國立臺灣大學管理學院財務金融學系

碩士論文

Department of Finance

College of Management

National Taiwan University

Master's Thesis

控制 FDR 和潛在因子下的共同基金表現

Examining Mutual Fund Performance with FDR-Control and Latent Factors

張育豪

Yu-Hao Chang

指導教授: 管中閔 博士

Advisor: Chung-Ming Kuan Ph.D.

中華民國 113 年 11 月

November, 2024

doi:10.6342/NTU202404642

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

日本共同基金表現實證

Japanese Mutual Fund Performance

本論文係張育豪君（R11723033）在國立臺灣大學財務金融學系暨研究所完成之碩士學位論文，於民國113年11月19日承下列口試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Finance on 19th November 2024 have examined a Master's Thesis entitled above presented by Chang, Yu-Hao (R11723033) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

管中閔　　莊惠菁　　許秉瑜

（指導教授 Advisor）

財金系（所）主任 Director: 石百達

Acknowledgements

I have received much help and support along the way in completing this thesis. First and foremost, I am deeply grateful to my advisor, Professor Chung-Ming Kuan, for his immense support and guidance during my master's thesis journey. From him, I not only gained the econometric knowledge and research skills learned from his life philosophy, which has profoundly influenced me. He has been the most impactful figure during my time in graduate school.

I would also like to thanks to Professors Hui-Ching Chuang, Dr. Yu-Chin Hsu, Professor Larry Y. Tzeng, and Professor Yan-Shing Chen for their invaluable assistance in addressing various questions related to my thesis. Additionally, I extend my deepest gratitude to my family for their steadfast support, which enabled me to devote myself wholeheartedly to completing this thesis.

Lastly, I I would like to thank to my friend Terasa for her insightful feedback on my thesis, and to my friends Ivy, Hannah, Bilis, Mavis, Peggy, Jacky, and many others for listening to my thoughts and enriching my graduate school experience.

I am sincerely grateful for all the support and encouragement that have made my master's studies a fulfilling and joyful experience.

doi:10.6342/NTU202404642

摘要

本文在考慮潛在因子與多重檢定的問題下檢驗日本共同基金的績效。我們主要遵循 Giglio, Liao, and Xiu (2021) 的方法來辨識潛在的定價因子，處理基金報酬資料中的缺失值問題，且運用 screening Benjamini and Hochberg procedure 控制偽發現率（false discovery rate, FDR）。結果顯示，在這些基金中僅有 0.47% 在長期內被辨識為具有顯著績效。然而，我們也發現在短期內有較高比例的基金展現出顯著的績效，而這些基金在不同子期間的表現持續優於其他基金，但並未成為長期具有顯著績效的基金。最後，我們在不同的 FDR 水準下構建了由顯著績效基金組成的投資組合，這些投資組合的樣本外績效均優於日經 225，顯示出這些基金在樣本內的表現能成功轉化為樣本外收益，並帶來顯著的經濟價值。

關鍵字：共同基金表現、多重檢定問題、偽發現率、主成分分析、矩陣完備化、自助重抽法

doi:10.6342/NTU202404642

Abstract

This paper examines the performance of Japanese mutual funds while addressing latent factors and the issue of multiple testing. We follow the methodology of Giglio, Liao, and Xiu (2021) to identify latent factors, handle missing values, and apply the screening Benjamini and Hochberg procedure to control the false discovery rate (FDR). Among these funds, only 0.47% are identified as outperforming funds. However, a greater proportion of mutual funds demonstrate superior performance in the short term, which continues to outperform others across different subperiods, though their performance does not sustain over the long term. Finally, we construct portfolios of outperforming funds controlled at varying FDR levels, all of which outperform the Nikkei 225 out-of-sample, indicating that these in-sample alphas successfully translate to out-of-sample returns and generate significant economic values.

Keywords: Mutual funds performance, Multiple testing problem, False discovery rate, Principal component analysis, Matrix completion, Bootstrap

doi:10.6342/NTU202404642

Contents

doi:10.6342/NTU202404642

doi:10.6342/NTU202404642

List of Figures

doi:10.6342/NTU202404642

List of Tables

doi:10.6342/NTU202404642

Chapter 1 Introduction

Compared to passive mutual funds, active mutual funds may potentially generate higher returns if it can outperform the index. Therefore, identifying outperforming mutual funds (i.e., funds with positive alphas) is a key concern for investors. Although there is a large number of literature evaluating mutual funds performance, most studies focus on US mutual funds (e.g., Kosowski et al., 2006; Fama and French, 2010; Barras et al., 2010; Cuthbertson et al., 2010). In contrast, relatively few studies examine Japanese mutual funds, and those that do (e.g., Cai et al., 1997; Pilbeam and Preston, 2019) primarily analyze early-period performance. This motivates us to investigate the performance of Japanese active mutual funds by updating the sample period and using a novel method to identify outperforming mutual funds.

The most common approach to identify outperforming mutual funds is to estimate their alphas using a benchmark model, followed by conducting many individual hypothesis tests to infer whether their true alphas are positive. However, due to the multiple testing problem, conducting these tests simultaneously can lead to many funds showing significant alphas even when their true alphas are non-positive. Furthermore, the complex dependency structure among these mutual funds may cause the pre-specified benchmark model to fail to capture all common risk factors (Fama and French, 2010), resulting in the omitted variable problem and potentially biasing the alpha estimator.

doi:10.6342/NTU202404642

To address the omitted variable problem, we adopt the methodology of Giglio, Liao, and Xiu (2021), considering a benchmark model with latent factors and applying the principal component analysis (PCA) proposed by Giglio and Xiu (2017) to identify these unobservable common factors. However, in practice, mutual fund data often include missing values, making it difficult to apply PCA directly. Therefore, we follow Giglio, Liao, and Xiu (2021) and employ matrix completion (Ma et al., 2011; Cai et al., 2010; Goldfarb and Ma, 2011) to impute missing values by approximating the observed data using a low-rank matrix.

Once alphas are estimated, we then evaluate their p-values for hypothesis testing. Although Giglio, Liao, and Xiu (2021) propose an asymptotically normal test statistic, its finite sample performance may be affected when the data have missing values. Therefore, we use wild bootstrap to evaluate the corresponding p-values. Unlike Giglio, Liao, and Xiu (2021), we bootstrap the test statistic rather than alpha since the test statistic is standardized by the standard deviation and the number of observations, providing better statistical properties (Kosowski et al., 2006).

Finally, to address the multiple testing problem, one may apply the Bonferroni method to control the family-wise error rate (FWER), which is defined as the probability of making at least one false rejection. However, the Bonferroni method becomes overly conservative as the number of tests increases. To improve test power, Holm (1976) proposes a stepwise procedure, while White (2000), Hansen (2005), Romano and Wolf (2005), and Hsu et al. (2010) develop bootstrap techniques that account for dependencies among tests. Nonetheless, methods that control the number of rejections tend to lack power when faced with thousands of hypothesis tests. As a result, some researchers prefer to control the false discovery rate (FDR), which is the expected proportion of false discoveries among all re-

doi:10.6342/NTU202404642

jections. The well-known Benjamini and Hochberg (1995; BH) procedure is commonly used to control the FDR. However, the BH procedure still lacks power when the number of tests is large. Therefore, in this paper, we follow the screening criterion proposed by Giglio, Liao, and Xiu (2021) to enhance the power of the BH procedure by filtering out funds with extremely negative alphas.

In summary, this paper examines the performance of 1483 Japanese mutual funds over the period from 2002 to 2023 using a benchmark model with latent factors and a screening BH procedure to control the luck. We find that, while only 0.47% of mutual funds outperform the benchmark when controlling the FDR at 10%, more mutual funds demonstrate superior performance in the short term, especially during the 2010−2019 and 2012–2021 subperiods, with 3.14% and 9.30% of mutual funds , respectively. However, these short-term superior performance tend to vanish quickly. We further examine their rank persistence and find that funds with short-term superior performance consistently outperform others but lack the strength to maintain their superior performance over the long term. Lastly, we investigate whether these in-sample positive alphas can translate into out-of-sample economic values by forming portfolios that control the FDR at levels of 10%, 15%, and 20%, and evaluating their cumulative wealth from 2012−2023. Our results demonstrate that all these portfolios exhibit superior performance compared to Nikkei 225, indicating these funds with in-sample positive alpha have the ability to beat the benchmark in out-of-sample performance.

The rest of this paper is organized as follows: Section 2 provides a literature review on the multiple testing problem and the bootstrap methods used to evaluate p-values in mutual fund data. Section 3 outlines the methodology employed in this study. Section 4 presents the empirical findings, while Section 5 offers the conclusions.

doi:10.6342/NTU202404642

Chapter 2 Literature Review

2.1 Multiple Testing Problem

Multiple hypothesis testing refers to situations where more than one null hypothesis is tested simultaneously. This often occurs in financial empirical studies. For example, one might be interested in identifying superior trading rules (or outperforming mutual funds) from thousands of technical strategies (or mutual funds). When there are multiple hypotheses under consideration, the multiple testing problem arises if each hypothesis is tested without properly controlling the type I error. For example, suppose we are testing 100 hypotheses and their test statistics are independent. With a significance level of 5% for each test, the probability of rejecting at least one true null hypothesis is $1 - 0.95^{100} = 99.4\%$, which is much larger than the pre-specified individual significance level. In this case, it is necessary to apply proper methods to avoid false rejections.

Consider a multiple hypothesis testing situation when there are M hypotheses. Table 2.1 illustrates the possible outcomes when testing these M hypotheses simultaneously. Suppose m_o of M hypotheses are true under the null, and M-m_o are true alternatives. In these M hypotheses, R of them have been rejected. Among these R rejected hypotheses, *FP* of them are falsely rejected (also called false rejections or false discoveries). Conversely, *FN* are true alternatives that have not been rejected (also called false negatives).

doi:10.6342/NTU202404642

Table 2.1: Outcome of hypothesis testing

	H_0 is true	H_1 is true	Total
Reject H_0	*FP*	*TP*	R
Not reject H_0	*TN*	*FN*	M-R
	m_o	M-m_o	M

To address the multiple testing problem, researchers aim to control the overall Type I error across a family of hypothesis tests. Suppose the parameter of interest is α_k for $k = 1, \cdots, M$, and the objective is to test whether $\alpha_k \leq 0$. One approach is to conduct a joint hypothesis test:

$$H : \alpha_k \leq 0 \ \ \forall\ k = 1, \cdots, M \ \ \text{vs.} \ \ H'_k : \exists\ k \text{ with } \alpha_k > 0, \tag{2.1}$$

which directly controls the overall Type I error across these M tests. However, joint hypothesis testing only determines whether at least one α_k rejects the null hypothesis, while researchers are often interested in identifying which specific hypotheses reject the null. Therefore, an alternative approach is to conduct M individual hypothesis tests (i.e., multiple tests):

$$H_k : \alpha_k \leq 0 \ \ \text{vs.} \ \ H'_k : \alpha_k > 0 \ \ \text{for } k = 1, \cdots, M, \tag{2.2}$$

using appropriate critical values for each individual test to control specific error measures across these M tests and determine which hypotheses reject the null. For example, one could control the probability of $FP > 0$ to avoid any false rejections among these M hypotheses (i.e., FWER). Alternatively, one might focus on controlling the proportion of false rejections relative to the total number of rejections (FP/R), such as the false discovery proportion (FDP) or the expectation of FDP, referred to as the false discovery rate (FDR). In the following sections, we will provide more details on these error measures

doi:10.6342/NTU202404642

and the methods for determining the appropriate critical values for both the joint test and multiple tests.

2.2 Family-wise Error Rate (FWER)

The most classical method used to deal with the multiple testing problem is to control the FWER. The FWER is defined as the probability of falsely rejecting at least one true null hypothesis:

$$\text{FWER} = \mathbb{P}\{FP \geq 1\}. \tag{2.3}$$

Once the FWER is controlled, the probability of Type I error does not increase when the number of hypotheses increases. One well-known method used to control the FWER is the Bonferroni method. To maintain the FWER at the level of γ, Bonferroni suggests setting the individual significance level at γ/M. This procedure is justified by the following inequality:

$$\text{FWER} = \mathbb{P}\{\bigcup_{k\in\mathcal{I}_0} \left(\text{Reject } H_0^k\right)\} \leq \sum_{k\in\mathcal{I}_0} \mathbb{P}\left(\text{Reject } H_0^k\right) \leq \sum_{k\in\mathcal{I}_0} \frac{\gamma}{M} \leq \gamma, \tag{2.4}$$

where $\mathcal{I}_0$ is the set of indices of the true null hypotheses and H_0^k is the k-th null hypothesis. Although the Bonferroni method can be used to control for the FWER, it becomes too conservative when M is large (i.e. γ/M is too small), resulting in few rejections.

To enhance the power of the Bonferroni method, Holm introduces a stepwise procedure. In the Holm method, the p-values of M statistics are initially ordered as $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(M)}$. To control for the FWER at the level of γ, the Holm method begins by

doi:10.6342/NTU202404642

testing the most significant hypothesis (i.e. $k = 1$) and rejecting the null hypothesis if

$$p_{(k)} < \frac{\gamma}{(M - k + 1)} \quad \text{for } k = 1, 2 \dots M. \tag{2.5}$$

If the null hypothesis corresponding to $p_{(k)}$ is rejected, the procedure continues to test subsequent hypotheses until a null hypothesis cannot be rejected. Notably, while testing the most significant hypothesis, the threshold of the Holm method is identical to that of the Bonferroni method. However, as it moves to less significant hypotheses, the threshold of the Holm method decreases. Therefore, the Holm method typically rejects more hypotheses than the Bonferroni method while controlling the same FWER, making it more powerful.

Note that although the Holm method improves the power of the Bonferroni method, it still lacks of power if M is large. Moreover, one drawback of both the Bonferroni and the Holm methods is that they do not consider the dependence structure of the test statistics (and hence the p-values), leading to overly stringent thresholds. For instance, if there is a dependence structure causing all p-values to be the same, the Bonferroni threshold should be adjusted from γ/M to γ. In practice, test statistics for multiple hypotheses are usually dependent, which diminishes the power of the Bonferroni and the Holm methods.

2.3 Joint test

To address the dependency among these hypotheses, White (2000) proposes the reality check to examine whether superior strategy exists. Let f_k for $k = 1, \cdots, M$ denote the performance measure of the strategy k compared with the benchmark. The null hypothesis

doi:10.6342/NTU202404642

that no superior strategy exists among these M strategies is:

$$H_{0k} : \mathbb{E}(f_k) \leq 0 \quad \text{for } k = 1, \ldots, M. \tag{2.6}$$

To test these hypotheses, White (2000) uses the concept of the least favorable configuration to the alternative and enforces the null hypothesis as $\mathbb{E}(f_k) = 0$ for all k. The performance of the best strategy can then be written as the maximum value of the normalized sample average of $f_{t,k}$:

$$\bar{V}_M \equiv \max_{k=1,\ldots,M} \sqrt{n}\bar{f}_k, \tag{2.7}$$

where $f_{t,k}$ is the return of k-th strategy at time t, and $\bar{f}_k = \sum_{t=1}^{n} f_{t,k}/n$ is its sample average. If the test statistic $\bar{V}_M$ is larger than the critical value, it implies that at least one superior strategy exists.

White (2000) suggests employing the stationary bootstrap proposed by Politis and Romano (1994) to determine the critical value of $\bar{V}_M$. Let $f_k^*(b)$ be the b-th bootstrap sample for f_k, with its sample average defined as $\bar{f}_k^*(b) = \sum_{t=1}^{n} f_{t,k}^*(b)/n$. Then, the empirical distribution of $\bar{V}_M^*$ is constructed by:

$$\bar{V}_M^*(b) \equiv \max_{k=1,\ldots,M} \sqrt{n}(\bar{f}_k^*(b) - \bar{f}_k) \quad \text{for } b = 1, \ldots, B. \tag{2.8}$$

Finally, we can determine the critical value of $\bar{V}_M$ by evaluating the percentile value of the empirical distribution $\bar{V}_M^*$ and infer whether a superior strategy exists.

It is important to note that White's reality check utilizes the stationary bootstrap to preserve the dependency structure of individual statistics. However, Hansen (2005) points out two drawbacks of this method. Firstly, the test statistic $\sqrt{n}\bar{f}_k$ is not studentized. Sec-

doi:10.6342/NTU202404642

ondly, White (2000) uses the least favorable configuration to the alternative, making the reality check conservative and susceptible to the inclusion of poor or irrelevant strategies. Therefore, Hansen (2005) introduces a new method for superior predictive ability (SPA) by studentizing the test statistic:

$$\bar{V}_M \equiv \max\left(\max_{k=1,\ldots,M} \frac{\sqrt{n}\bar{f}_k}{\hat{\sigma}_k}, 0\right), \tag{2.9}$$

where $\hat{\sigma}_k$ is the consistent estimator of the standard deviation of $\sqrt{n}\bar{f}_k$. The studentized statistic generally results in better power performance than the non-studentized one.

Furthermore, to address the problem arising from the least favorable configuration, Hansen (2005) suggests generating the empirical distribution $\tilde{V}_M^*$ by:

$$\tilde{V}_M^*(b) \equiv \max\left(\max_{k=1,\ldots,M} \frac{\sqrt{n}\bar{Z}_k^*(b)}{\hat{\sigma}_k}, 0\right) \quad \text{for } b = 1, \ldots, B, \tag{2.10}$$

where $\bar{Z}_k^*(b)$ is the sample average of the adjusted bootstrap performance measure $Z_{t,k}^*(b)$, and $Z_{t,k}^*(b)$ is defined as:

$$Z_{t,k}^*(b) \equiv \bar{f}_{t,k}^*(b) - \bar{f}_k \mathbb{1}_{\{\bar{f}_k \geq -\sqrt{(\hat{\sigma}_k^2/n)2\text{loglog}n}\}}. \tag{2.11}$$

In Hansen's bootstrap procedure, the bootstrap performance $\bar{f}_{t,k}^*(b)$ is not centered by the sample average performance $\bar{f}_k$ if the sample average is much too low, specifically less than less than $-\sqrt{(\hat{\sigma}_k^2/n)2\text{loglog}n}$. Therefore, the SPA test is not susceptible to the inclusion of poor or irrelevant strategies, making it more powerful than White's reality check.

The methods mentioned above are widely used in economics and finance. Sullivan et al. (1999) apply White's reality check and find that superior trading rules exist for the DJIA from 1897 to 1986, but the best strategy does not maintain its performance from

doi:10.6342/NTU202404642

1987 to 1996. Hsu and Kuan (2005) employ White's reality check and the SPA test with more comprehensive trading rules. They discover these rules outperform on NASDAQ but do not outperform on S&P 500 and DJIA. Qi and Wu (2006) examine trading rules across seven exchange rates and find that superior trading rules exist in five markets. Chen, Huang and Lai (2011) and Metghalchi et al. (2012) find that the 14-day MFI strategy outperform in Taiwan stock market.

2.4 k-FWER

In fact, researchers are not only concerned about false rejections but also about the ability to reject the null hypothesis when the null is false (i.e., the power of the test). When the number of hypotheses M is large, the FWER criterion becomes too stringent, making it difficult to reject hypotheses when they are false. Therefore, researchers relax the control of the FWER and propose the k-FWER, which allows for more false rejections. The k-FWER is defined as the probability of rejecting at least k true null hypotheses:

$$k\text{-FWER} = \mathbb{P}\{FP \geq k\}. \tag{2.12}$$

By allowing up to $k-1$ false rejections, the k-FWER is more powerful than the FWER and can reject more hypotheses.

Moreover, in many financial applications, such as identifying outperforming mutual funds and superior trading rules, researchers are not only interested in determining whether the best mutual fund (or trading rule) beats the benchmark but also in identifying all mutual funds (or trading rules) with superior performance. To address this problem, researchers use the stepwise procedure. The stepwise procedure is similar to the Holm procedure.

doi:10.6342/NTU202404642

It begins with a single-step approach and then continues by adjusting the critical values based on the remaining hypotheses in subsequent steps. This sequential adjustment allows the stepwise procedure to reject more hypotheses than the single-step approach while controlling the same error rate, thereby enhancing the power of the test.

Romano and Wolf (2007) propose the k-StepM, which modifies White's reality check by controlling the k-FWER and applying the stepwise procedure. With these modifications, the k-StepM can identify as many superior strategies as possible. The summary of k-StepM is as follows. Let $\mathcal{R}$ be a real number set and $|\mathcal{R}|$ denote the number of elements in set $\mathcal{R}$. For any subset $k \subseteq \{1, ..., M\}$, $\hat{c}_k(\gamma, k)$ is the γ-th quantile of k-max$\{\psi_j^b \mid j \in k\}$, where k-max$\{\mathcal{R}\}$ is the k-th largest value in $\mathcal{R}$ and $\{\psi_j^b \mid j \in k\}$ are the simulated distributions that can be constructed by the bootstrap procedure. The bootstrap procedure is similar to the Equation (2.8) (further details can be found in White (2000) and Romano and Wolf (2007)). The algorithm of the k-StepM is as follows:

1. Let $B_1 = \{1, ..., M\}$ be the initial set of hypotheses. For each hypothesis H_0^s with $s \in B_1$, reject H_0^s if

$$\sqrt{n}\hat{T}_s \geq \max\{\hat{c}_{B_1}(\gamma, k), 0\}, \tag{2.13}$$

   where $\hat{T}_s = \frac{\hat{f}_s}{\hat{\sigma}_s}$ is the studentized test statistic. Then, let R_1 be the set of indices of the rejected hypotheses. If $|R_1| \leq k$, stop the algorithm; otherwise, proceed to the next step.

2. Let B_2 be the set of the indices of the hypotheses that are not rejected in previous step, i.e., $B_2 = B_1 \backslash R_1$. For H_0^s with $s \in B_2$, reject H_0^s if

$$\sqrt{n}\hat{T}_s \geq \max_{I \subset R_1,\ |I| = k-1} \{\hat{c}_{B_2 \cup I}(\gamma, k), 0\}. \tag{2.14}$$

doi:10.6342/NTU202404642

3. Repeat step 2 by substituting R_1 and B_2 with R_{j-1} and B_j for $j \geq 3$ until there are no further rejections.

It is important to note that the k-StepM reassesses $k-1$ strategies that have already been rejected when evaluating the critical value at each step. This is because the k-StepM allows for up to $k-1$ false rejections. Since we cannot determine which rejected hypotheses are true alternatives, it is necessary to reconsider all possible subsets of rejected hypotheses to ensure that this method controls the k-FWER at every step.

Although the k-StepM enhances the power of White's reality check by applying the stepwise procedure and controlling the k-FWER, it still suffers from the drawback of the least favorable configuration. Therefore, Hsu, Kuan, and Yen (2014) propose the k-StepSPA, which improves the k-StepM by incorporating ideas from the SPA test. For each test s, define $\hat{u}_s$ as

$$\hat{u}_s = \hat{T}_s \mathbb{1}_{\sqrt{n}\hat{T}_s \leq -a_k} \text{ , where } a_k = \sqrt{2\text{loglog}n}. \tag{2.15}$$

Note that $\hat{u}_s$ is used to re-center the simulated distribution if the sample average performance $\sqrt{n}\hat{T}_s$ is much too low, specifically less than $-\sqrt{2\text{loglog}n}$. Consequently, the simulated distribution becomes $k\text{-max}\{\psi_j^b + \sqrt{n}\hat{u}_j \mid j \in K\}$. The simulation results of Hsu, Kuan, and Yen (2014) show that the k-StepSPA has greater power than the k-StepM.

2.5 False Discovery Rate (FDR)

Although k-FWER relaxes the constraint of the FWER by allowing up to $k-1$ false rejections, when there are thousands of tests and lots of them are true alternatives, limiting

doi:10.6342/NTU202404642

the number of false rejections to a fixed k may still be too stringent. Moreover, for some multiple hypothesis testing problems, the goal is to identify as many true alternatives as possible, and false discoveries are relatively less harmful. For example, in the mutual fund performance problem, the primary goal is to find outperforming mutual funds. Even if there are some false discoveries, they are usually zero-alpha funds, which do not cause significant losses to investors. To address such problems, researchers find it more appealing to control the FDR. The FDR is defined as the expected value of the FDP:

$$\text{FDR} \equiv \mathbb{E}(\text{FDP}) = \mathbb{E}(\frac{FP}{R} \mid R > 0)\mathbb{P}(R > 0). \tag{2.16}$$

Since the FDR is used to control the rate of false discoveries, it is more tolerant of false discoveries. Therefore, it can identify more true alternatives compared to the FWER.

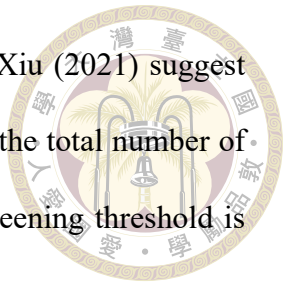
Benjamini and Hochberg (1995; BH) propose the first FDR-controlling method that utilizes the stepwise procedure. Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(M)}$ denote the ordered sequence of p-values corresponding to hypotheses $H_{(1)}$, $H_{(2)}$..., $H_{(M)}$. To control the FDR at the level of γ, the BH procedure begins by testing the least significant hypothesis (i.e. $k = M$) with the inequality:

$$p_{(k)} \leq \frac{k \times \gamma}{M}. \tag{2.17}$$

If $p_{(k)}$ does not satisfy the inequality, the procedure continues to test subsequent hypotheses until the first k^* satisfies the inequality. Finally, we can reject the hypotheses $H_{(i)}$ for all $i \leq k^*$.

In Equation (2.17), the critical value of the BH procedure depends on the number of hypotheses M, leading to a potential lack of power if many hypotheses have extremely high p-values (i.e., M becomes larger, but the order of p-values likely to be rejected

doi:10.6342/NTU202404642



remains almost the same). To address this issue, Giglio, Liao, and Xiu (2021) suggest screening the hypotheses that are obviously true nulls first to reduce the total number of hypotheses, thereby improving the power of the procedure. The screening threshold is defined as follows:

$$t_i > -c \cdot \log(\log T)\sqrt{\log M} \quad \text{for } i \leq M, \tag{2.18}$$

where t_i is the test statistic for i-th hypothesis, M is the number of tests, T is the time dimension, and c is a constant. Unlike the threshold in Giglio, Liao, and Xiu (2021), we add a constant c since we find they do so while implementing their screening method. Although c does not affect this method's asymptotic properties, it can improve the empirical results, and they set $c = 1/3$ in their implementation (see the Python code provided by Giglio, Liao, and Xiu, 2021).

Another way to improve the BH procedure is to estimate m_0. Note that under the assumption that the p-values corresponding to true nulls are independent, the BH procedure controls the FDR at the level γ by satisfying the following inequality (details in BH, 1995, Theorem 1):

$$\mathbb{E}\left(\frac{FP}{R}\right) \leq \frac{m_0}{M}\gamma \leq \gamma. \tag{2.19}$$

This inequality indicates that the FDR is controlled by the number of true nulls (m_0). However, the threshold in the BH procedure only uses M and does not include information about m_0, potentially reducing the procedure's power. Therefore, Storey (2002) proposes an estimator for the proportion of true nulls π_0 (i.e., m_0/M):

$$\hat{\pi}_0(\lambda) = \frac{\#\{\hat{p}_i > \lambda\}}{(1-\lambda)M}, \tag{2.20}$$

where $\lambda \in (0, 1)$ is the threshold that defines the boundary of true nulls (hypotheses with

doi:10.6342/NTU202404642

p-values greater than λ are considered true nulls). This estimator relies on two assumptions: first, p-values under true alternatives are close to zero; and second, p-values corresponding to true nulls are independent and uniformly distributed on the interval $[0, 1]$. Under these assumptions, approximately $m_0(1-\lambda)$ p-values lie in the interval $(\lambda, 1]$ if λ is sufficiently large. Storey (2002) suggests using $\#\{\hat{p}_i > \lambda\}$ to estimate $m_0(1-\lambda)$ and hence estimate m_0 by $\#\{\hat{p}_i > \lambda\}/(1-\lambda)$. This leads to the estimator of $\hat{\pi}_0(\lambda)$. It is important to note that the null hypothesis in Storey's method must be an equality; otherwise, the true nulls do not follow a uniform distribution.

To select λ that satisfies the assumptions mentioned above, we can plot the p-values on a histogram (see Figure 2.1) and choose λ for which the histogram of p-values becomes flat (i.e., satisfies uniform distribution). An alternative approach for choosing λ is to minimize the estimated mean square error (MSE) of $\hat{\pi}_0$ by using the bootstrap procedure (Storey, 2002). In practice, these two methods result in similar values of λ.

Once we have $\hat{\pi}_0$, we can estimate the FDR with the pre-specified threshold of the rejection region κ by using the following estimator:

$$\widehat{\text{FDR}}_\lambda(\kappa) = \frac{\hat{\pi}_0(\lambda) \cdot \kappa \cdot M}{\#\{\hat{p}_i \leq \kappa\}}. \tag{2.21}$$

Here, $\hat{\pi}_0(\lambda) \cdot \kappa \cdot M$ is used to estimate the number of false rejections by assuming true nulls are uniformly distributed, and $\#\{\hat{p}_i \leq \kappa\}$ is used to estimate the number of significant funds. Therefore, we obtain the estimator of the FDR and can select κ with an acceptable $\widehat{\text{FDR}}$.

To compare the power of Storey's method with the BH procedure, we rewrite Storey's method in a form similar to the BH procedure. Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(M)}$ be the ordered

doi:10.6342/NTU202404642

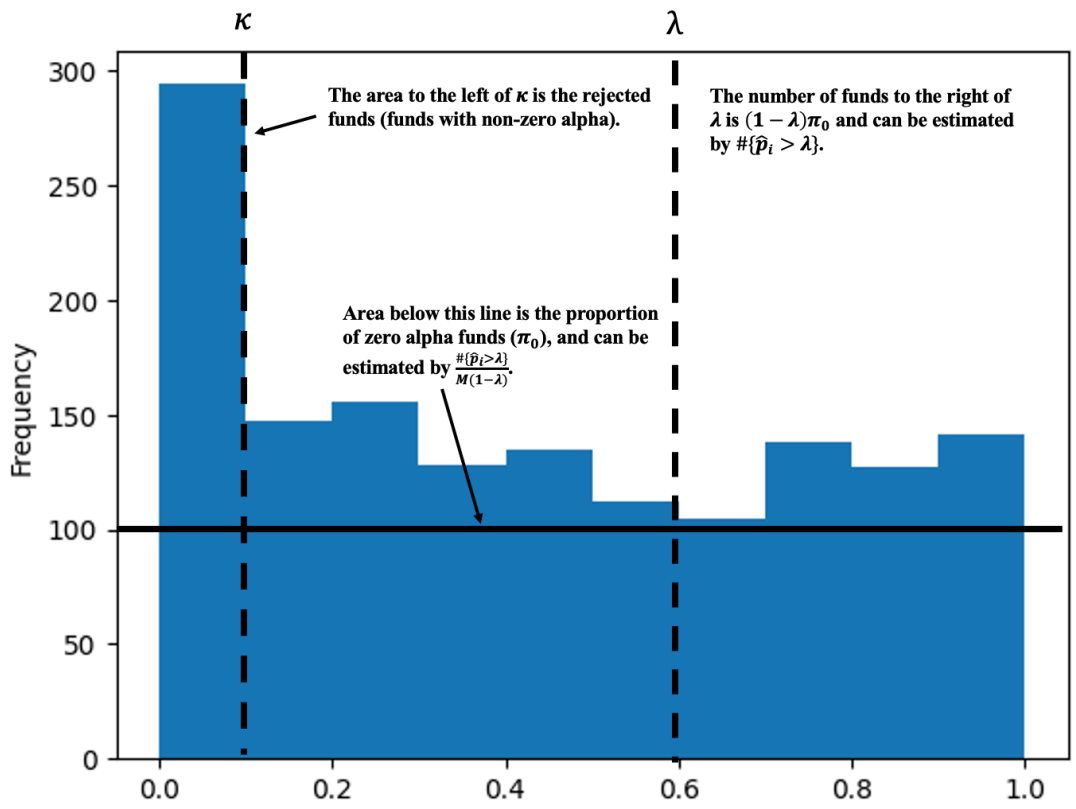


Figure 2.1: Histogram of mutual funds p-values. This figure uses 1483 Japanese mutual funds during $2002-2023$ and shows the histogram of p-values of the t-statistic of alpha from the Carhart's four-factor model.

sequence of p-values. Storey's method aims to find the largest rejected index $\hat{\ell}$ while controlling the FDR at the significance level γ:

$$\hat{\ell} = \max\{\ell : \widehat{\text{FDR}}(p_{(\ell)}) \leq \gamma\}, \tag{2.22}$$

where $\widehat{\text{FDR}}(p_{(\ell)}) = \hat{\pi}_0(\lambda) \cdot p_{(\ell)} \cdot M/\ell$. Therefore, Equation (2.22) can be written as:

$$\hat{\ell} = \max\{\ell : p_{(\ell)} \leq \frac{\ell}{M}\gamma \cdot \frac{1}{\hat{\pi}_0}\}. \tag{2.23}$$

Comparing this equation to the Equation (2.17), we have that $\hat{\ell} \geq \hat{k}$ since $1/\hat{\pi}_0 \geq 1$. This indicates that the threshold of the rejection region in Storey's method is larger than the one in the BH procedure. Therefore, Storey's method is more powerful than the BH procedure.

doi:10.6342/NTU202404642

Notably, unlike other approaches, Storey (2002) first fixes the rejection region and then estimates the FDR. Although this method may seem counterintuitive, it allows for leveraging prior knowledge to choose the rejection region appropriately, leading to better testing results.

FDR-based methods have been widely used across various fields of research. In medicine, Benussi et al. (2020) utilize the BH procedure to investigate the relationship between COVID-19 and neurologic diseases. They conclude that patients with both COVID-19 and neurological conditions have significantly higher in-hospital mortality compared to those without COVID-19. In ecology, Betts et al. (2017) study the threat to forest-exclusive species while controlling the FDR by applying the BH procedure. They find that the threat is associated with the interaction between forest loss and the proportions of initial forest cover. In finance, Barras et al. (2010) use Storey's method and find that only 0.6% of U.S. funds outperform the benchmark. Moreover, they observe a rapid decline in the proportion of funds with true positive alphas, which dropped from 14.4% in 1993 to 0.6% in 2006, while the proportion of funds with true negative alphas increased from 9.2% to 24.0% over the same period. Cuthbertson et al. (2012) also apply Storey's method to analyze UK equity mutual funds, finding that only 3.7% of mutual funds outperform the benchmark. Furthermore, they show that these outperformances are not persistent, while poor performances are persistent.

2.6 Bootstrap Approach

Another series of studies, including those by Kosowski et al. (2006; KTWW) and Fama and French (2010; FF), investigate whether outperforming mutual funds exist and

doi:10.6342/NTU202404642

address the multiple testing problem by employing different bootstrap approaches. The bootstrap approach is applied because KTWW find that the finite sample distribution of the t-statistics for individual fund alphas exhibits non-normality and dependence structures under the null hypothesis (where the true alpha is less than or equal to zero). Consequently, the bootstrap method is more appropriate for evaluating the p-values of the t-statistics.

Before applying the bootstrap procedure, we need to first measure the mutual fund performance. Both KTWW and FF do this by estimating alpha for each fund using the Carhart four-factor model and then recording the estimated alphas, factor loadings, and residuals. In each bootstrap iteration of KTWW's method, they independently resample the residuals for each fund and generate pseudo-time series return data of each fund by adding the resampled residuals to the original order of factor values and their corresponding factor loadings. Note that to generate the zero-alpha pseudo-time series data, KTWW set alpha equal to zero.

After generating the zero-alpha pseudo-time series data, KTWW re-estimate the Carhart four-factor model using these pseudo-time series data to evaluate the t-statistic under the null hypothesis (i.e., alpha is less than or equal to zero). They then infer the existence of outperforming mutual funds by comparing the t-statistics of alpha from the original data with those from the resampled zero-alpha data across different percentiles. If the t-statistics from the original data are large relative to the null distribution, we can conclude that some of the mutual funds outperform the benchmark. The reason for using the t-statistic of alpha rather than alpha itself is that the t-statistic is standardized by the mutual fund's standard deviation and the number of observations, providing better properties than alpha.

doi:10.6342/NTU202404642

However, since KTWW resamples each fund's residuals independently, this bootstrap procedure fails to preserve the cross-sectional dependency of residuals among mutual funds. To address this issue, FF modify KTWW's bootstrap procedure by resampling cross-sectionally. Specifically, at each bootstrap iteration, FF resample the time indexes and use the entire cross-sectional returns for that time period to construct pseudo-time series data. Similarly, FF set alpha to zero to generate zero-alpha pseudo-time series data and compare the t-statistics of alpha from the original data with those from the resampled zero-alpha data to infer the existence of outperforming mutual funds.

Since FF's bootstrap procedure resamples all mutual fund data simultaneously, it successfully preserves the dependency structure among mutual fund residuals. However, since mutual fund data often contain missing values, FF's bootstrap procedure may result in a different number of observations for each fund between the resampled data and the real-world data, potentially making the bootstrap distribution different from the real-world distribution.

Although both KTWW and FF apply their bootstrap methods to the same data, they reach opposite conclusions. KTWW find that outperforming mutual funds exist, while FF conclude that no mutual funds outperform the benchmark. To determine which method is more appropriate for real-world data, Harvey and Liu (2022) designed a simulation to test the Type I error and the power of these two bootstrap methods. The simulation results of Harvey and Liu (2022) show that KTWW's bootstrap procedure overrejects the null hypothesis because it fails to capture the dependency of residuals. In contrast, FF's bootstrap procedure lacks power due to discrepancies in the number of observations between the resampled and real-world data.

doi:10.6342/NTU202404642

To provide a more powerful method, Harvey and Liu (2022) modify FF's method using the idea of the interquartile range:

$$\widehat{band}(i) = [\widehat{Q}_{1,i} - \phi \times (\widehat{Q}_{3,i} - \widehat{Q}_{1,i}), \widehat{Q}_{3,i} + \phi \times (\widehat{Q}_{3,i} - \widehat{Q}_{1,i})] \tag{2.24}$$

where ϕ is the scale parameter that determines the width of bandwidth, $\widehat{Q}_{1,i}$ and $\widehat{Q}_{3,i}$ are the first and the third quantiles of the bootstrapped t-statistic distribution for mutual fund i, respectively. Harvey and Liu (2022) only consider the bootstrap t-statistics that fall within this bandwidth and remove the extreme values from the bootstrapped t-statistics, thereby enhancing the power of the test.

doi:10.6342/NTU202404642

Chapter 3 Methodology

Our objective is to evaluate the performance of Japanese mutual funds and identify outperforming funds. To achieve this, we begin by estimating alpha with a specified benchmark model and use alpha as the performance measure for each mutual fund. However, since these theoretical benchmark models may be stylized, they often fail to capture the full dependence structure of excess returns. This raises the problem of omitted variables (i.e., the existence of latent factors), which can potentially bias the alpha estimator.

To address this issue, we apply the asset pricing model proposed by Giglio, Liao, and Xiu (2021) to identify these latent factors. Their method utilizes the concepts of matrix completion and Principal Component Analysis (PCA) to handle the problem of missing data and omitted variables. Once these latent factors are identified, we can accurately estimate the alpha for each mutual fund.

Since some of our factors may be nontradable, we apply cross-sectional regression to estimate the risk premiums of the factors and alphas for each mutual fund. Then, to improve the performance of finite sample inference, we employ the wild bootstrap to evaluate the p-values for the t-statistics. Finally, we control the FDR and identify outperforming funds by applying the adjusted BH procedure. In the following sections, we will illustrate these methods in more detail.

doi:10.6342/NTU202404642

3.1 Mutual Fund Performance Measurement

To evaluate the performance of each mutual fund, we assume that the excess returns r_t follow the linear asset pricing model:

$$r_t = \alpha + \beta\lambda + \beta(f_t - \mathbb{E}(f_t)) + \varepsilon_t, \tag{3.1}$$

where λ is a $K \times 1$ vector denoting the risk premium of the factors, f_t is a $K \times 1$ vector denoting the factor values at time t, ε_t is a $M \times 1$ vector denoting the idiosyncratic factor, and β is a $M \times K$ matrix where the (i, j) -th element is the factor loading of the i-th mutual fund with respect to the j-th factor. Note that λ is identical to $\mathbb{E}(f_t)$ if f_t is tradable.

One well-known benchmark model for evaluating mutual fund performance is the Carhart (1997) four-factor model, defined as follows:

$$r_{i,t} = \alpha_i + \beta_{i,r_m} \cdot r_{m,t} + \beta_{i,SMB} \cdot SMB_t + \beta_{i,HML} \cdot HML_t + \beta_{i,MOM} \cdot MOM_t + \varepsilon_{i,t}, \tag{3.2}$$

where $r_{i,t}$ is the excess return of fund i at time t over the monthly risk-free rate (defined by the monthly U.S. T-bill rate). $r_{m,t}$ is the excess return of the Japan value-weighted market portfolio over the monthly risk-free rate at time t. SMB_t, HML_t, and MOM_t denote the size, book-to-market ratio, and momentum factors for Japan, respectively. All these data can be found on Kenneth French's website.

However, after applying the Carhart four-factor model, we find that the cross-sectional alphas are still correlated with each other. This suggests that the four-factor model does not capture all the dependency structures among excess returns. To prevent the estimator of alpha from being biased, we extend the four-factor model by including latent factors.

doi:10.6342/NTU202404642

The extended benchmark model that includes both observed factor $f_{o,t}$ and latent factor $f_{\ell,t}$ is as follows:

$$r_t = \alpha + \beta_o' \lambda_o + \beta_\ell' \lambda_\ell + \beta_o' \left[f_{o,t} - \mathbb{E}(f_{o,t}) \right] + \beta_\ell' \left[f_{\ell,t} - \mathbb{E}(f_{\ell,t}) \right] + \varepsilon_t, \tag{3.3}$$

where $f_{o,t}$ is a $K_o \times 1$ vector denoting the observed factors, and $f_{\ell,t}$ is a $K_\ell \times 1$ vector denoting the latent factors, with K_o and K_ℓ being the numbers of observed and latent factors, respectively.

Before explaining the estimation procedure for alpha, we first introduce some notations used in this paper. Matrices are denoted by uppercase italic letters, and their column vectors are denoted by lowercase italic letters (e.g., $X = (x_1, x_2, \ldots, x_T)$, where x_t represents a vector of data at time t). Let $\mathbf{1}_M$ be the $M \times 1$ vector of ones. For any $m \times n$ matrix X, we use $\mathbb{H}_X = X(X'X)^{-1}X'$ to denote its hat matrix and $\mathbb{M}_X = I_m - \mathbb{H}_X$ to denote its annihilator matrix. Since some return data contain missing values, we define $\mathcal{T}_i$ as the set of observed time periods with T_i elements for mutual fund i, and $\mathcal{M}_t$ as the set of observed mutual funds with M_t elements at time t. Let $F_{o,i}$ be the $K_o \times T_i$ matrix of $\{f_{o,t} : t \in \mathcal{T}_i\}$, denoting the observed factors for mutual fund i. Similarly, we define the matrix of latent factors for mutual fund i as $F_{\ell,i}$, representing $\{f_{\ell,t} : t \in \mathcal{T}_i\}$. Let R_i be the $T_i \times 1$ vector of $\{r_{it} : t \in \mathcal{T}_i\}$, denoting the excess returns for mutual fund i, and $\bar{r}_i = \sum_{t \in \mathcal{T}_i} r_{it}/T_i$ denotes the average excess return for mutual fund i over its observed time periods.

To estimate alpha using the benchmark model with latent factors as shown in Equation (3.3), we first need to estimate the factor loadings for both the observed and latent

doi:10.6342/NTU202404642

factors, $\beta_{o,i}$ and $\beta_{\ell,i}$, for each mutual fund through a time-series regression:

$$R_i = a_i + F'_{o,i}\beta_{o,i} + F'_{\ell,i}\beta_{\ell,i} + \varepsilon_i, \tag{3.4}$$

where a_i is the intercept of the time-series regression. However, since the latent factor $F_{\ell,i}$ is unobservable, we can only estimate $\beta_{o,i}$ first. The effect of $F'_{o,i}$ on R_i in this time-series regression can be obtained by regressing the residual from the regression of R_i on $\mathbf{1}_{T_i}$ (denoted as $\mathbb{M}_{\mathbf{1}_{T_i}} R_i$) on the residual from regression of $F'_{o,i}$ on $\mathbf{1}_{T_i}$ (denoted as $\mathbb{M}_{\mathbf{1}_{T_i}} F'_{o,i}$). The estimator of $\beta_{o,i}$ is then given by:

$$\hat{\beta}_{o,i} = (F_{o,i}\mathbb{M}_{\mathbf{1}_{T_i}} F'_{o,i})^{-1}(F_{o,i}\mathbb{M}_{\mathbf{1}_{T_i}} R_i). \tag{3.5}$$

It is important to note that $\hat{\beta}_{o,i} \xrightarrow{p} \beta_{o,i}$ only if $F_{o,i}$ is uncorrelated with $F_{\ell,i}$. However, since we do not impose this assumption here, the estimator of $\beta_{o,i}$ might be biased, which could potentially lead to bias in the alpha estimator. Therefore, we will de-bias the estimator of alpha in Section 3.3.

Subsequently, to estimate the factor loadings of the latent factors, we begin by extracting their effect from excess returns by subtracting the effects of alpha, risk premium, and observed factors from Equation (3.3):

$$z_{it} = r_{it} - \bar{r}_i - \hat{\beta}'_{o,i}(f_{o,t} - \bar{f}_{o,i}), \tag{3.6}$$

where $\bar{r}_i$ is the estimator of $\mathbb{E}(r_i)$, representing the effects of alpha and risk premium[1]. The term $\bar{f}_{o,i}$ denotes the average observed factors for mutual fund i over its observed time periods, defined as $\bar{f}_{o,i} = \sum_{t\in\mathcal{T}_i} f_{o,t}/T_i$. Consequently, the matrix $Z_{M\times T} = (z_{it})$

[1] We can derive this by taking the expectation on both sides of Equation (3.3) and have $\mathbb{E}(r_i) = \alpha_i + \beta'_{o,i}\lambda_{o,i} + \beta'_{\ell,i}\lambda_{\ell,i}$

doi:10.6342/NTU202404642

contains only the effect of latent factors and is referred to as the residual matrix. Note that z_{it} is defined only when r_{it} is observable; otherwise, it is treated as missing.

Once we obtain the residual matrix Z, we can decompose it by applying PCA to identify the latent factors $F_{\ell,i}$ and estimate their corresponding factor loadings β_ℓ. However, in practice, mutual fund data often contain missing values since some funds only have short lifespans while new funds frequently enter the market, making it difficult to apply PCA directly. To address this problem, in the next section, we apply the matrix completion used in Giglio, Liao, and Xiu (2021) to identify the latent factors and estimate their factor loadings.

3.2 Matrix Completion

Matrix completion is a technique for filling in missing values within observed data. The main assumption in the matrix completion approach used in this paper is that the observed residual matrix Z can be decomposed into a lower-rank matrix X, which captures the underlying structures of Z, and a matrix N, which contains only noise:

$$z_{ij} = X_{ij} + N_{ij} \quad \text{if } z_{ij} \text{ is observable.} \tag{3.7}$$

Under this assumption, the goal of matrix completion is to find a simplified matrix X that preserves the essential information from the residual matrix Z. Therefore, the objective function can be written as:

$$\min_X \quad \|(Z - X) \circ \Omega\|_F^2 + \text{rank}(X), \tag{3.8}$$

doi:10.6342/NTU202404642

where Ω is a $M \times T$ binary matrix that indicates the observed data, with the (i, j)-th element equals to 1 if z_{ij} is defined, $\circ$ denotes the Hadamard product, used for element-wise multiplication between two matrices, and $\|X\|_F$ is the Frobenius norm of the matrix X where $\|X\|_F = (\text{Tr}(X'X))^{1/2}$. Therefore, the first term of the objective function aims to minimize the noise while the second term is used for constraining the rank of X.

However, the rank minimization problem is NP-hard, and all known algorithms for solving it are exponential-time algorithms, which become inefficient for high-dimension inputs. Therefore, an alternative approach is to replace the rank with the nuclear norm. The intuition behind using the nuclear norm is that rank is the count of non-zero singular values, while the nuclear norm is the sum of these singular values, much like using their magnitude as an approximation of rank. Importantly, the nuclear norm is a convex function and can be solved efficiently. Therefore, we reformulate the objective function in Equation (3.8) by substituting the rank with the nuclear norm and have:

$$\min_X \quad \|(Z - X) \circ \Omega\|_F^2 + \lambda_{MT} \|X\|_n, \tag{3.9}$$

where $\lambda_{MT} > 0$ is a regularization parameter, and $\|X\|_n$ denotes the nuclear norm of matrix X.

We can then solve this objective function by using the optimal solution that proved by Ma et al. (2011). For any $\tau > 0$, the optimal solution $\hat{X}$ should satisfy the following equation:

$$\hat{X} = D_\nu(\hat{X} - \tau\Omega \circ (\hat{X} - Z)), \quad \nu = \tau\lambda_{MT}/2, \tag{3.10}$$

doi:10.6342/NTU202404642

where D_ν is a singular value thresholding operator defined as:

$$\mathcal{D}_\nu(Y) \coloneqq U\,\Sigma_\nu\,V', \quad \Sigma_\nu = \text{diag}(\max\{\sigma_{ii} - \nu, 0\}), \tag{3.11}$$

where $Y = U\Sigma V'$ is the result of singular value decomposition (SVD) of matrix Y and its singular value σ_{ii} are all positive. Finally, we can use the iterative algorithm proposed by Ma et al. (2011) to find $\hat{X}$. The algorithm is as follows:

1. Let $\nu = \tau\lambda_{MT}/2$, $k = 0$, and set the initial state $X_0 = Z \circ \Omega$.
2. Update $X_{k+1} = \mathcal{D}_\nu(X_k - \tau\Omega \circ (X_k - Z))$ and let $k = k + 1$.
3. Repeat the second step until X_k converges.

Note that this algorithm has two hyperparameters, τ and λ_{MT}. As mentioned previously, λ_{MT} is the regularization parameter controlling the balance between reducing the noise and enforcing the low-rank structure of X. The hyperparameter τ serves as a threshold for D_τ, acting as a cutoff for singular values, replacing low singular values with zero to help identify a low-rank matrix while minimizing noise. Following the suggestion of Giglio, Liao, and Xiu (2021), we choose $\tau = 0.9$ and $\lambda_{MT} = 2.2 \cdot \|\Omega \circ W\|_2$, where W is a noise matrix with $W \sim N(0, \Sigma_\varepsilon)$. The (i, j) -th element of Σ_ε is the estimated covariance between mutual fund i and j. We follow the estimation procedure with the Python code provided by Giglio, Liao, and Xiu (2021) to estimate Σ_ε through SVD.

Once we obtain the lower-rank matrix $\hat{X}$, we apply PCA to $\hat{X}$ to estimate the de-meaned latent factors, $v_{\ell,t}$, where $v_{\ell,t} = f_{\ell,t} - \sum_{t=1}^{T} f_{\ell,t}/T$, and their corresponding load-

doi:10.6342/NTU202404642

ings, $\beta_{\ell,t}$. The estimators for the latent factors and their loadings are as follows:

$$\hat{v}_{\ell,t} = \left(\sum_{i \in \mathcal{M}_t} b_i b_i' \right)^{-1} \sum_{i \in \mathcal{M}_t} b_i z_{it}, \quad t = 1,\ldots,T, \tag{3.12}$$

$$\hat{\beta}_{\ell,i} = \left(\sum_{i \in \mathcal{T}_i} \hat{v}_{\ell,t} \hat{v}_{\ell,t}' \right)^{-1} \sum_{i \in \mathcal{T}_i} \hat{v}_{\ell,t} z_{it}, \quad i = 1,\ldots,M, \tag{3.13}$$

where $b_i = \sqrt{M}\rho_i$ for $i \leq K_\ell$, and $(\rho_1, \ldots, \rho_{K_\ell})$ are the left singular vectors of $\hat{X}$ corresponding to the largest K_ℓ eigenvalues. Finally, we combine the observed factor loadings with latent factor loadings as $\hat{\beta} = \left(\hat{\beta}_o, \hat{\beta}_\ell\right)$, and also combine the demeaned factors as $\hat{v}_t = ((f_{o,t} - \bar{f}_o)', \hat{v}_{\ell,t}')'$, where $\bar{f}_o = \sum_{t=1}^{T} f_{o,t}/T$.

3.3 Estimate alpha and the test statistics

With the estimated factor loadings $\hat{\beta}$, we can estimate the risk premiums of the factors, $\lambda = (\lambda_o', \lambda_\ell')'$, by taking the expectation on both sides of Equation (3.3):

$$\mathbb{E}(r_t) = \alpha + \beta' \lambda. \tag{3.14}$$

We then estimate λ by regressing $\bar{r}$ on $\hat{\beta}$ using the cross-sectional regression:

$$\hat{\lambda} = (\hat{\beta}' \mathbb{M}_{\mathbf{1}_M} \hat{\beta})^{-1} (\hat{\beta}' \mathbb{M}_{\mathbf{1}_M} \bar{r}), \tag{3.15}$$

where $\bar{r} = (\bar{r}_1, \ldots, \bar{r}_M)'$. Finally, we can estimate alpha by subtracting the product of the risk premiums and factor loadings from the excess return for each mutual fund.

However, due to the bias in the $\hat{\beta}$ discussed in Section 3.1 and the unbalanced panel effect, the alpha estimator is biased. To address this problem, Giglio, Liao, and Xiu (2021)

doi:10.6342/NTU202404642

propose a de-biasing estimator of alpha, denoted as A_i. The unbiased estimator of alpha is as follows:

$$\hat{\alpha}_i = \bar{r}_i - \hat{\beta}_i \hat{\lambda} + \hat{A}_i, \quad i = 1, \ldots, M, \tag{3.16}$$

where

$$\hat{A}_i = \hat{\beta}'_{\ell,i}(\hat{H}_{o,i} - \hat{H}_o)\hat{\lambda}_o - \hat{\xi}'_i \hat{g}, \tag{3.17}$$

$$\hat{g}_i = \frac{1}{T_i}\sum_{t \in \mathcal{T}_i} \hat{v}'_t \hat{\beta}_i, \quad \hat{\xi}'_i = e'_i - \hat{\beta}'_i(\hat{\beta}' \mathbb{M}_{\mathbf{1}_M} \hat{\beta})^{-1} \hat{\beta}' \mathbb{M}_{\mathbf{1}_M}, \quad e'_i = (0, \ldots, 1, \ldots, 0),$$

$$\hat{H}_{o,i} = \hat{V}_{\ell,i} \mathbb{M}_{\mathbf{1}_{T_i}} F'_{o,i}(F_{o,i} \mathbb{M}_{\mathbf{1}_{T_i}} F'_{o,i})^{-1}, \text{ and } \hat{H}_o = \hat{V}_\ell \mathbb{M}_{\mathbf{1}_T} F'_o (F_o \mathbb{M}_{\mathbf{1}_T} F'_o)^{-1}.$$

Here, $\hat{V}_\ell$ is the $K_\ell \times T$ matrix of $\{\hat{v}_{\ell,t} : t \leq T\}$, and $\hat{V}_{\ell,i}$ is the $K_\ell \times T_i$ matrix of $\{\hat{v}_{\ell,t} : t \in \mathcal{T}_i\}$. For further details on this derivation, refer to Giglio, Liao, and Xiu (2021), Appendix, Proof of Theorem A.2.

Finally, Giglio, Liao, and Xiu (2021) propose an estimator for the variance of alpha to construct the t-statistic for evaluating the mutual performance. The variance estimator, denoted by $\hat{\sigma}_i^2$, for mutual fund i, is as follows:

$$\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t \in \mathcal{T}_i} \hat{u}_{it}^2 (1 - \hat{v}'_t \hat{\Sigma}_f^{-1} \hat{\lambda})^2 + \frac{1}{M^2} \widehat{\text{Var}}(\hat{\alpha}) \hat{\beta}'_i \hat{S}_{\hat{\beta}}^{-1} \hat{\beta}_i, \tag{3.18}$$

where $\hat{u}_{it} = r_{it} - \bar{r}_i - \hat{\beta}'_i \hat{v}_t$ is the residual of Equation (3.3), $\hat{\Sigma}_f = \sum_{t=1}^{T} \hat{v}_t \hat{v}'_t / T$, and $\hat{S}_{\hat{\beta}} = \hat{\beta}' \mathbb{M}_{\mathbf{1}_M} \hat{\beta} / M$. The first component in Equation (3.18) is the variance from time-series estimation (Equation (3.5) and (3.13)), while the second component is the variance from cross-sectional estimation (Equation (3.15)). Further details can be found in Giglio, Liao, and Xiu (2021), Appendix, Theorem A.1. The t-statistic of alpha is then calculated as:

$$\hat{t}_{\hat{\alpha}_i} = \sqrt{T_i} \cdot \frac{\hat{\alpha}_i}{\hat{\sigma}_i}, \quad i = 1, \ldots, M. \tag{3.19}$$

doi:10.6342/NTU202404642

Giglio, Liao, and Xiu (2021) show that $\hat{t}_{\hat{\alpha}_i}$ is asymptotically normal so that its p-values can be computed directly. However, due to the missing values in mutual fund data, the performance of asymptotically inference in finite sample may be a concern. Therefore, we follow the suggestion of KTWW (2006) but apply the wild bootstrap procedure to evaluate the p-values.

3.4 Wild Bootstrap procedure

To improve the finite sample performance, the bootstrap procedure is an appealing approach for evaluating the critical value of test statistics. Therefore, we apply the wild bootstrap proposed by Liu (1988) to evaluate the p-values since it can avoid the problem of discrepancies in the number of observations, making it suitable for assessing mutual fund performance. Moreover, as shown by Mammen (1993), the wild bootstrap can accommodate heteroskedasticity, reducing the false rejections of the null hypothesis. Let $\hat{\varepsilon}_{it} = r_{it} - \bar{r}_i - \hat{\beta}_i' \hat{v}_t$ be the residual of our asset pricing model. The wild-bootstrap procedure is as follows:

1. Generate a sequence of weighted residuals $\{w_{it} : i \leq M, t \leq T\}$ that satisfy $\mathbb{E}(w_{it}) = 0$ and $\text{Var}(w_{it}) = 1$. Mammen (1993) suggest using the following equation to generate w_{it}:

$$w_{it} = \frac{1}{\sqrt{2}} \eta_{it} + \frac{1}{2}(\gamma_{it}^2 - 1), \tag{3.20}$$

where η_{it} and γ_{it} are independent standard normal random variables. Then, we can generate the bootstrap return by:

$$r_{it}^* = \hat{\beta}_i' \hat{\lambda} + \hat{\beta}_i' \hat{v}_t + \hat{\varepsilon}_{it}^*, \quad \hat{\varepsilon}_{it}^* = \hat{\varepsilon}_{it} w_{it}, \quad \text{for } t \in \mathcal{T}_i. \tag{3.21}$$

doi:10.6342/NTU202404642

2. For each bootstrap iteration, estimate its factor loading and risk premium by:

$$\hat{\beta}_i^* = (\hat{V}_i \mathbb{M}_{\mathbf{1}_{T_i}} \hat{V}_i')^{-1} (\hat{V}_i \mathbb{M}_{\mathbf{1}_{T_i}} R_i^*), \tag{3.22}$$

where R_i^* is a $T_i \times 1$ vector, denoting the bootstrap return for mutual fund i, and $\hat{V}_i$ is a $K \times T_i$ matrix, denoting the factor values when mutual fund i exists. Then, we estimate the risk premium λ^* by substituting $\hat{\beta}$ and $\bar{r}$ in Equation (3.15) with $\hat{\beta}^*$ and $\bar{r}^*$.

3. For each mutual fund, estimate its alpha by the de-biased estimator:

$$\hat{\alpha}_i^* = \bar{r}_i^* - \hat{\beta}_i^{*'} \hat{\lambda}^* - \hat{\xi}_i^{*'} \hat{g}^*, \quad i = 1, \ldots, M, \tag{3.23}$$

where $\hat{\xi}_i^{*'} = e_i' - \hat{\beta}_i^{*'} (\hat{\beta}^{*'} \mathbb{M}_{\mathbf{1}_M} \hat{\beta}^*)^{-1} \hat{\beta}^{*'} \mathbb{M}_{\mathbf{1}_M}$, and $\hat{g}_i^* = \sum_{t \in \mathcal{T}_i} \hat{v}_t' \hat{\beta}_i^* / T_i$. Then, compute the bootstrap t-statistic of alpha, $\hat{t}_{\hat{\alpha}_i,b}^*$, by substituting $\hat{\beta}$, $\hat{\lambda}$, $\hat{\alpha}$ and $\hat{u}$ in Equation (3.18) and (3.19) with $\hat{\beta}^*$, $\hat{\lambda}^*$, $\hat{\alpha}^*$ and $\hat{\varepsilon}^*$.

4. Repeat steps 1 to 3 B times to obtain the bootstrapped distribution of $\hat{t}_{\hat{\alpha}_i,b}^*$ when the true t-statistic is zero. Finally, we evaluate the p-value of the test that $t_\alpha > 0$ for each fund by

$$p_i = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{\hat{t}_{\hat{\alpha}_i,b}^* > \hat{t}_{\hat{\alpha}_i}\}, \quad i = 1, \ldots, M, \tag{3.24}$$

where $\mathbb{1}$ is a $(1, 0)$ indicator variable, and $\hat{t}_{\hat{\alpha}_i}$ is obtained from Equation (3.19).

Note that since we use the estimated latent factors as the observed factors in the bootstrap procedure, the estimator to correct the bias of alpha in Equation (3.23) is different from the estimator in Equation (3.16) and (3.17).

It important to note that unlike the bootstrap procedure in Giglio, Liao, and Xiu

doi:10.6342/NTU202404642

(2021), we do not bootstrap on alpha; instead, we bootstrap on the t-statistic of alpha. It is because the t-statistic is standardized by the standard deviation and the number of observations, which provides better properties (KTWW, 2006). Finally, we apply the screening BH procedure (Equations (2.17) and (2.18)) to these p-values and infer the existence of ouperforming mutual funds.

doi:10.6342/NTU202404642

Chapter 4 Empirical Results

4.1 Mutual Fund Data

To analyze the performance of Japanese mutual funds, we use the adjusted closing price at the end of each month from January 2002 to December 2023, sourced from the Bloomberg database, to evaluate monthly returns. Since the returns are based on adjusted closing prices, they already account for stock dividends, splits, management fees, taxes, etc. Our analysis focuses on open-ended mutual funds that primarily invest in the Japanese equity market. We exclude funds with names containing "index", "idx", "ETF", or "tracker fund" to focus on active mutual funds. Furthermore, we select funds with a minimum of 60 months of return data to ensure precise alpha estimation. Our final database contains a total of 1483 mutual funds.

Table 4.1 presents the number of funds, descriptive statistics of their average monthly excess returns, and the proportion of missing values (denoted as "Missing Proportion") over the sample period (2002−2023) and overlapping ten-year subperiods. For each subperiod, only funds with at least 60 observations are included. The descriptive statistics include the mean, standard deviation, and key percentiles of the distribution of average return for each time period.

doi:10.6342/NTU202404642

Table 4.1: Summary Statistic.

Time Periods	N	Mean	Standard Deviation	50th Percentile	90th Percentile	Max	Missing Proportion
Full Sample	1483	0.47	0.0045	0.48	0.98	4.51	46.85
2002 – 2011	629	−0.22	0.0045	−0.20	0.26	1.47	15.91
2003 – 2012	678	−0.01	0.0043	0.06	0.48	2.21	17.31
2004 – 2013	700	0.18	0.0039	0.25	0.61	1.55	17.26
2005 – 2014	715	0.32	0.0050	0.31	0.92	3.04	16.88
2006 – 2015	743	0.25	0.0059	0.16	1.05	4.51	16.29
2007 – 2016	783	0.36	0.0056	0.26	1.12	4.51	16.13
2008 – 2017	803	0.63	0.0047	0.55	1.22	4.51	15.49
2009 – 2018	860	0.78	0.0040	0.76	1.16	4.51	16.27
2010 – 2019	924	0.77	0.0039	0.78	1.17	4.51	16.38
2011 – 2020	1017	0.74	0.0045	0.77	1.23	3.11	17.41
2012 – 2021	1065	0.93	0.0046	0.98	1.46	3.46	17.01
2013 – 2022	1073	0.74	0.0041	0.77	1.18	2.99	15.42
2014 – 2023	1069	0.60	0.0034	0.62	0.98	2.83	13.66

Notes: N represents the number of funds in each time period. The value of Mean, 50th percentile, 90th percentile, Max, and Missing Proportion are reported in percentage (%).

As shown in Table 4.1, the full sample contains 46.85% missing values, while each subperiod has approximately 16% missing values, showing the necessity of employing matrix completion to identify latent factors in the benchmark model. The lower proportion of missing values in each subperiod is because we only include funds with at least five years of data, which reduces the missing value proportion compared to the full sample. Table 4.1 also shows that the maximum average excess returns among these mutual funds are notably high, particularly from 2006−2015 to 2010−2019, with returns averaging 4.51%. However, we still need to control the risk factors and the FDR to determine whether these funds truly outperform the benchmark. Lastly, we observe that the mean, 50th percentile, 90th percentile, and maximum excess returns are low in the first few subperiods. They then slightly increase to 2012−2021 and decline in the last two subperiods. This suggests that the proportion of outperforming mutual funds may vary across different subperiods.

To estimate the alpha of each mutual fund, we use the Carhart four-factor as the ob-

doi:10.6342/NTU202404642

served factors and follow the methodology described in Sections 3.1 to 3.3. Figure 4.1 displays the scree plot of the top 15 eigenvalues from the residual covariance matrices for both the Carhart four-factor model and its extended model, which incorporates additional latent factors. The solid line represents the eigenvalues of the residual matrix from the Carhart four-factor model. The relatively high values of the first two eigenvalues, compared to the others, indicate that certain common patterns remain in the residuals even after applying the Carhart four-factor model. To address this, we apply the elbow method[1] and incorporate four additional latent factors. The dashed line, representing the eigenvalues of the residual matrix from the extended model, shows a significant drop in the first three eigenvalues, indicating that the latent factors effectively capture these common patterns. Furthermore, the flattening of the dashed line across all values suggests that no significant common patterns remain among the mutual funds after applying the model with latent factors.

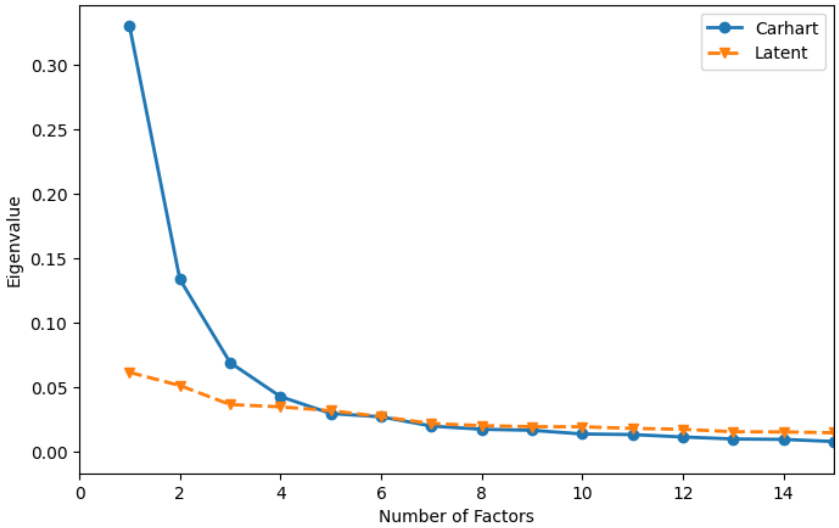


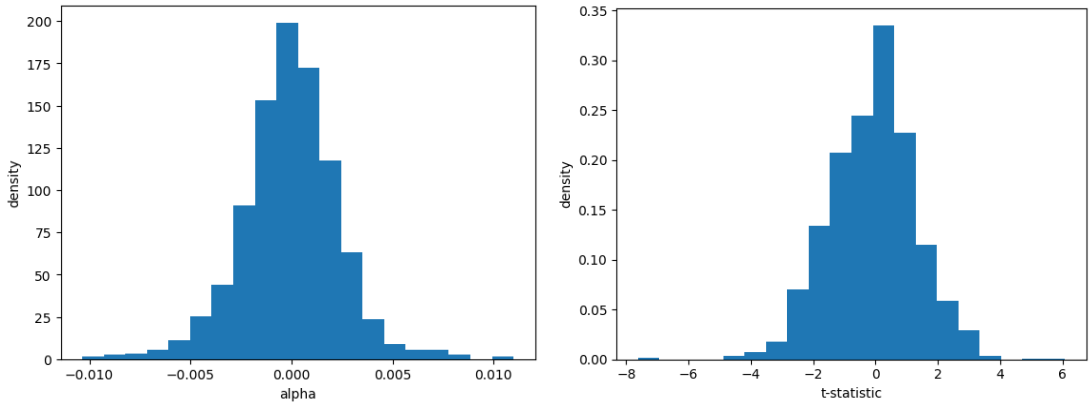
Figure 4.1: Scree plots of eigenvalues

Figure 4.2 displays the histogram of alpha and the t-statistic of alpha for Japanese mu-

[1]The elbow method is a heuristic technique used to determine the optimal number of principal components in PCA. This method involves plotting the ordered eigenvalues and identifying the ”elbow” point in the curve, which represents a significant change in the rate of decline. The number of principal components corresponding to this point is then selected, as it typically captures most of the data’s variance without including too many components that contribute little additional information.

doi:10.6342/NTU202404642

tual funds. Among the 1483 mutual funds, the average alpha is -0.644 basis points (bps), with 49.8% of them showing a positive alpha (and a positive t-statistic). Compared to the mean of average excess return shown in Table 4.1 (0.47%), the mean of positive alpha decreases significantly, indicating that some of these excess returns are due to risk-taking. Additionally, 6.47% of the mutual funds have a t-statistic greater than 1.96, suggesting that some of these funds have a statistically significant alpha. However, due to the multiple testing problems, some funds may outperform the benchmark purely by luck. Therefore, in the next section, we will control the FDR to identify the truly outperforming mutual funds.



(a) Histogram of alpha **(b)** Histogram of t-statistic of alpha

Figure 4.2: Histograms of mutual funds alpha and the t-statistic of alpha

4.2 Long-Term Mutual Fund Performance

In our empirical analysis, we begin by identifying outperforming mutual funds based on long-term performance during the period from 2002 to 2023. We use the t-statistic of alpha as the test statistic and implement 1000 wild bootstrap replications outlined in Section 3.4 to construct the null distribution and evaluate the p-values for each fund. Finally, we apply the screening BH procedures, as described in equations (2.17) and (2.18), to identify outperforming mutual funds. Following Barras et al. (2010) and Cuthbertson

doi:10.6342/NTU202404642

et al. (2012), we control the FDR at 10%, 15%, 20%. Table 4.2 presents the proportion of the outperforming funds, along with their average alphas.

Table 4.2: Long-Term Performance: Proportion of funds and average α.

	FDR (γ)		
	10%	15%	20%
Proportion of the outperforming funds (%)	0.47	0.74	0.94
Average alpha (bps/month)	56.75	57.49	51.83

Among the 1483 mutual funds, we find that although 10.99% have a p-value lower than 0.1 in individual tests (not reported in the table), only 0.47% are identified as outperforming funds (i.e., having a significant positive t-statistic for alpha) after controlling the FDR at 10%. This suggests that most Japanese mutual funds outperform the benchmark purely by luck and are filtered out by the screening BH procedure.

We also observe that as γ increases from 10% to 20%, the proportion of outperforming funds rises because more false discoveries are tolerated. For example, the proportion of outperforming funds grows from 0.47% to 0.94% as γ increases from 10% to 20%. Interestingly, the average alpha of outperforming funds reaches its peak at 57.49 bps when γ =15%. This suggests that the additional outperforming funds identified by raising γ to 15% generate higher alpha compared to those identified when γ =10%. However, when γ =20%, the average alpha declines to 51.83 bps, potentially due to the inclusion of funds with weaker performance.

Lastly, when comparing the screening BH procedure to the BH procedure, we find that the screening BH procedure identifies four more outperforming mutual funds when $\gamma = 20\%$ (not reported here). Although the increase is just a few, it indeed enhances the power of the test. Therefore, for the remainder of this paper, we use the screening BH procedure to control the FDR.

doi:10.6342/NTU202404642

Overall, the long-term performance analysis shows that only 0.47% mutual funds in Japan truly outperform the benchmark, with a relatively high average alpha of 56.75 bps per month (6.81% per year). Even when allowing for more false discoveries by increasing γ to 20%, still only 0.94% of the funds are identified as outperforming. However, it is possible that these mutual funds have the ability to outperform the benchmark over shorter periods, but their superior performance tend to vanish and fail to sustain over the long term. To further explore this probability, we will analyze their short-term performance across different time periods in the next section.

4.3 Short-Term Performance

To examine mutual fund performance in the short-term, we divide the data into thirteen overlapping 10-year groups, beginning with the period from 2002 to 2011 and ending with 2014 to 2023. For each group, we include only mutual funds with at least 60 months of data. We then apply the same methodology used in the long-term performance analysis to estimate each group's alpha independently. Table 4.3 presents the results of the short-term performance analysis. For each row of Table 4.3, we report the average alpha for each subperiod (denoted as "Full Avg α"), the proportion of outperforming funds (denoted as "Prop") and the average alpha of these outperforming funds (denoted as "Avg α") while controlling the FDR at 10%, 15%, and 20% using the screening BH procedure.

From Table 4.3, we first observe that in the beginning four time periods, the proportion of outperforming funds is lower than the long-term proportion (0.47%). However, during the subperiods 2010−2019 and 2012−2021, the proportion of outperforming funds increases to 3.14% and 9.30%, respectively, and suddenly decreases to 0.28% and

doi:10.6342/NTU202404642

Table 4.3: Short-Term Performance: Proportion of funds and average α

Time Periods	Full Avg α	$\gamma = 10\%$ Prop	$\gamma = 10\%$ Avg α	$\gamma = 15\%$ Prop	$\gamma = 15\%$ Avg α	$\gamma = 20\%$ Prop	$\gamma = 20\%$ Avg α
2002 – 2011	−22.00	0.00	0.00	0.00	0.00	0.00	0.00
2003 – 2012	−15.20	0.29	80.32	0.29	80.32	0.29	80.32
2004 – 2013	−7.73	0.29	86.80	0.29	86.80	0.29	86.80
2005 – 2014	−5.38	0.00	0.00	0.00	0.00	0.42	76.07
2006 – 2015	4.87	0.94	64.98	0.94	64.98	1.08	70.00
2007 – 2016	4.69	0.26	72.49	0.26	72.49	0.26	72.49
2008 – 2017	8.93	0.25	67.42	0.25	67.42	0.50	71.15
2009 – 2018	12.22	0.93	66.91	1.28	65.31	1.86	62.09
2010 – 2019	20.51	3.14	59.56	5.84	56.31	6.39	55.45
2011 – 2020	18.24	1.18	66.65	4.03	59.01	8.16	52.10
2012 – 2021	25.88	9.30	57.13	15.87	52.19	25.63	47.96
2013 – 2022	11.61	0.28	48.41	0.28	48.41	1.49	55.63
2014 – 2023	8.85	1.40	49.36	1.96	50.57	3.55	47.31

Note: The proportion of mutual funds (denoted as "Prop") is reported in percentage (%), while the average alpha of all mutual funds in each subperiod (denoted as "Full Avg α") and the average alpha of outperforming funds in each subperiod (denoted as "Avg α") are reported in bps/month.

1.40% in the subsequent subperiods. This suggests that Japanese mutual funds may outperform the benchmark in the short term, but their superior performances vanish quickly. Additionally, when the full sample average alpha is high, more outperforming funds are identified in those subperiods, particularly when γ is 15% and 20%. For example, in the subperiods 2010−2019 through 2012−2021, their average alpha is 20.51 bps, 18.24 bps, and 25.88 bps per month, with corresponding proportions of outperforming funds being 5.84%, 4.03%, and 15.87% when γ is 15% (6.39%, 8.16%, and 25.63% when γ is 20%). This may be because a higher full sample average alpha reflects more favorable market conditions for funds, leading to a higher proportion of outperforming funds.

Moreover, Table 4.3 also shows that the average alpha of outperforming funds is much higher than the full sample average alpha, indicating that our procedure indeed selects those with superior performance. It is also interesting to find that there is a dramatic

doi:10.6342/NTU202404642

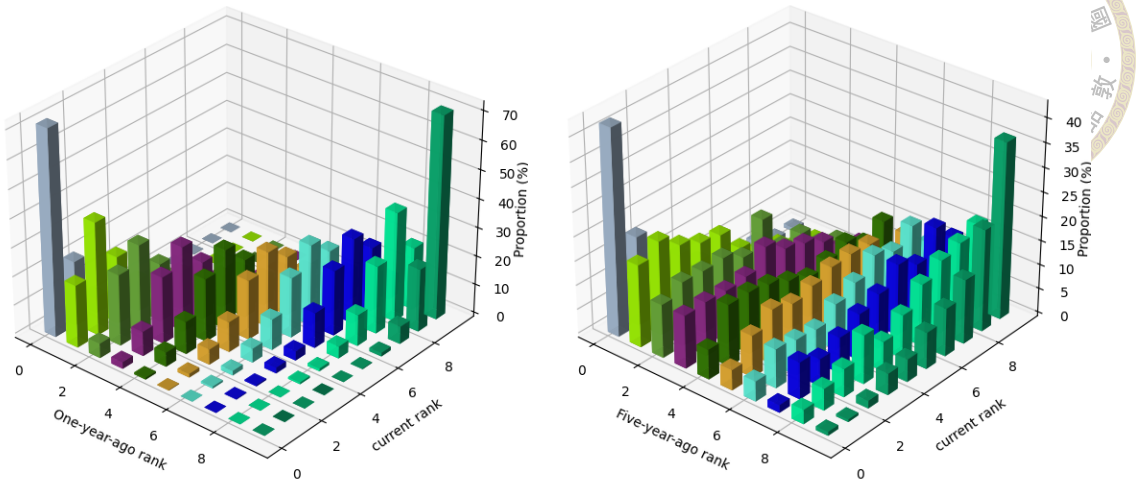
drop in the proportion of outperforming funds between the subperiods 2012−2021 and 2013−2022. This decline also reflects the decrease in average excess returns reported in Table 4.1.

Overall, although only 0.47% of mutual funds outperform the benchmark in the long term, we find that more mutual funds demonstrate superior performance in the short term, particularly in the subperiods 2010−2019 and 2012−2021. This raises the question of whether these outperforming funds with short-term superior performance consistently outperform others but fail to sustain sufficiently strong performance to outperform the benchmark over the long term or whether they perform well only in specific subperiods and underperform in others. To determine which explanation is more reliable, we will examine the rank persistence of these mutual funds in the next section.

4.4 Rank Persistence Analysis

To examine the persistence of rank among these mutual funds, we rank mutual fund performance in each subperiod defined in the previous section and analyze how these ranks evolve over time. Specifically, we divide the data into thirteen overlapping 10-year groups and apply a method similar to the one used in the short-term performance analysis to evaluate the p-values of the t-statistic of alpha. For each period, we rank the mutual funds into ten groups based on their p-values and evaluate the probability $\mathbb{P}(\text{Current rank} = j \mid \text{Previous rank} = i)$, where $1 \leq i, j \leq 10$, to determine how these ranks change over time. Finally, we present the results using a 3D bar plot in Figure 4.3, which compares the ranks from one and five years prior to the current rank. The value of the bar at position (i, j) represents $\mathbb{P}(\text{Current rank} = j \mid \text{Previous rank} = i)$.

doi:10.6342/NTU202404642



(a) One-year-ago rank compared to current rank **(b)** Five-year-ago rank compared to current rank

Figure 4.3: 3D bar plot comparing the performance rank from one year and five years prior to the current rank.

Panel (a) of Figure 4.3 shows that most mutual funds tend to remain in the same rank, especially for those with the best and worst performance. Specifically, 71.67% of the top-performing mutual funds remain in the top rank, while 70.54% of the worst-performing funds stay in the bottom rank. Moreover, almost no mutual funds experience dramatic rank changes, indicating that these mutual funds exhibit persistence over a one-year period. To examine the persistence over a longer period, Panel (b) of Figure 4.3 compares the ranks from five years ago to their current ranks. The results show that mutual funds with the best and worst performance continue to exhibit persistence, while funds in the middle ranks tend to have more random ranking over the five-year period.

The results above indicate that, although these short-term outperforming mutual funds do not consistently exhibit superior performance across all periods, they tend to remain in the top ranks over time. Furthermore, we are also interested in whether these in-sample superior performances can generate economic value out-of-sample. To investigate this, we will examine the out-of-sample performance of portfolios composed of these outperforming funds in the next section.

doi:10.6342/NTU202404642

4.5 Out-of-Sample Performance

In this section, we use the outperforming mutual funds identified in Section 4.3 to construct portfolios of outperforming funds, while controlling the FDR (γ) at levels of 10%, 15%, and 20% to examine their out-of-sample performance and determine whether these outperforming funds provide economic value to investors. Specifically, at each rebalancing date, we evaluate the p-values of the t-statistic of alpha using data from the previous ten years for each mutual fund. We then apply the screening BH procedure to these p-values to identify outperforming funds and use these funds to construct equally weighted portfolios. If no outperforming mutual funds are identified, the portfolio will be empty in the subsequent year. Finally, we evaluate their out-of-sample performance by examining their raw returns in the subsequent year. Each portfolio is held for one year and rebalanced annually at the end of each year. Only funds with at least 60 months of return data within the in-sample period are considered. The first portfolio is constructed at the end of 2011, using 2002 to 2011 as the in-sample period and testing performance in 2012, while the final portfolio is constructed at the end of 2022, using 2013 to 2022 as the in-sample period and testing performance in 2023.

We present the cumulative wealth of the FDR-controlled portfolios at various levels and the portfolio information in Figure 4.4 and Table 4.4, respectively. All portfolios start with an initial wealth value of 1. To evaluate the cumulative wealth, we multiply the raw return for each year by the portfolio's wealth at the beginning of that year, covering the period from 2012 through 2023. Since the analysis focuses on the performance of Japanese mutual funds, the Nikkei 225 index is used as the benchmark portfolio for comparison. It is important to note that during periods when no outperforming funds are identified and

doi:10.6342/NTU202404642

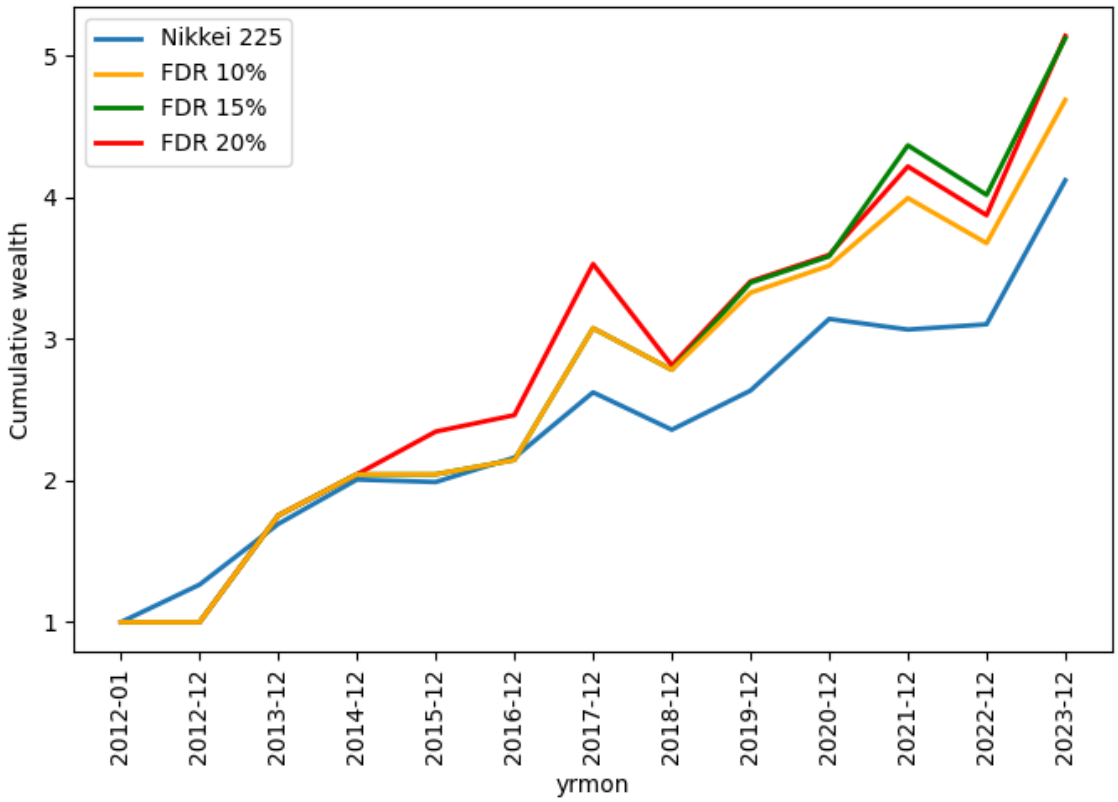


Figure 4.4: Cumulative wealth of different portfolios.

Table 4.4: Portfolio information

Out-of-Sample Periods	Number of funds in each portfolio $\gamma = 10\%$	$\gamma = 15\%$	$\gamma = 20\%$
2012	0	0	0
2013	2	2	2
2014	2	2	2
2015	0	0	3
2016	7	7	8
2017	2	2	2
2018	2	2	4
2019	8	11	16
2020	29	54	59
2021	12	41	83
2022	99	169	273
2023	3	3	16

doi:10.6342/NTU202404642

included in the portfolio, the cumulative wealth remains constant (flat) for that year.

From Figure 4.4, we first observe that over the out-of-sample periods, the cumulative wealth of these three FDR-controlled portfolios increases steadily, with only slight declines in 2018 and 2022. Moreover, these FDR-controlled portfolios generally generate higher returns than the Nikkei 225, indicating that the in-sample alphas successfully translate into out-of-sample economic value.

We also evaluate the Sortino ratios (Sortino and Price, 1994) to take risk into account, which are 1.90, 2.11, 1.38, and 1.90 for the portfolios with FDR controlled at 10%, 15%, 20%, and the Nikkei 225, respectively. This shows that FDR-controlled portfolios may still outperform the Nikkei 225 when risk is considered. However, this does not hold for all portfolios; for instance, the portfolio with FDR controlled at 20% performs worse than the Nikkei 225. This may be due to the higher tolerance for false discoveries, which increases the downside standard deviation and thereby affects its risk-adjusted performance.

It is interesting to note that, although a portfolio with a lower γ is expected to perform better due to a smaller proportion of false discoveries, Figure 4.4 shows that portfolios controlling the FDR at 15% and 20% consistently outperform the portfolio controlling the FDR at 10%. One possible explanation is that the in-sample alphas do not imply the out-of-sample returns; hence, the out-of-sample performance of the outperforming funds identified when $\gamma = 10\%$ does not always perform better compared to those only identified by increasing γ further. Another explanation is that while false discoveries have non-positive alphas, they may still generate positive returns out-of-sample (albeit with lower probability), leading to portfolios with higher γ generating more returns.

Finally, as shown in Table 4.4, while the portfolios constructed in 2022 includes a

doi:10.6342/NTU202404642

large number of outperforming mutual funds, their out-of-sample performance are negative. This suggests that a greater number of outperforming funds does not necessarily lead to superior out-of-sample performance in the Japanese FDR-controlled portfolios.

doi:10.6342/NTU202404642

Chapter 5 Conclusion

In this paper, we aim to examine mutual fund performance within a multiple testing framework. We begin by using alpha from a benchmark model as the performance measure for these mutual funds. To estimate these alphas accurately, we adopt the approach of Giglio and Xiu (2017), extending the Carhart four-factor model with latent factors to mitigate omitted variable bias. We then apply the matrix completion method used in Giglio, Liao, and Xiu (2021) to handle missing values in mutual fund return data and employ PCA to identify these latent factors. Finally, we implement the screening BH procedure to control luck and identify mutual funds with truly superior performance.

Our results indicate that Japanese mutual funds tend to outperform the benchmark in the short term, but their superior performance diminishes quickly, leaving only a small number of funds that maintain superior performance over the long term. Moreover, mutual funds with short-term superior performance tend to outperform others across all subperiods, though their performance is not strong enough to be classified as outperforming funds in the long term. Finally, we find that portfolios of outperforming mutual funds generate economic value and outperform the Nikkei 225, providing a profitable strategy for investors.

Interestingly, contrary to our expectations, our long-term and out-of-sample analyses

doi:10.6342/NTU202404642

indicate that when the FDR is 15%, the identified outperforming funds may achieve the highest long-term in-sample alpha and out-of-sample returns. This outcome may be due to mutual funds with non-positive alphas randomly generating positive out-of-sample returns or because the funds with best performance are not always identified when using a lower γ. Therefore, a possible direction for future research is to investigate whether these results arise from data noise or if there is an underlying pattern that can be captured by econometric methods, ultimately determining the optimal FDR threshold for control.

doi:10.6342/NTU202404642

References

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

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

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

Barras, L., Scaillet, O., and Wermers, R. (2019). Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? a reply. *Journal of Finance, Forthcoming, Swiss Finance Institute Research Paper*, (19-61).

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):405–416.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

doi:10.6342/NTU202404642

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Benussi, A., Pilotto, A., Premi, E., Libri, I., Giunta, M., Agosti, C., Alberici, A., Baldelli, E., Benini, M., Bonacina, S., et al. (2020). Clinical characteristics and outcomes of inpatients with neurologic disease and covid-19 in brescia, lombardy, italy. *Neurology*, 95(7):e910–e920.

Betts, M. G., Wolf, C., Ripple, W. J., Phalan, B., Millers, K. A., Duarte, A., Butchart, S. H., and Levi, T. (2017). Global forest loss disproportionately erodes biodiversity in intact landscapes. *Nature*, 547(7664):441–444.

Cai, J., Chan, K. C., and Yamada, T. (1997). The performance of Japanese mutual funds. *The Review of Financial Studies*, 10(2):237–274.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Chen, C.-W., Huang, C.-S., and Lai, H.-W. (2011). Data snooping on technical analysis: Evidence from the taiwan stock market. *Review of Pacific Basin Financial Markets and Policies*, 14(02):195–212.

Cuthbertson, K., Nitzsche, D., and O'Sullivan, N. (2010). Mutual fund performance: Measurement and evidence. *Financial Markets, Institutions & Instruments*, 19(2):95–187.

doi:10.6342/NTU202404642

Cuthbertson, K., Nitzsche, D., and O'Sullivan, N. (2012). False discoveries in UK mutual fund performance. *European Financial Management*, 18(3):444–463.

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

Giglio, S., Liao, Y., and Xiu, D. (2021). Thousands of alpha tests. *The Review of Financial Studies*, 34(7):3456–3496.

Giglio, S. and Xiu, D. (2021). Asset pricing with omitted factors. *Journal of Political Economy*, 129(7):1947–1990.

Goldfarb, D. and Ma, S. (2011). Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, 11(2):183–210.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380.

Harvey, C. R. and Liu, Y. (2022). Luck versus skill in the cross section of mutual fund returns: Reexamining the evidence. *The Journal of Finance*, 77(3):1921–1966.

Harvey, C. R., Liu, Y., and Saretto, A. (2020). An evaluation of alternative multiple testing methods for finance applications. *The Review of Asset Pricing Studies*, 10(2):199–248.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Hsu, P.-H., Hsu, Y.-C., and Kuan, C.-M. (2010). Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance*, 17(3):471–484.

doi:10.6342/NTU202404642

Hsu, P.-H. and Kuan, C.-M. (2005). Reexamining the profitability of technical analysis with data snooping checks. *Journal of Financial Econometrics*, 3(4):606–628.

Hsu, Y.-C., Kuan, C.-M., and Yen, M.-F. (2014). A generalized stepwise procedure with improved power for multiple inequalities testing. *Journal of Financial Econometrics*, 12(4):730–755.

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

Ma, S., Goldfarb, D., and Chen, L. (2011). Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1):321–353.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The annals of statistics*, 21(1):255–285.

Metghalchi, M., Chang, Y.-H., and Garza-Gomez, X. (2012). Technical analysis of the taiwanese stock market. *International Journal of Economics and Finance*, 4(1):90–102.

Pilbeam, K. and Preston, H. (2019). An empirical investigation of the performance of Japanese mutual funds: Skill or luck? *International Journal of Financial Studies*, 7(1):6.

Qi, M. and Wu, Y. (2006). Technical trading-rule profitability, data snooping, and reality check: Evidence from the foreign exchange market. *Journal of Money, Credit and Banking*, pages 2135–2158.

Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12).

doi:10.6342/NTU202404642

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008). Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2):404–447.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2010). Hypothesis testing in econometrics. *Annu. Rev. Econ.*, 2(1):75–104.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4):1378–1408.

Sortino, F. A. and Price, L. N. (1994). Performance measurement in a downside risk framework. *the Journal of Investing*, 3(3):59–64.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Sullivan, R., Timmermann, A., and White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The journal of Finance*, 54(5):1647–1691.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

doi:10.6342/NTU202404642