NBER WORKING PAPER SERIES

IS THERE A REPLICATION CRISIS IN FINANCE?

Theis Ingerslev Jensen
Bryan T. Kelly
Lasse Heje Pedersen

Working Paper 28432
http://www.nber.org/papers/w28432

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2021

We are grateful for helpful comments from Nick Barberis, Andrea Frazzini, Cam Harvey, Antti Ilmanen, Ronen Israel, John Liew, Toby Moskowitz, Stefan Nagel, Scott Richardson, Anders Rønn-Nielsen, and seminar and conference participants at AQR, Georgetown Virtual Fintech Seminar, Tisvildeleje Summer Workshop 2020, Yale, and 2020 CFA Institute European Investment Conference. We thank Tyler Gwinn for excellent research assistance. Jensen and Pedersen gratefully acknowledge support from the FRIC Center for Financial Frictions (grant no. DNRF102). AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w28432.ack

Is There A Replication Crisis In Finance?
Theis Ingerslev Jensen, Bryan T. Kelly, and Lasse Heje Pedersen
NBER Working Paper No. 28432
February 2021
JEL No. C11,C58,G02,G10,G11,G12,G15,G17

ABSTRACT

Several papers argue that financial economics faces a replication crisis because the majority of studies cannot be replicated or are the result of multiple testing of too many factors. We develop and estimate a Bayesian model of factor replication, which leads to different conclusions. The majority of asset pricing factors: (1) can be replicated, (2) can be clustered into 13 themes, the majority of which are significant parts of the tangency portfolio, (3) work out-of-sample in a new large data set covering 93 countries, and (4) have evidence that is strengthened (not weakened) by the large number of observed factors.

Theis Ingerslev Jensen
Copenhagen Business School
Department of Finance
Solbjerg Plads 3, A5
DK-2000 Frederiksberg
Denmark
tij.fi@cbs.dk

Lasse Heje Pedersen
Copenhagen Business School
Solbjerg Plads 3, A5
DK-2000 Frederiksberg
DENMARK
and NYU
Lhp001@gmail.com

Bryan T. Kelly
Yale School of Management
165 Whitney Ave.
New Haven, CT 06511
and NBER
bryan.kelly@yale.edu

Code and data is available at bryankellyacademic.org
Code repository is available at https://github.com/bkelly-lab/GlobalFactor
Online appendix is available at http://www.nber.org/data-appendix/w28432

Several research fields face replication crises (or credibility crises), including medicine (Ioannidis, 2005), psychology (Nosek et al., 2012), management (Bettis, 2012), experimental economics (Maniadis et al., 2017), and now also financial economics. Challenges to the replicability of finance research take two basic forms:

1. **No internal validity.** Most studies cannot be replicated with the same data (e.g., because of coding errors or faulty statistics) or are not robust in the sense that the main results cannot be replicated using slightly different methodologies and/or slightly different data.[1] E.g., Hou et al. (2020) state:

   "Most anomalies fail to hold up to currently acceptable standards for empirical finance"

2. **No external validity.** Most studies may be robustly replicated, but are spurious and driven by "p-hacking," that is, finding significant results by testing multiple hypotheses without controlling the false discovery rate. Such spurious results are not expected to replicate in other samples or time periods, in part because the sheer number of factors is simply too large, and too fast growing, to be believable. E.g., Cochrane (2011) asks for a consolidation of the "factor zoo," and Harvey et al. (2016) states:

   "most claimed research findings in financial economics are likely false."[2]

We examine both of these challenges theoretically and empirically. We conclude that neither criticism is tenable and that the collective body of factor research is both internally and externally valid.

We analyze replicability of the main finance factors using a Bayesian model and a new global data set of 153 factors across 93 countries. To help advance replication in finance,

[1] Hamermesh (2007) contrasts "pure replication" and "scientific replication." Pure replication is, "checking on others' published papers using their data," also called "reproduction" by Welch (2019). Scientific replication uses, "different sample, different population and perhaps similar, but not identical model." We focus on scientific replication. We propose a new modeling framework to jointly estimate factor alphas, we use robust factor construction methods that are applied uniformly to all factors, and we test both internal and external validity of prior factor research in several dimensions, including out-of-sample time series replication and international sample replication. In complementary and contemporaneous work, Chen and Zimmermann (2020) consider pure replication, attempting to use the same data and methods as the original papers for a large number of factors. They are able to reproduce nearly 100% of factors, but Hou et al. (2020) challenge the scientific replication and Harvey et al. (2016) challenge validity due to multiple testing.

[2] Similarly, Linnainmaa and Roberts (2018) state "the majority of accounting-based return anomalies, including investment, are most likely an artifact of data snooping."

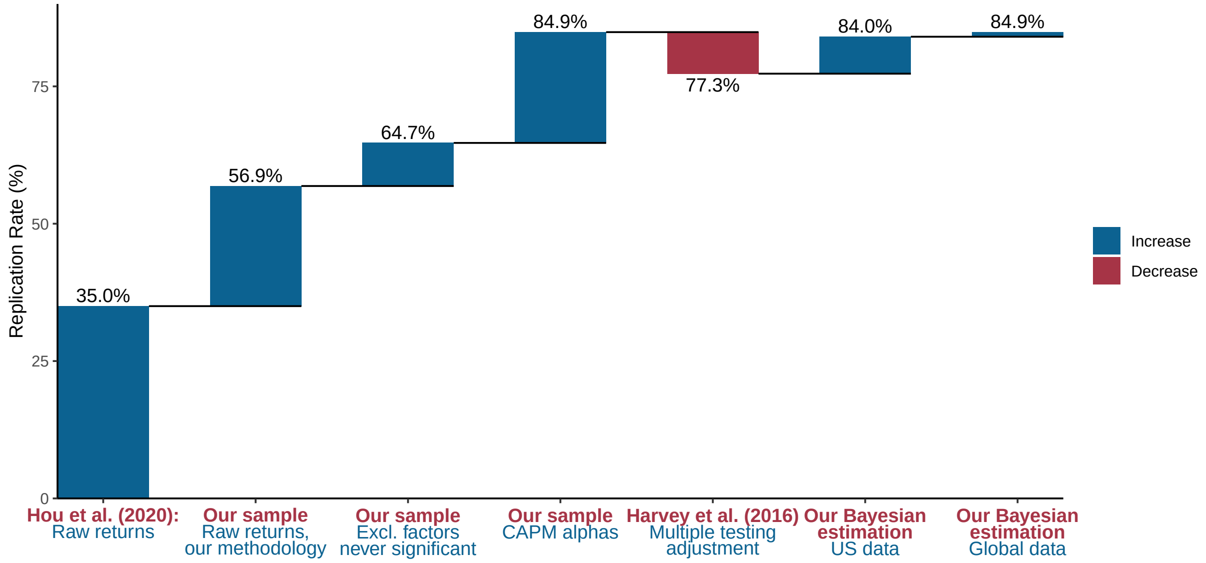


Figure 1: Replication Rates Versus the Literature

Note: This figure summarizes analyses throughout the paper. Refer to Section 3 for estimation details.

we have made this data set easily accessible to researchers via a direct open-source link to WRDS. We also include meticulous documentation of the data set and the underlying code base to reproduce it.

Our findings challenge the dire view of finance research. We find that the majority of factors do replicate, do survive joint modeling of all factors, do hold up out-of-sample, are strengthened (not weakened) by the large number of observed factors, are further strengthened by global evidence, and the number of factors can be understood as multiple versions of a smaller number of themes. At the same time, a non-trivial minority of factors fail to replicate in our data, but the overall evidence is much less disastrous than some people suggest. Further, we show that factors must be understood in light of economic theory and we develop a Bayesian model that offers a very different interpretation of the evidence on factor replication.

Figure 1 illustrates our main results and how they relate to the literature in a sequence of steps. It presents the "replication rate," that is, the percent of factors with a statistically significant average excess return. Our paper builds on the extraordinarily expansive and thorough factor replication study of Hou et al. (2020). The starting point of Figure 1—

shown as the first bar on the left—is the 35% replication reported by Hou et al. (2020) in their universe of 452 factors.

Differences in Sample and Factor Construction

The second bar in Figure 1 shows the replication rate in our main sample of US factors. It is based on significant OLS t-statistics for average raw factor returns, in direct comparability to the 35% calculation from Hou et al. (2020). While our factor data construction has minor differences versus Hou et al. (2020), we find a baseline replication rate of 56.9%, a difference of 21.9 percentage points.

This difference has the following decomposition. First, we use a longer sample, which contributes +4.3% to the difference in replication rate. Second, for each characteristic, Hou et al. (2020) construct three variations on each factor having either 1-month, 6-month, or 12-month holding periods. They treat these as separate factors so that their factor count essentially multiplies their characteristics count by a factor of three. In contrast, we focus on 1-month returns because this is the horizon of interest in almost all of the original studies (and we believe it is the most economically meaningful since it uses the most current data as theory dictates). Our focus on only the 1-month holding period factor for each characteristic contributes +4.0% to our replication rate.

Next, Hou et al. (2020) focus their analysis on value-weighted factors rather than the standard Fama and French (1993) methodology that gives half the weight to small stocks (or equal-weighting that gives even more weight to small stocks). However, pure value weighting sometimes leads to excessively concentrated portfolios that mask the behavior of factors.[3] We use a weighting scheme that we refer to as "capped value-weighting" that winsorizes market caps at the NYSE 80^{th} percentile. This weighting is a helpful compromise between pure value-weighting and the Fama-French method since our factors continue to emphasize large stocks, but the capped scheme avoids undue skewness toward a few mega stocks, which in turn produces more robust factor behavior over time and across countries. Capped value weights contribute +8.5% to our higher replication rate.[4]

[3] For example, Nokia stock accounted for more than 70% of the total market capitalization in Finland in 1999 and 2000.

[4] In Figure C.1 of the appendix, we show an alternative version of Figure 1 with factors constructed using

We add 15 factors to our sample that were previously studied in the literature but not studied by Hou et al. (2020), which contributes +2.4%. The remaining +2.7% difference in replication rates is due to minor (and conservative) factor construction details that we believe robustify factor behavior.[5] We discuss this decomposition further in Section 2, where we detail our factor construction choices and discuss why we prefer them.

The Hou et al. (2020) sample includes a number of factors that the original studies found to be insignificant.[6] We exclude these when calculating the replication rate. After we make this adjustment, the replication rate rises to 64.7%, shown in the third bar in Figure 1.

Alpha, Not Raw Return

Hou et al. (2020) analyze and test factors' raw returns. But if we wish to learn about "anomalies," economic theory dictates the use of risk-adjusted returns. Raw return gives a misleading inference for the factor if it differs from the alpha: When the raw return is significant, but the alpha is not, it simply means that the factor is taking risk exposure and the risk premium is significant, which does not indicate anomalous returns of the factor. Likewise, when the raw return is insignificant, but the alpha is significant, then the factor's efficacy is masked by its risk exposure. An example of this is the low-beta anomaly, where theory predicts that the alpha of a dollar-neutral low-beta factor is positive, but its raw return is negative or close to zero (Frazzini and Pedersen, 2014). In this case, the "failure to replicate" of Hou et al. (2020) is, in fact, support for the betting-against-beta theory.

We analyze alpha to the CAPM, which is the clearest theoretical benchmark model that is not mechanically linked to other so-called anomalies in the list of replicated factors. The fourth bar in Figure 1 shows that the replication rate rises to 84.9% based on tests of factors' CAPM alpha.

In their abstract, Hou et al. (2020) also emphasize a stunning 4% replication rate among factors that they group in a "trading frictions" category. For this same set of factors, we

straight (as opposed to capped) value weights. It shows that all of our conclusions are unaffected. Our ultimate replication rate in this case is 77.3% (based on global data and Bayesian model estimates).

[5] We use tercile spreads while they use deciles; we use tercile breakpoints from all stocks above the NYSE 20^{th} percentile (i.e., non-micro-caps), they use straight NYSE breakpoints; we always lag data four months, they use a mixture of updating schemes; we exclude IBES factor due to their relatively short history.

[6] We identify 34 factors from Hou et al. (2020) for which the original paper did not find a significant alpha or did not study factor returns (see appendix Table C.3).

find a replication rate of 63% based on CAPM alphas.

Multiple Testing and Bayesian Modeling

The first four bars in Figure 1 are based on individual ordinary least squares (OLS) t-statistics for each factor. But Harvey et al. (2016) rightly point out that this type of analysis suffers from a multiple testing (MT) problem. Harvey et al. (2016) recommend MT adjustments that raise the threshold for a t-statistic to be considered statistically significant. We report one such MT correction using a leading method proposed by Benjamini and Yekutieli (2001). Accounting for MT in this manner, we find that the replication rate drops to 77.3% (the fifth bar of Figure 1). For comparison, Hou et al. (2020) consider a similar adjustment and find that their replication rate drops from 35% with OLS to 18% after MT correction.

However, sticking with independent tests and adding an ex post p-value correction based on Bonferroni, family-wise error rate, or false discovery rate adjustments can be an unnecessarily crude solution to the multiple testing problem. This approach is best justified in environments where independence among subjects is a reasonable assumption, as is sometimes the case in biomedical research. If the data are dependent, on the other hand, independent testing with an MT correction fails to make efficient use the data.

Our handling of the MT problem is different. We propose a Bayesian framework and directly model the joint behavior of all the factors. There are two major benefits to our approach. First, we impose a prior that all alphas are expected to be zero. The role of the Bayesian prior is conceptually similar to that of frequentist MT corrections—it imposes conservatism on statistical inference and controls the false discovery rate.[7] As emphasized by Gelman et al. (2012), "the problem of multiple comparisons can disappear entirely when viewed from a hierarchical Bayesian perspective." We provide a detailed elaboration of this point in Section 1. In addition, Bayesian estimation produces a posterior distribution that describes all there is to know about alpha estimates. It provides a basis for p-values as in traditional hypothesis testing, but can tell us much more. For example, it can be used to calculate the probability of seeing a particular number of alphas larger than some threshold,

[7] A large statistics literature explains how Bayesian estimation naturally combats MT problems; see Gelman et al. (2013) and references therein.

or to calculate expected discovery rates of true and false positives (calculations we make later). The completeness with which the posterior describes model parameters is the basis of Harvey (2017)'s argument in favor of Bayesian approaches to factor evaluation.[8]

Second, we use a joint model of factors, which allows us to conduct inference for all factor alphas simultaneously. The joint structure among factors leverages dependence in the data in order to draw more informative statistical inferences (relative to conducting independent individual tests). Our zero-alpha prior shrinks alpha estimates of all factors, thereby raising p-values with similar conservatism as a frequentist MT correction. At the same, however, the model allows us to learn more about the alpha of any individual factor, borrowing estimation strength across all factors. This improves precision of alpha estimates for all factors, which lowers p-values all else equal. Which effect dominates when we construct our final Bayesian p-values—the conservative shrinkage to the prior or the improved precision of alphas—is an empirical question. In our sample, we find that the two effects almost exactly offset, which is why the Bayesian multiple testing view delivers a replication rate nearly identical to the OLS-based rate. The intuition behind this surprising result is simply that having many factors can be a strength rather than a weakness when assessing the replicability of factor research. For example, the better a factor has performed and the longer its time series, the stronger and tighter is our posterior, but our posterior is further tightened if similar factors have also performed well, and if additional data shows that these factors have performed well in many other countries.[9]

From our Bayesian approach to the MT problem, our estimated replication rate rises to 84.0% (the sixth bar of Figure 1). The results summarized in the first six bars of Figure 1 lead us to conclude that factor research, by and large, demonstrates internal validity.

[8] Our analysis differs from Harvey (2017) who focuses on MT adjustments via minimum Bayes factors, while we propose a complete Bayesian modeling and estimation scheme.

[9] Taking this intuition further, we can glean additional information from studying whether factors work in other asset classes, as has been done for value and momentum (Asness et al., 2013), betting against beta (Frazzini and Pedersen, 2014), time series momentum (Moskowitz et al., 2012), and carry (Koijen et al., 2018).

Global Replication

We investigate how our conclusions are affected when we extend the data to include all factors in a large global panel of 93 countries. The last bar in Figure 1, shows that, based on the global sample, the final replication rate rises slightly to 84.9%. This estimate is based on the Bayesian model extended to incorporate the joint behavior of international data. Because it accounts for the global correlation structure among factors, the model recognizes that international evidence is not independent out-of-sample evidence, and uses only the incremental global evidence to update the overall replicability assessment. And it continues to account for multiple testing. The global result reflects that factor performance in the US replicates well in an extensive cross section of countries. Serving as our final estimate, the global factor replication rate more than doubles that of Hou et al. (2020) by grounding our tests in economic theory and modern Bayesian statistics. We conclude from the global analysis that factor research demonstrates external validity in the cross section of countries.

Post-publication Performance

McLean and Pontiff (2016) find that US factor returns "are 26% lower out-of-sample and 58% lower post-publication," suggesting that "investors learn about mispricing from academic publications."[10] Our Bayesian framework shows that, given a prior belief of zero alpha but an OLS alpha ($\hat{\alpha}$) that is positive, then our posterior belief about alpha lies somewhere between zero and $\hat{\alpha}$. Hence, a positive but attenuated post-publication alpha is the expected outcome based on Bayesian learning, rather than a sign of non-reproducibility. Further, when comparing factors cross-sectionally, the prediction of the Bayesian framework is that higher pre-publication alphas, if real, should be associated with higher post-publication alphas on average. And that is exactly what we find. We contribute new cross-sectional out-of-sample evidence that factors with higher in-sample alpha generally have higher out-of-sample alpha, and our Bayesian model offers a logical interpretation of this evidence. We conclude from this analysis that factor research demonstrates external validity in the time series.[11]

[10] Extending the evidence to global stock markets, Jacobs and Müller (2020) find that "the United States is the only country with a reliable post-publication decline in long-short returns."

[11] Data prior to the sample used in original studies also constitutes out-of-sample evidence (Linnainmaa and Roberts, 2018; Ilmanen et al., 2019). Our external validity conclusions are the same when we also include

The Multidimensional Challenge

Harvey et al. (2016) challenge the sheer number of factors and Cochrane (2011) refers to as “the multidimensional challenge” when he asks “which characteristics really provide independent information...which are subsumed by others...how many of these new factors are really important?”

The factor research universe should not be viewed as hundreds of distinct factors. We show that factors cluster into a relatively small number of highly correlated themes, and this property features prominently in our Bayesian modeling approach. We propose a factor taxonomy that algorithmically classifies factors into 13 themes possessing a high degree of within-theme return correlation and economic concept similarity, and low across-theme correlation. The emergence of themes, in which factors are minor variations on a related idea, is intuitive. For example, each value factor is defined by a specific valuation ratio, but there are many plausible ratios. Considering their variations is not spurious alpha-hacking, particularly when the “correct” value signal approach is debatable.

We estimate a replication rate of greater than 75% in 10 of the 13 themes (based on the Bayesian model including MT adjustment), the exceptions being “seasonality,” “leverage,” and “size” factor themes. We also analyze which themes matter when simultaneously controlling for all other themes. To do so, we estimate the ex post tangency portfolio of 13 theme-representative portfolios. We find that 10 of the 13 themes enter into the tangency portfolio with significantly positive weights, where the three displaced themes are “profitability,” “investment,” and “size.”

Our factor theme analysis offers a different perspective on the multiplicity of factors.[12] At the most basic level, it shows 1) that many factors are highly correlated, well in excess of 50% on average within themes, and 2) that many themes contribute significantly to the tangency portfolio. This means that many different factors bear distinct information about the economy-wide risk-return tradeoff—in other words that most themes have alpha with respect to all others.

pre-original-study out-of-sample evidence.

[12] See Bryzgalova et al. (2019), Chordia et al. (2020), Kelly et al. (2019), Kozak et al. (2020), Green et al. (2017), and Feng et al. (2020) for other perspectives on high-dimensional asset pricing problems, and Chen (2020) for an argument why everything cannot be p-hacking.

Why, the profession asks, have we arrived at a "factor zoo"? The answer, evidently, is because the risk-return tradeoff is complex and difficult to measure. The complexity manifests in our inability to isolate a single, silver bullet characteristic that pins down the risk-return tradeoff. Classifying factors into themes, we trace the economic culprits to roughly a dozen concepts. This is already a multidimensional challenge, but it is compounded by the fact that within a theme there are many detailed choices for how to configure the economic concept, which results in highly correlated within-theme factors. Together, the themes (and the factors in them) each make slightly different contributions to our collective understanding of markets. A more positive take on the factor zoo is *not* as a collective exercise in data mining and false discovery; instead, it is a natural outcome of a decentralized effort in which researchers make contributions that are correlated with, but incrementally improve on, the shared body of knowledge.

Finally, we present a number of new stylized facts regarding factor performance and replication at the theme level, at the country level, and within size groups ranging from mega-caps to micro-caps. We end with a comment on finance replication in general and in our database in particular.

1 A Bayesian Model of Factor Replication

This section presents our Bayesian model for assessing factor replicability. We first draw out some basic implications of the Bayesian framework for interpreting evidence on individual factor alphas, then present a hierarchical structure for simultaneously modeling factors in a variety themes and across many countries.

1.1 Learning About Alpha: The Bayes Case

Posterior Alpha

We begin by considering an excess return factor f_t. A study of "anomalous" factor returns requires a risk benchmark, without which we cannot separate distinctive factor behavior from run of the mill risk compensation. We assume a CAPM benchmark due to its history as a factor research benchmark for decades, and because it is not mechanically related to any of the factors that we attempt to replicate (in contrast to, say, the model of Fama and French, 1993, which by construction explains size and value factors). The factor's net performance versus the excess market factor (r_t^m) is its α:

$$f_t = \alpha + \beta r_t^m + \varepsilon_t. \tag{1}$$

Our Bayesian prior is that the alpha is normally distributed with mean zero and variance τ^2, or $\alpha \sim N(0, \tau^2)$. The mean of zero implies that CAPM holds on average, and τ measures the potential deviations from CAPM. Intuitively, the higher the confidence in the prior, the lower is τ. The error term, $\varepsilon_t \sim N(0, \sigma^2)$, has volatility σ, is independent and identically distributed over time, and σ and β are observable.[13]

The risk-adjusted return, α, is estimated as the average market-adjusted factor return from T periods of data:

$$\hat{\alpha} = \frac{1}{T}\sum_t (f_t - \beta r_t^m) = \alpha + \frac{1}{T}\sum_t \varepsilon_t. \tag{2}$$

This observed ordinary least squares (OLS) estimate $\hat{\alpha}$ is distributed $N(\alpha, \sigma^2/T)$ given the true alpha, α. From Bayes' rule, we can compute the posterior distribution of the true alpha given the data evidence and prior. The posterior exhaustively describes the Bayesian's beliefs about alpha at a future time $t > T$ given the past experience, including the posterior expected factor performance,

$$E(\alpha|\hat{\alpha}) = E\left(f_t - \beta r_t^m \middle| \hat{\alpha}\right). \tag{3}$$

We derive the posterior alpha distribution via Bayes' rule (the derivation, which is standard,

[13] Here we seek to derive some simple expressions that illustrate the economic implications of Bayesian logic. In the empirical implementation, we use slightly richer model, taking into account that σ and β must be estimated, but this does not affect the economic points that we make in this section.

is shown in Appendix A). The posterior alpha is normal with mean

$$E(\alpha|\hat{\alpha}) = \ \kappa\hat{\alpha} \tag{4}$$

where κ is a shrinkage factor given by

$$\kappa = \frac{\tau^2}{\tau^2 + \sigma^2/T} = \frac{1}{1 + \frac{\sigma^2}{\tau^2 T}} \in (0, 1) \tag{5}$$

and the posterior variance is

$$\text{Var}(\alpha|\hat{\alpha}) = \kappa\frac{\sigma^2}{T} = \frac{1}{\frac{1}{\sigma^2/T} + \frac{1}{\tau^2}}. \tag{6}$$

The first insight from this posterior is that a Bayesian predicts future returns will have smaller alpha (in absolute value) than the OLS estimate $\hat{\alpha}$, because the posterior mean ($\kappa\hat{\alpha}$) must lie between $\hat{\alpha}$ and the prior mean of zero. Said differently, a large observed alpha might be due to luck and, given the prior, we expect that at least part of this performance indeed is luck. The more data we have (higher T), the less shrinkage there is (i.e., κ closer to 1). Likewise, the stronger is the prior of zero alpha (i.e., lower τ), the heavier is the shrinkage.

When evaluating out-of-sample evidence, a positive, but lower, alpha is sometimes interpreted as a sign of replication failure. But this is the expected outcome from the Bayesian perspective (i.e., based on the latest posterior), and can be fully consistent with a high degree of replicability. In fact, we show later that the comparatively low post-publication factor performance documented by McLean and Pontiff (2016) turns out to be consistent with the posterior a Bayesian would have formed given published results. Thus, post-publication results have tended to confirm the Bayesian's beliefs and as a result the Bayesian posterior alpha estimate has been extraordinarily stable over time (see Section 3.2).

Alpha-hacking

Because out-of-sample alpha attenuation is not generally a sign of replication failure, we may want a more direct probe for non-replicability. We can build such a test into our Bayesian framework by embedding scope for "alpha-hacking," or selectively reporting or manipulating data to artificially make the alpha seem larger. We represent this idea using the following distribution of factor returns in the in-sample time period $t = 1, \ldots, T$:

$$f_t = \alpha + \beta r_t^m + \underbrace{\tilde{\varepsilon}_t + u}_{\varepsilon_t}. \tag{7}$$

Here, $\tilde{\varepsilon}_t \sim N(0, \sigma^2)$ captures usual return shocks and $u \sim N(\bar{\varepsilon}, \sigma_u^2)$ represents return inflation due to alpha-hacking. The total in-sample return shock ε_t is normally distributed, $N(\bar{\varepsilon}, \bar{\sigma}^2)$, where $\bar{\varepsilon} \geq 0$ is the alpha-hacking bias, and the variance $\bar{\sigma}^2 = \sigma^2 + \sigma_u^2 \geq \sigma^2$ is elevated due to the artificial noise created by alpha-hacking.[14] Naturally, the false benefits of alpha-hacking disappear in out-of-sample data, or in other words $\varepsilon_t \sim N(0, \sigma^2)$ for $t > T$. The Bayesian accounts for alpha-hacking as follows:

Proposition 1 (Alpha-hacking) *The posterior alpha with alpha-hacking is given by*

$$E(\alpha|\hat{\alpha}) = \; -\kappa_0 + \kappa^{hacking}\hat{\alpha} \tag{8}$$

where $\kappa^{hacking} = \frac{1}{1+\frac{\bar{\sigma}^2}{\tau^2 T}} \leq \kappa$ *and* $\kappa_0 = \kappa^{hacking}\bar{\varepsilon} \geq 0$*. Further,* $\kappa^{hacking} \to 0$ *in the limit of "pure alpha-hacking,"* $\tau \to 0$ *or* $\bar{\sigma} \to \infty$*.*

The Bayesian posterior alpha accounts for alpha-hacking in two ways. First, the estimated alpha is shrunk more heavily toward zero since the factor κ^{hacking} is now smaller. Second, the alpha is further discounted by the intercept term κ_0 due to the bias in the error terms.

We examine alpha-hacking empirically in Section 3.2 in light of Proposition 1. We consider a cross-sectional regression of factors' out-of-sample (e.g., post-publication) alphas on their in-sample alphas, looking for the signatures of alpha-hacking in the form of a negative intercept term or a slope coefficient that is too small. Appendix A presents additional theoretical results characterizing alpha-hacking.

[14] We note that this elevated variance cannot be detected by looking at the in-sample variance of residual returns since the alpha-hacking term u does not depend on time t.

1.2 Hierarchical Bayesian Model

Shared Alphas: The Case of Complete Pooling

We now embed a critical aspect of factor research into our Bayesian framework: Factors are often correlated and conceptually related to each other. For concreteness, we begin with a setting in which the researcher has access to "domestic" evidence in (1) as well as "global" evidence from an international factor, f_t^g, with known exposure β^g to the global market index r_t^g:

$$f_t^g = \alpha + \beta^g r_t^g + \varepsilon_t^g. \tag{9}$$

Here, we assume that the true alpha for this global factor is the same as the domestic alpha. In other words, we have complete "pooling" of information about alpha across the two samples. As an alternative interpretation, the researcher could have access to two related factors, say two different value factors in the same country, and assume that they have the same alpha because they capture the same investment principle.

The global shock, ε_t^g, is normally distributed $N(0, \sigma^2)$, and ε_t^g and ε_t are jointly normal with correlation ρ.[15] The estimated alpha based on the global evidence is simply its market-adjusted return:

$$\hat{\alpha}^g = \frac{1}{T} \sum_t \left(f_t^g - \beta^g r_t^g \right). \tag{10}$$

To see the power of global evidence (or, more generally, the power of observing related strategies), we consider the posterior when observing both the domestic and global evidence.

Proposition 2 (The Power of Shared Evidence) *The posterior alpha given the domestic estimate, $\hat{\alpha}$, and the global estimate, $\hat{\alpha}^g$, is normally distributed with mean*

$$E(\alpha | \hat{\alpha}, \hat{\alpha}^g) = \kappa^g \left(\frac{1}{2}\hat{\alpha} + \frac{1}{2}\hat{\alpha}^g \right). \tag{11}$$

[15] The framework can be generalized to a situation where the global shocks have a different volatility and sample length. In this case, the Bayesian posterior puts more weight on the sample with lower volatility and longer sample.

The global shrinkage parameter is

$$\kappa^g = \frac{1}{1 + \frac{\sigma^2}{\tau^2 T} \frac{1+\rho}{2}} \in [\kappa, 1] \tag{12}$$

which decreases with the correlation ρ, attaining the minimum value, $\kappa^g = \kappa$, when $\rho = 1$. The posterior variance is lower when observing both domestic and global evidence:

$$\mathrm{Var}(\alpha|\hat{\alpha}) \geq \mathrm{Var}(\alpha|\hat{\alpha}, \hat{\alpha}^g). \tag{13}$$

Naturally, the posterior depends on the average alpha observed domestically and globally. Furthermore, the combined alpha is shrunk toward the prior of zero. The shrinkage factor κ^g is smaller (heavier shrinkage) if the markets are more correlated because the global evidence provides less new information. With low correlation, the global evidence adds a lot of independent information, shrinkage is lighter, and the Bayesian becomes more confident in the data and less reliant on the prior. The proposition shows that, if a factor has been found to work both domestically and globally, then the Bayesian expects stronger out-of-sample performance than a factor that has only worked domestically (or has only been analyzed domestically).

Two important effects are at play here, and both are important for understanding the empirical evidence presented below: The domestic and global alphas are shrunk both toward *each other* and toward *zero*. For example, suppose that a factor worked domestically but not globally, say $\hat{\alpha} = 10\% > \hat{\alpha}^g = 0\%$. Then the overall evidence points to an alpha of $\frac{1}{2}\hat{\alpha} + \frac{1}{2}\hat{\alpha}^g = 5\%$, but shrinkage toward the prior results in a lower posterior, say, 2.5%. Hence, the Bayesian expects future factor returns in both regions of 2.5%. That shared alphas are shrunk together is a key feature of a *joint* model, and it generally leads to different conclusions than when factors are evaluated independently. Next we consider a perhaps more realistic model in which factors are only partially shrunk toward each other.

Hierarchical Alphas: The Case of Partial Pooling

We now consider several factors, numbered $i = 1, ..., N$. Factor i has a true alpha given by

$$\alpha^i = c + w^i. \tag{14}$$

Here, c is the common component of all alphas, which has a prior distribution given by $N(0, \tau_c^2)$. Likewise, w^i is the idiosyncratic alpha component, which has a prior distribution given by $N(0, \tau_w^2)$, independent of c and across i. Said differently, we can imagine that nature first picks of the overall c from $N(0, \tau_c^2)$ and then picks the factor-specific α^i from $N(c, \tau_w^2)$.

This hierarchical model is a realistic compromise between assuming that all factor alphas are completely different (using equation (4) for each alpha separately) and assuming that they are all the same (using Proposition 2). Rather than assuming no pooling or complete pooling, the hierarchical model allows factors to have a common component and an idiosyncratic component.

Suppose we observe factor returns of

$$f_t^i = \alpha^i + \beta^i r_t^m + \varepsilon_t^i \tag{15}$$

where ε_t^i are normally distributed with mean 0 and variance σ^2 and $\mathrm{Cor}(\varepsilon_t^i, \varepsilon_t^j) = \rho \geq 0$ for all i, j.[16] Computing the observed alpha estimates as above, $\hat{\alpha}^i = \frac{1}{T}\sum_t (f_t^i - \beta^i r_t^m)$, we derive the posterior in the following result.[17]

Proposition 3 (Hierarchical Alphas) *The posterior alpha of factor i given the evidence*

[16] Alternatively, we can write the error terms in a similar way to how we write the alphas in (14), namely $\varepsilon_t^i = \sqrt{\rho}\,\tilde{\varepsilon}_t + \sqrt{1-\rho}\,\tilde{\varepsilon}_t^i$, where $\tilde{\varepsilon}_t^i$ are idiosyncratic shocks that are independent across factors and of the common shock $\tilde{\varepsilon}_t$, with $\mathrm{Var}(\tilde{\varepsilon}_t^i) = \mathrm{Var}(\tilde{\varepsilon}_t) = \sigma^2$. We note that we require (the empirically realistic case) that $\rho \geq 0$ since we cannot have an arbitrarily large number of normal random variables with equal negative correlation (because the corresponding variance-covariance matrix would not be positive semi-definite for large enough N).

[17] The general hierarchical model is used extensively in the statistics literature, see, e.g., Gelman et al. (2013), but to our knowledge the results in Proposition 3 are not in the literature.

on all factors is normally distributed with mean

$$E(\alpha^i|\hat{\alpha}^1,\dots,\hat{\alpha}^N) = \frac{1}{1+\frac{\rho\sigma^2}{\tau_c^2 T}+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{\tau_c^2 N}}\,\hat{\alpha}^{\cdot} + \frac{1}{1+\frac{(1-\rho)\sigma^2}{\tau_w^2 T}}\left(\hat{\alpha}^i - \frac{1}{1+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{(\tau_c^2+\rho\sigma^2/T)N}}\,\hat{\alpha}^{\cdot}\right), \tag{16}$$

where $\hat{\alpha}^{\cdot} = \frac{1}{N}\sum_j \hat{\alpha}^j$ *is average alpha. When the number of factors* N *grows, the limit is*

$$\lim_{N\to\infty} E(\alpha^i|\hat{\alpha}^1,\dots,\hat{\alpha}^N) = \frac{1}{1+\frac{\rho\sigma^2}{\tau_c^2 T}}\,\hat{\alpha}^{\cdot} + \frac{1}{1+\frac{(1-\rho)\sigma^2}{\tau_w^2 T}}\left(\hat{\alpha}^i - \hat{\alpha}^{\cdot}\right). \tag{17}$$

The posterior variance of factor i's alpha using the information in all factor returns is lower than the posterior variance when looking at this factor in isolation:

$$\operatorname{Var}(\alpha^i|\hat{\alpha}^1,\dots,\hat{\alpha}^N) < \operatorname{Var}(\alpha^i|\hat{\alpha}^i). \tag{18}$$

The posterior variance is decreasing in N *and, as* $N\to\infty$*, its limit is*

$$\operatorname{Var}(\alpha^i|\hat{\alpha}^1,\dots,\hat{\alpha}^N) \searrow \frac{\rho\sigma^2}{T}\frac{1}{1+\frac{\rho\sigma^2}{\tau_c^2 T}} + \frac{(1-\rho)\sigma^2}{T}\frac{1}{1+\frac{(1-\rho)\sigma^2}{\tau_w^2 T}}. \tag{19}$$

The main insight of this proposition is that having data on many factors is helpful for estimating the alpha of any of them. Intuitively, the posterior for any individual alpha depends on all of the other observed alphas because they are all informative about the common alpha component. That is, the other observed alphas tell us whether alpha exists in general or, said another way, tell us if the CAPM appears to be violated in general. Further, the factor's own observed alpha tells us whether this specific factor appears to be especially good or bad. Using all of the factors jointly reduces posterior variance for all alphas. In summary, the joint model with hierarchical alphas has the dual benefits of identifying the common component in alphas and tightening confidence intervals by sharing information among factors.

To understand the proposition in more detail, consider first the (unrealistic) case in which all factor returns have independent shocks ($\rho = 0$). In this case, we essentially know the overall alpha when we see many uncorrelated factors. Indeed, the average observed alpha becomes a precise estimator of the overall alpha with more and more observed factors, $\hat{\alpha}^{\cdot} \to c$. Since we essentially know the overall alpha in this limit, the first term in (17) becomes $1 \times \hat{\alpha}^{\cdot}$ when $\rho = 0$ meaning that we don't need any shrinkage here. The second term is the outperformance of factor i above the average alpha, and this outperformance is shrunk toward our prior of zero. Indeed, the outperformance is multiplied by a number less than one, and this multiplier naturally decreases in the return volatility σ and decreases in our conviction in the prior (increases in τ_w).

The posterior variance is also intuitive in the case of $\rho = 0$. The posterior variance is clearly lower compared to only observing the performance of factor i itself:

$$\operatorname{Var}(\alpha^i | \hat{\alpha}^1, \hat{\alpha}^2 \ldots) = \frac{\sigma^2}{T} \frac{1}{1 + \frac{\sigma^2}{\tau_w^2 T}} < \frac{\sigma^2}{T} \frac{1}{1 + \frac{\sigma^2}{(\tau_c^2 + \tau_w^2) T}} = \operatorname{Var}(\alpha^i | \hat{\alpha}^i) \tag{20}$$

based on (19) and (6). With partial pooling, the posterior variance decreases because the denominator on the left does not have τ_c^2, reflecting that uncertainty about the general alpha has been eliminated by observing many factors.

In the realistic case where factor returns are correlated ($\rho > 0$), we see that both the average alpha $\hat{\alpha}^{\cdot}$ and factor i's outperformance $\hat{\alpha}^i - \hat{\alpha}^{\cdot}$ are shrunk toward the prior of zero. This is because we cannot precisely estimate the overall alpha even with an infinite number of correlated factors—the correlated part never vanishes. Nevertheless, we still shrink the confidence interval, $\operatorname{Var}(\alpha^i | \hat{\alpha}^1, \ldots, \hat{\alpha}^N) \leq \operatorname{Var}(\alpha^i | \hat{\alpha}^i)$, since more information is always better than less.

Multi-level Hierarchical Model

The model development to this point is simplified to draw out its intuition. Our empirical implementation is based on a more realistic (and slightly more complex) model that takes into account that factors naturally belong to different economic themes and to different regions.

In our global analysis, we have N different characteristic signals (e.g., book-to-market) across K regions, for a total of NK factors (e.g., US, developed, and emerging markets versions of book-to-market). Each of the N signals belongs to a smaller number of J theme clusters, where one cluster consists of various value factors, another consists of various momentum factors, and so on. One level of our hierarchical model allows for partially shared alphas among factors in the same theme cluster. Another level allows for commonality across regions among factors associated with the same underlying characteristic, capturing for example the connection between the book-to-market factor in different markets.

Mathematically, this means that an individual factor i has an alpha of

$$\alpha^i = \alpha^o + c^j + s^n + w^i. \tag{21}$$

Concretely, suppose factor $i \in \{1, \ldots, NK\}$ is the book-to-market factor in the US region. Part of its alpha is driven by a component that is common to all factors, α^o, which we set to zero in the empirical implementation. In addition, this factor i belongs to the value cluster $j \in \{1, \ldots, J\}$, which contributes a cluster-specific alpha $c^j \sim N(0, \tau_c^2)$. Next, since factor i is based on book-to-market characteristic $n \in \{1, \ldots, N\}$, it has an incremental signal-specific alpha of $s^n \sim N(0, \tau_s^2)$ that is shared across regions—e.g., it's the common behavior among book-to-market factors regardless of geography. Finally, $w^i \sim N(0, \tau_w^2)$ is factor i's idiosyncratic alpha, namely the incremental alpha that is unique to the US version of book-to-market.

We write this model in vector form as[18]

$$\alpha = \alpha^o 1_{NK} + Mc + Zs + w \tag{22}$$

where $\alpha = (\alpha^1, \ldots, \alpha^{NK})'$, $c = (c^1, \ldots, c^J)'$, $s = (s^1, \ldots, s^N)'$, $w = (w^1, \ldots, w^{NK})'$, M is the $NK \times J$ matrix of cluster memberships, and Z is the $NK \times N$ matrix indicating the characteristic that factor i is based on. In particular, $M_{i,j} = 1$ if factor i is in cluster j and $M_{i,j} = 0$ otherwise. Likewise, $Z_{i,n} = 1$ if factor i is based on characteristic n and $Z_{i,n} = 0$

[18] The notation 1_N refers to an $N \times 1$ vector of ones and I_N is the $N \times N$ identity matrix.

otherwise. This hierarchical model implies that the prior variance of alpha, denoted Ω, is[19]

$$\Omega = \mathrm{Var}(\alpha) = MM'\tau_c^2 + ZZ'\tau_s^2 + I_{NK}\tau_w^2. \tag{23}$$

In some cases, we analyze this model within a single region, $K = 1$ (for example, in our US-only analysis). In this case, there is no difference between signal-specific alphas and idiosyncratic alphas, so we collapse one level of the model by setting $\tau^s = 0$ and $s^n = 0$ for $n \in \{1, \ldots, N\}$. In any case, the following result shows how to compute the posterior distribution of all alphas based on the prior uncertainty, Ω, and a general variance-covariance matrix of return shocks, $\Sigma = \mathrm{Var}(\varepsilon)$. This result is at the heart of our empirical analysis.

Proposition 4 *In the multi-level hierarchical model, the posterior of the vector of true alphas is normally distributed with posterior mean*

$$E(\alpha|\hat{\alpha}) = \left(\Omega^{-1} + T\Sigma^{-1}\right)^{-1}\left(\Omega^{-1}1_{NK}\alpha_0 + T\Sigma^{-1}\hat{\alpha}\right) \tag{24}$$

and posterior variance

$$\mathrm{Var}(\alpha|\hat{\alpha}) = \left(\Omega^{-1} + T\Sigma^{-1}\right)^{-1}. \tag{25}$$

1.3 Bayesian Multiple Testing and Empirical Bayes Estimation

Frequentist MT corrections embody a principle of conservatism that seeks to limit false discoveries, controlling the family-wise error rate (FWER) or the false discovery rate (FDR). Leading frequentist methods achieve this by widening confidence intervals and raising p-values, but do not alter the underlying point estimate.

[19] Stated differently, each diagonal element of Ω is $\tau_c^2 + \tau_s^2 + \tau_w^2$. Further, if $i \neq k$, then the $(i,k)^{th}$ element of Ω is $\tau_c^2 + \tau_s^2$ if i and k are constructed from the same signal in the same cluster in different regions, it is τ_c^2 if i and k are constructed from different signals in the same cluster, and it is 0 if i and k are in different clusters.

Bayesian Multiple Testing

A large statistics literature describes how Bayesian modeling is effective for making reliable inferences in the face of multiple testing.[20] Drawing on this literature, our hierarchical model is a prime example of how Bayesian methods accomplish their MT correction based on two key model features.

First is the model prior, which imposes statistical conservatism in analogy to frequentist MT methods. It anchors the researcher's beliefs to a sensible default (e.g., all alphas are zero) in case the data are insufficiently informative about the parameters of interest. Reduction of false discoveries is achieved first by shrinking estimates toward the prior. When there is no information in the data, the alpha point estimate is the prior mean and there are no false discoveries. As data evidence accumulates, posterior beliefs migrate away from the prior toward the OLS alpha estimate. In the process, discoveries begin to emerge, though they remain dampened relative to OLS. In the large data limit, Bayesian beliefs converge on OLS with no MT correction, which is justified because in the limit there are no false discoveries. In other words, the prior embodies a particularly flexible form conservatism—the Bayesian model decides how severe of an MT correction to make based on the informativeness of the data.

Second is the hierarchical structure that captures joint behavior of factors. Modeling factors jointly means that each alpha is shrunk toward its cluster mean (i.e., toward related factors), in addition to being shrunk toward the prior of zero. So, if we observe a cluster of factors in which most perform poorly, then this evidence reduces the posterior alpha even for the few factors with strong performance—another form of Bayesian MT correction. In addition to this Bayesian discovery control coming through shrinkage of the posterior mean alpha, the Bayesian confidence interval also plays an important role and changes as a function of the data. Indeed, having data on related factors leads to a contraction of the confidence intervals in our joint Bayesian model. So while alpha shrinkage often has the effect of reducing discoveries, the increased precision from joint estimation has the opposite effect of enhancing statistical power and thus increases discoveries.

[20] See Gelman et al. (2012); Berry and Hochberg (1999); Greenland and Robins (1991); Efron and Tibshirani (2002), among others. See Gelman (2016) for an intuitive, informal discussion of the topic.

In summary, a typical implementation of frequentist MT corrections estimates parameters independently for each factor, leaves these parameters unchanged, but inflates p-values to reduce the number of discoveries. In contrast, our hierarchical model leverages dependence in the data to efficiently learn about all alphas simultaneously. All data therefore helps to determine the center and width of each alpha's confidence interval (Propositions 3 and 4). This leads to more precise estimates with "built-in" Bayesian MT correction.

Empirical Bayes Estimation

Given the central role of the prior, it might seem problematic that the severity of the Bayesian MT adjustment is at the discretion of the researcher. A powerful (and somewhat surprising) aspect of a hierarchical model is that the prior can be learned in part from the data. This idea is formalized in the idea of "empirical Bayes (EB)" estimation, which has emerged as a major toolkit for navigating multiple tests in high-dimensional statistical settings (Efron, 2012).

The general approach to EB is to specify a multi-level hierarchical model, and then to use the dispersion of estimated effects within each level to learn about the prior parameters for that level. In our setting, the specific implementation of EB is dictated by Proposition 4. We first compute each factor's abnormal return, $\hat{\alpha}$, as the intercept in a CAPM regression on the market excess return. Next, we set the overall alpha prior mean, α^o, to zero to enforce conservatism in our inferences.

From here, the benefits of EB kick in: The realized dispersion in alphas across factors helps to determine the appropriate level of conviction for the prior (that is, the appropriate values for τ_c^2, τ_s^2, and τ_w^2). For example, if we compute the average alpha for each cluster, $\hat{c}^j$ (e.g., the average value alpha, the average momentum alpha, and so on), the cross-sectional variation in $\hat{c}^j$ suggests that $\tau_c^2 \cong \frac{1}{J-1}\sum_{j=1}^{J}(\hat{c}^j - \hat{c}^{\cdot})^2$. The same idea applies to τ_s^2. Likewise, variation in observed alphas after accounting for hierarchical connections is informative about $\tau_w^2 \cong \frac{1}{NK-N-J}\sum_{i=1}^{N}(\hat{w}^i)^2$, where $\hat{w} = \hat{\alpha} - M\hat{c} - Z\hat{s}$.

The above variances illustrate the point that EB can help calibrate prior variances using the data itself. But those calculations are too crude, because they ignore sampling variation coming from the noise in returns, ε, which has covariance matrix Σ. Empirical Bayes estimates the prior variances by maximizing the prior likelihood function of the observed alphas, $\hat{\alpha} \sim N(0, \Omega(\tau_c, \tau_s, \tau_w) + \hat{\Sigma}/T)$, where the notation emphasizes that Ω depends on τ_c, τ_s, and τ_w according to (23). The likelihood function accounts for sampling variation through the a plug-in estimate of the covariance matrix of factor return shocks, $\hat{\Sigma}$.[21]

Bayesian FDR and FWER

With the EB estimates on hand, we can compute the posterior distribution of the alphas from Proposition 4. From the posterior, we can in turn compute Bayesian versions of the FDR and FWER. Suppose that we consider a factor to be "discovered" if its z-score is greater than the critical value $\bar{z} = 1.96$:

$$\frac{E(\alpha^i|\hat{\alpha}^1, \ldots, \hat{\alpha}^N)}{\sqrt{\text{Var}(\alpha^i|\hat{\alpha}^1, \ldots, \hat{\alpha}^N)}} > \bar{z}. \tag{26}$$

That is, factor i is discovered if $p\text{-value}_i^{\text{Bayes}} < 2.5\%$,[22] where we use the posterior to compute

$$p\text{-value}_i^{\text{Bayes}} = Pr(\alpha^i < 0|\hat{\alpha}^1, \ldots, \hat{\alpha}^N). \tag{27}$$

This is a false discovery if the true alpha is actually non-positive. The Bayesian FDR is:

$$\text{FDR}^{\text{Bayes}} = E\left(\frac{\sum_i 1_{\{i \text{ false discovery}\}}}{\sum_i 1_{\{i \text{ discovery}\}}}\middle|\hat{\alpha}^1, \ldots, \hat{\alpha}^N\right) \tag{28}$$

where we condition on the data including at least one discovery (so the denominator is not zero), otherwise FDR is set to zero (see Benjamini and Hochberg, 1995).

The following proposition is a novel characterization of the Bayesian FDR, and shows that it is the average Bayesian p-value across all discoveries:

Proposition 5 (Bayesian FDR) *Conditional on the parameters of the prior distribution and data with at least one discovery, the Bayesian false discovery rate is the average Bayesian p-value:*

[21] We discuss the details of our EB estimation procedure in Appendix B.

[22] We use a critical value of 2.5% rather than 5% because the 1.96 cut-off corresponds to a 2-sided test, while false discoveries are only on one side in the Bayesian framework.

$$\text{FDR}^{\text{Bayes}} = \frac{1}{\#\text{discoveries}} \sum_{i \text{ discovery}} p\text{-value}_i^{\text{Bayes}} \leq 2.5\%. \tag{29}$$

This result shows explicitly how the Bayesian framework controls the false discovery rate without the need for additional MT adjustments. The definition of a discovery ensures that at most 2.5% of the discoveries are false according to the Bayesian posterior, which is exactly the right distribution for assessing discoveries from the perspective of the Bayesian. Further, if many of the discovered factors are highly significant (as is the case in our data), then the Bayesian FDR is much lower than 2.5%.[23]

We can also compute a Bayesian version of the family-wise error rate, which is the probability of making one or more false discoveries in total:

$$\text{FWER}^{\text{Bayes}} = Pr\left(\sum_i 1_{\{i \text{ false discovery}\}} \geq 1 \middle| \hat{\alpha}^1, \ldots, \hat{\alpha}^N\right). \tag{30}$$

If we define a discovery as in (26) using the standard critical value $\bar{z} = 1.96$, then we do not necessarily control the family-wise error rate, $\text{FWER}^{\text{Bayes}}$, since this is a harsh criterion that is concerned with the risk of a single false discovery without regard for the number of missed discoveries. In any case, $\text{FWER}^{\text{Bayes}}$ is a probability that we can compute from the posterior so it is straightforward to choose a critical value $\bar{z}$ to ensure $\text{FWER}^{\text{Bayes}} \leq 5\%$ or any other level one prefers. The main point is that the Bayesian approach to replication lends itself to any inferential calculation the researcher desires because the posterior is a complete characterization of Bayesian beliefs about model parameters.

A Comparison of Frequentist and Bayesian False Discovery Control

We illustrate the benefits of Bayesian inference for our replication analysis via simulation. We assume a factor generating process based on the hierarchical model above and, for simplicity, consider a single region (as in our empirical US-only analysis), removing s^n and τ_s^2 from equations (21) and (23). We analyze discoveries as we vary the prior variances τ_c and τ_w.

[23] Proposition 5 formalizes the argument of Greenland and Robins (1991) that "from the empirical-Bayes or Bayesian perspective, multiple comparisons are not really a 'problem.' Rather, the multiplicity of comparisons provides an opportunity to improve our estimates through judicious use of any prior information (in the form of model assumptions) about the ensemble of parameters being estimated."

The remaining parameters are calibrated to our estimates for the US region in our empirical analysis below.

We simulate an economy with 130 factors in 13 different clusters of 10 factors each, observed monthly over 68 years. We assume that the mean alpha, α^o, is zero. We then draw a cluster alpha from $c^j \sim N(0, \tau_c^2)$ and a factor-specific alpha as $w^i \sim N(0, \tau_w^2)$. Based on these alphas, we generate realized returns by adding Gaussian noise.[24]

We compute p-values separately using OLS with no adjustment or adjusting with the Benjamini-Yekutieli (BY) method. We also use EB to estimate the posterior alpha distribution, treating τ_c and τ_w as known in order to simplify simulations and focus on the Bayesian updating. For OLS and BY, a discovery occurs when the alpha estimate is positive and the two-sided p-value is below 5%. For EB, we consider it a discovery when the posterior probability that alpha is negative is less that 2.5%. For each pair of τ_c and τ_w, we draw 10,000 simulated samples, and report average discovery rates over all simulations.

Figure 2 reports alpha discoveries based on the OLS, BY, and EB approaches. For each method, we report the true FDR in the top panels (recall, we know the truth since this is a simulation) and the "true discovery rate"[25] in the bottom panels.

When idiosyncratic variation in true alphas is small (left panels with $\tau_w = 0.01\%$) and the variation in cluster alphas is also small (values of τ_c near zero on the horizontal axis), alphas are very small and true discoveries are unlikely. In this case, the OLS false discovery rate can be as high as 25% as seen in the upper left panel. However, both BY and EB successfully correct this problem and lower the FDR. The lower left panel shows that the BY correction pays a high price for its correction in terms of statistical power when τ_c is larger. In contrast, EB exhibits much better power to detect true positives while maintaining a similar false discovery control as BY. In fact, when there are more discoveries to be made in the data (as τ_c increases), EB becomes even more likely to identify true positives than OLS. This is due

[24] The noise covariance matrix has a block structure calibrated to our data, with a correlation of 0.55 among factors in the same cluster and a correlation of 0.03 across clusters. The return volatility for each factor is 10% per annum.

[25] We define the true discovery rate to be the number of significantly positive alphas according to, respectively, OLS, BY, and EB divided by the number of truly positive alphas. Given our simulation structure, half of the alphas are expected to be positive in any simulation. Some of these will be small (i.e., economically insignificant) positives, so a testing procedure would require a high degree of statistical power to detect them. This is why the true discovery rate is below one even for high values of τ_c.

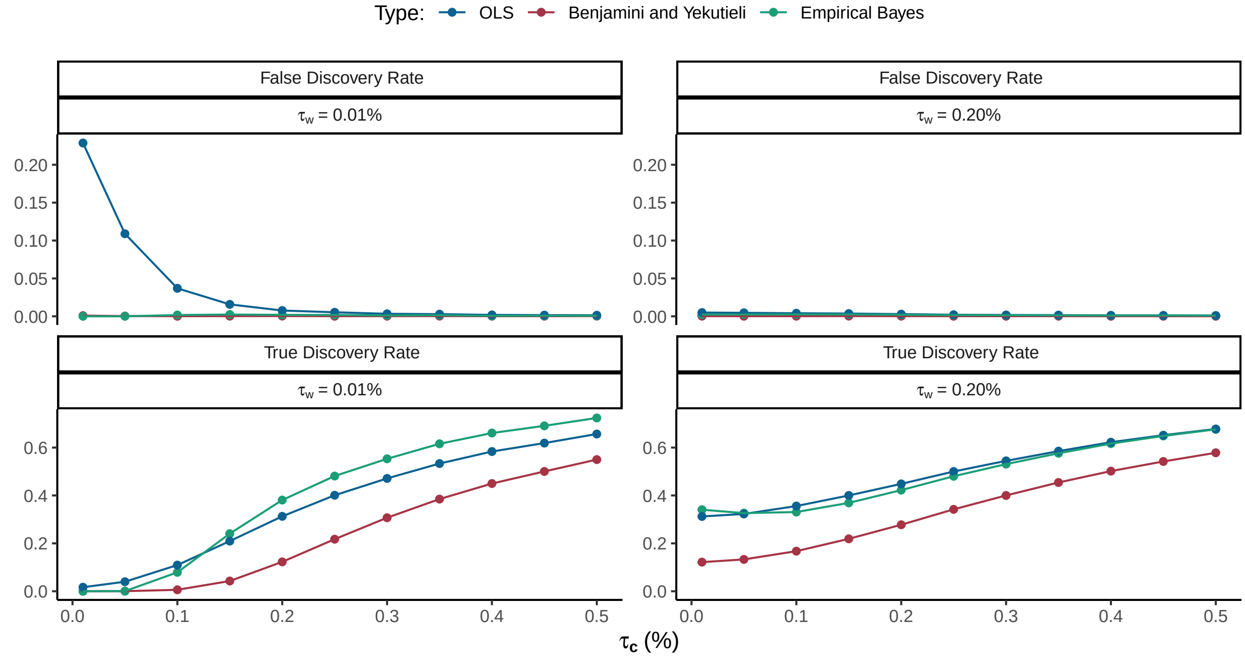


Figure 2: Simulation Comparison of False Discovery Rates

Note: The upper panels show the realized false discovery rate computed as the proportion of discovered factors for which the true alpha is negative, averaged over 10,000 simulations. The lower panels show the true discovery rate computed as the number of discoveries where the true alpha is positive divided by the total number of factors where the true alpha is positive. The left and right panels use low and high values of idiosyncratic variation in alphas (τ_w), respectively. The x-axis varies cluster alpha dispersion, τ_c.

to the joint nature of the Bayesian model, whose estimates are especially precise compared to OLS due to EB's ability to learn more efficiently from dependent data. This illustrates a point of Greenland and Robins (1991) that "Unlike conventional multiple comparisons, empirical-Bayes and Bayes approaches will alter and can improve point estimates and can provide more powerful tests and more precise (narrower) interval estimators." When the idiosyncratic variation is larger ($\tau_w = 0.20\%$), there are many more true discoveries to be made, so the false discovery rate tends to be low even for OLS with no correction. Yet in the lower right panel we continue to see the costly loss of statistical power suffered by the BY correction.

In summary, EB accomplishes a flexible MT adjustment by adapting to the data generating process. When discoveries are rare so that there is a comparatively high likelihood of false discovery, EB imposes heavy shrinkage and behaves similarly to the conservative BY correction. In this case, the benefit of conservatism costs little in terms of power exactly because true discoveries are rare. Yet when discoveries are more likely, EB behaves more like uncorrected OLS, giving it high power to detect discoveries and suffering little in terms of false discoveries because true positives abound.

The limitations of frequentist MT corrections are well studied in the statistics literature. Berry and Hochberg (1999) note that "these procedures are very conservative (especially in large families) and have been subjected to criticism for paying too much in terms of power for achieving (conservative) control of selection effects." The reason is that, while inflating confidence intervals and p-values indeed reduces the discovery of false positives, it also reduces power to detect true positives.

Much of the discussion around MT adjustments in the finance literature fails to consider the loss of power associated with frequentist corrections. But, as Greenland and Hofman (2019) point out, this tradeoff should be a first-order consideration for a researcher navigating multiple tests, and frequentist MT corrections tend to place an implicit cost on false positives that can be unreasonably large. Unlike some medical contexts for example, there is no obvious motivation for asymmetric treatment of false positives and missed positives in factor research. The finance researcher may be willing to accept the risk of a few false discoveries to avoid missing too many true discoveries. In statistics, this is sometimes discussed in terms of an (abstract) cost of Type I versus Type II errors,[26] but in finance we can make this cost concrete: We can look at the profit of trading on the discovered factors, where the cost of false discoveries is then the resulting extra risk and money lost (Section 3.3).

2 A New Public Data Set of Global Factors

We study a global dataset with 153 factors in 93 countries. In this section, we provide a brief overview of our data construction. We have posted the code along with extensive documentation detailing every implementation choice that we make for each factor.[27]

[26] As Greenland and Robins (1991) point out, "Decision analysis requires, in addition to the likelihood function, a loss function, which indicates the cost of each action under the various possible values for the unknown parameter (benefits would be expressed as negative costs). Construction of a loss function requires one to quantify costs in terms of dollars, lives lost, or some other common scale."

[27] It is available at www.bryankellyacademic.org and at https://github.com/bkelly-lab/GlobalFactor.

Factors

The set of factors we study is based on the exhaustive list compiled by Hou et al. (2020). They study 202 different characteristic signals from which they build 452 factor portfolios. The proliferation is due to treating 1, 6, and 12-month holding periods for a given characteristic as different factors, and due to their inclusion of both annual and quarterly updates of some accounting-based factors. In contrast, we focus on a 1-month holding period for all factors, and we only include the version that updates with the most recent accounting data (which could be either annual or quarterly). Lastly, we exclude a small number of factors for which data is not available globally. This gives us a set of 180 feasible global factors. For this set, we exclude factors based on industry or analyst data because they have comparatively short samples.[28] This leaves us with 138 factors. Finally, we add 15 factors studied in the literature that were not included in Hou et al. (2020).

For each characteristic, we build the 1-month holding period factor return within each country as follows. First, in each country and month, we sort stocks into characteristic terciles (top/middle/bottom third) with breakpoints based on non-micro stocks in that country.[29] For each tercile, we compute its "capped value weight" return, meaning that we weight stocks by their market equity winsorized at the NYSE 80th percentile. This construction ensures that tiny stocks have tiny weights and any one mega stock does not dominate a portfolio, seeking to create tradable, yet balanced, portfolios.[30] The factor is then defined as the high-tercile return minus the low-tercile return, corresponding to the excess return of a long-short zero-net-investment strategy. The factor is long (short) the tercile identified by the original paper to have the highest (lowest) expected return. Finally, we compute each factor's $\hat{\alpha}^i$ via an OLS regression on a constant and the corresponding region's market portfolio.

[28] Global industry codes (GICS) are only available from 2000 and I/B/E/S data from the mid-1980's (but coverage in early years is somewhat sparse).

[29] Specifically, we start with all non-micro stocks in a country (i.e., larger than NYSE 20th percentile) and sort them into three groups of equal numbers of stocks based on the characteristic, say book-to-market. Then we distribute the micro-cap stocks into the three groups based on the same characteristic breakpoints. This process ensures that the non-micro stocks are distributed equally across portfolios, creating more tradable portfolios.

[30] For robustness, Figure C.1 of the appendix reports our replication results to using standard, uncapped value weights to construct factors.

For a factor return to be non-missing, we require that it has at least 5 stocks in each of the long and short legs. We also require a minimum of 60 non-missing monthly observations for each country-specific factor for inclusion in our sample. When grouping countries into regions (US, developed ex. US, and emerging) we use the MSCI development classification as of January 7th 2021. When aggregating factors across countries, we use capitalization-weighted averages of the country-specific factors. For the developed and emerging market factors, we require that at least three countries have non-missing factor returns.

Clusters

We group factors into clusters using hierarchical agglomerative clustering (Murtagh and Legendre, 2014). We define the distance between factors as one minus their pairwise correlation and use the linkage criterion of Ward (1963). The correlation is computed based on CAPM-residual returns of US factors signed as in the original paper. Appendix Figure C.2 shows the resulting dendrogram, which illustrates the hierarchical clusters identified by the algorithm. Based on the dendrogram, we choose 13 clusters that demonstrate a high degree of economic and statistical similarity. The cluster names indicate the types of characteristics that dominate each group: Accruals*, Debt Issuance*, Investment*, Leverage*, Low risk, Momentum, Profit Growth, Profitability, Quality, Seasonality, Size*, Skewness*, and Value, where the star (*) indicates that these factors bet against the corresponding characteristic (e.g., accrual factors go long stocks with low accruals while shorting those with high accruals). Appendix Figure C.3 shows that the average within-cluster pairwise correlation is above 0.5 for 10 out of 13 clusters, and Table C.3 provides details on the cluster assignment, sign convention, and original publication source for each factor.

Data

Return data is from CRSP for the US (beginning in 1926) and from Compustat for all other countries (beginning in 1986 for most developed countries).[31] All accounting data is from Compustat. For international data, all variables are measured in US dollars (based on exchange rates from Compustat) and excess returns are relative to the US treasury bill

[31] Appendix Table C.5 shows start date and other information for all countries included in our dataset.

rate. To alleviate the influence of data errors in international data, we winsorize return and market equity across all countries in the international sample in each month at 0.1% and 99.9%.

We restrict our focus to common stocks that are identified by Compustat as the primary security of the underlying firm and assign stocks to countries based on the country of their exchange.[32] In the US, we include delisting returns from CRSP. If a delisting return is missing and the delisting is for a performance-based reason, we set the delisting return to −30% following Shumway (1997). In the global data, delisting returns are not available, so all performance-based delistings are assigned a return of −30%.

We create characteristics using both annual and quarterly accounting data, and assume accounting data is available to the public four months after the fiscal period end. When creating a factor we use the most recent data, which means that most accounting characteristics are updated four times per year when the quarterly release becomes available. An important choice in this implementation is that we aggregate quarterly income and cash flow items over the most recent four quarters to avoid distortions from seasonal effects in the underlying business. When creating valuation ratios, we always use the most recent price data following Asness and Frazzini (2013).

Empirical Bayes Estimation

We estimate the hyperparameters and the posterior alpha distributions of our Bayesian model via EB. Appendix B provides details on the EB methodology and the estimated parameters.

3 Empirical Assessment of Factor Replicability

We now report replication results for our global factor sample. We first present an internal validity analysis by studying US factors over the full sample. Then we analyze external validity in the global cross section and in the time series (post-publication factor returns).

[32] Compustat identifies primary securities in the US, Canada and rest of the world. This means that some firms can have up to three securities in our data set. In practice, the vast majority of firms (97%) only have one security in our sample at a given point in time.

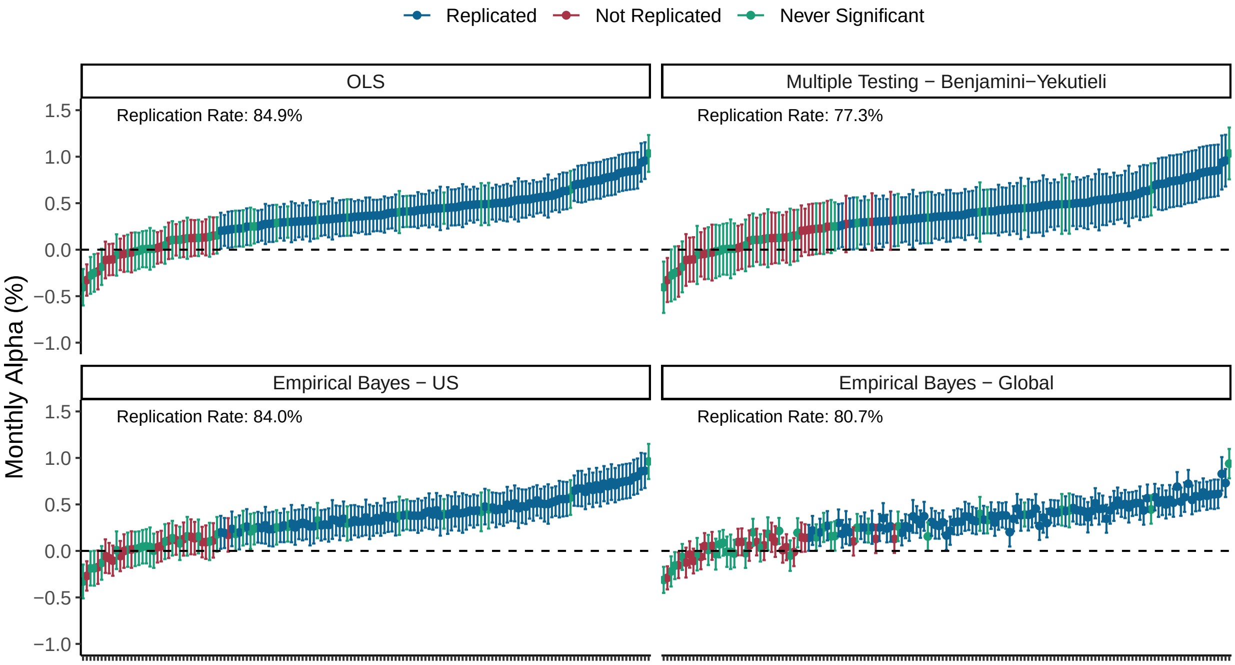


Figure 3: Alpha Distributions for US Factors

Note: The figure reports point estimates and confidence intervals for US factors. The upper left reports OLS estimates. The upper right uses the OLS point estimate but adjusts the confidence interval following the BY procedure. The lower left panel shows our EB posterior confidence intervals using only US data. The lower right continues to show EB results for US factors, but estimates the US factor posterior from global data rather than US-only data. Blue (red) confidence intervals correspond to factors that were significant in the original study in the literature and that we find significant (insignificant) based on the method in each panel. Green intervals correspond to factors that the original study find insignificant or do not evaluate in terms of average return significance. The order of factors is the same in all panels and is arranged from lowest OLS alpha to highest. Table C.4 shows the factor names arranged in the same order.

3.1 Internal Validity

We report full sample performance of US factors in Figure 3. Each panel illustrates the CAPM alpha point estimate of each factor corresponding to the dot at the center of the vertical bars. Vertical bars represent the 95% confidence interval for each estimate. Bar colors differentiate between three types of factors. Blue shows factors that are significant in the original study and remain significant in our full sample. Red shows factors that are significant in the original study but are insignificant in our test. Green shows factors that are not significant in the original study, but are included in the sample of Hou et al. (2020).

The four panels in Figure 3 differ in how the alphas and their confidence intervals are estimated. The upper left panel reports the simple OLS estimate of each alpha, $\hat{\alpha}_{\text{ols}}$, and the 95% confidence intervals based on unadjusted standard errors, $\hat{\alpha}_{\text{ols}} \pm 1.96 \times SE_{\text{ols}}$.[33] The factors are sorted by OLS $\hat{\alpha}$ estimate, and we use this ordering for the other three panels as well. We find that the OLS replication rate is 84.9%, computed as the number of blue factors (101) divided by the sum of red and blue factors (119). Based on OLS tests, factors are highly replicable.

The upper right panel repeats this analysis using the MT adjustment of Benjamini and Yekutieli (2001) (denoted BY), which is advocated by Harvey et al. (2016) and implemented by Hou et al. (2020). This method leaves the OLS point estimate unchanged, but inflates the p-value. We illustrate this visually by widening the alpha confidence interval. Specifically, we find the BY-implied critical value[34] in our sample to be a t-statistic of 2.75, and we compute the corresponding confidence interval as $\hat{\alpha}_{\text{ols}} \pm 2.75 \times SE_{\text{ols}}$. We deem a factor as significant according to the BY method if this interval lies entirely above zero. Naturally, this widening of confidence intervals produces a lower replication rate of 77.3%. However, the BY correction does not materially change the OLS-based conclusion that factors appear highly replicable.

The lower left panel is based on our empirical Bayes estimates using the full sample of US factors. For each factor, we use Proposition 4 to compute its posterior mean, $E(\alpha_i|(\hat{\alpha}_j)_{j \text{ any US factor}})$, shown as the dot at the center of the confidence interval. These dots change relative to the OLS estimates, in contrast to BY and other frequentist MT methods that only change the size of the confidence intervals. We also compute the posterior volatility to produce Bayesian confidence intervals, $E(\alpha_i|(\hat{\alpha}_j)_{j \text{ any US factor}}) \pm 1.96 \times \sigma(\alpha_i|(\hat{\alpha}_j)_{j \text{ any US factor}})$. The replication rate based on Bayesian model estimates is 84.0%, larger than BY and, coincidentally, similar to the OLS replication rate. This replication rate has a built-in conservatism from the zero-alpha prior, and it further accounts for the multiplicity of factors because each factor's posterior depends on *all* of the observed evidence in the US (not just own-factor performance).

The lower right panel again reports EB estimates for US factor, but now we allow the

[33] We define SE_{ols} as the diagonal of the alpha covariance matrix $\hat{\Sigma}$, which we estimate according to Appendix B.

[34] We compute the BY-implied critical value as the average of the t-statistic of the factor that is just significant based on BY (the factor with the highest BY-adjusted p-value below 5%) and the t-statistic of the factor that is just insignificant (the factor with the lowest BY-adjusted p-value above 5%).

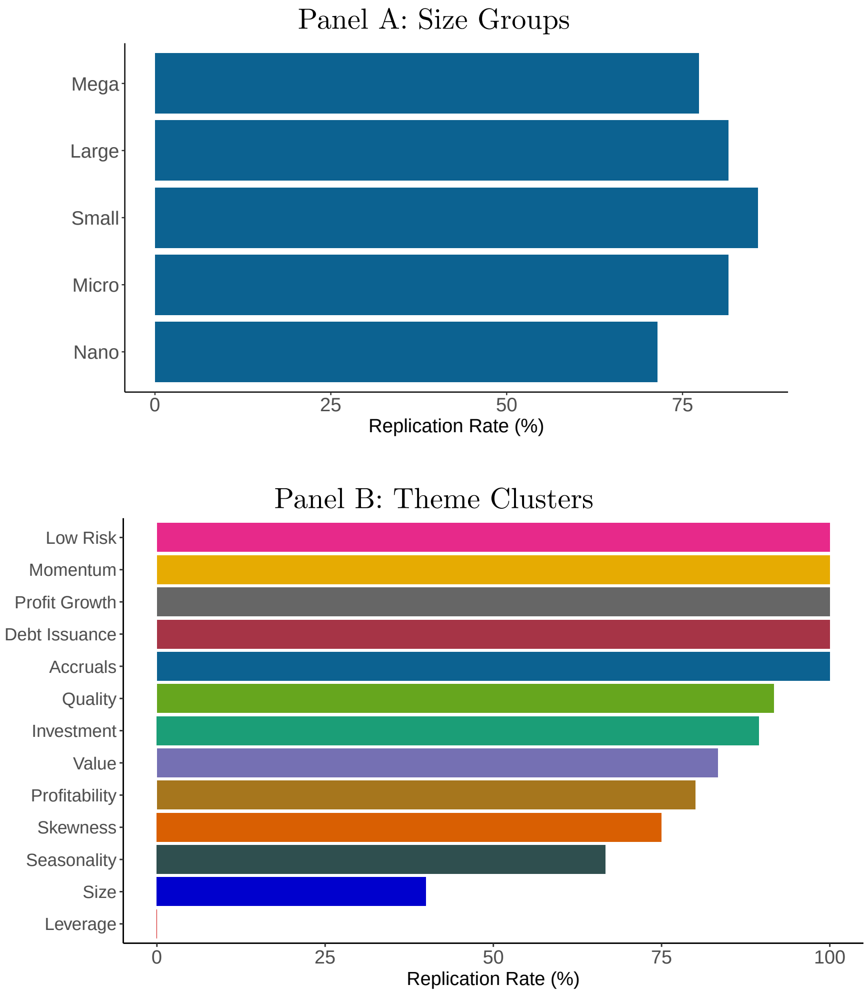
posterior to depend not just on US data, but on data from all over the world. That is, we compute the posterior mean and variance for each US factor conditional on the alpha estimates for all factors in all regions. The resulting replication rate is 80.7%, which is slightly lower than the EB replication rate using only US data. Some posterior means are reduced due to the fact that some factors have not performed as well outside the US, which affects posterior means for the US through the dependence among global alphas. For example, when the Bayesian model seeks to learn the true alpha of the "US annual sales growth" factor, the Bayesian's conviction regarding positive alpha is reduced by taking into account that the international version of this factor has underperformed the US version.[35]

To further assess internal validity, we investigate the replication rate for US factors when those factors are constructed from subsamples based on stock size. One of the leading criticisms of factor research replicability is that results are driven by illiquid small stocks whose behavior in large part reflects market frictions and microstructure as opposed to just economic fundamentals or investor preferences. In particular, Hou et al. (2020) argue that they find a low replication rate because they limit the influence of micro-caps. We find that factors demonstrate a high replication rate throughout the size distribution. Panel A of Figure 4 reports replication rates for US size categories shown in the five bars: mega stocks (largest 20% of stocks based on NYSE breakpoints), large stocks (market capitalization between the 80^{th} and 50^{th} percentile of NYSE stocks), small stocks (between the 50^{th} and 20^{th} percentile), micro stocks (between the 20^{th} and 1^{st} percentile), and nano stocks (market capitalization below the 1^{st} percentile).

We see that the EB replication rates in mega and large stock samples are 77.3% and 81.5%, respectively. This is only marginally lower than the overall US sample replication rate of 84.0%, indicating that criticisms of factor replicability based on arguments around stock size or liquidity are largely groundless. Even micro and nano stocks deliver replication rates of 81.5% and 71.4%, respectively.

In Panel B of Figure 4, we report US factor replication rates by theme cluster. 10 out

[35] To provide a few more details on this example, the US factor based on annual sales growth (sale_gr1) has a posterior volatility of 0.0987% using only US data and 0.0768% using global data, leading to a tighter confidence interval with the global data. However, the posterior mean is 0.264% using only US data and 0.128% using global data.



Panel A: Size Groups

Panel B: Theme Clusters

Figure 4: US Replication Rates By Size Group and Theme Cluster

Note: Panel A reports replication rates for US factors formed from subsamples defined by stocks' market capitalization using our EB method. Panel B reports replication rates for US factors in each theme cluster.

of 13 themes are replicable with a rate of 75% or better, with the exceptions being the seasonality, leverage, and size themes. To understand these exceptions, we note that size factors are stronger in emerging markets (bottom panel of Figure C.8) and among micro and nano stocks (bottom panels of Figure C.9). The theoretical foundation of the size effect is a compensation for market illiquidity (Amihud and Mendelson, 1986) and market liquidity risk (Acharya and Pedersen, 2005). Theory predicts that the illiquidity (risk) premium should be the same order of magnitude as the differences in trading costs and these differences are simply much larger in emerging markets and among micro stocks.

Another reason why some factors and themes appear insignificant is that we are not accounting for other factors. Factors published after 1993 are routinely benchmarked to the Fama-French three-factor model (and, more recently, to the updated five-factor model). Some factors are insignificant in terms of raw return or CAPM alpha, but their alpha becomes significant after controlling for other factors. This indeed explains the lack of replicability for the leverage theme. While CAPM alphas of leverage factors are insignificant, we find that leverage is one of the best performing themes once we account for multiple factors (see Section 3.4 below).

3.2 External Validity

We find a high replication rate in our full-sample analysis, indicating that the large majority of factors are reproducible at least in-sample. We next study the external validity of these results in international data and in post-publication US data.

Global Replication

Figure 5 shows corresponding replication rates around the world. We report replication rates from four testing approaches: (1) OLS with no adjustment; (2) OLS with Benjamini-Yekutieli MT adjustment; (3) the EB posterior conditioning only on factors within a region ("Empirical Bayes – Region"); and (4) EB conditioning on factors in all regions ("Empirical Bayes – All"). Even when using all global data to update the posterior of all factors, the reported Bayesian replication rate applies only to the factors within the stated region.

The first set of bars establishes a baseline by showing replication rates for the US sample, summarizing the results from Figure 3. The next two sets of bars correspond to the developed ex. US sample and the emerging markets sample, respectively.[36] Each region factor is a capitalization-weighted average of that factor among countries within a given region, and the replication rate describes the fraction of significant CAPM alphas for these regional factors.

OLS replication rates in developed and emerging markets are generally lower than in

[36] The developed and emerging samples are defined by the MSCI development classification and include 23 and 27 countries, respectively. The remaining 43 countries in our sample that are classified as neither developed nor emerging by MSCI do not appear in our developed and emerging region portfolios, but they are included in the "world" versions of our factor portfolios.

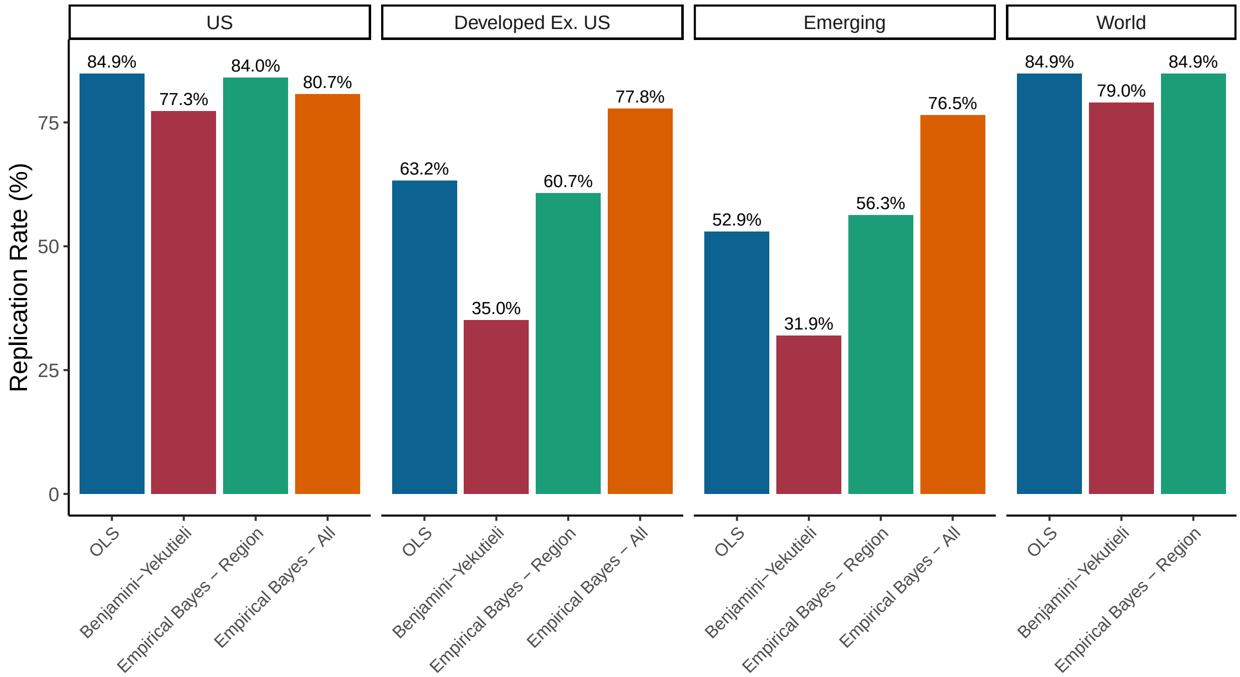


Figure 5: Replication Rates in Global Data

Note: We report replication rates for factors in three global regions (US, developed ex. US, and emerging) and for the world as a whole. A factor in a given region is the capitalization-weighted average factor for countries in that region. We report OLS replication rates with no adjustment and with Benjamini-Yekutieli multiple testing adjustment. We also report replication rates based on the empirical Bayes posterior. We consider two EB methods. In both methods, the replication rate refers only to factors within the region of interest, but the posterior is computed by conditioning either on data from that region alone ("Empirical Bayes – Region") or on the full global sample ("Empirical Bayes – All"). We deem a factor successfully replicated if its 95% confidence interval excludes zero for a given method.

the US, and the frequentist Benjamini-Yekutieli correction has an especially large negative impact on replication rate. This is a case in which the Bayesian approach to MT is especially powerful. Even though the alphas of all regions are shrunk toward zero, the global information set helps EB achieve a high degree of precision, narrowing the posterior distribution around the shrunk point estimate. We can see this in increments. First, the EB replication rate using region-specific data ("Empirical Bayes – Region" in the figure) is just below the OLS replication rate but much higher than the Benjamini-Yekutieli rate. When the posterior leverages global data ("Empirical Bayes – All" in the figure), the replication rate is higher still, reflecting the benefits of sharing information across regions, as recommended by the dependence among alphas in the hierarchical model.

Finally, we use the global model to compute, for each factor, the capitalization-weighted

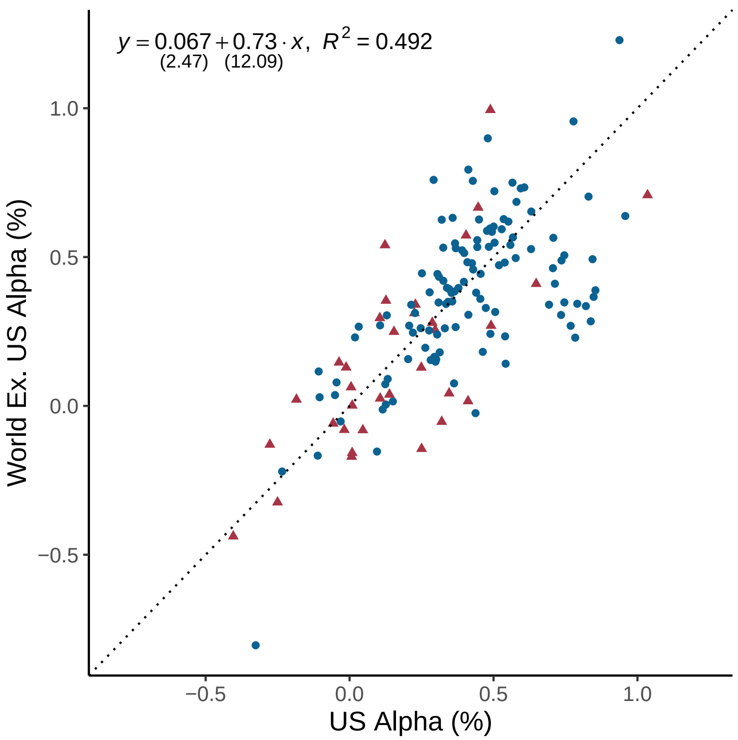


Figure 6: US Factor Alphas Versus World Ex. US

Note: The figure compares OLS alphas for US factors versus their international counterpart. Each world ex. US factor is a capitalization-weighted average of the factor in all other countries of our sample. Blue points correspond to factors that were significant in the original study in the literature, while red points are those for which the original paper did not find a significant effect (or did not study the factor in terms of average return significance). The dotted line is the 45° line. The figure also reports a regression of world ex. US alpha on US alpha.

average alpha across all countries in our sample ("World" in the figure). Using data from around the world, we find a Bayesian replication rate of 84.9%.

Why do international OLS replication rates differ from the US? This is due primarily to the the fact that foreign markets have shorter time samples. Point estimates are similar in magnitude for the US and international data. Figure 6 shows the alpha of each US factor against the alpha of the corresponding factor for the world ex. US universe. The data cloud aligns closely with the 45° line, demonstrating the close similarity of alpha magnitudes in the two samples. But shorter international samples widen confidence intervals, and this is the primary driver of the drop in OLS replication rates outside the US.

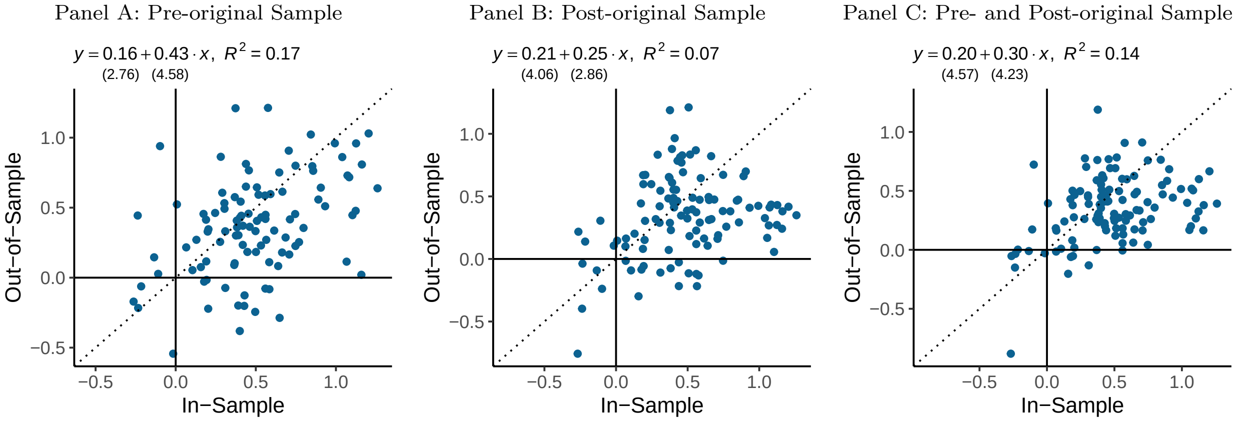


Figure 7: In-Sample versus Out-of-Sample Alphas for US Factors

Note: The figure plots OLS alphas for US factors during the in-sample period (i.e., the period studied in the original publication) versus out-of-sample alphas. In Panel A, out-of-sample is the time period before the in-sample period. In Panel B, out-of-sample is the time period before the in-sample period. In Panel C, out-of-sample includes both the time period before and after the in-sample period. We require at least five years of out-of-sample data for a factor to be included, amounting to 103, 109 and 113 factors in panel A, B and C. The figure also reports an OLS regression of out-of-sample alphas on in-sample alphas. The dotted line is the 45° line.

Time Series Out-of-Sample Evidence

McLean and Pontiff (2016) document the intriguing fact that, following publication, factor performance tends to decay. They estimate an average post-publication decline of 58% in factor returns. In our data, the average in-sample alpha is 0.45% per month and the average out-of-sample alpha is 0.31% when looking post-original sample (or 0.32% when looking pre- and post-original sample), implying a decline of about a third. We can get more economic intuition by looking at these findings cross-sectionally.

Figure 7 makes a cross-sectional comparison of the in-sample and out-of-sample alphas of our US factors. The in-sample period is the sample studied in the original reference. The out-of-sample period in Panel A is the time period before the start of in-sample period, while in Panel B it is the period following the in-sample period. Panel C defines out-of-sample as the combined data from the periods before and after the originally studied sample. We find that 84.3% of US factors have positive returns in the pre-original sample, 82.6% are positive in the post-original sample, and 88.5% are positive in the combined out-of-sample period. When we regress out-of-sample alphas on in-sample alphas, we find a slope coefficient of 0.43, 0.25, and 0.30 in Panels A, B, and C, respectively. The slopes are highly significant (ranging from $t = 2.9$ to $t = 4.6$) indicating that in-sample alphas contain something "real" rather than being the outcome of pure data mining, as factors that performed better in-sample also tend to perform better out-of-sample.

The significantly positive slope allows us to reject the hypothesis of "pure alpha-hacking," which would imply a slope of zero, as seen in Proposition 1. Another inference from this regression is that the intercept is positive, while alpha-hacking of the form studied in Proposition 1 would imply a negative intercept.

That the slope coefficient is positive and less than one is consistent with basic Bayesian logic of equation (4). As we emphasize in Section 1, a Bayesian would expect at least some attenuation in out-of-sample performance. This is because the published studies report the OLS, while Bayesian beliefs include shrinkage of the OLS toward the zero-alpha prior. More specifically, with no alpha hacking or arbitrage, the Bayesian expects a slope of approximately 0.9 using equation (5) and our EB hyperparameters (see appendix Table B.1).[37] Hence, the slope coefficients in Figure 7 are too low relative to this Bayesian benchmark. In addition to the moderate slope, there is evidence that the dots in Figure 7 have a concave shape (as seen more clearly in appendix Figure C.5). These results indicate that, while we can rule out pure alpha-hacking (or p-hacking), there is some evidence that the highest in-sample alphas may either be data-mined or arbitraged down.

From the Bayesian perspective, another interesting evaluation of time series external validity is to ask whether the new information contained in out-of-sample data moves the posterior alpha toward zero or not. Imagine a Bayesian observing the arrival of factor data in real time. As new data arrives, she updates her beliefs for all factors based on the information in the full cross section of factor data. In the top panel of Figure 8, we show how the Bayesian's average alpha posterior would have evolved in real time.[38] We focus on all US factors that are available since at least 1955. Starting in 1960, we re-estimate the hierarchical model using the empirical Bayes estimator in December of each year. The

[37] The slope is $\kappa = 1/(1 + \sigma^2/(T\tau^2)) = 0.9$, where $\sigma^2 = 10\%^2/12$, the average in-sample period length is $T = 420$ months, and $\tau^2 = \tau_c^2 + \tau_w^2 = (0.38\%)^2 + (0.21\%)^2 = (0.43\%)^2$.

[38] Here we keep τ_c and τ_w fixed at their full-sample values of 0.37% and 0.21% to mimic the idea of given decision maker who starts with a given prior and updates this view based on new data, while keeping the prior fixed. Figure C.6 shows that the figure is almost the same with rolling estimates of τ_c and τ_w, and Figure C.7 shows that this consistency arises because the rolling estimates are relatively stable.

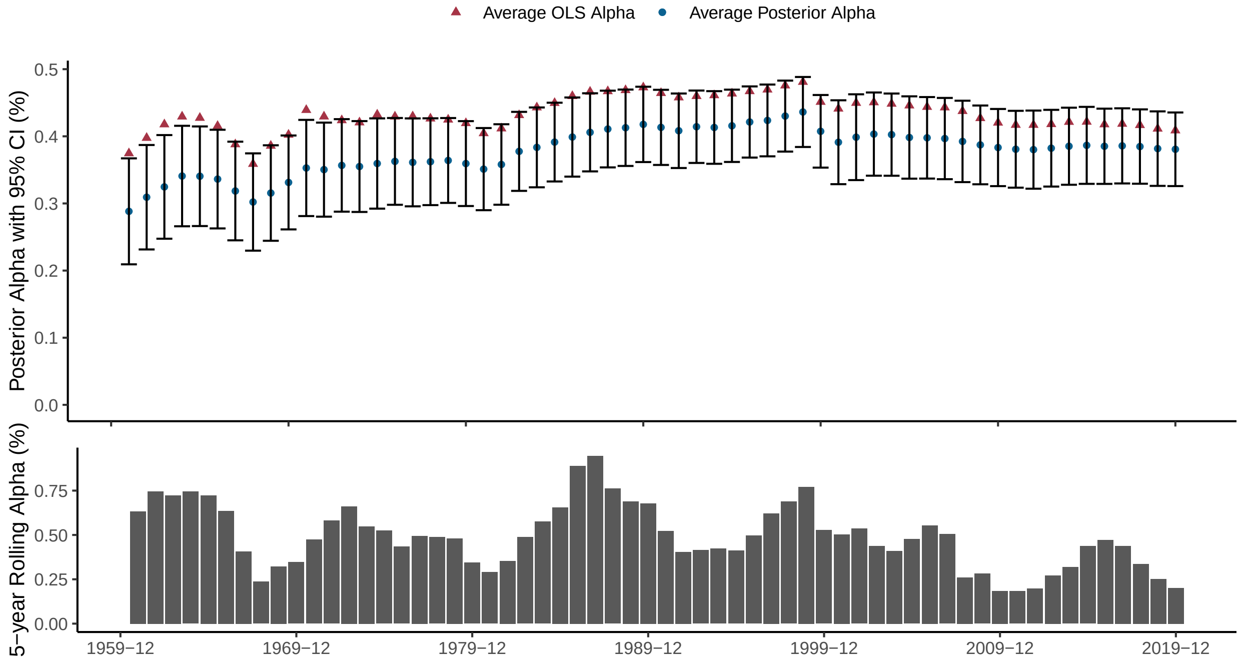


Figure 8: US Factor Alpha Posterior Distribution Over Time

Note: The top panel reports the average 95% posterior confidence interval for the average monthly alpha of US factors based on EB posteriors re-estimated in December each year. That is, each blue dot is $E(\frac{1}{N}\sum_i \alpha^i|$ data until time $t)$ and the vertical lines are ± 2 times the posterior volatility. Triangles show average OLS alpha at each point in time, $\frac{1}{N}\sum_i \hat{\alpha}^i_{t^i_0,t}$, estimated using data through date t. The bottom panel reports the average monthly alpha for all factors in a rolling 5-year window.

plot shows the average 95% confidence interval for the average CAPM alpha among all US factors. The posterior mean alpha is relatively stable around 0.3% to 0.4% per month during our sample. And, as data evidence has accumulated over time, the confidence interval has narrowed by a third, from about 0.16% wide in 1960 to less than 0.11% in 2019.

To understand the posterior alpha, Figure 8 also shows the average OLS alpha as triangles. We see that the EB posterior is below the OLS estimate, which occurs for several reasons. First, the Bayesian posterior is shrunk toward the zero prior. Second, the Bayesian method gives more weight to factors with longer time series evidence, and these factors have lower estimated alphas. Over time, the OLS estimate moves nearer to the Bayesian posterior mean.

Naturally, good performance increases the posterior mean (as well as the OLS estimate), while poor performance pulls down the posterior. We see a general increase in the mean alpha up until the end of 1998 as the factor evidence mounts, and a decline thereafter as some factors disappoint out-of-sample.

The bottom panel in Figure 8 provides additional insight into the posterior dynamics by reporting average monthly alpha among all factors in a rolling 5-year window. Consistent with Figure 7, it shows that average factor performance has weakened over the past decade. Changes in the posterior mean are obviously much smaller than the variation in performance. This is because, while the out-of-sample OLS alphas have been notably lower than in-sample for many factors, this lower performance is partly expected by a Bayesian. Hence, we see only a moderate updating as Bayesian beliefs have been mostly confirmed throughout the sample.

A simple example helps develop intuition for this phenomenon. Suppose a researcher has $T = 10$ years of data for factors with an OLS alpha estimate of $\hat{\alpha} = 10\%$ standard error $\sigma/\sqrt{T}$. Further, assume their zero-alpha prior is equally as informative as their 10-year sample (i.e., $\tau = \sigma/\sqrt{T}$). Then the shrinkage factor is $\kappa = 1/2$ using equation (5). So, after observing the first ten years with $\hat{\alpha} = 10\%$, the Bayesian expects a future alpha of $E(\alpha|\hat{\alpha}) = 5\%$ (equation (4)). What happens if this Bayesian belief is confirmed by additional data, namely that the factor realizes an alpha of 5% over the next 10 years? In this case, the full-sample OLS of alpha is $\hat{\alpha} = 7.5\%$, but now the shrinkage factor becomes $\kappa = 2/3$ because the sample length doubles, $T = 20$. This results in a posterior alpha of $E(\alpha|\hat{\alpha}) = 7.5\% \cdot 2/3 = 5\%$. Naturally, when beliefs are confirmed by additional data, the posterior mean does not change. Nevertheless, we learn something from the additional data, because our conviction increases as the posterior variance is reduced. If $\sigma = 0.1$, the posterior volatility $\sqrt{\text{Var}(\alpha|\hat{\alpha})} = \sigma\sqrt{\frac{\kappa}{T}}$ goes from 2.2% with 10 years of data to 1.8% with 20 years of data, and the confidence interval, $[E(\alpha|\hat{\alpha}) \pm 2\sqrt{\text{Var}(\alpha|\hat{\alpha})}]$, is reduced from $[0.5\%, 9.5\%]$ to $[1.3\%, 8.7\%]$.

3.3 Bayesian Multiple Testing

A great advantage of Bayesian methods for tackling challenges in multiple testing is that, from the posterior distribution, we can make explicit probability calculations for essentially any inferential question. We use our EB posterior to investigate the false discovery and family-wise error rates in our data. We define a false discovery as a factor where we claim that the alpha is positive, but where the true alpha is negative.[39]

First, based on Proposition 5, we calculate the Bayesian FDR in our sample as the average p-value among all discoveries. The p-values are based on the EB posterior using the world factors. In particular, we find $\text{FDR}^{\text{Bayes}} = 0.12\%$, meaning that we expect roughly one discovery in 1000 to be a false positive given our Bayesian hierarchical model estimates. In other words, the model generates a highly conservative MT adjustment in the sense that once a factor is found to be significant, we can be very confident the effect is genuine.

We can also use the posterior to make other inference calculations. We compute the FWER, which we define as the probability of at least one false discovery. We simulate 1,000,000 draws of the 153 × 1 vector of alphas from the EB posterior and compute

$$\text{FWER}^{\text{Bayes}} = \frac{1}{1,000,000} \sum_{s=1}^{1{,}000{,}000} 1_{\{n_s \geq 1\}} = 11.96\%$$

where n_s is the number of false discoveries in simulation s. In other words, the probability of at least one alpha having the wrong sign is 11.96%. The $\text{FWER}^{\text{Bayes}}$ is naturally much higher than the $\text{FDR}^{\text{Bayes}}$ given the extreme conservatism built into the FWER's definition of false discovery. Whether it is too high is subjective. A nice aspect of our approach is that a researcher can control the $\text{FWER}^{\text{Bayes}}$ as desired. For example, using a t-statistic threshold of 2.78 rather than 1.96 leads to $\text{FWER}^{\text{Bayes}} = 2.1\%$.

Economic Benefits of More Powerful Tests

MT adjustments should ultimately be evaluated based on whether they lead to better decisions. It is important to balance the relative costs of false positives versus false negatives, and the appropriate tradeoff depends on the context of the problem (Greenland and Hofman, 2019). We apply this general principle in our context by directly measuring costs in terms of investment performance.

Specifically, we can compute the difference in out-of-sample investment performance from investing using factors chosen with different methods. We compare two alternatives. One is

[39] In particular, we define a discovery as a factor for which the posterior probability of the true alpha being negative is less than 2.5%. With this definition, 118 out of 153 world factors are discoveries.

the BY decision rule advocated by Harvey et al. (2016), which is a frequentist MT method that successfully controls false discoveries relative to OLS, but in doing so sacrifices power (the ability to detect true positives). The second alternative is our EB method, whose false discovery control typically lies somewhere between BY and unadjusted OLS. EB uses the data sample itself to decide whether its discoveries should behave more similarly to BY or to unadjusted OLS.

For investors, the optimal decision rule is the one that leads to the best performance out-of-sample. For the most part, the set of discovered factors for BY and EB coincide. It is only in marginal cases where they disagree which, in our sample, occurs when EB makes a discovery that BY deems insignificant. Therefore, to evaluate MT approaches in economic terms, we track the out-of-sample performance of factors included by EB but excluded by BY. If the performance of these is negative on average, then the BY correction is warranted and preferred by the investor.

We find that the out-of-sample performance of factors discovered by EB but not BY is positive on average and highly significant. The monthly alpha for these marginal cases is 0.42% per month among US factors ($t = 4.2$).[40] This estimate suggests that the BY decision rule is too conservative. An investor using the rule would fail to invest in factors that subsequently have a high out-of-sample return.

Another way to see this comes from the connection between Sharpe ratio and t-statistics: $t = \mathrm{SR}\sqrt{T}$. If we have a factor with an annual Sharpe ratio of 0.5, an investor using the 1.96 cutoff would in expectation invest in the factor after 15 years. An investor using the 2.78 cutoff, would not start investing until observing the factor for 31 years.

Unobserved Factors: Addressing Publication Bias

A potential concern with our replication rate is that the set of factors that make it into the literature is a selected sample. In particular, researchers may have tried many different factors, some of which are observed in the literature, while others are unobserved because they never got published. Unobserved factors may have worse average performance if poor

[40] For the developed ex. US sample, the monthly alpha for marginal cases is 0.28% per month ($t = 4.3$), and for the emerging sample it is 0.34% ($t = 3.4$), in favor of the EB decision rule. Appendix Table C.2 reports additional details for this analysis.

performance makes publication more difficult or less desirable. Alternatively, unobserved factors could have strong performance if people chose to trade on them in secret rather than publishing them. Either way, we next show how unobserved factors can be addressed in our framework.

The key insight is that the performance of factors across the universe of observed and unobserved factors is captured in our prior parameters τ_c, τ_w. Indeed, large values of these priors correspond to a large dispersion of alphas (that is, a lot of large alphas "out there") while small values means that most true alphas are close to zero. Therefore, smaller τ's lead to a stronger shrinkage toward zero for our posterior alphas, leading to fewer factor "discoveries" and a lower replication rate. Figure 9 shows how our estimated replication rate depends on the most important prior parameter, τ_c, based on the τ_w that we estimated from the data.[41]

In Figure 9, we show how the replication rate varies with τ_c in precise quantitative terms. Note that while the replication rate indeed rises with τ_c, the differences are small in magnitude across a large range of τ_c values, demonstrating robustness of our conclusions about replicability.

This stable replication rate in Figure 9 also suggests that the replication rate among the observed factors would be similar even if we had observed the unobserved factors. The figure highlights several key values of τ_c: Both the value of τ_c that we estimated from the observed data (as explained in Appendix B) and values that adjust for unobserved data in different ways.

We adjust τ_c for unobserved factors as follows. We simulate a data set that proxies for the full set of factors in the population (including those unobserved), and then estimate the τ's that match this sample. One set of simulations is constructed to match the baseline scenario of Harvey et al. (2016) (Table 5.A, row 1), which estimates that researchers have tried $M = 1,297$ factors, of which 39.6% of have zero alpha and the rest have a Sharpe ratio of 0.44. We also consider the more conservative scenario of Harvey et al. (2016) (Table 5.B, row 1), which implies that researchers have tried $M = 2,458$ factors, of which 68.3% have zero alpha. The appendix has more details on these simulations. The result, as seen

[41] Figure C.4 in the appendix shows that the results are robust to alternative values of τ_w.

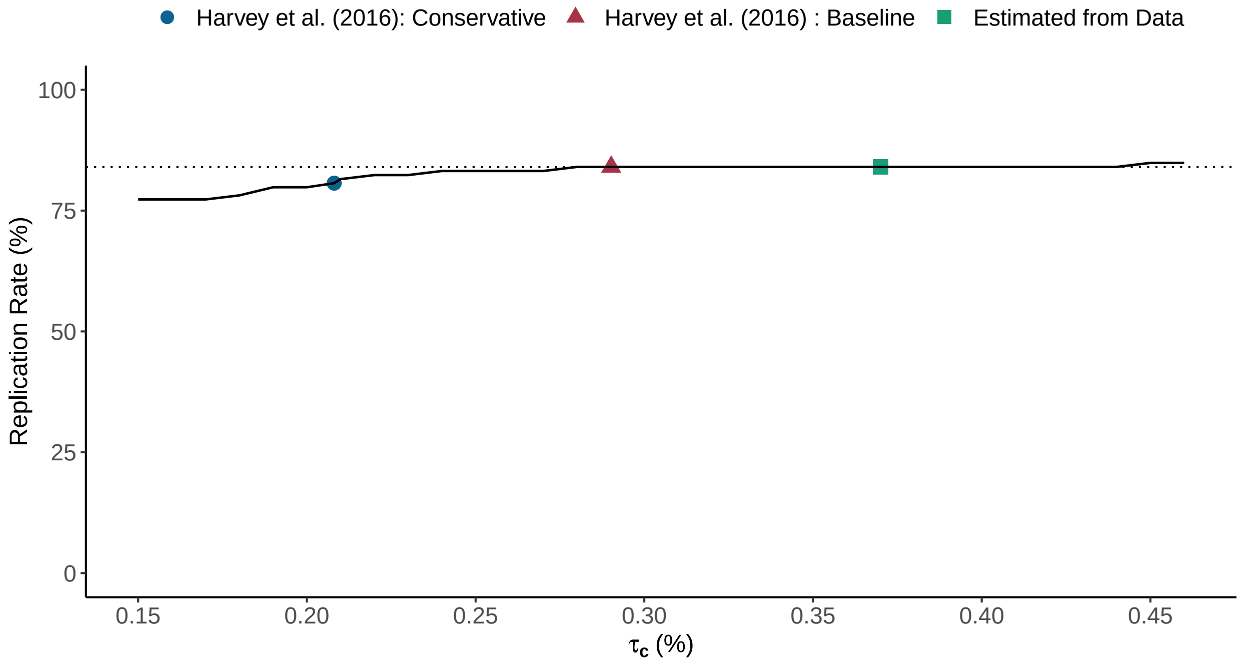


Figure 9: Replication Rate with Prior Estimated in Light of Unobserved Factors

Note: The figure shows how the replication rate in the US varies when changing the τ_c parameter. The τ_w parameter is fixed at the estimate value of 0.21%. The dotted line shows our replication rate of 84%. The green square, highlights the value estimated in the data $\tau_c = 0.37\%$. The red triangle and the blue circle highlights values that are found by estimating the empirical Bayes model according to assumptions about unobserved factors from Harvey et al. (2016). The values are $\tau_c = 0.29\%$ in the baseline scenario and $\tau_c = 0.21\%$ in the conservative scenario. A description of this approach can be found in the appendix, section A.

in Figure 9, is that values of τ_c that correspond to these scenarios from Harvey et al. (2016) still lead to a conclusion of a high replication rate in our factor universe.

3.4 Economic Significance of Factors

Which factors (and which themes) are the most impactful anomalies in economic terms? We investigate this question by identifying which factors matter most from an investment performance standpoint.

Figure 10 shows the alpha confidence intervals for all world factors, sorted by the median posterior alpha within clusters. This illustration is similar to Figure 3, but now we focus on the world instead of the US factors, and here we sort factors into clusters. We also focus on factors that the original studies conclude are significant. We see that world factor alphas

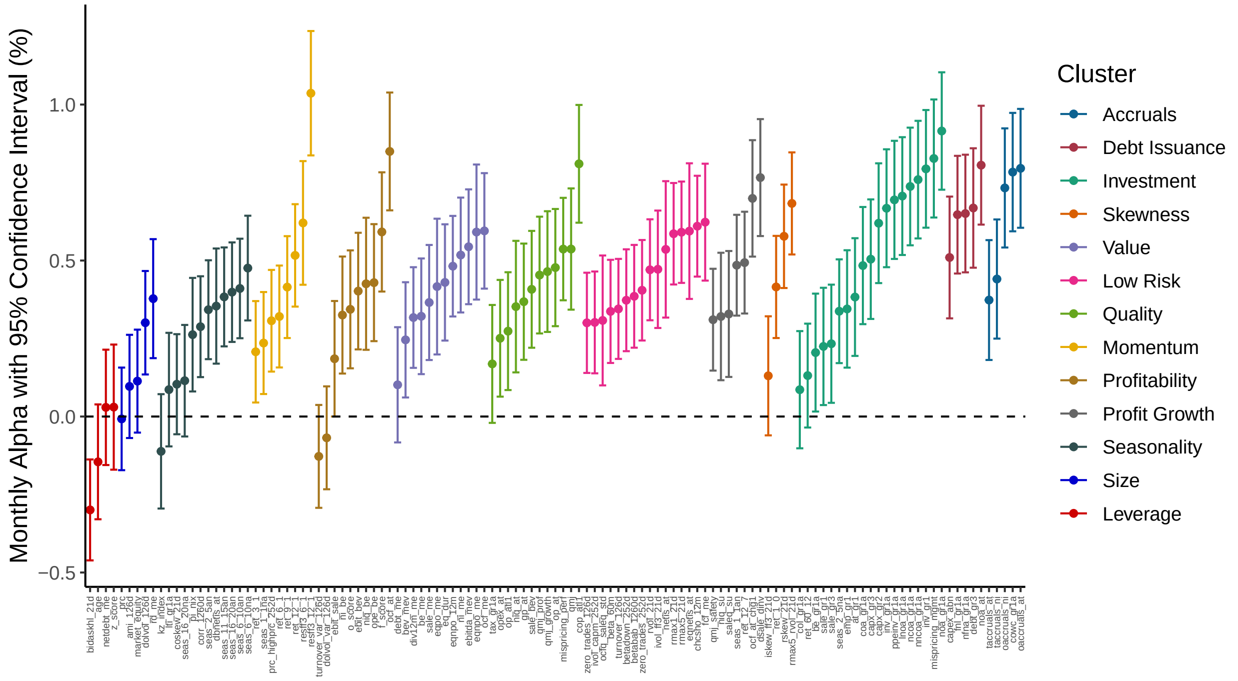


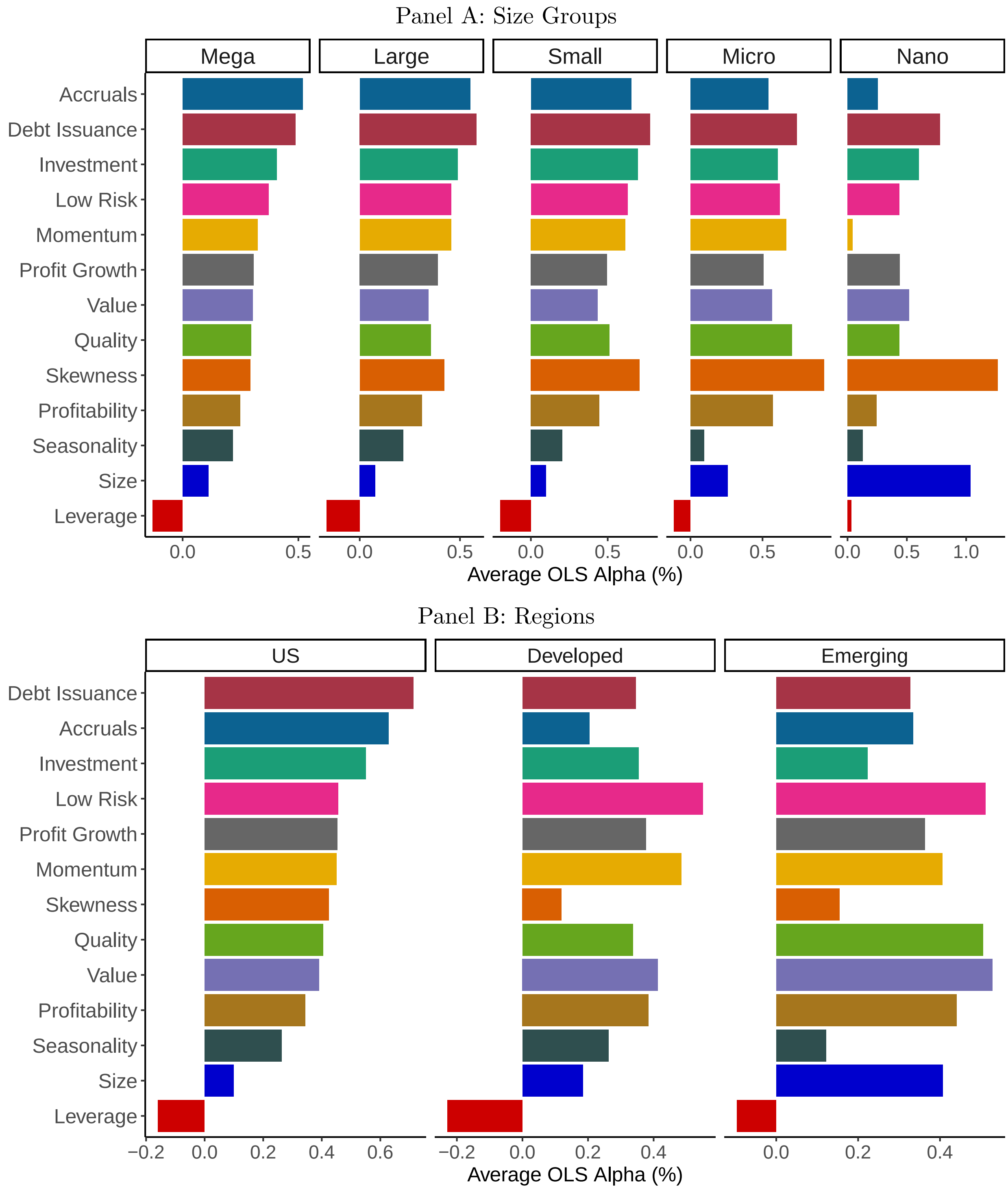
Figure 10: World Alpha Posterior By Factor and Cluster

Note: The figure reports the EB posterior 95% confidence interval for the true alpha of a world factor create as a capitalization weighted average of all country specific factors in our dataset. We only include factors that the original paper finds significant.

tend to be economically large, often above 0.3% per month, and tend to be highly significant, in most clusters. The exception is the leverage cluster, where we also saw a low replication rate in preceding analyses.

By Region and By Size

We next consider which factors are most economically important across global regions and across stock size groups. In Panel A of Figure 11, we construct factors using only stocks in the five size subsamples presented earlier in Figure 4; namely mega, large, small, micro, and nano stock samples. For each sample, we calculate cluster-level alphas as the equal-weighted average alpha of factors within the cluster. We see, perhaps surprisingly, that the ordering and magnitude of alphas is broadly similar across size groups. The Spearman rank correlation of alphas for mega caps versus micro caps is 73%. Only the nano stock sample, defined as stocks below the 1^{st} percentile of the NYSE size distribution (which amounted to 1051 out of 5256 stocks in the US at the end of 2019), exhibits notable deviation from the



Panel A: Size Groups

Panel B: Regions

Figure 11: Alphas By Geographic Region and Stock Size Group

Note: The figure reports average cluster-level alphas for factors formed from subsamples defined by different stock market capitalization groups (Panel A) and regions (Panel B).

other groups. The Spearman rank correlation between alphas of mega caps and nano caps is 35%.

Panel B of Figure 11, shows cluster-level alphas across regions. Again, we find some consistency in alphas across the globe, with the obvious standout being the size theme, which is much more important in emerging markets than in developed markets. US factor alphas share a 71% Spearman correlation with the developed ex. US sample, and a 48% correlation with the emerging markets sample.

Controlling for Other Themes

We have focused so far on whether factors (or clusters) possess significant positive alpha relative to the market. The limitation of studying factors in terms of CAPM alpha is that it does not control for duplicate behavior other than through the market factor. Economically important factors are those that have large impact on an investor's overall portfolio, and this requires understanding which clusters contribute alpha while controlling for all others.

To this end, we estimate cluster weights in a tangency portfolio that invests jointly in all cluster-level portfolios. We test the significance of the estimated weights using the method of Britten-Jones (1999). In addition to our 13 cluster-level factors, we also include the market portfolio as a way of benchmarking factors to the CAPM null. Lastly, we constrain all weights to be non-negative (because we have signed the factors to have positive expected returns according to the findings of the original studies).

Figure 12 reports the estimated tangency portfolio weights and their 90% bootstrap confidence intervals. When a factor has a significant weight in the tangency portfolio, it means that it matters for an investor, even controlling for all the other factors. We see that all but three clusters are significant in this sense. We also see that conclusions about cluster importance change when clusters are studied jointly. For example, value factors become stronger when controlling for other effects because of their hedging benefits relative to momentum, quality, and leverage. More surprisingly, the leverage cluster becomes one of the most heavily weighted clusters, in large part due to its ability to hedge value and low risk factors. The hedging performance of value and leverage clusters is clearly discernible in Appendix table C.3, which shows the average pairwise correlations among factors within

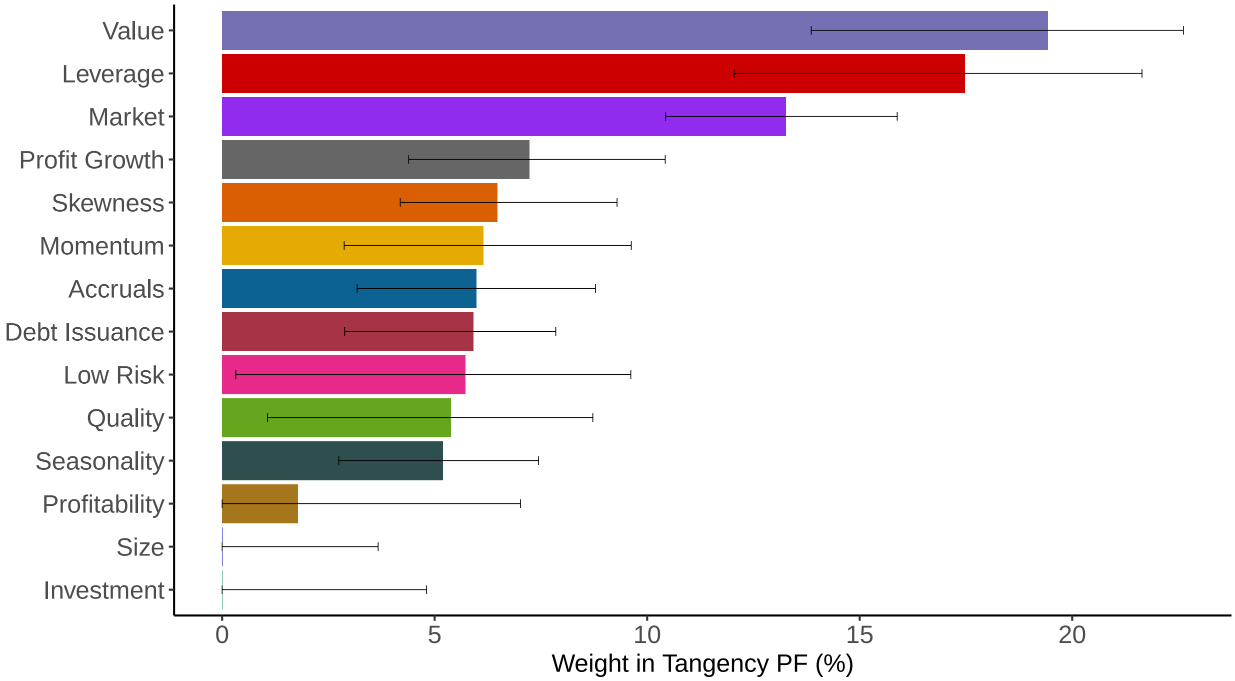


Figure 12: Tangency Portfolio Weights

Note: The return are from the world portfolio. We compute the cluster return as the equal weighted return of all factors with data available at a given point in time. We further add the Global market return. We estimate the tangency weights following the method of Britten-Jones (1999) with a non-negativity constraint. The error bars are the 90% confidence intervals based on 10,000 bootstrap samples and the percentile method. The data starts in 1952 to ensure that all cluster have non-missing observations.

and across clusters.[42]

4 Conclusion: Finance Research Posterior

We introduce a hierarchical Bayesian model of alphas that emphasizes the joint behavior of factors and provides an alternative, and evidently more powerful, multiple testing adjustment than common frequentist methods. Based on this framework, we re-visit the evidence on replicability in factor research and come to substantially different conclusions versus the prior literature. We find that US equity factors have a high degree of internal validity in the sense that over 80% of factors remain significant after modifications in factor construction that make all factors consistent, more implementable, while still capturing the original signal (Hamermesh, 2007) and after accounting for multiple testing concerns (Harvey et al., 2016;

[42] Appendix Tables C.10 and C.11 show how tangency portfolio weights vary by region and by size group.

Harvey, 2017).

We also provide new evidence demonstrating a high degree of external validity in factor research. In particular, we find highly similar qualitative and quantitative behavior in a large sample of 153 factors across 93 countries as we find in the US. We also show that, within the US, factors exhibit a high degree of consistency in their behavior between their published in-sample periods and in out-of-sample data not considered in the original studies. We show that some out-of-sample factor decay is to be expected in light of Bayesian posteriors based on publication evidence. Therefore, the new evidence from post-publication data largely confirms the Bayesian's beliefs, which has led to relatively stable Bayesian alpha estimates over time.

In addition to providing a powerful tool for replication, our Bayesian framework has several additional applications. For example, the model can be used to correctly interpret out-of-sample evidence, look for evidence of alpha-hacking, compute the expected number of false discoveries and other relevant statistics based on the posterior, analyze portfolio choice taking into account both estimation uncertainty and return volatility, and evaluate asset pricing models.

Finally, the code, data, and meticulous documentation for our analysis are available online. Our large global factor data set and the underlying stock-level characteristics are easily accessible to researchers by using our publicly available code and its direct link to WRDS. We are maintaining a living code and database, updated regularly with the new data releases and code improvements. We hope that our methodology and data will help promote credible finance research.

References

Acharya, V. and L. H. Pedersen (2005). Asset pricing with liquidity risk. *Journal of Financial Economics 77*, 375–410.

Amihud, Y. and H. Mendelson (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics 17*, 223–249.

Asness, C. and A. Frazzini (2013). The devil in HML's details. *The Journal of Portfolio Management 39*(4), 49–68.

Asness, C., T. Moskowitz, and L. H. Pedersen (2013). Value and momentum everywhere. *The Journal of Finance 68*(3), 929–985.

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

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*(4), 1165–1188.

Berry, D. A. and Y. Hochberg (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference 82*(1-2), 215–227.

Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal 33*(1), 108–113.

Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance 54*(2), 655–671.

Bryzgalova, S., J. Huang, and C. Julliard (2019). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Available at SSRN*.

Chen, A. Y. (2020). The limits of p-hacking: Some thought experiments. *The Journal of Finance, forthcoming*.

Chen, A. Y. and T. Zimmermann (2020). Open source cross-sectional asset pricing. *Working paper, Board of Governors of the Federal Reserve System*.

Chordia, T., A. Goyal, and A. Saretto (2020). Anomalies and false rejections. *The Review of Financial Studies 33*(5), 2134–2179.

Cochrane, J. H. (2011). Presidential address: Discount rates. *Journal of Finance 66*, 1047–1108.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Volume 1. Cambridge University Press.

Efron, B. and R. Tibshirani (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology 23*(1), 70–86.

Elton, E. J., M. J. Gruber, and J. Spitzer (2006). Improved estimates of correlation coefficients and their impact on optimum portfolios. *European Financial Management 12*(3), 303–318.

Engle, R. and B. Kelly (2012). Dynamic equicorrelation. *Journal of Business & Economic Statistics 30*(2), 212–228.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance 75*(3), 1327–1370.

Frazzini, A. and L. H. Pedersen (2014). Betting against beta. *Journal of Financial Economics 111*(1), 1 – 25.

Gelman, A. (2016, Aug). Bayesian inference completely solves the multiple comparisons problem.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. CRC press.

Gelman, A., J. Hill, and M. Yajima (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness 5*(2), 189–211.

Green, J., J. R. Hand, and X. F. Zhang (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies 30*(12), 4389–4436.

Greenland, S. and A. Hofman (2019). Multiple comparisons controversies are about context and costs, not frequentism versus bayesianism. *European journal of epidemiology 34*(9), 801–808.

Greenland, S. and J. M. Robins (1991). Empirical-bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology*, 244–251.

Hamermesh, D. S. (2007). Replication in economics. *Canadian Journal of Economics/Revue canadienne d'économique 40*(3), 715–733.

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

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

Hou, K., C. Xue, and L. Zhang (2020). Replicating anomalies. *The Review of Financial Studies 33*(5), 2019–2133.

Ilmanen, A., R. Israel, T. J. Moskowitz, A. K. Thapar, and F. Wang (2019). How do factor premia vary over time? a century of evidence. *AQR Working Paper*.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine 2*(8), e124.

Jacobs, H. and S. Müller (2020). Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics 135*(1), 213–230.

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Koijen, R. S., T. J. Moskowitz, L. H. Pedersen, and E. B. Vrugt (2018). Carry. *Journal of Financial Economics 127*(2), 197–225.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Linnainmaa, J. T. and M. R. Roberts (2018). The history of the cross-section of stock returns. *The Review of Financial Studies 31*(7), 2606–2649.

Maniadis, Z., F. Tufano, and J. A. List (2017). To replicate or not to replicate? exploring reproducibility in economics through the lens of a model and a pilot study. *The Economic Journal 127*, F209–F235.

Maritz, J. S. (2018). *Empirical Bayes methods with applications.* CRC Press.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–32.

Moskowitz, T. J., Y. H. Ooi, and L. H. Pedersen (2012). Time series momentum. *Journal of financial economics 104*(2), 228–250.

Murtagh, F. and P. Legendre (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification 31*(3), 274–295.

Newey, W. K. and K. D. West (1987, May). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica 55*(3), 703–708.

Nosek, B. A., J. R. Spies, and M. Motyl (2012). Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science 7*(6), 615–631.

Shumway, T. (1997). The delisting bias in crsp data. *The Journal of Finance 52*(1), 327–340.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*(301), 236–244.

Welch, I. (2019). Reproducing, extending, updating, replicating, reexamining, and reconciling. *Critical Finance Review 8*(1-2), 301–304.