross alphas in Pástor et al. (2015), we follow the main literature (e.g., Barras et al. (2010); Ferson and Chen (2020)) and use the net alphas to illustrate our idea. Different from Pástor et al. (2015), our alphas are time-varying rather than time-invariant.

in Security Prices (CRSP) Survivor-Bias-Free Mutual Fund data base for the Feb1962-Sept2017 period. The sample period of ours is the same as the ones from Harvey and Liu (2018) and Ferson and Chen (2020) for comparison reasons. We exclude the index funds. To mitigate omission bias (Elton et al. (2001)) and incubation and back-fill bias (Evans (2010)), we exclude observations prior to the reported year when the mutual funds first entered into the data base, and the funds which do not report a year of organization. We only include the funds which have initial total net assets (TNA) above $10 million and more than 80% of their holdings in equity markets. To avoid the look-ahead bias, we do not exclude funds whose TNA subsequently falls below $10 million. These screens leave us with a sample of 7123 (5862) mutual funds with at least 8 (30) months of returns data for the Feb1962-Sept2017 period.[14]

Our cross-section sample of hedge funds is similar to Ferson and Chen (2020). To be specific, we obtain U.S. equity-oriented hedge funds data from Lipper TASS for the Jan1994-Apr2017 period. The sample period of ours is the same as the ones from Ferson and Chen (2020) for comparison reasons. To mitigate back-fill bias, we remove the first 24 months of returns and returns before the dates when funds first entered the data base, and funds with missing values in the field for the add date (Ferson and Chen (2020)). We only include those categorized for a given month as either Dedicated Short bias, Event-driven, Equity market neutral, Fund-of-Funds or Long/short equity hedge. Similar to the mutual fund sample, we require that a fund has initial total net assets (TNA) above $10 million as of the first date. These screens leave us with a sample of 3463 (2371) hedge funds with at least 8 (30) months of returns data for the Jan1994-Apr2017 period.

Table 1 presents summary statistics of the mutual fund and hedge fund data in our study.

[14]Similarly, Harvey and Liu (2018) and Ferson and Chen (2020) have obtained a sample of 3619 and 3716 mutual funds with at least 8 months of returns over the same period, respectively. We use 30 months as our threshold as it adds robustness to our results (Hunter et al., 2014; Harvey and Liu, 2020)

We find that they share similar characteristics with the data sample used in Ferson and Chen (2020). The main characteristics are listed as follows.

- The range of average returns across funds is much greater in the hedge fund sample ($-0.120 \sim 0.165$) than that in the mutual fund sample ($-0.069 \sim 0.119$).
- The median of estimated alpha from the Fung-Hsieh seven-factor (Fung and Hsieh (1997, 2001) model for the hedge funds is positive, while the one from the Fama-French-Carhart four-factor (Carhart (1997)) for the mutual funds it is slightly negative. The tails of the cross-sectional alpha distributions extend to larger values for the hedge funds. For example, the upper 5% tail value for the alphas in the hedge fund sample is 1.0% per month, while for the mutual funds it is only 0.3%. In the left tail these two types of funds also present different alpha distributions, with a thicker lower tail for the alphas in the hedge fund sample.
- The sample volatility of the median hedge fund return (2.8% per month) is smaller than for the median mutual fund (3.7%). The range of volatilities across the hedge funds is greater, with more mass in the lower tail. Between the 10% and 90% quantiles, the volatility range is 1.2% - 7.2% (1.2% - 6.7% in Ferson and Chen (2020)) for hedge funds, and 1.1% - 6.2% (4.2% - 7.0% in Ferson and Chen (2020)) for mutual funds.
- The return autocorrelation is slightly higher for the hedge funds than mutual funds. The median autocorrelation for the hedge (mutual) funds is 0.127 (0.074), although the range of return autocorrelation is larger for mutual funds (from -0.843 to 0.978) than for hedge funds (from -0.794 to 0.814).

3.2. Benchmark factor models for mutual funds and hedge funds

Before using the high-dimensional cross-sectional tests to evaluate fund performance, we have to specify the benchmark factor models to estimate fund alpha. For mutual funds, we consider the Fama-French-Carhart four-factor model including the Market excess return (MKT) factor, the Small-Minus-Big (SMB) size factor, the High-Minus-Low (HML) value factor and the Momentum (MOM) factor. If anything, this choice of Fama-French-Carhart four-factor model as our benchmark factor model can only bias our results against finding positive and significant fund alphas, and we try alternative benchmarks in the robustness section.

Our benchmark factor model for hedge funds is the analogous Fung-Hsieh seven-factor model (c.f., Fung and Hsieh (1997, 2001)), instead of the Fama-French-Carhart four-factor model. These seven-factors (i.e., Bond Trend-Following Factor, Currency Trend-Following Factor, Commodity Trend-Following Factor, Equity Market Factor, Size Spread Factor constructed from Russell 2000 index and S&P500, Bond Market Factor and Credit Spread Factor) based on Fung and Hsieh (1997, 2001) are arguably more suitable for the hedge funds than the Fama-French-Carhart four-factors. For robustness, we have also tried the Fama-French-Carhart four-factor model for the hedge funds, and the results are available upon request.

4. Mutual funds and hedge funds performance evaluation

In this section, we evaluate the fund performance for both mutual funds and hedge funds. To shed light into the debates between Kosowski et al. (2006) and Fama and French (2010), we first test the high-dimensional vector ($H_0 : \alpha = 0$) against sparse alternatives. After that, we take a further step to gauge the ratio of significant non-zero alphas for both mutual funds and hedge funds.

4.1. Testing the high-dimensional zero-alpha hypothesis

In this subsection, we test the high-dimensional vector ($H_0 : \alpha = 0$) against sparse alternatives.

4.1.1. High-dimensional test for alphas in linear factor pricing models

We evaluate the performance of active mutual funds in the U.S. using the method based on Pesaran and Yamagata (2018) and to control for CSD. We implement the test on a rolling window basis, as we can not only mitigate the impact of time-varying factor loadings and sampling biases, but also make sure that we have a panel with the cross-sectional dimension larger than the time-series dimension, which is 60 here. Our choice of window-width (i.e., 60) is supported by the simulation exercise we provided in Appendix A.

Figure 1 and Figure 2 report the results for mutual funds and hedge funds, respectively. In each figure, the left (right) top graph shows the J test statistics (and corresponding p-values) of the alphas for all the funds in each window; The left (right) middle graph shows J test statistics (and corresponding p-values) for the positive alphas in each window; The left (right) bottom graph shows J test statistics (and corresponding p-values) for the negative alphas in each window.

According to the top two graphs of Figure 1, from the early 1980s until 2017, the null hypothesis of zero alphas based on each estimation window is rejected at the conventional 5% significance level. The J statistics in the left top graph, peaked during both the 1997 Asian Financial Crisis and the GFC. The J statistics in the left bottom graph, based on the negative alphas in each estimation window, share a similar pattern, which indicates that the negative alphas in each estimation window dominate the performance of all the mutual funds. The p-values during the crisis periods are almost zero, which implies that the negative alphas of mutual funds during crises are statistically significant at any conventional level.

Interestingly, according to the J statistics in the left middle graph, the positive alphas of mutual funds stay above 5 during the GFC and once surges to 17, with a p-value of zero all over the crisis period, which suggests that some mutual funds have also reaped considerable profits from the GFC.

Figure 2 presents a very different picture for hedge funds. According to the three graphs in the left column of Figure 2, positive alphas dominate the performance of all the hedge funds in each estimation window. Among the three graphs in the left column of Figure 2, only the J statistics in the left top and the middle left graphs increase sharply during the GFC, and the corresponding p-values are almost zero, which implies that the good performance of hedge fund during crises is statistically significant at any conventional level. Put differently, the GFC is a good opportunity for hedge funds. We also reject the null hypothesis of zero alphas for almost all the estimation windows.

To sum up, the J test tends to reject the null hypothesis of zero alphas during periods of major financial disruptions for both mutual funds and hedge funds. It mainly implies the negative and statistically significant alphas for mutual funds, but positive and statistically significant alphas for hedge funds. However, some hedge funds and a smaller portion of mutual funds have contributed to the increased positive alphas during the GFC.

4.1.2. Power enhancement method in high-dimensional test

Alternatively, we apply the power enhancement method for high-dimensional cross-sectional tests to our mutual fund and hedge fund sample. We also implement it in a rolling fashion with a window-width of 60. As we have 668 months from Feb1962 to Sept2017 for our mutual fund sample, we have 609 (=668-60+1) tests in total. For each test (window), we first select funds without missing observations in the past five years, which yields a balanced panel. Then we evaluate the screening set $\widehat{S}$ using the preceding 60 months' data.

We also apply the same test to our hedge fund sample. As our data covers 280 months from Jan1994 to Apr2017, so we have 221 (=280-60+1) tests in total. The procedure is similar to that for mutual funds, which we omit here for brevity.

Table 2 reports the summary statistics calculated for the funds selected via the method based on Fan et al. (2015). Panel A and B present the results for mutual funds and hedge funds, respectively. During each 60-month window, the maximum number of mutual (hedge) fund is 3068 (518), while the minimum is 33 (0). On average, among the 617 mutual funds during each window, on average 34 mutual funds are selected by the threshold in Fan et al. (2015)), which yields a percentage of 5.5%. Among 309 hedge funds during each 60-month window, on average of 38 mutual funds are selected by the threshold in Fan et al. (2015), which yields a percentage of 12.3%. In terms of average alpha, not only hedge funds (0.4825%) outperform mutual funds (0.2444%), but also the selected hedge funds (0.8106%) outperform the selected mutual funds (0.3512%).

4.1.3. High-dimensional test for time-varying alphas

Now we evaluate the performance of both active mutual funds and hedge funds in the U.S. using the method based on Gagliardini et al. (2016).

For mutual funds, our baseline asset pricing model is the Fama-French-Carhart four-factor model below with $f_t = (MKT_t, SMB_t, HML_t, MOM_t)^\top$. We take the instruments $Z_t = (1, Z_t^{*\top})^\top$, where bivariate vector Z_t^* includes the term spread, proxied by the difference between yields on 10-year Treasury and 3-month T-bill, and the default spread, proxied by the yield difference between Moody's Baa-rated and Aaa-rated corporate bonds. We take a scalar $Z_{i,t}$ corresponding to the fund return of fund i. We refer to Avramov and Chordia (2006) for convincing theoretical and empirical arguments in favor of the chosen conditional specification.

For hedge funds, our baseline asset pricing model is the Fung-Hsieh seven-factor model, instead of the Fama-French-Carhart four-factor model. The setting of instruments for hedge funds is the same as the case of mutual funds. Since this study is mainly about fund performance evaluation, in the following main text, we focus on the results of fund alphas.

Panel A of Table 3 gathers the results for testing the asset pricing restrictions (i.e., no-arbitrage condition) within mutual funds using factor models with time-invariant coefficients. For mutual funds, the test statistics reject both null hypotheses $H_0 : \alpha_i = \beta_i^\top \nu$ and $\alpha_i = 0$ for the three specifications at any conventional significant level (e.g., 1%, 5%, and 10%). Likewise, Panel B of Table 3 gathers the results for testing the asset pricing restrictions within mutual funds in time-varying specifications. Similarly, for mutual funds, the test statistics reject both null hypotheses $H_0 : \alpha_i = \beta_i^\top \nu$ and $\alpha_i = 0$ for the three specifications at any conventional significant level. Altogether, both our time-invariant and time-varying models suggest that mutual fund alphas are not always zero, and there are some arbitrage opportunities among mutual fund alphas for fund of mutual funds.

Panel A of Table 4 gathers the results for testing the asset pricing restrictions (i.e., no-arbitrage condition) within hedge funds using factor models with time-invariant coefficients. For hedge funds, the test statistic rejects the null hypothesis $H_0 : \alpha_i = 0$ at any conventional significant level while it cannot reject the null hypothesis $H_0 : \alpha_i = \beta_i^\top \nu$. Panel B of Table 4 gathers the results for testing the asset pricing restrictions within hedge funds using factor models with time-varying coefficients. Similarly, for hedge funds, the test statistic rejects the null hypothesis $H_0 : \alpha_i = 0$ at any conventional significant level while it cannot reject the null hypothesis $H_0 : \alpha_i = \beta_i^\top \nu$. Altogether, both our time-invariant and time-varying models suggest that hedge fund alphas are not always zero, but there are few arbitrage opportunities among hedge fund alphas for fund of hedge funds.

4.1.4. High-dimensional test for long-run alphas

Now, we focus on testing long-run alphas for both mutual funds and hedge funds in the U.S. using the method based on Ma et al. (2020). We also implement it in a rolling fashion with a window-width of 60 to evaluate funds' overall performance during a specific time period. For each window, we first select funds without missing observations in the past five years, which yields a balanced panel.

Figure 3 plots the p-values of long-run alphas across the 609 (221) tests for mutual (hedge) funds in each 60-month window. From this figure, it can be seen that the p-values fluctuate dramatically over time. This strong evidence indicates that even the average of fund alphas across time is time-varying, which implies that time-varying factor model is more suitable than the time-invariant factor model for evaluating fund performance. For both mutual funds and hedge funds, only some of the p-values are lower than the 5% significance level, which implies that the average of fund alphas is not always zero.

4.2. Ratio of non-zero alphas

From the previous subsection, we conclude that the null hypothesis of all alphas being zero should be rejected at 5% level of significance with all the four high dimensional test methods. The next natural question would be: what is the average ratio of significantly non-zero alphas? To answer this question, we conduct the following exercises based on time-invariant factor model (1) and time-varying factor model (5), respectively. Moreover, to deal with the multiple testing problem, we use the threshold of t-statistic 3 to establish the significance. We also justify the choice of this threshold from a simulation-based experiment which is given in Appendix A.

4.2.1. Results based on time-invariant factor model (1)

Similar to Section 4.1, we compute the significantly non-zero alphas in a rolling fashion with a window-width of 60. For ease of exposition, we use ‘screening set’ here to denote the set of significant alphas, and we use ‘significant’ and ‘selected’ interchangably in this section. Figure 4 and Figure 5 present the results using a threshold of t-statistic of 3 for our sample of mutual funds and hedge funds, respectively. The left (right) top graph shows the average value (number) of selected positive alphas and negative alphas in the screening set. The left middle graph shows the ratio of selected positive and negative alphas in the screening set while the right middle graph shows the ratio of selected positive and negative alphas in the whole panel size during the test period. The left (right) bottom graph shows the alpha (t-statistic of alpha) of each fund during each 60-month window.

According to the top two graphs in Figure 4, the average value and number of the selected negative alphas slightly exceed the ones of the selected positive alphas using our sample of mutual funds. This is clearer when we look at the middle two graphs, which show that there are more negative alphas than positive alphas for mutual funds except for our early sample. Interestingly, it seems that there is an increase (decrease) in the ratio of selected negative (positive) alphas and the ratio of selected negative (positive) alphas over the panel size during both the 1997 Asian Financial Crisis and the late 2000s Global Financial Crisis (GFC), which suggests that the mutual funds suffer from crises. This message is clearer when we plot the alpha and especially the t-statistic of alpha of each fund during each 60-month window in the bottom two graphs. The deteriorated performance of mutual funds during crises is more due to the sharp increase of negative t-statistic of the alphas rather than the decrease of positive t-statistic of the alphas, which is clearly shown in the bottom right graph.

Next, we shift our attention to the performance of the hedge funds, which is in stark contrast to the performance of the mutual funds. For instance, the left top graph in Figure 5 suggests that there are few negative alphas selected for hedge funds while few positive alphas selected for mutual funds during our early sample, which suggests that hedge funds perform better than mutual funds. Moreover, the top right graph shows that there is an increase in the number of selected positive alphas while the number of selected negative alphas stays stable during the GFC, which has been further supported by the right middle graph in Figure 5. In the bottom two graphs, there are more selected positive alphas than selected negative alphas over our sample, and the t-statistic of the selected positive alphas for hedge funds peaked during the GFC. Together, these results provide fresh evidence supporting the extant literature conjecturing the hedge funds as arbitrageurs.

Based on the threshold of t=3 and model (1), we find that on average there are 9.7% (9.06% positive and 0.64% negative) significant hedge fund alphas, and 6.8% (2.91% positive and 3.89% negative) significant mutual fund alphas.

4.2.2. Results based on time-varying factor model (5)

In this subsection, we compute the significantly non-zero alphas at each time t based on the time-varying factor model (5) in Gagliardini et al. (2016) using a threshold of t-statistic of 3. For computation details of $\widehat{\alpha}_{it}$ and its standard deviations, please refer to subsection 2.3. Figures 6 and 7 present the results for our sample of mutual funds and hedge funds, respectively. The left (right) top graph shows the average value (number) of positive alphas and negative alphas in the screening set. The left middle graph shows the ratio of positive and negative alphas in the screening set while the right middle graph shows the ratio of selected positive and negative alphas in the whole panel size. The left (right) bottom graph shows the alpha (t-statistic of alpha) of each fund. Perhaps the most striking difference between Figure 4 and Figure 6 is the ratio of the right middle graph, which is due to a substantially increased portion of extreme alphas (especially during crises) revealed by Gagliardini et al. (2016) but not captured by the rolling windows of time-invariant factor model (1). Overall, it demonstrates the advantage of the time-varying property of Gagliardini et al. (2016).

According to Figure 6, during the late 2000s Global Financial Crisis (GFC), there is an increase in the ratio (number) of selected positive alphas for mutual funds, as the performance of mutual funds diverges. For the selected funds with negative alphas, in spite of the further plummeting alpha values, actually there is also a decrease in the ratio (number) of selected negative alphas for them, which further suggests that some mutual funds which performed poorly during non-crisis periods improved their position during the GFC. Except for the GFC and early Dot-com crisis periods, the number (ratio) of negative alphas are comparable with or even larger than the ones of positive alphas for mutual funds during non-crisis periods.

However, the performance of hedge funds differs from that of mutual funds. For instance, there are few negative alphas selected for hedge funds and the number (ratio) of selected positive alphas dominates the number (ratio) of selected negative alphas. This result suggests that hedge funds perform better than mutual funds and is consistent with our previous results from the rolling window analyses of model (1). Also, both the average selected positive alpha value and the number/ratio of selected positive alphas in the whole panel size increases during the GFC, which re-affirms our previous results.[15]

Based on the threshold of t=3 and time-varying high-dimensional method based on Gagliardini et al. (2016), we find that on average there are 10.32% (8.97% positive and 1.35% negative) significant hedge fund alphas, and 5.33% (3.34% positive and 1.99% negative) significant mutual fund alphas, which strongly corroborate with our previous results of 9.7% (6.8%) significant hedge (mutual) fund alphas using the time-invariant coefficient model with the same threshold.

To ensure the robustness of our results, we further consider the 'smoothing approach' adopted in Ma et al. (2020), which imposes no parametric assumptions on the time variation of alphas and betas. The results are respectively presented in Figures 8 and 9, from which we find similar results as those using the rolling windows of time-invariant factor model (1). More importantly, the 'smoothing approach' cannot capture the presence of a substantially increased portion of extreme alphas during financial crises but revealed by our model based on Gagliardini et al. (2016). This is consistent with our expectation because time-varying α_{it} in the setting of Gagliardini et al. (2016) can capture potential structural breaks which are ruled out by the 'smoothing approach' as it requires α_{it} change smoothly over time.

[15]The decrease in the ratio of selected positive alphas during the GFC in the left middle graph is a mechanic result of the sharp increase in the ratio of selected negative alphas at the same time.

5. Robustness

We focus on the net alphas only in the main text. While net alphas measure the abnormal return earned by fund investors, gross fund alphas measure the return the fund earns, and the value-added measure from Berk and van Binsbergen (2015) evaluates the money/value that the fund extracts from capital markets.[16] In this section we perform robustness checks to test whether our results are sensitive to the length of estimation window, altering the number of factors in our benchmark factor model, adding back fees and expenses to fund returns, using the value-added measure from Berk and van Binsbergen (2015) instead of fund alphas, using the Vanguard index fund as the passive benchmark portfolio alternative to the traditional Fama-French factors, as well as sub-sample analyses. We report these results in online appendices for brevity. There are other benchmarks, fund skill indicators, and data sources in the literature as well. Since the main contribution of this paper is methodological, we leave them as a future research direction.

5.1. Altering the length of estimation window

When we use the method based on Fan et al. (2015), we have focused on the estimation window of 60 months in the main analysis, simply because it is the most popular length of estimation window in the literature on fund performance evaluation. We do, however, fully acknowledge that there may be a certain degree of arbitrariness in choosing such a length of estimation window. To allay the concerns that our conclusion may hinge on the choice of length of the estimation window, we use an alternative length of estimation of 36 months. The results are available in Section 1 of our online appendix, which are almost identical.

[16]All of them have all been used as indicators of fund skill, depending on whether the researchers take the perspective of the fund investors, fund managers, etc. By no means, we had the intention to be involved in the re-heated debate on which measure is the right/better measure of fund skills.

5.2. Altering the number of factors in the benchmark factor model

The mutual fund performance results from our main section are mainly the Fama-French-Carhart 4-factor model, which is the most popular benchmark in the literature on mutual fund performance evaluation. However, even this benchmark factor model may not suit all funds. For instance, there are almost no passive exposures to value and/or small stocks in the early sub-sample such as 1960s, and the momentum factor, which involves heavy transaction costs, may not suit all fund managers. Hence, the hurdle for the net return of an active fund to beat such a hypothetical benchmark may be arguably high for some fund managers. To check whether our results are sensitive to the choice of benchmark factor model, we substitute our benchmark factor model of Fama-French-Carhart 4-factor model with CAPM and Fama-French 3-factor model. The results available in Section 2 of our online appendix suggest that our main results stay qualitatively unchanged.

5.3. Sub-sample analysis

Do the alphas of mutual funds and hedge funds vary over our sample period? To answer this question, we follow Cai et al. (2018) and implement a subsample analysis for the pre-2001 and post-2001 sub-periods since Harvey and Liu (2018) find '*noticeable differences between the parameter estimates for 1984–2001 period and for full sample period*'. Our results remain qualitatively similar and are available in Section 3 of our online appendix.

5.4. Adding back fees and expenses to fund returns

Since a recent strand of literature (e.g., Fama and French (2010)) has proposed the gross alphas in place of net alphas as the indicator of fund skills, we use the procedures of Fama and French (2010) and add back fees and expenses to net fund returns and re-run our tests. To be specific, we compute our first gross measure of returns (i.e., $grossreturn1$) as the net returns plus 1/12th of the fund's annual expense ratio. When a fund's annual expense ratio is missing, we follow Fama and French (2010) and assume it is the same as the average of other active mutual funds with similar Assets Under Management (AUM). Put differently, our first measure of gross fund returns (i.e., $grossreturn1$) include the costs in expense ratios but exclude other costs such as trading costs, in light of the highlighted measurement issues in trading costs in the appendix A of Fama and French (2010). We find qualitatively similar results and present them in Section 4 of the online appendix.

5.5. Using the value-added measure instead of fund alphas

In the literature, there are non-alpha fund skill indicators that we cannot ignore. The strongest competitor so far to net alphas and gross alphas is perhaps the value-added measure based on Berk and van Binsbergen (2015), although on page 4 immediately after their equation 1, they admit '*The most commonly used measure of skill in the literature is the unconditional mean of ε_{it}, or the net alpha, ...*'. To be specific, Berk and van Binsbergen (2015) have proposed using the value-added measure, which is the product of AUM and gross alphas, to measure fund performance. They argue that it is a better measure than both net alphas and gross alphas. Based on their suggestion, we construct our first value-added measure (i.e., $valueadded1$) as the product of the gross alphas (estimated from the Fama-French-Carhart 4-factor model, Fama and French 3-factor model and CAPM, respectively) and the natural logarithm of Total Net Asset Value (i.e., the variable '*Mtna*' in CRSP database) for each mutual fund in our sample[17]. We find qualitatively similar results and present them in Section 5 of the online appendix.

[17] Strictly speaking, we should have used lagged AUM to make our value-added measure identical to Berk and van Binsbergen (2015). However, due to the persistence in AUM, the sample correlation coefficient between our value-added measure and the measure constructed from lagged AUM is above 0.98 (no matter whether we lag one month or one year) and hence we ignore this very subtle difference in this section.

5.6. *Using Vanguard index fund as the benchmark*

While the standard practice is to adjust for risk with the traditional factor models (i.e., the Fama-French-Carhart 4-factor model, Fama and French 3-factor model or the CAPM), Berk and van Binsbergen (2015) propose constructing an alternative passive investment opportunity itself for a couple of reasons. On the one hand, it remains debatable whether these factor models can accurately adjust for risk and which factor model is the best. On the other hand, sometimes these (e.g., Fama-French) factors may either be unknown/unavailable to fund investors (in the early periods), or involve too much transaction costs (e.g., the momentum factor).

We use the net return of the Vanguard S&P 500 index fund as the alternative benchmark to the traditional factor models to explore this concern. We only use the Vanguard S&P 500 index fund instead of all 11 Vanguard index funds in Berk and van Binsbergen (2015), as we follow the mainstream literature (e.g., Harvey and Liu (2018); Ferson and Chen (2020)) and focus on CRSP funds that hold mainly U.S. equity. We construct our second set of measures for net alpha ($netreturn2$), gross alpha ($grossreturn2$) and value-added measure ($valueadded2$) via CAPM (using the net return of Vanguard S&P 500 index fund as the new MKT factor). Overall, we find qualitatively similar results and present them in Section 6 of our online appendix.

5.7. *Improved mutual fund performance in the post-crisis era*

One puzzling finding in our previous analysis is the improved performance of mutual funds relative to hedge funds in the post-crisis era. While it requires further investigation in empirical and theoretical research to explain this phenomenon, we simply would like to check the existence of it via traditional methods. To this end, we follow the approach of Akbas et al. (2015) to use the flows of both mutual funds and hedge funds to explain the mispricing metric in Stambaugh et al. (2012, 2015). Due to data availability of mispricing metric in Stambaugh et al. (2012, 2015), we can only extend the sample period of 1994-2012 in Akbas et al. (2015)

to the period of 1994-2016. We find that mutual fund flows no longer exacerbate mispricings once we add the 4 more recent years, while the estimated coefficients for hedge funds stay qualitatively the same. The results are presented in Section 7 of our online appendix.

5.8. Time-varying risk premia of funds

We use the high-dimensional test for time-varying alphas to investigate the time-varying risk premia of U.S. active equity-oriented mutual funds and hedge funds and present the results in Section 8 of our online appendix. In consistency with the equity risk premia results in Gagliardini et al. (2016), we find risk premia of both mutual funds and hedge funds indeed fluctuate over time. Furthermore, we cannot reject the hypothesis that the factors conditional expectations are time-invariant, but we reject the hypothesis that the risk premia due to the dynamics induced by the cross-sectional parameter ν for both mutual funds and hedge funds are time-invariant.

6. Concluding remarks

To shed light into both the zero-alpha hypothesis and the ratio of non-zero alphas, we recognize the panel nature of fund industries and tailor high-dimensional cross-sectional tests (Pesaran and Yamagata, 2018; Fan et al., 2015; Gagliardini et al., 2016; Ma et al., 2020). We are particularly keen on the high-dimensional test for time-varying alphas proposed by Gagliardini et al. (2016), as it accounts for the unbalanced nature of the panel of individual funds, and allows for time-variation in betas driven both by common variables as well as asset-specific variables. Nevertheless, the results from the time-invariant methods corroborate the ones from our time-varying method and provide additional information regarding reliability and robustness. Relative to the literature, we are not only able to test the zero-alpha hypothesis, but also show the time-varying ratio of non-zero alphas.

To be specific, we collect monthly mutual fund return data over the Feb1962-Sept2017 period, and hedge fund return data over the Jan1994-Apr2017 period. Applying these tests to these two real data sets, our results strongly reject the zero-alpha joint hypothesis (Jensen, 1968; Carhart, 1997; Fama and French, 2010; Blake et al., 2014, 2017). This observation is especially prominent during the late 2000s Global Financial Crisis, when the performance of both mutual funds and hedge funds diverges.

Moreover, the hedge funds outperform both the market and mutual funds on average. In addition, we find some evidence supporting the existence of some arbitrage opportunities among mutual funds, but little evidence supporting the existence of arbitrage opportunities among hedge funds, which implied a better profitability for funds of mutual funds than funds of hedge funds. Altogether, we provide fresh evidence supporting the claim that *'hedge funds are the investor class that most closely resembles textbook arbitrageurs'* (e.g., Ben-David et al. (2012); Gao et al. (2018)).

Reference

Akbas, F., Armstrong, W. J., Sorescu, S. and Subrahmanyam, A. (2015), 'Smart money, dumb money, and capital market anomalies', *Journal of Financial Economics* **118**(2), 355–382.

Andrikogiannopoulou, A. and Papakonstantinou, F. (2019), 'Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power?', *Journal of Finance* **74**(5), 2667–2688.

Ang, A. and Kristensen, D. (2012), 'Testing conditional factor models', *Journal of Financial Economics* **106**(1), 132–156.

Aragon, G. O., Martin, J. S. and Shi, Z. (2019), 'Who benefits in a crisis? evidence from hedge fund stock and option holdings', *Journal of Financial Economics* **131**(2), 345–361.

Avramov, D. and Chordia, T. (2006), 'Asset pricing models and financial market anomalies', *fReview of Financial Studies* **19**(3), 1001–1040.

Barras, L. (2019), 'A large-scale approach for evaluating asset pricing models', *Journal of Financial Economics* **134**(3), 549–569.

Barras, L., Gagliardini, P. and Scaillet, O. (2018), 'The cross-sectional distribution of fund skill measures', *Swiss Finance Institute Research Paper* (18-66).

Barras, L., Scaillet, O. and Wermers, R. (2010), 'False discoveries in mutual fund performance: Measuring luck in estimated alphas', *Journal of Finance* **65**(1), 179–216.

Beltratti, A. and Stulz, R. M. (2012), 'The credit crisis around the globe: Why did some banks perform better?', *Journal of Financial Economics* **105**(1), 1–17.

Ben-David, I., Franzoni, F. and Moussawi, R. (2012), 'Hedge fund stock trading in the financial crisis of 2007–2009', *Review of Financial Studies* **25**(1), 1–54.

Berk, J. B. and van Binsbergen, J. H. (2015), 'Measuring skill in the mutual fund industry', *Journal of Financial Economics* **118**(1), 1–20.

Blake, D., Caulfield, T., Ioannidis, C. and Tonks, I. (2014), 'Improved inference in the evaluation of mutual fund performance using panel bootstrap methods', *Journal of Econometrics* **183**(2), 202–210.

Blake, D., Caulfield, T., Ioannidis, C. and Tonks, I. (2017), 'New evidence on mutual fund performance: A comparison of alternative bootstrap methods', *Journal of Financial and Quantitative Analysis* **52**(03), 1279–1299.

Blake, D., Rossi, A. G., Timmermann, A., Tonks, I. and Wermers, R. (2013), 'Decentralized investment management: Evidence from the pension fund industry', *Journal of Finance* **68**(3), 1133–1178.

Cai, B., Cheng, T. and Yan, C. (2018), 'Time-varying skills (versus luck) in U.S. active mutual funds and hedge funds', *Journal of Empirical Finance* **49**(12), 81–106.

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

Cheng, T. and Yan, C. (2017), 'Evaluating the size of the bootstrap method for fund performance evaluation', *Economics Letters* **156**, 36–41.

Chordia, T., Goyal, A. and Saretto, A. (2020), 'Anomalies and false rejections', *Review of Financial Studies* **33**(5), 2134–2179.

Elton, E. J., Gruber, M. J. and Blake, C. R. (2001), 'A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases', *Journal of Finance* **56**(6), 2415–2430.

Evans, R. B. (2010), 'Mutual fund incubation', *Journal of Finance* **65**(4), 1581–1611.

Fama, E. F. and French, K. R. (2010), 'Luck versus skill in the cross-section of mutual fund returns', *Journal of Finance* **65**(5), 1915–1947.

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), 603–680.

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

Ferson, W. and Chen, Y. (2020), How many good and bad fund managers are there, really?, *in* C. Lee, ed., 'Handbook of Financial Econometrics, Mathematics, Statistics, and Technology', World Scientific Publishing, p. Ch. 108.

Fung, W. and Hsieh, D. A. (1997), 'Empirical characteristics of dynamic trading strategies: The case of hedge funds', *Review of Financial Studies* **10**(2), 275–302.

Fung, W. and Hsieh, D. A. (2001), 'The risk in hedge fund strategies: Theory and evidence from trend followers', *Review of Financial Studies* **14**(2), 313–341.

Gagliardini, P., Ossola, E. and Scaillet, O. (2016), 'Time-varying risk premium in large cross-sectional equity data sets', *Econometrica* **84**(3), 985–1046.

Gao, G. P., Gao, P. and Song, Z. (2018), 'Do hedge funds exploit rare disaster concerns?', *Review of Financial Studies* **31**(7), 2650–2692.

Gibbons, M. R., Ross, S. and Shanken, J. (1989), 'A test of the efficiency of a given portfolio', *Econometrica* **57**(5), 1121–52.

Giglio, S., Liao, Y. and Xiu, D. (2019), 'Thousands of alpha tests', *Chicago Booth Research Paper* (18-09).

Gungor, S. and Luger, R. (2016), 'Multivariate tests of mean-variance efficiency and spanning with a large number of assets and time-varying covariances', *Journal of Business & Economic Statistics* **34**(2), 161–175.

Harvey, C. R. (2017), 'Presidential address: The scientific outlook in financial economics', *Journal of Finance* **72**(4), 1399–1440.

Harvey, C. R. and Liu, Y. (2018), 'Detecting repeatable performance', *Review of Financial Studies* **31**(7), 2499–2552.

Harvey, C. R. and Liu, Y. (2019*a*), 'Cross-sectional alpha dispersion and performance evaluation', *Journal of Financial Economics* **134**(2), 273–296.

Harvey, C. R. and Liu, Y. (2019*b*), 'Lucky factors', *SSRN working paper 2528780* .

Harvey, C. R. and Liu, Y. (2020), 'False (and missed) discoveries in financial economics', *Journal of Finance* **forthcoming**.

Harvey, C. R., Liu, Y. and Saretto, A. (2020), 'An evaluation of alternative multiple testing methods for finance applications', *Review of Asset Pricing Studies* **forthcoming**.

Harvey, C. R., Liu, Y. and Zhu, H. (2016), '. . . and the cross-section of expected returns', *Review of Financial Studies* **29**(1), 5–68.

Hau, H. and Lai, S. (2017), 'The role of equity funds in the financial crisis propagation', *Review of Finance* **21**(1), 77–108.

Hunter, D., Kandel, E., Kandel, S. and Wermers, R. (2014), 'Mutual fund performance evaluation with active peer benchmarks', *Journal of Financial Economics* **112**(1), 1–29.

Ioannidis, J., Stanley, T. and Doucouliagos, H. (2017), 'The power of bias in economics research', *Economic Journal* **127**(605), F236–F265.

Jensen, M. C. (1968), 'The performance of mutual funds in the period 1945–1964', *Journal of Finance* **23**(2), 389–416.

Jiao, Y. (2012), 'Hedge funds and equity prices', *Review of Finance* **17**(3), 1141–1177.

Jotikasthira, C., Lundblad, C. and Ramadorai, T. (2012), 'Asset fire sales and purchases and the international transmission of funding shocks', *Journal of Finance* **67**(6), 2015–2050.

Kokkonen, J. and Suominen, M. (2015), 'Hedge funds and stock market efficiency', *Management Science* **61**(12), 2890–2904.

Kosowski, R., Naik, N. Y. and Teo, M. (2007), 'Do hedge funds deliver alpha? a bayesian and bootstrap analysis', *Journal of Financial Economics* **84**(1), 229–264.

Kosowski, R., Timmermann, A., Wermers, R. and White, H. (2006), 'Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis', *Journal of Finance* **61**(6), 2551–2595.

Ledoit, O. and Wolf, M. (2004), 'A well-conditioned estimator for large-dimensional covariance matrices', *Journal of Multivariate Analysis* **88**(2), 365–411.

Ma, S., Lan, W., Su, L. and Tsai, C.-L. (2020), 'Testing alphas in conditional time-varying factor models with high-dimensional assets', *Journal of Business & Economic Statistics* **38**(1), 214–227.

Mamaysky, H., Spiegel, M. and Zhang, H. (2007), 'Improved forecasting of mutual fund alphas and betas', *Review of Finance* **11**(3), 359–400.

Mamaysky, H., Spiegel, M. and Zhang, H. (2008), 'Estimating the dynamics of mutual fund alphas and betas', *Review of Financial Studies* **21**(1), 233–264.

Pástor, L. and Stambaugh, R. F. (2012), 'On the size of the active management industry', *Journal of Political Economy* **120**(4), 740–781.

Pástor, L., Stambaugh, R. F. and Taylor, L. A. (2015), 'Scale and skill in active management', *Journal of Financial Economics* **116**(1), 23–45.

Pesaran, M. H. and Yamagata, T. (2018), 'Testing for alpha in linear factor pricing models with a large number of securities', *CESifo Working Paper Series No. 6432* .

Petersen, M. A. (2009), 'Estimating standard errors in finance panel data sets: Comparing approaches', *Review of Financial Studies* **22**(1), 435–480.

Raddatz, C. and Schmukler, S. L. (2012), 'On the international transmission of shocks: Micro-evidence from mutual fund portfolios', *Journal of International Economics* **88**(2), 357–374.

Ross, S. A. (1976), 'The arbitrage theory of capital asset pricing', *Journal of Economic Theory* **13**(3), 341–60.

Stambaugh, R. F., Yu, J. and Yuan, Y. (2012), 'The short of it: Investor sentiment and anomalies', *Journal of Financial Economics* **104**(2), 288–302.

Stambaugh, R. F., Yu, J. and Yuan, Y. (2015), 'Arbitrage asymmetry and the idiosyncratic volatility puzzle', *Journal of Finance* **70**(5), 1903–1948.

Yan, C. and Cheng, T. (2019), 'In search of the optimal number of fund subgroups', *Journal of Empirical Finance* **50**(1), 78–92.

Yan, X. and Zheng, L. (2017), ‘Fundamental analysis and the cross-section of stock returns: A data-mining approach’, *Review of Financial Studies* **30**(4), 1382–1423.

Zhang, H. and Yan, C. (2018), ‘A skeptical appraisal of the bootstrap approach in fund performance evaluation’, *Financial Markets, Institutions and Instruments* **27**(2), 49–86.

Table 1: **Summary statistics.** Monthly returns are summarized for mutual funds over the Feb1962-Sept2017 period (top panel) and hedge funds over the Jan1994-Apr2017 period (bottom panel), measured in excess of the one-month return of a three-month Treasury bill. The values at the cutoff points for various quantiles of the cross-sectional distributions of the sample of funds are reported. Each column is sorted on the statistic shown. Nobs is the number of available monthly returns, where a minimum of 8 are required for the left top (and left bottom) panel, while a minimum of 30 are required for the right top (and right bottom) panel. Mean is the sample mean return, Std is the sample standard deviation of return, and Rho1 is the first order sample autocorrelation. The alpha estimates are based on OLS regressions using the Fama-French-Carhart four-factors (Carhart (1997)) for mutual funds, while the Fung-Hsieh seven-factors (Fung and Hsieh (1997, 2001) are used for the hedge funds.

Quantiles	Mutual funds (minimum 8 obs)					Mutual funds (minimum 30 obs)				
	Nobs	Mean	Std	Rho1	$\widehat{\alpha}_{ols}$	Nobs	Mean	Std	Rho1	$\widehat{\alpha}_{ols}$
Top	668	0.119	0.640	0.978	0.106	668	0.040	0.640	0.978	0.040
1%	416	0.023	0.111	0.805	0.006	483	0.015	0.106	0.816	0.006
5%	224	0.014	0.073	0.391	0.003	228	0.012	0.071	0.409	0.003
10%	200	0.011	0.062	0.284	0.002	211	0.010	0.061	0.287	0.002
20%	155	0.009	0.052	0.197	0.001	169	0.008	0.053	0.204	0.001
30%	130	0.008	0.046	0.157	0.001	141	0.007	0.048	0.165	0.001
Median	90	0.005	0.037	0.074	-0.000	106	0.005	0.039	0.093	-0.000
30%	46	0.003	0.026	-0.052	-0.001	68	0.003	0.029	-0.012	-0.001
20%	32	0.001	0.019	-0.106	-0.002	56	0.002	0.022	-0.074	-0.002
10%	20	-0.001	0.011	-0.188	-0.004	42	-0.000	0.011	-0.132	-0.003
5%	11	-0.006	0.005	-0.280	-0.005	34	-0.003	0.006	-0.178	-0.005
1%	8	-0.023	0.000	-0.517	-0.011	31	-0.014	0.000	-0.260	-0.010
Bottom	8	-0.069	0.000	-0.843	-0.116	30	-0.046	0.000	-0.560	-0.027
Quantiles	Hedge funds (minimum 8 obs)					Hedge funds (minimum 30 obs)				
	Nobs	Mean	Std	Rho1	$\widehat{\alpha}_{ols}$	Nobs	Mean	Std	Rho1	$\widehat{\alpha}_{ols}$
Top	259	0.165	1.474	0.814	0.868	259	0.165	1.474	0.814	0.178
1%	221	0.027	0.176	0.579	0.021	229	0.022	0.156	0.568	0.018
5%	160	0.013	0.095	0.440	0.010	174	0.012	0.086	0.438	0.009
10%	129	0.009	0.072	0.372	0.007	148	0.009	0.068	0.372	0.007
20%	98	0.006	0.051	0.282	0.005	113	0.006	0.050	0.286	0.005
30%	75	0.004	0.041	0.224	0.003	96	0.004	0.041	0.234	0.003
Median	47	0.002	0.028	0.127	0.002	69	0.002	0.028	0.146	0.002
30%	28	-0.001	0.020	0.023	-0.000	50	0.000	0.020	0.057	0.000
20%	20	-0.004	0.016	-0.045	-0.001	43	-0.001	0.017	0.005	-0.001
10%	13	-0.010	0.012	-0.142	-0.004	36	-0.004	0.013	-0.082	-0.003
5%	10	-0.017	0.010	-0.237	-0.008	33	-0.007	0.010	-0.143	-0.006
1%	8	-0.036	0.006	-0.460	-0.021	30	-0.018	0.007	-0.311	-0.016
Bottom	8	-0.120	0.000	-0.794	-0.254	30	-0.035	0.001	-0.575	-0.039

Table 2: **Test statistics based on Fan et al. (2015) when T = 60.** This table reports the summary statistics calculated for the funds selected by applying the method based on Fan et al. (2015) to monthly mutual fund return data over the Feb1962-Sept2017 period, and hedge fund return data over the Jan1994-Apr2017 period. The alpha estimates are based on OLS regressions using the Fama-French-Carhart four-factors (Carhart (1997)) for mutual funds, while the Fung-Hsieh seven-factors (Fung and Hsieh (1997, 2001) are used for the hedge funds. N denotes the panel size, and $\widehat{S}$ denotes the screening set. Panel A and B present the results for mutual funds and hedge funds, respectively.

Variables	Mean	Std	Median	Min	Max
Panel A: Mutual funds with Fan et al. (2015) screening rule					
N	617.1905	954.4452	101	33	3068
$\widehat{S}$	33.5599	59.2543	8	1	360
$\lvert\widehat{\alpha}_j\rvert(\%)$	0.2444	0.3042	0.1688	0.0000	21.4154
$\lvert\widehat{\alpha}_j\rvert_{j\in\widehat{S}}(\%)$	0.3512	0.4072	0.2824	0.0013	4.7316
Panel B: Hedge funds with Fan et al. (2015) screening rule					
N	309.0226	175.4781	410	0	518
$\widehat{S}$	37.6172	19.9913	31	0	111
$\lvert\widehat{\alpha}_j\rvert(\%)$	0.4825	1.5493	0.3035	0.0000	57.9476
$\lvert\widehat{\alpha}_j\rvert_{j\in\widehat{S}}(\%)$	0.8106	0.5595	0.6478	0.0032	3.9126

Table 3: **No-arbitrage restrictions test results for mutual funds based on Gagliardini et al. (2016) .** In Panel A, we compute the statistics $\tilde{\Sigma}_{\xi}^{-1/2}\widehat{\xi}_{NT}$ based on $\widehat{Q}_e$ and $\widehat{Q}_a$ to test the null hypotheses $H_0 : \alpha_i = \beta_i^{\top} \nu$ and $\alpha_i = 0$, respectively. In panel B, we compute the statistics $\tilde{\Sigma}_{\xi}^{-1/2}\widehat{\xi}_{NT}$ to test the null hypotheses $H_0 : b_{1,i} = b_{3,i}\nu$ and $H_0 : b_{1,i} = 0$. The trimming levels are $\chi_{1,T} = 15$ and $\chi_{2,T} = T/240$. The table reports the p-values of the statistics in parentheses.

	Test for $H_0 : \alpha_i = \beta_i^{\top} \nu$	Test for $H_0 : \alpha_i = 0$
Panel A: Time-invariant model (1)		
Fama-French-Carhart Four-factor model		
Test statistics	86.2282	82.9258
p-value	(0.0000)	(0.0000)
Fama-French Three-factor model		
Test statistics	94.8126	93.3542
p-value	(0.0000)	(0.0000)
CAPM		
Test statistics	95.8000	95.7806
p-value	(0.0000)	(0.0000)
Panel B: Time-varying model (5)	Test for $H_0 : \alpha_i = \beta_i^{\top} \nu$	Test for $H_0 : \alpha_i = 0$
Fama-French-Carhart Four-factor model		
Test statistics	22.6605	24.6587
p-value	(0.0000)	(0.0000)
Fama-French Three-factor model		
Test statistics	28.0767	30.7408
p-value	(0.0000)	(0.0000)
CAPM		
Test statistics	27.6421	29.0969
p-value	(0.0000)	(0.0000)

Table 4: **No-arbitrage restrictions test results for hedge funds based on Gagliardini et al. (2016) .** In Panel A, we compute the statistics $\tilde{\Sigma}_{\xi}^{-1/2}\widehat{\xi}_{NT}$ based on $\widehat{Q}_e$ and $\widehat{Q}_a$ to test the null hypotheses $H_0 : \alpha_i = \beta_i^{\top} \nu$ and $\alpha_i = 0$, respectively. In panel B, we compute the statistics $\tilde{\Sigma}_{\xi}^{-1/2}\widehat{\xi}_{NT}$ to test the null hypotheses $H_0 : b_{1,i} = b_{3,i}\nu$ and $H_0 : b_{1,i} = 0$. The trimming levels are $\chi_{1,T} = 15$ and $\chi_{2,T} = T/240$. The table reports the p-values of the statistics in parentheses.

Panel A: Time-invariant model (1)		
	Test for $H_0 : \alpha_i = \beta_i^{\top} \nu$	Test for $H_0 : \alpha_i = 0$
Fung-Hsieh Seven-factor model		
Test statistics	-1.6504	17.1068
p-value	(0.9506)	(0.0000)
Panel B: Time-varying model (5)		
	Test for $H_0 : \alpha_i = \beta_i^{\top} \nu$	Test for $H_0 : \alpha_i = 0$
Fung-Hsieh Seven-factor model		
Test statistics	0.3378	6.2219
p-value	(0.3678)	(0.0000)

Figure 1: **Results of mutual fund using the method based on Pesaran and Yamagata (2018).** The left (right) top graph shows the J test statistics (and corresponding p-values) of the alphas for all the funds in each 60-month window; The left (right) middle graph shows J test statistics (and corresponding p-values) for the positive alphas in each 60-month window; The left (right) bottom graph shows J test statistics (and corresponding p-values) for the negative alphas in each 60-month window.

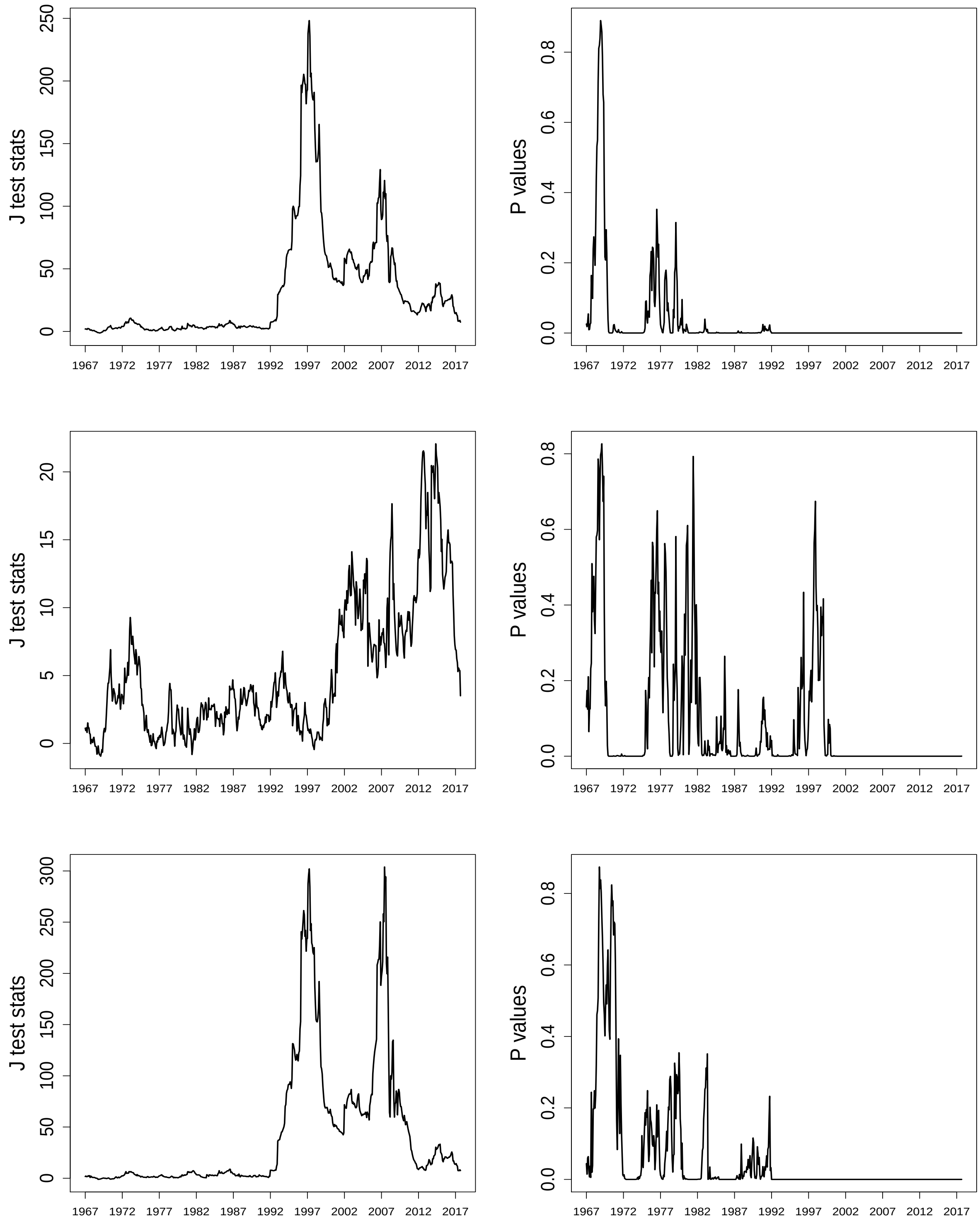


Figure 2: **Results of hedge fund using the method based on Pesaran and Yamagata (2018).** The left (right) top graph shows the J test statistics (and corresponding p-values) of the alphas for all the funds in each 60-month window; The left (right) middle graph shows J test statistics (and corresponding p-values) for the positive alphas in each 60-month window; The left (right) bottom graph shows J test statistics (and corresponding p-values) for the negative alphas in each 60-month window.

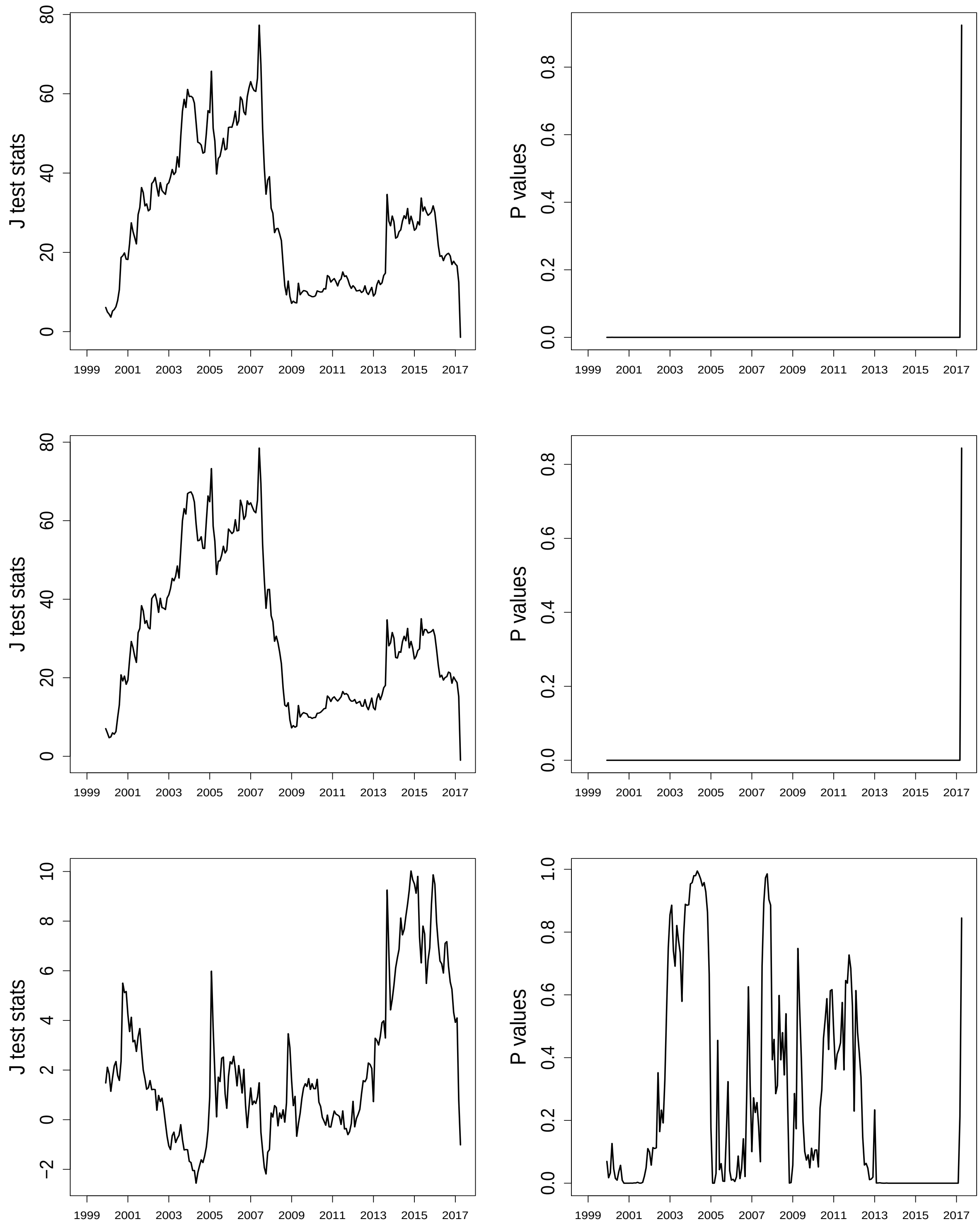


Figure 3: **Test statistics based on Ma et al. (2020) when $T = 60$.** The left (right) graph shows the p-values of the long-run alphas for mutual (hedge) funds in each 60-month window.

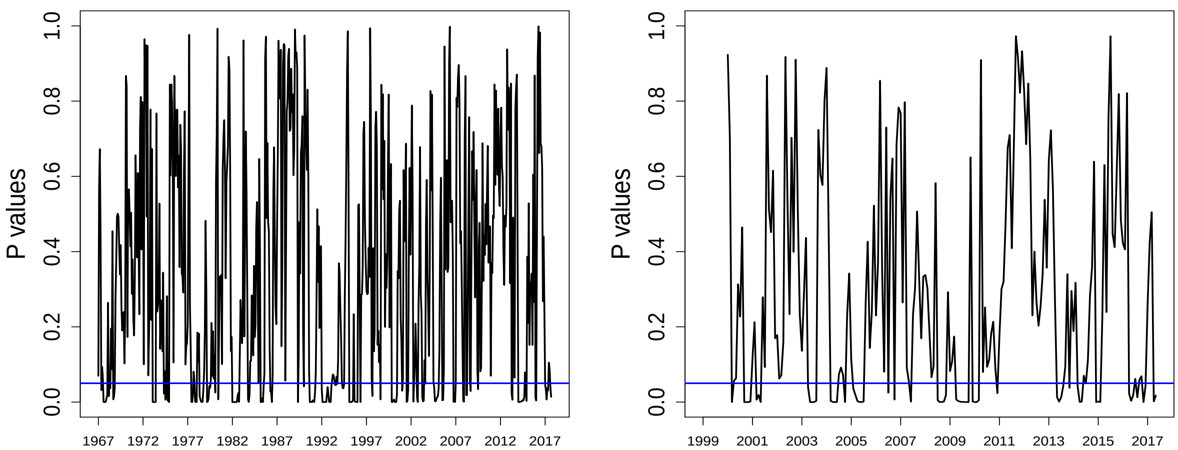


Figure 4: **Results of mutual funds based on time-invariant factor model (1) with the threshold of $|t| = 3$ and 60-month window.** The left (right) top graph shows the average value (number) of selected positive alphas and negative alphas. The left middle graph shows the ratio of selected positive and negative alphas in the screening set while the right middle graph shows the ratio of selected positive and negative alphas in the whole panel size during the test period. The left (right) bottom graph shows the alpha (t-statistic of alpha) of each fund during each 60-month window.

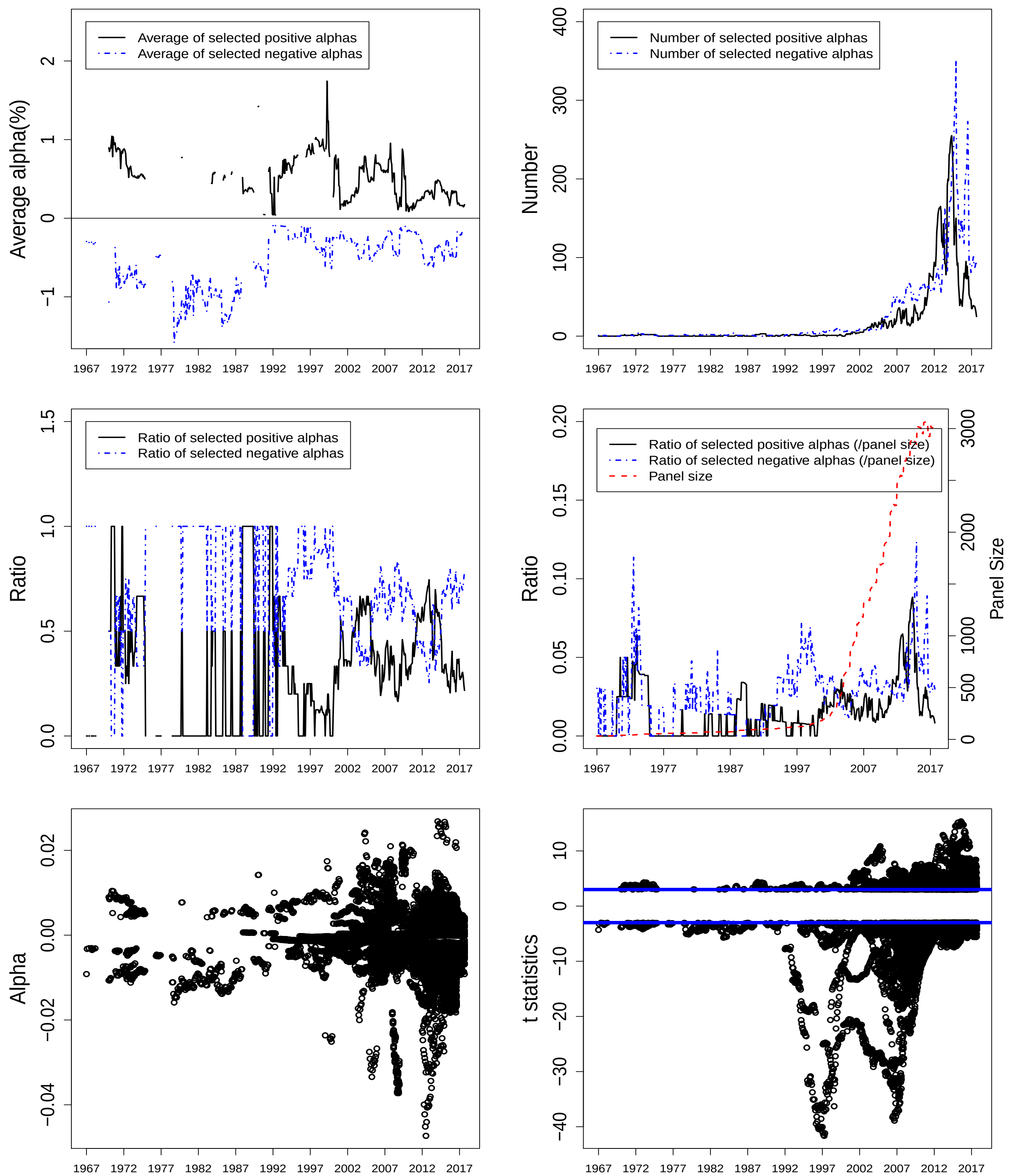


Figure 5: **Results of hedge funds based on time-invariant factor model (1) with the threshold of $|t| = 3$ and 60-month window.** The left (right) top graph shows the average value (number) of selected positive alphas and negative alphas. The left middle graph shows the ratio of selected positive and negative alphas in the screening set while the right middle graph shows the ratio of selected positive and selected negative alphas in the whole panel size during the test period. The left (right) bottom graph shows the alpha (t-statistic of alpha) of each fund during each 60-month window.

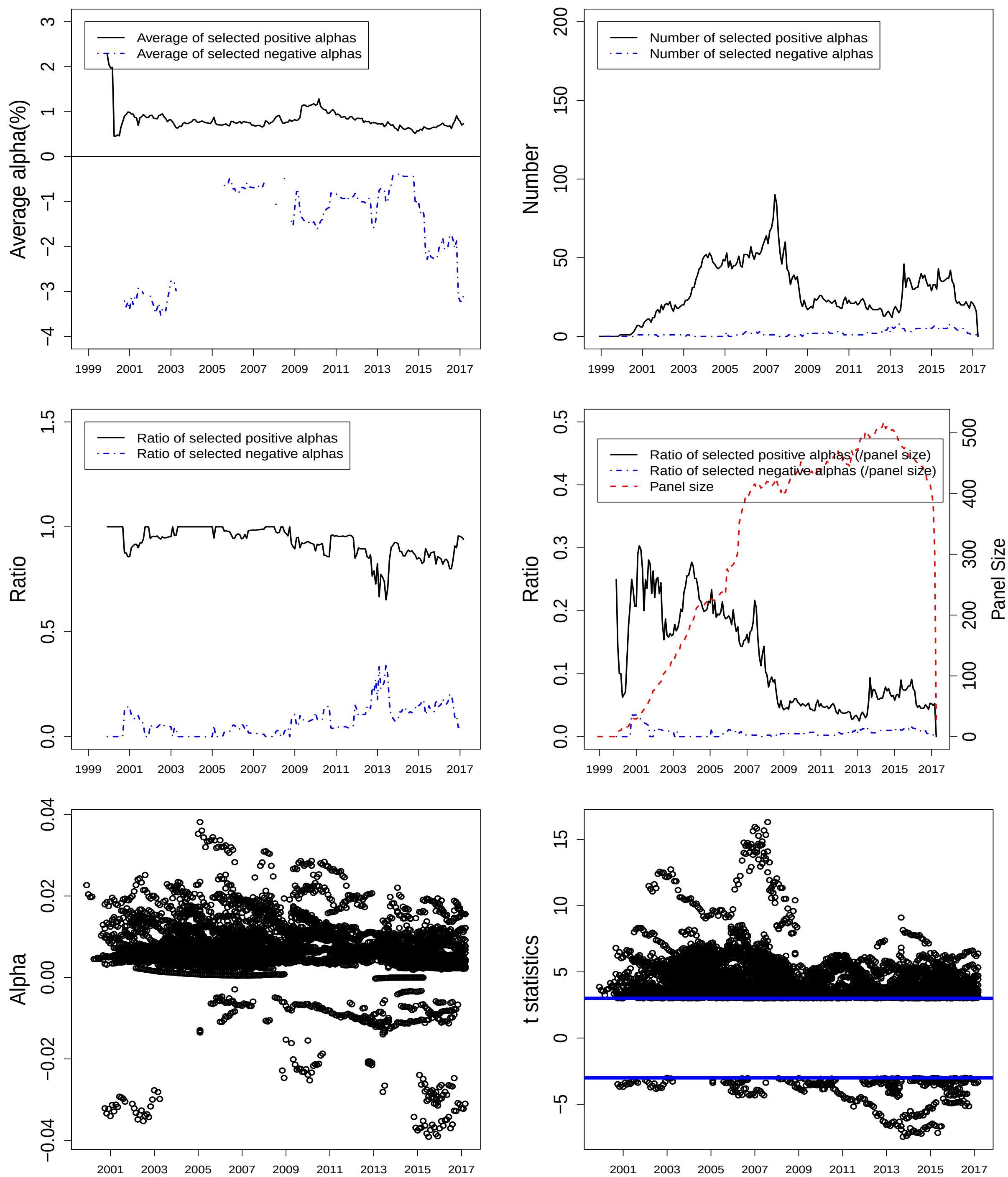


Figure 6: **Results of mutual funds based on Gagliardini et al. (2016) with the threshold of $|t| = 3$.** The left (right) top graph shows the average value (number) of selected positive alphas and negative alphas. The left middle graph shows the ratio of selected positive and negative alphas in the screening set while the right middle graph shows the ratio of selected positive and selected negative alphas in the whole panel size. The left (right) bottom graph shows the alpha (t-statistic of alpha) of each fund.

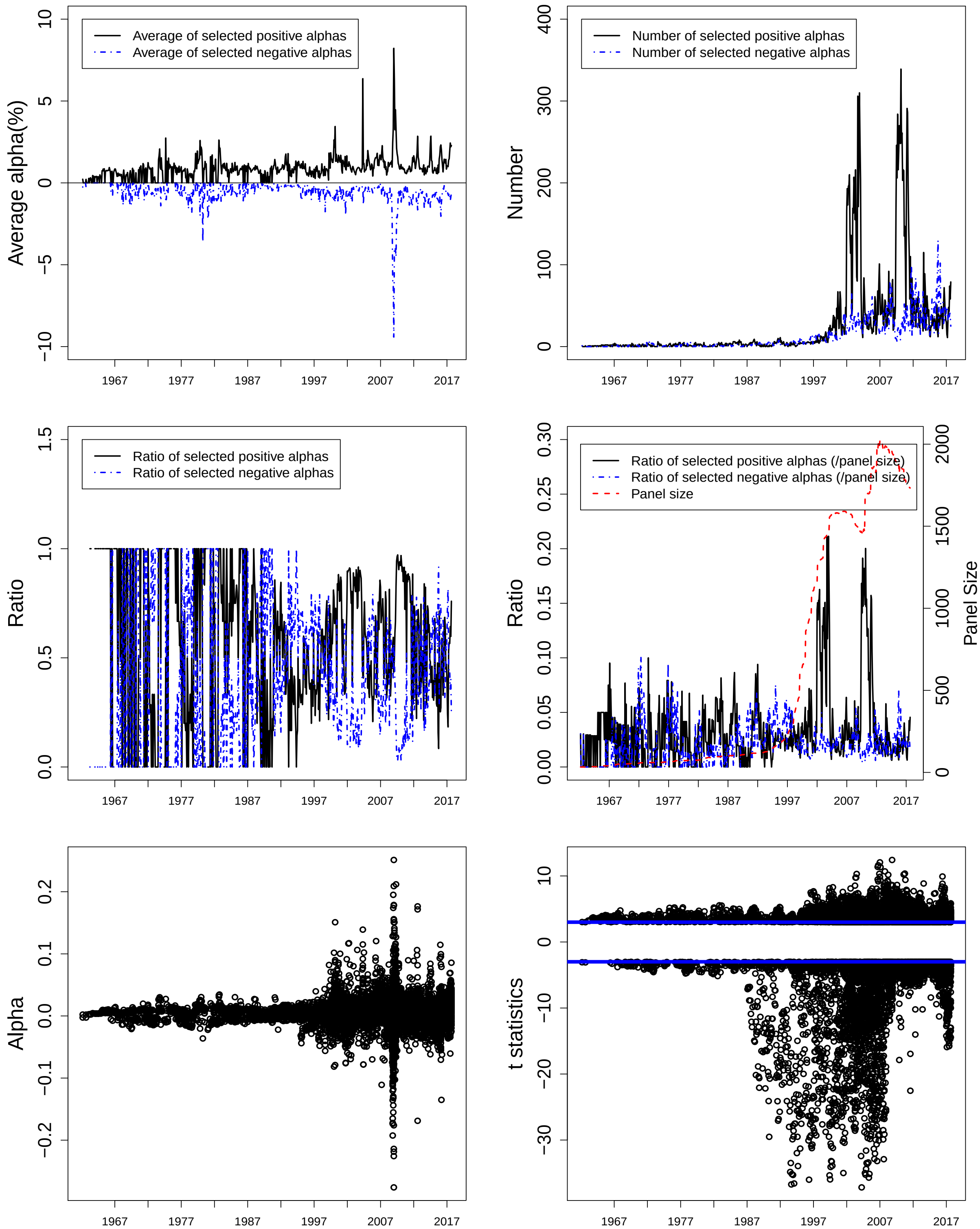


Figure 7: **Results of hedge funds based on Gagliardini et al. (2016) with the threshold of $|t| = 3$.** The left (right) top graph shows the average value (number) of selected positive alphas and negative alphas. The left middle graph shows the ratio of selected positive and negative alphas in the screening set while the right middle graph shows the ratio of selected positive and selected negative alphas in the whole panel size. The left (right) bottom graph shows the alpha (t-statistic of alpha) of each fund.

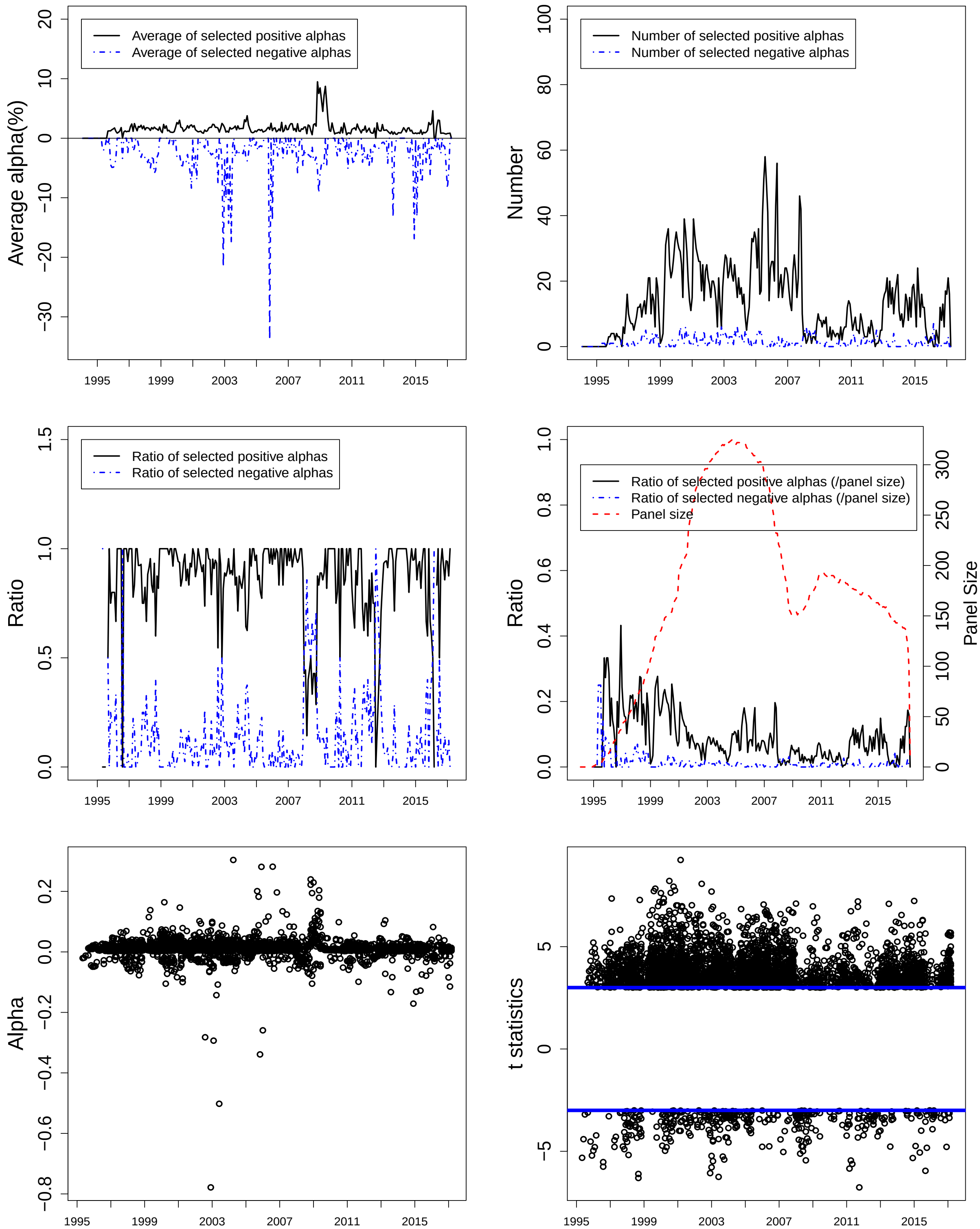


Figure 8: **Results of mutual funds based on Ma et al. (2020) with the threshold of $|t| = 3$.** The left (right) top graph shows the average value (number) of selected positive alphas and negative alphas. The left middle graph shows the ratio of selected positive and negative alphas in the screening set while the right middle graph shows the ratio of selected positive and selected negative alphas in the whole panel size. The left (right) bottom graph shows the alpha (t-statistic of alpha) of each fund.

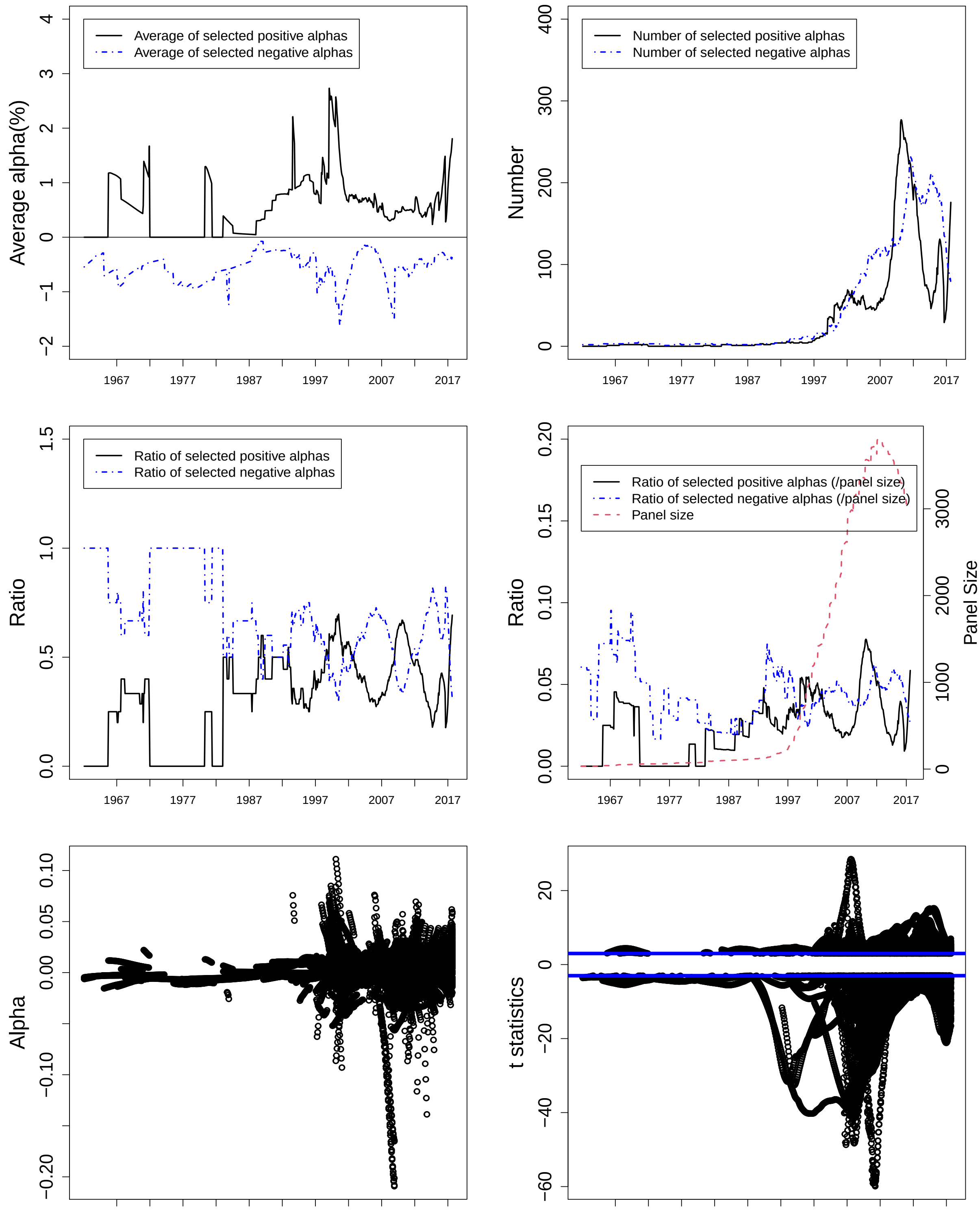
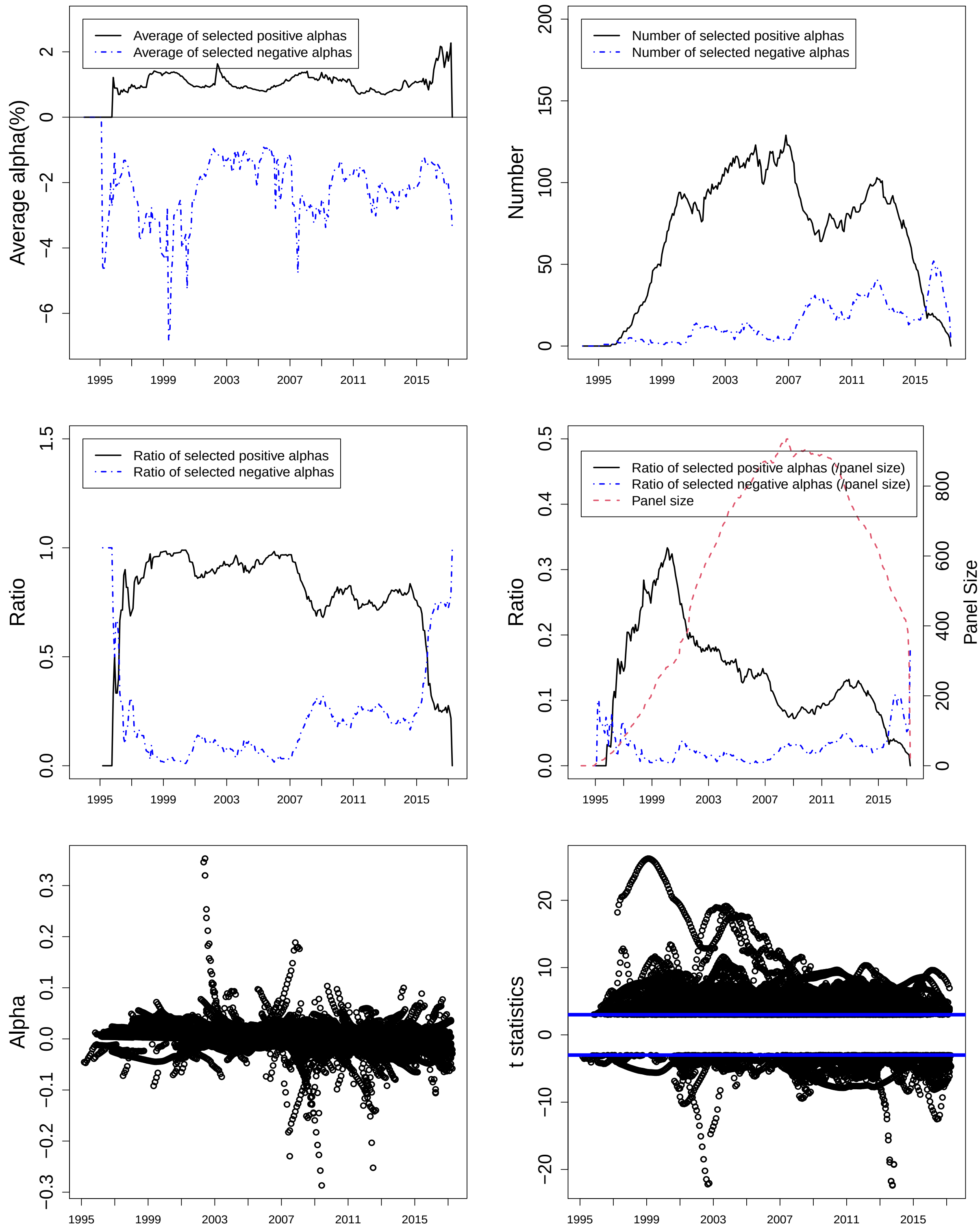


Figure 9: **Results of hedge funds based on Ma et al. (2020) with the threshold of $|t| = 3$.** The left (right) top graph shows the average value (number) of selected positive alphas and negative alphas. The left middle graph shows the ratio of selected positive and negative alphas in the screening set while the right middle graph shows the ratio of selected positive and selected negative alphas in the whole panel size. The left (right) bottom graph shows the alpha (t-statistic of alpha) of each fund.



Appendix A. Simulation-based choice of t-statistic threshold

In our paper, we calculate the ratio of zero alphas and non-zero alphas by using the t-statistic threshold of 3. Now in this appendix, we justify this choice and provide a guide to the choice of thresholds. To this end, we present simulation evidence on the relative performance of different thresholds. We provide researchers with suggestions as to which threshold is best for the particular data set at hand. For rigorous proof, we leave it for future research.

For alpha estimation, it was routine to consider a t-statistic of 1.96 or 2.0 as the threshold for statistical significance. Due to the multiple testing problem, these views have changed recently, as the traditional threshold of t-statistic of 1.96 or 2.0 is not enough to establish significance. Therefore, we need to make an adjustment to deal with the problem. Here we consider increasing the t-statistic threshold, which is based on the principle that the Type I error can be reduced by increasing the t-statistic threshold and reduce the false positives.

Assume that fund returns are simulated from the market model below.

$$r_{it} = \alpha_i + \beta_i x_t + e_{it}, \quad i = 1, 2, \cdots, N, \quad t = 1, 2, \cdots, T \tag{A.1}$$

where $\beta_i = 1$ for $i = 1, 2, \cdots, N$, x_t is generated from a Gaussian distribution N(0.006, 0.043) and e_{it} is generated from a Gaussian distribution N(0, $\sqrt{0.05^2/12}$). We assume that there are 10% non-zero alphas. For simplicity, we set 10% non-zero alphas to be 0.12/12. The rest of alphas are set to be zeros. The above setting is based on real data characteristics and is similar to the one in Harvey et al. (2020). We consider ten pairs of (N, T): $(N, T) = (100, 30)$, $(100, 60)$, $(100, 100)$, $(100, 200)$ and $(100, 500)$ for cases 1-5 while cases 6-10 are set to be the same as cases 1-5 except that $N = 500$. For each pair of (N, T), we do 1000 replications.

We aim to get the ratios of non-zero alphas including the ratios of positive and negative alphas. The procedure is as follows. We first get the estimate of alpha by ordinary least squares method and then test the null hypothesis that the true alpha is zero. Hypotheses will be evaluated by comparing t-statistic with some given threshold. If the absolute value of the t-statistic is larger than the threshold, the corresponding null hypothesis will be rejected. From these test results, we can further compute the ratios of non-zero alphas and also the ratios of positive and negative alphas. In this study, for each case, we try four different thresholds of t-statistic: 1.96, 2.5, 3 and 3.5. According to Tables A.5, we can see that

- In case 1 and case 6 ($T = 30$), using the threshold of t-statistic 2.5, the detected positive ratios are much closer to the true positive ratios compared with the results with other thresholds. Though by applying the thresholds 3 or 3.5, we can achieve a smaller type I error (as the ratios of negative alphas is closer to the true value: zero), the type II error is too large and as a result, the test does not have enough power to detect the skilled manager. Therefore, in these cases, a threshold of 2.5 is more suitable.

- In case 2 and case 7 ($T = 60$), the threshold of t-statistic 3 is more suitable. It is easy to see that the thresholds 1.96 and 2.5 have no advantage compared with the thresholds 3 and 3.5. Using the threshold of 3.5, the test has a smaller type I error but a larger type II error, which means that it does not have enough test power to reject the false null hypothesis or it lacks the power to detect outperforming managers.

- In cases 3-5 and cases 8-10, when T is larger, the threshold of 3.5 has the absolute advantage compared with the other three thresholds.

To sum up, we find that with a different number of observations, the threshold should be different. In addition, we see that when the length of time-series is about 60, the threshold of t-statistic should be around 3.

Table A.5: **Simulation results of fund-by-fund OLS method based on** $t = 1.96$, $t = 2.5$, $t = 3$ **and** $t = 3.5$. This table reports the averages (Mean) and standard deviations (Std.) of ratios of non-zero alphas. Ratios of positive (Pos.) and negative (Neg.) non-zero alphas using OLS method based on $t = 1.96$, $t = 2.5$, $t = 3$ and $t = 3.5$ for 1000 replications.

		t=1.96			t=2.5			t=3			t=3.5		
		Overall	Pos.	Neg.	Overall	Pos.	Neg.	Overall	Pos.	Neg.	Overall	Pos.	Neg.
True value		0.1	0.1	0	0.1	0.1	0	0.1	0.1	0	0.1	0.1	0
Case 1	Mean	0.1551	0.1243	0.0309	0.1105	0.1001	0.0104	0.0848	0.0816	0.0032	0.0572	0.0561	0.0011
	Std.	0.0247	0.0185	0.0170	0.0169	0.0134	0.0103	0.0155	0.0145	0.0056	0.0163	0.0158	0.0033
Case 2	Mean	0.1536	0.1267	0.0270	0.1154	0.1073	0.0080	0.1031	0.1011	0.0020	0.0969	0.0965	0.0004
	Std.	0.0224	0.0160	0.0165	0.0132	0.0092	0.0092	0.0071	0.0055	0.0045	0.0067	0.0065	0.0021
Case 3	Mean	0.1502	0.1252	0.0250	0.1143	0.1071	0.0072	0.1032	0.1016	0.0016	0.1007	0.1003	0.0004
	Std.	0.0215	0.0157	0.0156	0.0121	0.0087	0.0080	0.0059	0.0041	0.0041	0.0030	0.0021	0.0021
Case 4	Mean	0.1478	0.1232	0.0246	0.1130	0.1065	0.0065	0.1028	0.1015	0.0012	0.1008	0.1004	0.0004
	Std.	0.0218	0.0153	0.0156	0.0116	0.0080	0.0083	0.0053	0.0039	0.0035	0.0029	0.0021	0.0020
Case 5	Mean	0.1459	0.1232	0.0228	0.1115	0.1054	0.0060	0.1027	0.1014	0.0013	0.1004	0.1002	0.0002
	Std.	0.0209	0.0157	0.0148	0.0106	0.0075	0.0076	0.0052	0.0038	0.0036	0.0020	0.0014	0.0015
Case 6	Mean	0.1555	0.1244	0.0312	0.1106	0.1005	0.0100	0.0870	0.0837	0.0033	0.0622	0.0613	0.0009
	Std.	0.0112	0.0086	0.0079	0.0074	0.0062	0.0045	0.0067	0.0061	0.0026	0.0073	0.0071	0.0014
Case 7	Mean	0.1525	0.1263	0.0262	0.1151	0.1073	0.0078	0.1031	0.1011	0.0020	0.0978	0.0973	0.0005
	Std.	0.0098	0.0070	0.0072	0.0056	0.0041	0.0039	0.0031	0.0025	0.0021	0.0028	0.0027	0.0010
Case 8	Mean	0.1494	0.1247	0.0247	0.1135	0.1067	0.0069	0.1032	0.1016	0.0016	0.1007	0.1003	0.0004
	Std.	0.0097	0.0068	0.0072	0.0050	0.0037	0.0036	0.0025	0.0018	0.0018	0.0012	0.0009	0.0008
Case 9	Mean	0.1473	0.1237	0.0236	0.1127	0.1063	0.0064	0.1029	0.1013	0.0016	0.1006	0.1003	0.0003
	Std.	0.0097	0.0069	0.0070	0.0051	0.0036	0.0035	0.0024	0.0016	0.0018	0.0011	0.0007	0.0008
Case 10	Mean	0.1458	0.1229	0.0229	0.1118	0.1059	0.0059	0.1026	0.1012	0.0013	0.1005	0.1002	0.0002
	Std.	0.0094	0.0067	0.0066	0.0049	0.0034	0.0035	0.0022	0.0015	0.0017	0.0009	0.0007	0.0007

Appendix B. Technical details of the method based on Gagliardini et al. (2016)

Appendix B.1. Linearity in the transformed model

To account for time-variation, Gagliardini et al. (2016) propose a structural econometric framework to take into account the no-arbitrage restrictions to conduct formal frequentist inference for multi-factor models in large cross-sectional equity data sets. Under this framework, the time-varying alpha α_{it} is a quadratic form in lagged instruments Z_{t-1} and $Z_{i,t-1}$ given by $\alpha_{it} = Z_{t-1}^\top B_i^\top(\Lambda - D)Z_{t-1} + Z_{i,t-1}^\top C_i^\top(\Lambda - D)Z_{t-1}$. As a result, model (5) can be written as

$$r_{it} = Z_{t-1}^\top B_i^\top(\Lambda - D)Z_{t-1} + Z_{i,t-1}^\top C_i^\top(\Lambda - D)Z_{t-1} + Z_{t-1}^\top B_i^\top f_t + Z_{i,t-1}^\top C_i^\top f_t + \epsilon_{it}. \tag{B.1}$$

Rewrite model (B.1) as a model that is linear in transformed parameters and new regressors. The regressors include $x_{2,i,t} = (f_t^\top \otimes Z_{t-1}^\top, f_t^\top \otimes Z_{i,t-1}^\top)^\top \in R^{d_2}$ with $d_2 = K(p+q)$ and $x_{1,i,t} = (\text{vech}[X_t]^\top, Z_{t-1}^\top \otimes Z_{i,t-1}^\top)^\top \in R^{d_1}$ with $d_1 = p(p+1)/2 + pq$, where the symmetric matrix $X_t = [X_{t,k,l}] \in R^{p\times p}$ is such that $X_{t,k,l} = Z_{t-1,k}^2$, if $k = l$, and $X_{t,k,l} = 2Z_{t-1,k}Z_{t-1,l}$, otherwise, for $k, l = 1, ..., p$. Then

$$r_{it} = b_i^\top x_{i,t} + \epsilon_{it}, \tag{B.2}$$

where $x_{i,t} = (x_{1,i,t}^\top, x_{2,i,t}^\top)^\top$ has dimension $d = d_1 + d_2$, and $b_i = (b_{1,i}^\top, b_{2,i}^\top)^\top$ is such that

$$b_{1,i} = ((N_p[(\Lambda - D)^\top \otimes I_p]\text{vec}[B_i^\top])^\top, ([(\Lambda - D)^\top \otimes I_q]\text{vec}[C_i^\top])^\top)^\top,\ N_p = \frac{1}{2}D_p^+(W_p + I_{p^2}),$$
$$b_{2,i} = (\text{vec}[B_i^\top]^\top, \text{vec}[C_i^\top]^\top)^\top,$$

where the matrix D_p^+ is the $p(p+1)/2 \times p^2$ Moore-Penrose inverse of the duplication matrix D_p and $W_p := W_{pp}$ is the commutation matrix.

Appendix B.2. Two-pass estimation

In equation (B.3), the $d_1 \times 1$ vector $b_{1,i}$ is a linear transformation of the $d_2 \times 1$ vector $b_{2,i}$. The coefficients of the linear transformation depend on matrix $\Lambda - D$. We rewrite the parameter restrictions as

$$b_{1,i} = b_{3,i}\nu, \;\; \nu = \text{vec}[\Lambda^\top - D^\top], \\ b_{3,i} = ([N_p(B_i^\top \otimes I_p)]^\top, [W_{pq}(C_i^\top \otimes I_p)]^\top)^\top. \tag{B.3}$$

Furthermore, we can relate the $d_1 \times Kp$ matrix $b_{3,i}$ to the vector $b_{2,i}$

$$\text{vec}[b_{3,i}^\top] = J_a b_{2,i}, \tag{B.4}$$

where the $d_1 pK \times d_2$ block-diagonal matrix of constants J_a is $J_a = \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}$ with diagonal blocks $J_1 = W_{p(p+1)/2,pK}(I_K \otimes [(I_p \otimes N_p)(W_p \otimes I_p)(I_p \otimes \text{vec}[I_p])])$ and $J_2 = W_{pq,pK}(I_K \otimes [(I_p \otimes W_{p,q})(W_{p,q} \otimes I_p)(I_q \otimes \text{vec}[I_p])])$. The link (B.4) is instrumental in deriving the asymptotic results.

Consider a two-pass approach building on Equations (B.2) and (B.3). The first pass consists in computing time-series OLS estimators $\hat{b}_i = (\hat{b}_{1,i}^\top, \hat{b}_{2,i}^\top)^\top = \hat{Q}_{x,i}^{-1} \frac{1}{T_i} \sum_t I_{i,t} x_{i,t} r_{i,t}$, for $i = 1, \dots, N$, where $\hat{Q}_{x,i}^{-1} = \frac{1}{T_i} \sum_t I_{i,t} x_{i,t} x_{i,t}^\top$. In available unbalanced panels, the random sample size T_i for fund i can be small, and the inversion of matrix $\hat{Q}_{x,i}$ can be numerically unstable. This can yield unreliable estimates of b_i. To address this, we need a trimming device: $I_i^\chi = I\{CN(\hat{Q}_{x,i}) \leq \chi_{1,T}, \tau_{i,T} \leq \chi_{2,T}\}$, where $CN(\hat{Q}_{x,i}) = \sqrt{eig_{\max}(\hat{Q}_{x,i})/eig_{\min}(\hat{Q}_{x,i})}$ and $\tau_{i,T} = T/T_i$, and the two sequences $\chi_{1,T}$ and $\chi_{2,T}$ diverge asymptotically.

The second pass consists of a cross-sectional estimator of ν by regressing the $\hat{b}_{1,i}$ on the

$\widehat{b}_{3,i}$ keeping untrimmed funds only. The Weighted Least Squares (WLS) estimator is

$$\widehat{\nu} = \widehat{Q}_{b_3}^{-1} \frac{1}{N} \sum_i \widehat{b}_{3,i}^\top \widehat{\omega}_i \widehat{b}_{1,i}, \tag{B.5}$$

where $\widehat{Q}_{b_3} = \frac{1}{N}\sum_i \widehat{b}_{3,i}^\top \widehat{\omega}_i \widehat{b}_{3,i}$ and $\widehat{\omega}_i = I_i^\chi (\text{diag}[\widehat{\nu}_i])^\top$. We use the estimates $\widehat{\nu}_i = \tau_{i,T} C_{\widehat{\nu}_1}^\top \widehat{Q}_{x,i}^{-1} \widehat{S}_{ii} \widehat{Q}_{x,i}^{-1} C_{\widehat{\nu}_1}$, where $\widehat{S}_{ii} = \frac{1}{T_i}\sum_t I_{i,t} \widehat{\epsilon}_{it}^2 x_{i,t} x_{i,t}^\top$, $\widehat{\epsilon}_{it} = r_{it} - \widehat{b}_i^\top x_{it}$, $C_{\widehat{\nu}_1} = (E_1^\top - (I_{d_1} \otimes \widehat{\nu}_1^\top) J_a E_2^\top)^\top$ and $\widehat{\nu}_1 = (\sum_i I_i^\chi \widehat{b}_{3,i}^\top \widehat{b}_{3,i})^{-1} \times (\sum_i I_i^\chi \widehat{b}_{3,i}^\top \widehat{b}_{1,i})$, a first-step estimator with unit weights. The final estimator of the risk premia is $\widehat{\lambda}_t = \widehat{\Lambda} Z_{t-1}$, where we deduce Λ from the relationship $\text{vec}[\widehat{\Lambda}^\top] = \widehat{\nu} + \text{vec}[\widehat{D}^\top]$ with the estimator $\widehat{D}$ obtained by a regression of factors on lagged instruments Z_{t-1}: $\widehat{D} = \sum_t f_t Z_{t-1}^\top (\sum_t Z_{t-1} Z_{t-1}^\top)^{-1}$.

Appendix B.3. Tests of pricing restrictions

From (B.3), the null hypothesis underlying the asset pricing restriction $\alpha_{it} = \beta_{it}^\top \nu_t$ is H_0: there exists $\nu \in R^{pK}$ such that $b_{1,i} = b_{3,i}\nu$, for most all i. Under H_0, we have $E[(b_{1,i} - b_{3,i}\nu)^\top (b_{1,i} - b_{3,i}\nu)] = 0$. Since we estimate ν via the WLS cross-sectional regression of the estimates $\widehat{b}_{1,i}$ on the estimates $\widehat{b}_{3,i}$, we suggest a test based on the weighted Sum of Squared Residuals (SSR) of the cross-sectional regression. The weighted SSR is $\widehat{Q}_e = \frac{1}{N}\sum_i \widehat{e}_i^\top \widehat{\omega}_i \widehat{e}_i$, with $\widehat{e}_i = \widehat{b}_{1,i} - \widehat{b}_{3,i}\widehat{\nu} = C_{\widehat{\nu}}^\top \widehat{b_i}$. Define the statistic $\widehat{\xi}_{nT} = T\sqrt{N}(\widehat{Q}_e - \frac{1}{T}\widehat{B}_\xi)$, where the recentering term simplifies to $\widehat{B}_\xi = d_1$ thanks to the weighting scheme. Under the null hypothesis H_0, we prove that $\widehat{\xi}_{NT} \Rightarrow N(0, \Sigma_\xi)$, where $\Sigma_\xi = 2\lim_{N\to\infty} E[\frac{1}{N}\sum_{i,j} \frac{\tau_i^2 \tau_j^2}{\tau_{ij}^2} tr[(C_\nu^\top Q_{x,i}^{-1} S_{ij} Q_{x,j}^{-1} C_\nu)\omega_j (C_\nu^\top Q_{x,j}^{-1} S_{ji} Q_{x,i}^{-1} C_\nu)\omega_i]]$ as $N, T \to \infty$. Then, a feasible testing procedure exploits the consistent estimator $\tilde{\Sigma}_\xi = 2\frac{1}{N}\sum_{i,j} \frac{\tau_{i,T}^2 \tau_{j,T}^2}{\tau_{ij,T}^2} \times tr[(C_{\widehat{\nu}}^\top \widehat{Q}_{x,i}^{-1} \tilde{S}_{ij} \widehat{Q}_{x,j}^{-1} C_{\widehat{\nu}}) \widehat{\omega}_j (C_{\widehat{\nu}}^\top \widehat{Q}_{x,j}^{-1} \tilde{S}_{ji} \widehat{Q}_{x,i}^{-1} C_{\widehat{\nu}}) \widehat{\omega}_i]$ of the asymptotic variance Σ_ξ.

Finally, we can derive a test for $H_0 : \alpha^\top = (\alpha_1, \cdots, \alpha_N) = 0$, which equals to $H_0 : b_{1,i} = 0$ for almost all i against the alternative hypothesis $H_1 : E[b_{1,i}^\top b_{1,i}] > 0$. We only have to substitute $\widehat{Q}_a = \frac{1}{N}\sum_i \widehat{b}_{1,i}^\top \widehat{w}_i \widehat{b}_{1,i}$ for $\widehat{Q}_e$, and $E_1 = (I_{d_1} : 0_{d_1 \times d_2})^\top$ for $C_{\widehat{\nu}}$.

Appendix B.4. Asymptotic properties of risk premium

Define $\tau_{ij,T} = T/T_{ij}$, where $T_{ij} = \sum_t I_{ij,t}$, $I_{ij,t} = I_{i,t}I_{j,t}$ for $i,j = 1,...,N$, $\tau_{ij} = \text{plim}_{T\to\infty}\tau_{ij,T} = E[I_{ij,t}]^{-1}$. Let $Q_{b_3} = E[b_{3,i}^\top \omega_i b_{3,i}]$, $Q_z = E[Z_t^\top Z_t]$, $S_{ij} = \text{plim}_{T\to\infty}\frac{1}{T}\sum_t \sigma_{ij,t}x_{i,t}x_{j,t}^\top = E[\epsilon_{i,t}\epsilon_{j,t}x_{i,t}x_{j,t}^\top]$, $S_{Q,ij} = Q_{x,i}^{-1}S_{ij}Q_{x,j}^{-1}$, $\nu_{3,i} = \text{vec}[b_{3,i}^\top\omega_i]$, and $S_{\nu_3} = \text{a.s.}-\lim_{N\to\infty}\frac{1}{N}\sum_{i,j}\frac{\tau_i\tau_j}{\tau_{ij}}[S_{Q,ij}\otimes \nu_{3,i}\nu_{3,j}^\top]$.

Under reasonable assumptions, Gagliardini et al. (2016) proved that $\sqrt{NT}(\widehat{\nu}-\nu-\frac{1}{T}\widehat{B}_\nu) \Rightarrow N(0,\Sigma_\nu)$, where $\widehat{B}_\nu = \widehat{Q}_{b_3}^{-1}J_b\frac{1}{N}\sum_i \tau_{i,T}\text{vec}[E_2^\top \widehat{Q}_{x,i}^{-1}\widehat{S}_{ii}\widehat{Q}_{x,i}^{-1}C_{\widehat{\nu}_1}\widehat{\omega_i}]$ and $\Sigma_\nu = (\text{vec}[C_\nu^\top]^\top \otimes Q_{b_3}^{-1})S_{\nu_3}(\text{vec}[C_\nu^\top]^\top \otimes Q_{b_3}^{-1})$, with $J_b = (\text{vec}[I_{d_1}]^\top \otimes I_{Kp})(I_{d_1}\otimes J_a)$; $\sqrt{T}\text{vec}[\widehat{\Lambda}^\top - \Lambda^\top] \Rightarrow N(0,\Sigma_\Lambda)$, where $\Sigma_\Lambda = (I_K \otimes Q_z^\top)\Sigma_u(I_K\otimes Q_z^\top)$, $\Sigma_u = E[u_tu_t^\top \otimes Z_{t-1}Z_{t-1}^\top]$ and $u_t = f_t - DZ_{t-1}$, when $N,T\to\infty$ such that $N = O(T^{\bar{\gamma}})$ for $0<\bar{\gamma}<3$.

Since $\lambda_t = \Lambda Z_{t-1} = (Z_{t-1}^\top \otimes I_K)W_{pK}\text{vec}[\Lambda^\top]$, we have $\sqrt{T}(\widehat{\lambda}_t - \lambda_t) \Rightarrow N(0,(Z_{t-1}^\top Q_z^{-1}\otimes I_K)E[Z_{t-1}Z_{t-1}^\top \otimes u_tu_t^\top](Z_{t-1}Q_z^{-1}\otimes I_K))$. Replace the unknown quantities $Q_{x,i}$, Q_z, Q_{b_3} and ν by their empirical counterparts $\widehat{Q}_{x,i}$, $\widehat{Q}_z$, $\widehat{Q}_{b_3}$ and $\widehat{\nu}$, and Σ_u is estimated by a standard HAC estimator $\widehat{\Sigma}_u$. Hence, the construction of confidence of $\widehat{\Lambda}$ is straightforward.

The construction of confidence interval for the components of ν is more difficult. The variance-covariance matrix Σ_ν through S_{ν_3} involves a limiting double sum over S_{ij} scaled by N and not N^2. To handle this, we rely on consistent estimation of large-dimensional sparse covariance matrices by hard thresholding. Let us introduce the thresholded estimator $\tilde{S}_{ij} = \widehat{S}_{ij}I\left\{\|\widehat{S}_{ij}\| \geq \kappa\right\}$ of S_{ij}, which we refer to as $\widehat{S}_{ij}$ threshold at $\kappa = \kappa_{N,T}$. We can drive an asymptotically valid confidence interval for the components of ν based on the $\tilde{\Sigma}_\nu = (\text{vec}[C_{\widehat{\nu}}^\top]^\top \otimes \widehat{Q}_{b_3}^{-1})\tilde{S}_{\nu_3}(\text{vec}[C_{\widehat{\nu}}^\top]^\top \otimes \widehat{Q}_{b_3}^{-1})$, with $\tilde{S}_{\nu_3} = \frac{1}{N}\sum_{i,j}\frac{\tau_i\tau_j}{\tau_{ij}}[\tilde{S}_{Q,ij}\otimes\widehat{\nu}_{3,i}\widehat{\nu}_{3,j}^\top]$, and $\tilde{S}_{Q,ij} = \widehat{Q}_{x,i}^{-1}\tilde{S}_{ij}\widehat{Q}_{x,i}^{-1}$. Then we have $\tilde{\Sigma}_\nu^{-1/2}\sqrt{NT}(\widehat{\nu}-\nu-\frac{1}{T}\widehat{B}_\nu) \Rightarrow N(0,I_K)$.