Superior or Spurious? Mutual Fund Alpha in Japan with Latent Factors and False Discovery Control

Yu-Hao Chang[a], Hui-Ching Chuang[b,*], Chung-Ming Kuan[a]

[a] *Department of Finance, National Taiwan University*
[b] *Department of Statistics, National Taipei University*

Abstract

This paper evaluates the performance of Japanese equity mutual funds over a period characterized by substantial changes in Japan's financial and macroeconomic environment. Using 1,483 funds from 2002 to 2023, we account for latent common factors and multiple-testing bias within a two-pass asset-pricing framework. Over the full sample, evidence of superior performance is limited: only 0.20% of funds are classified as outperformers when the false discovery rate (FDR) is controlled at levels from 5% to 20%. Over shorter subperiods, however, outperformance is more common, and fund performance rankings exhibit weak persistence. Portfolios formed from Benjamini–Hochberg-selected funds subsequently earn compound annual returns of 15.08%–15.61%, compared with 10.31% for TOPIX, and generate significant factor-adjusted monthly alphas of approximately 31 basis points. These portfolios also exhibit higher Sharpe and Sortino ratios and smaller maximum drawdowns than TOPIX.

Keywords: Mutual fund performance; Latent factors; Matrix completion; False discovery rate; Wild bootstrap
JEL: C12; C38; G11; G23

*Corresponding author: Hui-Ching Chuang (hcchuang@gm.ntpu.edu.tw). Address: Department of Statistics, National Taipei University, No. 151, University Rd., Sanxia Dist., New Taipei City 23741, Taiwan.
Email addresses: yuhaochang1212@gmail.com (Yu-Hao Chang), hcchuang@gm.ntpu.edu.tw (Hui-Ching Chuang), ckuan@ntu.edu.tw (Chung-Ming Kuan)

1. Introduction

The rapid expansion of Japan's investment-trust industry has made it increasingly important for household wealth accumulation and institutional portfolio management. The number of contractual-type publicly offered investment trusts more than doubled from 2,718 in 2002 to 5,913 in 2023, and their total assets increased from JPY 36.0 trillion to JPY 196.9 trillion.[1] Over the same period, Japan experienced some major changes in its monetary and household-investment policies. The Bank of Japan began purchasing ETFs in December 2010, introduced quantitative and qualitative monetary easing (QQE) in April 2013, and adopted a negative interest rate policy in January 2016 and yield curve control in September 2016.[2] On the household-investment side, the Japanese government launched tax-advantaged programs that changed investors' incentives to hold investment funds, including the Nippon Individual Savings Account (NISA) in 2014 and Tsumitate NISA in 2018.[3] It is then of interest to know how Japanese mutual fund performance has evolved in response to changing market environments.

We address this question by evaluating the performance of 1,483 actively managed Japanese equity mutual funds from January 2002 through December 2023. To study fund performance over time, two empirical issues arise. First, evaluating 1,483 funds simultaneously is a multiple-testing problem; conventional fund-by-fund inference is misleading because it may identify outperforming funds by chance. To this end, we apply the screening Benjamini and Hochberg (1995) procedure, which controls the false discovery rate (FDR), as suggested by Giglio et al. (2021). Second, to compute fund performance measures (alphas), conventional factor models may omit common variation in fund returns, causing estimated alphas to reflect unobserved factor exposures rather than managerial ability (Giglio and Xiu, 2021; Ardia et al., 2024). Following Giglio et al. (2021), we augment the observed-factor model with latent common factors. Because the fund-return panel is highly unbalanced, we apply matrix completion and principal component analysis to estimate the latent factor structure, and a two-pass asset-pricing procedure to estimate fund alphas.

In our empirical study, we first evaluate fund performance over the full sample and then examine shorter-run performance based on five-year rolling windows. The full-sample results provide scant

[1] Industry statistics are from JITA/IMAJ, *Factbook of Japanese Investment Trusts*, December 2018 edition, and December 2023 edition (retrieved September 8, 2026).

[2] For the monetary-policy chronology, see `https://www.boj.or.jp/en/mopo/outline/ref_qqe.htm`.

[3] The original NISA provided tax exemptions on returns from eligible investments, including mutual funds. The Tsumitate NISA further encouraged regular, long-term, and diversified investment through a restricted set of eligible investment products. In 2024, after the end of our sample period, these schemes were replaced by the new NISA, which combines a Tsumitate Investment Quota and a Growth Investment Quota. See `https://www.fsa.go.jp/policy/nisa2/know/` for details.

evidence of long-run superior performance. Over 2002–2023, only 0.20% of funds are selected as outperformers at the 5%, 10%, and 20% FDR levels, substantially below the share suggested by conventional fund-by-fund inference. This finding is consistent with earlier empirical evidence from Japan. Cai et al. (1997) document substantial underperformance of Japanese mutual funds during 1981–1992, while Brown et al. (2001) and Brown et al. (2003) show that Japan's tax and institutional arrangements materially affected measured fund performance. More recently, Pilbeam and Preston (2019) find little evidence of outperformance of Japanese mutual funds during 2011–2016. It is, however, not clear how fund performance evolves over a longer period spanning major changes in Japan's markets.

For the short-run analysis based on five-year rolling windows, we find that fund performance varies quite substantially over time. The shares of funds selected at the 5% FDR level remain low (at most 1.27%) in the first six windows, from 2002–2006 through 2007–2011. Yet, this share jumps to 71.58% in 2010–2014 and 41.44% in 2011–2015, falls to 2.14% in 2012–2016, and rebounds to 36.89% in 2013–2017. Selection signals are weaker in several subsequent windows but strengthen again in 2017–2021 and 2019–2023, where the selection shares are 10.65% and 10.79%, respectively. Although several windows with high selection shares coincide with the stock-market rise beginning in late 2012 and the introduction of QQE in 2013, the selection share drops sharply to 2.14% in 2012–2016, a period that also covers much of the same monetary-policy and market environment. Hence, there is no simple/direct relation between particular policy regimes and fund performance. Nevertheless, we find that rankings of fund performance exhibit weak persistence across adjacent windows.

We further examine whether the selected funds carry economic value out of sample. At each year-end, we select funds from the preceding ten-year data using the screening Benjamini–Hochberg procedure, form an equal-weighted portfolio, and evaluate its performance in the following calendar year. We consider a ten-year estimation window because it provides a more comprehensive return history for fund selection. We find that, from 2012 to 2023, these portfolios earn compound annual returns of 15.08%–15.61%, compared with 10.31% for TOPIX, with volatility similar to that of TOPIX. They also exhibit higher Sharpe and Sortino ratios and smaller maximum drawdowns. Relative to TOPIX, these portfolios generate significant monthly alphas of 39.08–41.08 basis points; these alphas remain significant at 30.75–31.26 basis points after controlling for the Japanese size, value, and momentum factors.

To summarize, this paper contributes to the mutual fund literature by providing updated evidence on the evolution of Japanese active fund performance over time. Compared with earlier

empirical evidence, e.g., Cai et al. (1997), we show that long-run superior performance is rare, while shorter-run performance varies substantially across rolling windows without a simple pattern tied to economic or institutional changes. Despite this short-run variation, fund performance rankings exhibit weak persistence. We also find that the portfolios formed by historically selected funds generate significant factor-adjusted alphas out of sample. We note that the screening Benjamini and Hochberg (1995) procedure and factor models with latent factors are two crucial empirical tools for establishing these empirical findings.

The remainder of the paper is organized as follows. Section 2 presents the screening Benjamini–Hochberg procedure and the methods for alpha estimation based on the observed- and latent-factor model. Section 3 describes the sample and reports the full-sample, rolling-sample, rank-persistence, and out-of-sample results. Section 4 concludes.

2. Methods for Evaluating Fund Performance

In this section we describe the econometric methods used for evaluating fund performance. We first discuss the screening Benjamini and Hochberg (1995) procedure that controls FDR. We then introduce the observed- and latent-factor model considered in Giglio et al. (2021) and its estimation strategy. Technical details of estimating the latent-factor model are deferred to Appendix A.

2.1. The Screening Benjamini–Hochberg Procedure

Let θ_i denote the performance measure of fund i relative to a benchmark. We want to test:

$$H_{i,0} : \theta_i \leq 0 \quad \text{against} \quad H_{i,A} : \theta_i > 0, \quad i = 1, \ldots, N.$$

Rejecting the null hypothesis provides evidence that fund i outperforms its benchmark. Testing whether the performance of a large number of individual funds exceeds zero is inherently a multiple-testing problem. Although each test may control its Type I error at the nominal level, the probability of obtaining false rejections increases rapidly as the number of tests grows. Consequently, some funds may appear to exhibit superior performance purely by chance, rather than because of genuine managerial skill.

One approach is to control the family-wise error rate (FWER), defined as the probability of making at least one false rejection among candidate hypotheses. The Bonferroni correction provides a simple way to control the FWER, while more powerful single-step and stepwise procedures have

been developed by White (2000), Hansen (2005), Romano and Wolf (2005), and Hsu et al. (2010).[4] However, tests that control FWER may be overly conservative when the number of hypotheses is large, resulting in low testing power to identify genuinely superior funds (Harvey et al., 2016).

An alternative is to control the FDR, the expected proportion of false rejections among all rejected hypotheses. Specifically,

$$\text{FDR} = \mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right],$$

where R is the number of rejected null hypotheses and V the number of false rejections. Thus, controlling FDR at the level γ limits the expected share of false selections among the funds classified as outperformers. It is common to set $\gamma \in \{5\%, 10\%, 20\%\}$; see, e.g., Barras et al. (2010) and Harvey et al. (2016). Note that lower γ values impose more stringent control over false discoveries, whereas higher γ values permit better power to detect superior funds. The choice of γ in effect determines a balance between limiting false discoveries and retaining power to detect genuine outperformance. This approach has been applied to large-scale testing problems in empirical asset pricing (e.g., Harvey et al., 2016; Chordia et al., 2020), and multiple-testing methods have been used to evaluate mutual fund performance (e.g., Barras et al., 2010; Tuzov and Viens, 2011; Harvey and Liu, 2022).

Let t_i denote the standardized test statistic for θ_i and p_i its one-sided p-value. Given the ordered p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$, define

$$k = \max\left\{i : p_{(i)} \leq \frac{i\gamma}{N}\right\}.$$

Benjamini and Hochberg (1995) propose to reject the hypotheses with p-values less than or equal to $p_{(k)}$; if no such k exists, no fund is selected. As the probability threshold above would be very small when N is very large, this procedure may not be powerful enough to reject any hypothesis. To improve test power, Giglio et al. (2021) suggest retaining funds that satisfy

$$t_i > -c\log(\log T)\sqrt{\log N}, \tag{1}$$

where c is a positive constant, N is the number of funds, and T is the time-series dimension.[5]

[4] FWER-based multiple-testing procedures have been widely applied in finance, including tests of trading-strategy profitability (e.g., Hsu and Kuan, 2005; Kuang et al., 2014; Goyal and Wahal, 2015; Hsu et al., 2016; Chuang et al., 2024) and comparisons of forecasting and volatility models (e.g., Hansen and Lunde, 2005; Bao et al., 2006; Liu et al., 2015). Related extensions and applications include procedures controlling the false discovery proportion and k-FWER; see Romano et al. (2008); Chuang and Kuan (2020).

[5] In our application, we set $c = 1/3$, as in the supplementary code accompanying Giglio et al. (2021). We assess

Because the cut-off value in (1) is negative, the screening step removes funds with very negative statistics that are unlikely to be identified as outperformers. Let N' denote the number of retained funds and

$$k^* = \max\left\{ i : p_{(i)} \leq \frac{i\gamma}{N'} \right\}.$$

The screening Benjamini–Hochberg procedure then rejects the hypotheses with $p_{(1)}, \ldots, p_{(k^*)}$. As $N' \leq N$, the cut-off value for p_i is at least as large, so that $k^* \geq k$. That is, this procedure tends to reject more hypotheses than does the original Benjamini and Hochberg (1995) procedure.

2.2. Estimation of the Observed- and Latent-Factor Model

In our empirical application, the performance measure is fund-specific alpha from an asset-pricing factor model. To mitigate concerns about omitted factors, we follow the latent-factor framework of Giglio et al. (2021) and evaluate alpha relative to both observed and latent common return components. For fund i in month t, we specify

$$r_{it} = \alpha_i + f'_{o,t}\beta_{o,i} + f'_{\ell,t}\beta_{\ell,i} + \varepsilon_{it}, \tag{2}$$

where r_{it} is the fund's excess return, $f_{o,t}$ contains the observed factors, $f_{\ell,t}$ contains the latent factors, and α_i is the fund-specific abnormal return. We adopt the four-factor model of Carhart (1997), which includes the market, size, value, and momentum factors, and augment it with latent factors that capture common return variation not explained by the observed factors. Note that including latent factors provides a conservative benchmark for fund performance. If skilled managers follow similar strategies, part of their shared performance may be captured by the estimated latent factors rather than appearing in α_i. Thus, α_i in Equation (2) measures fund-specific abnormal performance relative to both the observed factors and estimated latent common components.

We first estimate the observed-factor model separately for each fund and remove the fitted observed-factor component from its return. Because funds are observed over different sets of months, the resulting residual panel is unbalanced, and the funds do not share a common time dimension T. To accommodate these unequal return histories, we use nuclear-norm *matrix completion* to recover the low-rank common structure from the observed residuals across funds and months. We then apply principal component analysis to the completed residual matrix and retain the leading components as latent factors. For each fund, we estimate its exposure to the latent

its sensitivity using $c \in \{1, 3/4, 2/3, 1/2, 1/3, 1/4\}$. We find that the selected share remains 0.20% at every value and FDR level, indicating that the baseline screening threshold does not drive our main conclusion.

factors using only the months in which that fund is observed, and combine the observed- and latent-factor loadings in the subsequent two-pass alpha estimation. The scree plot and explained variance ratios provide the empirical basis for selecting the number of latent factors.

Finally, we use the (de-biased) two-pass estimator of Giglio et al. (2021) to estimate α_i. In the first pass, we estimate each fund's exposures to the observed and latent factors using its available monthly returns. In the second pass, we regress average fund returns on the estimated factor loadings to estimate factor risk premia and obtain alpha as the component of average return not explained by these exposures. Because the second-pass regression includes estimated factor loadings as regressors, particularly those associated with the latent factors, the resulting alpha estimates may be affected by estimation error. To circumvent this problem, we apply the de-biasing adjustment proposed in Giglio et al. (2021). See Appendix A for details on matrix completion, estimation of latent-factor loadings, and the two-pass estimator.

3. Japan's Mutual Fund Performance

3.1. Sample and Related Fund Information

We obtain the Japanese fund data from Bloomberg's Fund Screening (FSRC) function. We consider open-end equity funds domiciled in Japan and available for sale in Japan. We mitigate survivorship bias by including active as well as inactive, delisted, liquidated, acquired, and suspended funds. This sample is therefore not restricted to surviving funds, which reduces the concern that unsuccessful funds disappear from the analysis.[6] Moreover, we focus on actively managed funds and exclude products whose names contain "index," "idx," "ETF," or "tracker fund." Finally, following Kosowski et al. (2006), we require each fund to have at least 60 months of return observations over 2002–2023 to be included in our sample.[7] When multiple share classes are available for the same fund, we retain only the share class designated by Bloomberg as the primary share class; this avoids treating the funds classified into different share classes as separate observations. These lead to a sample of 1,483 funds from 2002 to 2023.

We compute monthly fund returns from Bloomberg-adjusted closing prices, which incorporate recorded distributions and corporate actions such as dividends, splits, spin-offs, and rights offer-

[6]The sample-screening procedure is as follows. First, using Bloomberg's Fund Screening (FSRC) function, we retain funds satisfying: (i) Fund Type = Open-End Funds; (ii) Fund Asset Class Focus = Equity; (iii) Country/Territory of Domicile = Japan; (iv) Country/Territory of Availability = Japan; and (v) Fund Primary Share Class = Yes.

[7]In the short-run analysis, we estimate models based on the data in five-year windows and further require funds in our sample to have at least 12 monthly observations within the window.

ings. We do not reconstruct returns before operating expenses, nor do we observe investor-specific sales loads or trading costs. As a result, the estimated alpha should not be interpreted as gross managerial value added or the realized return earned by every investor after all implementation costs. Rather, it measures abnormal performance reflected in published adjusted fund prices.

Figure 1 summarizes the evolution of our empirical sample. The average monthly number of eligible funds rises from 428 in 2002 to a peak of 1,069 in 2018, then declines to 830 in 2023. This pattern parallels developments in the wider Japanese investment-trust industry. JITA's industry-wide series shows that the number of contractual-type publicly offered investment trusts rose from 3,333 in 2008 to 5,843 in 2015, peaked at 6,152 in 2017, and subsequently declined to 5,913 in 2023.[8] The similar long-run pattern suggests that our Bloomberg sample is in line with broader industry dynamics. Figure 1 also shows substantial variation in average monthly fund returns over time. These returns fall sharply in 2008 during the global financial crisis and rebound strongly in 2013. Notably, average returns remain positive throughout the core Abenomics period, 2013–2017, but decline sharply in 2018.

Table 1 summarizes fund-level returns over the full sample and 18 overlapping five-year windows used in the short-run performance analysis. Within each five-year window, we retain funds with at least 12 observed monthly returns and report the cross-sectional distribution of their average monthly returns, calculated over each fund's observed months. The first window, 2002–2006, contains 562 eligible funds, and the final window, 2019–2023, contains 1,029, consistent with Figure 1. The mean fund-level return falls from 1.08% in 2002–2006 to −0.91% in 2007–2011, rises to a peak of 1.29% in 2013–2017, and declines to 0.37% in 2018–2022. The largest fund-level average is 7.87% in 2012–2016.

We also note that there is a substantial proportion of missing observations in the fund-month return matrix, as shown in the last column of Table 1. The proportion of unobserved entries in the return matrix is defined as $100[1 - O/(NT)]$, where N is the number of eligible funds, T is the number of months in the estimation period, and O is the number of observed returns. The full 2002–2023 sample contains 208,109 observed monthly returns for 1,483 funds over 264 months; thus, 46.84% of the entries are missing. For rolling five-year windows, the share of missing observations ranges from 8.08% to 20.84%. We deal with this missing-data problem by invoking the matrix-completion procedure of Giglio et al. (2021).

[8]The fund-count series and the industry asset figures reported in the Introduction are from JITA/IMAJ, *Factbook of Japanese Investment Trusts*, December 2018 and December 2023 editions.

3.2. Factor Alpha Analysis

To account for observed systematic risk, we adopt the Carhart (1997) four-factor model, which is widely used in mutual fund performance evaluation (Kosowski et al., 2006; Fama and French, 2010).[9] To capture common variation not explained by the four-factor model, we examine the eigenvalues of the residual covariance matrix from this model. Figure 2 reports the 15 largest eigenvalues as shares of total residual variance. The number of latent factors is determined using three criteria: (i) cumulative explained variance exceeding 60%; (ii) marginal explained variance exceeding 3%; and (iii) a clear elbow in the scree plot. These criteria support retaining five latent factors.[10] Our full-sample performance model is thus based on the four-factor model with five estimated latent factors.

Because the latent factors are obtained from estimation, rather than constructed as traded portfolios, they do not have an immediate economic interpretation. To examine their economic content, we calculate the correlations between latent factors and a set of Japanese macroeconomic variables.[11] Table 2 reports the correlations of 5 latent factors with these variables. Apart from the strong association between the first latent factor and the yen/dollar exchange rate, the latent factors show only weak correlations with individual macroeconomic series. Specifically, the first latent factor correlates with several variables, with the strongest correlation with the yen and U.S. dollar exchange rate (-0.83) and the second strongest with household financial assets (-0.36). Each of the other latent factors also has some weaker correlations with several economic variables. This shows that latent factors carry mixed economic content.

Figure 3 and Table 3 report the distributions of estimated alphas and corresponding t-statistics from our baseline model. The de-biased alphas are centered below zero, with a mean of -2.62 basis points (bps) per month, a median of -2.36 bps, and an interquartile range from -14.96 to 10.55 bps. The mean alpha of -2.62 bps is close to zero relative to the mean monthly return

[9]These factors are the Japanese three factors and the momentum factor. These factors and the risk-free rates are obtained from Kenneth French's Data Library: `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

[10]The five latent factor choice pertains to the full 2002–2023 sample. For the short-run analysis with five-year windows, we allow the number of latent factors to vary across windows by reapplying the same factor-selection rule separately within each period.

[11]The CPI is taken from the Statistics Bureau of Japan; the ten-year government bond yield from the OECD; the basic discount and loan rate, the yen and U.S. dollar exchange rate at month-end, broad money, and household financial assets from the BOJ Time-Series Data Search; industrial production from METI; and GDP from the Cabinet Office/e-Stat National Accounts. Broad money comprises currency in circulation, current deposits, and other deposits, including relatively less liquid forms such as savings and time deposits. CPI, the exchange rate, broad money, industrial production, household financial assets, and GDP are expressed as log differences, while the ten-year government bond yield and the basic discount and loan rate are measured in levels. Monthly series are matched by calendar month, and quarterly GDP observations are matched to the corresponding quarter-end month.

of 0.58% reported in Table 1, so average returns largely reflect exposure to observed and latent common factors rather than abnormal performance. Further, both distributions have long tails: alphas range from −115.87 to 108.03 bps per month, and t-statistics range from −7.45 to 6.51.

3.3. Selection of Outperformers in the Full Sample

To select outperformers from a set of 1,483 funds, we need a test that properly addresses the multiple-testing problem, as discussed in Section 2.1. From the distribution of estimated alphas, funds in the far left tail are unlikely to belong to the positive-alpha group we want to identify. We therefore apply the screening Benjamini–Hochberg procedure that controls the FDR (Giglio et al., 2021). For each of the 1,483 funds, we use the de-biased alpha t-statistic as the test statistic and obtain a one-sided p-value from 1,000 wild-bootstrap replications. With $N = 1{,}483$, $T = 264$, and the baseline $c = 1/3$, the screening threshold in Equation (1) is approximately -1.55. Thus, funds with $t_i \leq -1.55$ are removed before the step-up rule is applied to the remaining tests at $\gamma \in \{5\%, 10\%, 20\%\}$.

Table 4 reports the full-sample results for identified outperformers during 2002–2023. At each FDR level, only 0.20% of funds (3 funds) are selected, with an average alpha of 59.06 basis points per month. These results indicate that persistent outperformance is rare, regardless of FDR levels. This 0.20% share is substantially below the share of funds whose conventional t-statistics exceed 1.645 (8.02%). That is, many funds that appear significant under individual fund-by-fund tests are no longer classified as outperformers. The full-sample results, however, need not imply that opportunities for active management are absent throughout the sample. In fact, fund performance may vary in different periods.

3.4. Fund Performance Across Different Market Environments

We now examine the short-run performance of funds across different market environments. We apply the same screening Benjamini–Hochberg procedure to 18 overlapping five-year windows, advancing by one year at a time from 2002–2006 through 2019–2023. Table 5 reports the mean alpha of all eligible funds (those with at least 12 monthly observations), the share of selected outperforming funds, and their mean alpha at FDR levels $\gamma \in \{5\%, 10\%, 20\%\}$.

It can be seen from Table 5 that fund performance exhibits substantial time variation across these windows. In the first 6 windows (the last one ending in 2011), which span Japan's low-growth years, the global financial crisis, and the 2011 Great East Japan Earthquake, the shares of selected outperforming funds at 5% and 10% FDR levels are very low: 5 out of 6 shares are less than 0.5%,

and the other one is higher but still less than 1.5% (2003–2007). The corresponding mean alphas of all funds are negative, except in the first window (2002–2006). At 5% FDR level, the selection share starts increasing from the window 2008–2012 and jumps to 71.58% in 2010–2014 and 41.44% in 2011–2015, with selected-fund mean alphas 100.57 and 82.37 basis points per month, respectively. This share then falls sharply to 2.14% in 2012–2016 and rebounds back to 36.89% in 2013–2017, with the selected-fund mean alpha 85.68 basis points. In the last 6 windows, the selection share remains very low in 4 out of 6 windows and is higher in the other two. This variation pattern also holds for 10% and 20% FDR levels.

Note that strong fund performance (high selection share) appears during the core Abenomics period, during which we observe aggressive monetary easing, yen depreciation, the introduction of NISA, and the Bank of Japan's Negative Interest Rate Policy. Yet, such strong performance does not persist. Due to large variation in selection shares across different market environments, we fail to find a simple relation between particular policy regimes and superior fund performance.

3.5. Persistence of Performance Ranks

Although high selection shares of funds do not persist across different periods, we now examine whether the rank of funds exhibits some persistence. To mitigate the concern that overlapping observations in adjacent five-year windows may mechanically increase measured persistence,[12] we consider seven five-year windows: 2002–2006, 2005–2009, 2008–2012, 2011–2015, 2014–2018, 2017–2021, and 2019–2023, which overlap by two years, except the last two windows (which overlap by three years). For each window, we retain funds with at least 12 monthly observations and compute one-sided wild-bootstrap p-values for the alpha t-statistics using the same estimation procedure as in the preceding sub-section. We then compare each window with the next one. For each of the six pairs of consecutive windows, we keep the funds observed in both windows and, separately for each of the two windows, sort these funds by their p-values into ten equal-sized deciles, so that decile 1 contains the funds with the smallest p-values, that is, the best performance. The deciles of the current window are further grouped into three categories: High (deciles 1–3), Medium (4–7), and Low (8–10). For each pair, we then estimate the rank-transition probabilities

$$\Pr\big(\text{Current rank category} = j \mid \text{Previous rank} = i\big)\,,$$

[12] We thank an anonymous referee for raising this concern.

where $j \in \{\text{High}, \text{Medium}, \text{Low}\}$ denotes the current-window category and $i = 1, \ldots, 10$ denotes the decile in the previous window. Each probability is estimated as the proportion of funds in decile i of the previous window that fall into category j in the current window. Table 6 reports the simple average of these probabilities over the six pairs, so each column sums to 100%, and Figure 4 presents the same results in a three-dimensional bar plot. If performance ranks were unrelated across windows, a fund would fall into the High, Medium, and Low categories with probabilities of 30%, 40%, and 30%, respectively, regardless of its previous decile.

These results reveal some relations between more extreme ranks. The probability of entering the High group in the next period is 55.34% for funds previously in the first decile, whereas this probability is only 8.11% for those previously in the tenth decile. Similarly, the probability of entering the Low group in the next period is only 11.86% for funds previously in the first decile, whereas this probability is 61.52% for those previously in the 10th decile. For funds in the 2nd and 3rd (8th and 9th) deciles, it is more likely that they will be in the High or Medium (Low or Medium) group in the next period. For funds in the middle deciles, they will be more evenly distributed across performance groups in the next period. The relations between more extreme ranks may be understood as weak persistence in ranks. As a robustness check, we also consider four fully non-overlapping periods and find similar (but weaker) relations between extreme ranks.

3.6. Out-of-Sample Performance

We now examine whether the selected funds carry economic value beyond the estimation sample. At each year-end, we estimate fund alphas and their p-values using the data in the preceding ten years and apply the screening Benjamini–Hochberg procedure to select superior funds. We then form an equally weighted portfolio of the selected funds and evaluate its monthly returns in the following calendar year. We compute the monthly portfolio returns using the observed returns in that month. The Tokyo Stock Price Index (TOPIX) is used as the market benchmark, because its broad coverage represents the Japanese equity opportunity set. TOPIX is a free-float-adjusted, market-capitalization-weighted price-return index obtained from Investing.com.

Table 7 reports the number of funds identified during each out-of-sample year. Portfolio size increases with the FDR level and varies substantially over time. At 5% FDR level, the number of selected funds ranges from 3 in 2014 to 259 in 2022; at 20% FDR level, it ranges from 3 to 540. Selection is generally more extensive for holding years whose preceding ten-year estimation windows include much of the core Abenomics period, particularly 2016–2019 and 2022. This pattern echoes the rolling-window evidence in Table 5 and indicates that both portfolio composition and

diversification vary across years.

Table 8 examines whether conventional factor exposures explain the subsequent returns of the selected portfolios. Relative to TOPIX alone, the portfolios (selected from 3 FDR levels) generate monthly alphas of 39.08–41.08 basis points per month, with t-statistics ranging from 3.146 to 3.667. After controlling for the Japanese size, value, and momentum factors, these alphas reduce to 30.75–31.26 basis points per month, but their t-statistics are somewhat similar (between 3.284 and 3.776). All 6 alpha estimates are statistically significant at the 1% level. The inference is based on the Newey–West standard errors with 12 monthly lags for accommodating dependence within the annual holding cycle; the conclusions remain unchanged when 4 or 6 lags are used.

Figure 5 plots the cumulative wealth of four portfolios over 2012–2023, and Table 9 summarizes their returns and risk measures. These portfolios are equally-weighted portfolios with funds selected from 3 FDR levels. In terms of cumulative wealth, it can be seen that the selected portfolios perform much better than TOPIX. Moreover, the selected portfolios earn compound annual returns of 15.08%–15.61%, compared with 10.31% for TOPIX. Their annualized volatilities, ranging from 15.80% to 16.40%, are similar to (or slightly higher than) the 15.71% volatility of TOPIX. Nevertheless, their Sharpe ratios of 0.908–0.915 and Sortino ratios of 1.522–1.533 exceed the corresponding TOPIX values of 0.643 and 0.997. Their maximum drawdowns of 20.56%–22.29% are also smaller than the 25.56% drawdown of TOPIX.

Note that out-of-sample performance is not monotonic in the FDR level. A stricter FDR level reduces the expected proportion of false selections and leads to a smaller and potentially less diversified portfolio. In contrast, a loose FDR level permits more selected funds and more diversified portfolios. Performance differences across the three portfolios reflect not only the difference in selected funds but also diversification of funds. Moreover, TOPIX is measured using a price-return index, whereas the portfolio returns are based on adjusted fund prices and do not deduct trading costs or investor-specific sales loads. These out-of-sample results suggest that proper fund selection helps improve subsequent fund performance and generate higher returns.

4. Conclusions

This paper evaluates 1,483 actively managed Japanese equity mutual funds over 2002–2023 using the screening Benjamini–Hochberg procedure with a latent-factor structure. After accounting for latent common variation and multiple testing, we find that long-run outperforming funds are rare: only 0.20% across FDR levels from 5% to 20%. Our short-run analysis shows that the

incidence of superior fund performance varies substantially over time and is not persistent. Such performance is not necessarily related to changes in Japan's macroeconomic and institutional environments. Nevertheless, ranking of funds is found to exhibit weak persistence across adjacent windows. We also show that the selected funds carry economic value out of sample: portfolios formed from rolling historical selections would accumulate greater wealth than does TOPIX, exhibit higher Sharpe and Sortino ratios, experience smaller maximum drawdowns, and generate significant factor-adjusted alphas. This suggests that fund selection based on the methods adopted in this study may be useful for practitioners.

References

Ardia, D., Barras, L., Gagliardini, P., Scaillet, O., 2024. Is it alpha or beta? Decomposing hedge fund returns when models are misspecified. Journal of Financial Economics 154, 103805.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.

Bao, Y., Lee, T.H., Saltoglu, B., 2006. Evaluating predictive performance of Value-at-Risk models in emerging markets: A reality check. Journal of Forecasting 25, 101–128.

Barras, L., Scaillet, O., Wermers, R., 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. The Journal of Finance 65, 179–216.

Benjamini, Y., 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, 289–300.

Brown, S.J., Goetzmann, W.N., Hiraki, T., Otsuki, T., Shiraishi, N., 2001. The Japanese open-end fund puzzle. The Journal of Business 74, 59–77. doi:10.1086/209663.

Brown, S.J., Goetzmann, W.N., Hiraki, T., Shiraishi, N., 2003. An analysis of the relative performance of Japanese and foreign money management. Pacific-Basin Finance Journal 11, 393–412. doi:10.1016/S0927-538X(03)00046-5.

Cai, J., Chan, K.C., Yamada, T., 1997. The performance of Japanese mutual funds. The Review of Financial Studies 10, 237–273.

Carhart, M.M., 1997. On persistence in mutual fund performance. The Journal of Finance 52, 57–82.

Chordia, T., Goyal, A., Saretto, A., 2020. Anomalies and false rejections. The Review of Financial Studies 33, 2134–2179.

Chuang, H.C., Kuan, C.M., 2020. Identifying and assessing superior mutual funds: An application of the new step-wise data-snooping-bias-free test. Review of Securities & Futures Markets 32, 1–32.

Chuang, O.C., Chuang, H.C., Wang, Z., Xu, J., 2024. Profitability of technical trading rules in the chinese stock market. Pacific-Basin Finance Journal 84, 102278.

Fama, E.F., French, K.R., 2010. Luck versus skill in the cross-section of mutual fund returns. The Journal of Finance 65, 1915–1947.

Giglio, S., Liao, Y., Xiu, D., 2021. Thousands of alpha tests. The Review of Financial Studies 34, 3456–3496.

Giglio, S., Xiu, D., 2021. Asset pricing with omitted factors. Journal of Political Economy 129, 1947–1990.

Goyal, A., Wahal, S., 2015. Is momentum an echo? Journal of Financial and Quantitative Analysis 50, 1237–1267.

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

Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: Does anything beat a GARCH (1, 1)? Journal of Applied Econometrics 20, 873–889.

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

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

Hsu, P.H., Hsu, Y.C., 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, 471–484.

Hsu, P.H., Kuan, C.M., 2005. Reexamining the profitability of technical analysis with data snooping checks. Journal of Financial Econometrics 3, 606–628.

Hsu, P.H., Taylor, M.P., Wang, Z., 2016. Technical trading: Is it still beating the foreign exchange market? Journal of International Economics 102, 188–208.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer Series in Statistics. 2 ed., Springer, New York.

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

Kuang, P., Schröder, M., Wang, Q., 2014. Illusory profitability of technical analysis in emerging foreign exchange markets. International Journal of Forecasting 30, 192–205.

Liu, L.Y., Patton, A.J., Sheppard, K., 2015. Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. Journal of Econometrics 187, 293–311.

Mammen, E., 1993. Bootstrap and wild bootstrap for high dimensional linear models. The Annals of Statistics 21, 255–285.

Pilbeam, K., Preston, H., 2019. An empirical investigation of the performance of Japanese mutual funds: skill or luck? International Journal of Financial Studies 7, 6.

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

Romano, J.P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73, 1237–1282.

Satopää, V., Albrecht, J., Irwin, D., Raghavan, B., 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior, in: 2011 31st International Conference on Distributed Computing Systems Workshops (ICDCSW), IEEE. pp. 166–171.

Tuzov, N., Viens, F., 2011. Mutual fund performance: False discoveries, bias, and power. Annals of Finance 7, 137–169.

White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.

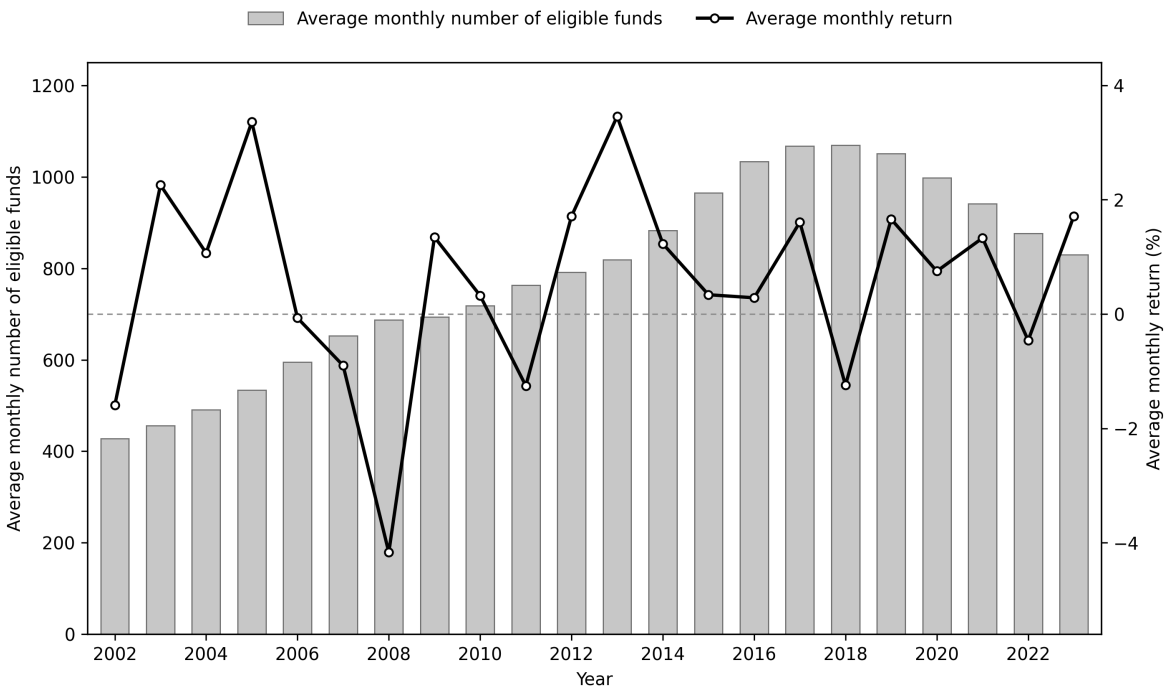


Figure 1: Number of sample funds and average return. The bars show the average monthly number of Japanese equity mutual funds in our sample in each year. For each month, we first calculate the cross-sectional mean return across eligible funds; the solid line shows the average of these 12 monthly means within each year.

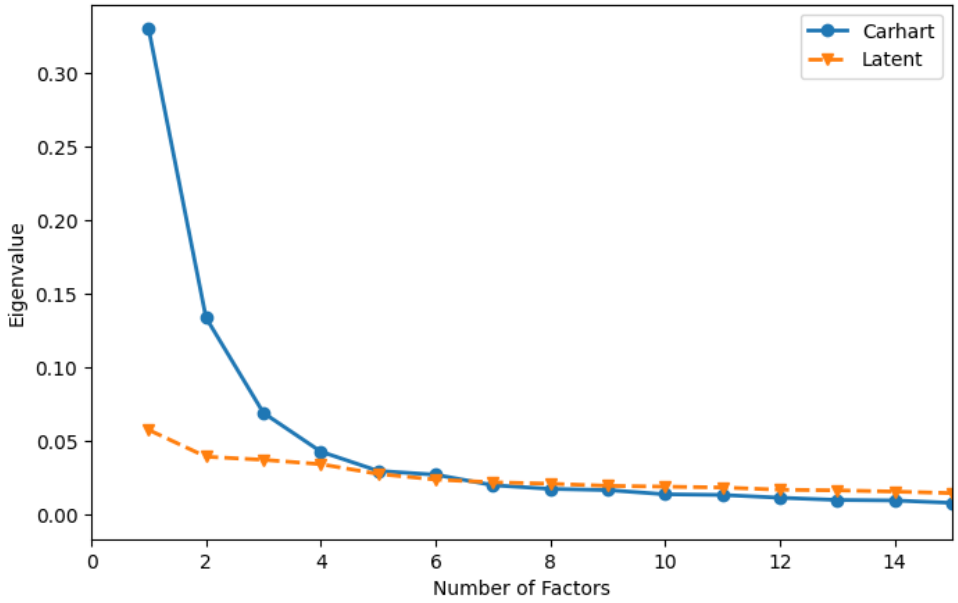
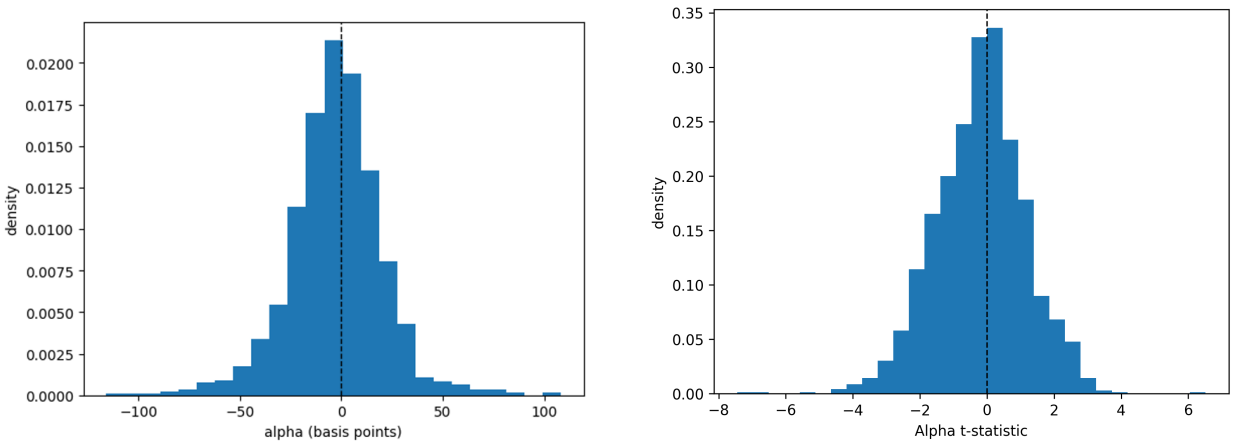


Figure 2: Scree plots before and after latent-factor augmentation. The figure plots the leading eigenvalues, expressed as shares of total variance, of the observed-factor residual covariance matrix and of the residual covariance matrix after adding latent factors. The knee point and the subsequent flattening of the eigenvalue sequence support retaining five latent factors.



(a) Estimated alpha

(b) Alpha t-statistic

Figure 3: Distributions of de-biased monthly alphas and corresponding t-statistics for 1,483 Japanese equity mutual funds over 2002–2023. Estimates are obtained from the Carhart four-factor model augmented with five latent factors. Alphas are reported in basis points per month; t-statistics are unitless.

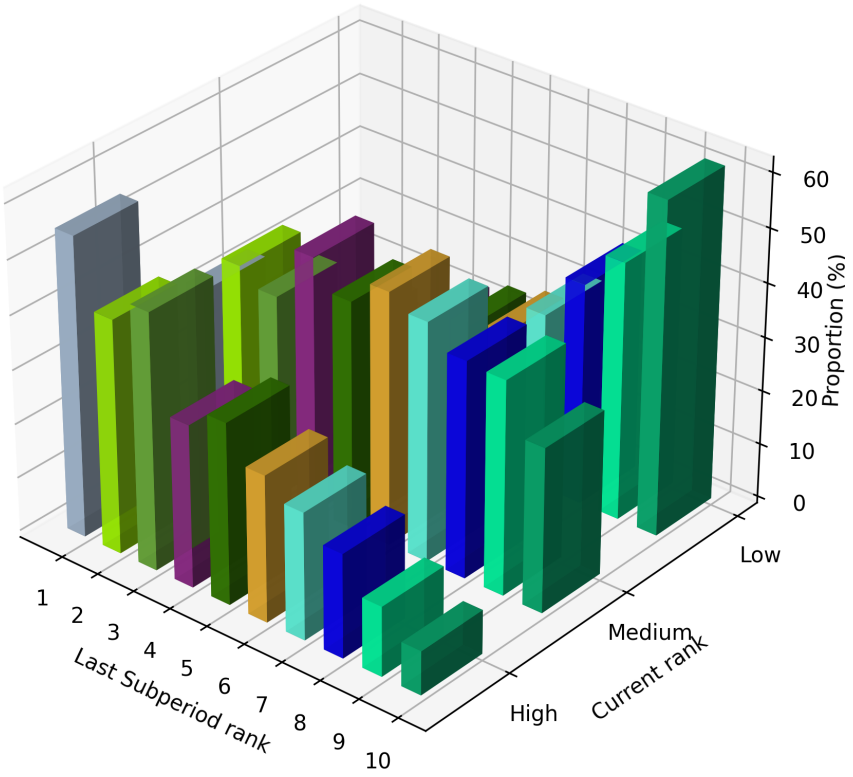


Figure 4: Performance-rank transitions. The three-dimensional bars report transition probabilities averaged across the six consecutive pairs formed from seven five-year windows spanning 2002–2006 to 2019–2023. Adjacent windows overlap by two years, except for the final pair, which overlaps by three years. The horizontal axis gives the previous-window performance decile, with 1 denoting the best-performing funds, while the depth axis groups current-window performance into High (deciles 1–3), Medium (4–7), and Low (8–10).

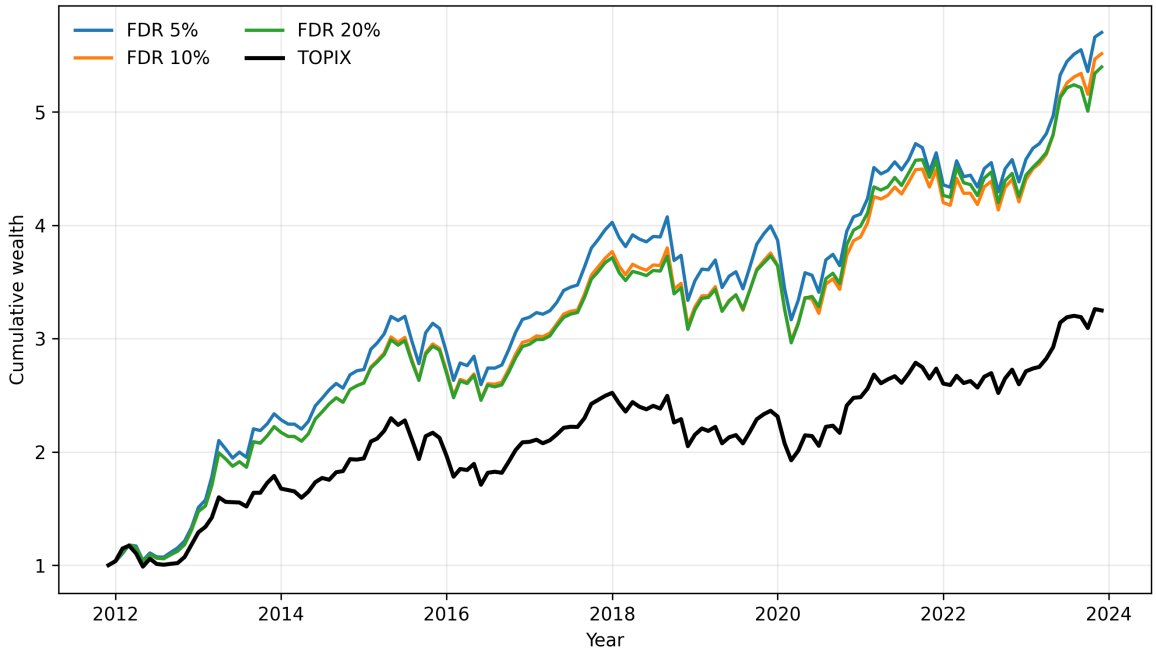


Figure 5: Cumulative wealth of Benjamini–Hochberg-selected portfolios and the TOPIX price-return index, 2012–2023. At each year-end, funds are selected based on the preceding ten years. During the following year, the monthly portfolio return is the equal-weighted mean of the raw returns of selected funds observed in that month. Wealth is normalized to one at the end of December 2011.

Table 1: Summary statistics of returns

Time Periods	N	Mean	SD	Min	P25	Median	P75	Max	Missing
2002 – 2006	562	1.08	0.58	−2.77	0.79	1.03	1.38	3.41	12.10
2003 – 2007	629	0.89	0.76	−3.15	0.71	1.04	1.25	3.98	14.28
2004 – 2008	693	−0.66	1.28	−6.47	−0.62	−0.25	−0.04	1.78	15.18
2005 – 2009	731	−0.32	0.70	−2.64	−0.65	−0.19	0.02	2.71	14.25
2006 – 2010	808	−0.60	0.80	−3.46	−1.02	−0.74	−0.36	2.86	17.80
2007 – 2011	860	−0.91	0.70	−5.62	−1.24	−1.00	−0.64	1.77	19.16
2008 – 2012	906	−0.38	0.98	−8.53	−0.70	−0.42	0.00	3.75	19.95
2009 – 2013	893	1.18	0.68	−1.25	0.90	1.11	1.41	5.04	16.77
2010 – 2014	964	1.17	0.68	−1.16	0.89	1.09	1.50	4.25	18.78
2011 – 2015	1040	1.03	0.73	−3.02	0.84	1.15	1.39	3.96	20.01
2012 – 2016	1122	1.24	0.82	−1.36	0.76	1.36	1.64	7.87	20.84
2013 – 2017	1152	1.29	0.74	−0.78	0.82	1.36	1.66	6.30	18.16
2014 – 2018	1169	0.43	0.46	−3.89	0.23	0.44	0.63	5.27	14.76
2015 – 2019	1195	0.52	0.45	−1.88	0.30	0.53	0.74	2.94	13.32
2016 – 2020	1161	0.62	0.54	−2.04	0.29	0.54	0.85	3.37	10.35
2017 – 2021	1117	0.79	0.52	−1.97	0.53	0.75	1.00	4.02	8.55
2018 – 2022	1069	0.37	0.45	−2.23	0.19	0.36	0.60	2.12	8.08
2019 – 2023	1029	0.99	0.53	−2.44	0.73	1.02	1.22	4.32	9.16
Full Sample	1483	0.58	0.44	−1.71	0.35	0.58	0.83	4.51	46.84

Notes: This table reports summary statistics for the full sample period and each five-year subperiod. Subperiods advance by one year, from 2002–2006 through 2019–2023. The second column shows the number of funds in each period (N). For each period, we first calculate each eligible fund's average return over its observed months. The table reports the cross-sectional distribution of these fund-level average monthly returns. The subsequent columns present the mean, standard deviation (SD), minimum, 25th percentile (P25), Median, 75th percentile (P75), and maximum. The last column reports the proportion of missing observations. A fund is eligible in a subperiod if it has at least 12 monthly observations within that subperiod. All return statistics and missing-data proportions are expressed as percentages (%).

Table 2: Correlations between latent factors and Japanese macroeconomic variables

Variable	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5
CPI	-0.03	0.01	0.09	-0.03	-0.06
Ten-year government bond yield	0.03	0.06	-0.02	0.03	-0.02
Basic discount and loan rate	0.09	-0.07	-0.09	0.02	0.1
Yen per U.S. dollar, month end	-0.83	-0.25	0.05	0.05	0.02
Broad money	-0.16	0.02	-0.01	0.05	-0.07
Industrial production (mining and manufacturing)	-0.06	-0.08	0.05	0.01	0.08
Household financial assets	-0.36	-0.06	0.01	0.02	-0.03
GDP	-0.16	0.07	0.13	-0.06	0.07

Notes: We obtain CPI from the Statistics Bureau of Japan, the ten-year government bond yield from the OECD, the basic discount and loan rate, the yen per U.S. dollar exchange rate at month-end, broad money, and household financial assets from the BOJ Time-Series Data Search, industrial production from METI, and GDP from the Cabinet Office/e-Stat National Accounts. Broad money comprises currency in circulation, current deposits, and other deposits, including relatively less liquid forms such as savings and time deposits. We express CPI, the exchange rate, broad money, industrial production, household financial assets, and GDP as log differences. We also measure the ten-year government bond yield and the basic discount and loan rate in levels. We match monthly series by calendar month and quarterly series to quarter-end months. The sample period is 2002–2023.

Table 3: Distribution of full-sample alpha and t-statistics

	Mean	SD	Min	P25	Median	P75	Max
Alpha (bps/month)	-2.62	23.00	-115.87	-14.96	-2.36	10.55	108.03
t-statistic	-0.21	1.36	-7.45	-1.06	-0.15	0.64	6.51

Notes: This table reports the distribution of de-biased fund alphas and corresponding t-statistics for 1,483 Japanese equity mutual funds over 2002–2023. Alphas are estimated using the Carhart four-factor model augmented with five latent factors and are reported in basis points per month. The t-statistics are the standardized statistics used to construct the one-sided wild-bootstrap p-values.

Table 4: Long-term performance

	FDR (γ)		
	5%	10%	20%
Proportion of outperforming funds (%)	0.20	0.20	0.20
Average alpha (bps/month)	59.06	59.06	59.06

Notes: This table reports the proportion of funds selected as outperformers at different FDR control levels. The evaluation period is 2002–2023, and the sample includes 1,483 Japanese equity mutual funds. The second row reports the average estimated alpha among the selected funds.

Table 5: Rolling five-year screening FDR results

Period	All-fund mean α	$\gamma = 5\%$		$\gamma = 10\%$		$\gamma = 20\%$	
		Share (%)	Selected-fund mean α	Share (%)	Selected-fund mean α	Share (%)	Selected-fund mean α
2002–2006	9.47	0.36	176.79	0.36	176.79	0.36	176.79
2003–2007	-15.70	1.27	92.07	1.43	95.69	1.91	103.42
2004–2008	-42.97	0.00	–	0.00	–	0.00	–
2005–2009	-24.56	0.27	94.41	0.27	94.41	0.41	87.70
2006–2010	-4.62	0.25	94.88	0.25	94.88	0.62	103.75
2007–2011	-27.83	0.23	157.81	0.23	157.81	0.35	161.91
2008–2012	5.70	2.21	98.71	6.51	76.93	9.82	67.68
2009–2013	50.17	5.94	98.50	17.92	86.30	49.50	72.03
2010–2014	89.39	71.58	100.57	79.05	101.91	85.79	100.55
2011–2015	58.68	41.44	82.37	50.58	81.50	63.46	79.79
2012–2016	14.87	2.14	95.97	2.32	97.06	3.39	89.07
2013–2017	63.21	36.89	85.68	47.05	85.64	64.24	80.32
2014–2018	10.30	2.22	74.41	3.25	71.64	5.05	64.97
2015–2019	9.45	0.42	87.47	0.42	87.47	0.59	82.38
2016–2020	-2.46	0.09	137.92	0.09	137.92	0.09	137.92
2017–2021	36.19	10.65	73.12	25.78	62.57	48.79	56.50
2018–2022	2.34	0.09	151.10	0.09	151.10	0.09	151.10
2019–2023	33.89	10.79	58.64	19.05	60.71	31.00	60.65

Notes: Each row corresponds to a five-year estimation window that advances by one year, from 2002–2006 through 2019–2023. All-fund mean α is the cross-sectional mean of estimated fund alphas across all eligible funds before FDR selection. For each FDR level, Share is the percentage of funds selected as outperformers and Selected-fund mean α is the cross-sectional mean alpha among the selected funds. Alphas are reported in basis points per month. A fund is eligible in a window if it has at least 12 monthly observations within that window. A dash indicates that no fund is selected.

Table 6: Persistence of mutual fund performance

Current \ Previous decile	1	2	3	4	5	6	7	8	9	10
High	55.34	43.16	47.39	30.02	33.47	27.05	23.49	19.18	12.87	8.11
Medium	32.80	40.12	37.26	47.27	42.05	46.67	44.14	40.05	39.56	30.38
Low	11.86	16.72	15.35	22.71	24.48	26.27	32.37	40.77	47.57	61.52

Notes: Each entry reports the percentage of funds in a current performance group conditional on their decile in the previous five-year window; hence, each column sums to 100. Decile 1 denotes the best-performing funds and Decile 10 the worst. The current performance groups are High (deciles 1–3), Medium (deciles 4–7), and Low (deciles 8–10). The reported transition probabilities are averaged across the six consecutive pairs of five-year windows. Adjacent windows overlap by two years, except for the final pair (2017–2021 and 2019–2023), which overlaps by three years.

Table 7: Number of selected funds in each out-of-sample portfolio

	Number of selected funds		
Holding year	$\gamma = 5\%$	$\gamma = 10\%$	$\gamma = 20\%$
2012	5	7	9
2013	6	10	11
2014	3	3	3
2015	19	41	61
2016	133	199	288
2017	62	131	214
2018	156	206	337
2019	186	267	412
2020	72	101	214
2021	50	99	202
2022	259	375	540
2023	46	65	130

Notes: This table reports the number of funds included in each out-of-sample portfolio at FDR levels of 5%, 10%, and 20%. For each holding year, fund performance is estimated using the preceding ten years of data, with funds required to have at least 60 monthly observations. We then apply the screening BH procedure to the wild-bootstrap p-values and form an equal-weighted portfolio of the selected funds. The resulting portfolio is held throughout the indicated calendar year.

Table 8: Out-of-sample monthly alpha of Benjamini–Hochberg-selected portfolios, 2012–2023

	TOPIX only		TOPIX + SMB, HML, WML	
Portfolio	Alpha (bps/month)	t (p-value)	Alpha (bps/month)	t (p-value)
FDR 5%	41.08	3.146 (0.002)	30.75	3.284 (0.001)
FDR 10%	39.92	3.667 (< 0.001)	31.26	3.776 (< 0.001)
FDR 20%	39.08	3.505 (< 0.001)	30.95	3.517 (< 0.001)

Notes: The dependent variable is the portfolio's monthly excess return. The first specification uses a one-factor model with TOPIX as the market factor, while the second uses a four-factor model that includes TOPIX, SMB, HML, and WML. Alphas are reported in basis points per month. The t-statistics and p-values use Newey–West HAC covariance estimates with Bartlett weights, twelve monthly lags to accommodate dependence within the annual holding cycle. Results remain significant with four or six lags. Each regression contains 144 monthly observations.

Table 9: Out-of-sample return and risk measures, 2012–2023

Portfolio	CAGR	Volatility	Sharpe	Sortino	Maximum drawdown
TOPIX	10.31%	15.71%	0.643	0.997	25.56%
FDR 5%	15.61%	16.40%	0.909	1.533	22.29%
FDR 10%	15.29%	15.91%	0.915	1.526	21.56%
FDR 20%	15.08%	15.80%	0.908	1.522	20.56%

Notes: CAGR and maximum drawdown are calculated from the cumulative wealth index constructed using monthly raw portfolio returns from January 2012 through December 2023. Annualized volatility is computed as the standard deviation of monthly raw portfolio returns multiplied by $\sqrt{12}$. The Sharpe ratio is calculated using monthly excess portfolio returns over the risk-free rate. In contrast, the Sortino ratio uses the root mean squared negative monthly excess return as the downside deviation. Selected portfolios are equally weighted across funds observed in each month. TOPIX is used as the benchmark and is measured using its price-return series.

Appendix A. Matrix completion, latent factors, and bootstrap inference

This appendix provides implementation details for the estimation procedure described in Section 2. Starting from the fund-return model in Equation (2), we first estimate each fund's exposure to the observed Japanese market, size, value, and momentum factors using only its available monthly returns. We then remove the fitted observed-factor component and apply matrix completion to the resulting unbalanced residual panel. The completed residual matrix is used to estimate the latent factors and their fund-specific loadings, which are combined with the observed-factor loadings in the two-pass de-biased alpha estimation. Finally, we construct wild-bootstrap p-values for the multiple-testing analysis. The implementation builds on the unbalanced-panel framework of Giglio et al. (2021).

Appendix A.1. Observed-factor residuals and matrix completion

Let $\mathcal{T}_i$ denote the set of months for which fund i has an observed return, and let $T_i = |\mathcal{T}_i|$ denote the number of such observations. Define the fund-specific sample means

$$\bar{r}_i = \frac{1}{T_i} \sum_{t \in \mathcal{T}_i} r_{it}, \qquad \bar{f}_{o,i} = \frac{1}{T_i} \sum_{t \in \mathcal{T}_i} f_{o,t}.$$

Using only the observations in $\mathcal{T}_i$, we estimate the fund's exposure to the observed factors as

$$\hat{\beta}_{o,i} = \left[\sum_{t \in \mathcal{T}_i} (f_{o,t} - \bar{f}_{o,i})(f_{o,t} - \bar{f}_{o,i})' \right]^{-1} \sum_{t \in \mathcal{T}_i} (f_{o,t} - \bar{f}_{o,i})(r_{it} - \bar{r}_i).$$

We then remove the fitted observed-factor component and construct

$$z_{it} = r_{it} - \bar{r}_i - \hat{\beta}'_{o,i}(f_{o,t} - \bar{f}_{o,i}), \qquad t \in \mathcal{T}_i.$$

When r_{it} is unobserved, z_{it} is also left missing. Stacking these residuals across funds and months gives the incomplete residual matrix

$$Z = [z_{it}] \in \mathbb{R}^{N \times T}.$$

Let

$$\Omega = [\Omega_{it}] \in \{0,1\}^{N \times T}, \qquad \Omega_{it} = \begin{cases} 1, & z_{it} \text{ is observed,} \\ 0, & z_{it} \text{ is missing.} \end{cases}$$

Following Giglio et al. (2021), we recover a low-rank matrix $X \in \mathbb{R}^{N\times T}$ by solving

$$\min_{X} \ \|(Z - X) \circ \Omega\|_F^2 + \lambda_{NT}\|X\|_*, \tag{A.1}$$

where $\lambda_{NT} > 0$ controls the strength of the nuclear-norm penalty and $\circ$ denotes elementwise multiplication. For any matrix $A = [A_{it}]$, the Frobenius norm is $\|A\|_F = \left(\sum_{i,t} A_{it}^2\right)^{1/2}$, while $\|X\|_*$ denotes the nuclear norm, equal to the sum of the singular values of X. Thus, the first term in Equation (A.1) measures fit over the observed entries of the residual matrix, whereas the second term favors a low-rank common component.

We solve Equation (A.1) using a singular-value-thresholding iteration. Let X_k denote the current estimate of the completed low-rank matrix after iteration k. We initialize $X_0 = Z \circ \Omega$, so that missing entries are initially set to zero, and update

$$X_{k+1} = \mathcal{D}_\nu \left[X_k - \tau\, \Omega \circ (X_k - Z)\right], \qquad \nu = \frac{\tau \lambda_{NT}}{2}, \tag{A.2}$$

where $\tau > 0$ is the step size, ν is the implied singular-value threshold, and $\mathcal{D}_\nu(\cdot)$ denotes the singular-value soft-thresholding operator. To define this operator, consider the singular-value decomposition

$$Y = U\, \mathrm{diag}(d_1, \ldots, d_q) V',$$

where q is the number of singular values, the columns of U and V are the corresponding left and right singular vectors, V' denotes the transpose of V, and d_j is the jth singular value of Y. The thresholding operator is

$$\mathcal{D}_\nu(Y) = U\, \mathrm{diag}\left\{(d_1 - \nu)_+, \ldots, (d_q - \nu)_+\right\} V',$$

where $(a)_+ = \max\{a, 0\}$. Hence, each iteration reduces the singular values by ν and sets those below the threshold to zero, thereby favoring a low-rank solution.

We iterate Equation (A.2) until convergence. Following our implementation, we set $\tau = 0.9$, and $\lambda_{NT} = 2.2\|\Omega \circ W\|_2$, where $W \in \mathbb{R}^{N\times T}$ is a simulated noise matrix constructed using the estimated residual covariance. The notation $\|\cdot\|_2$ denotes the spectral norm, equal to the largest singular value of a matrix. We denote the converged matrix-completion estimate by $\hat{X}$.

Appendix A.2. Latent factors and loadings

Let K_ℓ denote the number of retained latent factors, and let $\rho_1, \ldots, \rho_{K_\ell}$ denote the corresponding leading left singular vectors of the completed matrix $\hat{X}$. For fund i, define

$$b_i = \sqrt{N}(\rho_{1i}, \ldots, \rho_{K_\ell i})'.$$

Let $\mathcal{M}_t$ denote the set of funds observed in month t. Using the observed residuals z_{it}, we estimate the latent factors as

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

and the fund-specific latent-factor loadings as

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

The observed- and latent-factor loadings are then stacked as $\hat{\beta}_i = (\hat{\beta}_{o,i}', \hat{\beta}_{\ell,i}')'$. We further define $\bar{f}_o = T^{-1} \sum_{t=1}^{T} f_{o,t}$ and $\hat{v}_t = ((f_{o,t} - \bar{f}_o)', \hat{v}_{\ell,t}')'$, where $\bar{f}_o$ is the full-sample mean of the observed factors and $\hat{v}_t$ stacks the demeaned observed factors and the estimated latent factors.

We determine K_ℓ using cumulative explained variance, marginal explained variance, and the scree-plot knee point (Jolliffe, 2002; Bai and Ng, 2002; Satopää et al., 2011).

Appendix A.3. De-biased alpha and wild-bootstrap inference

Let $\bar{r}_i$ denote fund i's average return and let $\hat{\lambda}$ denote the estimated vector of factor risk premia obtained from the cross-sectional regression of average fund returns on the stacked factor loadings $\hat{\beta}_i$. We compute the de-biased fund alpha as

$$\hat{\alpha}_i = \bar{r}_i - \hat{\beta}_i' \hat{\lambda} + \hat{A}_i,$$

where $\hat{A}_i$ denotes the bias adjustment used in our implementation following Giglio et al. (2021). The corresponding standardized alpha statistic is

$$\hat{t}_i = \frac{\sqrt{T_i} \hat{\alpha}_i}{\hat{\sigma}_i},$$

where $\hat{\sigma}_i$ is the estimated standard error associated with fund i's alpha.

We obtain one-sided inference using a wild bootstrap that preserves each fund's observed return pattern. Let

$$\hat{\varepsilon}_{it} = r_{it} - \bar{r}_i - \hat{\beta}_i' \hat{v}_t, \qquad t \in \mathcal{T}_i,$$

denote the fitted residual from the model containing the observed and estimated latent factors. For bootstrap replication b, we draw independent Mammen-type weights (Mammen, 1993),

$$w_{it}^{(b)} = \frac{\eta_{it}^{(b)}}{\sqrt{2}} + \frac{\left(\zeta_{it}^{(b)}\right)^2 - 1}{2},$$

where $\eta_{it}^{(b)}$ and $\zeta_{it}^{(b)}$ are independent standard normal variables. For each observed fund-month, define $\hat{\varepsilon}_{it}^{*(b)} = w_{it}^{(b)} \hat{\varepsilon}_{it}$. The corresponding bootstrap return is

$$r_{it}^{*(b)} = \hat{\beta}_i' \hat{\lambda} + \hat{\beta}_i' \hat{v}_t + \hat{\varepsilon}_{it}^{*(b)}, \qquad t \in \mathcal{T}_i.$$

Entries that are missing in the original sample remain missing in the bootstrap sample. Within each replication, we re-estimate the factor loadings, factor risk premia, de-biased alphas, and standardized alpha statistics using the same procedure as in the original sample.

Let $\hat{t}_{i,b}^*$ denote the resulting bootstrap standardized alpha statistic for fund i in replication b. Using $B = 1{,}000$ bootstrap replications, we compute the one-sided p-value as

$$p_i = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left\{\hat{t}_{i,b}^* > \hat{t}_i\right\},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Re-estimating the model within each bootstrap replication allows the resulting p-values to reflect residual variation as well as uncertainty associated with factor loadings, factor risk premia, and alpha estimation.