NBER WORKING PAPER SERIES

WHAT THRESHOLD SHOULD BE APPLIED TO TESTS OF FACTOR MODELS?

Campbell R. Harvey
Alessio Sancetta
Yuqian Zhao

Working Paper 34898
http://www.nber.org/papers/w34898

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2026

I have read the NBER disclosure and attest that this acknowledgment discloses all sources of funding and all material and relevant financial relationships. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

What Threshold Should be Applied to Tests of Factor Models?
Campbell R. Harvey, Alessio Sancetta, and Yuqian Zhao
NBER Working Paper No. 34898
February 2026
JEL No. C12, C58, G11, G12

ABSTRACT

Researchers generally acknowledge that statistical tests must be adjusted when hundreds of factors and trading strategies have been examined. But how should these adjustments be made? Existing methods are often misunderstood or misapplied. We show that proper inference requires accounting for dependence across tests, correctly specifying the null distribution, and mitigating sample-selection bias. We develop a simple framework that avoids assumptions about the total number of tests run and yields a lower bound on valid significance thresholds - implying that researchers should employ a *t*-statistic cutoff of at least 3.0. In addition, we advocate using the local False Discovery Rate, which provides the probability that the null hypothesis is true for a given test-statistic realization - information that a conventional *p*-value cannot supply.

Campbell R. Harvey
Duke University
Fuqua School of Business
and NBER
cam.harvey@duke.edu

Yuqian Zhao
University of Sussex
Yuqian.Zhao@sussex.ac.uk

Alessio Sancetta
Royal Holloway, University of London (RH)
asancetta@gmail.com

1 Introduction

The finance profession has good reason to be puzzled by the recent debate on whether there is a replication crisis in the profession. Some researchers have argued that a large proportion of published findings in the factor literature are likely false and suggest that p-values need to be substantially inflated to reduce Type I errors (Harvey, Liu and Zhu (HLZ hereafter), 2016; Linnainmaa and Roberts, 2018; Harvey and Liu (HL), 2020; Hou, Xue and Zhang (HXZ), 2020; Chordia, Goyal, and Saretto (CGS), 2020). Others, while acknowledging that some adjustment needs to be made for multiple tests, offer minor inflation, thereby suggesting that the overwhelming number of published discoveries is robust (Chen and Zimmermann, 2018; Chen, 2021; Jensen, Kelly, and Pedersen (JKP), 2023; Chen, 2025). Hence, researchers fundamentally disagree on how to interpret a given set of empirical results.

Most researchers use the classical statistical testing framework, employing a critical value of 2.0 for t-statistics when testing the null hypothesis, for example, testing whether the premium on a proposed factor is zero. However, this approach is only valid if the p-value is derived with a single test. Given that hundreds of factors have been published, along with an unknown number of unpublished tests, we naturally encounter a multiple hypothesis testing problem. That is, with a large number of tests, some factors will seem "significant" purely as a result of luck leading to Type I errors. Test statistics such as p-values need to be inflated, and the question is by how much. The goal of our paper is to try to sort out this problem and to provide a practical guide for researchers making statistical inference.

Consider the extent of the disagreement. HLZ (2016) offer three different adjustment methods – all of which suggest substantial inflation of p-values. In contrast, JKP (2023) apply a Bayesian approach, leading to conclusions that yield the same critical values as conventional single-test p-values. Further, JKP look at the out-of-sample performance of some marginally significant factors and find comparable performance in the in-sample and out-of-sample periods, which strengthens their claim that there is no replication crisis. But, as we will show later, JKP's novel approach is contingent on a key assumption that influences their results. In addition, Chen (2025) also argues that the overwhelming number of factors are true. He presents analyses based on both simulated and published factors, suggesting only a minor inflation in p-values. Indeed, his tightest bound suggests that 91% of the published factors are "true".

We replicate these previous results and reconcile the findings. We identify potential drivers of the contradictory evidence. To begin with, it is crucial to take into account any correlation among the factors. Correlation is important because

adjustment methods in the statistics literature generally assume the independence of the tests. Although the correlation issue can be somewhat mitigated by using abnormal factor returns (i.e., adjusting returns for the market factor), this is not sufficient. Dependence in factor returns can cause the empirical distribution of t-statistics under the null hypothesis – that the factor premium is zero – to deviate significantly from the theoretical standard normal (see Efron 2007a, 2007b, Fan et al., 2012). Ignoring this can lead to a large number of false discoveries. Therefore, accurately approximating the empirical null distribution is crucial for reliable inference.

In addition, sample selection can impact the results. It is important to control for the number of hypotheses, including published and unpublished tests. HLZ (2016), HXZ (2020), JKP (2023), and Chen (2024) only consider published factors. Hence, they ignore sample selection bias. This can substantially change the distribution of the test statistics and invalidate the control of false discovery rate and related quantities as the number of hypotheses is unknown.

Our paper has multiple goals. First, we seek to clarify the statistical methods that are used in the profession. We have already mentioned how correlation among tests and distributional assumptions can impact inference when using standard methods. There are other fundamental issues, such as the specification of the probability that the null hypothesis is true. We introduce a simple method for identifying this key input. We also provide insight on the False Discovery Rate (FDR), a term that is used frequently but is not well understood. Unlike the family-wise error rate (FWER), which measures the probability of making at least one false discovery, the FDR is the expected proportion of false discoveries among the results deemed significant. The two measures capture different types of error and are best viewed as complementary rather than competing. We also introduce the *local* FDR (LFDR), which estimates the chance that a specific finding is a false positive. This provides a simple, intuitive way to assess individual factor candidates, still controlling for multiple testing. In fields where a single false result can have severe consequences, FWER remains the norm, while in finance or clinical settings, some tolerance for occasional errors makes FDR-based tools more appropriate.[1]

The LFDR provides a direct estimate of the probability that a discovery is false,

[1]The k-FWER is an extension of the FWER that allows for up to k false positives (Romano and Wolf, 2007). It provides a valid alternative to FDR control, though k-FWER procedures are typically less powerful than FDR-based methods and do not scale as easily to very large numbers of tests because of their computational demands.

which a p-value does not convey.[2] We recommend that researchers report the LFDR. We introduce a method that avoids specifying a full prior density, making the use of the LFDR essentially as straightforward as reporting a p-value. Although our method still requires specifying the unconditional probability of the null, $\mathbb{P}(H_0)$, the results in Table 3 and the analysis in Section A.3 of the Appendix show that the findings are not highly sensitive to its choice. Moreover, reporting the LFDR for different values of $\mathbb{P}(H_0)$ is advantageous, as researchers may reasonably lean toward smaller values of $\mathbb{P}(H_0)$ when a finding is well supported by theory.

Second, we reconcile existing results in the literature. In particular, we explain what drives Chen's (2025) result that a very large number of factors are likely true. We also revisit the JKP paper.

Third, and most importantly, we provide guidance for researchers in finance as to what cutoff should be used to assess statistical significance. We provide an objective method that does not depend on the specification of the number of tests. Of course, any method requires some assumptions. We make the case that our assumptions are minimal and we provide a grid of possible choices so that the researcher can select what they are comfortable with. Importantly, across our grid, there is a common message: p-values need to be inflated (the t-statistic threshold needs to increase). We derive a lower bound that suggests that the minimum t-statistic is approximately 3.0 for factor research. Methods to control the FDR require knowledge of the number of hypotheses tested because they rely on the empirical distribution of the t-statistics (Benjamini and Hochberg, 1995, Efron et al., 2001, Genovese and Wasserman, 2004). We further discuss this in our Online Supplement. The LFDR requires an estimator of the density of the t-statistics. However, we bypass this challenge by deriving a lower bound that can be implemented without knowledge of this density

Finally, we reflect on the plausibility of the assumptions embedded in our method as well as some important outstanding issues that go beyond our current research. For example, we sketch an approach that might be useful for dealing with factors that appear significant but are just duplicates of other factors. In addition, our approach does not allow for differential costs of Type I and Type II errors. Finally, we also discuss some limitations of our proposed approach in the context of theory-based versus data-mined factors.

Our paper is organized as follows. The second section explores the meaning and interpretation of the false discovery rate. The third section details a number of statistical issues with the current literature. In particular, the assumption

[2]The American Statistical Association's statement on p-values (Wasserstein and Lazar, 2016), "p-values do not measure the probability that the studied hypothesis is true," and "by itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis."

of a standard normal distribution combined with the assumption that factors are uncorrelated leads to an inflation in the number of factors that are declared "significant". The fourth section offers some guidance for researchers. Here we introduce the local false discovery rate and detail a method to get cutoffs to control error rates that do not depend on the specification of the number of tests. In the fifth section, we detail some unresolved issues that go beyond the paper. In addition, we critically assess all the assumptions necessary for our method to produce the recommended cutoffs. Some concluding remarks are offered in the final section.

2 Setting

We begin by describing the nature of hypothesis tests, the inputs required, and the concept of the false discovery rate (FDR).

2.1 Factor models and tests

Since many statistical methods assume zero cross-correlation, our first step will be to remove some of the common variation from the factor returns. Given the market excess return r_t^m and N factor anomalies, estimate the following market model:

$$f_{t,i} = \alpha_i + \beta_i r_t^m + \varepsilon_{t,i}, \tag{1}$$

where $f_{t,i}$ is the excess return at time t on the i^{th} factor, α_i and β_i are unknown parameters and $\varepsilon_{t,i}$ is the error term. In existing studies, time is usually measured at monthly frequencies over multiple decades, and the number N of factors is large, possibly thousands, when examining ratio-based combinations of trading signals (e.g., Yan and Zheng, 2017), or on the order of hundreds when evaluating significant factors from existing published work (e.g., JKP, 2023).

The factor discovery process leads researchers to test the null hypothesis H_0^i : $\alpha_i = 0$ for $i = 1, 2, ..N$. The alternative hypothesis H_1^i can be one sided ($H_1^i : \alpha_i > 0$) or two sided ($H_1^i : \alpha_i \neq 0$). Rejection of the null hypothesis in favor of the alternative is interpreted as identifying a significant factor. Some studies, such as JKP (2023), use published factors that are assumed to carry positive alphas and, as a result, are already signed. In this case, their hypotheses are one-sided, meaning that only true positive alphas are considered discoveries. Other studies, such as HXZ (2020) and Chen (2025), consider two-sided hypotheses, that is, a large negative t-statistic is considered a discovery because the sign can always be flipped by reversing the long and short positions.

Allowing for sign flipping with N candidate factors is conceptually equivalent to testing $2N$ factors, since each factor can be taken in either a long or short position. We ignore this in what follows.[3] In what follows, we refer to null t-statistics (or p-values) as those corresponding to $\alpha_i = 0$. Accordingly, in this paper we use two-sided alternatives, as this simplifies the arguments, and we disregard the fact that we are cherry-picking the sign of the factors.

Relying on p-values of t-statistics at conventional levels to test the assertion $\alpha_i \neq 0$ can be misleading. This is because the probability of observing a large t-statistic increases with the number of factors tested. It is inappropriate to use thresholds based on a single test framework when many tests are conducted. This is not a new issue, and many different approaches have been proposed in the statistics literature. The key question is: What threshold should be used?

To overcome this problem, out-of-sample validation can be performed. For example, McLean and Pontiff (2016) evaluate the robustness of predictive variables using out-of-sample data collected after their publication. JKP probe the devitalization of alpha using out-of-sample data post-publication. Similarly, Baltussen et al. (2023) examine the cross-section of stock returns using out-of-sample data that historically precede publications. Linnainmaa and Roberts (2018) investigate the validity of accounting-based return anomalies using out-of-sample data, allowing for time to move both backward and forward. Although out-of-sample testing typically provides reliable validation results, HLZ (2016) highlight its impracticality due to the scarcity of real-time data. Alternatively, a more feasible approach to handle multiple hypothesis testing is to control the false discovery rate using statistical frameworks.

2.2 The false discovery rate

The objective is to find a procedure capable of controlling the expectation of the False Discovery Proportion (FDP). The False Discovery Rate (FDR) is the expectation of the FDP,

$$\text{FDP} := \frac{\sum_{i=1}^{N} \delta_i \left(1 - A_i\right)}{\sum_{i=1}^{N} \delta_i}. \tag{2}$$

Here, the numerator counts the total number of false discoveries, where δ_i equals 1 if hypothesis H_0^i is rejected by a data-driven testing procedure and zero otherwise, while A_i is a binary random variable equal to 1 when the alternative hypothesis H_1^i

[3] While ignoring this increases the number of discoveries, when the number of tests is large the effect is negligible.

is true and zero otherwise. This is just the ratio of false positives divided by the sum of both the false and true positives. The definition in (2) cannot be used directly because the random variables A_i are unobservable. If we knew $A_i = 1$, we would then know that the i^{th} hypothesis is true. This is exactly what we are trying to infer from a testing procedure. However, the definition in (2) is the starting point for much theory and approximations, as we shall see.

The common data-driven method sets $\delta_i = 1$ if the t-statistic is greater than a prespecified cutoff point. The challenge lies in estimating A_i, as it requires knowing whether hypothesis H_0^i is genuinely rejected. In Section A.1.1 of the Appendix, we detail a method to estimate this quantity. For a rejection rule based on the t-statistics, researchers aim to choose the cutoff so that the expectation of FDP, that is, FDR, is less than a number in $(0, 1)$. Thus, having fixed a number, say 0.05, the goal is to choose a cutoff as small as possible but such that

$$\text{FDR} = \mathbb{E}\text{FDP} \leq 0.05. \tag{3}$$

In this case, the test procedure is said to control the FDR at a level 0.05; 0.1 is another common choice. The interpretation when the level is 0.05 is that it chooses a t-statistic cutoff point such that no more than 5% of the expected total rejections are false discoveries.

To provide some intuition, let us consider the case of independent identically distributed (i.i.d.) hypotheses and define FDP in (3) to be zero when the denominator is zero.[4] Then, using $|T|$ to denote the absolute value of the t-statistic for any of the two-sided hypotheses, we have that

$$\text{FDR} = \frac{\mathbb{P}\left(|T| \geq \tau \middle| H_0\right)\mathbb{P}\left(H_0\right)}{\mathbb{P}\left(|T| \geq \tau\right)} = \mathbb{P}\left(H_0 \middle| |T| \geq \tau\right), \tag{4}$$

where the first equality follows from Storey (2003, Theorem 1) and the second by Bayes' rule.[5] That is, for a Bayesian, the FDR is the probability that the null is true given that the observed t-statistics is greater than τ. Because of the i.i.d. condition, we drop the reference to a specific hypothesis i in both the t-statistic

[4]The i.i.d. condition is not required – we assume it for convenience, but the results hold for a large number N of hypotheses, under more general conditions, as discussed in Section S.1 of the Online Supplement.

[5]In this framework, the hypotheses are random variables, and the probability that the i^{th} null is true is $\mathbb{E}[1 - A_i] = \mathbb{P}(H_0^i)$. This probability is assumed to be the same for all i, so we drop the superscript. From a frequentist perspective, $\mathbb{P}(H_0)$ can be interpreted as the proportion of true nulls, which we assume is approximately constant when the number of hypotheses is large (see Section S.1 in the Online Appendix for details)

T_i and the null hypothesis H_0^i. The quantity τ is the critical value used to reject the null hypothesis. Hence, to compute FDR for a given critical value, we need estimates of the probability of the null hypothesis being true, the probability of the null t-statistics and the denominator in (4), which is the unconditional probability of $|T|$ being greater than τ. Failure to accurately estimate any of these quantities leads to incorrect inference about the FDR implied by a given critical value.

The role of Type I and Type II errors is evident by noting that the denominator in (4) can be written:

$$\mathbb{P}\left(|T| \geq \tau\right) = \underbrace{\mathbb{P}\left(|T| \geq \tau \middle| H_0\right)}_{\text{Type I error rate}} \mathbb{P}\left(H_0\right) + \underbrace{\mathbb{P}\left(|T| \geq \tau \middle| H_1\right)}_{\text{Power}} \mathbb{P}\left(H_1\right), \tag{5}$$

which follows by the basic rules of conditional probability. Substituting (5) in the first equality in (4) and rearranging, it can be shown that:[6]

$$\text{FDR} = \left(1 + \frac{\mathbb{P}\left(|T| \geq \tau \middle| H_1\right)\mathbb{P}\left(H_1\right)}{\mathbb{P}\left(|T| \geq \tau \middle| H_0\right)\mathbb{P}\left(H_0\right)}\right)^{-1} = \left(1 + \frac{\text{Power} \times \mathbb{P}\left(H_1\right)}{(\text{Type I error rate}) \times \mathbb{P}\left(H_0\right)}\right)^{-1}.$$

The above display shows that a large cutoff will reduce the FDR.[7] A large cutoff reduces the Type I error (the probability of rejecting the null when the null is true). Since power is the probability of avoiding Type II errors (the probability of not rejecting the null when it is, in fact, false), the large cutoff increases a Type II error, so the term in large parentheses increases. Large Type II errors can lead to economic losses, as emphasized by CGS (2020) and HL (2020). So, a reasonable goal is to find a cutoff that is as small as possible, but still large enough so that the FDR is less than 5%, as emphasized in Equation (3).

The second equality in (4) highlights an important distinction between two related concepts: the False Discovery Rate (FDR) and the *local* false discovery rate (LFDR). The FDR is defined as the expected proportion of false discoveries among all hypotheses with test statistics exceeding a threshold τ. That is, it conditions on the event $\{|T| \geq \tau\}$ and assesses the average false discovery proportion among all rejections beyond this cutoff. As a result, it does not take into account the specific value of the test statistic observed in a given instance. In contrast, the LFDR – a closely related concept introduced next – is defined for a specific value of the test statistic. It estimates the probability that a hypothesis is null given that we observe

[6] We divide the numerator and denominator in (4) by $\mathbb{P}\left(|T| \geq \tau \middle| H_0\right)\mathbb{P}\left(H_0\right)$.

[7] This occurs because $|T|$ tends to take larger values under H_1 than under H_0, so the probability of $\{|T| \geq \tau\}$ decreases faster under H_0 than under H_1 as τ increases.

a particular value, $|T| = \tau$. The local FDR (denoted by LFDR) is formally defined as:

$$\text{LFDR} = \mathbb{P}(H_0 \big| \, |T| = \tau). \tag{6}$$

This reveals the probability that the null hypothesis is true, given the observed test statistic. Hence, it is a quantity in its own right and has been used as a competitor to p-values (Berger and Delampady, 1987; see Section 4 for more details).

When assessing a specific hypothesis among many, we set τ equal to the realized value of $|T|$ and compute the corresponding LFDR. In this sense, the FDR for a fixed cutoff (e.g., $\tau = 3$) can be interpreted as the average of the LFDR values for all cases where $|T| \geq 3$. Since it is an average, it is smaller than the corresponding LFDR when the LFDR is decreasing in τ.

In Section 3, we discuss the problems in the existing literature and focus on the FDR, in line with previous studies where FDR control is the standard approach. We return to the use of the local FDR in Section 4. Table 1 summarizes the main notation used in the paper.

Table 1: Main Notation. In the text, we may suppress the subscript i when the hypotheses are identically distributed.

Symbol	Meaning	Additional Comments
H_0^i	the i^{th} null hypothesis	$\alpha_i = 0$ (α_i in (1))
H_1^i	the i^{th} alternative hypothesis	$\alpha_i \neq 0$
A_i	the unobservable binary random variable equal to 1 if H_1^i is true and zero otherwise	this is the quantity we are after: if $A_i = 1$ we know that the i^{th} factor is an anomaly
a	the expectation of A_i, i.e. $a = \mathbb{E}A_i = \mathbb{P}(H_1^i)$	for ease of exposition we drop the i subscript, assuming that this is the same across hypotheses; otherwise it can be interpreted as fraction of true alternative hypotheses
T_i	the random variable representing the i^{th} t-statistic	used to test the i^{th} hypothesis
τ	a constant or a variable	used to denote a threshold or a value taken by $\lvert T_i \rvert$
P_i	the random variable representing the i^{th} p-value	obtained using the distribution of $\lvert T_i \rvert$ under the null H_0^i
N	the total number of hypotheses	number of factors
n	the sample size of testing factors	number of time-series observations

3 The Statistical Issues with Current Literature

We begin with the dataset from Chen et al. (2024), which contains 29,314 randomly generated factors. This dataset is constructed using 242 variables, including 241 Compustat accounting variables studied in Yan and Zheng (2017) and the CRSP market equity return. Generation rules for simple ratios (X/Y) and first differences scaled by a lagged denominator ($\Delta X/\text{lag}(Y)$) are applied to create a total of 29,314 ratios. Specifically, the numerators are selected from the full set of 242 variables, while the denominators are limited to 65 variables that exhibit positive values for at least 25% of firms in 1963, with corresponding CRSP data, ensuring the avoidance of normalizing by zeros or negative values. The dataset spans from July 1952 to

December 2022, forming an unbalanced panel with missing observations included. For convenience, we refer to it as the 29K dataset below. Since the anomalies are generated from portfolios sorted on randomly selected characteristics, the sample is effectively randomized by construction and therefore not subject to sample selection bias. As a result, the publication bias issues highlighted in HLZ (2016) and Chen (2024) are not operational, and the estimated FDR remains unbiased.

We replicate Chen (2025)'s procedure and obtain the proportion of factor discoveries as $\mathbb{P}(|T| \geq 2) \simeq 0.33$ from this dataset, based on the frequency of absolute t-statistics of equal-weighted raw returns exceeding 2.[8] Chen (2025) argues that at least 91% of these discoveries are true (Chen, 2025, Figure 2). This is derived using his "easy bound", where the standard normal distribution is used in the numerator of (4). That is, he substitutes 0.05 for $\mathbb{P}\left(|T| \geq 2 \middle| H_0\right)$ and $\mathbb{P}(|T| \geq 2) = 0.33$:

$$\text{FDR} = \frac{0.05 \times \mathbb{P}(H_0)}{\mathbb{P}(|T| \geq 2)}. \tag{7}$$

He then estimates $\mathbb{P}(H_0) \simeq 0.56$ using a visual heuristic similar to the method of Storey et al. (2004); substituting this value into the above equation yields FDR $\simeq$ 0.09.[9] This finding is encouraging because it suggests that using a threshold for t-statistics as low as 2.0 is associated with a low (9%) false discovery rate.

3.1 Do not trust the standard normal distribution under the null

The promising findings of Chen (2025) rely on the key assumption that the t-statistics follow a standard normal distribution under the null hypothesis, such that his easy bound (7) holds.

Chen (2025) claims that the assumption of standard normality for the t-statistic under the null can be verified via bootstrap methods, as shown in Chen (2021), which shows that the distribution of bootstrapped t-statistics is indistinguishable from the standard normal. However, this reflects a misunderstanding of both the proper use of the bootstrap and, more importantly, the impact of correlated hypotheses. We address these issues in what follows.

First, we explain why correlation among hypotheses is problematic and why the 0.05 value in the numerator of Chen's easy bound in (7) is inappropriate

[8] Chen considers $\mathbb{P}(|T| > 2)$ rather than the more conventional $\mathbb{P}(|T| \geq 2)$; this makes no practical difference.

[9] The estimator for $\mathbb{P}(H_0)$ is reviewed in detail in Section A.1.1 of the Appendix.

for controlling the FDR under dependence. We believe that the impact of correlated hypotheses is not fully appreciated in the finance literature, leading to misunderstandings such as those found in Chen (2025).[10] Furthermore, we show that Chen's (2025) reliance on the bootstrap is not appropriate in this context. As is well known in theory, the bootstrap recovers the critical values of the standard normal distribution. However, as we show below, these values are not appropriate under dependence, as the resulting cutoffs are too low and lead to an excessive number of Type I errors.

Consider an illustrative model for the N factor returns in which the i^{th} factor at time t is given by

$$f_{t,i} = \alpha_i + \beta_i r_t^m + b_i v_t + \eta_{t,i},$$

where v_t is an unobserved common latent component that induces correlation across factors, and $\eta_{t,i}$ is an idiosyncratic term. This decomposition captures the portion of factor comovement that is not driven by the market.[11] To simplify the discussion, assume factors orthogonalized against the market return, which is equivalent to setting $\beta_i = 0$. The model then reduces to:

$$f_{t,i} = \alpha_i + \underbrace{b_i v_t + \eta_{t,i}}_{\varepsilon_{t,i}}, \tag{8}$$

where v_t has mean zero, b_i is its loading for factor i, and $\eta_{t,i}$ is mean zero, uncorrelated across i and uncorrelated with v_t. Suppose v_t and $\eta_{t,i}$ are Gaussian with mean zero, the variance of each factor is known and denoted by $\sigma^2_{f,i}$, while the variance of $\eta_{t,i}$ is $\sigma^2_{\eta,i}$. The t-statistic for the mean of the i^{th} factor in: (8) is

$$T_i = \frac{\sqrt{n}\,(\alpha_i + b_i\bar{v} + \bar{\eta}_i)}{\sigma_{f,i}}, \tag{9}$$

where $\bar{v} = n^{-1}\sum_{t=1}^{n} v_t$ and $\bar{\eta}_i = n^{-1}\sum_{t=1}^{n} \eta_{t,i}$.[12] For simplicity, we use the population standard deviation $\sigma_{f,i}$ in the denominator.

[10] For example, Chen (2025) claims that FDR bounds are too conservative for correlated hypotheses, citing Schwartzman and Lin (2011), who study a bias that can arise for the estimator of Benjamini and Hochberg (see Section S.2 in the Online Supplement for a clarification of this point).

[11] A multivariate latent component could capture a general correlation structure; we use a scalar component for illustration.

[12] Under the null $\alpha_i = 0$, the t-statistic equals the sample mean of the factor $(b_i\bar{v} + \bar{\eta}_i)$ divided by the standard error $\sigma_{f,i}/\sqrt{n}$.

Under the null hypothesis $H_0^i: \alpha_i = 0$, T_i is unconditionally standard normal. However, because v_t is common across factors, the null t-statistics T_i share dependence through $\bar{v}$. Conditional on a realization of $\bar{v}$, we have:

$$T_i \mid \bar{v} \sim \mathcal{N}\left(\underbrace{\frac{\sqrt{n}\, b_i \bar{v}}{\sigma_{f,i}}}_{\mu_i}, \underbrace{\frac{\sigma_{\eta,i}^2}{\sigma_{f,i}^2}}_{\sigma_i^2} \right), \tag{10}$$

where $\mathcal{N}(\mu_i, \sigma_i^2)$ denotes the normal distribution with mean μ_i and variance σ_i^2. Thus, the conditional mean is shifted by $b_i \bar{v}$, while the remaining variation arises from $\bar{\eta}_i$. Thus, even under the null, the t-statistics need not be centered at zero nor have unit variance once we condition on the common component.

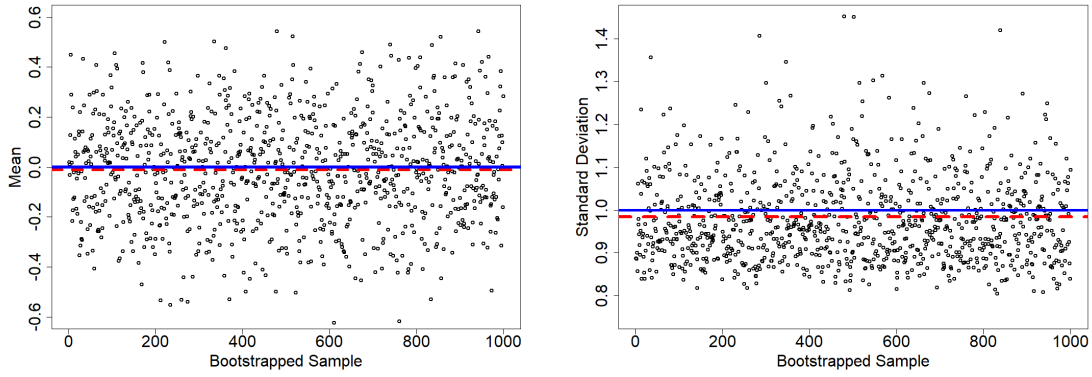
Conditional on $\bar{v}$, the cross-sectional dispersion of the t-statistics depends on the signs of $b_i \bar{v}$ for $i = 1, \ldots, N$. If all loadings $b_i \bar{v}$ share the same sign, all T_i shift in the same direction, producing a cross-sectional variance smaller than one. In contrast, if the signs differ across i, some statistics shift left and others shift right, increasing their cross-sectional variance above one. In the latter case, the overall mean may remain close to zero, but the cross-sectional spread is larger. The impact of such latent dependence on the empirical mean and variance of t-statistics has been emphasized in the multiple-testing literature (see, inter alia, Efron, 2004, 2007a, 2007b; Fan et al., 2012).

Under the standard assumption that $\sqrt{n}\, \alpha_i$ diverges under the alternative, equation (9) implies that non-null t-statistics shift into the tails, while null t-statistics remain near the center (though not necessarily standard normal). This separation underlies the methods used in Section 3.2.

To demonstrate that, under the null, the bootstrap does not recover a distribution similar to (10) under dependence, we demean the data so that all true alphas are zero and apply the bootstrap procedure used in Chen (2021). This procedure resamples, with replacement, the entire cross-section of returns in each time period to preserve the cross-sectional dependence between funds and benchmarks. In this setup, the 5% bootstrap threshold for a two-sided test should reject about 5% of the t-statistics in each replication if the method were valid. However, our results (Figure 1) show that this is not the case. In each bootstrap replication, the mean of the t-statistics across the 29,314 factors deviates from zero, sometimes above and sometimes below, due to correlation among the t-statistics, as predicted by (10). When we take absolute values, this shift inflates the entire distribution of t-statistics, making large absolute values of t-statistics much more common than expected under

the null. Importantly, the t-statistics that a researcher observes in practice are like those from a single bootstrap replication: they exhibit this dependence-induced shift, leading to larger absolute values even though the true alphas are zero. Similarly, the standard deviations vary above and below one, reflecting the mix of positive and negative correlations among the t-statistics. Yet, by averaging over many bootstrap replications, the bootstrap procedure produces a reference distribution that appears standard normal (mean zero, variance one), as expected. This masks the dependence-driven shifts that occur in any single realization. To see this, consider the simple dependence structure that results in (10). Within the same bootstrap replicate, the sample mean of the factors depends on a realization of $\bar{v}$ which is usually non-zero. However, across replicates its expectation is zero, so the bootstrap effectively integrates over $\bar{v}$ and recovers the unconditional distribution of T_i. As a result, using the bootstrap to approximate the null-distribution can lead to excessive false discoveries, because it understates the probability of large absolute t-statistics under the null. This holds only if the variance of the realized null t-statistics is not much smaller than one, which we find to be the case in our empirical analysis.

Figure 1: The Inadequacy of the Bootstrap Method. The Fama and French (2010) bootstrap is applied to demeaned, equal-weighted returns from the 29,314-factor dataset to replicate the results in Chen (2025). For each bootstrap replication (x-axis), the figure shows a dot representing the mean (left panel) and the standard deviation (right panel) of the 29,314 t-statistics in that replication. For reference, the mean (left panel) and standard deviation (right panel) of the standard normal distribution are indicated by a horizontal blue line. The average of the t-statistics (left panel) and the square root of the average variance (right panel) across the 1,000 bootstrap replications are indicated by a horizontal dashed red line.



To mitigate the concern of non-standard normality under the null, one can exploit the information coming from the large number of t-statistics to obtain more precise estimates of the null distribution. Throughout the paper, we shall refer to the "distribution under the null" as the empirical (conditional) distribution of the test statistics given the realized dependence structure, rather than the unconditional standard normal. As a solution, we recommend two approaches based on Efron (2007b): (a) the matching estimator and (b) the maximum likelihood estimator (MLE).[13] The matching estimator fits a smooth density estimate of the overall distribution of t-statistics, then approximates its logarithm near zero with a quadratic curve, and finally matches this to the log of a normal density to estimate the empirical null. The MLE estimator assumes the null t-statistics follow a normal distribution and maximizes a truncated likelihood over the central region of the data, where non-null cases are assumed to be absent, to estimate the empirical null parameters.[14]

These methods do not require a balanced panel, are simple to estimate, and the code is available. They reasonably assume that all non-null t-statistics lie in the tails of the distribution. Section 3.2 demonstrates these methods using empirical examples.

3.2 Assessing the impact of the standard normal assumption

To illustrate the impact of inappropriately imposing standard normality under the null and its implications for inference in finance, we first analyze and seek to understand the findings of Chen (2025), in particular those based on (7) which he calls the "easy bound". Following Chen's empirical setting, we estimate the proportion of true nulls, $\mathbb{P}(H_0) = 0.59$, using the standard method of Storey et al. (2004).[15]

The estimate of 0.59 is slightly higher than the 0.56 obtained by Chen (2025, eq. 21), who derived it using visual inspection. Based on our estimate, we would conclude that 12,019 of the 29,314 randomly generated factors (41%) are anomalies with a true non-zero alpha. To identify these anomalies, we follow Chen's procedure, testing each factor and applying a rule that rejects the null hypothesis of zero alpha if the absolute t-statistic exceeds 2.0. We then identify 9,537 significant factors out of 29,314 or 33%. To compute the FDR for this finding, we apply Chen's easy bound

[13] Efron (2007a) also suggests a method based on a latent variable to capture dependence, which can be estimated from the data.

[14] Details can be found in Section A.1.2 of the Appendix.

[15] The details of this method are provided in Section A.1.1 of the Appendix.

in (7):

$$\text{FDR} = \frac{0.05 \times \mathbb{P}(H_0)}{\mathbb{P}(|T| \geq 2)} \simeq \frac{0.05 \times 0.59}{0.33} \simeq 0.09, \tag{11}$$

which aligns with the result in Chen (2025). It is reassuring that, by performing the same analysis as in Chen (2025) and making only a slight adjustment to $\mathbb{P}(H_0)$, we obtain consistent results.

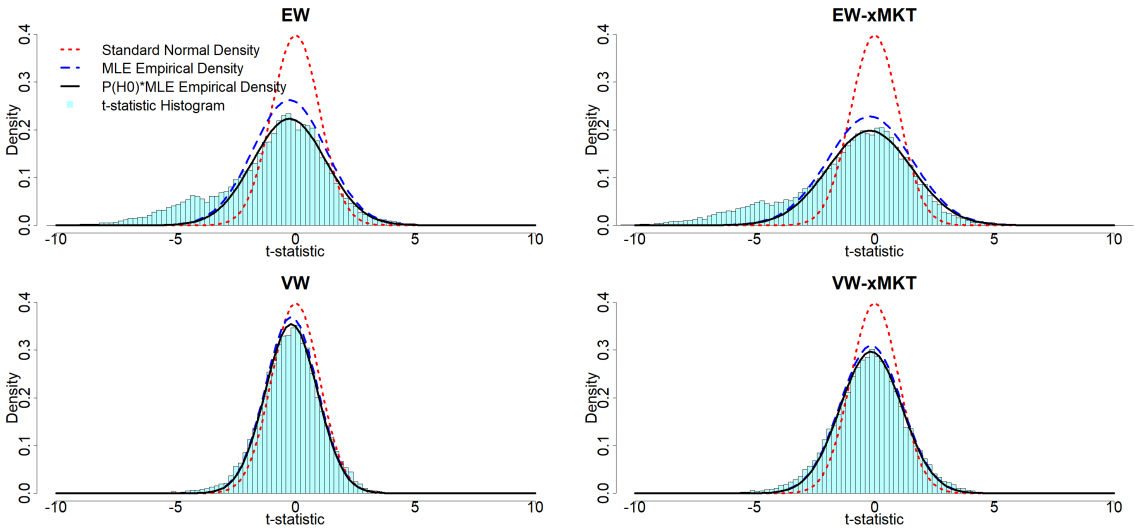
The results of this simple analysis are indeed remarkable, as 9,537 factors are identified with absolute t-statistics greater than two, with no more than 9% expected to be false discoveries.

The large number of discoveries can be explained by the effect discussed in Section 3.1. Due to the nature of the factor generation process, the t-statistics of these factors are correlated, causing the empirical distribution of null t-statistics to deviate significantly from a standard normal distribution. The null t-statistics are still well approximated by a normal distribution, but with mean and variance different from zero and one, as shown in (10). Consequently, the true FDR is expected to differ substantially from the bound presented in the above analysis, as the probability of an absolute t-statistic exceeding 2.00 could be significantly higher than 0.05. According to (10), there is a mean effect that always leads to more false rejections (whether positive or negative) because we are interested in absolute values of the t-statistics. The dispersion of the conditional means across different t-statistics produces a variance effect, which can lead to over-rejection when the means are highly dispersed or under-rejection when they are closely clustered together. It is the combined effect that matters. For this dataset, we will find that the variance is estimated to be greater than one, further increasing the number of false rejections.

For the same reason, the assumption that p-values follow a uniform distribution under the null no longer holds, and the estimate of the proportion of true alternative hypotheses will be wrong. This will, erroneously, lead us to conclude that $\mathbb{P}(H_1)$ is higher than it actually is. As we will show, accounting for the effect described in (10) leads to $\mathbb{P}(H_1)$ dropping from 0.41 to 0.15 and as low as 0.02, depending on the method used for estimation. We are interested in the discovery frequency $\mathbb{P}(H_1) = 1 - \mathbb{P}(H_0)$ and report this throughout. A nonzero null mean and a large null variance explain the inflated value $\mathbb{P}(H_1) = 1 - 0.59$ in (11).

We now turn to an analysis designed to obtain reliable inference, and the results demonstrate that common sense and statistical theory go hand in hand. We examine both the equal-weighted and value-weighted returns from the 29K dataset, although most factor applications rely on value-weighted returns. Furthermore, we also remove the effect of the market (which was not done in preliminary analysis in Equation (11) nor in Chen (2025)). This allows us to assess factors in a market beta neutral way,

Figure 2: Density Plot of t-statistics for Factor Returns. The histogram of the t-statistics for the 29,314 factors is plotted, with the top two subplots computed from equal-weighted returns and the bottom two subplots computed from value-weighted returns. The right two subplots present t-statistics from raw returns, while the left two subplots show t-statistics each orthogonalized with respect to the market return. The dashed blue line depicts the estimated null distribution of the t-statistics, the solid black line shows the approximation to the center of the density obtained by multiplying the estimated null density by $\mathbb{P}(H_0)$, where all the quantities are estimated by the MLE method. For reference, the standard normal density is shown as a dotted red line.



while also mitigating the problems of correlation and asymmetry in the t-statistics, thereby enhancing power. To be clear, we look at four approaches: equally weighted (EW), EW–xMKT, value weighted (VW), and VW–xMKT where "xMKT" denotes the dataset after removing the market factor, as in (1). We then compute the t-statistic for the alpha of each factor. The null hypothesis assumes a zero alpha, whereas the alternative is a two-sided test for a nonzero alpha.

Figure 2 displays the density plots for the four approaches. Each subplot presents a histogram of the t-statistics, superimposed with the standard normal density, the estimated null distribution, and the null component of the mixture density. The null component is modeled as the product of the normal density (with mean μ and variance σ^2) and the probability $\mathbb{P}(H_0)$, estimated using the MLE approach introduced as method (b) in Section 3.1 and detailed in Appendix A.1.2.

Figure 2 shows the resulting empirical densities of value-weighted (VW) returns are much more symmetric compared to the density of t-statistics obtained from equally-weighted (EW) returns. Despite this, the estimated null distribution from the MLE deviates from the standard normal for all four approaches. Taking VW–xMKT as an example, it shows a shift with an estimated mean of -0.15 and a standard deviation of 1.29.[16] The empirical distribution obtained by the matching method (referred to as method (a) in Section 3.1) is approximately the same as shown in Table 2, particularly for VW returns, so we suppress it in the figure.

All the superimposed approximate null distributions in the four panels of Figure 2 show a negative mean and a variance greater than one. This is particularly evident for the EW and EW–xMKT portfolios. These features arise from the underlying correlation structure of the t-statistics, which can produce clusters of heavily correlated tests. This intuition aligns with the findings in JKP, who document strong within-cluster correlations and weak dependence across clusters. Intra-cluster correlation induces joint shifts in the t-statistics, often all positive or all negative due to sign sharing, leading the mean of the distribution to deviate from zero (in this case, toward the negative). In contrast, weak or negative dependence across clusters increases overall dispersion, resulting in a variance greater than one. The 29K dataset contains both positively and negatively correlated factors, making such a structure plausible. Together, these effects explain the shifted mean and over-dispersion observed in the null distributions.

Rather than computing the FDR directly, our objective is to find a critical cutoff value τ that controls the FDR at a desired level, in particular ensuring that the FDR does not exceed 0.05.[17] Searching for such a cutoff is useful because researchers can use it as a guideline in future studies. The cutoff τ can be deduced using (4) by setting the FDR to 5%. We need three quantities from the first equality in (4): $\mathbb{P}(H_0) = 1 - \mathbb{P}(H_1)$, $\mathbb{P}(|T| \geq \tau \mid H_0)$, and $\mathbb{P}(|T| \geq \tau)$. We employ the matching and MLE estimators described in Section 3.1. Both methods require specifying the central region of the distribution from which the null is estimated, and we determine this region using adaptive procedures (see Appendix A.1.2 for details and Section S.3 of the Online Appendix for robustness checks). As a benchmark, we also consider the simple case where the null distribution is assumed to follow a standard normal, as in Chen (2025). The matching and MLE methods directly produce estimates for

[16] Details of the methodology are provided in Section A.1.2 of the Appendix, and a sensitivity analysis is presented in Section S.3 of the Online Appendix. The t-statistics are computed using the usual standard errors.

[17] If the FDR is 0.05 and not 0.09 as in Equation (11), the corresponding cutoff must necessarily be greater than 2.00, even when using the easy bound in (7).

$\mathbb{P}(H_1)$. Storey's estimator also produces an estimate for $\mathbb{P}(H_1)$ using the p-values derived from the null distribution.[18] The matching and MLE methods allow us to estimate a null distribution $\mathbb{P}(|T| \geq \tau \mid H_0)$, which is normal with mean and variance potentially different from zero and one to account for correlation. Finally, $\mathbb{P}(|T| \geq \tau)$ is obtained using the empirical distribution of the absolute t-statistics.

Table 2 presents the results. The upper panel reports the analysis based on the EW and EW–xMKT returns, respectively. The number of rejections under the standard normal null is broadly consistent with the preliminary analysis at the beginning of this section, with EW–xMKT yielding the highest rejections. We then use two empirical distribution estimation methods to find that $\mathbb{P}(H_1)$ reduces significantly, with MLE suggesting 0.14 and 0.12 for EW and EW–xMKT, respectively, which is much lower than $\mathbb{P}(H_1) = 0.41$ suggested at the beginning of the section.

Since the null t-statistics may have an estimated mean μ different from zero, we report left and right cutoffs. The number of rejections corresponds to T lying outside the interval, rounded to two decimal places, $\big(\text{Cutoff (left)}, \text{Cutoff (right)}\big)$. [19]

Assuming the standard normal distribution, the cutoffs for EW and EW–xMKT are ± 2.26 and ± 2.16, which are roughly consistent with Chen (2025) accounting for the fact that we use an FDR of 0.05 as opposed to his 0.09. The matching and MLE estimators provide a much different picture. Notice that the estimated standard deviation is at least 50% larger than that of a standard normal distribution. The MLE estimator also has a large negative shift in the mean. The Storey estimator of the fraction of true alternatives is substantially reduced. The cutoffs are dramatically higher. Focusing on the MLE, the t-statistic cutoffs are $(-4.48, 3.97)$ for the EW and $(-5.16, 4.75)$ for the EW–xMKT.

There are good reasons not to focus our analysis on EW returns. Aside from the fact that it is unusual to put the same dollar investment into every stock, the transactions costs for the monthly rebalanced EW portfolio are much larger. Ignoring these transaction costs inflates the average factor returns leading to t-statistics that are higher than could be practically obtained. As such, we prefer to focus on the VW returns. The lower panel of Table 2 reports the results of VW and VW–xMKT.

Consistent with the EW analysis, we find that the proportion of true alternative hypotheses is substantially reduced when using the matching and MLE estimators. The estimated proportions of true non-null hypotheses, $\mathbb{P}(H_1)$, are consistent across

[18] We also apply Storey's estimator using the null distributions estimated by the matching and MLE methods to check for consistency across methods.

[19] The formulas for FDR (4) and LFDR (6) hold as they are with a single τ if we replace $|T|$ by $|T - \mu|$. We could report a single cutoff for $|T - \mu|$, but found this less informative.

Table 2: Cutoff Points Using Different Procedures. Results are reported for equal-weighted raw returns (EW) and ex market returns (EW–xMKT) in the upper panel, and value-weighted raw returns (VW) and ex market returns (VW–xMKT) in the lower panel. The rows "$\mathbb{P}(H_1)$ (Storey)" and "$\mathbb{P}(H_1)$ (Matching/MLE)" present the estimated proportion of true alternative hypotheses using the Storey or the Matching/MLE methods, respectively. The rows "μ" and "σ" report estimates of the mean and standard deviation of the t-statistics under the null for the matching and MLE estimators. For the other methods, they are set to 0 and 1. The "N rejected" row displays the number of factors rejected out of 29,314 total using the cutoffs specified in the "Cutoff (left)" and "Cutoff (right)" rows.

		EW			EW–xMKT	
	$\mathcal{N}(0,1)$	Matching	MLE	$\mathcal{N}(0,1)$	Matching	MLE
$\mathbb{P}(H_1)$ (Storey)	0.41	0.10	0.14	0.47	0.07	0.12
$\mathbb{P}(H_1)$ (Matching/MLE)	NA	0.09	0.15	NA	0.05	0.13
μ	0.00	-0.08	-0.26	0.00	0.03	-0.20
σ	1.00	1.62	1.52	1.00	1.89	1.75
Cutoff (left)	-2.26	-4.70	-4.48	-2.16	-5.51	-5.16
Cutoff (right)	2.26	4.55	3.97	2.16	5.57	4.75
N rejected	8264	2308	2754	9599	1919	2393
		VW			VW–xMKT	
$\mathbb{P}(H_1)$ (Storey)	0.12	0.04	0.05	0.24	0.04	0.05
$\mathbb{P}(H_1)$ (Matching/MLE)	NA	0.05	0.06	NA	0.05	0.06
μ	0.00	-0.17	-0.18	0.00	-0.15	-0.15
σ	1.00	1.09	1.08	1.00	1.31	1.29
Cutoff (left)	-3.32	-4.08	-4.00	-2.89	-5.06	-4.97
Cutoff (right)	3.32	3.73	3.64	2.89	4.76	4.66
N rejected	471	205	219	1697	96	111

Storey's estimator and those derived from the matching and MLE methods.[20]

The most realistic results are for VW–xMKT. Both the matching and MLE methods suggest a null distribution with a mean of -0.15 and a standard deviation of approximately 1.3. We obtain $\mathbb{P}(H_1) \simeq 0.05$, whether the value is obtained directly using the matching and MLE estimators or using Storey's method with p-

[20]The matching and MLE methods are designed to produce similar results. Broadly, the matching estimator is more non-parametric, while the MLE is more parametric, and each carries the strengths and weaknesses of its respective approach. The MLE can be sensitive to the interval chosen to estimate the null distribution. Ideally, both methods should agree and be used in conjunction; disagreement indicates that the distribution of the t-statistics is particularly difficult to approximate.

values based on the estimated null. This differs substantially from the value of 0.24 obtained using Storey's method with p-values based on the standard normal null distribution, thus ignoring the correlation. Using the 0.05 estimate for $\mathbb{P}(H_1)$, we have that 1,466 (=0.05*29,314) of the 29,314 hypotheses represent true alternatives. This is still sizable, yet far smaller than Chen's conclusion that most discoveries are true.

As reported in Table 2, focusing on the VW and VW–xMKT returns, the cutoffs are ± 3.32 and ± 2.89, respectively, assuming a standard normal null. This results in 471 and 1,697 hypotheses being rejected, which is much lower than the equal-weighted analysis. Our approaches again imply higher cutoffs, with the matching and MLE methods yielding $(-5.06, 4.76)$ and $(-4.97, 4.66)$ for VW–xMKT, respectively, resulting in only 96 to 111 hypotheses being classified as significant.

These thresholds may seem very high. Even if we change the FDR to 10%, the cutoffs for VW-xMKT are $(-4.66, 4.36)$ (matching) and $(-4.55, 4.24)$ (MLE). The high cutoffs reflect the nature of the problem. The datasets contain 29,314 data-mined factors, and the cutoff would vary with a different population, especially if the published factors were not purely data-mined. This will be discussed later.

3.3 Revisiting the analysis of JKP

This subsection revisits the recent work of JKP (2023), who analyze 153 published factors, including 138 from the HXZ Q-library and 15 additional ones. Using a Bayesian framework that conditions on the data and accounts for cross-sectional correlations, they control the FDR in testing CAPM alphas. Their analysis covers both U.S. and global datasets and concludes there is no replication crisis.

The factors in JKP are signed as in the original published studies before running the market model regression in (1). The resulting alpha is not guaranteed to be positive, but it is in practice for most factors (136 out of 153). JKP select a subset of 119 factors from the 153 original ones, considering only those that were significant in the original studies. Their goal is to show that these 119 factors are replicable. To this end, they apply their Bayesian methodology to all 153 factors and then take the intersection with the pre-selected 119 factors. The replication rate is computed as the number of discovered factors in this intersection divided by 119. They employ a hierarchical Bayesian methodology that they claim controls the FDR at the 5% level. They conclude that the replication rate in the U.S. sample is 82.4%. However, their analysis contains two serious issues. First, they use a sample of 153 factors that suffers from sample selection bias: it includes published factors, the vast majority of which are significant. Due to this bias, the null t-statistics cannot be standard

normal even if the factors are independent. Second, their Bayesian methodology controls the FDR only under assumptions that are unlikely to hold in practice, as we show next.

JKP's method conflates the posterior probability of a hypothesis with the posterior probability that the parameter α_i lies in a certain region (see their equation (27)). JKP consider one-sided hypotheses of the form $H_0^i : \alpha_i \leq 0$ vs. $H_1^i : \alpha_i > 0$, which makes the conceptual distinction more subtle and contributes to the confusion.[21] However, on closer inspection, the distinction becomes clear. The posterior probability that $\alpha_i \leq 0$ given the data can be written as

$$\begin{aligned}\mathbb{P}\left(\alpha_i \leq 0 \middle| \text{ Data}\right) = \mathbb{P}\left(\alpha_i \leq 0 \middle| \text{ Data}, H_0^i\right) \mathbb{P}\left(H_0^i \middle| \text{ Data}\right) \\ + \mathbb{P}\left(\alpha_i \leq 0 \middle| \text{ Data}, H_1^i\right) \mathbb{P}\left(H_1^i \middle| \text{ Data}\right).\end{aligned}$$

The assumption in JKP (equation (27)) is that the left-hand side of the above display equals the posterior probability of the null hypothesis. By definition, this posterior probability is $\mathbb{P}(H_0^i \mid \text{Data})$, which appears as one of the terms on the right-hand side. It is then clear that these two quantities, $\mathbb{P}(\alpha_i \leq 0 \mid \text{Data})$ and $\mathbb{P}(H_0^i \mid \text{Data})$, are generally not the same. They coincide only under highly restrictive prior assumptions: the prior for α_i under H_0^i places all its mass on $\alpha_i \leq 0$, and the prior under H_1^i places all its mass on $\alpha_i > 0$. If, in addition, the overall prior for α_i is symmetric around zero —as in JKP— then it follows that $\mathbb{P}(H_0^i) = 1/2$ (see Section A.2 of the Appendix for details). In practice, JKP's approach relies on treating $\mathbb{P}(\alpha_i \leq 0 \big| \text{ Data})$ as if it were the posterior probability of the null. This leads to a decision rule that, in large samples, is effectively equivalent to using standard unadjusted OLS p-values.[22] We demonstrate this equivalence both theoretically (in Section A.2 of the Appendix) and through simulations (in Section S.4 of the Online Supplement). Hence, it is no wonder that they obtain high replication rate.

A key challenge in the analysis of JKP is sample selection bias. The set of published factors represents only a small, non-random subset of the unknown set of all possible signals that researcher might have tested. Because researchers tend to focus on promising or statistically significant results, the distribution becomes biased toward large absolute t-statistics. JKP argue that sample selection is not a problem once their Bayesian multilevel approach is applied. Their argument

[21] JKP signs the factors, rather than the alpha in (1), which also causes some difficulties in the discussion.

[22] In the framework of JKP, the posterior distribution of α_i is normal and, as the sample size increases, its mean and variance converge to the sample mean and variance. Consequently, $\mathbb{P}(\alpha_i \leq 0 \mid \text{Data})$ converges to the usual frequentist p-value. This can be seen by taking large-sample limits of equations (17) and (19) in JKP.

is that by calibrating the variance of the within-cluster alpha prior using data simulated to mimic a desired population size and selection bias, one can account for publication bias within the Bayesian procedure (JKP, Section C2 and Section IV in their Appendix). However, it is easy to see that this argument is problematic. As the sample size increases, the influence of the prior vanishes, and the posterior alpha converges to the usual MLE estimator (JKP, Proposition 3, taking the sample size in the time dimension to infinity). This is equivalent to claiming that, as the sample size increases while the number of hypotheses N remains fixed, the effect of sample selection bias disappears, which is clearly false. The bias arises from the cross-sectional selection of the most significant factors among many candidates, not from sampling error in the time series; therefore, it persists when the sample size increases.

Due to sample selection bias, determining an appropriate FDR-controlling cutoff for the JKP dataset is challenging. The approach in Section 3.2 cannot be applied directly, as the null distribution parameters cannot be reliably estimated in a selection-biased sample. One way to address this is to treat Chen's 29K dataset as a proxy for the entire factor population, that is, by adding 153 JKP factors (U.S., July 1951–December 2022) to the 29K set and re-applying the analysis from Section 3.2.[23] Using the MLE cutoff of (-4.97,4.66) from Table 2, we find 159 VW–xMKT factors exceeding the threshold under 5% FDR control, of which 48 are from the 153-factor JKP dataset. However, this is only an approximation. One may argue that adding 153 factors to the 29K dataset is overly conservative for the number of observable factors, as existing literature considers far fewer unobservable tests — for example, HLZ imply that researchers have tested around 2,500 factors.

Recall that the formula for the FDR in (4) requires estimation of three quantities: the probability of the null hypothesis and the null distribution of $|T|$ in the numerator, and $\mathbb{P}(|T| \geq \tau)$ in the denominator. The latter is estimated by the number of rejections divided by the number of hypotheses tested. Hence, we can still test the 153 factors while controlling for the FDR when the true number of tests is greater than 153, as long as we fix (a) $\mathbb{P}(H_1)$, (b) the mean and variance parameters for the approximation of the null distribution, and (c) the unknown number of hypotheses tested. In fact, to estimate $\mathbb{P}(|T| \geq \tau)$ when the true number of tests exceeds the 153 observed factors, we compute the number of rejections from the 153 observed t-statistics and divide by the postulated true number of tests $N > 153$. We therefore conduct an analysis to examine how different parameter choices affect the cutoff calculation for controlling the FDR. Specifically, we analyze the effect of varying

[23]There are four overlapping factors between the two datasets, resulting in a merged dataset of 29,463 unique factors.

three parameters: the proportion of true alternatives, $\mathbb{P}(H_1)$; the standard deviation of the t-statistics under the null, σ (for simplicity, the mean is fixed to zero); and the unknown number of tests conducted. We shall refer to the latter as the true number of tests. The simple analysis at the beginning of Section 3.2 suggests that $\mathbb{P}(H_1) = 0.41$, whereas a more refined analysis using VW, shown in Table 2, indicates that $\mathbb{P}(H_1)$ may be closer to 0.05. Here, we consider a range of scenarios where $\mathbb{P}(H_1) = 0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$. The standard deviation is set to $\sigma = 1, 1.25, 1.5$, and the size of the population is scaled to be 1, 10, and 100 times that of the JKP dataset, i.e., 153, 1,530, and 15,300.

Figure 3: Number of Rejected JKP Factors Under Varying Population Size, Standard Deviation of the Null, and Proportion of True Factors. From left to right, the plots show the number of factors rejected when the true number of tests is 1, 10, and 100 times 153 (the number of factors in JKP), with the corresponding derived cutoff shown as a dashed reference line. The x-axis labels indicate the proportion of true alternatives, $\mathbb{P}(H_1) = 0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$. Three lines in each subfigure present the results for different standard deviations of the null, with $\sigma = 1$, 1.25, and 1.5. For comparison, we also report the number of discoveries (122) obtained using the hierarchical Bayesian method of JKP (solid black line).

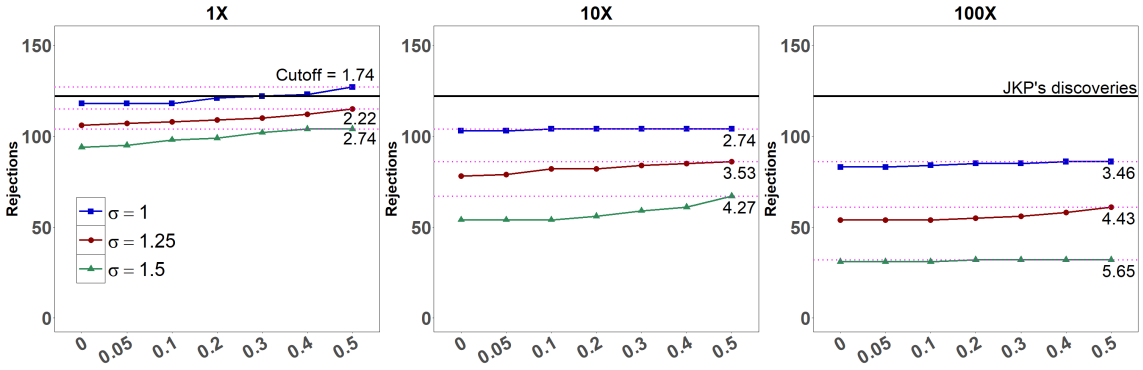


Figure 3 displays the number of significant factors and the corresponding cutoffs when controlling the FDR at the 5% level. In particular, in a large population with minimal selection biases, the number of discoveries sharply decreases. This trend worsens further when the null distribution of t-statistics deviates from the standard normal. For example, when the population size is 100 times larger and σ under the null equals 1.25 (1.5), only 61 (32) factors are rejected using the cutoff of 4.43

(5.56).[24] This result does not vary substantially with the choice of $\mathbb{P}(H_1)$, particularly when the unknown population size is large. The effect of $\mathbb{P}(H_0)$ or $\mathbb{P}(H_1)$ on cutoff is discussed in Section A.3 of the Appendix. Building on the empirical analysis in Section 3.2, we consider a nearby case using a cutoff of 4.43. These cutoffs are much larger than the OLS ones — using a cutoff of 1.96, we obtain 122 discoveries.[25] For comparison, the method of JKP also produces 122 factor discoveries, the same as OLS. JKP's assumptions reduce their method to OLS in large samples. As such, their results suggest that, in many situations, no adjustment to the thresholds implied by OLS is necessary.

We make two points. First, their Bayesian model equates two key quantities: the posterior probability of a hypothesis with the posterior probability that the alpha lies in a certain region (JKP, equation (27)). Second, this equality and the assumption of a normal distribution—a symmetric distribution—for the prior of α (JKP, page 9 of the Appendix) imply a prior probability of 0.50 for the null hypotheses (see Section A.2 of the Appendix for further details).[26]

As a result, their model is not generalizable. In contrast, our analysis of VW–xMKT in Table 2 produces a probability close to 0.95. However, it is important to consider that JKP's factors are published ones and not randomly generated, as are the factors in Chen (2025). Hence, it is reasonable to set the probability of the null to be lower. Nevertheless, the impact is somehow negligible, especially if we assume that the unknown number of original hypotheses tested by researchers is larger than the 153 published factors – as illustrated in Figure 3 (Section A.3 in the Appendix provides a theoretical argument for this). Furthermore, correlation among hypotheses may require assuming a standard deviation greater than one to produce a more realistic approximation to the null distribution of the t-statistics. This, coupled with sample selection bias, requires a threshold higher than the one implied by JKP.

In Figure 3, we carry out calculations using assumed parameter values for a hypothetical null distribution of the dataset. We can only hypothesize because—due to sample selection bias—the methodology in Section 3.2 is not valid. The sample selection bias is obvious from Figure 4. To approximate the dispersion of the

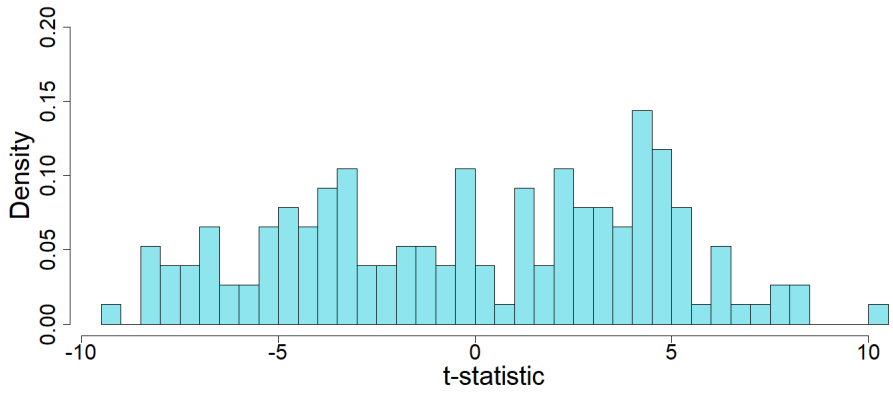
[24] Given that the mean is fixed at zero under the null, the cutoffs are symmetric, and we report only a single value for the two-sided tests.

[25] To be consistent with the rest of the paper, we use a two-sided test on the alphas, whereas JKP employ a one-sided test after signing the factors and then running the regression in (1). This difference has minor consequences but explains why they reject 119 factors instead of 122.

[26] The simulations in JKP are also based on this assumption. Modifying the design to produce a larger proportion of nulls shows that the procedure fails to control the FDR, as we show in Section S.4 of the Online Appendix.

possible null t-statistics, we compute the standard deviation for those in the interval $[-1.25, 1.25]$. In particular, 20 of the 153 t-statistics fall in this interval, and we assume that all of these are null. Given the small number of observations found in the central body of the distribution, a normal fit in this region is not precise. The standard deviation of the 20 t-statistics is derived assuming a normal density with mean zero truncated to the interval $[-1.25, 1.25]$. The sample standard deviation of the 20 t-statistics is 0.69. Using the relationship between the variance of a mean-zero truncated normal distribution and that of the corresponding normal distribution, we estimate a standard deviation of 1.52, which is the upper limit of the possible values of σ used in Figure 3. This is greater than the value estimated in Section 3.2, but consistent with it.

Figure 4: Histogram of t-statistics for the 153 JKP Factors.



In summary, our critique has two angles. First, JKP claim that their Bayesian method controls the false discovery rate (FDR) (it may do so only under unrealistic assumptions) and that it also solves the selection bias problem. We disagree and have provided arguments inconsistent with their claim. Second, when we challenge their Bayesian method by assuming (a) a larger total number of tests and (b) a null distribution that may differ from the standard normal (based on our experience with the 29K dataset), we find that the correct threshold would have to be higher — leading to fewer discoveries than both their method and OLS produce.

4 What Researchers Should Do?

For the Chen (2025) dataset, which consists of purely data-mined factors, our results in Section 3.2 suggest a cutoff between 3.7 and 5, depending on the method used (see Table 2). However, researchers are primarily interested in determining the appropriate cutoff for identifying a newly discovered factor as significant while controlling for false discoveries. Specifically, since the left-hand side of (4) equals 5%, what is the optimal threshold τ that researchers should use?

In this framework, we cannot leverage information from multiple t-statistics, as we assume that only one is reported. We therefore use the standard normal density under the null, which is asymptotically correct unconditionally (as illustrated in equation (10)) and represents the best feasible choice in this setting. However, we still require a procedure that accounts for the possibility of multiple tests conducted by the researcher. To assess whether a new factor is significant, we therefore suggest using the LFDR rather than the FDR, for reasons that will become apparent in due course.

We then apply Bayesian arguments to derive a lower bound for the LFDR, enabling us to determine the minimum cutoff or threshold necessary to control for multiple tests, even without knowing the number of tests. Moreover, despite the simplicity of the formula, we tabulate the values to eliminate the need for calculations, helping researchers better understand the implications of different cutoffs and assumptions. Based on the results of this section, we are able to answer the main question of this paper: the cutoff for a t-statistic for factor research lies between 3.0 and 3.5.

The argument we use to derive this result, which is agnostic to the number of tests conducted by the researcher, is not available for the FDR. By the end of this section, the reader will see that the LFDR is the appropriate quantity to control when assessing a new factor proposed by researchers, and will understand how the level at which we control the LFDR relates to the odds ratio in favor of the alternative, conditional on $|T|$.

For clarity of exposition, recall the definitions of FDR and LFDR from Section 2.2, making explicit the dependence on τ:

$$\text{FDR}\,(\tau) = \frac{\mathbb{P}(H_0)\mathbb{P}(|T| \geq \tau \big| H_0)}{\mathbb{P}(|T| \geq \tau)} = \mathbb{P}(H_0 \big| |T| \geq \tau), \tag{12}$$

and define the local FDR (denoted by LFDR) as

$$\text{LFDR}\,(\tau) = \frac{\mathbb{P}(H_0)\text{pdf}(\tau \big| H_0)}{\text{pdf}(\tau)} = \mathbb{P}(H_0 \big| |T| = \tau), \tag{13}$$

where $\text{pdf}(\tau \big| H_0)$ is the null probability density function (pdf) of $|T|$ evaluated at τ, and $\text{pdf}(\tau)$ is the marginal pdf of $|T|$. We write the pdf of $|T|$ as

$$\text{pdf}(\tau) = \text{pdf}(\tau \big| H_0)\mathbb{P}(H_0) + \text{pdf}(\tau \big| H_1)\mathbb{P}(H_1), \tag{14}$$

with obvious notation; this can be obtained from (5) by differentiating with respect to τ.

From the right-hand sides of (12) and (13), we have the following.

$$\text{FDR}(\tau) = \frac{\mathbb{E}\left[\mathbb{P}(H_0 \big| |T|) \times 1_{\{|T|\geq\tau\}}\right]}{\mathbb{P}(|T| \geq \tau)} = \frac{\mathbb{E}\left[\text{LFDR}(|T|) \times 1_{\{|T|\geq\tau\}}\right]}{\mathbb{P}(|T| \geq \tau)}, \tag{15}$$

where $\text{LFDR}(|T|) = \mathbb{P}(H_0 \big| |T|)$ is the probability of the null H_0 given $|T|$, and $1_{\{|T|\geq\tau\}}$ is one when $|T| \geq \tau$ and zero otherwise. There are two key takeaways from this display. First, it shows how the LFDR contributes to the overall FDR: the FDR is a population average of the LFDR evaluated at values greater than τ. Thus, the LFDR has the advantage of illustrating how an individual hypothesis contributes to the overall FDR.

A simple illustrative example can aid intuition. Suppose we have N i.i.d. realizations $\{\tau_1, \tau_2, \ldots, \tau_N\}$ of $|T|$. Consider the average LFDR among those realizations satisfying $\tau_i \geq 3$:

$$\frac{1}{N_\tau} \sum_{i=1}^{N} \text{LFDR}(\tau_i)\, 1_{\{\tau_i \geq 3\}}, \tag{16}$$

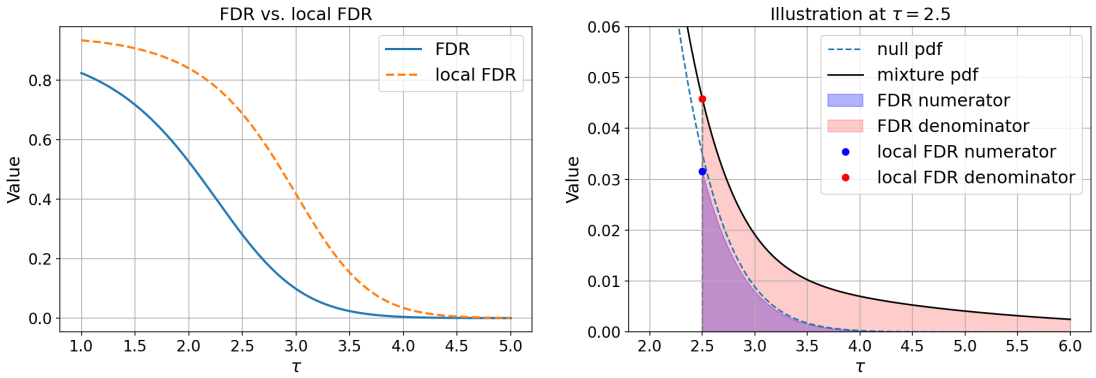
where $N_\tau = \sum_{i=1}^{N} 1_{\{\tau_i \geq 3\}}$ is the number of realizations with $\tau_i \geq 3$. As $N \to \infty$, the average converges in probability to its expectation, which, using (15), is FDR(3).

The second takeaway is that $\text{FDR}(\tau) \leq \text{LFDR}(\tau)$ in the usual situation where $\text{FDR}(\tau)$ decreases with τ.[27] This means that the LFDR tends to be higher than the overall FDR at the same point. While FDR involves setting a threshold and looking at all test statistics beyond it, the LFDR evaluates the likelihood that a single test statistic is a false discovery.

Figure 5 illustrates the previous points. The left panel shows that the LFDR is always greater than the FDR. The right panel plots the components used to compute the FDR and LFDR for $\tau = 2.5$. The mixture pdf follows (5). The LFDR equals the ratio of the two pdfs evaluated at $\tau = 2.5$, multiplied by the null probability, whereas the FDR equals the ratio of their areas for $\tau \geq 2.5$, also multiplied by the null probability.

[27] Since it depends on pdfs, the LFDR may not necessarily be monotonically decreasing in τ.

Figure 5: FDR and Local FDR. The right panel shows the FDR and LFDR as functions of τ, assuming $|T|$ follows an exponential distribution with mean 2 under H_1 and that $\mathbb{P}(H_0) = 0.9$. The left panel shows the null and mixture densities of $|T|$. At $\tau = 2.5$, the LFDR is computed as $0.9 \times \text{pdf}(2.5 \big| H_0)/\text{pdf}(2.5) \simeq 0.9 \times 0.03506/0.04588 \simeq 0.69$. The FDR at the same threshold is the ratio of tail probabilities: $0.9 \times \mathbb{P}(|T| \geq 2.5 \big| H_0)/\mathbb{P}(|T| \geq 2.5) \simeq 0.9 \times 0.01242/0.03983 \simeq 0.28$. The blue and red dots highlight the numerator and denominator in the LFDR calculation, while the shaded areas correspond to the FDR components. The figure illustrates that LFDR evaluates individual observations, whereas FDR averages over all more extreme values. A similar calculation for $\tau = 3.4$ appears in the main text.



When evaluating individual t-statistics, a natural question is: what is the probability that the null is true given the observed value τ? In this setting, we argue that the LFDR is more appropriate than the FDR, as it conditions on the observed value of $|T|$. LFDR-based procedures typically reject fewer hypotheses. This is because they rely on a more informative measure: the probability that a specific finding is a false discovery. To ensure reliable, case-by-case assessments of significance for each hypothesis being tested, it may be reasonable to control the LFDR at higher levels (e.g., 10-20%) than the FDR (5-10%). The two measures target different errors so that they are not directly comparable: the FDR controls the expected proportion of false discoveries, while the LFDR relates to the posterior odd ratios of the alternative, and tells us the probability that the null is true, as we show next.

To interpret the meaning of a given target level for the LFDR (e.g., 20%), we can use its relation to the posterior odds ratio:

$$\frac{\mathbb{P}(H_1 \big| \, |T| = \tau)}{\mathbb{P}(H_0 \big| \, |T| = \tau)} = \frac{1 - \text{LFDR}\,(\tau)}{\text{LFDR}\,(\tau)}. \tag{17}$$

This equation follows directly from (13). Then, an LFDR of 0.2 corresponds to posterior odds of 4-to-1 in favor of H_1, conditional on the observed value of $|T|$. Furthermore, p-values do not inherently account for multiple testing, which can lead to inflated false positive rates in large-scale studies. Furthermore, unlike p-values, the LFDR inherently accounts for multiple testing, thereby avoiding inflated false positive rates in large-scale studies.

We work out a simple example to compare the FDR and the LFDR. We need to specify the distributions under the null and the alternative. We assume that the null pdf of T is standard normal, so that the null pdf of $|T|$ evaluated at $\tau > 0$ is $2\varphi(\tau)$, denoting by φ the standard normal pdf.[28] For definiteness, let the non-null pdf of $|T|$ be exponential with mean 2, i.e., $\text{pdf}(\tau \big| \, H_1) = 2^{-1} e^{-\tau/2}$.[29] In practice, $\text{pdf}(\tau \big| \, H_1)$ is unknown, and we shall suggest a solution to this, but it is instructive to go over an analytical example first to put the problem into the right perspective. Based on the results in Section 3.2, we assume that $\mathbb{P}\,(H_0) = 0.9$. To obtain a numerical value, suppose that we have a factor with a t-statistic of 3.4. Then,

$$\text{LFDR}\,(3.4) = \frac{2\varphi(3.4) \times 0.9}{2\varphi(3.4) \times 0.9 + 2^{-1} e^{-3.4/2} \times (1 - 0.9)} \simeq 0.2$$

[28] This follows by symmetry around zero of the standard normal density.

[29] A similar choice is used in Chen (2025).

The FDR is related to LFDR (3.4) by the formula in (15), but it is easier to compute it directly from (12). Denoting the standard normal distribution function at τ by $\Phi(\tau)$, we have

$$\text{FDR}\,(3.4) = \frac{2\,(1-\Phi(3.4))\times 0.9}{2\,(\Phi(3.4)-1)\times 0.9 + e^{-3.4/2}\times(1-0.9)} \simeq 0.03$$

where we have used the formula for $\mathbb{P}(|T|\geq\tau)$ in (5), with $\mathbb{P}(|T|\geq\tau\big|\,H_0) = 2\,(1-\Phi(\tau))$ and $\mathbb{P}(|T|\geq\tau\big|\,H_1) = e^{-\tau/2}$. This example illustrates that the LFDR and the FDR differ substantially even when evaluated at the same t-statistic. For instance, an LFDR of 0.2 corresponds to an FDR of only 0.03 in this case. The example also illustrates that computing the LFDR requires specifying the density of $|T|$ at τ under both the null and alternative hypotheses. While it is common to assume a normal distribution under the null when only a single t-statistic is observed, the distribution under the alternative is problem-specific and often difficult to specify. Similarly, unless all t-statistics from the conducted tests are available—in which case the empirical distribution function can be used, as in Section 3.2—the FDR requires specifying the tail distribution under the alternative. Fortunately, as we show next, this difficulty can be addressed within the context of LFDR estimation.

To avoid specifying a subjective density we consider a general upper bound for the density under the alternative, which in turn produces a lower bound for the LFDR (Theorem 1). This is because increasing the denominator in (13) reduces the value of the LFDR. Working with a lower bound on the LFDR is appropriate when the distribution of $|T|$ under the alternative is unknown. In such cases, it is desirable to consider the alternative that provides the strongest possible evidence against the null, thereby yielding the smallest value for the LFDR. This approach results in the lowest threshold that still satisfies a given LFDR level. While this may increase the Type I error rate, it reduces the Type II error, which is especially relevant in discovery-oriented settings where researchers often lack complete information about all tests that may have been run. In the absence of precise information about selective reporting or p-hacking, it is justifiable to err on the side of facilitating discovery, subject to controlling the posterior probability of a false positive.

The lower bound holds under a technical condition on the non-null density, which is satisfied by most distributions commonly encountered in finance. Intuitively is holds when the non-null density of $|T|$ is decreasing and exhibits heavier tails than the normal density; examples include Pareto and exponential distributions, among many others (see Section A.4 in the Appendix for the details).[30] The detailed argument,

[30]Under the alternative it is natural to assume that $|T|\to\infty$ with the sample size, which in turn implies that the tails must be thicker than a normal

whose proof is given in the Appendix, implies the following.

Theorem 1 *Suppose that the density of $|T|$ under the alternative satisfies the technical condition given in Section A.4 of the Appendix. Then, we have that:*

$$\text{LFDR}(\tau) \geq \frac{\text{bf}(\tau)\,\mathbb{P}(H_0)}{\text{bf}(\tau)\,\mathbb{P}(H_0) + [1 - \mathbb{P}(H_0)]}$$

where $\text{bf}(\tau) = -e \times p(\tau) \ln(p(\tau))$ *for* $p(\tau) = 2[1 - \Phi(\tau)]$ *and any cutoff* τ *such that* $0 < p(\tau) < \frac{1}{e}$*, where* $e \simeq 2.72$*.*

In Theorem 1, $p(\tau)$ is the p-value corresponding to $|T| = \tau$ under the standard normal null. The term $\text{bf}(\tau)$ is a lower bound on the ratio of the pdf under the null to the pdf under the alternative, evaluated at τ. That ratio is also known as the Bayes factor, which is a way to quantify the strength of evidence provided by the statistic for one model over the other. In this case, a Bayes factor greater than one indicates evidence in favor of the null model in the numerator (the null), and less than one favors the model in the denominator (the alternative).

Theorem 1 derives a lower bound for the LFDR (which is not possible for the FDR), providing an objective benchmark. The resulting cutoff is the smallest we can expect across all possible non-null distributions with monotonically decreasing densities that have heavier tails than the normal density (see Section A.4 in the Appendix for details).

Theorem 1 provides a nearly universal lower bound on the LFDR, given a two-sided p-value derived under a standard normal null and a prior on the probability under the null $\mathbb{P}(H_0)$. For the sake of clarity, we consider a numerical example. Choose a threshold $\tau = 3.16$, and suppose that $\mathbb{P}(H_0) = 0.9$. This means that

$$\text{bf}(3.16) = -e \times 2[1 - \Phi(3.16)] \times \ln(2[1 - \Phi(3.16)]) \simeq 0.028,$$

so that

$$\text{LFDR}(3.16) \geq \frac{0.028 \times 0.9}{0.028 \times 0.9 + 0.1} \simeq 0.2.$$

Hence, a threshold of $\tau = 3.16$ yields a lower bound on the LFDR equal to 0.2. This implies that if we use a threshold smaller than 3.16, the LFDR will exceed 0.2 even in the best-case scenario. Put differently, if our goal is to achieve an LFDR of 0.2, then we require $|T|$ to be at least 3.16 if not much larger. Hence, our result imposes a minimum requirement on the threshold value. The strength of this result is that it does not depend on the number of tests conducted or require specification of the distribution under the alternative.

Using the formula, we can solve for the threshold τ to obtain critical values corresponding to various combinations of $\mathbb{P}(H_0)$ and target LFDR levels. The resulting thresholds are reported in Table 3. As previously remarked, an LFDR of 0.2 corresponds to a posterior odds ratio of 4-to-1 in favor of H_1 – a reasonable level of evidence in support of the alternative. An LFDR of 0.05 corresponds to a posterior odds ratio of 19-to-1 in favor of the alternative, which should be considered overwhelming evidence. However, these thresholds only control a lower bound on the LFDR. Therefore, assuming $\mathbb{P}(H_0) = 0.9$ – based on the results from Section 3.2– it is reasonable to consider thresholds closer to 3.65 rather than 3.16.

Table 3: Critical values of $|T|$ for Controlling a Lower Bound on the Local FDR. The critical values are reported at levels 0.2, 0.1, and 0.05, for various values of $\mathbb{P}(H_0)$.

LFDR	$\mathbb{P}(\boldsymbol{H_0}) = \mathbf{0.5}$	**0.6**	**0.7**	**0.8**	**0.9**	**0.95**
0.20	2.24	2.44	2.64	2.86	3.16	3.39
0.10	2.63	2.80	2.97	3.16	3.41	3.64
0.05	2.93	3.08	3.23	3.39	3.64	3.92

Section A.5 of the Appendix provides user-friendly procedures for researchers to test factors under both multiple and single testing frameworks. In addition, we provide code and usage examples for all methods in the GitHub repository `https://github.com/yzhao7322/Threshold-in-Factor-Models.git`.[31]

Setting $\mathbb{P}(H_0) = 0.95$ by controlling the LFDR at 10%, we identify 16.3% (4,786 factors) of the 29,314 factors that exceed the threshold of 3.64 in the EW–xMKT dataset, and 2.2% (636 factors) in the VW–xMKT dataset. These proportions are substantially lower than the standard-normal–null results reported in Table 2 in Section 3.2, namely 32.7% (9,599 factors) for EW–xMKT and 5.8% (1,697 factors) for VW–xMKT, while yielding more rejections than the matching and MLE results. Applying the same cutoff to the JKP dataset results in 83 factor rejections, which is 54.2% of the 153 JKP factors. Appendix Section A.6 presents simulation studies comparing the standard normal null with the lower-bound method and other FDR and LFDR procedures. Our simulations detail both the Type I and Type II errors across the different methods. The results highlight the unreliability of the $\mathcal{N}(0,1)$ null assumption. In contrast, the lower-bound approach yields superior overall performance by providing a balanced trade-off between Type I and Type II errors.

[31] The `user_guide.R` file in the GitHub repository provides step-by-step guidance with examples for testing factors under both the multiple-testing framework and the single-factor discovery framework.

5 Limitations and Outstanding Issues

5.1 Assumptions

The factor discovery literature faces a number of challenges, which can only be addressed by making certain assumptions. It is important to keep these assumptions in mind. In our empirical analysis, the primary assumption is that all t-statistics near the center of the distribution correspond to null hypotheses. This allows us to use these t-statistics to estimate the null distribution, provided the number of hypotheses is large. This assumption holds when the absolute values of the non-null t-statistics diverge to infinity, which is necessary for the power of the tests to approach one. However, the assumption fails if a factor's alpha is small relative to the noise level. Specifically, based on the model in (1), the assumption fails if $\sqrt{n}\alpha_i/\sqrt{\sigma_{\varepsilon,i}}$ is not large, where n is the sample size and $\sigma_{\varepsilon,i}$ is the standard deviation of the error term $\varepsilon_{t,i}$. This may occur when the alpha or sample size is small, or the factor is very noisy.

Our empirical analysis is also complemented with guidance on selecting an appropriate threshold when a researcher proposes a new factor (Section 4). Our main results (Theorem 1 and Table 3) provide a lower bound on the threshold needed to achieve a desired level for the LFDR. These results rely on an assumption that the tails of the non-null distribution of the absolute t-statistics are thicker than those of the Gaussian distribution (see Section A.4 Appendix for details). This is plausible in settings where the power of the tests tends to one as the sample size increases.

In summary, the main assumptions in this paper rest on the idea that non-null t-statistics tend to be large in absolute value, such that they typically fall outside the interval used to estimate the null distribution.

5.2 Relative cost of Type I and Type II errors

Setting a higher threshold allows us to reduce Type I errors, but at the cost of higher Type II errors. Optimizing the trade-off of Type I and Type II errors involves the estimation of the relative costs of these errors. This is explored in Harvey and Liu (2020).

To motivate, consider two non-finance applications. First, there is a quality control process for the manufacture of home water heaters. There is a trade-off for the manufacturer. To reduce the chance that a defective water heater is sold to the customer, quality control would flag (and recycle) potentially defective units. Ex post, some of the flagged units would not have been defective, and recycling the units

is very costly to the firm. As a result, the company's quality control does not flag that many units. Second, consider the manufacturing of a crucial airline part, say the fuel cutoff switch. The manufacturer has a quality control system and, similar to the water heater, every potentially defective unit that is taken out of production is very costly for the manufacturer. However, the quality control system will be implemented differently in these two situations.

It is important to fully analyze the costs. For the water heater, the more units pulled from production, the higher the cost. However, if fewer units are pulled, there is another cost - servicing and replacing customers' defective units. In addition, annoyed customers can give the manufacturer a bad reputation as a result of their experience. As such, the water heater manufacturer will optimize quality control to balance these costs.

For the airline part, the calculus is much different. There is a cost of pulling the switch from the production line and scrapping or recycling it. However, that cost is trivial compared to the cost of failure that could potentially cause a jet to crash. The optimization of quality control means that many switches are pulled off the line with a huge error rate (switches were fine but falsely classified as defective), to eliminate the chance that even one defective switch makes its way into an aircraft. Put bluntly, there is a vastly different cost to the inconvenience of a cold shower versus death.

This type of analysis rarely makes its way into financial economics applications – though it is common in engineering. Publishing a factor that turns out to be false might negatively impact both the author and journal's reputation, but it is very difficult to measure that cost. Authors may prefer lower thresholds because discovery is difficult and time-consuming. Editors may want lower thresholds because the cost of a Type II error (rejecting a future seminal paper) is high. Perhaps this partially explains the finance profession's hesitancy to move away from the single test threshold of two standard errors.

Incentives may be different in the practice of finance. It might be that the factor is offered as an investment product. If the asset manager is a hedge fund, there are clear costs of Type I errors (launching a false product). First, if the product does not make any money, that means the manager will get a zero incentive fee. Second, reputation is damaged by the launch of a false product. Third, managers are rewarded for the performance of their existing products and are not punished for what they did not launch. This means that the Type II error is less important for the asset manager. This perhaps explains why only a small subset of the hundreds of asset pricing factors that have been published in academic journals make their way into investment products.

The relative cost of Type I and Type II errors is important and is understudied.

Although it goes beyond our paper, it is a promising research area.

5.3 Other causes of false discoveries

Our paper focuses on how thresholds need to be changed due to multiple testing. However, there are many sources of false discoveries that are unrelated to multiplicity. For example, model misspecification can lead to false discoveries. Jagannathan and Wang (1998) show that estimates of premia could be biased if there are omitted factors, and Giglio and Xiu (2021) propose a solution. López de Prado and Zoonekynd (2024) and López de Prado, Lipton and Zoonekynd (2025) examine the role of specification errors.

There is also the case where there is discovery of a factor premium using historical data but the discovery does not replicate out of sample (see McLean and Pontiff, 2016). This is consistent with a false discovery due overfitting the historical data. However, it is also consistent with a true discovery but once noticed, the premium disappears because many investors begin to harvest the premium – driving the price up and reducing the expected return. The effect is arbitraged away. The effect may also disappear, independent of arbitrage, if there is a structural change or intervention – for example, a change in the regulatory environment.

5.4 Failing to account for market frictions

It is routine in the factor literature to assume zero transactions costs. Imposing real world costs changes inference. Transactions cost reduce the factor return and reduce its significance. Indeed, it is not clear what it means to declare a factor "significant" if it is impossible to implement in an applied setting. Transactions costs include regular trading costs, bid-ask spread, market impact costs and borrowing costs. Indeed, the borrowing costs might be very important for smaller capitalization stocks.

However, transactions costs are complicated. While it is mechanical to calculate the transactions costs for a particular factor, very few hold the factor naked. That is, the factor is part of a portfolio. Hence, the transactions cost associated with the factor depend on the investor's portfolio. For example, suppose an investor buys the market portfolio and then overlays a particular factor. This might mean that there are no costly shorts in the portfolio. The factor overlay just reduces weights. Another investor might have a multi-factor portfolio. Here, there may be cancelations, i.e., one factor says to rebalance by buying a particular stock and the other factor is saying sell that stock. The trades net out. A useful transaction cost indicator is the covariance of the existing portfolio weight changes with the factor weight changes.

The bottom line is that it makes no sense to ignore real world frictions. While transactions costs are investor dependent, they unambiguously reduce performance and therefore significance.

5.5 Theory-based vs. data mined factors

Our paper blends both Bayesian and classical statistics perspectives. For example, using Bayes rule, we interpret the FDR for a certain level as the probability that the null is true given that the observed t-statistics is greater than τ. The LFDR sharpens the interpretation and delivers the probability that the null is true for the specific realization of a t-statistic, i.e., τ = absolute value of the t-statistic. However, this is not a full Bayesian analysis since prior beliefs are not incorporated.

Harvey (2017) describes one way of looking at the importance of prior beliefs. Suppose a researcher is engaged in a data mining expedition and tries millions of different factors and discovers one with a t-statistic of 5.0. Another researcher is a theorist and develops a model from first principles without looking at the data. This researcher tests the theoretical prediction and estimates the t-statistic to be 1.9. Although economists generally agree that the economic mechanism is plausible, the t-statistic is marginal. How should we judge these findings? Furthermore, suppose that the economic mechanism in the t-statistic of 5.0 is implausible.

A good example is in CGS who examine more than two million factors.[32] Among their list of top factors is a variable called: (csho-cshpri)/mrc4. The numerator is common shares outstanding minus common shares used to calculate earnings per share (basic). It is not obvious what the economic interpretation of the numerator is. However, the denominator is especially puzzling. It represents rental payment obligations four years in the future. This is a factor that has no economic interpretation, but "works" in terms of statistical tests. Surely, this changes the way that we view the high t-statistic.

The message in Harvey (2017) is that the proper deployment of statistical tools is necessary for the interpretation of the data. However, often it is not sufficient to rely on statistics alone. The economic foundation is also important. The challenge is that the assessment of the economic foundation is somewhat subjective as are the imposition of prior beliefs.

[32]The details are in Table A2 of the working paper version, CGS (2020).

6 Conclusions

Some researchers argue thresholds need to be raised in the factor literature, while others believe that the traditional two standard errors from zero suffices. Our primary goal is to sort out this disagreement. Our analysis suggests that thresholds need to be raised. Further, we offer a method that is easily implemented by researchers, and, importantly, our method does not require the specification of the number of tests conducted.

As part of our analysis, we clear up some statistical confusion in the literature. We show both analytically and by simulation that the correlation among factors and the deviation from standard normality substantially impact inference. Techniques that ignore correlation and non-normality produce thresholds that are far too low leading to large Type I errors.

We also provide a deep dive on the false discovery rate – again, with the goal of clearing up misunderstandings about this important concept. We cast the FDR in a Bayesian context. At a particular level, say 0.05, the Bayesian interpretation tells us the probability that the null is true if the absolute value of the t-statistic exceeds the cutoff. We advocate for an alternative concept called the local FDR. This metric uses more information and conveniently provides the probability that the null is true for an observed t-statistic. That is, the LFDR focuses on the particular test - not all tests within a region.

Our use of the LFDR does not relocate the requirement from "tell me how many you tried" to "show me the entire distribution of what you tried." In finance, this distribution is often unavailable, especially with survivorship, publication bias, and hidden p-hacking. Crucially, utilizing the LFDR allows us to derive a lower bound that is independent of the number of tests. To be clear, our contribution is to show that we can avoid having to estimate the distribution of the t-statistics; however, to do so, we can only derive a lower bound – which we believe is a useful measure for researchers. The bound suggests that those conducting anomaly work should use thresholds in the range of 3.1-3.5 for t-statistics.

Our results are sharply different from some recent papers which argue that thousands of factors are "true". However, common sense and external validity suggest that this is implausible. For example, in randomly generating tens of thousands of factors from Compustat data, it does not make sense that there is a 40% chance that any one is a profitable trading strategy. Again, 40% is driven by counterfactual assumptions that lead to a very high false positive rate.

Our findings also exhibit some external validity. For example, there was an explosion of anomaly research in the 1990s and early 2000s. Many of these factors

have been implemented by asset managers and ETF products have been launched. Performance is unremarkable in this out-of-sample experiment. For example, the GSAM multifactor index is flat for the past 10 years.[33] Of course, this could be bad luck. Similarly, Brightman et al. (2015) and Ben-David et al. (2023) look at ETF launches of popular factors. In the SEC-required published backtest, the results look very promising. The performance in live trading is null on average. These findings are consistent with the important paper by McLean and Pontiff (2016). The external evidence is consistent with our results, but not conclusive.

Another piece of external evidence is to examine the number of factors that major investment managers use (or equivalently the factors that major institutional investors allocate to). The number is very small. For example, $800 billion Dimensional Fund Advisors appears to focus on two: size and value. Again, this is consistent with our results but not conclusive.

Finally, what about the provocative evidence in JKP that the factors that are marginal in previous research appear to replicate in out-of-sample tests? This is a powerful finding that goes beyond any statistical adjustments of a p-value. These results raise another issue not comprehensively addressed in the literature but first mentioned by Asness and Frazzini (2013). Academic factors ignore transaction costs, including the important market impact costs. Any reasonable allowance for transactions costs will reduce significance (because the alpha must decrease). So, one interpretation of the JKP findings is that the particular factors they examine are not investable in-sample or out-of-sample. This speaks to the very meaning of 'replication'. Does replication mean finding a factor retains its premium in an unrealistic setting of zero transactions costs? Indeed, we might mistakenly believe that there is no replication crisis because important real-world frictions are not taken into account. As finance is an applied field and given the deep academic research on market frictions, it seems inappropriate to ignore real-world costs.

Our paper is not intended to be the final word on testing. Any method, including ours, involves assumptions. Indeed, we carefully analyze each of our assumptions. Although we believe our assumptions are plausible, they are transparent, and the reader can decide. Our paper does not formally address the issue of factor duplication, a situation where a factor appears significant but is a noisy combination of other factors (see Supplement S.5). In addition, our framework does not deal

[33]Bloomberg GSAM US Equity Multi Factor Index (BGSUSEMF:IND) has recorded a total return over the last 10 years of -0.35% (as of August 21, 2025). Further, a footnote in the factsheet makes it clear that the index returns do not include transactions costs. So the net return is even lower. https://assets.bbhub.io/professional/sites/10/Bloomberg-GSAM-US-Equity-Multi-Factor-Index-Fact-Sheet.pdf

with how the variation in the relative cost of Type I and Type II errors impacts inferences. Finally, while we believe our new method to establish improved thresholds for statistical testing will be useful to researchers, it is important to recognize the limitation of any statistical model. Researchers should use the statistical model's result as one input in decision making. The economic foundation of the result must also be carefully weighed.

References

Asness, C. and A. Frazzini (2013) The devil in HML's details. *Journal of Portfolio Management* 39, 49-68.

Baltussen, G., B. van Vliet and P. van Vliet (2023) The Cross-Section of Stock Returns before CRSP. *Available at SSRN* 3969743.

Ben-David, I., Franzoni, F., B. Kim and R., Moussawi (2023) Competition for attention in the ETF space. *The Review of Financial Studies* 36, 987-1042.

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 B* 57, 289-300.

Bickel, P.J. and E. Levina (2008) Covariance Regularization by Thresholding. *Annals of Statistics* 36, 2577-2604.

Brightman, C., F. Li and X. Liu (2015) Chasing performance with ETFs. *Research Affiliates* (November).

Chen, A.Y. (2021) The Limits of P-Hacking: Some Thought Experiments. *The Journal of Finance* 76, 2447-2480.

Chen, A.Y. (2024) Do T-Stat Hurdles Need to Be Raised? *Management Science*. Forthcoming.

Chen, A.Y. (2025) Most Claimed Statistical Findings in Cross-sectional Return Predictability Are Likely True. *Journal of Finance: Insights*. Forthcoming.

Chen, A.Y. and T. Zimmermann (2020) Publication Bias and the Cross-Section of Stock Returns. *Review of Asset Pricing Studies* 10, 249-289.

Chen, A.Y., A., Lopez-Lira and T. Zimmermann (2024) Does Peer-reviewed Research Help Predict Stock Returns? *CFR Working Paper* (No. 24-02).

Chordia, T., A. Goyal and A. Saretto (2020) Anomalies and False Rejections. *The Review of Financial Studies* 33 2134-2179.

Chudik, A., G. Kapetanios and M.H. Pesaran (2018) A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models. *Econometrica* 86, 1479-1512.

Efron, B. (2004) Large-scale Simultaneous Hypothesis Testing: the Choice of A Null Hypothesis. *Journal of the American Statistical Association* 99, 96-104.

Efron, B. (2007a) Correlation and Large-Scale Simultaneous Significance Testing. *Journal of the American Statistical Association* 102, 93-103.

Efron, B. (2007b) Size, Power and False Discovery Rates. *Annals of Statistics* 35, 1351-1377.

Efron, B., R. Tibshirani, J.D. Storey and V. Tusher (2001) Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association* 96, 1151-1160.

Fama, E.F. and K.R. French (2021) The Value Premium. *Review of Asset Pricing Studies* 11, 105-121.

Fan, J. and J. Lv (2008) Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society Series B* 70, 849-911.

Fan, J., X. Han and W. Gu (2012) Estimating False Discovery Proportion Under Arbitrary Covariance Dependence. *Journal of the American Statistical Association* 107, 1019-1035.

Giglio, S. and D. Xiu (2021) Asset Pricing with Omitted Factors. *Journal of Political Economy* 129, 1947-1990.

Harvey, C.R. (2017) Presidential Address: The Scientific Outlook in Financial Economics. *The Journal of Finance* 72, 1399-1440.

Harvey, C.R. and Y. Liu (2020) False (and Missed) Discoveries in Financial Economics. *The Journal of Finance* 75, 2503-2553.

Harvey, C.R., Y. Liu and H. Zhu (2016). ...And the Cross-Section of Expected Returns. *The Review of Financial Studies* 29, 5-68.

Hou, K., C. Xue and L. Zhang (2020) Replicating Anomalies. *The Review of Financial Studies* 33, 2019-2133.

Jensen, T.I., B. Kelly and L.H. Pedersen (2023) Is There a Replication Crisis in Finance? *The Journal of Finance* 78, 2465-2518.

Jagannathan, R. and Z. Wang (1998) An Asymptotic Theory for Estimating Beta-Pricing Models Using Cross-Sectional Regression. *Journal of Finance* 53, 1285-1309.

Linnainmaa, J.T. and M.R. Roberts (2018) The History of the Cross-Section of Stock Returns. *The Review of Financial Studies* 31, 2606-2649.

López de Prado, M. and M.J. Lewis (2019) Detection of False Investment Strategies Using Unsupervised Learning Methods. *Quantitative Finance* 19, 1555-1565.

López de Prado, M., A. Lipton and V. Zoonekynd (2024) The Case for Causal Factor Investing. *The Journal of Portfolio Management* 51, 146-158.

López de Prado, M. and V. Zoonekynd (2025) Why has Factor Investing Failed?: The Role of Specification Errors. The Role of Specification Errors. FEB-RN Research Paper No. 101/2025, Available at SSRN 4697929.

McLean, R.D. and J. Pontiff (2016) Does Academic Research Destroy Stock Return Predictability?. *The Journal of Finance* 71, 5-32.

Meinshausen, N. and P. Bühlmann (2006) High-Dimensional Graphs and Variable Selection with the Lasso. *Annals of Statistics* 34, 1436-1462.

Novy-Marx, R. and M. Z. Velikov (2025). Assaying Anomalies. *Available at SSRN* 4338007.

Romano, J.P. and M. Wolf (2007). Control of Generalized Error Rates in Multiple Testing. *Annals of Statistics* 35, 1378-1408.

Sellke, T., M.J. Bayarri and J. Berger (2001) Calibration of p-Values for Testing Precise Null Hypotheses. *American Statistician* 55, 62-71.

Shwartzman, A. and X. Lin (2011) The Effect of Correlation in False Discovery Rate Estimation. *Biometrika* 98, 199-214.

Storey, J.D. (2003) The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *Annals of Statistics* 31, 2013-2035.

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 B* 66, 187-205.

Yan, X. and L. Zheng (2017) Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach. *The Review of Financial Studies* 30, 1382-1423.

Wasserstein, R.L. and N.A. Lazar (2016) The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 70, 129–133.

Appendix

A.1 Estimating Parameters in Section 3.2

A.1.1 Finding the proportion of true null hypotheses

Determining the proportion of null hypotheses $\mathbb{P}(H_0)$, i.e., the expectation of $1 - A_i$ in (2), is essential for reducing Type II errors. This section applies the method of Storey et al. (2004) and revisits (5) using p-values. Working with p-values clarifies the assumptions typically made when computing the FDR. For simplicity, assume identically distributed hypotheses and let P denote the common p-value (suppressing the i subscript). By the rules of probability, the unconditional distribution of P is

$$\mathbb{P}(P \geq p) = \mathbb{P}(P \geq p|H_0)\,\mathbb{P}(H_0) + \mathbb{P}(P \geq p|H_1)\,\mathbb{P}(H_1). \tag{A.1}$$

The null hypotheses are $H_0^i : \alpha_i = 0$ against two-sided alternatives. Non-null p-values converge to zero as sample size grows.[34] Thus, in finite samples it is reasonable to treat sufficiently large p-values (e.g., $P \geq 0.5$) as null, implying that the second term in (A.1) is zero. Under the standard assumption that null p-values are uniformly distributed, we have $\mathbb{P}(P \geq p \mid H_0) = 1 - p$.[35] Hence, for a choice such as $p = 0.5$, Equation (A.1) can be approximated as

$$\mathbb{P}(P \geq p) \simeq (1 - p)\,\mathbb{P}(H_0).$$

The only unknown term is $\mathbb{P}(P \geq p)$ on the left-hand side, which we estimate by the sample proportion of p-values exceeding p. For a large number of hypotheses N, this yields the estimator for the proportion of true nulls:

$$\widehat{\mathbb{P}(H_0)} = \frac{\#\{P_i \geq p\}/N}{(1-p)}, \tag{A.2}$$

where $\#\{P_i \geq p\}$ denotes the number of p-values greater than or equal to p. The choice of p in (A.2) reflects a bias/variance trade-off: larger p reduces bias by excluding non-null p-values but increases variance by using fewer observations. The goal is to select p large enough to exclude non-null p-values without discarding too many nulls. A simple, intuitive choice is $p = 0.5$.

We illustrate the estimation procedure with simulated data resembling the setting in Section 3.2. Consider 29,314 hypotheses, 90% of which correspond to zero-alpha

[34] Non-null t-statistics diverge, so their two-sided p-values converge to zero.

[35] For a $[0,1]$-uniform random variable U, $\mathbb{P}(U \leq u) = u$, $u \in [0,1]$.

factors, so that $\mathbb{P}(H_0) = 0.9$. The remaining 10% have nonzero alpha with a Sharpe ratio of 0.5, and the sample size is $n = 600$ (roughly 50 years of monthly data).[36] We assume that the market effect has been removed via (1) and that errors are i.i.d. normal. This implies that the non-null t-statistics have a mean of $\mu = 3.53$.[37]

Figure A.1 illustrates an example of the estimation procedure. With $p = 0.5$, the left panel shows that the red sorted p-values are almost entirely null, with only four contaminated by non-nulls (black crosses). The right panel indicates that this corresponds to treating factors with t-statistics between -0.55 and 0.55 as null. Thus, to estimate the proportion of true (non-null) hypotheses a in this simulated dataset, we count $\#\{P_i \geq 0.5\}$ and obtain:

$$\widehat{\mathbb{P}(H_1)} = 1 - \frac{\#\{P_i \geq 0.5\}/N}{(1-p)} = 1 - \frac{13178/29,314}{(1-0.5)} \simeq 0.10.$$

At no point do we use the knowledge that 10% of the hypotheses are true, yet the procedure recovers this proportion precisely.

A.1.2 Estimation of empirical distribution of t-statistics under the null

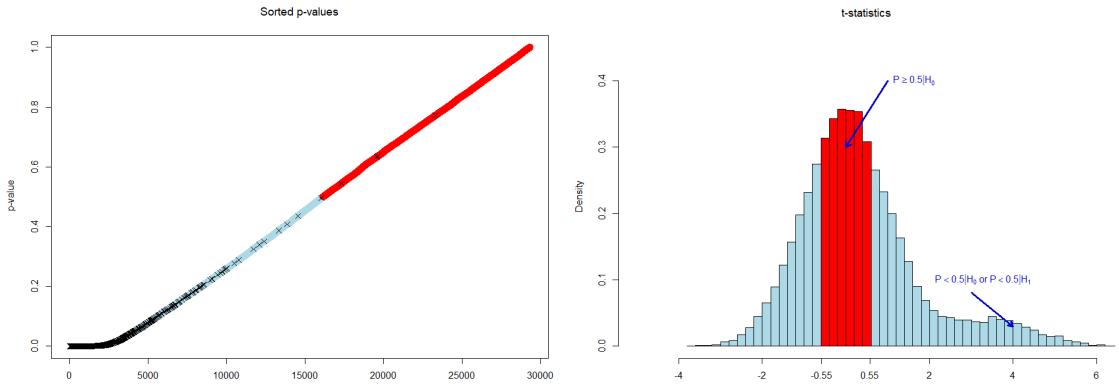
Both the matching and MLE methods share a common assumption: the center of the distribution of the t-statistics only contains null cases. The same principle is also applied in the approach described in Section A.1.1 for estimating $\mathbb{P}(H_0)$, as shown in the right subplot of Figure A.1. This is an identification condition, as otherwise it is impossible to separate the null from the non-null t-statistics. Efron refers to it as the "zero assumption". The center of the distribution is denoted by an interval $[x_l, x_u]$. In the R package `locfdr` the actual values of the lower bound x_l and upper bound x_u are method dependent.[38] In particular, for matching the default is x_l and x_u equal to the 25th and 75th quantile of the t-statistics, which for VW–xMKT in Section 3.2 correspond to $x_l = -1.13$ and $x_u = 0.73$. For the MLE, the default is data-driven and defined via an iterative procedure that yields $x_l = \hat{\mu} - b\hat{\sigma}$ and $x_u = \hat{\mu} + b\hat{\sigma}$,

[36] Working with Sharpe ratios avoids specifying factor volatility. A Sharpe ratio of 0.5 is representative of long-run estimates, such as for the Fama-French HML portfolio, and $n = 600$ corresponds to the median sample size in our empirical setting.

[37] Under these assumptions, non-null t-statistics follow a normal distribution with mean $\sqrt{n/12}\,\mathrm{SR}$ and variance one. When $n = 600$ and $\mathrm{SR} = 0.5$, this yields a mean of 3.53.

[38] From a theoretical point of view $[x_l, x_u]$ could be chosen the same for both methods, but the default software implementation provides good initial values ensuring that the estimation uses enough datapoints.

Figure A.1: Sorted p-values and Density of t-statistics from Simulated Data. The sorted p-values from $N = 29,314$ generated t-statistics are plotted in the left-hand subplot. These p-values are obtained by computing the absolute value of the t-statistic and assuming a standard normal distribution for the t-statistics under the null hypothesis. The non-null p-values are computed from t-statistics with an average of $\mu = 3.53$, corresponding to factors with positive alpha and a Sharpe ratio of 0.50, based on a sample of $n = 600$ monthly returns (50 years). The non-null p-values are indicated by black cross points. The sorted p-values in red represent p-values greater than $p = 0.5$. The sorted p-values in light blue mix both null and non-null p-values. The right-hand subplot displays the corresponding density of t-statistics. The light blue area is dominated by non-null p-values, whereas the central red region is dominated by null p-values.



where $\hat{\mu}$ and $\hat{\sigma}$ are intermediate parameter estimates of the null density $\mathcal{N}(\mu, \sigma^2)$, $b = 4.3 \exp\{-0.26 \log_{10}(N)\}$, and N is the number of t-statistics.[39] In our analysis of Chen's 29K, $b \simeq 1.35$. These values produced a reasonable fit to the center of the density as illustrated in Figure 2. We conducted sensitivity checks around the default values for both methods. For VW and VW–xMKT, we found that different choices that fit the center of the distribution well had little impact on our results. By contrast, the results for EW and EW–xMKT are more sensitive to the choice of the center interval, in part because these distributions are more complex. For further details on the sensitivity checks, we refer to Section S.3 in the Online Supplement. The two methods differ substantially in how they are constructed and details are provided next.

[39]The formula for b is a heuristic internal to the package, designed to provide a reasonable trade-off between bias and variance, resulting in $b \simeq 2$ for $N = 1000$ and $b \simeq 1.5$ for $N = 10000$.

Matching Estimator

The matching estimator is based on the assumption that the null distribution is normal but allows for its mean and variance to differ from the standard normal. This is a reasonable assumption, as discussed in HLZ and Efron (2004, 2007b). Assuming that the empirical distribution of the null t-statistics is normal with mean μ and variance σ^2, we can use information around the peak of the empirical distribution to estimate μ and σ^2, as most t-statistics near the peak are null (the "zero assumption").

There are three steps involved. First, approximate the distribution of the t-statistics using a flexible distribution modeled by an exponential polynomial. The estimation is performed by Poisson regression, first binning the t-statistics to obtain count data for each bin position; the result produces an approximation to the density of the t-statistics. Second, the log of the estimated density evaluated at the bin midpoints is regressed onto a quadratic polynomial, using only bin positions around the peak. Third, the estimated coefficients from the quadratic regression are matched with their corresponding expressions in the logarithm of the normal distribution's density function, thereby obtaining estimators for the parameters μ and σ^2. The first step produces a global fit to the density of the t-statistics. The second step produces a local fit of the density. Under the "zero assumption," this approximates the density of the null t-statistics around the peak because non-null t-statistics are away from the peak. The third step allows us to map the coefficients in the quadratic approximation to the coefficients in the log of the normal density.

Specifically, the first step approximates the whole, or most, of the distribution of the t-statistics by first binning them: for each bin k compute the count y_k and the bin position via its midpoint x_k for $k = 1, 2, \ldots, K$, where K is a predefined number smaller than the number of t-statistics.[40] The method supposes that the t-statistics have density $f(z)$, which can be approximated by the following model in the natural exponential family:

$$f(z) = \exp\left\{\sum_{i=0}^{I} \beta_i z^i\right\}, \tag{A.3}$$

The coefficients β_i are estimated by maximum likelihood, fitting the model to the binned data (y_k, x_k), $k = 1, 2, \ldots, K$, via Poisson regression, and I is an integer.[41]

[40] In the analysis of Section 3.2, we use the default value of 120 break points in the R package `locfdr`, which implies $K = 119$ bins.

[41] The function `locfdr` in the R package `locfdr` sets $I = 7$ as the default value. If the density of the t-statistics is not well approximated by the default choice, the software suggests increasing the value of I. This was the case for the datasets EW and EW–xMKT in Section 3.2, in which case we set $I = 14$. The polynomial inside the exponential can be replaced by a natural spline basis, though no substantive difference in results is observed in practice (Efron, 2007b).

In the second step, the density function is approximated by a quadratic polynomial using only bin midpoint positions in the interval $[x_l, x_u]$. The details are outlined next. Denote the estimator for the density $f(z)$ by $\hat{f}(z)$. We run a quadratic regression to obtain a local quadratic approximation of the log-density:

$$\ln \hat{f}(z) = \hat{\beta}_0 + \hat{\beta}_1 z + \hat{\beta}_2 z^2, \tag{A.4}$$

where the regression evaluates z at x_k for all k such that x_k is in the center of the distribution. This is a local regression in the sense that only bin midpoints inside $[x_l, x_u]$ are used.

The third step uses the assumption that the t-statistics within this interval are null (the "zero assumption"), though their mean μ and variance σ^2 can be different from 0 and 1 either because the central limit theorem does not perform well or because of correlation. We estimate the parameters μ and σ^2 from $\hat{\beta}_i$ in the above expression. To do so, we first use the assumption that all t-statistics in the interval $[x_l, x_u]$ are null (the "zero assumption"), so that $f(z) = (1-a)\,\varphi(z; \mu, \sigma^2)$ for $z \in [x_l, x_u]$, where $\varphi(\cdot; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2 and $1 - a = \mathbb{P}(H_0)$ using the notation in Table 1. Then, we equate (A.4) to

$$\ln\left((1-a)\varphi(z;\mu,\sigma^2)\right) = \ln(1-a) - \frac{1}{2}\left[\ln\left(2\pi\sigma^2\right) + \sigma^{-2}\mu^2\right] + \frac{\mu}{\sigma^2}z - \frac{1}{2\sigma^2}z^2, \tag{A.5}$$

Both (A.4) and (A.5) are quadratic polynomials in z, so we can match the coefficients and solve for the unknown parameters μ and σ^2. The methodology also produces an estimator for $\mathbb{P}(H_0)$.

Maximum Likelihood Estimator

The maximum likelihood estimator is developed based on a similar principle, focusing on the center of the empirical distribution of the t-statistics, but it is a more straightforward one-step estimator. This also exploits the "zero assumption": within the interval $[x_l, x_u]$ around the peak, all the t-statistics are null. Specifically, we can express the probability that N_0 of the N t-statistics fall within this interval as $\theta^{N_0}(1-\theta)^{N-N_0}$, where $\theta(a) := (1-a)\,\mathbb{P}(Z \in [x_l, x_u])$ and Z is a normally distributed variable with mean μ and variance σ^2. Thus, conditioning on N_0, the t-statistics are truncated normal with mean μ and variance σ^2, whose density we denote by

$$\psi\left(z; \mu, \sigma^2\right) = \frac{\varphi\left((z-\mu)/\sigma\right)}{\Phi\left((z_u-\mu)/\sigma\right) - \Phi\left((z_l-\mu)/\sigma\right)},$$

where φ and Φ represent the standard normal density and cumulative distribution functions, respectively.

The joint likelihood of N and the t-statistics in $[x_l, x_u]$ is

$$L(\mu, \sigma, 1-a) = \theta(a)^{N_0} (1-\theta(a))^{N-N_0} \prod_{i:T_i \in [x_l, x_u]} \psi\left(T_i; \mu, \sigma^2\right)$$

(Efron, 2007b, page 1366). Recall that T_i is the i^{th} t-statistic. Consequently, the parameters can be estimated by maximum likelihood. Even if the t-statistics are dependent, the above expression can be seen as a quasi-likelihood, and the estimators are still consistent. However, the standard errors will require adjustment.

This estimator is less noisy than the matching estimator, but it relies more heavily on the choice of the interval around the peak.

A.2 JKP Does Not Control the FDR

In this section, we explain why the JKP method does not appropriately control the FDR. Conditioning on the data and on observing at least one rejection, equation (3) in the Bayesian framework can be rewritten as

$$\text{FDR} = \frac{\sum_{i=1}^{N} \delta_i \left[1 - \mathbb{P}\left(H_1^i | \text{Data}\right)\right]}{\sum_{i=1}^{N} \delta_i},$$

where $1 - \mathbb{P}\left(H_1^i|\text{Data}\right) = 1 - \mathbb{E}\left[A_i|\text{Data}\right]$, and A_i is as in (2). This is just the probability of the null hypothesis i conditioning on the data.

JKP use the rejection rule $\delta_i = I\left\{\mathbb{P}\left(\alpha_i \leq 0|\text{Data}\right) \leq 0.025\right\}$, where $\mathbb{P}\left(\alpha_i \leq 0|\text{Data}\right)$ is the posterior probability of the alpha for the i^{th} factor. Hence, they reject the null if the posterior probability of the alpha being less than zero is less than 2.5%. In their method, they equate the posterior probability $\mathbb{P}\left(\alpha_i \leq 0|\text{Data}\right)$ to $1 - \mathbb{P}\left(H_1^i|\text{Data}\right)$ (which is equal to $\mathbb{P}\left(H_0^i|\text{Data}\right)$).[42] However, it is not true that the posterior probability of the alpha being less than zero is equal to the posterior probability of the null. Note that

$$\begin{aligned} \mathbb{P}\left(\alpha_i \leq 0|\text{Data}\right) = {} & \mathbb{P}\left(\alpha_i \leq 0|H_0^i, \text{Data}\right) \mathbb{P}\left(H_0^i|\text{Data}\right) \\ & + \mathbb{P}\left(\alpha_i \leq 0|H_1^i, \text{Data}\right) \mathbb{P}\left(H_1^i|\text{Data}\right). \end{aligned} \tag{A.6}$$

It is thus evident that $\mathbb{P}\left(\alpha_i \leq 0|\text{Data}\right) \neq \mathbb{P}\left(H_0^i|\text{Data}\right)$ in general.

[42] JKP (2023) do not derive their method from the definition of the FDR in (4). They compute the posterior distribution of the alphas (JKP, Eq.27) and implicitly assume that this is the posterior of the null. This is true only under restrictive and unrealistic conditions as discussed in this section.

Here we show that the two terms are equal only if the prior for α_i assigns zero weight to positive values under the null and zero weight to negative values under the alternative, i.e., $\mathbb{P}(\alpha_i \geq 0|H_0^i) = 0$ and $\mathbb{P}(\alpha_i \leq 0|H_1^i) = 0$. In this case, if we also assume that the prior of α_i is symmetric around zero – as JKP do on page 9 of their appendix– we have that $\mathbb{P}(H_0^i) = 0.5$. By Bayes' rule, the probability of the null given the data is

$$\mathbb{P}\left(H_0^i|\text{Data}\right) = \frac{\mathbb{P}\left(\text{Data}|H_0^i\right)\mathbb{P}\left(H_0^i\right)}{\mathbb{P}\left(\text{Data}\right)}. \tag{A.7}$$

JKP explicitly assume that $\mathbb{P}(H_0^i) = \mathbb{P}(H_1^i) = 1/2$, and implicitly assume the priors $F_0(x) := \mathbb{P}(\alpha_i \leq x|H_0^i)$ and $F_1(x) := \mathbb{P}(\alpha_i \leq x|H_1^i)$ to be truncated normal distributions in $(-\infty, 0]$ and $[0, \infty)$, respectively (there is no loss of generality in assuming the variances are the same in what follows). By the rules of conditional probability,

$$\mathbb{P}\left(\text{Data}|H_0^i\right) = \int_{-\infty}^{0} \mathbb{P}\left(\text{Data}|\alpha_i = x, H_0^i\right) dF_0(x).$$

Substituting the above into (A.7), we conclude that

$$\begin{aligned}\frac{\mathbb{P}\left(\text{Data}|H_0^i\right)\mathbb{P}\left(H_0^i\right)}{\mathbb{P}\left(\text{Data}\right)} =& \frac{\int_{-\infty}^{0} \mathbb{P}\left(\text{Data}|\alpha_i = x, H_0^i\right) dF_0(x)\,\mathbb{P}\left(H_0^i\right)}{\mathbb{P}\left(\text{Data}\right)} \\ =& \mathbb{P}\left(\alpha_i \leq 0|\text{Data}, H_0^i\right)\mathbb{P}\left(H_0^i\right) = \mathbb{P}\left(\alpha_i \leq 0|\text{Data}\right).\end{aligned}$$

The last equality follows from (A.6) and the fact that $\mathbb{P}(\alpha_i \leq 0|H_1^i) = 0$, because the prior under the alternative assigns zero probability to negative alphas. This condition also means that

$$\mathbb{P}(\alpha_i \leq 0) = \mathbb{P}(\alpha_i \leq 0 \mid H_0^i)\,\mathbb{P}(H_0^i) = \mathbb{P}(H_0^i),$$

where the second equality follows from the fact that the above calculations also imply $\mathbb{P}(\alpha_i \leq 0 \mid H_0^i) = 1$; the null density is zero for positive α_i. Since JKP assume that the prior of α_i is normal with mean zero, it is symmetric around zero, so the left-hand side of the above display equals $1/2$. Hence, the crucial assumptions in JKP are that $\mathbb{P}(H_0^i) = \mathbb{P}(H_1^i) = 1/2$ and that the priors under the null and alternative hypotheses assign support only to negative and positive α_i, respectively. This is the most stringent part of the assumptions, as it implies that positive α_i can only be observed under the alternative. This is a very restrictive condition.

JKP report good performance in simulations, as their design produces half of the simulated alphas to be positive, implicitly implying that about 50% of the alphas are true discoveries. However, as shown in the simulations in Section S.4 of the Online

Supplement, their method fails to properly control the FDR when the proportion of true discoveries falls below 50%. Empirically, Figure 6 in JKP shows that their Bayesian method yields results nearly identical to those obtained via OLS.

A.3 The Effect of $\mathbb{P}(H_0)$ on the Cutoff

This section verifies that the choice of cutoff is not highly sensitive to the value of $\mathbb{P}(H_0)$. Assuming that $\mathbb{P}(T \geq \tau | H_0)$ is the standard normal survival function, we have that $\mathbb{P}(T \geq \tau | H_0)$ is approximately a constant multiple of $e^{-\tau^2/2}/\tau$ for large τ. From (4) and by inverting $\mathbb{P}(T \geq \tau | H_0)$, we can then deduce that τ is approximately equal to

$$\sqrt{\ln\left(\mathbb{P}(H_0) / \left[\mathbb{P}(T \geq \tau) \times \text{FDR}\right]\right)},$$

which implies that the effect of $\mathbb{P}(H_0)$ on τ is of the order of $\sqrt{\ln(\mathbb{P}(H_0))}$. We are discarding the effect of the $1/\tau$ term in $e^{-\tau^2/2}/\tau$ as it is secondary. Note that the term FDR is constant and set by the researcher (e.g., 0.05), and that $\mathbb{P}(T \geq \tau)$ has a second-order effect on τ, assuming that $\mathbb{P}(T \geq \tau | H_0)$ is of smaller order than $\mathbb{P}(T \geq \tau)$ for large τ.

To see that this is plausible, recall the following equation:

$$\mathbb{P}(T \geq \tau) = \mathbb{P}(T \geq \tau | H_0)\,\mathbb{P}(H_0) + \mathbb{P}(T \geq \tau | H_1)\,\mathbb{P}(H_1).$$

We assume that $\mathbb{P}(T \geq \tau | H_1)$ has tails thicker than those of a normal random variable, for example, exponential tails. In this case, for large τ, the term $\mathbb{P}(T \geq \tau | H_1)\,\mathbb{P}(H_1)$ dominates, so under the exponential assumption, $\mathbb{P}(T \geq \tau)$ is approximately $\mathbb{P}(H_1)\,e^{-c\tau}$ for some $c > 0$. For simplicity, set FDR $= 0.05$. Then, we have:

$$\sqrt{\ln\left(\frac{\mathbb{P}(H_0)}{\left[\mathbb{P}(T \geq \tau) \times \text{FDR}\right]}\right)} = \sqrt{\ln(\mathbb{P}(H_0)) - \ln(0.05) - \ln(\mathbb{P}(T \geq \tau))}.$$

This simplifies to:

$$\sqrt{\ln\left(\frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}\right) - \ln(0.05) + c\tau}.$$

Hence, to be more specific, the effect of $a = 1 - \mathbb{P}(H_0)$ on the cutoff point τ is of the order of $\sqrt{\ln\left(\frac{(1-a)}{a}\right)}$, which is the square root of the log odds ratio.

A.4 Proof of Theorem 1

Using (4) and (14) together with Bayes' theorem, we have that

$$\text{LFDR}(\tau) = \mathbb{P}\left(H_0 \middle| |T| = \tau\right) = \frac{\text{pdf}\left(\tau | H_0\right)\mathbb{P}\left(H_0\right)}{\text{pdf}\left(\tau\right)}.$$

Define $\text{BF}(\tau) = \text{pdf}\left(\tau \middle| H_0\right)/\text{pdf}\left(\tau \middle| H_1\right)$ to be the Bayes factor in favor of the null. Using this definition and some algebra, the posterior probability of the null given $|T| = \tau$ is:

$$\mathbb{P}\left(H_0 \middle| |T| = \tau\right) = \frac{\text{BF}(\tau)\mathbb{P}\left(H_0\right)}{\text{BF}(\tau)\mathbb{P}\left(H_0\right) + \left[1 - \mathbb{P}\left(H_0\right)\right]}.$$

Note that $x/\left(x+a\right) = 1/\left[1 + \left(a/x\right)\right]$ is increasing in x for any $a > 0$, implying that the above expression increases with $x = \text{BF}(\tau)\mathbb{P}\left(H_0\right)$. The Vovk–Sellke bound on the Bayes factor is $\text{BF}(\tau) \geq -e \times p\left(\tau\right)\ln\left(p\left(\tau\right)\right)$ for any $0 < p\left(\tau\right) < \frac{1}{e}$, where $p\left(\tau\right) = 2\left[1 - \Phi\left(\tau\right)\right]$. The bound holds as long as the ratio of the hazard rate of the t-statistic under the alternative to that under the null is non-increasing (Sellke et al., 2001, eq. 18). We elaborate on the definition of hazard function and the condition at the end of this proof. Defining $\text{bf}\left(\tau\right) = -e \times p\left(\tau\right)\ln\left(p\left(\tau\right)\right)$ and substituting into the expression above, we obtain the statement of the theorem.

Remarks on the assumption. We remark on the conditions under which the ratio of the hazard rate function of $|T|$ under the alternative to that under the null is non-increasing. If $X \geq 0$ has distribution function G and probability density function g, then the hazard rate function at $x > 0$ is given by $g(x)/[1 - G(x)]$. When g is the standard normal density, the hazard rate function at x is approximately equal to x for sufficiently large x, e.g., $x > 1$. Therefore, we require that the hazard rate under the alternative, evaluated at x, does not grow faster than x. This condition is satisfied for any density that is monotonically decreasing and has tails of the form ce^{-ax^b} for $a, b, c > 0$ with $b \leq 2$, or of the form ax^{-b} for $a, b > 0$.

A.5 Testing Steps

We summarize feasible testing procedures for researchers dealing with either multiple or single factor tests. Algorithm 1 outlines the procedure for testing multiple factors, as implemented in Section 3.2. Details on estimating $\mathbb{P}(H_1)$ and the null distribution using matching and MLE methods in Step 2 are provided in Section A.1. Estimation of $\mathbb{P}(H_1)$ using Storey's estimator is described in Section

A.1.1. When all competing estimators are implemented, disagreement among the results suggests that the distribution of the t-statistics may be complex and that the matching estimator could require a larger number of terms I in the model (A.3). In such cases, the MLE is generally preferred (Efron, 2007b, Section 4).

Algorithm 1 Testing Multiple Factors

Input: A set of value-weighted test factors assumed not to be affected by sample selection bias; and a desired level for FDR control (e.g., 5% or 10%).

Step 1: Estimate model (1) for each factor to obtain the t-statistics of the estimated alphas.
Step 2: Estimate the null distribution $\mathcal{N}(\mu, \sigma)$ and $\mathbb{P}(H_1)$ using either the matching or MLE estimators. *(Estimation of $\mathbb{P}(H_1)$ via Storey's method should give a similar result if p-values are computed using the null density estimated from the matching or MLE method.)*
Step 3: For a given threshold τ, reject all factors with t-statistics whose absolute value is at least τ.
Step 4: Estimate $\mathbb{P}(|T| \geq \tau)$ as the number of rejections divided by the total number of tests.
Step 5: Set the FDR in equation (4) to the desired control level, replace the unknown quantities with their estimates from Steps 2 and 4, and solve for τ.
Step 6: Apply the rejection rule from Step 3 using the threshold τ obtained in Step 5. This yields factor discoveries with the FDR controlled at the desired level.

Output: The set of factors identified as discoveries in Step 6.

When a researcher proposes a new factor, readers are typically unable to assess how many tests were conducted. For this reason, when reporting empirical results on a new factor, researchers should report a lower bound on the local FDR corresponding to the observed absolute value of the t-statistic for different values of $\mathbb{P}(H_0)$, such as 0.5, 0.75, and 0.9. The local FDR represents the probability that the null hypotheses is true given the observed value of the t-statistic. For example a local FDR of 20% corresponds to 4-to-1 odds in favor of the alternative (see (17)). This is a more meaningful quantity than a p-value and it inherently accounts for multiple testing. The lower bound on the local FDR can be computed using Table 3 or the software we provide. Our replication package includes user-friendly functions and illustrative examples to guide users.

Similarly, hypothesis testing for a new factor should be conducted using a local FDR-based procedure. This consists of setting a desired local FDR control level, e.g., 20%. A lower bound threshold can then be derived by varying $\mathbb{P}(H_0)$ over a reasonable range, such as $[0.5, 0.95]$. For example, selecting $\mathbb{P}(H_0) = 0.95$ yields a threshold of 3.39 in Table 3. If the observed t-statistic exceeds this threshold, the factor can be considered a discovery at the 20% local FDR level under the assumed prior. The `user_guide.R` file in the GitHub repository provides step-by-step guidance, including examples, for testing factors under both the multiple-testing framework controlling the FDR and the single-factor discovery framework controlling the lower bound of the LFDR.

A.6 Simulations on FDR versus LFDR

We conduct simulations to illustrate the performance of the FDR and LFDR procedures discussed in Section 3 and Section 4.

Since our argument relies on the fact that, conditioning on a sample realization, the distribution of the t-statistics is distorted in the presence of correlation (recall (10)), we cannot simply simulate data. This would lead to the same issues highlighted in Figure 1: averaging across replicates is equivalent to removing the conditioning effect of $\bar{v}$ in (10)) and yields a standard normal distribution. Hence, we need to use an approach that conditions on specific realizations rather than unconditional simulation of the t-statistics.

Specifically, we apply the Fama and French (2010) bootstrap to the EW, EW–xMKT, VW, and VW–xMKT datasets, generating 10,000 bootstrap replications for each and retaining only those replicates satisfying specific criteria. Hence, this procedure resembles rejection sampling (von Neumann, 1951), and proceeds as follows. First, we bootstrap demeaned factors, and compute the 29K t-statistics for each bootstrap. This step mirrors the exercise in Figure 1, yielding 10,000 bootstrap samples of t-statistics under H_0. Second, from these 10,000 replicates, we retain only those for which the absolute sample mean and standard deviation of the t-statistics within that replicate exceed target levels. The choice of target levels ($|\mu| > \mu^\star$ and $\sigma > \sigma^\star$) is guided by the empirical results for the null density in Table 2 and is as follows: for EW $\mu^\star = 0.1$ and $\sigma^\star = 1.4$, EW–xMKT $\mu^\star = 0.1$ and $\sigma^\star = 1.4$, VW $\mu^\star = 0.1$ and $\sigma^\star = 1.2$, and VW–xMKT $\mu^\star = 0.1$ and $\sigma^\star = 1.2$. This step ensures that the retained bootstrap replicates adequately reproduce the characteristics of the original 29K dataset. Finally, for each selected bootstrap replicate, we add back the mean effect to a proportion $\mathbb{P}(H_1)$ of the t-statistics–specifically, those corresponding to the largest means obtained in the first step, so that these observations represent

factors under H_1. The number of selected samples depends on selected target levels $\mu^\star$ and $\sigma^\star$ as well as on the datasets. Increasing the bootstrap size would yield additional samples. Given the large cross-sectional dimension (29,314), we argue that additional samples would not materially alter the results reported here.

We compare four methods for controlling Type I and II errors. The first two control the FDR at 5% using (4) and (16), denoted EQ4 and EQ16, respectively. In (16) we replace the threshold 3 with the appropriate cutoff τ. The third method controls the LFDR using (13), denoted EQ13. We estimate the null distribution using the matching and MLE estimators. The last method, denoted LB, controls the lower bound of the LFDR using the cutoff values in Table 3, for $\mathbb{P}(H_1) = 0.1$. The LFDR is controlled at the 10% and 20% levels. For each method, we compute the number of true positives (TP), the number of false positives (FP), the true positive rate $\text{TP}/(N \times \mathbb{P}(H_1))$ (TPR) where N is the number of tests, the false discovery rate (FDR), and accuracy (ACC), defined as the sum of correctly rejected non-nulls and correctly non-rejected nulls over N. A high FDR indicates a larger proportion of Type I–like errors, whereas a high TPR corresponds to greater statistical power.

Table A.1 reports the results for the EW and VW datasets. The key findings can be summarized as follows. First, the results show that imposing a standard normal null leads to higher TPR but at the cost of substantially inflated FDR, particularly in the EW dataset. For example, the FDR is a massive **56%** for the EW and **30%** for the VW data. This pattern appears across all procedures that use the standard normal null, whether controlling FDR or LFDR. Second, estimating the null using the matching or MLE methods effectively controls the FDR without materially reducing the TPR in the EW datasets, yielding higher overall accuracy. However, these methods can markedly reduce the TPR in the VW datasets. This pattern suggests that the MLE and matching estimators perform well when the null distribution deviates substantially from the standard normal, but when the null t-statistics are close to standard normal, these methods may increase Type II errors. To provide further insight, Figure A.3 displays the fitted null densities obtained using the MLE, the matching estimator, and the standard normal density. The histogram of the t-statistics is well approximated by the MLE and the matching estimator, while the standard normal exhibits a noticeably poorer fit. By poorer fit we mean that the body of the distribution diverges substantially from the standard normal. By construction of the simulation, the t-statistics in the body of the distribution are essentially all null.

Lastly, the lower-bound method achieves a more balanced trade-off between TPR and FDR, and controlling the LFDR at 10% yields the highest overall accuracy across all EW and VW datasets. These findings are broadly consistent in the EW–

xMKT and VW–xMKT datasets, as shown in Table A.2. In summary, when testing dependent hypotheses, assuming a standard normal null leads to invalid results. Instead, the null density can be estimated using either the MLE or matching method. Preferably, our lower-bound approach offers an interpretable balance between Type I and Type II errors, as discussed in Section 4.

We conduct additional bootstrap simulations by varying the target levels for the absolute sample mean, standard deviation, and the proportion of true non-nulls $\mathbb{P}(H_1)$. These results are omitted for brevity, and we summarize only the main takeaways here. Clearly, by selecting less or more extreme values for the target levels of $\mu^\star$ and $\sigma^\star$, we can make the standard normal fit more or less appropriate. For example, under $\mathbb{P}(H_1) = 0.1$, setting the selection target to $\mu^\star = 0.1$ and $\sigma^\star = 1.6$ causes EQ4, when used with the $\mathcal{N}(0, 1)$-null, to yield an FDR as high as 0.68 in the EW dataset.

The results show that sample draws can produce considerable distortion under correlation. Note that if the data are not correlated, no replicates can produce samples with null t-statistics with sample mean and variance substantially different from the theoretical $\mathcal{N}(0, 1)$. This is because the average of 29,314 independent mean zero t-statistics has mean zero with a standard error $1/\sqrt{29314} = 0.006$. This is precisely what can be observed when bootstrapping without preserving cross-sectional dependence: none of the resulting bootstrap samples deviate meaningfully from the standard normal distribution, as shown in Figure A.2. This stands in stark contrast to the results in Figure 1, where both the mean and standard deviations of the t-statistics within each replicate deviate substantially from zero and one. We also note that the 29K dataset is not a balanced panel. Consequently, when resampling, many draws correspond to periods in which certain factors are missing, yielding NAs in the bootstrap sample. These missing observations reduce the amount of overlapping information across factors and thereby mute the cross-sectional correlation in the resulting t-statistics.

We also examine the effect of increasing the proportion of true non-nulls $\mathbb{P}(H_1)$ to 0.4, for example. Using the EW dataset yields an FDR of 0.31, compared with an FDR of 0.56 when $\mathbb{P}(H_1) = 0.1$, using the same selection targets for the replicates (i.e., $\mu^\star = 0.1$ and $\sigma^\star = 1.4$). In contrast, the 10% lower-bound method yields an FDR of 0.04, evidencing its superior performance across different values of $\mathbb{P}(H_1)$.

The objective of our simulation is to assess how multiple-testing procedures perform in samples resembling the observed realization of the 29K dataset. The design is deliberately conditional on matching key empirical features of the 29K distribution, rather than on an unconditional draw from the data-generating process. The unbalanced panel of the 29K dataset makes this matching process

more challenging. To be concrete, more than 40% of factors (about 12,000 of 29,000) have at least 40% missing observations. Restricting attention to the remaining 60% (about 17,000 factors) with the fewest missing observations increases the number of selected samples that meet the target level, for example, from 41 to 53 for EW and from 111 to 138 for VW. Furthermore, only 1,941 factors form a balanced panel in the 29K dataset; under this restriction, the target levels are met in 269 bootstrap samples for EW and 748 for VW. Although these changes are modest, they show that an unbalanced panel materially affects bootstrap outcomes. Under the bootstrap-imposed null, reducing the number of factors should, in principle, lessen distortions by lowering the chance of drawing extreme factors. Yet, the opposite occurs: more samples satisfy the condition, indicating that missing observations continue to distort the empirical null even after dimensionality is reduced. Using the above selected samples with fewer missing observations still produces the same comparative results across multiple-testing procedures.

In addition, we repeat the simulation procedure for EW and VW using less extreme distortion criteria for additional robustness. Table A.3 reports the comparative results based on bootstrapping 17,768 factors from the 29K dataset with fewer than 40% missing values. From the 10,000 bootstrap replicates, we retain those in which the absolute sample mean and standard deviation of the t-statistics exceed the less extreme targets of $\mu^\star = 0$ and $\sigma^\star = 1.1$ for both EW and VW, yielding 1,096 samples for EW and 1,121 for VW, about 10% of the bootstrap draws. These selected samples do not resemble the actual 29K dataset, as the true distortion driven by correlation is more severe. Nevertheless, Table A.3 shows that even under these milder distortions, the standard-normal method still underperforms by exceeding the FDR control for both EW and VW. The comparative patterns across methods remain robust.

Figure A.2: The Bootstrap without Preserving Cross-Sectional Dependence. For each bootstrap replication (x-axis), the figure shows a dot representing the mean (left panel) and the standard deviation (right panel) of the 29,314 t-statistics in that replication. For reference, the mean (left panel) and standard deviation (right panel) of the standard normal distribution are indicated by a horizontal blue line. The average of the t-statistics (left panel) and the square root of the average variance (right panel) across the 1,000 bootstrap replications are indicated by a horizontal dashed red line.

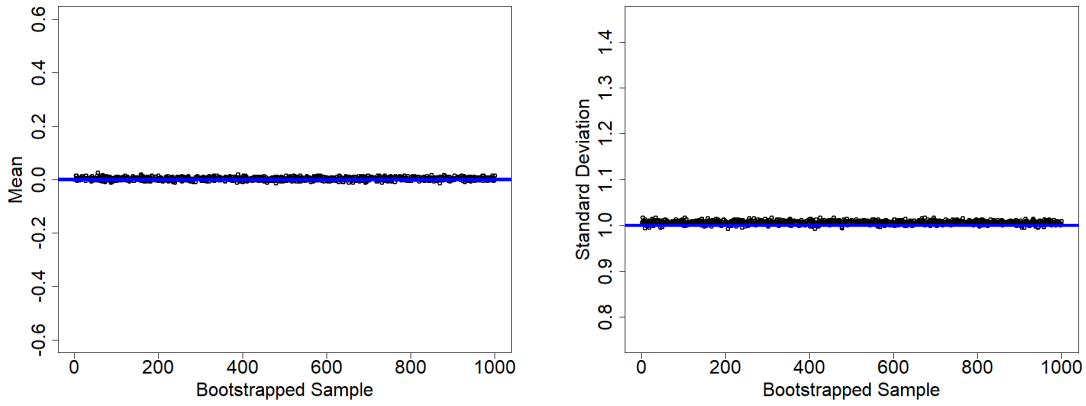


Table A.1: Simulation Results for the Bootstrapped t-statistics from the EW and VW 29K Datasets. The table reports the true positive rate (TPR), the false discovery rate (FDR), and overall accuracy (ACC). We also report the cutoffs and the estimated proportion of true non-nulls, $\mathbb{P}(H_1)$. The columns labeled "μ" and "σ" provide the estimated mean and standard deviation of the null t-statistics for the matching and MLE estimators; for all other methods, these values are fixed at 0 and 1. We set the true $\mathbb{P}(H_1) = 0.1$. From 10,000 bootstrap replications, we select samples satisfying $\mu^\star = 0.1$ and $\sigma^\star = 1.4$ for EW (41 samples) and $\sigma^\star = 1.2$ for VW (111 samples) for use in the simulation study.

EW

Method	Est	Type	Cutoff	μ	σ	$\mathbb{P}(H_1)$	TPR	FDR	ACC
EQ4	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.39, 2.39]	0.00	1.00	0.40	0.88	0.56	0.88
EQ4	Matching	FDR ≤ 0.05	[-4.01, 5.16]	0.57	1.58	0.05	0.69	0.02	0.97
EQ4	MLE	FDR ≤ 0.05	[-4.21, 5.13]	0.46	1.61	0.06	0.68	0.01	0.97
EQ16	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.73, 2.35]	0.00	1.00	0.34	0.85	0.55	0.88
EQ16	Matching	FDR ≤ 0.05	[-3.69, 6.03]	0.57	1.58	0.05	0.70	0.05	0.97
EQ16	MLE	FDR ≤ 0.05	[-3.90, 5.93]	0.46	1.61	0.06	0.68	0.03	0.97
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.10	[-2.96, 2.67]	0.00	1.00	0.34	0.83	0.43	0.92
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.20	[-2.58, 2.17]	0.00	1.00	0.34	0.87	0.61	0.86
EQ13	Matching	LFDR ≤ 0.10	[-4.67, 6.63]	0.57	1.58	0.05	0.59	0.00	0.96
EQ13	Matching	LFDR ≤ 0.20	[-4.20, 6.29]	0.57	1.58	0.05	0.65	0.01	0.96
EQ13	MLE	LFDR ≤ 0.10	[-4.90, 6.59]	0.46	1.61	0.06	0.57	0.00	0.96
EQ13	MLE	LFDR ≤ 0.20	[-4.45, 6.28]	0.46	1.61	0.06	0.63	0.01	0.96
LB	lwr	LFDR ≤ 0.10	[-3.41, 3.41]	0.00	1.00	0.10	0.78	0.15	0.97
LB	lwr	LFDR ≤ 0.20	[-3.16, 3.16]	0.00	1.00	0.10	0.81	0.25	0.96

VW

EQ4	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.70, 2.70]	0.00	1.00	0.31	0.67	0.30	0.94
EQ4	Matching	FDR ≤ 0.05	[-5.19, 5.24]	0.03	1.55	-0.03	0.15	0.00	0.92
EQ4	MLE	FDR ≤ 0.05	[-4.66, 4.91]	0.12	1.46	0.02	0.21	0.00	0.92
EQ16	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.67, 8.39]	0.00	1.00	0.28	0.42	0.14	0.94
EQ16	Matching	FDR ≤ 0.05	[-4.85, 8.39]	0.03	1.55	-0.03	0.15	0.00	0.92
EQ16	MLE	FDR ≤ 0.05	[-4.39, 8.39]	0.12	1.46	0.02	0.20	0.00	0.92
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.10	[-3.15, 8.39]	0.00	1.00	0.28	0.36	0.04	0.94
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.20	[-2.77, 8.39]	0.00	1.00	0.28	0.41	0.11	0.94
EQ13	Matching	LFDR ≤ 0.10	[-5.17, 8.39]	0.03	1.55	-0.03	0.13	0.00	0.91
EQ13	Matching	LFDR ≤ 0.20	[-4.81, 8.39]	0.03	1.55	-0.03	0.16	0.00	0.92
EQ13	MLE	LFDR ≤ 0.10	[-4.76, 8.39]	0.12	1.46	0.02	0.16	0.00	0.92
EQ13	MLE	LFDR ≤ 0.20	[-4.45, 8.39]	0.12	1.46	0.02	0.20	0.00	0.92
LB	lwr	LFDR ≤ 0.10	[-3.41, 3.41]	0.00	1.00	0.10	0.49	0.05	0.95
LB	lwr	LFDR ≤ 0.20	[-3.16, 3.16]	0.00	1.00	0.10	0.55	0.13	0.95

Table A.2: Simulation Results for the Bootstrapped t-statistics from the EW–xMKT and VW–xMKT 29K Datasets. The table reports the true positive rate (TPR), the false discovery rate (FDR), and overall accuracy (ACC). We also report the cutoffs and the estimated proportion of true non-nulls, $\mathbb{P}(H_1)$. The columns labeled "μ" and "σ" provide the estimated mean and standard deviation of the null t-statistics for the matching and MLE estimators; for all other methods, these values are fixed at 0 and 1. We set the true $\mathbb{P}(H_1) = 0.1$. From 10,000 bootstrap replications, we select samples satisfying $\mu^\star = 0.1$ and $\sigma^\star = 1.4$ for EW–xMKT (28 samples) and $\sigma^\star = 1.2$ for VW–xMKT (82 samples) for use in the simulation study.

EW–xMKT

Method	Est	Type	Cutoff	μ	σ	$\mathbb{P}(H_1)$	TPR	FDR	ACC
EQ4	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.41, 2.41]	0.00	1.00	0.39	0.93	0.53	0.89
EQ4	Matching	FDR ≤ 0.05	[-3.42, 4.44]	0.51	1.40	0.11	0.83	0.07	0.98
EQ4	MLE	FDR ≤ 0.05	[-3.83, 4.74]	0.46	1.51	0.09	0.80	0.03	0.98
EQ16	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.77, 2.35]	0.00	1.00	0.32	0.91	0.52	0.90
EQ16	Matching	FDR ≤ 0.05	[-3.04, 4.86]	0.51	1.40	0.11	0.84	0.12	0.97
EQ16	MLE	FDR ≤ 0.05	[-3.46, 5.20]	0.46	1.51	0.09	0.81	0.07	0.98
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.10	[-3.03, 2.72]	0.00	1.00	0.32	0.89	0.38	0.94
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.20	[-2.69, 2.24]	0.00	1.00	0.32	0.91	0.56	0.88
EQ13	Matching	LFDR ≤ 0.10	[-4.01, 5.66]	0.51	1.40	0.11	0.76	0.02	0.97
EQ13	Matching	LFDR ≤ 0.20	[-3.56, 5.29]	0.51	1.40	0.11	0.80	0.06	0.98
EQ13	MLE	LFDR ≤ 0.10	[-4.51, 6.04]	0.46	1.51	0.09	0.71	0.00	0.97
EQ13	MLE	LFDR ≤ 0.20	[-4.06, 5.69]	0.46	1.51	0.09	0.76	0.02	0.97
LB	lwr	LFDR ≤ 0.10	[-3.41, 3.41]	0.00	1.00	0.10	0.86	0.13	0.97
LB	lwr	LFDR ≤ 0.20	[-3.16, 3.16]	0.00	1.00	0.10	0.88	0.23	0.96

VW–xMKT

EQ4	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.62, 2.62]	0.00	1.00	0.32	0.80	0.33	0.94
EQ4	Matching	FDR ≤ 0.05	[-4.64, 4.62]	-0.01	1.50	0.02	0.41	0.00	0.94
EQ4	MLE	FDR ≤ 0.05	[-4.43, 4.62]	0.09	1.48	0.04	0.43	0.00	0.94
EQ16	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.84, 2.45]	0.00	1.00	0.29	0.79	0.35	0.94
EQ16	Matching	FDR ≤ 0.05	[-4.60, 9.12]	-0.01	1.50	0.02	0.25	0.00	0.93
EQ16	MLE	FDR ≤ 0.05	[-4.38, 9.12]	0.09	1.48	0.04	0.28	0.00	0.93
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.10	[-3.20, 2.87]	0.00	1.00	0.29	0.73	0.21	0.95
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.20	[-2.80, 2.41]	0.00	1.00	0.29	0.79	0.37	0.93
EQ13	Matching	LFDR ≤ 0.10	[-5.04, 9.12]	-0.01	1.50	0.02	0.20	0.00	0.92
EQ13	Matching	LFDR ≤ 0.20	[-4.73, 9.12]	-0.01	1.50	0.02	0.24	0.00	0.93
EQ13	MLE	LFDR ≤ 0.10	[-4.86, 9.12]	0.09	1.48	0.04	0.22	0.00	0.92
EQ13	MLE	LFDR ≤ 0.20	[-4.57, 9.12]	0.09	1.48	0.04	0.26	0.00	0.93
LB	lwr	LFDR ≤ 0.10	[-3.41, 3.41]	0.00	1.00	0.10	0.66	0.07	0.96
LB	lwr	LFDR ≤ 0.20	[-3.16, 3.16]	0.00	1.00	0.10	0.71	0.14	0.96

Figure A.3: Density Plots of Bootstrapped t-statistics. The histograms for the EW, VW, EW–xMKT, and VW–xMKT return series, using the first bootstrap replicate as an illustrative example, are shown together with estimates of the null component, $\mathbb{P}(H_0) \times \text{pdf}(\cdot \mid H_0)$, in (14). The dashed blue line represents the null component estimated via matching (left column) and MLE (right column), while the dotted red line shows the null component under the theoretical standard normal null. In the latter case, the quantity $\mathbb{P}(H_0)$ is estimated using Storey's method.

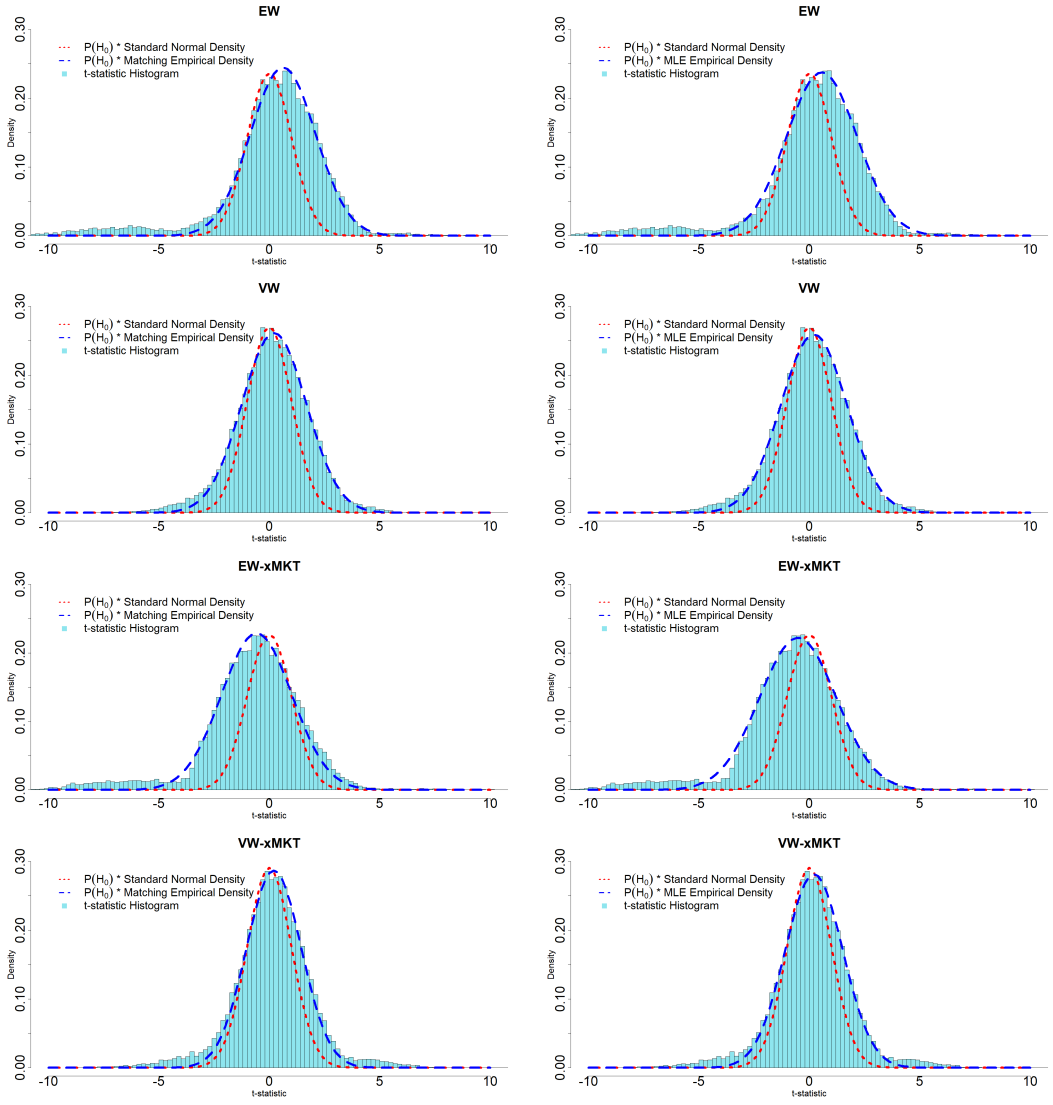


Table A.3: Simulation Results for the Bootstrapped t-statistics from the EW and VW 29K Datasets (17,768 factors with fewer than 40% missing observations). The table reports the true positive rate (TPR), the false discovery rate (FDR), and overall accuracy (ACC). We also report the cutoffs and the estimated proportion of true non-nulls, $\mathbb{P}(H_1)$. The columns labeled "μ" and "σ" provide the estimated mean and standard deviation of the null t-statistics for the matching and MLE estimators; for all other methods, these values are fixed at 0 and 1. We set the true $\mathbb{P}(H_1) = 0.1$. From 10,000 bootstrap replications, we select samples satisfying $\mu^\star = 0$ and $\sigma^\star = 1.1$ for both EW (1,096 samples) and VW (1,121 samples) for use in the simulation study.

EW

Method	Est	Type	Cutoff	μ	σ	$\mathbb{P}(H_1)$	TPR	FDR	ACC
EQ4	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.61, 2.61]	0.00	1.00	0.32	0.97	0.22	0.98
EQ4	Matching	FDR ≤ 0.05	[-2.55, 3.94]	0.70	1.17	0.08	0.96	0.05	0.99
EQ4	MLE	FDR ≤ 0.05	[-2.47, 3.74]	0.63	1.12	0.12	0.96	0.07	0.99
EQ16	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.99, 2.48]	0.00	1.00	0.18	0.95	0.26	0.98
EQ16	Matching	FDR ≤ 0.05	[-2.19, 4.88]	0.70	1.17	0.08	0.95	0.12	0.99
EQ16	MLE	FDR ≤ 0.05	[-2.14, 4.60]	0.63	1.12	0.12	0.96	0.13	0.99
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.10	[-3.30, 3.37]	0.00	1.00	0.18	0.92	0.03	0.99
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.20	[-3.04, 2.62]	0.00	1.00	0.18	0.94	0.20	0.98
EQ13	Matching	LFDR ≤ 0.10	[-3.11, 5.40]	0.70	1.17	0.08	0.90	0.01	0.99
EQ13	Matching	LFDR ≤ 0.20	[-2.75, 5.20]	0.70	1.17	0.08	0.93	0.03	0.99
EQ13	MLE	LFDR ≤ 0.10	[-3.01, 5.12]	0.63	1.12	0.12	0.91	0.01	0.99
EQ13	MLE	LFDR ≤ 0.20	[-2.66, 4.94]	0.63	1.12	0.12	0.94	0.04	0.99
LB	lwr	LFDR ≤ 0.10	[-3.41, 3.41]	0.00	1.00	0.10	0.91	0.02	0.99
LB	lwr	LFDR ≤ 0.20	[-3.16, 3.16]	0.00	1.00	0.10	0.93	0.05	0.99

VW

EQ4	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.88, 2.88]	0.00	1.00	0.24	0.37	0.40	0.95
EQ4	Matching	FDR ≤ 0.05	[-4.25, 4.10]	-0.08	1.23	0.08	0.12	0.06	0.95
EQ4	MLE	FDR ≤ 0.05	[-4.47, 4.33]	-0.07	1.27	0.06	0.10	0.03	0.95
EQ16	$\mathcal{N}(0,1)$	FDR ≤ 0.05	[-2.73, 3.09]	0.00	1.00	0.22	0.38	0.40	0.95
EQ16	Matching	FDR ≤ 0.05	[-3.95, 4.79]	-0.08	1.23	0.08	0.16	0.04	0.95
EQ16	MLE	FDR ≤ 0.05	[-4.12, 5.19]	-0.07	1.27	0.06	0.13	0.03	0.95
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.10	[-2.93, 3.33]	0.00	1.00	0.22	0.34	0.33	0.95
EQ13	$\mathcal{N}(0,1)$	LFDR ≤ 0.20	[-2.54, 2.87]	0.00	1.00	0.22	0.41	0.48	0.94
EQ13	Matching	LFDR ≤ 0.10	[-4.10, 5.19]	-0.08	1.23	0.08	0.13	0.03	0.95
EQ13	Matching	LFDR ≤ 0.20	[-3.68, 4.38]	-0.08	1.23	0.08	0.21	0.09	0.95
EQ13	MLE	LFDR ≤ 0.10	[-4.26, 5.71]	-0.07	1.27	0.06	0.11	0.01	0.95
EQ13	MLE	LFDR ≤ 0.20	[-3.83, 4.95]	-0.07	1.27	0.06	0.18	0.05	0.95
LB	lwr	LFDR ≤ 0.10	[-3.41, 3.41]	0.00	1.00	0.10	0.27	0.21	0.95
LB	lwr	LFDR ≤ 0.20	[-3.16, 3.16]	0.00	1.00	0.10	0.31	0.29	0.95

Online Supplement

S.1 Validity of Equation (4) Under Non-i.i.d. Hypotheses

We review the standard empirical Bayesian FDR approach. We start from the FDP in (2) and divide both the numerator and denominator by N. For simplicity, we assume that the t-statistics have been transformed into p-values. Then, the rejection rule $\delta_i = \delta_i(p)$ is defined as rejecting if the i^{th} p-value is less than p. We made the dependence of δ_i on p explicit by writing $\delta_i(p)$.

Following Efron et al. (2001) and Genovese and Wasserman (2004), the empirical Bayesian FDR is computed using the threshold p, which is computed as the FDR conditioning on the data, and satisfies

$$\mathbb{E}[\text{FDR} \mid \text{Data}] \to (1-a)\, p/S(p), \tag{S.1}$$

in probability, where (p) is the probability limit of $\sum_{i=1}^{N} \delta_i(p)/N$, with δ_i equal to one if the i^{th} p-value is less than p, and zero otherwise. Also, $p(1-a)$ is the probability limit of $\sum_{i=1}^{N} \delta_i(p)(1-A_i)/N$.

Note that $p = \mathbb{E}[\delta_i(p) \mid A_i = 0]$, i.e., under the null, the expectation that a p-value is less than p is equal to p. This is the same as saying that the null p-values are uniformly distributed. Moreover, $(1-a)$ is the probability that the null hypothesis is true, where $a = \mathbb{E}A_i$.

For a desired FDR level equal to 0.05, choose p such that

$$\frac{p(1-a)}{\sum_{i=1}^{m} \delta_i(p)/m} = 0.05.$$

It is crucial to recognize that convergence may fail if the data exhibit dependency, rendering the aforementioned procedure invalid in controlling the FDR, as noted by Shwartzman and Lin (2011). Note that when $a = 0$, the above procedure reduces to the Benjamini-Hochberg (BH) procedure. This results in a conservative method, implying that all null hypotheses are considered true, so all discoveries are expected to be false. Therefore, the Benjamini-Hochberg procedure might be too conservative when we expect some alpha discoveries to be true. On the other hand, when $a \to 1$, it implies that all null hypotheses are false, suggesting that all factors exhibit abnormal returns, which contradicts empirical evidence.

S.2 Problems with the BH Estimator and (4)

Under strong dependence (e.g., exchangeable correlation), the law of large numbers fails, so it is no longer true that in Equation (S.1) we have $S_N \to S$ in probability, where $S_N := \sum_{i=1}^{N} \delta_i / N$, dropping dependence on the p-value cutoff p. Even when convergence holds, it may be slow, so that S_N exhibits high variance. This can have implications when the number of hypotheses is small. Let us consider the FDR estimator for (4) corresponding to a p-value cutoff of 0.05:

$$\widehat{\text{FDR}} = \frac{0.05 \times \mathbb{P}(H_0)}{S_N}$$

The BH procedure assumes $\mathbb{P}(H_0) \simeq 1$. What matters is that the numerator is a constant. If S_N is consistent but converges slowly (i.e., it is highly variable), $\widehat{\text{FDR}}$ tends to over-estimate the FDR. In fact,

$$\mathbb{E}\widehat{\text{FDR}} \geq \frac{0.05 \times \mathbb{P}(H_0)}{\mathbb{E}S_N}$$

by Jensen's inequality, because the map $x \mapsto 1/x$ is convex. Since S_N is an unbiased estimator of the frequency of rejections, the right hand side is just the FDR in (4) when the rejection rule is for $|T| \geq \tau$. We could then say that the estimator over-estimates the FDR. However, the above argument does not apply to (S.1) (or the expectation of (2) in (4)) in general, because in (S.1) also the numerator is random and we cannot appeal to Jensen's inequality.

These remarks show that the situation is nuanced, but difficulties can arise in practice. We justify the use of (4) on two grounds. First, in a large dataset like the 29K, correlations are likely to be strong within certain clusters of factors but weaker between clusters, so the strong law of large numbers still applies. Second, Equation (4) is simple and can be taken as a definition in its own right — namely, the probability that the null holds given that the t-statistics exceed the cutoff in absolute value — and it is supported by a rich literature.

S.3 Sensitivity Checks for Matching and MLE methods

In this section, we conduct a sensitivity analysis to examine how varying the parameters $[x_l, x_u]$ used to determine the center of the distribution affects the results under the Matching and MLE approaches.

We first focus on the Matching method. Let q_s denote the sth quantile of the t-statistic. For example, $q_{0.25}$ is the value such that 25% of the distribution lies below it and 75% lies above it. Using the notation in Section A.1.2, the default is $[x_l, x_u] = [q_{0.25}, q_{0.75}]$. We analyze alternative values $[q_s, q_{1-s}]$ for $s = 0.20, 0.30, 0.35, 0.40, 0.45$. Table S.1 reports the results for EW, EW–xMKT, VW, and VW–xMKT.

Table S.1: Cutoff Points Using the Matching Procedure with Different Choices of $[x_l, x_u]$. Detailed results are reported for EW and EW–xMKT in the upper panel, and for VW and VW–xMKT in the lower panel. For each dataset, the columns labeled 0.2-0.45 correspond to the results based on $[q_s, q_{1-s}]$ with $s = 0.2, 0.3, 0.35, 0.4, 0.45$. The "Default" column reports the baseline results presented in Table 2.

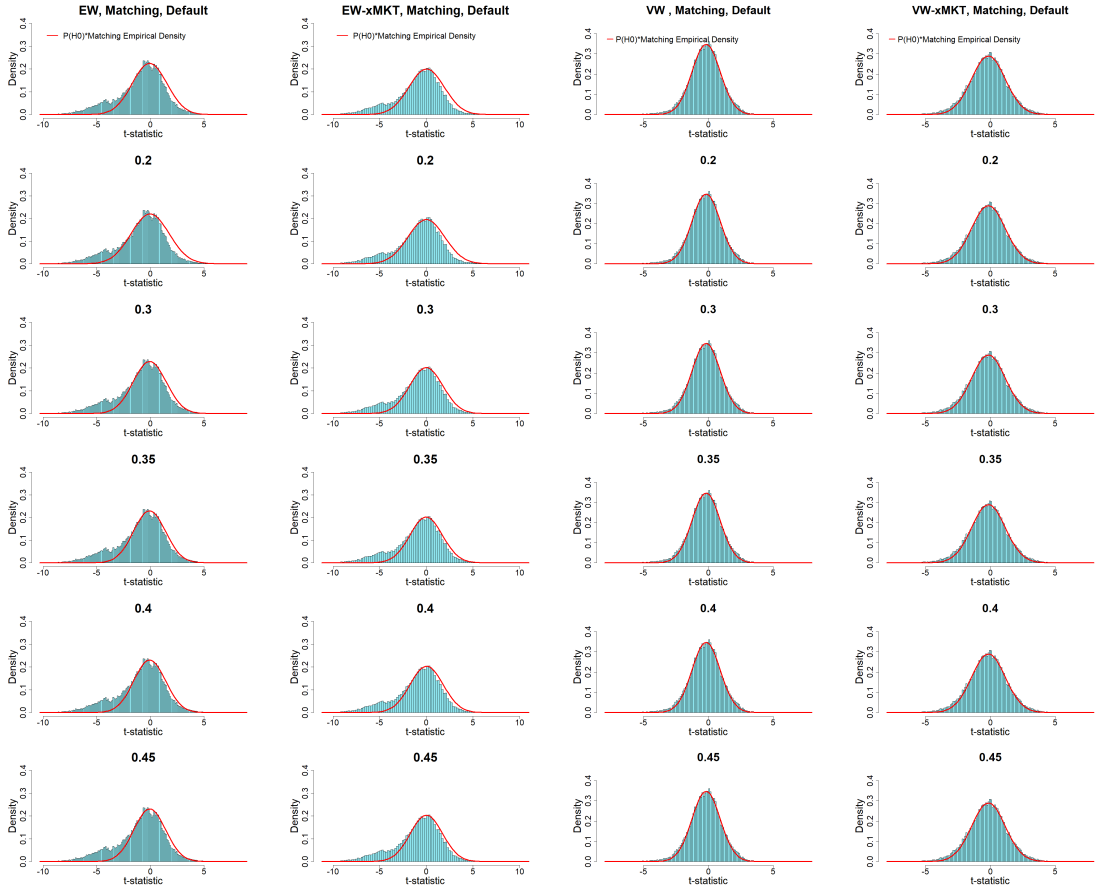
	EW						EW–xMKT					
	Default	0.2	0.3	0.35	0.4	0.45	Default	0.2	0.3	0.35	0.4	0.45
$\mathbb{P}(H_1)$	0.09	0.02	0.12	0.14	0.14	0.15	0.05	0.03	0.10	0.12	0.10	0.12
μ	-0.08	-0.02	-0.07	-0.08	-0.06	-0.06	0.03	0	-0.01	-0.01	0.03	0.01
σ	1.62	1.77	1.54	1.49	1.48	1.47	1.89	1.97	1.78	1.73	1.77	1.72
Cutoff (left)	-4.70	-5.38	-4.33	-4.10	-4.04	-4.02	-5.51	-5.96	-5.04	-4.82	-4.93	-4.75
Cutoff (right)	4.55	5.33	4.20	3.94	3.91	3.91	5.57	5.95	5.03	4.80	4.99	4.78
N rejected	2308	1416	2955	3445	3548	3580	1919	1454	2499	2856	2650	2957
	VW						VW–xMKT					
	Default	0.2	0.3	0.35	0.4	0.45	Default	0.2	0.3	0.35	0.4	0.45
$\mathbb{P}(H_1)$	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06
μ	-0.17	-0.17	-0.18	-0.18	-0.18	-0.18	-0.15	-0.16	-0.15	-0.15	-0.15	-0.15
σ	1.09	1.10	1.09	1.09	1.09	1.09	1.31	1.31	1.30	1.30	1.30	1.30
Cutoff (left)	-4.08	-4.10	-4.05	-4.05	-4.04	-4.04	-5.06	-5.14	-5.02	-4.99	-4.99	-4.96
Cutoff (right)	3.73	3.75	3.70	3.70	3.69	3.69	4.76	4.83	4.71	4.68	4.68	4.66
N rejected	205	196	210	210	212	212	96	81	104	108	108	112

To appreciate the results in Table S.1, note that changing the value of s produces different numbers of rejections but also different approximations, which can be visually inspected. Figure S.1 shows the histogram plots of the t-statistics together with the estimated null density component, i.e., a normal density $\mathcal{N}(\mu, \sigma^2)$ multiplied by $\mathbb{P}(H_0)$.

For both EW and EW–xMKT, it is clear that the fit of the null component is not optimal due to a strong asymmetry around zero.[43] Conversely, visual inspection of the plots for VW and VW–xMKT shows a good fit across the choice of intervals. For example, assuming that all the t-statistics in an interval as small as $[q_{0.45}, q_{0.55}]$ or

[43] We could, of course, choose $[q_s, q_r]$ where r is smaller than $1 - s$, so that the density is more shifted to the left to improve the approximation, but we avoid doing so to keep the discussion as simple as possible.

Figure S.1: Density Plots of t-statistics for Factor Returns. The histogram of the t-statistics for the 29,314 factors is plotted. The solid red line shows the approximation to the center of the density obtained by multiplying the estimated null density by $\mathbb{P}(H0)$, where all the quantities are estimated by the Matching method. Each row represents a different choice of intervals $[q_s, q_{1-s}]$ with $s = 0.2, 0.3, 0.35, 0.4, 0.45$. The default interval is as described in Section 3.2. Columns from left to right report results for EW, EW–xMKT, VW, and VW–xMKT.



slightly larger, such as $[q_{0.40}, q_{0.60}]$, are null is highly likely.[44] The results in Table S.1 confirm that this produces little variability. We conclude that the asymmetry in the histogram of the t-statistics for EW and EW–xMKT may require an asymmetric choice of quantiles, whereas for VW and VW–xMKT, we can regard the results in

[44]These are tight intervals around zero that contain only 10% and 20% of the t-statistics, respectively.

Table 2 as trustworthy.

We carry out a similar analysis for the MLE method. The default approach defines the central interval as $[\hat{\mu} \pm b\hat{\sigma}]$, where $\hat{\mu}$ and $\hat{\sigma}$ are intermediate parameter estimates of the null density $\mathcal{N}(\mu, \sigma^2)$, $b = 4.3 \exp\{-0.26 \log_{10}(N)\}$, and N denotes the number of t-statistics (see Section A.1.2 in the Appendix for more details). As alternatives, we consider $[x_l, x_u] = [q_{0.50} - z, q_{0.50} + z]$ for $z = 1.1, 1.2, 1.3, 1.4, 1.5$ – recall that $q_{0.50}$ is the median. Table S.2 presents the results.

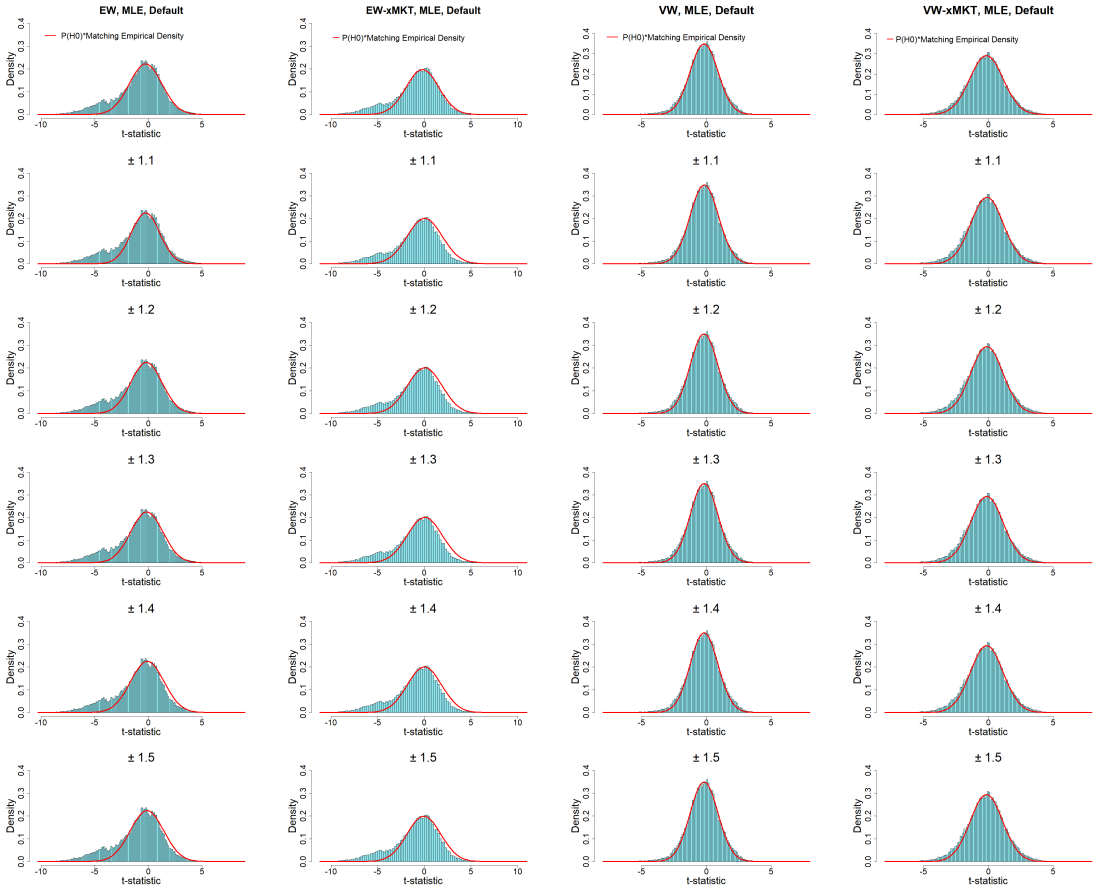
Table S.2: Cutoff Points Using the MLE Procedure with Different Choices of $[x_l, x_u]$. Detailed results are reported for EW and EW–xMKT in the upper panel, and for VW and VW–xMKT in the lower panel. For each dataset, the columns labeled ±1.1-±1.5 correspond to the results based on $[q_{0.50} - z, q_{0.50} + z]$ for $z = 1.1, 1.2, 1.3, 1.4, 1.5$. The "Default" column reports the baseline results presented in Table 2.

	EW						EW–xMKT					
	Default	±1.1	±1.2	±1.3	±1.4	±1.5	Default	±1.1	±1.2	±1.3	±1.4	±1.5
$\mathbb{P}(H_1)$	0.15	0.25	0.18	0.15	0.12	0.13	0.13	0.05	0.06	0.06	0.08	0.09
μ	-0.26	-0.27	-0.19	-0.15	-0.1	-0.13	-0.20	0.04	0.05	0.05	0.01	-0.05
σ	1.52	1.33	1.46	1.51	1.56	1.56	1.75	1.88	1.87	1.87	1.84	1.81
Cutoff (left)	-4.48	-3.68	-4.13	-4.26	-4.44	-4.48	-5.16	-5.45	-5.36	-5.39	-5.29	-5.22
Cutoff (right)	3.97	3.15	3.75	3.96	4.23	4.22	4.75	5.53	5.45	5.48	5.31	5.13
N rejected	2754	4403	3441	3137	2759	2695	2393	1987	2078	2047	2151	2270
	VW						VW–xMKT					
	Default	±1.1	±1.2	±1.3	±1.4	±1.5	Default	±1.1	±1.2	±1.3	±1.4	±1.5
$\mathbb{P}(H_1)$	0.06	0.04	0.06	0.06	0.06	0.05	0.06	0.07	0.08	0.07	0.07	0.08
μ	-0.18	-0.16	-0.18	-0.18	-0.18	-0.18	-0.15	-0.11	-0.12	-0.12	-0.13	-0.15
σ	1.08	1.1	1.07	1.07	1.07	1.08	1.29	1.25	1.25	1.26	1.26	1.25
Cutoff (left)	-4.00	-4.1	-3.97	-3.94	-3.94	-4.01	-4.97	-4.67	-4.65	-4.72	-4.71	-4.69
Cutoff (right)	3.64	3.78	3.61	3.58	3.59	3.66	4.66	4.45	4.40	4.47	4.46	4.40
N rejected	219	195	227	238	237	217	111	151	154	148	149	149

To understand the fit based on the choice of interval, we repeat the previous analysis and plot the histograms together with the estimated null component in Figure S.2.

For EW, both the default setting and the interval $[q_{0.50} - 1.1, , q_{0.50} + 1.1]$ indicate that the null component lies within the histogram and provide a good approximation of the body of the distribution. However, the two intervals produce very different standard deviations for the null density, which results in different numbers of rejections. The situation is clearer for VW and VW–xMKT. For these datasets, the fit appears to be consistently good across the choice of intervals, thus requiring less judgment when choosing one. This is reflected in the results in Table S.2, which are consistent across intervals.

Figure S.2: Density Plots of t-statistics for Factor Returns. The histogram of the t-statistics for the 29,314 factors is plotted. The solid red line shows the approximation to the center of the density obtained by multiplying the estimated null density by $\mathbb{P}(H0)$, where all the quantities are estimated by the MLE method. Each row represents a different choice of intervals $[q_{0.5} - z, q_{0.5} + z]$ for $z = 1.1, 1.2, 1.3, 1.4, 1.5$. The default interval is as described in Section 3.2. Columns from left to right report results for EW, EW–xMKT, VW, and VW–xMKT.



In summary, we find that the EW and EW–xMKT datasets are more difficult to analyze and require judgment in the choice of interval. Rather than attempting to over-engineer the way we choose the interval for these datasets, our goal is to highlight this difficulty. As argued in the paper, factor discovery for equally weighted portfolios should be regarded as dubious due to a small-cap bias, which results in many stocks not being actually tradable at the published prices.

S.4 Simulations of JKP Method

We use a simulation to demonstrate that JKP (2023)'s method does not control the False Discovery Rate (FDR) unless their simulation design is used - a design that we will argue is unrealistic. For comparison, we use ordinary least squares (OLS) unadjusted p-values, as well as the Benjamini–Hochberg (BH) procedure. The objective is to control the FDR at the 5% level.

To ensure consistency with the JKP setup, we consider one sided hypotheses: H_0^i : $\alpha_i = 0$ against $H_1^i : \alpha_i > 0$. Using the notation in (2), the number of true discoveries is defined as $N_{\text{TD}} := \sum_{i=1}^{N} \delta_i A_i$. The data-driven rejection rule δ_i for hypothesis i depends on the method used for multiple testing control. For JKP, the decision rule is defined in Section A.2, with priors discussed in Section S.4.2.1. For OLS, we set $\delta_i = 1$ if the p-value is below 0.05. For BH, we apply the Benjamini–Hochberg procedure to adjust the p-values for multiple testing. In both the OLS and BH procedures, p-values are computed using the test statistic T (not $|T|$), as we consider one-sided alternatives.

The number of false discoveries is defined as $N_{\text{FD}} := N_D - N_{\text{TD}}$, where $N_D = \sum_{i=1}^{N} \delta_i$ is the total number of discoveries. Hence, the False Discovery Proportion (FDP) is given by $\text{FDP} = N_{\text{FD}}/N_D$, and the False Discovery Rate is defined as $\text{FDR} = \mathbb{E}[\text{FDP}]$. The expectation is approximated using Monte Carlo methods over 10,000 simulations.

S.4.1 General simulation design

For all designs, we simulate n independent and identically distributed $N \times 1$ vectors $f_t = \alpha + \varepsilon_t$. The vectors f_t, α, and ε_t represent the factors, the alphas, and the noise terms, respectively, as in (1) with $\beta = 0$. The designs will differ according to how α is constructed. Here, we outline the elements common to all designs. The noise term ε_t is a Gaussian random vector with mean zero and homoskedastic variance $Var(\varepsilon_{t,i}) = \sigma_\varepsilon^2$, for $i = 1, 2, \ldots, N$.

The correlation matrix of ε_t has a cluster-block structure: factors within each cluster have correlation ρ_w, while correlations between clusters are ρ_b. Following JKP, we set $\rho_w = 0.55$ and $\rho_b = 0.3$. There are 13 clusters, each containing 10 distinct factors (see JKP, Eq. 23 and Section C.4 for details). JKP's method is computed under the assumption that ρ_w and ρ_b are known.

We set the sample size to $n = 840$ and the number of factors to $N = 130$, which results in a factor volatility of 10% per annum. Assuming monthly data, this implies a monthly volatility of $\sigma_\varepsilon = 0.1/\sqrt{12}$. The simulation therefore spans 70

years ($12 \times 70 = n$). This is exactly the same setup as in JKP.

For convenience, we introduce the following notation. Let m_0 denote the number of true null hypotheses (i.e., non-positive alphas), and let m_1 denote the number of true alternatives (i.e., strictly positive alphas). Then $N = m_0 + m_1$, and the proportion of true null hypotheses is defined as $\mathbb{P}(H_0) = m_0/N$. We reproduce the results in JKP using their exact design and contrast them with different designs that vary in how the vector α is constructed. The details of the designs are described next.

S.4.2 Designs and results

Each design corresponds to a different construction of α. For each design, we briefly describe the setup.

S.4.2.1 JKP design

This is the design used by JKP: we use the same parameters to replicate their results and compare them to other designs.[45] Following Section C.4 of JKP, we generate α using a hierarchical model, treating it as a random variable that varies across the 10,000 simulations.

Given a factor i belonging to cluster $j \in \{1, 2, \ldots, 13\}$, its alpha is decomposed as:

$$\alpha_i = c^j + w^i, \tag{S.2}$$

where c^j and w^i are independent random variables representing the cluster-specific and idiosyncratic alpha components, respectively. Both are normally distributed with mean zero and variances τ_c^2 and τ_w^2. We report results for $\tau_c \in \{0.01, 0.5\}$ to capture different levels of cluster-wide alpha variation, and for $\tau_w \in \{0.01, 0.2\}$ to represent low and high idiosyncratic alpha variability. These are the same values considered by JKP, who also report results for additional values of τ_c between 0.01 and 0.5.

In this setup, approximately 50% of the hypotheses are null, i.e., $m_0/N \simeq 1/2$, though never exactly equal to 1/2, since the alphas are randomly generated with equal probability of being positive or negative. This is equivalent to stating that, on average, half of the hypotheses are non-null.

[45] We also use their code, with slight modifications to allow the inclusion of additional designs. We are grateful to JKP for proving their code.

To implement JKP's procedure, we need to specify the priors for c^j and w^i and to have estimate of the covariance matrix of the error terms ε_t.[46] As in JKP, we assume that the priors and their hyperparameters match those used in the data-generating process, and that the covariance matrix of the errors – described in Section S.4.1 – is known. For this design, JKP's method successfully controls the FDR. Interestingly, OLS also controls the FDR, except when $\tau_c = 0.1$ and $\tau_w = 0.01$, that is, when the overall variance of the alphas is small and all alphas are close to zero. Detailed results are reported in Table S.3, which replicates the results from JKP.

S.4.2.2 Strong alphas design

In this design, we set $\alpha_i = 3\sigma_\varepsilon/\sqrt{n}$ for $i \in \{1, 2, \ldots, m_1\}$ and $\alpha_i = 0$ otherwise. This implies that nonzero alphas are three times the standard error of their OLS estimates, making them difficult to confuse with noise. For this design, $\mathbb{P}(H_0) = m_0/N$ is fixed across simulations. In particular, we consider $\mathbb{P}(H_0) \in \{0.5, 0.75, 0.9\}$.

To implement the JKP's method, we use the same priors as specified in Section S.4.2.1. In this case, however, the alphas are not generated according to JKP's priors. This allows us to examine how the choice of hyperparameters affects performance when the data-generating process is more realistic, permitting fewer than 50% true alternatives. We continue to assume that the covariance matrix of the errors is known.

From the results we observe that both OLS and JKP perform worse when $\mathbb{P}(H_0) = 0.9$ compared to 0.75 and 0.5. A large number of true nulls increases the difficulty of identifying true signals, and both methods perform poorly. The results also show that JKP's method is particularly sensitive to the choice of τ_c, which controls the within-cluster variability of alphas. Detailed results are provided in Table S.4.

S.4.2.3 Decaying alphas design

We choose $\alpha_i = 2.6 c_i \sigma_\varepsilon/\sqrt{n}$, where $c_i = -\ln\left(1 - \frac{i}{m_1+1}\right)$ for $i \in \{1, 2, \ldots, m_1\}$, and zero otherwise. Hence, α_i is equal to $2.6c_i$ times the standard error of the alpha estimator. This choice of c_i implies that the non-zero alphas take different values, ranging from relatively small to relatively large. For example, when $m_1 = 10$, we have:

$$c_i \in \{0.095, 0.201, 0.318, 0.452, 0.606, 0.788, 1.012, 1.299, 1.705, 2.398\}$$

[46] The covariance matrix is needed to derive the posterior probability of the alphas.

This means that some alphas under the alternative can be difficult to distinguish from noise when c_i is small. All other design elements follow those in Section S.4.2.2. Discovery under this design is more challenging because the alphas are not as well separated as in the strong alphas design. Results show further deterioration in FDR control for both OLS and JKP, relative to the strong alphas design. Once again, the JKP's method appears sensitive to the choice of hyperparameters and can fail to control the FDR even when half of the hypotheses are true alternatives. Detailed results are reported in Table S.5.

S.4.3 Discussion

The results suggest that JKP's method is effectively equivalent to using unadjusted OLS p-values, although low values of τ_c appear to make it more conservative – but only in finite samples. This behavior is expected, as JKP's method conflates the posterior probability of the null with the posterior probability of the alphas (see Section A.2 for the derivation). It is well known that, as the sample size n increases, the posterior estimator of a statistic converges to the maximum likelihood estimator, provided the prior is suitably specified (e.g., see Equations (16) and (17) in JKP). This also implies that Proposition 5 in JKP holds only under specific design settings, contrary to the generality claimed.

Our simulations show that JKP's method can fail to control the FDR under the decaying alpha design, even when half of the hypotheses are non-null—except for special values of the hyperparameters (Table S.5) that are difficult to choose in practice. In fact, choosing an inappropriate set of hyperparameters can cause JKP's method to perform worse than OLS. As shown in Section A.2, having 50% non-nulls is a necessary but not sufficient condition for FDR control in theory. A higher value of $\mathbb{P}(H_0)$ renders JKP's conclusions wrong for both the strong and decaying alpha designs.

In summary, when $\mathbb{P}(H_0) = 0.5$, there may be less need to control for multiple testing, and in such cases, even unadjusted OLS p-values may provide some FDR control. However, as discussed in Section 3.2 and even in Chen (2025), $\mathbb{P}(H_0)$ is likely to be much higher than 0.5. This suggests that the assumptions underlying JKP's original design are highly specific and based on beliefs not widely shared by other researchers. Any value of $\mathbb{P}(H_0)$ above 0.5 implies that thresholds must be increased.

S.5 Factor Duplication

A new factor with an alpha t-statistic exceeding a desired threshold may still not represent a genuinely new discovery, as it can be a duplication or noisy combination of existing factors. However, identifying such duplication is challenging. The main difficulty lies in defining the set of existing true discoveries against which to compare the new factor. Importantly, the rationale for duplication checks is to determine whether the newly discovered factor exhibits zero alpha after accounting for existing factors.[47] These considerations motivate a pragmatic approach to duplication assessment. Given that true discoveries are inherently unobservable, we could define a set of factors that likely contains the true factors and refer to them as pseudo-true factors. This set may include widely discovered factors such as those from Fama and French (2015), as well as the discoveries documented in Sections 3.2 and 3.3 – for example, the 111 factors that pass the cutoffs reported in Table 2. Given that the market excess return is treated as a factor, we compute the new factor and the pseudo-true factors ex-market. In what follows, we take the pseudo-true factors as given.[48]

Many of the pseudo-true factors may show no significant correlation with the new factor and can therefore be excluded. This filtering can be implemented by computing pairwise correlations between the new factor and each pseudo-true factor, and testing the null hypothesis of zero correlation. Because the new factor is tested for correlation with each pseudo-true factor, the procedure requires adjustment for multiple testing. Only pseudo-true factors that are statistically significantly correlated with the new factor will be retained for further analysis. The number of retained factors is expected to be much smaller than the number in the original set.

Next, we can apply a spanning regression of the new factor on an intercept and the selected pseudo-true factors. We assume that the factor is signed to have a positive alpha. Hence, to improve the power of the test, we conduct a one-sided test for a significantly positive intercept (alpha). If the t-statistic of the intercept is not greater than 1.645 (for a one-sided test at the 5% level of significance), we do not reject the null hypothesis that the new factor is a duplicate. As we focus on a single new factor, no adjustment for multiple testing is required.

In summary, a new factor can be regarded as a genuine discovery if its alpha is significantly positive in the spanning regression. Note that, in the population, the

[47] In that case, the factor's expected return equals its beta exposure to the existing factors, implying it is priced by them.

[48] Some of these pseudo-true factors are likely to be highly correlated and could be pooled. For the moment we ignore this, but remark on this at the end of the section.

regression of the new factor on a constant and all the pseudo-true factors yields a zero beta for any pseudo-true factor that is uncorrelated with the new factor. This justifies using correlations to first eliminate uncorrelated pseudo-true factors before performing the spanning regression.[49]

Pseudo-true factors defined from discoveries are often highly correlated, and it may be desirable to pool together those that are strongly related. For example, we find that the 111 factor discoveries based on the MLE cutoffs for VW–xMKT in Section 3.2 can be grouped into six clusters using hierarchical clustering, as in JKP (2023).[50] In this case, the pseudo-true factors within each cluster can be replaced by the within-cluster average. The exact methodology for pooling together highly correlated pseudo-true factors is beyond the scope of this paper, but it may involve hierarchical clustering (JKP, 2023) or correlation-based heuristics.[51] Chib et al. (2025) examine the issue of factor duplication using a Bayesian procedure that controls the expected false discovery rate in spanning tests. They consider the market factor together with the Fama-French six factors as pseudo-true factors, and, as a robustness check, replace the Fama-French factors with an alternative set of six factors – profitability, investment, liquidity, betting-against-beta, and alternative value factors.

References

Chib, S., S., Xiao and L. Zhao (2025) Factors or Fake? A New Look at Anomalies and the Replication Crisis. *Working Paper*.

Efron, B., R. Tibshirani, J.D. Storey and V. Tusher (2001) Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association* 96, 1151-1160.

Fama, E.F. and K.R. French (2015). A Five-Factor Asset Pricing Model. *Journal of Financial Economics* 116, 1-22.

[49]Such a two-step methodology is common in econometrics and statistics (Fan and Lv, 2008; Chudik et al., 2018).

[50]We selected the number of clusters by maximizing the average silhouette score—a data-driven method that evaluates how well each data point fits within its cluster relative to others. López de Prado and Lewis (2019) recommend this approach for investment strategies, whereas JKP (2023) uses visual inspection.

[51]Most procedures rely on estimators of the covariance matrix. One should exercise caution when estimating high-dimensional covariance matrices, as robust estimation is only feasible under certain sparsity assumptions on the covariance matrix or its inverse (e.g., Meinshausen and Bühlmann, 2006; Bickel and Levina, 2008).

Genovese, C. and L. Wasserman (2004) A Stochastic Process Approach to False Discovery Control. *Annals of Statistics* 32, 1035-1061.

Jensen, T.I., B. Kelly and L.H. Pedersen (2023) Is There a Replication Crisis in Finance? *The Journal of Finance* 78, 2465-2518.

Shwartzman, A. and X. Lin (2011) The Effect of Correlation in False Discovery Rate Estimation. *Biometrika* 98, 199-214.

Table S.3: Simulation Results for JKP Design. The dataset is generated under the JKP design, with alphas defined by equation S.2, using various combinations of $\tau_c = \{0.01, 0.5\}$ and $\tau_w = \{0.01, 0.2\}$. The priors for JKP's method match those of the data-generating process, using the same values of τ_c and τ_w. The OLS and BH methods are used as benchmarks. The results report the false discovery rate (FDR), the true positive rate (TPR), the number of false discoveries (FP), and the number of true discoveries (TP). An FDR value below the nominal control level of 0.05 indicates that the procedure successfully controls the FDR, while a value above suggests that FDR is not controlled. Ideally, the FDR would be exactly equal to the nominal level of 0.05.

τ_c	τ_w	Method	FDR	TPR	FP	TP
0.01	0.01	JKP	0	0	0	0
		OLS	0.36	0.06	2.59	4.17
		BH	0.01	0	0.03	0.07
0.5	0.01	JKP	0	0.80	0.16	52.02
		OLS	0.01	0.75	0.22	48.60
		BH	0	0.68	0.06	44.16
0.01	0.2	JKP	0.01	0.48	0.32	31.52
		OLS	0.02	0.46	0.52	29.68
		BH	0	0.28	0.08	17.91
0.5	0.2	JKP	0	0.76	0.19	49.60
		OLS	0	0.76	0.20	49.66
		BH	0	0.70	0.06	45.67

Table S.4: Simulation Results for the Strong Alphas Design. The dataset is generated under the strong alphas design for $\mathbb{P}(H_0) = 0.9$ (13 true factors), 0.75 (32 true factors) and 0.5 (65 true factors), where the non-zero alphas are set to $3(0.1/\sqrt{12})/\sqrt{840}$. The JKP method is implemented under the hierarchical structure specified in equation (S.2), assuming Gaussian priors with mean zero and variances $\tau_c \in \{0.01, 0.5\}$ and $\tau_w \in \{0.01, 0.2\}$. The OLS and BH methods are used as benchmarks. The results report the false discovery rate (FDR), the true positive rate (TPR), the number of false discoveries (FP), and the number of true discoveries (TP). An FDR value below the nominal control level of 0.05 indicates that the procedure successfully controls the FDR, while a value above suggests that FDR is not controlled. Ideally, the FDR would be exactly equal to the nominal level of 0.05.

$\mathbb{P}(H_0)$	τ_c	τ_w	Method	FDR	TPR	FP	TP
0.9	0.01	0.01	JKP	0	0	0	0
	0.5	0.01	JKP	0.26	0.86	6.51	10.27
	0.01	0.2	JKP	0.27	0.55	2.53	6.63
	0.5	0.2	JKP	0.29	0.91	5.39	10.98
			OLS	0.31	0.91	5.93	10.95
			BH	0.04	0.58	0.47	7.02
0.75	0.01	0.01	JKP	0	0	0	0
	0.5	0.01	JKP	0.12	0.94	5.45	29.96
	0.01	0.2	JKP	0.10	0.47	1.70	14.97
	0.5	0.2	JKP	0.12	0.92	4.43	29.33
			OLS	0.13	0.91	4.93	29.17
			BH	0.04	0.73	1.07	23.39
0.5	0.01	0.01	JKP	0	0	0	0
	0.5	0.01	JKP	0.07	0.96	5.53	62.12
	0.01	0.2	JKP	0.03	0.41	0.77	26.90
	0.5	0.2	JKP	0.04	0.92	2.90	59.49
			OLS	0.05	0.91	3.28	59.26
			BH	0.02	0.83	1.50	53.89

Table S.5: Simulation Results for the Decaying Alphas Design. The dataset is generated under the decaying alphas design for $\mathbb{P}(H_0) = 0.9$ (13 true factors), 0.75 (32 true factors), and 0.5 (65 true factors), where the nonzero alpha for factor i is set as $\alpha_i = 2.6\, c_i\, \sigma_\varepsilon/\sqrt{n}$, with $c_i = -\ln\left(1 - \frac{i}{m_1+1}\right)$ for $i \in \{1, 2, \ldots, m_1\}$. The JKP method is implemented under the hierarchical structure specified in equation (S.2), assuming Gaussian priors with mean zero and variances $\tau_c \in \{0.01, 0.5\}$ and $\tau_w \in \{0.01, 0.2\}$. The OLS and BH methods are used as benchmarks. The results report the false discovery rate (FDR), the true positive rate (TPR), the number of false discoveries (FP), and the number of true discoveries (TP). An FDR value below the nominal control level of 0.05 indicates that the procedure successfully controls the FDR, while a value above suggests that FDR is not controlled. Ideally, the FDR would be exactly equal to the nominal level of 0.05.

$\mathbb{P}(H_0)$	τ_c	τ_w	Method	FDR	TPR	FP	TP
0.9	0.01	0.01	JKP	0	0	0	0
	0.5	0.01	JKP	0.40	0.65	8.34	7.79
	0.01	0.2	JKP	0.31	0.42	2.56	5.02
	0.5	0.2	JKP	0.38	0.57	5.44	6.83
			OLS	0.41	0.57	5.93	6.79
			BH	0.04	0.34	0.30	4.13
0.75	0.01	0.01	JKP	0	0	0	0.02
	0.5	0.01	JKP	0.28	0.62	9.38	19.81
	0.01	0.2	JKP	0.14	0.32	1.76	10.29
	0.5	0.2	JKP	0.18	0.56	4.53	17.99
			OLS	0.20	0.56	4.93	17.97
			BH	0.04	0.40	0.62	12.79
0.5	0.01	0.01	JKP	0	0	0	0
	0.5	0.01	JKP	0.15	0.63	7.69	40.93
	0.01	0.2	JKP	0.04	0.31	0.87	20.19
	0.5	0.2	JKP	0.07	0.56	3.07	36.44
			OLS	0.08	0.56	3.28	36.61
			BH	0.02	0.45	0.84	29.12