The Review of Financial Studies

Thousands of Alpha Tests

Stefano Giglio
Yale School of Management, NBER and CEPR

Yuan Liao
Department of Economics, Rutgers University

Dacheng Xiu
Booth School of Business, University of Chicago

Data snooping is a major concern in empirical asset pricing. We develop a new framework to rigorously perform multiple hypothesis testing in linear asset pricing models, while limiting the occurrence of false positive results typically associated with data snooping. By exploiting a variety of machine learning techniques, our multiple-testing procedure is robust to omitted factors and missing data. We also prove its asymptotic validity when the number of tests is large relative to the sample size, as in many finance applications. To improve the finite sample performance, we also provide a wild-bootstrap procedure for inference and prove its validity in this setting. Finally, we illustrate the empirical relevance in the context of hedge fund performance evaluation. (*JEL* C12, C55, G12, G23)

Received March 4, 2019; editorial decision July 5, 2020 by Editor Wei Jiang. Authors have furnished an Internet Appendix and code, which are available on the Oxford University Press Web site next to the link to the final published paper online.

Multiple testing is pervasive in empirical finance. It takes place, for example, when trying to identify which among hundreds of factors add explanatory power for the cross-section of returns, relative to an existing model. It also appears

We benefited tremendously from discussions with Simona Abis (discussant), Vikas Agarwal, Campbell Harvey, Chris Hansen, Bryan Kelly, Olivier Scaillet, Rossen Valkanov (discussant), and Michael Wolf. We also appreciate helpful comments from seminar and conference participants at Kepos Capital, Two Sigma, the GSU/RFS FinTech Conference, American Finance Association 2020 Annual Meeting, SoFiE Annual Meeting in Shanghai, European Finance Association 46th Annual Meetings, NBER/NSF Time Series Conference at CUHK, Econometric Society Asian Meeting, NUS Workshop on Asset Pricing and Risk Management, and Big Data, Machine Learning and AI in Economics Conference at Tsinghua University. We thank Kangying Zhou, Michael Bian, and Allen Hu for excellent research assistance. We are grateful to Andrew Sinclair for sharing his TASS code with us. Supplementary data can be found on *The Review of Financial Studies* web site. Send correspondence to Stefano Giglio, 165 Whitney Avenue, New Haven, CT 06520, USA. Email: stefano.giglio@yale.edu.

The Review of Financial Studies 34 (2021) 3456–3496

doi:10.1093/rfs/hhaa111 Advance Access publication September 24, 2020

when trying to identify which funds are able to produce positive alphas (i.e., have "skill") among thousands of existing funds—a central question in asset management. In all these examples, the standard approach is to perform many individual statistical tests on the alphas of the factors or funds relative to the benchmark model, and then make a selection based on the significance of these individual tests.

With multiple testing comes the concern—closely related to data snooping—that as more and more tests are performed, an increasing number of them will be positive purely due to chance. Even if each test individually has a low probability of being due to chance alone, a potentially large fraction of the tests that ex post appear positive will be "false discoveries." A high "false discovery proportion" (FDP) decreases the confidence we have in the testing procedure; in the extreme case, if the number of false discoveries of a procedure reaches 100%, the significance of the individual tests becomes completely uninformative. Controlling the FDP ex ante or, more precisely, its expectation, the false discovery rate (FDR) is more involved than controlling the size of an individual test. The size only depends on evaluating the probability of rejection under the null hypothesis, whereas the FDR is associated with multiple tests and it depends on the true (unknown) data-generating process. In the case of fund performance evaluation, both the FDP and FDR depend on the true underlying distribution of alphas across funds.

In this paper, we propose a rigorous framework to address the data-snooping issue that arises in a specific, but fundamental, finance setting: testing for multiple alphas in linear asset pricing models. We base our methodology on the false discovery control approach introduced by Benjamini and Hochberg (1995) (hereinafter B-H). The idea of the FDR control procedure is to optimally set different significance thresholds across different individual tests in such a way that the overall false discovery rate of the procedure is bounded below a prespecified level, for example, 5%. The objective of our paper is to extend the false discovery control procedure like B-H to the asset pricing context. This involves combining different methods together (to deal with omitted factors, missing data, and so on) and requires additional estimation steps and new asymptotic theory, which we develop. Specifically, our paper makes four main contributions.

First of all, we formally address the threat of omitted factors to the performance of the B-H procedure. Anomaly returns or fund returns in excess of the standard benchmarks appear to be highly cross-sectionally correlated, suggesting that anomalies potentially have exposures to unknown risk factors, and that fund managers may trade common factors that are not observable. This means that there are plausibly omitted factors from the benchmark that, generally speaking, could bias the resulting alpha estimates, and in turn lead to more rejections if alphas are overestimated or less otherwise. Even if there were no bias, omitted factors can still produce large standard errors in the resulting alpha estimates and hence larger p-values corresponding to the true

alternatives, resulting in a power loss. Furthermore, leaving factors in the residuals produces strong correlation among the alpha test statistics, which invalidates the independence assumption of the standard B-H procedure and leads to an increase in the standard error of the FDP. We provide valid asymptotic theory on multiple testing of alphas, which accounts for latent factors estimated using principal component analysis (PCA). While in a recent study Giglio and Xiu (2017) advocate the use of PCA to address the issue of omitted factor bias in the context of risk premium estimation, they do not consider multiple testing of many alphas, which involves a high-dimensional vector of parameters and is thereby a fundamentally different statistical inference problem. Key to being able to tackle the challenge of omitted factors is the fact that we exploit the "blessing" of dimensionality—the other side of the "curse" of dimensionality (Donoho 2000)—that obtains as both $N, T \to \infty$.[1]

Second, we adopt the *matrix completion* method from the recent machine learning literature, see, for example, Koltchinskii et al. (2011), Ma et al. (2011), and Candès and Tao (2010), to recover missing entries in the matrix of asset returns. A notable application of this method is the so-called Netflix problem—predicting customers' ratings of movies based on existing ratings. Because most customers only rate a small set of movies, a large number of ratings are missing from the customer-movie matrix. The key assumption behind this algorithm is that the full matrix is approximately low-rank. This approach is particularly relevant to empirical asset pricing, because many time series of returns have short histories or missing records, and these returns likely follow a factor model. Our use of matrix completion exploits this structure of returns to recover the low-rank part of the returns without missing entries, which in turn leads to estimates of latent factors and their betas. Notably, this matrix completion approach would yield the same result as PCA if there were no missing data. When data are missing, the common approach in the literature is to adopt an EM algorithm (e.g., Stock and Watson 2002; Su et al. 2019). In contrast, our matrix completion approach is much faster; hence it is particularly appealing for large dimensional return matrices.[2]

To improve the finite sample performance of our procedure and prompted by the popularity of the bootstrap approach in this literature, we also develop a wild-bootstrap procedure for multiple testing of alphas, which we prove robust to missing factors and missing data. Notably, the empirical setting suffers from a severe missing data problem, which substantially reduces the effective sample size. As a concrete example, over 70% of entries are missing from the hedge

[1] In most finance applications, the number of hypotheses to test is overwhelmingly large relative to the available sample size. For example, in our empirical analysis we use a couple hundred monthly returns to evaluate the performance of thousands of hedge funds. It is more reasonable to cast our analysis in a large N and large T setting.

[2] The matrix completion method has also attracted attention in the recent econometrics literature on panel data, for example, Athey et al. (2018), Bai and Ng (2019), and Moon and Weidner (2018). Our paper is a first attempt to apply this technique to asset pricing.

fund returns we investigate in this paper because of their short lifespans. The same issue also occurs to individual stock and mutual fund returns. As shown from our simulations, the wild-bootstrap method outperforms the asymptotic approach; both methods appear to control the FDR well. On the contrary, standard widely used bootstrap procedures based on alphas from fund-by-fund time-series regressions fail to control the FDR.

Last but not least, in the context of testing for inequality null hypotheses (e.g., that alphas are non-negative, a null relevant for fund performance evaluation), another contribution of the paper is that we design a screening method to improve the power of the B-H procedure. Holding the count of true alternatives constant, the number of rejections according to the B-H procedure tends to decrease as the total number of hypotheses increases (because the critical value of the B-H procedure drops), hence reducing its average power.[3] In our context, the B-H procedure becomes increasingly less able to detect skilled fund managers as the count of unskilled funds rises. Effectively, our screening approach safely eliminates a set of very unskilled funds in a data-driven way, so that they no longer affect the critical value of the B-H procedure. The alphas of these funds are "deep in the null," and we show theoretically that ignoring them improves the average power compared to the usual B-H procedure, while maintaining the desired FDR control.[4]

In a nutshell, our false discovery control test proceeds as follows. We first adopt matrix completion to interpolate missing entries in asset returns, which also produces estimates of latent factors and betas of both observable and latent factors. We then use the cross-section of average returns to estimate the risk premiums of all factors and obtain all the individual alphas from the regression residuals. Next, we compute asymptotic t-statistics and their p-values or use a wild-bootstrap procedure to construct these p-values directly. Before we plug these p-values into the B-H procedure, we apply the alpha screening step to eliminate rather negative alphas. Finally, we apply the B-H procedure.

We illustrate this procedure using the Lipper TASS data of hedge fund returns. We show empirically that hedge fund returns are highly correlated in the cross-section, even after controlling for the standard models, like the Fung-Hsieh seven-factor model. This is perhaps not surprising, as it is to be expected that many hedge fund strategies load on factors beyond these standard ones, but it needs to be accounted for when measuring funds' alphas. Our procedure—which bounds the false discovery rate below 5%—is able to select, among the universe of funds, a subset of funds that consistently beats the benchmarks both in and out of sample. Compared to other methodologies

[3] According to Benjamini and Liu (1999), the average power of a multiple hypothesis test is defined as the ratio of the expectation of the number of correct rejections (a random variable) and the number of hypotheses for which the alternative holds (assumed to be known).

[4] This screening technique was previously adopted in different contexts by Hansen (2005), Chernozhukov et al. (2013), and Romano and Wolf (2018).

proposed to deal with multiple testing, our procedure is able to identify a larger number of "skilled" funds, achieving superior out-of-sample alpha even with significantly larger portfolios (both in terms of number of funds included and assets under management [AUM]). We also show that our results are robust with respect to using different portfolio sorting procedures, using different observable benchmarks, and using different hedge fund data sets (Evestment instead of TASS).

We emphasize that our methodology is by no means limited to the evaluation of hedge fund performance. Instead, our procedure (and all the inference techniques we derive) can be adapted to different contexts relevant to asset pricing research. For example, it could also be applied to the evaluation of mutual fund performance (e.g., Kosowski et al. 2006, Barras et al. 2010), the detection of anomalies (Harvey et al. 2015, Yan and Zheng 2017; Chordia et al. 2020), and the search for predictive signals of risk-adjusted returns (Green et al. 2013). Different components of our methodology may be applied separately to these contexts. For example, the matrix completion step is important if the underlying data have missing values, but not otherwise. We derive our theory in the most general case, so that it is relevant for all potential applications mentioned above.

The existing literature in asset pricing is aware of the data-snooping concern with multiple testing (e.g., Lo and MacKinlay 1990) and has taken in response two alternative approaches. One has been to abandon the multiple-testing problem altogether: for example, rather than trying to identify *which* funds or factors have alphas, an alternative is to ask whether *any* fund beats the benchmark, or whether funds *on average* beat the benchmark (see, e.g., White 2000, Kosowski et al. 2006, Fama and French 2010). This approach can overcome the multiple-testing problem, since it replaces a multitude of null hypotheses (one per fund) with one joint null hypothesis, but it throws the baby out with the bathwater, as it cannot tell us which of the funds actually produce alpha. The second approach, proposed by Barras et al. (2010), Bajgrowicz and Scaillet (2012), Harvey et al. (2015), etc., imports statistical methods that directly control the family-wise error rate or the false discovery rate. While the recent statistical advances on multiple testing have been successful in many fields (like biology and medicine), their general applicability to finance is still not well understood. The main issue at play is that many of the assumptions on which these methods are based are clearly violated in finance contexts because of omitted factors and missing data. Recently, Harvey and Liu (2018) propose a double-bootstrap approach to control FDR, while also considering other quantities such as false negative rate and odds ratio to trade off false and missed discoveries. Our approach to compute t-statistics can also apply to their context so that the combined approach correctly accounts for omitted factors and missing data.

Data snooping has been a central topic in statistics ever since the early 1950s. Earlier work mainly focuses on using Bonferroni-type procedures to

control the family-wise error rate (FWER) (see, e.g., Simes 1986; Holm 1979). These procedures guard against any single false discovery and hence are overly conservative, in particular when testing many hypotheses. Instead of targeting the question of whether any error was made, the approach of FDR control, developed by Benjamini and Hochberg (1995), takes into account the number of erroneous rejections and controls the expected proportion of errors among the rejected hypotheses. While this seminal work relies on the independence assumption among test statistics, follow-up studies such as Benjamini and Yekutieli (2001) and Storey et al. (2004) demonstrate the robustness of this procedure to certain forms of dependence. Nevertheless, the literature, for example Schwartzman and Lin (2011) and Fan et al. (2012), has recognized the drawbacks of the standard FDR approach in the presence of dependence, including excessive conservativeness, high variance of the FDP, etc., and has proposed alternative procedures, for example., Leek and Storey (2008), Romano and Wolf (2005), and Fan and Han (2016). These methods, however, do not exploit the factor structure of asset returns, nor are they directly applicable to an unbalanced panel. There is also a burgeoning body of research that applies machine learning methods to push the frontiers of empirical asset pricing (see, e.g., Kozak et al. 2017; Freyberger et al. 2017; Giglio and Xiu 2017; Feng et al. 2020; Kelly et al. 2017; Gu et al. 2018). These papers employ variable selection or dimension reduction techniques to analyze a large number of covariates. These procedures, however, do not directly control the number of false discoveries.

Finally, our empirical results directly speak to a long literature dedicated to evaluating the performance of the hedge fund industry. Contrary to the case of mutual funds, for which net alpha is estimated to be zero or negative for the vast majority of funds (with some exceptions, evidenced in the recent work of Berk and Van Binsbergen [2015]), there is more evidence that some hedge funds are able to generate alpha. An important first step in this empirical exercise has been the exploration of hedge fund strategies and their risk exposures (Fung and Hsieh 1997, 2001, 2004, Agarwal and Naik 2000, 2004, Agarwal et al. 2009, Patton and Ramadorai 2013, Bali et al. 2014); we use several of the benchmarks proposed in this literature as observable factors in our analysis. At the same time, the literature has explored whether hedge funds are able to produce alpha in excess of these benchmarks, using different statistical methodologies (Liang 1999, Ackermann et al. 1999, Liang 2001, Mitchell and Pulvino 2001, Baquero et al. 2005, Kosowski et al. 2007, Fung et al. 2008, Jagannathan et al. 2010, Aggarwal and Jorion 2010, Bali et al. 2011). While the majority in this literature focus on using alpha as the key performance measure, Chen (2019) proposes to identify skilled fund managers by evaluating their probability of outperforming their unskilled counterfactuals.

The paper proceeds as follows. Section 1 discusses the detailed procedure for our FDR control. Section 2 presents Monte Carlo simulations, followed by

an empirical study in Section 3. Section 4 concludes. The Internet Appendix provides the asymptotic theory and technical details.

1. Methodology

Our framework is based on a combination of three key ingredients, each essential to execute multiple testing correctly in our asset pricing context: factor analysis, Fama-MacBeth regressions, and false discovery control. These three ingredients are complemented by additional procedures that help us tackle issues specific to the asset pricing applications (e.g., the presence of missing data).

At a broad level, our methodology proceeds as follows. In a first step, we use time-series regressions to estimate fund exposures to (observable) benchmark factors. Since these benchmarks do not fully capture the comovement of fund returns, hiding, for example, unobservable risk exposures, we further apply PCA or matrix completion to the residuals to recover the missing commonalities. This results in a model where, effectively, both observable and estimated latent factors coexist. Next, we implement cross-sectional regressions like Fama-MacBeth to estimate the risk premiums of the factors and the alphas relative to the augmented benchmark model that includes both observable and estimated latent factors. Importantly, in the presence of missing data, the estimated alphas need to be debiased. Finally, we build t-statistics for these alphas and apply the B-H procedure for the FDR control.

In what follows, we describe each ingredient of our procedure in detail, and we discuss how we tackle complications due to missing data and how to enhance the power of our procedure using alpha screening. We also provide an alternative wild-bootstrap approach to construct statistical tests. We derive the statistical properties of our methodology, and show formally that it indeed achieves the desired false discovery rate control in multiple tests for alpha.

1.1 Model setup

We begin with a description of the model. We assume the $N \times 1$ vector of excess returns r_t follows a linear factor model:

$$r_t = \alpha + \beta\lambda + \beta(f_t - \mathbb{E}(f_t)) + u_t, \tag{1}$$

where f_t is a $K \times 1$ vector of factors and u_t is the idiosyncratic component. The parameter λ is a $K \times 1$ vector of factor risk premiums, which is identical to the expected return of f_t only if f_t is tradable.[5]

[5] Throughout, we impose an unconditional factor model in which both α and β are time-invariant. This corresponds to a practical trade-off between efficiency and robustness (to model misspecification) in light of the limited sample size in our empirical analysis. That said, it is straightforward to extend our procedure to conditional models à la Ang and Kristensen (2012), although the theory would become less transparent. In practice, we follow the literature and apply our procedure on a sequence of rolling windows.

The objective is to find individual funds with truly positive alphas. To do so, we formulate a collection of null hypotheses, one for each fund:

$$\mathbb{H}_0^i : \alpha_i \leq 0, \quad i = 1, \ldots, N. \tag{2}$$

While our primary target is on inequalities, almost all techniques we introduce below work for equality nulls (except for alpha screening, which is designed for inequalities):

$$\mathbb{H}_0^i : \alpha_i = 0, \quad i = 1, \ldots, N. \tag{3}$$

Importantly, the alpha testing problems we consider are fundamentally different from the standard GRS test, in which the null hypothesis is a single statement that

$$\mathbb{H}_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = 0. \tag{4}$$

The former is a multiple-testing problem that addresses which funds have significantly positive alphas. In contrast, the latter addresses whether there exists (at least one) fund whose alpha is significantly different from zero. While the latter is the natural way to test asset pricing models (which imply that all alphas should be zero), it is not the right one if the objective is to identify which funds are able to generate positive alpha.[6]

Simultaneous testing of multiple hypotheses—like the test we propose—is prone to a false discovery problem, also referred to as data-snooping bias: the possibility that many of the tests will look significant by pure chance, even if their true alpha is zero. To understand why, recall that for each 5%-level test, there is a 5% chance that the corresponding null hypothesis is falsely rejected. This is the so-called Type I error. In other words, there is a 5% chance that a fund with no alpha realizes a significant test statistic and is therefore falsely recognized as one with real positive alpha. This error exacerbates substantially when testing many hypotheses. For example, suppose there are 1,000 funds available, with only 10% of them having positive alphas. Conducting 1,000 tests independently would yield $1,000 \times (1 - 10\%) \times 5\% = 45$ false positive alphas, in addition to $1,000 \times 10\% = 100$ true positive alphas (assuming ideally a zero Type II error). Consequently, among the $100 + 45 = 145$ "skilled" fund managers we find, almost one-third of them are purely due to luck.

The multiple-testing problem is one of the central concerns in statistics and machine learning. One of the classical approaches is to control the probability of *one or more* false rejections, that is, the FWER, instead of the Type I Error. One such approach is the Bonferroni procedure, which suggests rejecting the null of the individual hypothesis at the $5\%/N$ level, where N is the total number of tests. However, this method is overly conservative in that the level of the test shrinks to zero asymptotically. To ensure that the probability that even just

[6] Common tests in the literature also include the test for positive average alphas, that is, $\mathbb{H}_0 : \mathrm{E}(\alpha_i) \leq 0$, and the test for existence of at least one fund with a positive alpha, that is, $\mathbb{H}_0 : \max_{1 \leq i \leq N}(\alpha_i) \leq 0$.

one of the N tests is a false discovery stays below a certain level, say 5%, the procedure needs to adopt a higher and higher threshold as the number of tests N increases; this will result in an unfeasibly high bar for the t-statistic of each test.

A more suitable procedure in this scenario is to control the FDR instead, that is, the expected fraction of false rejections. This is the purpose of the original B-H procedure, which has been the most popular since it was introduced and has been widely used across disciplines. We now turn to describing the B-H procedure and showing under what conditions it can be applied in an asset pricing context.

1.2 Controlling the false discovery rate

We start by setting up some notation. Suppose t_i is a test statistic for the null $\mathbb{H}_0^i$ (often taken as the t-statistic) and a corresponding test that rejects the null whenever $t_i > c_i$ under a prespecified cutoff c_i. Let $\mathcal{H}_0 \subset \{1, ..., N\}$ denote the set of indices for which the corresponding null hypotheses are true. In addition, let $\mathcal{R}$ be the total number of rejections in a sample, and let $\mathcal{F}$ be the number of false rejections in that sample:

$$\mathcal{F} = \sum_{i=1}^{N} 1\{i \leq N : t_i > c_i \text{ and } i \in \mathcal{H}_0\},$$

$$\mathcal{R} = \sum_{i=1}^{N} 1\{i \leq N : t_i > c_i\}.$$

Both $\mathcal{F}$ and $\mathcal{R}$ are random variables. Note that, in a specific sample, we can obviously observe $\mathcal{R}$, but we cannot observe $\mathcal{F}$. However, we can design a procedure to effectively limit how large $\mathcal{F}$ is relative to $\mathcal{R}$ in expectation. More formally, we write the FDP and its expectation, FDR, as

$$\text{FDP} = \frac{\mathcal{F}}{\max\{\mathcal{R}, 1\}}, \quad \text{FDR} = \mathbb{E}(\text{FDP}).$$

For comparison, we can also write the per-test error rate, $\mathbb{E}(\mathcal{F})/N$, and the FWER, $\mathbb{P}(\mathcal{F} \geq 1)$. The naive procedure that tests each individual hypothesis at a predetermined level $\tau \in (0, 1)$ guarantees that $\mathbb{E}(\mathcal{F})/N \leq \tau$. But note that it does not guarantee any limits on the false discovery rate, which can be much larger than τ. The Bonferroni procedure, instead, tests each hypothesis at a level τ/N. This guarantees that $\mathbb{P}(\mathcal{F} \geq 1) \leq \tau$ and implies a false discovery rate below τ, at the cost of reducing the power of the test in detecting the true alphas (in the limit, if a test is so strict that it never rejects, the false discovery rate is zero! But that test will have no power.).

1.2.1 The B-H algorithm. The FDR control procedure strikes a balance between these two approaches. It accepts a certain number of false discoveries

as the price to pay to gain power in detecting true rejections. Analogously to standard individual tests (that control the size of Type I error), this procedure controls the size of the FDR: it ensures that FDR $\leq \tau$.

We now describe the details of the B-H procedure.

Algorithm 1. (B-H procedure)

S1. Sort in ascending order the collection of p-values, $\{p_i : i = 1, \ldots, N\}$, of the individual test statistics $\{t_i\}$. Denote $p_{(1)} \leq \ldots \leq p_{(N)}$ as the sorted p-values.

S2. For $i = 1, \ldots, N$, reject $\mathbb{H}_0^i$ if $p_i \leq p_{(\widehat{k})}$, where $\widehat{k} = \max\{i \leq N : p_{(i)} \leq \tau i / N\}$.

Benjamini and Hochberg (1995) establish the validity of their procedure under the condition that the test statistics are independent.[7] A natural question is how this procedure can correctly control the FDR given that the FDR depends on the unobservable distribution of alphas. Below we provide a brief discussion of the intuition of this procedure.

Recall that we aim to identify a critical value p^*, so that the null $\mathbb{H}_0^i$ is rejected for all $p_i < p^*$. To increase power, p^* should be equal to the largest value $p \in (0, 1)$ such that the FDP is controlled with high probability:

$$\frac{\mathcal{F}(p)}{\max\{\mathcal{R}(p), 1\}} \leq \tau, \tag{5}$$

where $\mathcal{F}(p)$ denotes the number of false discoveries and $\mathcal{R}(p)$ the number of significant tests. They are the same as the aforementioned $\mathcal{F}$ and $\mathcal{R}$, but are written in terms of a given p:

$$\begin{aligned} \mathcal{F}(p) &= \sum_{i=1}^{N} 1\{i \leq N : p_i < p \text{ and } \alpha_i \leq 0\}, \\ \mathcal{R}(p) &= \sum_{i=1}^{N} 1\{i \leq N : p_i < p\}. \end{aligned}$$

Note that for any given p, $\mathcal{R}(p)$ is known. While $\mathcal{F}(p)$ is not, it can be bounded from above using the data. Let N_0 be the number of true null hypotheses. We have the following approximation (with high probability):

$$\mathcal{F}(p) \overset{(a)}{\approx} N_0 \mathbb{P}(p_i < p | \alpha_i \leq 0) \overset{(b)}{\leq} N_0 \mathbb{P}(p_i < p | \alpha_i = 0) \overset{(c)}{=} N_0 p, \tag{6}$$

where approximate equality (a) follows from the independence assumption of individual test statistics; inequality (b) follows from the fact that p-values p_i's

[7] Benjamini and Yekutieli (2001) revise the use of $\widehat{k}$ in S2 by $\widehat{k} = \max\{i \leq N : p_{(i)} \leq \tau i / (N C_N)\}$., where $C_N = \sum_{i=1}^{N} i^{-1}$, which they show guarantees FDR control under a certain form of dependence. Nonetheless, because of $C_N \approx \log(N) + 0.5$, $C_{1,000} \approx 7.4$, the method remains too conservative, limiting the power of the procedure.

are larger under $\alpha_i \leq 0$ than under $\alpha_i = 0$; equality (c) follows since under the null of $\alpha_i = 0$, p-values are uniformly distributed. We still do not know N_0, so we replace it with some upper bound M. The choice of M determines the degree of conservativeness of an FDR control procedure, which we shall discuss later. We then have, with high probability,

$$\mathcal{F}(p) \leq Mp.$$

Replace $\mathcal{F}(p)$ with such upper bound. Therefore, inequality (5) is preserved so long as:[8]

$$p \leq \frac{\tau \mathcal{R}(p)}{M} = \frac{\tau \sum_{i=1}^{N} 1\{i \leq N : p_i < p\}}{M}. \tag{7}$$

We can then find p^* as the largest p to satisfy inequality (7):

$$p^* = \max \left\{ p \in (0,1) : p \leq \frac{\tau \sum_{i=1}^{N} 1\{i \leq N : p_i < p\}}{M} \right\},$$

which equals the usual B-H critical value $p_{(\widehat{k})}$ if $M = N$, a rather conservative estimate of N_0. The above derivation sheds light on the sources of conservativeness in the B-H procedure: testing inequalities as shown from part (b) of inequality (6), as well as overestimating the number of true nulls, N_0. In the next section, we focus on alleviating the conservativeness of the procedure and enhancing its power.

1.2.2 Alpha screening. Based on the above discussion, not surprisingly, the count of negative alphas adversely affects the power of the B-H procedure. To see this, holding the number of non-negative alphas constant, as the count of negative alphas increases, the critical value $\tau\widehat{k}/N$ in the B-H procedure shrinks (because N increases and the relative order of p-values corresponding to non-negative alphas remains approximately the same), making it more difficult to detect true positive alphas.

We tackle this problem by using a simple yet powerful dimension reduction technique—the screening method in the context of testing for inequalities. The idea is that when some of the alphas are "overwhelmingly negative" (which we call "deep in the null"), their corresponding hypotheses could be simply eliminated from the set of candidate hypotheses, because it is safe to accept them. This would reduce the total count of hypotheses and thereby improve the power of FDR control. Based on this idea, we propose to reduce the set of funds to

$$\widehat{\mathcal{I}} = \left\{ i \leq N : t_i > -\log(\log T)\sqrt{\log N} \right\},$$

where the threshold depends on the sample size and the cross-sectional dimension. Our theory (presented in the Internet Appendix) shows that with

[8] If $\mathcal{R}(p) = 0$, then $\mathcal{F}(p) = 0$, so Equation (5) holds trivially.

probability approaching one, for any i such that $\widehat{\alpha}_i \notin \widehat{\mathcal{I}}$, the true $\alpha_i < 0$. We thereby can safely consider a smaller set of funds, $\widehat{\mathcal{I}}$, for FDR control. Also, we formally show in the Internet Appendix that this screening-based B-H procedure has a larger power to detect positive alphas, and it can consistently identify *all* positive alphas with reasonably strong signal strength.

Algorithm 2. (alpha screening B-H procedure) Let $|\widehat{\mathcal{I}}|$ denote the number of elements in $\mathcal{I}$.

S1. Sort the p-values, $p_{(1)} \leq \ldots \leq p_{(|\widehat{\mathcal{I}}|)}$, for $\{p_i : i \in \mathcal{I}\}$.

S2. For $i \in \mathcal{I}$, reject $\mathbb{H}_0^i$ if $p_i \leq p_{(\widehat{k})}$, where $\widehat{k} = \max\{i \in \widehat{\mathcal{I}} : p_{(i)} \leq \tau i / |\widehat{\mathcal{I}}|\}$. Accept all other $\mathbb{H}_0^i$.

Alternatively, Barras et al. (2010) apply a simple adjustment proposed by Storey (2002) to improve the power of the B-H procedure. Specifically, they suggest replacing $\widehat{k}\tau/N$ in the cutoff value by $\widehat{k}\tau/N_0$, where N_0 is the number of true null hypotheses that can be estimated using $\widehat{N}_0 = (1-\lambda)^{-1} \sum_{i=1}^{N} \{p_i > \lambda\}$, where $\lambda \in (0,1)$ is a tuning parameter. The intuition behind this adjustment is that under the zero-alpha nulls, the p-values are uniformly distributed on $(0,1)$; therefore one would expect $N_0(1-\lambda)$ of the p-values to lie within the interval $(\lambda, 1)$ for any sufficiently large λ. Replacing N with $N_0 < N$ thereby increases the power of the procedure. As shown in the previous section, this amounts to setting $M = \widehat{N}_0$, which still controls the FDR, making it less conservative.

However, this adjustment is not applicable in the context of testing null hypotheses that are *inequalities*, under which the p-values are no longer uniformly distributed. The deviation from the uniform distribution becomes very severe when many alphas are very negative, which would substantially overestimate N_0, eventually resulting in conservativeness.[9] In contrast, the Algorithm 2 we propose replaces $i \leq N$ in the definition of $\widehat{k}$ with $i \in \widehat{\mathcal{I}}$ and sets $M = |\widehat{\mathcal{I}}|$, which directly targets the conservativeness due to inequality null hypotheses. By eliminating the true negatives, the remaining null alphas are close to zero, so that we can safely increase the critical value, and consequently enhance the power of the procedure.

In a setting similar to ours, Harvey and Liu (2018) propose to increase statistical power by dropping funds that appear only for a small number of periods. This approach shares the same spirit as our alpha screening step and is typically used in the literature (e.g., Fung and Hsieh [1997] require at least 36 months of data). As we discuss below, this requirement is also important for

[9] Although truly negative alphas conform with the null hypotheses, their corresponding p-values are larger than those with zero alphas. In the case with many negative alphas, according to the estimator by Storey (2002), $\widehat{N}_0 = (1-\lambda)^{-1} \sum_{i=1}^{N} \{p_i > \lambda\}$, more p-values are greater than λ, resulting in an overestimate of N_0. In our simulations, $\widehat{N}_0$ can even be greater than N.

dealing with missing values, so we impose it in our analysis. However, in our study we also apply the alpha screening procedure.

Finally, it is interesting to think about our screening procedure in relation to the problem of selection bias in hedge fund reporting, which also affects our empirical application. A well-known problem with standard hedge fund data sets (e.g., Agarwal et al. 2013), is that "bad" funds, those with particularly negative alpha, will likely not report to the data set. This is of course an important issue for understanding the *average* alpha (denoted by α_0) of hedge funds, which would be biased upwards. But when the objective is to identify funds with skill, this bias is much less relevant. The fact that funds with a truly negative alpha and funds that would have displayed a negative t-statistic anyway are excluded from consideration if they do not report to the data sets, has the same effect as our screening step: it increases the power of the methodology to identify good funds among those that do report.

The above discussion on FDR control assumes the existence of valid test statistics that are approximately uncorrelated. In the next section, we explain how we construct alpha estimates and the corresponding t-statistics (or p-values) that satisfy this condition.

1.3 Estimating alpha

To properly estimate the alphas of asset or fund returns, we first need to specify a benchmark model. As discussed in detail in Cochrane (2009), when the benchmark includes nontradable factors, estimating Equation (1) requires two-pass Fama-MacBeth regressions. The first stage estimates β using time-series regressions of individual fund returns onto the benchmark factors, and the second stage involves a cross-sectional regression of average returns onto the estimated β, where the residuals of this regression yield estimates of alpha, denoted as $\widehat{\alpha}$.

The classical setting assumes a fixed dimension N, so that the asymptotic theory is developed under $T \to \infty$ only, which only holds under the GRS null hypothesis $\mathbb{H}_0$: $\alpha_1 = \alpha_2 = \ldots = \alpha_N = 0$. A close scrutiny of the Fama-MacBeth estimator shows that in a more general setting that allows for non-zero alphas, we have

$$\widehat{\alpha} - \alpha = -\beta(\beta' \mathbb{M}_{1_N} \beta)^{-1} \beta' \mathbb{M}_{1_N} \alpha + O_P(T^{-1/2}), \tag{8}$$

where $\mathbb{M}_{1_N} = \mathrm{I}_N - N^{-1} 1_N 1_N'$, and the first term dominates and hence prohibits a consistent estimator of α.[10] Importantly, $\widehat{\alpha}_i$ is inconsistent even if $\alpha_i = 0$ for any fixed i, as long as $\beta_i \neq 0$. The bias arises from the cross-sectional correlation between betas and alphas. This is not surprising since the cross-sectional regression requires an exact orthogonality condition between residual

[10] We give a formal statement on the inconsistency of $\widehat{\alpha}$ for fixed N setting in Proposition A.7 in the Internet Appendix.

and regressors, which is not satisfied if some alphas are not zero. Consequently, when the dimension N is fixed, the two-pass regression cannot be applied to the tests of multiple hypotheses unless all alphas are zero. Since we cannot exclude that some of the alphas are actually nonzero when testing individual hypotheses, this creates fundamental obstacles to the estimation and testing of alphas.

A potential solution to this problem is to estimate and test alphas using only time-series regressions. This approach appears viable but only in the case in which all factors are excess returns; this assumption is violated if the observable factors are nontradable.

A second concern relates to the choice of the benchmark and the possibility that some important factors are omitted, thus attributing to alpha what truly is just exposure to the omitted risk factors. More explicitly, consider a specific example of Equation (1):

$$r_t = \alpha + \begin{bmatrix} \beta_o & \beta_l \end{bmatrix} \begin{bmatrix} f_{o,t} \\ f_{l,t} \end{bmatrix} + u_t = \underbrace{\alpha + \beta_l \lambda_l}_{\text{“alpha”}} + \beta_o f_{o,t} + \underbrace{\beta_l (f_{l,t} - \mathbb{E} f_{l,t}) + u_t}_{\text{“idiosyncratic” error}}, \quad (9)$$

where $f_{o,t}$ is the observed benchmark model and $f_{l,t}$ is the vector of omitted factors missing from the benchmark. To make things simple in this example, assume that f_o is an excess return (this condition is not required in our general specification below). Equation (9) indicates at least three challenges due to the omitted factors. First, the “alpha” computed relative to the benchmark model that just includes f_o (and thus omits f_l) includes the risk premium associated with the missing factor f_l. As long as the latent factors in f_l contribute to the total risk premiums, then a bias $\beta_l \lambda_l$ would arise in the estimated “alpha.” Even if f_l is not priced, omitting f_l would still lead to an omitted variable bias for α when f_l is correlated with f_o.[11] The bias of alpha can result in more rejections if alphas are overestimated, or less because of underestimation. Secondly, f_l plays the role of “idiosyncratic” error. Since the idiosyncratic error covariance matrix contributes to the asymptotic covariance matrix of the alpha estimates, the presence of f_l in the residuals increases the standard errors of the alpha estimates, making it more difficult to separate the nulls from the alternatives. For instance, the p-values corresponding to the true positive alphas become larger, resulting in a loss of power. Thirdly, leaving f_l in the residuals produces strong correlation among the alpha test statistics, which invalidates the independence assumption of the standard B-H procedure. A consequence of this is that the standard error of the FDP will increase, so the control on FDR (the mean of FDP) becomes unstable, as shown by Efron (2010).

In what follows, we explain how our new test statistics overcome these obstacles by exploiting the blessings of dimensionality. To better explain the

[11] Explicitly, the bias equals $\beta_l \lambda_l - \beta_l \mathbb{E} f_l f_o' (\mathbb{E} f_o f_o')^{-1} \mathbb{E} f_o A$, where $A = (1 - \mathbb{E} f_o' (\mathbb{E} f_o f_o')^{-1} \mathbb{E} f_o)^{-1}$.

intuition, we start with the balanced panel and two special cases of Equation (1). For convenience, we introduce some additional notation. We use capital letter A to denote the matrix $(a_1, a_2, \ldots, a_T)$, where a_t is a time series of vectors. We use $\mathbb{M}_B = \mathrm{I}_p - B(B'B)^{-1}B$ to denote the annihilator matrix for any $p \times q$ matrix B. Let F be the $K \times T$ matrix of $\{f_t : t \leq T\}$, V be the $K \times T$ matrix of $\{f_t - \mathbb{E} f_t : t \leq T\}$, R be the $N \times T$ matrix of $\{r_t : t \leq T\}$, and U be the $N \times T$ matrix of $\{u_t : t \leq T\}$. Let $\bar{r} = \frac{1}{T}\sum_t r_t$. We use excess returns for r_t throughout.

1.3.1 Observable factors only. When all factors are observable (not necessarily tradable), we can directly estimate α using the classical two-pass regression:

Algorithm 3. (observable factors only)

S1a. Run time-series regressions and obtain the OLS estimator $\widehat{\beta}$.

$$\widehat{\beta} = (R\mathbb{M}_{1_T}F')(F\mathbb{M}_{1_T}F')^{-1}. \tag{10}$$

S2. Run a cross-sectional regression of $\bar{r}$ on the estimated $\widehat{\beta}$ and a constant regressor 1_N to obtain the slopes $\widehat{\lambda}$:

$$\widehat{\lambda} = (\widehat{\beta}'\mathbb{M}_{1_N}\widehat{\beta})^{-1}(\widehat{\beta}'\mathbb{M}_{1_N}\bar{r}). \tag{11}$$

S3. Estimate α by subtracting the estimated risk premiums from average returns:

$$\widehat{\alpha} = \bar{r} - \widehat{\beta}\widehat{\lambda}. \tag{12}$$

Since r_t is a vector of excess returns, risk premiums are usually estimated without using an intercept in the cross-sectional regression of S2. That is because the assumption of zero alphas is typically imposed for risk premiums estimation. In contrast, including an intercept term here allows for a possibly nonzero cross-sectional mean for α, denoted by α_0. Its estimator can be written explicitly as $\widehat{\alpha}_0 = N^{-1}1_N'\widehat{\alpha}$. As a side note, it is also interesting to test if α_0 is non-negative, which addresses whether on average hedge fund alphas are positive, though it is not the objective we pursue in this paper.

In a special case that all observable factors are tradable, factors' risk premiums are equal to their expectations, so we simply estimate the risk premiums by taking their time-series averages. As such, $\widehat{\lambda}$ in Equation (11) can be replaced by $\widehat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} f_t$, whereas the remaining steps are identical.

1.3.2 Latent factors only. In the case that some factors are missing from an observable factor model, the first-step time-series regressions are no longer consistent. We could instead consider a model with all factors being latent. Such a model is in fact quite general, in that we can always assume all factors being latent even if there were observable factors, and estimate all factors from

the data altogether. In this case, we follow Giglio and Xiu (2017) and proceed by rewriting Equation (1) into a statistical factor model:

$$\bar{R} = \beta \bar{V} + \bar{U}, \tag{13}$$

where $\bar{A} = A\mathbb{M}_{1_T}$ for $A = R$, V, and U.

Simply replacing S1a in Algorithm 3 with S1b below leads to a new algorithm for estimating α in this scenario:

Algorithm 4. (latent factors only)

S1b. Let $S_R = \frac{1}{T}\bar{R}\bar{R}'$ be the $N \times N$ sample covariance matrix of R. Conduct the principal component analysis of S_R: set

$$\widehat{\beta} = \sqrt{N}(b_1, \ldots, b_K),$$

where $b_1, \ldots, b_K$ are the K eigenvectors of S_R, corresponding to its largest K eignvalues.

S2 & S3 are the same as in Algorithm 3.

This procedure therefore uses the principal components of returns as factors and uses them as a benchmark to estimate the alphas. Note that Algorithm 4 requires the number of latent factors as an input, which can be estimated using a variety of procedures in the literature, such as those based on information criteria (Bai and Ng 2002), or based on eigenvalue ratios (Ahn and Horenstein 2013), etc. Alternatively, we can treat the number of latent factors as a tuning parameter, which can be selected based on the eigenvalue scree plot. We adopt this procedure in practice for convenience.

1.3.3 General case. We now present the most general case, where we assume

$$r_t = \alpha + \begin{bmatrix} \beta_o & \beta_l \end{bmatrix} \begin{bmatrix} \lambda_o \\ \lambda_l \end{bmatrix} + \begin{bmatrix} \beta_o & \beta_l \end{bmatrix} \begin{bmatrix} f_{o,t} - \mathbb{E} f_{o,t} \\ f_{l,t} - \mathbb{E} f_{l,t} \end{bmatrix} + u_t, \tag{14}$$

where $f_{o,t}$ is a $K_o \times 1$ vector of observable factors, and $f_{l,t}$ is a $K_l \times 1$ vector of latent factors, respectively. Both factors can be nontradable. Note that we do not assume that latent and observable factors are uncorrelated, or that the betas with respect to observable and latent factors are cross-sectionally uncorrelated.

While Algorithm 4 still works without using observable factors, using these factors is expected to deliver better performance. To estimate α in this case, we combine S1a and S1b, and then proceed with S2 and S3 as in Algorithm 3. Specifically, we first obtain $\widehat{\beta}_o$ from time-series regressions using observable factors alone, and then obtain $\widehat{\beta}_l$ by applying PCA to the covariance matrix of residuals from time-series regressions. The estimated $\widehat{\beta}_o$ and $\widehat{\beta}_l$ are stacked together as $\widehat{\beta}$. The algorithm is summarized as follows.[12]

[12] In Internet Appendix A.1.1, we provide a simpler version of this algorithm in the special case of tradable observable factors, in which we directly use the time-series average of these factors as the estimates for their risk premiums. This, however, does not affect the asymptotic behavior of the estimator as N and T increase.

Algorithm 5. (estimating α in model (14))

S1. a. Run time-series regressions and obtain the OLS estimator $\widehat{\beta}_o$ and residual matrix Z:

$$\widehat{\beta}_o = (R\mathbb{M}_{1_T} F_o')(F_o \mathbb{M}_{1_T} F_o')^{-1}, \quad Z = \bar{R} - \widehat{\beta}_o \bar{F}_o, \tag{15}$$

where $F_o = (f_{o,1}, f_{o,2}, \ldots, f_{o,T})$.

b. Let $S_Z = \frac{1}{T} ZZ'$ be the $N \times N$ sample covariance matrix of Z. Let

$$\widehat{\beta}_l = \sqrt{N}(b_1, \ldots, b_{K_l}),$$

where $b_1, \ldots, b_{K_l}$ are the K_l eigenvectors of S_Z, corresponding to its largest K_l eigenvalues.

The resulting $\widehat{\beta}$ is given by

$$\widehat{\beta} = (\widehat{\beta}_o, \widehat{\beta}_l).$$

S2 & S3. The same as S2 & S3 in Algorithm 3.

It is worth mentioning that $\widehat{\beta}_o$ is a consistent estimator of β_o only if f_o and f_l are uncorrelated. In our general setting, this condition is not imposed, so $\widehat{\beta}_o$ is possibly inconsistent due to the omitted variable (latent factors) bias. However, one of our theoretical contributions is to show that the presence of such bias does not affect the inference for alphas, thanks to the invariance of alpha to the rotation of the factors. Formally, we can show that $\widehat{\beta}_o \xrightarrow{P} \beta_o + \beta_l H_1$ for some matrix H_1, where $\beta_l H_1$ denotes the omitted variable bias. Hence the probability limit of $\widehat{\beta}_o$ is still spanned by $\beta = (\beta_o, \beta_l)$. Note that the probability limit of the PCA estimator $\widehat{\beta}_l$ is also spanned by β_l. As a result, we have established that there is a rotation matrix H such that[13]

$$\widehat{\beta} = (\widehat{\beta}_o, \widehat{\beta}_l) \xrightarrow{P} \beta H.$$

The resulting alpha estimate remains consistent because it is invariant to rotations (the rotation matrix H is canceled with its inverse in $\widehat{\lambda}$) and is thus not affected by the omitted variable bias.

[13] As a side note, a close scrutiny of the matrix H shows that it has the following structure:

$$H = \begin{pmatrix} \mathbb{I} & 0 \\ H_1 & H_2 \end{pmatrix}.$$

The rotation invariance of $\beta\lambda$ implies that $\widehat{\lambda}$ converges to $H^{-1}\lambda$, which in turn yields that $\widehat{\lambda}_o$ is a consistent estimator of λ_o. This conclusion echoes the result of Giglio and Xiu (2017).

1.4 Dealing with missing data

It is not uncommon in finance applications to deal with unbalanced panels. For example, many hedge funds last for short periods of time, then liquidate, and many new funds pop up. It is therefore important that the estimators we propose work in the presence of missing data. In this section, we describe how to use a *matrix completion* algorithm to handle missing data within our procedure.

1.4.1 Matrix completion. The matrix completion approach relies on a critical assumption that the full matrix can be written as a noisy low-rank matrix. This assumption is naturally justified for de-meaned realized returns in our context (see Equation (13)).

We now present the matrix completion algorithm in a generic setting. The goal is to recover an $N \times T$ low-rank matrix X. Suppose that Z is an $N \times T$ matrix (the "noisy version" of X), which can be written as $Z = X + E$, and E is the noise. In addition, suppose Z is not fully observed and Ω is an $N \times T$ matrix whose (i,t)-th element $\omega_{it} = 1\{z_{it} \text{ is observed}\}$. Using this notation, econometricians can only observe $Z \circ \Omega$ and Ω, where $\circ$ represents the element-wise matrix product.

We employ the following nuclear-norm penalized regression approach to recover X:

$$\widehat{X} = \arg\min_{X} ||(Z - X) \circ \Omega||^2 + \lambda_{NT} ||X||_n, \quad [14] \tag{16}$$

where $||X||_n$ denotes the matrix nuclear norm and $\lambda_{NT} > 0$ is a tuning parameter. By penalizing the singular values of X, the algorithm achieves a low-rank matrix as the output.[15] The latent factors and betas can then be estimated via the associated singular vectors of $\widehat{X}$.

We now apply this algorithm to our model (Equation (14)) and update the steps of Algorithm 5. For this purpose, we need some additional notation. Let $\mathcal{N}_t$ denote the set of funds that are observed at time t and $\mathcal{T}_i$ denote the collection of time points on which the i-th fund return is observed:

$$\mathcal{N}_t = \{i \in \{1, \ldots, N\} : r_{it} \text{ is observed}\}, \quad \mathcal{T}_i = \{t \in \{1, \ldots, T\} : r_{it} \text{ is observed}\}.$$

We first estimate an observable factor model to obtain loadings of observable factors, as we do in S1a of Algorithm 5. This step allows us to calculate the residual matrix Z for all periods and all funds (of course with missing entries), which will then serve as the input for the matrix completion algorithm. With the estimated factors and loadings, we then proceed with S2 and S3 of Algorithm 5, except that in the case of missing data, the estimated α_i has a bias. So we add a de-biasing step before we can use the estimates for testing. The detailed steps are given as follows:

---

[14] The *nuclear-norm* $||X||_n := \sum_{i=1}^{\min\{N,T\}} \psi_i(X)$, where $\psi_1(X) \geq \psi_2(X) \geq \ldots$ are the sorted singular values of X.

[15] We leave the details on how we solve Equation (14) to Algorithm A.2 of the Internet Appendix.

Algorithm 6. (**estimating α in model** (14) **via matrix completion**)

S1. a. Obtain $\widehat{\beta}_o$ and the residual matrix $Z=(z_{it})_{N\times T}$:

$$\begin{aligned}\widehat{\beta}_{o,i} &= (F_{o,i}\mathbb{M}_{1_{T_i}}F_{o,i}')^{-1}(F_{o,i}\mathbb{M}_{1_{T_i}}R_i),\\ z_{it} &= r_{it}-\bar{r}_i-\widehat{\beta}_{o,i}'(f_{o,t}-\bar{f}_{o,i}) \text{ when } r_{it} \text{ is observable;}\\ &\quad \text{otherwise } z_{it} \text{ is missing,}\end{aligned}$$

where $\bar{r}_i=\frac{1}{T_i}\sum_{t\in\mathcal{T}_i}r_{it}$ is the average return for each fund i at its observed time points, $\bar{f}_{o,i}=\frac{1}{T_i}\sum_{t\in\mathcal{T}_i}f_{o,t}$ is the average of observable factors at time points on which fund i is observed (note that we assume no data are missing for observable factors), R_i is the $T_i\times 1$ vector of $\{r_{it}:t\in\mathcal{T}_i\}$, and $F_{o,i}$ is the $K_o\times T_i$ matrix of $\{f_{o,t}:t\in\mathcal{T}_i\}$.

b. Conduct matrix completion, with Z in Equation (14) constructed above, and obtain a low-rank matrix $\widehat{X}$. Estimate the latent factors and their loadings using $\widehat{X}$:

$$\begin{aligned}\widehat{v}_{l,t} &= \left(\sum_{i\in\mathcal{N}_t}b_ib_i'\right)^{-1}\sum_{i\in\mathcal{N}_t}b_iz_{it}, \quad t=1,...,T,\\ \widehat{\beta}_{l,i} &= \left(\sum_{t\in\mathcal{T}_i}\widehat{v}_{l,t}\widehat{v}_{l,t}'\right)^{-1}\sum_{t\in\mathcal{T}_i}\widehat{v}_{l,t}z_{it}, \quad i=1,...,N,\end{aligned}$$

where $(b_1,...,b_{K_l})$ is the top K_l left singular-vectors of $\widehat{X}$. The resulting $\widehat{\beta}$ is given by

$$\widehat{\beta}=(\widehat{\beta}_o,\widehat{\beta}_l),$$

and the factors $\widehat{v}_t=(f_{o,t}-\bar{f}_o,\widehat{v}_{l,t}')'$, where $\bar{f}_o=\frac{1}{T}\sum_{t=1}^{T}f_{o,t}$.

S2 is the same as in Algorithm 3 with inputs $\widehat{\beta}$ from above and $\bar{r}_i$, which yields the estimate $\widehat{\lambda}$.

S3. Estimate and de-bias the estimates of α:

$$\widehat{\alpha}_i=\bar{r}_i-\widehat{\beta}_i'\widehat{\lambda}+\widehat{A}_i, \quad i=1,...,N, \tag{17}$$

where, writing $\widehat{\xi}_i'=e_i'-\widehat{\beta}_i'(\widehat{\beta}'\mathbb{M}_{1_N}\widehat{\beta})^{-1}\widehat{\beta}'\mathbb{M}_{1_N}$, $e_i'=(0,...0,1,0,...,0)$, $\widehat{g}_i=\frac{1}{T_i}\sum_{t\in\mathcal{T}_i}\widehat{v}_t'\widehat{\beta}_i$, $\widehat{H}_{o,i}=\widehat{V}_{l,i}\mathbb{M}_{1_{T_i}}F_{o,i}'(F_{o,i}\mathbb{M}_{1_{T_i}}F_{o,i}')^{-1}$, and $\widehat{H}_o=\widehat{V}_l\mathbb{M}_{1_T}F_o'(F_o\mathbb{M}_{1_T}F_o')^{-1}$,

$$\widehat{A}_i=\widehat{\beta}_{l,i}'(\widehat{H}_{o,i}-\widehat{H}_o)\widehat{\lambda}_o-\widehat{\xi}_i'\widehat{g}.$$

Here $\widehat{V}_l$ is the $K_l\times T$ matrix of $\{\widehat{v}_{l,t}:t\le T\}$ and $\widehat{V}_{l,i}$ is the $K_l\times T_i$ matrix of $\{\widehat{v}_{l,t}:t\in\mathcal{T}_i\}$.

The additional term $\widehat{A}_i$ in Equation (17) is introduced to de-bias the estimated alphas due to an unbalanced panel. It is worth noting that for balanced panels (i.e., $\mathcal{T}_i = \{1, ..., T\}$ for all i), $\widehat{A}_i = 0$ and our matrix completion algorithm is equivalent to the usual PCA.

1.5 Constructing valid test statistics for false discovery control

Having described how we obtain the alpha estimates, we now turn to the construction of the test statistics. One of our theoretical contributions is that we formally show in Theorem A.1 of Internet Appendix A.2 that the alpha estimates satisfy, in the case of balanced panel, for each $i \leq N$, as $N, T \to \infty$,

$$\begin{aligned} \sigma_{i,NT}^{-1}(\widehat{\alpha}_i - \alpha_i) &\xrightarrow{d} \mathcal{N}(0,1), \\ \sigma_{i,NT}^2 &= \frac{1}{T}\mathrm{Var}(u_{it}(1 - v_t'\Sigma_f^{-2}\lambda)) + \frac{1}{N}\mathrm{Var}(\alpha_i)\frac{1}{N}\beta_i' S_\beta^{-1}\beta_i, \quad (18) \end{aligned}$$

where $v_t := f_t - \mathbb{E}f_t$, $\Sigma_f := \mathrm{Cov}(f_t)$ and $S_\beta = \frac{1}{N}\beta'\mathbb{M}_{1_N}\beta$. This formula holds true for all three cases: observable factors only, latent factors only, and the general case. The asymptotic result (18) can be used for inference about each individual alpha. Note that the variance $\sigma_{i,NT}^2$ consists of two components: in addition to the $1/T$ term that arises from time-series estimation, the second term $\frac{1}{N}\mathrm{Var}(\alpha_i)\frac{1}{N}\beta_i' S_\beta^{-1}\beta_i$ directly reflects the estimation errors from the cross-sectional regression. This second component will result in cross-sectional dependence among the test statistics, jeopardizing the FDR control. That said, as long as $T \log N = o(N)$, the second term is dominated by the first, and so is its impact on FDR.

In the general case of an unbalanced panel, if $\max_i(T_i)\log N = o(N)$, Theorem A.2 of Internet Appendix A.2 shows that for $i = 1, 2, \ldots, N$,

$$\sqrt{T_i}(\widehat{\alpha}_i - \alpha_i) = \frac{1}{\sqrt{T_i}}\sum_{t \in \mathcal{T}_i} u_{it}(1 - v_t'\Sigma_f^{-1}\lambda) + o_P\left(\frac{1}{\sqrt{\log N}}\right). \quad (19)$$

Based on Equation (19), Algorithm 7 below constructs the corresponding t-statistics. Note that this asymptotic approximation also holds in the case where all factors are observable and tradable, in which time-series regressions are used for alpha estimation. Because the idiosyncratic error u_{it}'s are weakly dependent, these t-statistics are weakly dependent, using which we then apply the proposed alpha screening B-H procedure (Algorithm 2) to select the positive alphas.

Algorithm 7. (construction of the test statistics)

S1 & S2 & S3 are the same as those in Algorithms 3, 4, 5, and 6.

S4. Calculate the standard error as follows:

$$\text{se}(\widehat{\alpha}_i) \quad = \quad \frac{1}{\sqrt{T_i}}\widehat{\sigma}_i, \quad \widehat{\sigma}_i^2 = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i}\widehat{u}_{it}^2(1-\widehat{v}_t'\widehat{\Sigma}_f^{-1}\widehat{\lambda})^2, \tag{20}$$

where $\widehat{u}_{it} = r_{it} - \bar{r}_i - \widehat{\beta}_i'\widehat{v}_t$ is the residual, and $\widehat{\Sigma}_f = \frac{1}{T}\sum_{t=1}^{T}\widehat{v}_t\widehat{v}_t'$.

S5. Calculate the t-statistics and p-values:

$$t_i = \frac{\widehat{\alpha}_i}{\text{se}(\widehat{\alpha}_i)}, \quad p_i = 1-\Phi(t_i), \quad i = 1,\ldots,N,$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function.

Obviously, while this algorithm is presented in the context of unbalanced panel, the balanced panel is a special case by setting $\mathcal{T}_i = \{1,...,T\}$ and $T_i = T$ for all $i \le N$. Another remark worth mentioning is that in the context of testing for the equalities in Equation (3), we can simply replace the calculation of p-values in S5. by $p_i = 2(1-\Phi(|t_i|))$ for $i = 1,2,\ldots,N$.

1.6 Constructing valid p-values using bootstrap

While the asymptotic inference is straightforward, its finite sample performance may become a concern in scenarios where a large amount of data are missing. The bootstrap is a popular approach that has been used frequently in this context to compute critical values for various test statistic (e.g., Kosowski et al. 2006, Fama and French 2010, Harvey and Liu 2018; Chordia et al. 2020). It substitutes computation for asymptotic approximations, yet delivers potentially better small sample performance. That said, the bootstrap is not a panacea as it is known to fail in extensive examples (Horowitz 2001). The choice of bootstrap algorithms also matters in that less effective algorithms can lead to a dramatic loss of power and level of accuracy of a test (Hall and Wilson 1991). Distinct from what has been commonly employed in asset pricing, we propose here a wild-bootstrap algorithm, originally introduced by Liu (1988) and Mammen (1993), and prove its validity in the presence of omitted factors and missing data.

Recall that Algorithm 6 produces the estimated factors $\widehat{v}_t$, their loadings $\widehat{\beta}_i$, their risk premiums $\widehat{\lambda}$, and $\widehat{\alpha}_i$ for each fund, which in turn yield $\widehat{u}_{it} := r_{it} - \bar{r}_i - \widehat{\beta}_i'\widehat{v}_t$ if r_{it} is not missing. Our bootstrap algorithm below produces the p-values of the alpha test statistics that will serve as inputs for the B-H procedure in Algorithm 2.

Algorithm 8. (bootstrapping p-values)

S0. Generate a bootstrap sample of r_{it}^* by resampling weighted residuals:

$$r_{it}^* = \widehat{\beta}_i'\widehat{\lambda} + \widehat{\beta}_i'\widehat{v}_t + \widehat{u}_{it}^*, \quad \widehat{u}_{it}^* = \widehat{u}_{it}w_{it}, \quad \text{for } t \in \mathcal{T}_i, \tag{21}$$

where $\{w_{it}: i \le N, t \le T\}$ is a sequence of i.i.d. random variables, satisfying $\mathbb{E}w_{it}=0$ and $\mathrm{Var}(w_{it})=1$.[16]

S1. Obtain $\widehat{\beta}^*=(\widehat{\beta}_1^*,...,\widehat{\beta}_N^*)'$:

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

where R_i^* is the $T_i\times 1$ vector of $\{r_{it}^*: t\in\mathcal{T}_i\}$, and $\widehat{V}_i$ is the $K\times T_i$ matrix of $\{\widehat{v}_t: t\in\mathcal{T}_i\}$.

S2 is the same as in Algorithm 6, using $\bar{r}_i^*$ and $\widehat{\beta}^*$, which yields the estimate $\widehat{\lambda}^*$.

S3. Estimate and de-bias the estimates of α:

$$\widehat{\alpha}_i^*=\bar{r}_i^*-\widehat{\beta}_i^{*'}\widehat{\lambda}^*-\widehat{\xi}_i^{*'}\widehat{g}^*,\quad i=1,...,N, \tag{22}$$

with $\widehat{\xi}_i^{*'}=e_i'-\widehat{\beta}_i^{*'}(\widehat{\beta}^{*'}\mathbb{M}_{1_N}\widehat{\beta}^*)^{-1}\widehat{\beta}^{*'}\mathbb{M}_{1_N}$ and $e_i'=(0,...0,1,0,...,0)$, $\widehat{g}_i^*=\frac{1}{T_i}\sum_{t\in\mathcal{T}_i}\widehat{v}_t'\widehat{\beta}_i^*$.

S4. Repeat S0-S3 for B times and denote the estimates from S3 as $\{\widehat{\alpha}_{i,b}^*: i=1,\ldots,N, b=1,\ldots,B\}$. Compute the bootstrap p-values as

$$p_i=\frac{1}{B}\sum_{b=1}^{B}1\{\widehat{\alpha}_{i,b}^*>\widehat{\alpha}_i\},\quad i=1,\ldots,N,$$

where $\widehat{\alpha}_i$ is given by S3 of Algorithm 6.

A few points are worth mentioning. First of all, Hall and Wilson (1991) suggest that the bootstrap resampling should reflect the null hypothesis to preserve the power of a test. In our context, we test a large number of individual hypotheses with different null hypotheses. We prove that it is sufficient to impose the joint null hypothesis that all alphas are zero for resampling, instead of generating separate bootstrap samples imposing each one of the null hypotheses. This substantially simplifies the bootstrap procedure. Second, bootstrap resampling should maintain the same missing pattern as that of the original sample; that is, r_{it}^* is resampled if and only if r_{it} is not missing ($t\in\mathcal{T}_i$). Needless to say, this algorithm also works if no data are missing. Last but not least, compared to S1b of Algorithm 6, the corresponding step in Algorithm 8 does not involve matrix completion as it directly uses the estimated latent factors as if they were observed in the bootstrap samples. For the same reason, the bias correction in S3 is also simpler for the bootstrap approach. This strategy again simplifies and accelerates the calculation. Furthermore, we prove in Theorem A.4 of Internet Appendix A.2 that our wild-bootstrap method is valid in this context. Finally, if the object of interest is multiple testing for equality nulls as in Barras et al. (2010), then the only modification to make in the above

[16] Mammen (1993) suggested using $w_{it}=\frac{1}{\sqrt{2}}\eta_{it}+\frac{1}{2}(\gamma_{it}^2-1)$, where η_{it} and γ_{it} are independent standard normal.

algorithm is to replace the one-sided bootstrap p-values in S4 with two-sided ones:

$$p_i = \frac{1}{B}\sum_{b=1}^{B} 1\{|\widehat{\alpha}^*_{i,b}| > |\widehat{\alpha}_i|\}, \quad i = 1, \ldots, N.$$

The (standard) bootstrap algorithm, widely adopted in the literature, differs from Algorithm 8 in Step S0. Specifically, this approach generates a bootstrap sample with replacement $\mathcal{S} = \{t_1, ..., t_T\}$ from $\{1, ..., T\}$ using

$$r^*_{it} = \widehat{\beta}_i'\widehat{\lambda} + \widehat{\beta}_i'\widehat{v}^*_t + \widehat{u}^*_{it}, \tag{23}$$

where $\{(\widehat{v}^*_t, \widehat{u}^*_{it}) : t = 1, ..., T\} = \{(\widehat{v}_t, \widehat{u}_{it}) : t \in \mathcal{S}\}$. In the case of a severe missing data problem, we find (not reported here for reasons of space) that this approach performs inadequately.[17] In contrast, our wild-bootstrap procedure is immune to idiosyncrasies due to missing data.

2. Simulations

In this section, we examine the finite sample performance of the tests and their asymptotic approximations developed in Internet Appendix A.2. With respect to the data-generating process, we consider a five-factor model for hedge fund returns, with factors calibrated to match a latent factor model estimated using the empirical data. We then resample the estimates of beta and idiosyncratic volatility of individual funds in our data, so that the summary statistics (e.g., time series R^2s, volatilities) of the simulated fund returns match their empirical counterparts. To examine the impact of missing data on our procedures, we also resample the exact missing pattern of these funds.[18] Throughout the simulations, we fix $T = 240$ and $N = 1{,}000$, and on average over 70% of all records are missing.

We vary the simulated cross-sectional distribution of alphas to check the performance of the FDR control under various fractions of true null hypotheses. As illustrated earlier, the conservativeness of an FDR procedure depends on the percentage of true alternatives. To show this, we simulate the alphas from a mixture of two Gaussian distributions, $\mathcal{N}(-2\sigma, \sigma^2)$ and $\mathcal{N}(2\sigma, \sigma^2)$, with mixture probabilities p_1 and p_2, plus a point mass at zero. Recall that we test multiple inequalities (alphas less than or equal to zero). The individual t-tests

[17] One reason is that for certain fund with limited data, the degree of freedom in the resampled factors is rather limited because resampling is only possible at time points when this fund's return is not missing. In extreme cases, this leads to singularity in beta estimation. Alternatively, we might resample residuals alone, but this requires stronger assumptions (e.g., strict exogeneity) on the dependence between factors and residuals.

[18] In the hedge fund literature, it is typical to remove funds that only survive for short periods of time. In the empirical study, we follow the literature and require a fund to have at least 36 months of tracking record; the remaining sample includes about 1,700 funds over 290 months. This is quite a challenging environment for the various procedures we discuss, as more than 70% of the observations are missing. When the missing data problem becomes more severe, the finite sample performance deteriorates. In Section 3, we instead explore imposing a tighter constraint that mitigates the issue at the cost of removing additional funds.

become increasingly conservative when there are more negative true alphas (larger p_1). Also, the FDR procedure tends to be more conservative when the number of true null hypotheses is lower (larger p_2). We vary their mixture probabilities p_1 and p_2 to demonstrate the conservativeness of our procedure. Different values of p_1 and p_2 also mimic different data sets. According to Barras et al. (2010), most mutual fund alphas are not significant, in which case p_2 would be very small.[19] In contrast, for hedge fund returns in the TASS sample, p_1 tends to be small, potentially due to a self-reporting (selection) bias, whereas p_2 appears to be much larger (around 20%, as we discuss below).

In Table 1 we report the FDR (i.e., the sample average of the FDP) and the standard error of the FDP for different procedures in many scenarios. In addition, we report the average power as well as the false negative rates (FNRs), defined as the average percentage of false acceptance among all accepted tests in multiple testing, because both measures reflect the power of a multiple-testing procedure. We consider in total 11 different scenarios. For (a), we apply Algorithms 1 and 3 to all five observable factors. There are no missing factors in this case. In (b) we apply the same procedure to four observable factors alone. The comparison between (a) and (b) demonstrates the effect of missing factors. We then apply Algorithm 4 with five latent factors and no observable factors in (c), and Algorithm 5 with four observable factors and one latent factor in (d). Next, in (e) we conduct alpha screening (Algorithm 2) on top of the algorithms used in (d). To investigate the effect of missing data and missing factors separately, we use a balanced panel in the above settings. In (f), we use an unbalanced panel and adopt matrix completion (Algorithm 6) instead of PCA (Algorithm 5), and construct t-statistics using asymptotic standard errors. The next setting, (g), is identical, except that there we adopt the wild-bootstrap procedure (Algorithm 8) for inference. For both (f) and (g), we also apply alpha screening as in (e). For comparison, we then implement a list of common procedures in the literature. To start with, in (h) and (i), we apply the standard bootstrap procedure to fund-by-fund time-series regression estimates of α using four factors with/without a balanced panel, respectively. In (j) we report the results based on the bootstrap procedure of Barras et al. (2010).[20] Finally, for (k), we simulate the ideal setting where neither factors nor data are missing, but without using any FDR control methods.

We first verify the central limit results developed in Theorem A.1 of Internet Appendix A.2. Figure 1 provides histograms of the standardized alpha estimates for an arbitrary fund in each of the three scenarios (b), (c), and (d). The estimator in the scenario (d) (based on Algorithm 3) is inconsistent, as verified from the

[19] Our theory does not allow for the case of $p_2 = 0$; see Assumption A.4 in the Internet Appendix, in which the number of true alternatives is zero.

[20] We follow the same procedure and choice of tuning parameters as theirs, in which we estimate the FDR over a pre-selected grid of significance levels and determine a significance level that provides the estimated FDR as close as possible to the 5% target.

Table 1
Monte Carlo simulation results

			(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
# of observable factors			5	4	0	4	4	4	4	4	4	4	5
# of latent factors			0	0	5	1	1	1	1	0	0	0	0
Missing data								✓	✓		✓	✓	
p_1	p_2												
0.1	0.1	FDR	5.20	8.14	4.81	4.85	5.55	8.36	5.58	47.78	27.83	45.28	35.27
		FDP std.	3.50	15.27	3.23	3.27	3.52	5.11	4.51	6.70	8.69	6.17	7.21
		Avg. power	64.81	53.62	64.62	64.58	65.45	49.85	42.70	60.98	41.17	51.45	79.88
		FNR	3.79	4.95	3.80	3.81	3.72	5.32	6.00	4.44	6.26	5.38	2.31
0.1	0.2	FDR	4.38	6.18	4.12	4.13	4.76	6.55	5.45	33.41	20.99	26.81	19.56
		FDP std.	2.48	11.24	2.41	2.43	2.61	3.44	3.18	4.94	5.32	4.73	5.31
		Avg. power	69.02	57.68	68.78	68.72	69.63	54.70	51.72	64.19	47.44	51.78	80.43
		FNR	7.15	9.53	7.20	7.21	7.03	10.14	10.72	8.77	11.81	11.11	4.83
0.1	0.3	FDR	3.67	4.88	3.44	3.44	3.92	5.19	4.59	24.62	16.07	17.91	12.31
		FDP std.	2.02	8.56	1.90	1.93	2.02	2.53	2.42	3.87	3.80	4.09	3.84
		Avg. power	71.29	60.14	71.10	71.03	71.96	57.47	55.48	66.18	50.39	52.25	80.43
		FNR	10.86	14.44	10.92	10.94	10.64	15.32	15.89	13.52	17.84	17.40	7.94
0.2	0.1	FDR	4.50	7.24	4.17	4.21	5.31	7.87	5.39	43.60	24.84	42.10	32.12
		FDP std.	3.33	14.22	3.01	3.07	3.52	4.92	4.62	6.81	8.35	6.22	7.44
		Avg. power	63.40	52.36	63.18	63.14	64.55	49.28	43.17	59.53	40.04	50.63	78.54
		FNR	4.03	5.19	4.05	4.05	3.91	5.50	6.10	4.66	6.50	5.57	2.50
0.2	0.2	FDR	3.79	5.41	3.53	3.53	4.42	5.97	5.13	29.77	18.26	24.06	17.29
		FDP std.	2.32	10.20	2.20	2.22	2.51	3.18	3.00	4.76	4.98	4.57	5.16
		Avg. power	68.15	56.97	67.92	67.84	69.33	54.56	51.90	63.35	46.67	51.35	79.66
		FNR	7.43	9.79	7.48	7.49	7.18	10.30	10.82	8.96	12.06	11.28	5.05
0.3	0.1	FDR	3.90	6.46	3.63	3.64	5.04	7.11	5.08	39.19	21.89	38.58	28.93
		FDP std.	3.02	13.11	2.80	2.83	3.40	4.70	4.43	6.90	8.05	6.28	7.34
		Avg. power	62.07	51.28	61.85	61.84	63.69	48.75	43.47	58.12	38.90	49.82	77.25
		FNR	4.27	5.42	4.29	4.29	4.09	5.68	6.22	4.88	6.75	5.76	2.70

The table reports the false discovery rate (FDR), the standard error of false discovery proportion (FDP std.), the average power (Avg. power) of different procedures, and the false negative rate (FNR), in simulation settings with different choices of mixture probabilities p_1 and p_2. The total number of factors in the DGP is five. The total number of observable and latent factors determines whether a procedure omits any factors. A checkmark "✓" in the row "Missing Data" indicates that the simulated panels of individual fund returns are unbalanced. The number of Monte Carlo repetitions is 1,000. All numbers in Columns a-k are percentages. We consider in total 11 different scenarios. Column a applies Algorithms 1 and 3 to all five observable factors. Column b applies the same procedure to four observable factors alone. Column c applies Algorithm 4 with five latent factors and no observable factors, and d applies Algorithm 5 with four observable factors and one latent factor. Column e conducts alpha screening (Algorithm 2) on top of the algorithms used in d. The above settings use a balanced panel. Column f uses an unbalanced panel, adopts matrix completion (Algorithm 6) instead of PCA (Algorithm 5), and constructs t-statistics using asymptotic standard errors. Column g applies to the same setting, except that it adopts the wild-bootstrap procedure (Algorithm 8) for inference. Alpha screening is employed in both f and g. For comparison, columns h and i apply the standard bootstrap procedure to fund-by-fund time-series regression estimates of α using four factors with/without a balanced panel, respectively. Column j reports the results based on the bootstrap procedure of Barras et al. (2010).

left panel, because it adopts a misspecified four-factor model. In (c), Algorithm 5 is designed to take into account the omitted factor in the regression residual. Not surprisingly, it works well. The estimator in the scenario (b) (based on Algorithm 4) ignores all observable factors, but it estimates a five-factor latent factor model, which also corrects the omitted factor bias; its histogram thereby matches the asymptotic distribution.

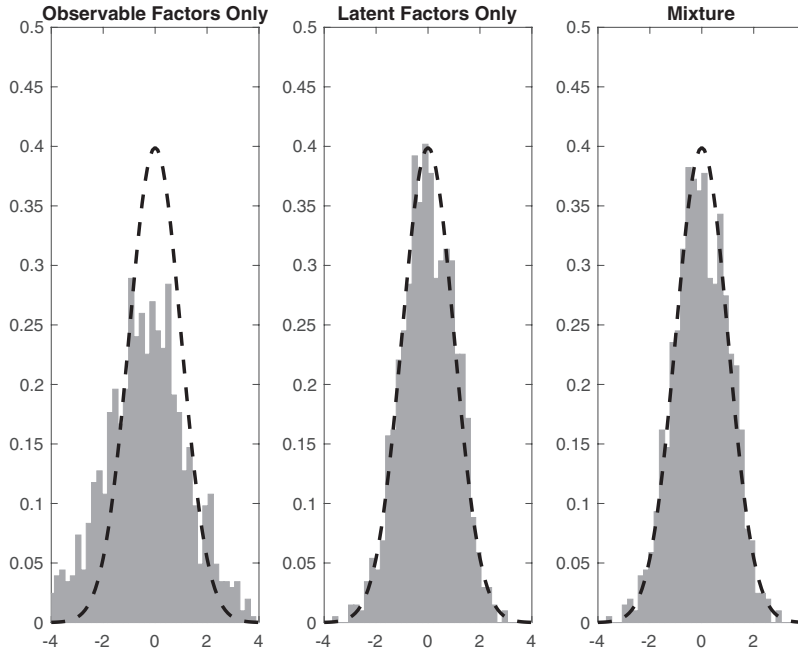

Figure 1
Histograms of the standardized individual alpha tests
The figure plots the histograms of the standardized alpha estimates for one fixed fund using Algorithms 3 (observable factors only), 4 (latent factors only), and 5 (mixture), respectively. The true data-generating process is a five-factor model. p_1 and p_2 are fixed at .1. The number of Monte Carlo repetitions is 1,000.

There are quite a few conclusions we can draw from the comparison results in Table 1. First of all, we find it critical to use alpha tests that take into account omitted factors, by comparing columns c and d with column b. Comparing columns c and d with a, using latent factors in place of the omitted factors works well, as if the omitted factors were known. Second, comparing columns d and e, we find alpha screening less conservative and more powerful, in particular when p_1, the percentage of unskilled funds, is large. Third, the standard bootstrap method in various scenarios does not cope well with missing data and missing factors. In contrast, our wild-bootstrap approach works well, as our theoretical analysis shows. The wild-bootstrap algorithm also improves over the asymptotic inference in f, which is perhaps not surprising. Fourth, columns b, h, i, and j clearly show that missing a factor tends to increase the standard errors of the FDP. This observation agrees with our intuition and earlier discussion. Finally, without any B-H type control, the false discovery rate can exceed 35% among the experiments we consider even in the most ideal setting, even though, not surprisingly, its false negative rate hits the lowest value, given that this procedure overly rejects by a large margin compared to the other procedures.

Overall, the alpha screening B-H procedure (Algorithm 2), together with matrix completion for alpha estimation (Algorithm 6) and wild-bootstrap for inference (Algorithm 8), performs stably well in all scenarios we consider. We thereby choose it as our benchmark in the following empirical analysis.

3. Empirical Analysis: Hedge Fund Alphas

3.1 Hedge fund returns data

To illustrate a potential use of our methodology, we apply it to the Lipper TASS hedge funds data set, covering the time period 1994–2018. The data set contains a panel of returns and assets under management. The Lipper TASS data set is subject to a number of potential biases. We follow closely the bias correction and data-cleaning procedures of Sinclair (2018), who kindly shared his code with us; these are in turn mostly based on the procedures detailed in Getmansky et al. (2015). We describe the main concerns with the data and the data-cleaning procedures in Internet Appendix C. As is standard in this literature, we only focus on funds that have a sufficiently long time series. This is particularly important in our setting, because a large amount of missing data affects our ability to estimate the risk premiums of latent factors and deteriorates the finite sample performance of the estimator. Based on our simulations, we choose a minimum period of 36 months (i.e., a three-year tracking record), and we explore robustness below.

After applying these filters, we are left with 1,761 funds in our data set. Panel A of Figure 2 reports the histogram of average monthly excess returns, which shows large dispersion across funds. The cross-sectional mean of average excess returns in our sample is 19 bp per month. In total, around 73% of records are missing, so our matrix completion step plays an important role in this context.

3.2 Benchmark models

We consider two standard benchmark models. Our baseline model will be the Fung and Hsieh (2004) seven-factor model, a well-known model proposed specifically to benchmark hedge funds. The model includes market, size, a bond factor, a credit risk factor, and three trend-following factors (related to bonds, currencies, and commodities). As an alternative, we also consider the model proposed by Agarwal and Naik (2004), which includes the Fama-French-Carhart four factors (market, size, value, and momentum factors) plus two option-based factors (an out-of-the-money call and an out-of-the-money put factor).

3.3 Factor structure of hedge fund returns

To get a sense of the factor structure of hedge fund returns, the blue line in panel B of Figure 2 shows the first 15 eigenvalues of the excess returns in our panel. There clearly are important common components driving hedge

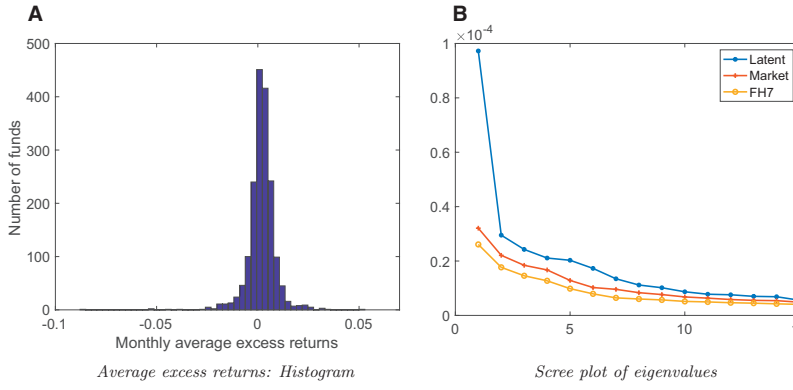

Figure 2
Properties of hedge fund excess returns
Panel A shows the histogram of average monthly excess returns for the 1,761 funds in our full sample. Panel B reports the first 15 eigenvalues of the covariance matrix of excess returns, denoted as "Latent," sorted from highest to lowest, as well as eigenvalues of the residual covariance matrices relative to two benchmarks: the Market and the FH7 model. Sample period is 1994–2018.

fund returns. The figure also plots the eigenvalues of the residual covariance matrices of two benchmark models: the CAPM and the FH7 model. It is evident that observable factors indeed help capture certain common variation in the cross-section, because the largest few eigenvalues shrink substantially. The largest gain comes from the market factor, which shrinks the largest eigenvalue by about two-thirds. The marginal contribution by the remaining observable factors is less significant. Importantly, there remains common variation in the residuals of the FH7 model, which will be captured in our empirical analysis by the additional latent factors. Based on the scree plot, we choose three or five factors in our analysis.

3.4 In-sample analysis

We begin with an in-sample analysis in which we compare the funds selected by our FDR control methodology to those selected using different methodologies. For 10 different fund selection procedures (one in each column), Table 2 reports the average alphas in bp per month (first row) and the average t-stat (second row) for the selected funds, as well as the fraction of funds selected out of 1,761 (third row). The table also reports in the fourth row the p-value of the test that the average alpha is equal to zero, and in the fifth row the p-value for the difference in average alpha between each methodology and our full-fledged FDR procedure (reported in column 1, computed as described in Internet Appendix A.3.2). To make the results comparable across methodologies, alphas for funds selected using different procedures are computed using a common benchmark, which includes the FH7 factors plus three latent factors.

The first column applies our methodology to select funds: our FDR control with three latent factors in addition to FH7, using the wild bootstrap to compute

Table 2
In-sample results

	Mixed FDR		Only observable			No screen	No screen	Only latent	Alternative models	
	Bootstrap	Avar	FDR	$p < .05$	$p > .05$	FDR	observ. FDR	FDR	HL	BSW
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Average alpha	70.4	68.7	66.2	62.9	4.8	70.6	66.4	61.4	65.8	65.4
Average t-stat	4.6	4.2	5.8	4.3	.3	4.7	5.9	4.3	18.8	5.4
Fraction selected	.19	.24	.10	.20	.80	.18	.10	.19	.01	.12
p-val. alpha =0	<.01	<.01	<.01	<.01	.04	<.01	<.01	<.01	<.01	<.01
p-val., alpha = (1)	–	.04	<.01	<.01	<.01	.66	<.01	<.01	.05	<.01

The table reports the results of the multiple alpha tests for the 1,761 hedge funds in our sample using different methodologies. The first row reports the average alpha of the funds selected; the second row the average t-stat of the funds selected; the third row the fraction of funds selected; the fourth row the p-value of the test that the average alpha is equal to zero; the fifth row the p-value of the test that the average alpha is equal to the one obtained using the full FDR methodology (first column). Each column corresponds to a different selection procedure: (1) our FDR with seven observable factors (FH7) and three latent factors, and bootstrap standard errors; (2) our FDR with seven observable factors (FH7) and three latent factors, and asymptotic standard errors; (3) FDR with only seven observable factors (FH7); (4) funds with individual p-value below .05, using only observable factors; (5) funds with individual p-value above 0.05, using only observable factors and alpha screening; (6) our FDR without alpha screening; (7) standard FDR (with only observable factors) without alpha screening; (8) our FDR with ten latent factors and no observable factors; (9) the methodology in Harvey and Liu (2018); (10) the methodology in Barras et al. (2010). The benchmark model used to compute the alphas is FH7 with three additional latent factors. The sample periods is 1994–2018.

p-values. Out of a total of 1,761 funds, our procedure deems 19% to have positive alpha. The average fund selected has an alpha of 70 bp per month, highly statistically significant. Column 2 also applies our methodology, but using asymptotic standard errors to compute the p-values of the funds, instead of the wild bootstrap. As expected, the two methodologies give similar results, with bootstrap performing better, as expected from the simulations.

Columns 3–5 apply the standard methods of this literature (e.g., Kosowski et al. 2007), in which alphas and p-values for the fund selection are calculated against the FH7 model without any latent factors. Given that the benchmark model has no latent factors and all observable factors are tradable, we use time-series estimation for the alphas in columns 3–5, as prescribed in Section 1.3.1, with alpha screening employed in column 3.[21] Specifically, column 4 selects all funds with p-values below .05, and therefore corresponds to the standard selection procedure that ignores the multiple-testing problem. This procedure selects about 20% of the funds, with an average alpha of 63 bp per month. Column 3 uses these p-values to compute the FDR control, which reduces the fraction of funds selected to 10% but has little effect on the alpha (66 bp per month). Note that while the differences are economically small (5–10 bp), they are often statistically significant (due to the nontrivial overlap in the different portfolios). Finally, column 5 reports for comparison the characteristics of funds with p-values *above* .05, which is therefore the complement of the funds selected in column 4. These funds, which represent 80% of the fund population, have an estimated average alpha of only 5 bp per month.

Two broad comments on these in-sample results are worth emphasizing. First, the fact that, ex post, funds with a p-value below .05 seem to have a much larger alpha than those with a p-value above .05 (columns 4 and 5) should not be surprising, given that we are effectively selecting on the alpha estimated ex post in this in-sample exercise. The difference will naturally be more muted once we move to the out-of-sample analysis in the next section. That said, these results are still interesting because they give us an overall sense of the fraction of "skilled" funds in our universe of funds: around 20% of the total. The results suggest that a nontrivial subset of hedge funds does seem to produce a significant alpha, even after accounting for the multiple-testing problem.

A second note relates to the evaluation and comparison of different methodologies. Different methodologies imply different degrees of conservativeness in the choice of funds. Making the bar for selection stricter will mechanically go in the direction of selecting fewer (even if on average better) funds. A methodology produces a clear improvement over alternatives if it achieves better performance (alpha) without sacrificing investment opportunities (i.e., without reducing the number of funds selected and corresponding investible AUM); or, alternatively, if it maintains the same performance but allows investments to be scaled up

[21] While some of the FH7 are based on changes in yields, which are not exactly tradable, they have been treated by the existing literature as tradable, and we do the same here for comparability.

(i.e., it selects more funds); or, ideally, both, achieving better performance on a larger portfolio of funds. Given our theoretical results and corresponding simulations, we would expect the power of our FDR to yield improvements in both dimensions.

That is precisely what we find in Table 2. Consider, for example, columns 1 and 3: both use the FDR control, but column 1 uses our approach with latent factors to estimate the alphas and test statistics, whereas column 3 omits latent factors. Our procedure selects twice as many funds, while achieving a higher average alpha, and it therefore represents a clear improvement over the FDR that omits latent factors. This highlights the importance of accounting for latent factors when testing for alpha.

Columns 6–8 consider a few variations of our approach. Column 6 removes the alpha screening step from our full FDR procedure. Comparing it to column 1, it is clear that this makes little difference in our hedge fund data. Similarly, column 7 removes the alpha screening step from the case with observable factors (column 3), and again, the performance is very close. This is actually not surprising, given that the main use of alpha screening is to remove deep-in-the-null funds (that is, funds with an extremely low alpha). But unskilled funds are less likely to report to TASS, and are also more likely to be removed by our filters (described in Internet Appendix C).[22] Of course, this does not mean that the alpha screening step is not useful in general. We propose a methodology that can be applied in other contexts as well, and different components of each methodology can be more or less useful in different contexts. Column 8 uses 10 latent factors instead of seven observable ones and three latent ones. The performance decreases to the level of the models that only use observable factors. These results show that the best-performing model is the one that mixes observable and latent factors, suggesting that in practice both are useful to properly select funds. Economically motivated observable factors are important to capture dimensions of risk that are not easily captured by factors estimated via machine learning tools, especially in a setting like the one we study here, in which many funds appear only for a few years and in which risk exposures might change significantly over time as funds change their strategies.[23]

Finally, columns 9 and 10 compare our procedure to two others that have tried addressing the multiple testing problem: Harvey and Liu (2018) and Barras et al. (2010), respectively. These methodologies differ from ours in several respects; one thing that is common to both is that they do not employ latent factors or deal directly with the issue of missing data. As the table shows, both methodologies perform worse than our FDR procedure. Harvey and Liu (2018) achieve an

[22] Given that the purpose of our test is to select good funds, it is not a problem for our exercise if "bad" funds do not appear in the data set, whereas it would obviously be more of a problem for studies whose goal is to compute the *average* performance of funds.

[23] Note that even if omitted factors have zero risk premiums, they might still be correlated with the existing factors, so omitting them would still produce biased alphas. So in this setting, it is not sufficient to only focus on priced factors.

average alpha similar to that of Barras et al. (2010), but their procedure selects only a tiny fraction of funds.[24]

Overall, the in-sample analysis shows that the use of FDR control procedures has a significant effect on the fraction of funds selected, and the addition of latent factors allows us to increase the average alpha of the selected funds without sacrificing investment opportunities. Of course, it is hard to judge the effectiveness of the selection procedures without looking out of sample. We turn to that analysis next.

3.5 Out-of-sample analysis and robustness

In this section, we study the out-of-sample performance of portfolios of funds selected using our FDR procedure and using the other methodologies presented above. For each portfolio at each rebalancing date, we use the previous 10 years of data to estimate the alphas and implement the selection, and we only focus on funds with a 36-month track record during that window. We then compute the value-weighted out-of-sample alpha of each portfolio using the estimated time-varying betas and factor risk premiums.

Tables 3 and 4 report all the results of our out-of-sample analysis. In Table 3, we find the alphas in panel A; the p-values for the test that the alphas are equal to zero in panel B; and the p-values for the difference between the alphas of the various methodologies and our full FDR methodology (column 1) in panel C. In Table 4, we report the characteristics of the funds in each portfolio (average fraction of funds selected and average AUM in the two panels). Panel A of Table 4 also reports the average number of funds that are alive at each rebalancing date, as well as the average number of funds used in the estimation (this also includes funds that disappear sometime during the 10 years before rebalancing but were still used to estimate the model). To show that our results are robust to various ways of conducting the analysis, all the panels in these tables contain several rows, corresponding to different specifications.

The first row of each panel contains our baseline specification, which corresponds to the one we used for the in-sample analysis; here, portfolios are rebalanced at the end of each year (we show monthly rebalancing a few rows below). Looking at the first row of each panel in Tables 3 and 4, we can see how the in-sample results from Table 2 change once we go out of sample. The first clear result is that—not surprisingly—it is much harder to identify skilled funds ex ante than ex post. To see this, compare the alphas in columns 4 and 5 of these tables, that is, the alpha of portfolios that select funds with a p-value $< .05$ versus $> .05$. The alphas of the two columns of Table 3 are much closer to each other (11 vs 12 bp) out of sample compared to the same columns

[24] The methodology of Harvey and Liu (2018) requires taking a stand ex ante on the number of true positive alphas in the data. We calibrate this parameter to 20% based on our baseline results in column 1; results are similar when choosing 10% (calibrating this parameter to column 3).

Table 3
Out-of-sample results: Performance

A. Alphas

	Mixed FDR		Only observable			No screen	No screen	Only latent	Alternative models	
	Bootstrap	Avar	FDR	$p < .05$	$p > .05$	FDR	Observ. FDR	FDR	HL	BSW
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Baseline	22.3	21.0	15.1	10.7	12.5	21.8	16.8	13.8	10.4	11.6
Yearly rebal., 1-month lag	22.6	20.8	18.2	12.4	12.8	22.3	17.9	14.2	8.9	14.3
Yearly rebal., 3-month lag	20.9	20.7	17.9	12.2	13.4	20.5	18.7	13.7	9.0	12.0
Monthly rebal.	27.3	26.7	21.9	18.7	14.7	26.9	18.2	23.1	43.5	21.7
Monthly rebal., 1-month lag	25.9	22.6	20.5	16.0	12.8	26.1	21.7	21.1	33.2	20.1
Monthly rebal., 3-month lag	24.9	21.6	19.3	15.5	12.5	24.4	20.8	20.3	−16.1	19.2
Min 60 months	25.1	24.6	18.3	15.7	13.4	26.6	19.6	16.9	14.3	15.7
5 latent factors	20.5	18.3	15.5	10.5	12.3	20.4	18.0	13.1	11.4	10.7
Agarwal and Naik (2004)	22.5	17.0	18.4	2.6	17.0	22.0	18.1	11.3	−27.6	9.2
Evestment, annual rebal.	24.7	20.7	7.4	13.9	16.3	24.6	8.2	17.6	−10.0	14.0
Evestment, monthly rebal.	35.5	39.4	18.6	32.6	3.5	35.8	16.0	25.5	16.0	25.6
Only U.S.	25.4	21.6	15.4	10.3	16.0	24.8	15.4	11.9	10.7	14.4
Equal-weighted returns	25.7	25.0	19.8	16.6	6.8	25.4	19.9	20.8	18.9	19.6

(*Continued*)

Table 3
(Continued)

	Mixed FDR		Only observable			No screen	No screen	Only latent	Alternative models	
	Bootstrap	Avar	FDR	$p < .05$	$p > .05$	FDR	Observ. FDR	FDR	HL	BSW
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
B. p-values										
Baseline	.02	.04	.12	.29	.30	.02	.08	.18	.40	.24
Yearly rebal., 1-month lag	.01	.04	.05	.20	.30	.02	.06	.16	.47	.13
Yearly rebal., 3-month lag	.02	.04	.06	.20	.28	.02	.05	.18	.47	.21
Monthly rebal.	<.01	.01	.03	.06	.25	<.01	.07	.02	<.01	.02
Monthly rebal., 1-month lag	<.01	.04	.05	.11	.33	.01	.03	.05	<.01	.04
Monthly rebal., 3-month lag	.01	.05	.06	.12	.35	.02	.04	.06	.70	.05
Min 60 months	.01	.02	.07	.11	.30	<.01	.06	.11	.10	.11
5 latent factors	.03	.07	.11	.30	.31	.03	.06	.20	.36	.28
Agarwal and Naik (2004)	.07	.19	.10	.83	.27	.07	.11	.38	.65	.45
Evestment, annual rebal.	.16	.23	.66	.49	.39	.17	.63	.29	.52	.48
Evestment, monthly rebal.	.05	.03	.31	.10	.88	.05	.38	.13	.32	.22
Only U.S.	<.01	.03	.10	.30	.16	<.01	.10	.23	.29	.14
Equal-weighted returns	.01	.02	.06	.14	.57	.01	.06	.07	.04	.07
C. p-values, difference from FDR (column 1)										
Baseline	–	.46	.05	<.01	.14	.12	.14	<.01	.34	<.01
Yearly rebal., 1-month lag	–	.38	.21	<.01	.14	.24	.18	<.01	.27	.02
Yearly rebal., 3-month lag	–	.91	.40	.01	.26	.23	.55	<.01	.33	.01
Monthly rebal.	–	.81	.18	.08	.11	.57	.03	.18	.12	.21
Monthly rebal., 1-month lag	–	.30	.19	.04	.10	.93	.17	.18	.58	.20
Monthly rebal., 3-month lag	–	.20	.14	.03	.10	.41	.22	.14	.30	.18
Min 60 months	–	.72	.03	.01	.08	.02	.20	<.01	.06	.02
5 latent factors	–	.32	.15	<.01	.19	.93	.51	<.01	.47	<.01
Agarwal and Naik (2004)	–	.02	.55	<.01	.47	.30	.53	<.01	.40	.08
Evestment, annual rebal.	–	.07	<.01	.21	.62	.47	<.01	.30	<.01	.19
Evestment, monthly rebal.	–	.07	<.01	.66	.10	.15	<.01	.06	.09	.23
Only U.S.	–	.01	<.01	<.01	.11	.25	<.01	<.01	.14	<.01
Equal-weighted returns	–	.64	.13	.02	<.01	.71	.13	.04	.50	.13

This table reports the out-of-sample alphas (panel A) and p-values for the alphas (panel B) for portfolios of hedge funds selected using different methodologies. Panel C reports p-values for the test of the difference between the alphas of each methodology relative to the bootstrap FDR (column 1). Methodologies are described in Table 2. Each row is a different robustness test. The baseline (first row) uses yearly rebalancing, three latent factors, and funds that are alive for at least 36 months. The other robustness tests add a one or three-month lag between portfolio formation and evaluation, add monthly rebalancing, use only funds with at least 60 months of data, use the Agarwal and Naik (2004) model as a benchmark instead of FH7, use the Evestment data set, use only U.S. funds, and use equal-weighted returns. The benchmark model against which alphas are calculated is FH7 with three latent factors. The sample period is 1994–2018.

Table 4
Out-of-sample results: Selection

A. Fraction selected

	Mixed FDR		Only observable			No α-scr.	No α-scr.	Only lat.	Alt. models		# funds	# funds
	Bootstrap	Avar	FDR	$p<.05$	$p>.05$	FDR	Obs. FDR	FDR	HL	BSW	used	alive
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Baseline	.25	.27	.17	.29	.71	.25	.17	.33	.06	.20	982.8	408.0
Yearly rebal., 1-month lag	.25	.27	.17	.29	.71	.25	.17	.33	.06	.20	982.8	408.0
Yearly rebal., 3-month lag	.25	.27	.17	.29	.71	.25	.17	.33	.06	.20	982.8	408.0
Monthly rebal.	.29	.31	.17	.29	.71	.28	.23	.35	.07	.20	475.8	391.7
Monthly rebal., 1-month lag	.28	.31	.17	.29	.71	.28	.17	.35	.07	.20	958.7	392.7
Monthly rebal., 3-month lag	.28	.30	.17	.29	.71	.27	.17	.35	.06	.20	959.4	394.3
Min 60 months	.37	.38	.23	.35	.65	.37	.17	.44	.15	.30	960.5	246.1
5 latent factors	.27	.29	.17	.29	.71	.26	.17	.35	.06	.20	982.8	408.0
Agarwal and Naik (2004)	.23	.27	.08	.22	.78	.22	.08	.28	.01	.12	1066.8	442.0
Evestment, annual rebal.	.45	.45	.36	.47	.53	.45	.35	.58	.22	.53	297.2	244.8
Evestment, monthly rebal.	.46	.45	.38	.48	.52	.46	.38	.55	.23	.57	274.5	230.3
Only U.S.	.31	.34	.22	.34	.66	.30	.21	.43	.07	.23	784.4	326.1
Equal-weighted returns	.26	.27	.17	.29	.71	.25	.17	.33	.06	.20	982.8	408.0

B. Average AUM

	Mixed FDR		Only observable			No screen	No screen	Only latent	Alternative models	
	Bootstrap	Avar	FDR	$p<.05$	$p>.05$	FDR	Obs. FDR	FDR	HL	BSW
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Baseline	53.4	57.8	43.1	63.1	55.9	52.8	41.0	65.6	11.1	47.1
Yearly rebal., 1-month lag	52.6	57.9	42.0	63.1	55.9	51.9	41.6	65.5	11.1	47.2
Yearly rebal., 3-month lag	52.4	57.9	42.8	62.7	56.3	51.8	40.5	65.6	11.0	47.4
Monthly rebal.	55.7	61.0	42.4	62.7	55.6	55.0	32.8	66.1	11.0	47.1
Monthly rebal., 1-month lag	55.5	60.9	42.3	63.0	55.5	54.8	41.4	65.9	11.2	47.8
Monthly rebal., 3-month lag	55.3	60.6	42.7	63.1	55.7	54.6	41.3	65.9	11.1	48.1
Min 60 months	46.2	47.9	33.0	46.7	32.2	46.0	41.4	51.2	17.1	40.8
5 latent factors	56.5	60.7	42.7	63.8	55.2	55.5	40.5	67.8	11.1	47.7
Agarwal and Naik (2004)	52.5	61.1	25.1	58.2	66.0	52.2	24.5	61.1	1.1	37.3
Evestment, annual rebal.	177.4	174.9	129.0	172.8	135.5	176.9	123.7	206.9	73.3	169.0
Evestment, monthly rebal.	162.6	161.1	118.7	156.1	137.6	162.1	116.1	192.3	67.9	164.8
Only U.S.	51.5	56.7	43.8	62.2	46.6	50.7	42.9	67.6	12.5	45.7

The table reports the average fraction selected (panel A) and average portfolio AUM (panel B) for portfolios of hedge funds selected using different methodologies. Rows and columns are the same as in Table 3.

in Table 2. This suggests that using individual p-values estimated on past data to select funds does not actually produce a good selection out of sample.

Applying the FDR control helps. Column 3—which uses the FDR control but no latent factors—shows an improvement in the alpha to 15 bp per month, even ignoring latent factors. Our procedure, which also incorporates three latent factors, further increases the alpha of the portfolio to 22 bp per month. So our procedure produces a portfolio with an alpha that is almost double that produced by standard procedures that ignore latent factors and multiple selection. In addition, Table 4 shows that the portfolio based on our FDR control selects a larger number of funds, and invests in funds with larger AUM, compared to the FDR without latent factors, all while achieving better out-of-sample performance. So, just like in the in-sample case studied before, adding latent factors improves the selection along both dimensions, even though the economic magnitudes of these differences are relatively small, which once again highlights the difficulty of predicting hedge fund performance.

The remaining results for the baseline specification (first row of each panel of Tables 3 and 4) are similar to the in-sample analysis. To summarize without delving into details: alpha screening improves only minimally in our context; using a mixed-factor model dominates using only observable or only latent factors, allowing for a larger investment while maintaining (in fact, improving in almost all cases) the average alpha; finally, our FDR procedure also improves over the selection by the alternative models of Harvey and Liu (2018) and Barras et al. (2010).

The remaining rows of the two tables consider alternative ways to construct the portfolios, alternative observable benchmarks, and alternative data sets altogether. We first consider allowing for a lag between the data used for fund selection and the investment of the portfolio (one month or three months) and increasing the frequency of rebalancing to monthly. We next consider a battery of additional robustness tests for our baseline specification: (a) restricting the estimation only to those funds that have data for at least five years; (b) using five instead of three latent factors; (c) using the Agarwal and Naik (2004) benchmark model (Fama-French-Carhart four factors plus two option factors); (d) using the Evestment data, an entirely different hedge fund data set; (e) restricting to only U.S.-based hedge funds; (f) using equal-weighted instead of value-weighted alphas.

Looking across the various rows, we see that the main patterns of the analysis are robust to these different specifications. Our FDR control procedure with a mixed observable-latent model achieves a higher alpha than the alternatives quite consistently. For example, looking at columns 1 and 3 shows that when comparing our procedure to the FDR control without latent factors, our portfolio selects substantially more funds, and with larger AUM, while achieving a higher alpha (though only slightly higher in some cases). This illustrates the power of our procedure and the importance of accounting for latent factors when applying the FDR control.

It is also remarkable how stable the alphas are for our methodology across the various specifications: all between 20.5 and 35.5 bp per month, in contrast with many of the alternatives, whose performance appears much more variable across rows (the most extreme case of this appears to be Harvey and Liu [2018], which also tends to select a much smaller number of funds).

Overall, these results show that our FDR methodology performs well out of sample, robustly selecting a larger number of funds with better performance, compared to alternative methodologies.

4. Conclusion

This paper presents a rigorous framework to address the data-snooping concerns that arise when applying multiple testing in the asset pricing context. In situations in which many tests are performed, many "false discoveries" should be expected: cases in which the significance of some of the tests is obtained by pure chance. The rate of false discoveries is hard to evaluate ex ante; and it can grow unboundedly when standard statistical tests are used as the number of tests performed increases.

Statistical theory has proposed different methods that aim to control and mitigate this data-snooping problem, like the so-called "false discovery rate" control test of Benjamini and Hochberg (1995). But these methods do not work in the standard asset pricing context, in which some of the main assumptions for the procedures are violated. In the paper, we show that the FDR control test can be extended and generalized to be valid under a much broader range of assumptions, specifically those that appear crucial when thinking about testing for alphas in linear factor models.

Our paper exploits the "blessing of dimensionality" to build an FDR control test that is valid when the benchmark includes nontradable factors whose risk premiums need to be estimated, and is robust to the presence of omitted factors and an unbalanced data panel. In addition, contrary to existing multiple-testing methods, our test is built explicitly to handle large cross-sections; this makes it particularly suitable for many finance applications, in which the size of the cross-section N can be large relative to the sample size T.

We illustrate this procedure by applying it to the evaluation of hedge fund performance, a typical example where multiple testing issues arise. We show empirically that hedge fund returns are highly correlated in the cross-section, even after controlling for the standard models. We show that our procedure—which allows for such correlation and bounds the false discovery rate to a predetermined level—produces superior in- and out-of-sample results compared to several standard methodologies, including some that have been designed to specifically deal with multiple testing.

The last few years have seen a burgeoning strand of literature on the applications of machine learning techniques to high-dimensional problems in asset pricing, in which data snooping leads to potentially numerous false

discoveries. Our paper proposes a way to rigorously account for the data-snooping bias, taking into account explicitly the specific properties of the finance context to which it is applied. There remain many other settings in which our high-dimensional multiple-testing framework can be applied: for example, the evaluation of multiple potential new factors against an existing asset pricing model. We leave the study of these applications to future research.

References

Ackermann, C., R. McEnally, and D. Ravenscraft. 1999. The performance of hedge funds: Risk, return, and incentives. *The Journal of Finance* 54:833–74.

Agarwal, V., G. Bakshi, and J. Huij. 2009. Do higher-moment equity risks explain hedge fund returns? Technical Report, Georgia State University.

Agarwal, V., V. Fos, and W. Jiang. 2013. Inferring reporting-related biases in hedge fund databases from hedge fund equity holdings. *Management Science* 59:1271–89.

Agarwal, V., and N. Y. Naik. 2000. Multi-period performance persistence analysis of hedge funds. *Journal of Financial and Quantitative Analysis* 35:327–42.

———. 2004. Risks and portfolio decisions involving hedge funds. *The Review of Financial Studies* 17:63–98.

Aggarwal, R. K., and P. Jorion. 2010. The performance of emerging hedge funds and managers. *Journal of Financial Economics* 96:238–56.

Ahn, S., and A. Horenstein. 2013. Eigenvalue ratio test for the number of factors. *Econometrica* 81:1203–27.

Ang, A., and D. Kristensen. 2012. Testing conditional factor models. *Journal of Financial Economics* 106:132–56.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. 2018. Matrix completion methods for causal panel data models. Technical Report, National Bureau of Economic Research.

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

———. 2019. Rank regularized estimation of approximate factor models. *Journal of Econometrics* 212:78–96.

Bajgrowicz, P., and O. Scaillet. 2012. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics* 106:473–91.

Bali, T. G., S. J. Brown, and M. O. Caglayan. 2011. Do hedge funds' exposures to risk factors predict their future returns? *Journal of Financial Economics* 101:36–68.

———. 2014. Macroeconomic risk and hedge fund returns. *Journal of Financial Economics* 114:1–19.

Baquero, G., J. Ter Horst, and M. Verbeek. 2005. Survival, look-ahead bias, and persistence in hedge fund performance. *Journal of Financial and Quantitative Analysis* 40:493–517.

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

Benjamini, Y., and Y. Hochberg. 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.

Benjamini, Y., and W. Liu. 1999. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* 82:163–70.

Benjamini, Y., and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29:1165–88.

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

Candès, E. J., and T. Tao. 2010. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory* 56:2053–80.

Chen, Y. 2019. Individual stock-picking skills in active mutual funds. Available at: SSRN 3324672.

Chernozhukov, V., D. Chetverikov, and K. Kato. 2013. Testing many moment inequalities. Technical Report, Massachusetts Institute of Technology.

Chordia, T., A. Goyal, and A. Saretto. 2020. Anomalies and false rejections. *Review of Financial Studies* 33:2134–79.

Cochrane, J. H. 2009. *Asset pricing: Revised edition*. Princeton NJ: Princeton University Press.

Donoho, D. L. 2000. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture* 1–32.

Efron, B. 2010. Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association* 105:1042–55.

Fama, E. F., and K. R. French. 2010. Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance* 65:1915–47.

Fan, J., and X. Han. 2016. Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79:1143–64.

Fan, J., X. Han, and W. Gu. 2012. Estimating False Discovery Proportion Under Arbitrary Covariance Dependence. *Journal of the American Statistical Association* 107:1019–35.

Feng, G., S. Giglio, and D. Xiu. 2020. Taming the factor zoo: A test of new factors. *The Journal of Finance* 75:1327–70.

Freyberger, J., A. Neuhierl, and M. Weber. 2017. Dissecting characteristics nonparametrically. Technical Report, University of Wisconsin-Madison.

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

———. 2001. The risk in hedge fund strategies: Theory and evidence from trend followers. *The Review of Financial Studies* 14:313–41.

———. 2004. Hedge fund benchmarks: A risk-based approach. *Financial Analysts Journal* 60:65–80.

Fung, W., D. A. Hsieh, N. Y. Naik, and T. Ramadorai. 2008. Hedge funds: Performance, risk, and capital formation. *The Journal of Finance* 63:1777–803.

Getmansky, M., P. A. Lee, and A. W. Lo. 2015. Hedge funds: A dynamic industry in transition. *Annual Review of Financial Economics* 7:483–577.

Giglio, S., and D. Xiu. 2017. Asset pricing with omitted factors. Technical Report, Yale University and University of Chicago.

Green, J., J. R. M. Hand, and X. F. Zhang. 2013. The supraview of return predictive signals. *Review of Accounting Studies* 18:692–730.

Gu, S., B. Kelly, and D. Xiu. 2018. Empirical asset pricing via machine learning. Technical Report, University of Chicago.

Hall, P., and S. R. Wilson. 1991. Two guidelines for bootstrap hypothesis testing. *Biometrics* 47:757–62.

Hansen, P. 2005. A test for superior predictive ability. *Journal of Business and Economic Statistics* 23:365–80.

Harvey, C. R., and Y. Liu. 2018. False (and missed) discoveries in financial economics. Technical Report, Duke University.

Harvey, C. R., Y. Liu, and H. Zhu. 2015. ... and the cross-section of expected returns. *Review of Financial Studies* 29:5–68.

Holm, S. 1979. A Simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65–70.

Horowitz, J. L. 2001. The bootstrap. In J. J. Heckman and E. Leamer, eds., *Handbook of Econometrics*, vol. ed. 5, 3159–228. Amsterdam: North Holland/Elsevier.

Jagannathan, R., A. Malakhov, and D. Novikov. 2010. Do hot hands exist among hedge fund managers? An empirical evaluation. *The Journal of Finance* 65:217–55.

Kelly, B., S. Pruitt, and Y. Su. 2017. Some characteristics are risk exposures, and the rest are irrelevant. Technical Report, University of Chicago.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov. 2011. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* 39:2302–29.

Kosowski, R., N. Y. Naik, and M. Teo. 2007. Do hedge funds deliver alpha? A Bayesian and bootstrap analysis. *Journal of Financial Economics* 84:229–64.

Kosowski, R., A. Timmermann, R. Wermers, and H. White. 2006. Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. *The Journal of Finance* 61:2551–95.

Kozak, S., S. Nagel, and S. Santosh. 2017. Shrinking the cross section. Technical Report, University of Michigan.

Leek, J. T., and J. D. Storey. 2008. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* 105:18718–23.

Liang, B. 1999. On the performance of hedge funds. *Financial Analysts Journal* 55:72–85.

———. 2001. Hedge fund performance: 1990-1999. *Financial Analysts Journal* 57:11–8.

Liu, R. Y. 1988. Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics* 16:1696–708.

Lo, A. W., and A. C. MacKinlay. 1990. Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies* 3:431–67.

Ma, S., D. Goldfarb, and L. Chen. 2011. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming* 128:321–53.

Mammen, E. 1993. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* 21:255–85.

Mitchell, M., and T. Pulvino. 2001. Characteristics of risk and return in risk arbitrage. *The Journal of Finance* 56:2135–75.

Moon, H. R., and M. Weidner. 2018. Nuclear norm regularized estimation of panel regression models. arXiv, preprint, arXiv:1810.10987.

Patton, A. J., and T. Ramadorai. 2013. On the high-frequency dynamics of hedge fund risk exposures. *The Journal of Finance* 68:597–635.

Romano, J. P., and M. Wolf. 2005. Stepwise multiple testing as formalized data snooping. *Econometrica* 73:1237–82.

———. 2018. Multiple testing of one-sided hypotheses: Combining bonferroni and the bootstrap. In *International conference of the Thailand Econometrics Society*, ed. 753, 78–94. Cham: Springer.

Schwartzman, A., and X. Lin. 2011. The effect of correlation in false discovery rate estimation. *Biometrika* 98:199–214.

Simes, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73:751–4.

Sinclair, A. J. 2018. The allocative role of prime brokers. Technical Report, The University of Hong Kong.

Stock, J., and M. Watson. 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* 20:147–62.

Storey, J. D. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64:479–98.

Storey, J. D., J. E. Taylor, and D. Siegmund. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66:187–205.

Su, L., K. Miao, and S. Jin. 2019. On factor models with random missing: Em estimation, inference, and cross validation. Technical Report, Singapore Management University.

White, H. 2000. A reality check for data snooping. *Econometrica* 68:1097–126.

Yan, X., and L. Zheng. 2017. Fundamental analysis and the cross-section of stock returns: A data-mining approach. *Review of Financial Studies* 30:1382–423.