How many factors?

Jonathan Lewellen
Tuck School of Business at Dartmouth

June 2022

Tel: 603-646-8650; email: jon.lewellen@dartmouth.edu. I am grateful to Ken French and Juhani Linnainmaa for helpful comments and suggestions.

How many factors?

Abstract

This paper studies the costs and benefits of adding factors to empirical asset-pricing models. I argue that, for many purposes, the literature's preference for models with fewer factors is misplaced. Including extra factors in a model, even redundant ones, can improve estimates of individual alphas and increase the power of asset-pricing tests. I provide empirical examples to illustrate these results.

In recent decades, asset-pricing studies have explored the link between expected stock returns and hundreds of firm characteristics (Harvey, Liu, and Zhu 2016). The proliferation of findings has, in turn, inspired a renewed effort to identify a small number of factors that can explain the cross-section of expected returns and new methods of comparing different factor models (e.g., Hou, Xue, and Zhang 2015; Fama and French 2015, 2018; Barillas and Shanken 2017, 2018; Hou, Mo, Xue, and Zhang 2019; Barillas, Kan, Robotti, and Shanken 2020; Kozak, Nagel, and Santosh 2020; Harvey and Liu 2021). The spirit of this literature is summarized well by Barillas and Shanken (2017): "Given the variety of portfolio-based factors that have been examined by researchers, it is important to understand how best to combine them in a parsimonious asset pricing model for expected returns, one that excludes redundant factors" (p. 715). Similarly, Fama and French (2018) note that "if factor modeling is not to degenerate into meaningless dredging for the ex post MVE [mean-variance-efficient] portfolio, the number of factors in models must also be limited. Establishing ground rules, however, awaits more experience" (p. 248).

In this paper, I offer a contrarian perspective on the literature. I consider two related questions: (i) What are the costs and benefits of dropping factors from a model? (ii) How many factors can a model have, i.e., how many factors are too many? My results lead me to conclude that, for many purposes and within fairly generous bounds, the benefits of including additional factors exceed the costs. It follows that searching for a parsimonious model with just a handful of factors can actually be counterproductive. By extension, questions like "Which investment factor is best?" or "Do size, profitability, and investment subsume the value premium?" are often just a sideshow.

To frame the analysis, we need to have an objective function in mind: Why do we want a factor model? What will it be used for? These questions could be answered in a variety of ways. My working assumption in this paper is that the underlying goal is either (i) to measure abnormal returns (alphas) on stocks, portfolios, or mutual funds, or (ii) to estimate properties of the minimum-variance stochastic discount factor (SDF) and, in particular, the volatility of the SDF. The main point of my paper is that, with these goals in mind, there may be little benefit from reducing the number of factors in a model; in fact, dropping even redundant factors, that do not contribute to the SDF, can be suboptimal.

My analysis has three parts. The first part focuses on estimates of individual alphas. Consider a typical time-series regression of asset i's excess returns in month t, R_{it}, on a set of factors F_t, the elements of which are either excess returns or long-short returns on traded portfolios:

$$R_{it} = \alpha_i + \beta_i' F_t + \varepsilon_{it}. \quad (1)$$

The sampling variance of the estimated alpha is, from standard results,

$$\text{var}(a_i) = (1 + sh^2(F))/T \times \text{var}(\varepsilon_i), \quad (2)$$

where $sh^2(F)$ is the sample maximum squared Sharpe ratio of the factors and T is the number of months. The costs and benefits of including an extra factor in the regression are immediate. First, if the factor is not redundant—it has a nonzero alpha, in population, when regressed on the other factors—including the factor in (1) will generally affect the asset's true alpha (unless the asset's loading on the factor is zero). On the other hand, if the factor is redundant, including it has no impact on α_i but will affect the standard error of the estimate through $sh^2(F)$ or $\text{var}(\varepsilon_i)$. The first effect is necessarily bad (including the factor cannot lower $sh^2(F)$ and, hence, will generally increase the standard error) while the second effect is necessarily good (including the factor cannot increase $\text{var}(\varepsilon_i)$ and, hence, will generally reduce the standard error). However, the first effect is likely to be small because squared Sharpe ratios are typically small numbers (e.g., the market's squared Sharpe ratio is around 0.01 monthly), and a redundant factor, by definition, is not expected to contribute significantly to $sh^2(F)$. The second effect, on $\text{var}(\varepsilon_i)$, can be either large or small depending on how highly correlated the asset is with the factor. It follows that including even a redundant factor in the regression can be useful because it might well improve the precision of the alpha. On balance, I argue that we should err on the side of including more, not fewer, factors.

It might be useful to expand on this point. The essence of the argument is that, while including a redundant factor in the regression does not affect the true alpha, it can help to estimate the alpha more accurately. For example, suppose the redundant factor is uncorrelated with the other factors and, therefore, has a true mean of zero. In sample, however, the factor's average return is 0.5% monthly. That information will help us estimate the alpha of any asset correlated with the factor. If an asset's loading on the factor is, say, 2.0, one percentage point of the asset's return can be attributed to the factor's unexpected return during the sample, and including

the factor in the regression appropriately reduces the estimated alpha by that amount. Including the factor in the regression is only detrimental if estimation error in the factor loading is so high that it increases the noise in alpha, an effect captured by the term $sh^2(F)$. An interesting corollary is that, as the sampling frequency of returns increases (time intervals get shorter), sampling error in factor loadings and the impact on $sh^2(F)$ both go to zero, so including redundant factors is always a net positive in continuous time if they are correlated with the test asset.

The argument here has implications beyond the analysis of factor models. For example, suppose we estimate the alpha of a mutual fund that tilts toward a particular industry or industries. The convention in asset pricing would be to regress the mutual fund return on a small set of asset-pricing factors, effectively ignoring the industry tilt because industries are not thought of as priced factors. My analysis implies this logic is faulty: Even if industry returns are not priced factors, adding industry returns to the regression can help absorb residual variation and will improve the estimate of alpha. More generally, there is no inherent reason to include only priced factors in the regression.

The second part of my analysis focuses on the properties of asset-pricing tests: How well do factors explain the cross-section of expected returns? Barillas and Shanken (2017) observe that, to compare models with different factors, the key question is which set of factors has a higher squared Sharpe ratio. Suppose we have factor models F_1 and F_2, and our goal is to understand which produces smaller alphas, specifically, a smaller value of $\alpha'\Sigma^+\alpha$, where α is the vector of alphas for all assets (test assets R and all factors in F_1 or F_2) and Σ^+ is the pseudoinverse of the residual covariance matrix for a given model. (The pseudoinverse is needed here because some factors are included as both 'right-hand-side' (RHS) and 'left-hand-side' (LHS) assets, implying the residual covariance matrix has rows and columns of zero; the quadratic is the same if the RHS factors are dropped from the LHS assets.) Gibbons, Ross, and Shanken (1989) show that $\alpha'\Sigma^+\alpha$ measures the difference between the population maximum squared Sharpe ratio of all assets, $SH^2(R_{all})$, and the maximum squared Sharpe ratio of the factors included in the model, $SH^2(F_i)$ (uppercase 'SH' denotes a population Sharpe ratio). Thus, comparing models:

$\alpha'\Sigma^{+}\alpha$ for model 1 = $SH^2(R_{all}) - SH^2(F_1)$. (3)

$\alpha'\Sigma^{+}\alpha$ for model 2 = $SH^2(R_{all}) - SH^2(F_2)$. (4)

Since the first terms in the equations are the same, it is immediate that the better model is the one with factors that produce the higher squared Sharpe, $SH^2(F_i)$. The test assets themselves are irrelevant. This idea is consistent with the recent trend of comparing models by testing whether the factors in one model explain ('span') the factors in the other model.

In the case of two nested models, suppose that F_2 includes the factors in F_1 plus additional factors that turn out to be redundant: the extra factors have insignificant alphas when regressed on F_1 and, therefore, $sh^2(F_2)$ is insignificantly different from $sh^2(F_1)$. The implication is that the smaller set of factors F_1 capture all of the pricing information and the extra factors in F_2 can be dropped.

But what can we do with the information that F_1 works as well as F_2? If we want to test whether F_1 explains the cross-section of expected returns, it will be tempting to run the usual test of whether

$SH^2(R,F_1) - SH^2(F_1) = 0$. (5)

The problem is that F_1 can appear to be a good model based on (5) even if it does not appear to be a good model based on (3). Put differently, the test assets in R and the extra factors in F_2 might, individually, have alphas that are insignificantly different from zero even if putting R and F_2 together into a test leads to a strong rejection of the model. Thus, if the question is "Does F_1 explain the expected returns on all assets?", the extra factors in F_2 cannot be dropped from the tests. In a sense, the factors need to be included as either RHS or LHS assets. One of my key results is that including them on the RHS, in the model, always leads to more powerful tests.

Let me offer a concrete example. Suppose we have two factors, R_1 and R_2, and a third test asset R_3. R_2 and R_3 have the same volatility, are uncorrelated with R_1, and are negatively correlated with each other. (R_1 could be the market portfolio, R_2 a value factor, and R_3 a momentum factor.) R_2 and R_3 also have the same insignificant alpha when regressed on R_1. Thus, in the first step, when we compare model $F_1 = R_1$ with model $F_2 = (R_1, R_2)$,

R_2 appears to be redundant and can be dropped. In the second step, F_1 also appears to explain the returns on R_3. At the same time, it is clear that the portfolio ½ R_2 + ½ R_3 could have a highly significant alpha when regressed on R_1 since its alpha is the same as the (common) alpha of R_2 and R_3 but has a lower standard error. This 'anomaly' will only be discovered if R_2 is included either as a test asset in the second step or as a factor in the model, i.e., R_2 cannot be dropped from the tests even though it appears to be redundant. And I show that power is always higher when R_2 is included in the model. The implication, again, is that we should err on the side of including more, not fewer, factors.

The third part of my analysis focuses on estimating the variance of the SDF or, equivalently, the maximum squared Sharpe ratio attainable from a given set of factors. This metric is often used as a summary measure of a model's performance. I study how well it can be estimated as a function of the number of factors in the model. Given a set of K normally distributed factors, the sample statistic

$$p = sh^2(F) \times (T–K)/K \tag{6}$$

has a noncentral F distribution with degrees of freedom K and T–K and noncentrality parameter $T \times SH^2(F)$ (e.g., Morrison 1990). I describe how to use these facts to obtain a confidence interval for the population $SH^2(F)$ and then study how the confidence interval varies as we add additional factors, depending on whether the additional factors are redundant or not.

In this case, including redundant factors tends to be detrimental: sampling error in $sh^2(F)$ increases, and the expected width of the confidence interval for $SH^2(F)$ goes up, if redundant factors are added to the model. However, the impact on the confidence interval is surprisingly small for typical sample sizes encountered in the literature. For example, with 40 years of data and a true monthly $SH^2(F)$ of 0.05, a 90% confidence interval is expected to be [0.021, 0.091] if we have four factors, [0.019, 0.094] if we have 12 factors, and [0.012, 0.101] if we have 32 factors.

Adding non-redundant factors is more complicated because the factors affect not only the sampling error in $sh^2(F)$ but also the true $SH^2(F)$. Intuitively, adding priced factors makes the confidence interval wider but

shifts it upward toward the true value of $SH^2(F)$. For example, suppose we start with four factors that have a $SH^2(F)$ of 0.04 and add eight factors that increase $SH^2(F)$ to 0.06. With 40 years of data, a 90% confidence interval is expected to be [0.015, 0.077] using the first four factors and [0.026, 0.107] using all 12 factors. The latter is wider but centered closer to the true $SH^2(F)$ of the full model.

Overall, my results suggest that, from an empirical standpoint, the benefits of including extra factors in a model may well outweigh the costs, even if the factors are redundant. Models with many factors can work better for many applications than parsimonious models with only a few factors. There is, of course, a limit to this argument—the marginal benefits of adding factors are likely to go down and the costs are likely to go up as the number increases. However, my results suggest the optimal number may be much larger—perhaps an order of magnitude greater—than the current fashion in the literature.

1. Estimating alphas

One important use of factor models is to estimate alphas on stocks, portfolios, or mutual funds. We might want to test whether a stock performs well on a 'risk-adjusted' basis, whether a proposed trading strategy generates abnormal profits relative to existing strategies, or whether a mutual fund manager has stock-picking skill after adjusting for factor tilts. In this section, I study how the estimation of alpha changes as we increase the number of factors in the model.

An asset's alpha is estimated in the time-series regression of the asset's excess returns on a set of K factors, assumed to be either excess returns or long-short returns on traded portfolios:

$$R = \alpha + \beta'F + \varepsilon. \quad (7)$$

I omit subscripts here, but R and F should be interpreted as excess returns in month t. The sampling variance of the estimated alpha is, from standard results (e.g., Gibbons, Ross, and Shanken 1989),

$$\text{var}(a) = (1 + sh^2(F))/T \times \text{var}(\varepsilon), \quad (8)$$

where T is the number of months and $sh^2(F)$ is the sample maximum squared Sharpe ratio of the factors (in this calculation, the sample variance does not include a degree-of-freedom adjustment; I assume $T > K+5$). The

main question here is how the alpha in (7) and the sampling variance in (8) change if we include or exclude a factor from the model.

For concreteness, suppose we add one factor F_{K+1} to the model. The impact is easiest to evaluate if we first orthogonalize F_{K+1} with respect to the other factors. The orthogonalized factor, OF_{K+1}, equals the intercept, α_{FK+1}, plus the residual, ε_{FK+1}, when F_{K+1} is regressed on F. (F_{K+1} is 'priced' if $\alpha_{FK+1} \neq 0$ and 'redundant' if $\alpha_{FK+1} = 0$.) Adding OF_{K+1} to the model, the regression becomes:

$$R = \alpha^* + \beta' F + \beta_{K+1} OF_{K+1} + \varepsilon^*, \tag{9}$$

where $\alpha^* = \alpha - \beta_{K+1}\alpha_{FK+1}$ and $\varepsilon^* = \varepsilon - \beta_{K+1}\varepsilon_{FK+1}$. ($\alpha^*$ and ε^* are the same if F_{K+1} is added to the regression instead.) The impact on alpha depends on how much of the asset's expected return is explained by the new factor (the loading β_{K+1} times the mean of OF_{K+1}), while the new residual excludes variation in the asset's return that is explained by OF_{K+1}. The variance of the estimate of α^* is

$$\text{var}(a^*) = (1 + sh^2(F,F_{K+1}))/T \times \text{var}(\varepsilon^*) \tag{10}$$

or

$$\text{var}(a^*) = [1 + sh^2(F) + sh^2(OF_{K+1})]/T \times [\text{var}(\varepsilon) - \beta_{K+1}{}^2\text{var}(OF_{K+1})]. \tag{11}$$

In this equation, $sh^2(OF_{K+1})$ should be interpreted as the squared Sharpe ratio of the in-sample orthogonalized version of F_{K+1}. The first substitution in (11), compared to (10), follows from basic properties of maximum squared Sharpe ratios for orthogonal assets (e.g., MacKinlay 1995) and the second substitution follows from the fact that ε^* is uncorrelated with OF_{K+1}.

Equations (7)–(11) allow us to evaluate the trade-offs associated with adding a factor to the model. First, if the factor is not redundant ($\alpha_{FK+1} \neq 0$) and the asset loads on the factor ($\beta_{K+1} \neq 0$), the true alpha changes because the benchmark for evaluating the asset's abnormal return changes, i.e., part of the asset's return is attributed to its loading on the new factor. That, of course, is the whole point of including a priced factor in the regression.

On the other hand, if the new factor is redundant ($\alpha_{FK+1} = 0$), adding it to the regression has no effect on the true alpha ($\alpha^* = \alpha$), so the main consideration comes from the impact on the standard error of the estimate.

The latter depends on two effects, the change in the sample maximum squared Sharpe ratio of the factors and the impact on the residual variance. Consider each in turn:

Squared Sharpe ratio. From (11), the change in the maximum squared Sharpe ratio equals $sh^2(OF_{K+1})$. This term arises because estimation error in the factor loading on F_{K+1} increases estimation error in alpha (it would drop out if β_{K+1} were known). The term is expected to be small because monthly Sharpe ratios are typically small numbers. For example, the market portfolio's Sharpe ratio is around 0.11 (excess return around 0.5% and standard deviation around 4.5%), so adding an uncorrelated factor that has twice the market's sample Sharpe ratio would add only $sh^2(OF_{K+1}) = 0.05$ to the first term in (11). And, since $1+sh^2(F)$ cannot be less than one, the percentage increase in the sampling variance is even smaller.

More formally, if the factor is redundant, $sh^2(OF_{K+1})$ is expected to be close to zero because the population $SH^2(OF_{K+1})$ is exactly zero. Gibbons, Ross, and Shanken's (1989) results imply that, if returns are normally distributed, the statistic

$$g = sh^2(OF_{K+1}) \times (T–K–1)/(1+sh^2(F)) \tag{12}$$

has a central F distribution with degrees of freedom 1 and T–K–1. This implies that the percentage increase in the first term in (11), $sh^2(OF_{K+1})/(1+sh(F))$, has a mean of 1/(T–K–3) and standard deviation of $[2(T–K–2)/(T–K–5)]^{1/2}/(T–K–3)$ (e.g., Mood, Graybill, and Boes 1974). As long as the number of factors K is substantially less than the length of the time-series T, the percentage increase should be close to zero. For example, with 40 years of data, the expected increase is 0.2% if K = 25 and 0.3% if K = 100.

These results generalize easily if we consider adding multiple factors simultaneously to the model. Define $sh^2(OF_{K+q})$ as the sample maximum squared Sharpe ratio of orthogonalized factors F_{K+1} through F_{K+q} (orthogonalized relative to F; I assume T > K+q+4). If returns are normally distributed and the extra factors are all redundant, the statistic

$$g = sh^2(OF_{K+q})/q \times (T–K–q)/(1+sh^2(F)) \tag{13}$$

has a central F distribution with degrees of freedom q and T–K–q. This implies that the percentage increase in

the first term in (11), $sh^2(OF_{K+q})/(1+sh(F))$, has a mean of $q/(T–K–q–2)$ and standard deviation of $[2q(T–K–2)/(T–K–q–4)]^{1/2}/(T–K–q–2)$. If we have 40 years of data and start with 10 factors, the expected increase in the variance of alpha coming from this effect is 2% if add 10 redundant factors to the model and 4.5% if we add 20 redundant factors. (The impact on the standard error of alpha is roughly half as large.) These results suggest that, in the worst case scenario—the extra factors are all redundant and do not absorb any residual variation in the test asset's return—adding even dozens of factors to the model has only a modest effect on our ability to estimate alpha precisely (unless K approaches T).

Of course, we cannot know whether the extra factors are truly redundant. If the determination is made in sample, the results above may actually overstate the impact of adding redundant factors since the factors will have already been found to contribute insignificantly to the model's squared Sharpe ratio, i.e., $sh^2(OF_{K+q})$ would be insignificantly different from zero. Thus, in practice, the impact of adding redundant factors may be less than the ex ante expected value of $sh^2(OF_{K+q})$.

Residual variance. The second effect on the sampling variance of alpha arises if the extra factors absorb residual variation in the asset's return, an effect that is necessarily beneficial regardless of whether the factors are redundant or not. This effect, via the second bracketed term in (11), is hard to quantify in general terms because it depends on how highly correlated the asset is with the factors; that is, it depends on how much the factors raise the regression R^2. The important point is that even a non-priced factor can help us to estimate alpha more precisely.

Examples. Table 1 illustrates these results. I estimate alphas for momentum, profitability, and asset-growth deciles using several factor models: (i) the CAPM; (ii) the Fama-French (1993) three-factor model; (iii) an expanded version of the Fama-French model that directly uses the six underlying size-B/M portfolios (thus, a six-factor model); (iv) a second extension that uses all 25 Fama-French size-B/M portfolios as factors (a 25-factor model); and (v) a model that includes the three Fama-French factors plus returns on Fama and French's 30 industry portfolios (a 33-factor model). The motivation here is to illustrate how alpha estimates and

standard errors change as we add factors to the model. Intuitively, going from model (i) to model (ii) illustrates the impact of adding two seemingly-priced factors (SMB and HML) to the model. Going from model (ii) to models (iii) and (iv) illustrates the impact of adding either a few or many seemingly-redundant factors to the model, i.e., three combinations of the six or 25 size-B/M portfolios do a good job capturing the pricing information in all of the portfolios, so three of the six portfolios are redundant in model (iii) and 22 of the 25 portfolios are redundant in model (iv). The fifth model, with 30 industry factors added to the three-factor model, illustrates the impact of adding returns on portfolios that would generally not be thought of as priced factors. All of these factors come from Ken French's website (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/).

Momentum is measured by a stocks' return over the past year (skipping the final month); profitability is measured as operating profits divided by total assets in the prior year; and asset growth is the percentage increase in total assets in the prior year. The sample includes all common stocks on CRSP and Compustat with necessary data. All portfolios are value-weighted and the sample period is July 1963–June 2021.

The alphas in Table 1 change somewhat as factors are added to the model, but the main point of the table is that models with many factors work just as well as models with a few factors, even though the extra factors would often be interpreted as redundant (and are often highly correlated with each other). For example, standard errors from the six- and 25-factor extensions of the three-factor model are nearly identical to those from the three-factor model itself. The biggest change in standard errors comes from adding 30 industry factors to the three-factor model in Panel B: adding industry factors reduces the standard errors by an average of 16% for the profitability deciles and 30% for the high-minus-low profitability portfolio. This reflects that fact that industry returns explain some of the variation in returns on profitability portfolios, presumably because profitability tends to vary across industries.

Table 1 also shows results for a 40-factor model using the anomaly portfolios of Kozak, Nagel, and Santosh (2020). I start with the 57 factors provided on Serhiy Kozak's website (https://www.serhiykozak.com/) and, for simplicity, drop factors that begin after July 1963 and three factors ('mom12', 'prof', and 'growth') that,

based on the descriptions provided, seem most similar to my momentum, profitability, and asset-growth portfolios. The 40-factor model includes both equal- and value-weighted market returns and a range of value, momentum, reversal, growth, issuance, and leverage strategies (along with few others such as trading volume and firm age). The results, in the final column of the table, again show that including a large number of overlapping factors presents no special problem for estimating alpha; the standard errors are comparable to or slightly lower than those for the other models.

The bottom line is that adding a potentially large number of factors to a model isn't inherently problematic for estimating alphas, even if the factors are redundant. In fact, adding redundant factors that are correlated with the test asset can be just as beneficial as adding priced factors: they help absorb some of the random variation in returns and, consequently, improve the precision of alpha.

2. Asset-pricing tests

The analysis above focuses on estimating a single asset's alpha in isolation. A closely related question is how well a model explains the cross-section of expected returns on all assets. In this section, I study how adding factors to a model affects such asset-pricing tests, especially as they relate to the 'model comparison' or 'spanning' tests that have become popular in the literature.

Asset-pricing tests consider whether alphas are zero for all assets when their excess returns R_{all} (a vector) are regressed on the factors:

$$R_{all} = \alpha + \beta F + \varepsilon. \tag{14}$$

Barillas and Shanken (2017) emphasize that the factors included in a model should price any factors excluded from the model (as well as other assets), so I assume R_{all} includes all of the assets available, encompassing a set of N test assets with excess returns R and all of the included and excluded factors. Alphas and residuals when F is regressed on F are zero, so including F on the LHS of (14) does not affect the tests except that the residual covariance for the full set of residuals, $\text{var}(\varepsilon) = \Sigma$, is singular and the statistics below use pseudo-inverses (Σ^{+}) rather than simple inverses.

Gibbons, Ross, and Shanken (1989) show that alphas in (14) are all zero if and only if some combination of the factors equals the tangency portfolio attainable from R_{all}. In particular,

$$\alpha'\Sigma^{+}\alpha = SH^2(R_{all}) - SH^2(F), \tag{15}$$

where 'SH' indicates a population Sharpe ratio. As noted by Barillas and Shanken (2017), an immediate consequence of (15) is that the relative pricing performance of two models, F_1 and F_2, can be evaluated simply by comparing their squared Sharpe ratios, $SH^2(F_1)$ and $SH^2(F_2)$. The test assets in R turn out to be irrelevant if the question is "Which model is better?" as measured by $\alpha'\Sigma^{+}\alpha$.

In the case of nested models, where F_2 includes the factors in F_1 plus some additional factors, the smaller model performs as well as the larger model, $SH^2(F_1) = SH^2(F_2)$, if the extra factors in F_2 are all redundant. This suggests that redundant factors can be dropped from the model with no loss in pricing power, consistent with the spanning tests in the literature and the search for parsimonious models. But that conclusion is less useful than it might appear:

First, it is important to emphasize that 'dropped from a model' is not the same as 'dropped from the tests.' If the goal is to understand how well F_1 explains the cross-section of expected returns, the extra factors in F_2 need to be included in the tests as LHS variables even if they seem to be redundant. (Formally, the result that $SH^2(F_1) = SH^2(F_2)$ does not imply that $SH^2(R,F_1) = SH^2(R,F_2)$.) Thus, regardless of whether the extra factors are redundant or not, asset-pricing tests need to include the full set of returns. This, in turn, suggests there is little benefit from an asset-pricing perspective of identifying 'redundant' factors: redundant factors can not only improve estimates of individual alphas, as discussed earlier, but identifying redundant factors doesn't actually reduce the 'multidimensionality' problem (e.g., Cochrane 2011).

Second, for asset-pricing tests, it is better to include redundant factors in the model on the RHS of the regression than on the LHS. The standard test, from Gibbons, Ross, and Shanken (1989), is based on the sample counterpart to (15). For a given model, the statistic is

$$g = a'\hat{\Sigma}^{+}a/d \times (T-K-d)/(1+sh^2(F)), \tag{16}$$

or

$$g = [sh^2(R_{all}) - sh^2(F)]/d \times (T-K-d)/(1+sh^2(F)), \quad (17)$$

where d equals the total number of assets minus the number of factors in F and $\hat{\Sigma}$ is the residual covariance matrix estimated without a degree-of-freedom adjustment. Conditional on the realized factor returns, and assuming normality, this statistic has a noncentral F-distribution with degrees of freedom d and T–K–d and noncentrality parameter $[SH^2(R_{all}) - SH^2(F)] \times T/(1+sh^2(F))$. (The unconditional distribution, integrating over factor realizations, is more complicated because the noncentrality parameter will vary from sample-to-sample unless $SH^2(R_{all}) - SH^2(F) = 0$.)

Suppose we consider nested models F_1 and F_2, where F_2 includes the K factors in F_1 and q additional factors that are redundant in population, so that $SH^2(F_1) = SH^2(F_2)$. (Since the two Sharpe ratios are identical, I drop subscripts in what follows.) Conditional on the realized returns of F_1, $g(F_1)$ has an F-distribution with degrees of freedom N+q and T–K–N–q and noncentrality parameter $[SH^2(R_{all}) - SH^2(F)] \times T/(1+sh^2(F_1))$. Conditional on the realized returns of F_2, $g(F_2)$ has an F-distribution with degrees of freedom N and T–K–N–q and noncentrality parameter $[SH^2(R_{all}) - SH^2(F)] \times T/(1+sh^2(F_2))$. Notice that these distributions are the same except for (i) the first degree-of-freedom parameter and (ii) what is expected to be a small difference in the denominator of the noncentrality parameter (see my discussion in Section 1).

The key issue is how the power of tests based on $g(F_1)$ and $g(F_2)$ differ. In the absence of any effect on the noncentrality parameter, the answer is both simple and surprising: tests based on $g(F_2)$, the model that includes redundant factors, are always more powerful than tests based on $g(F_1)$. Intuitively, tests based on $g(F_2)$ need to test whether N alphas are zero, while tests based on $g(F_1)$ need to test whether N+q alphas are zero and the extra q alphas only add noise to the tests (the alphas are zero in population because the extra factors are redundant). I establish this result formally in the appendix.

The generality of this conclusion is quite remarkable: It does not depend on the number of redundant factors, the length of the time series, or even the correlation between the redundant factors and the other assets (which,

as explained earlier, does affect the precision of individual alphas). In all cases, if the goal is to test whether the model explains the cross-section of expected returns, it is always better to include redundant factors in the model, on the RHS of the regression, than on the LHS. Dropping redundant factors from the model is unambiguously suboptimal.

The impact on the noncentrality parameter is expected to be small and doesn't seem to alter the conclusion. Since the unconditional distributions of the statistics are nonstandard (they are a mixture of F distributions with different noncentrality parameters), I rely on numerical results to illustrate the effects. Table 2 reports the probability of rejecting the models at a 5% significance level given a variety of different parameters. I assume we start with five factors in F_1 that have a maximum squared Sharpe ratio of $SH^2(F_1) = 0.05$ and consider adding up to 32 redundant factors to the model. Twenty-five test assets are also included in the tests. The unexplained squared Sharpe ratio, $SH^2(R_{all}) - SH^2(F)$, ranges from zero (the null that the model works perfectly) to 0.10. The latter represents substantial mispricing, implying that some portfolio can be found that is uncorrelated with the factors and has a Sharpe ratio of 0.32 monthly (1.10 annualized). I report results for T = 240 or T = 480 months.

Table 2 confirms the inferences above: Tests based on $g(F_2)$, the model that includes redundant factors, are always more powerful than tests based on $g(F_1)$. The tests reject 5% of the time when the models work perfectly, in the first column, but with greater probability when $SH^2(R_{all}) > SH^2(F)$. The main result for our purposes is that, for a given T (time series), q (number of redundant factors), and SH^2(all) (mispricing), the rejection probabilities are always greater for $g(F_2)$ than for $g(F_1)$. The differences are typically modest, but the important point is that dropping redundant factors from a model never helps asset-pricing tests— in fact, it is actually detrimental.

To be clear, comparing the top row of each panel with the subsequent rows, including redundant factors at all in the tests is worse than dropping them completely *if the redundant factors do not contribute anything to* $SH^2(R_{all})$. However, there is no reason to believe the italicized condition is true: redundant factors do not

contribute to the model's squared Sharpe ratio, $SH^2(F_1) = SH^2(F_2)$, but that says nothing about how much they contribute to $SH^2(R_{all})$. The example in the introduction, with two factors and one test asset, illustrates this point (it is described in terms of sample moments but could easily be adapted for population moments). The potential impact on $SH^2(R_{all})$ explains why the relevant comparison is $g(F_1)$ versus $g(F_2)$, not a comparison of the numbers in each panel as q changes.

In sum, if the goal is to test a given factor model, identifying redundant factors in the model isn't helpful: redundant factors cannot be dropped from the tests and, as a rule, it is better to include them on the RHS not the LHS of the regression.

3. Estimating Sharpe ratios

The analysis above focuses on testing whether a factor model fully explains the cross-section of expected returns. An alternative metric of performance, relevant even if a model is not perfect, is simply to focus on the model's maximum squared Sharpe ratio, $SH^2(F)$, or, equivalently, the variance of the model-implied SDF (Hansen and Jagannathan 1991). Barillas and Shanken (2017, 2018) advocate using this metric to compare models, and the literature often reports it as a summary measure of a model's performance.

The issue I consider here is how including or excluding factors affects estimates of $SH^2(F)$. As noted earlier, for a model with K normally distributed factors, the sample statistic

$$p = sh^2(F) \times (T–K)/K \tag{18}$$

has a noncentral F distribution with degrees of freedom K and T–K and noncentrality parameter $T \times SH^2(F)$. The moments of a noncentral F imply that $sh^2(F)$ has a mean of

$$E[sh^2(F)] = (SH^2(F)+K/T) \times T/(T–K–2). \tag{19}$$

The sample $sh^2(F)$ is, of course, an upward biased estimate of the true $SH^2(F)$, and the bias increases with the number of factors. Suppose, for example, that $SH^2(F) = 0.05$ and $T = 480$. The mean of $sh^2(F)$ is 0.061 with 5 factors, 0.073 with 10 factors, 0.096 with 20 factors, and 0.146 with 40 factors. This link between the number of factors and the bias in $sh^2(F)$ seems to provide, at least in part, the preference for models with fewer factors

in the literature (e.g., Fama and French 2018, p. 235).

In principle, we could also use (18) to obtain the variance of $sh^2(F)$ and combine it with (19) to derive an approximate two-standard-deviation confidence interval for $SH^2(F)$. However, the distribution of $sh^2(F)$ is not symmetric, and the mean and variance are both functions of the true $SH^2(F)$, so a better approach is to get an exact confidence interval by directly determining the set of $SH^2(F)$ that cannot be rejected by the data (using whatever confidence level is desired).

The idea is illustrated in Fig. 1. The graph shows the sampling distribution of $sh^2(F)$ on the y-axis (specifically, the 5th, 50th, and 95th percentiles) plotted against the true $SH^2(F)$ given 480 months of data for four factors on the left and 16 factors on the right. The lines are upward sloping because a higher value of $SH^2(F)$ leads to higher sample Sharpe ratios and, conversely, a higher sample Sharpe ratio makes it more likely that the true value of $SH^2(F)$ is greater. Given an observed sample $sh^2(F)$, a 90% confidence interval for $SH^2(F)$ can be obtained by finding values of $SH^2(F)$ for which the sample $sh^2(F)$ is between the 5th and 95th percentiles of the sampling distribution. The upper and lower bounds are identified in the graph where a horizon line drawn at y = "observed $sh^2(F)$" intersects the 5th and 95th percentiles. Put differently, the sampling distribution for $sh^2(F)$ is found by fixing $SH^2(F)$ and scanning up, while a confidence interval for $SH^2(F)$ is found by fixing $sh^2(F)$ and scanning across. (Lewellen, Nagel, and Shanken (2010) suggest a similar approach to get confidence intervals for unexplained squared Sharpe ratios.)

For example, if the sample $sh^2(F) = 0.10$, a 90% confidence interval for $SH^2(F)$ is determined in Fig. 1 by where a horizontal line at 0.10 intersects the 5th and 95th percentiles. The confidence interval is [0.052, 0.146] with 4 factors and [0.028, 0.115] with 16 factors. The latter is shifted downward to reflect, implicitly, the greater bias in $sh^2(F)$ when there are more factors in the model.

My main interest here is how adding factors to a model affects our ability to estimate $SH^2(F)$ depending on whether the extra factors are redundant or not. Suppose, again, we have two nested models F_1 and F_2, where F_2 includes the factors in F_1 plus additional factors. If the extra factors are all redundant, the models have the

same maximum squared Sharpe ratio $SH^2(F)$, so adding factors affects the sampling distribution of p in (18) only through the degree-of-freedom parameters. As noted above, the extra factors lead to a stronger upward bias in the sample Sharpe ratio, $E[sh^2(F_2)] > E[sh^2(F_1)]$, but the confidence intervals inherently adjust for that. The more important issue, for my purposes, is how the extra factors affect the width of the confidence interval.

A complication is that the width of the confidence interval depends on the realized value of $sh^2(F)$, an effect clearly visible in Fig. 1. Therefore, we either need to consider the width for different values of $sh^2(F)$, recognizing that adding factors affects the likely values of $sh^2(F)$, or the expected width integrating over $sh^2(F)$. For brevity, I follow the spirit of the latter approach and, in particular, report the confidence interval if $sh^2(F)$ falls at its 50^{th} percentile for a given K and $SH^2(F)$. (The results are similar for the expected width of the confidence interval, averaging over different values of $sh^2(F)$.)

Table 3 illustrates how confidence intervals change as the number of factors increases, given T = 480 monthly observations and true $SH^2(F)$ ranging from 0.00 to 0.10. (Note that $SH^2(F)$ is not actually used to get the confidence interval; its only role in Table 3 is to change the median $sh^2(F)$ for which I report the confidence interval.) The top panel shows the median value of $sh^2(F)$ for a given K and $SH^2(F)$. The second and third panels show the lower and upper bounds of the 90% confidence interval for $SH^2(F)$ if the sample $sh^2(F)$ equals its median value, and the bottom panel reports the width of the confidence interval.

For example, if a model has four factors and the true $SH^2(F)$ equals 0.05, the median value of the sample $sh^2(F)$ equals 0.057 and, if that value is observed in the data, the resulting confidence interval for $SH^2(F)$ is [0.021, 0.091] with a width of 0.070. If the model has 32 factors and the true $SH^2(F)$ equals 0.05, the median value of $sh^2(F)$ equals 0.123 and, if that value is observed in the data, the resulting confidence interval is [0.012, 0.101] with a width of 0.089. Notice, again, that the confidence interval implicitly adjusts for the bias in the sample $sh^2(F)$. The bias with 32 factors is so large, in fact, that the confidence interval falls entirely below the sample estimate.

A key result in Table 3 is that including redundant factors in a model has a relatively modest effect on the

width of the confidence intervals (a proxy for the precision of the estimate). For example, if the true $SH^2(F)$ is 0.02, the median width of the confidence interval is 0.045 for a model with four factors, 0.055 for a model with 16 factors, and 0.062 for a model with 32 factors. Similarly, if the true $SH^2(F)$ is 0.08, the median width of the confidence interval is 0.088 for a model with four factors, 0.095 for a model with 16 factors, and 0.105 for a model with 32 factors. The results suggest that, for the purposes of estimating $SH^2(F)$, including redundant factors in a model is detrimental but not prohibitively so.

This finding is actually rather surprising: It is well known that estimating the tangency portfolio itself—that is, the combination of the factors that attains $SH^2(F)$—is challenging when the number of factors is large because average returns are so noisy (e.g., Kozak, Nagel, and Santosh 2020). Table 3 suggests, however, that the higher estimation error in portfolio weights with more factors only translates into modestly more error in estimates of the portfolio's Sharpe ratio.

The discussion above assumes that the true $SH^2(F)$ does not change as more factors are added to a model (the extra factors are all redundant). Table 3 also illustrates how confidence intervals change if the extra factors are instead priced, which implies that $SH^2(F)$ increases as K increases. For example, if we start with four factors in the model that have $SH^2(F_1) = 0.03$ and add eight factors that increase the true $SH^2(F_2)$ to 0.06, the median confidence interval shifts from [0.009, 0.063] to [0.027, 0.107]. The latter is wider but centered at closer to the true value of $SH^2(F_2)$.

It is also interesting to consider how the confidence intervals change if we include or exclude factors that appear to be redundant *in sample*. This thought experiment is relevant in practice since factors are often judged as redundant or not based on their sample returns. The example considered above in discussing Fig. 1 illustrates this scenario: If a four-factor model and 16-factor model produce the same sample $sh^2(F)$ of 0.10, the confidence interval for $SH^2(F)$ is [0.052, 0.146] for the four-factor model and [0.028, 0.115] for the 16-factor model. The confidence interval with 16 factors is actually narrower in this case, 0.087, compared with 0.094 for the four-factor model.

This thought experiment can be interpreted as an illustration of how data mining affects estimates of Sharpe ratios: If we start with 16 factors that produce a sample $sh^2(F)$ of 0.10 but drop 12 factors that are redundant in sample, the confidence interval for $SH^2(F)$ shifts upward from [0.028, 0.115], the interval based appropriately on the full set of 16 factors, to [0.052, 0.146], the interval implied by the data-mined four-factor model. The midpoint of the confidence interval, which can be taken as a point estimate of $SH^2(F)$, shifts upward from 0.072 to 0.099.

Not surprisingly, the problem is more severe if we search over more factors. Continuing with the same example, if we started with 40 factors and observe the same sample $sh^2(F)$ of 0.10, the 90% confidence interval for $SH^2(F)$ should be [0.000, 0.052] with a midpoint of 0.026. This 'unbiased' confidence interval lies entirely below the one for the data-mined four-factor model, and the the midpoint is nearly 75% lower. This example is extreme because the extra 36 factors are assumed to be completely redundant in sample (they contribute nothing to $sh^2(F)$) but illustrates how data mining can affect the inferences. It is important to note that the problem here does not come from looking at many factors per se but, rather, from ignoring that fact when evaluating the reduced model.

Table 4 provides an application of the use of confidence intervals. The table reports the sample $sh^2(F)$ and a 90% confidence interval for $SH^2(F)$ for the six factor models considered earlier: the CAPM; the Fama-French (1993) three-factor model; two expanded versions of the model with either the six size-B/M portfolios used to construct the Fama-French factors or Fama and French's full set of 25 size-B/M portfolios; the three-factor model augmented with 30 industry factors; and a 40-factor model with factors from Kozak, Nagel, and Santosh (2020). I also report statistics for the square root of $SH^2(F)$ for each model.

As expected, the sample $sh^2(F)$ increases as factors are added to a model, rising from 0.017 for the CAPM to 0.035 for the FF three-factor model, 0.175 for the 25-factor model, 0.135 for the 'FF3+30 industry' model, and 0.627 for the KNS40 model. But just as important, the confidence intervals for the true $SH^2(F)$ also shift to the right and are reasonably tight even for the models with many factors. In fact, for the maximum Sharpe ratios (not squared), the confidence intervals are only slightly wider for models with more factors. For example, the

confidence interval for SH(F) is [0.071, 0.192] for the CAPM, [0.114, 0.241] for the three-factor model, [0.298, 0.435] for the 25-factor model, and [0.655, 0.804] for the 40-factor model. The last result is remarkable, both because it is incredibly high (implying a Sharpe ratio about 5 times that of the market portfolio) and surprisingly narrow given that the model has 40 factors.

4. Conclusions

The empirical asset-pricing literature often advocates models with only a handful of factors, even as the count of potential factors has grown to dozens or even hundreds. It is common today to use models with only market, size, B/M, profitability, and investment factors, sometimes supplemented with a momentum factor or the 'mispricing' factors of Stambaugh and Yuan (2016).

The preference for models with only a few factors may be rooted in theoretical models with a small number of factors (e.g., the CAPM). My results suggest that, from an empirical perspective, the preference for fewer factors may be misplaced: If the goal is to estimate alphas or to test whether a model explains the cross-section of expected returns, adding factors to a model—even completely redundant ones—can be beneficial, not costly. Extra factors can improve estimates of individual alphas and increase the power of asset-pricing tests. The impact on the sampling error in individual alphas depends on how highly correlated the extra factors are with the asset being considered, while the benefits for asset-pricing tests hold generally regardless of the number of factors, the length of the time series, or the correlation between the extra factors and the other assets included in the tests.

My results have several implications. The most immediate is that the empirical literature should be willing to entertain models with potentially many more factors. A corollary is that, for applications, there is little advantage from answering questions like "Does earnings momentum subsume return momentum?" or "Do size, profitability, and investment factors subsume the B/M effect?". In the absence of other considerations (such as data availability), my results suggest that all of the factors can be included in a model, regardless of the answer to those questions. To be clear, it is interesting to know whether, say, a cash-profitability factor

subsumes an operating-profitability factor. My results simply say that the answer is not important if the goal is to use the factor model to estimate alphas or to test whether the model explains the cross-section of expected returns, because all of the factors might as well be included in those tests.

As a side benefit, including more factors in a model might also improve the search for new anomalies. The standard approach to testing whether a proposed characteristic is associated with expected returns is to regress returns for a proposed long-short portfolio on a few well-accepted factors. If the model expands to include more factors, not only can that lower the standard error of the estimated alpha but it can also reduce the chance of rediscovering old anomalies in new disguises.

My results also imply that, to estimate alphas, the choice of factors to include in a model could be determined at least in part by the (expected) correlation between the factors and the assets, not just a belief about whether the factors are priced. For example, suppose a researcher wants to estimate a mutual fund's alpha. If the fund is known to favor, say, energy stocks, it can make sense to include an energy-industry factor in the regression even if that factor is not thought to be priced. In general, the factors to include will depend on the question being asked: Does the fund outperform the market?, or Does the fund outperform the market controlling for its industry tilt? Both questions are potentially interesting, regardless of whether industry returns are priced factors.

Appendix

Section 2 studies asset-pricing tests for nested models F_1 and F_2, where F_2 includes the K factors in F_1 plus q redundant factors. The models have the same true $SH^2(F_i)$, denoted $SH^2(F)$. The Gibbons, Ross, and Shanken (1989) F-statistic, testing whether the models explain the cross-section of expected returns on all test assets and factors, $R_{all} = (R, F_2)$, are denoted $g(F_1)$ and $g(F_2)$. This appendix shows that tests based on $g(F_2)$ are more powerful than tests based on $g(F_1)$.

Conditional on the realized returns of F_1, $g(F_1)$ has a noncentral F-distribution with degrees of freedom N+q and T–K–N–q and noncentrality parameter $[SH^2(R_{all}) - SH^2(F)] \times T/(1+sh^2(F_1))$. Conditional on the realized returns of F_2, $g(F_2)$ has a noncentral F-distribution with degrees of freedom N and T–K–N–q and noncentrality parameter $[SH^2(R_{all}) - SH^2(F)] \times T/(1+sh^2(F_2))$. The key observation is that these distributions are the same except for (i) the first degree-of-freedom parameter and (ii) what is expected to be a small difference in the denominator of the noncentrality parameter.

Ignoring the second effect (it is discussed in the text), the only difference comes from the first degree-of-freedom parameter. Thus, the key issue is how that parameter affects the power of the tests. Unfortunately, the cumulative distribution function of the noncentral F distribution is unwieldy, so it is challenging to prove any results mathematically.

As a simple but admittedly inelegant solution, I confirm the result holds numerically for essentially the entire empirically relevant parameter space. Generically, $g(F_i) \sim G(v_1, v_2, \lambda)$, where G is the noncentral F distribution and, in the applications here, parameters v_1 and v_2 are positive integers. To establish the result, we need to show that power is decreasing in v_1 holding the other parameters constant. I calculate power (for a test of size 5%) for v_1 ranging from 1 to 200 (representing the total number of assets minus the number of RHS factors), v_2 ranging from 12 to 960 in increments of 12 (representing the number of months minus the total number of assets), and noncentrality parameters that correspond, roughly, to $SH^2(R_{all}) - SH^2(F)$ ranging from 0.01 to 0.30 in increments of 0.01 (λ equals these values times v_2). I then check numerically that power is decreasing over

the full range of v_1 given every combination of v_2 and λ. A few representative cases are illustrated in Fig. A1. (The procedure is implemented in SAS's matrix language using the built-in functions for the noncentral F distribution.)

As noted in the text, the intuition is that $g(F_2)$ tests whether N alphas are zero, while $g(F_1)$ tests whether N+q alphas are zero and the extra alphas simply add noise to the test. More formally, from standard results on noncentral F distributions, $g(F_1)$ can be expressed as a noncentral chi-squared variable, C, with N+q degrees of freedom and noncentrality parameter $\lambda = [SH^2(R_{all}) - SH^2(F)] \times T/(1+sh^2(F_1))$, divided by an independent central chi-squared variable, W, with T–K–N–q degrees of freedom:

$$g(F_1) = \frac{C/(N+q)}{W/(T-K-N-q)}. \quad \text{(A1)}$$

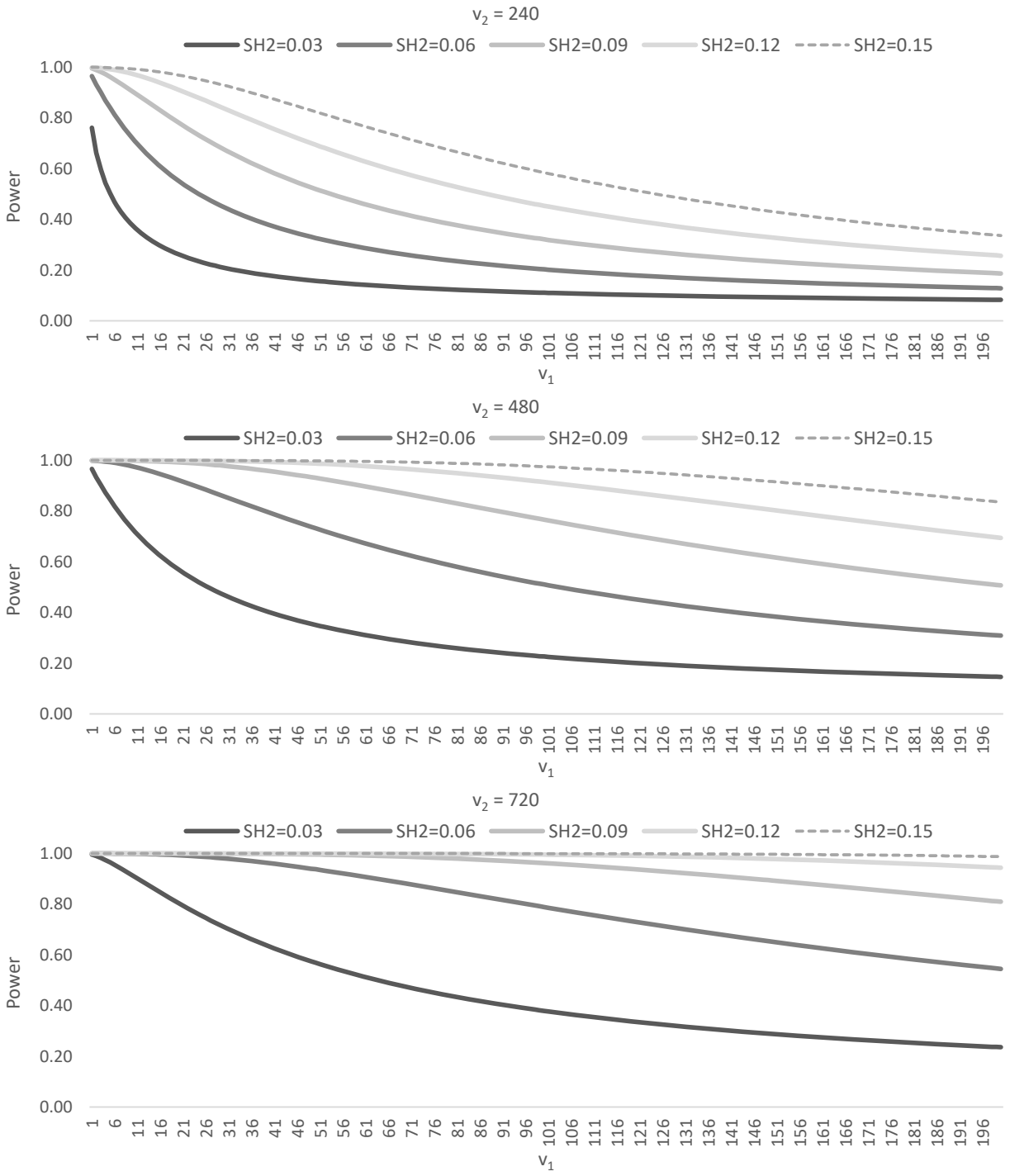
In turn, C can be expressed as the sum of two independent random variables: C_1, a noncentral chi-squared variable with N degrees of freedom and noncentrality parameter λ (the same as C), and C_2, a central chi-squared with q degrees of freedom (and, thus, whose distribution does not depend on λ):

$$\begin{aligned} g(F_1) &= \frac{(C_1+C_2)/(N+q)}{W/(T-K-N-q)} \\ &= \frac{C_1/N}{W/(T-K-N-q)}\frac{N}{N+q} + \frac{C_2/q}{W/(T-K-N-q)}\frac{q}{N+q} \\ &= g_{11}\frac{N}{N+q} + g_{12}\frac{q}{N+q}. \end{aligned} \quad \text{(A2)}$$

Eq. (A2) formalizes the idea that the extra alphas in $g(F_1)$ simply add noise to the test: $g(F_1)$ is a weighted average of a variable whose distribution depends on λ (g_{11}, which is noncentral F with degrees of freedom N and T–K–N–q and noncentrality parameter λ) and a second variable whose distribution does not (g_{12}, whch is central F with degrees of freedom q and T–K–N–q). All of the information in $g(F_1)$ about the parameter we care about (λ or $SH^2(R_{all}) - SH^2(F)$) comes from g_{11}, which, crucially, has the same distribution as $g(F_2)$ (again, ignoring the small difference in noncentrality parameters). Thus, $g(F_1)$ is basically a noisy version of $g(F_2)$, so we learn more from $g(F_2)$ than from $g(F_1)$.

Figure A1

This figures plots the probability of rejecting at a 5% significance level the null hypothesis that $\lambda = 0$ (i.e., a factor model explains the cross-section of expected returns) as a function of v_1, v_2, and λ. The sample statistic is distributed noncentral F with degrees of freedom v_1 and v_2 and noncentrality parameter λ, where v_1 corresponds to the total number of assets minus the number of factors, v_2 corresponds to the number of months minus the total number of assets, and λ roughly corresponds to the magnitude of mispricing (in the graph, λ equals v_2 times the unexplained squared Sharpe ratio of the model, which is labeled 'SH2').



References

Barillas, Francisco, Raymond Kan, Cesare Robotti, and Jay Shanken, 2020. Model comparison with Sharpe ratios. *Journal of Financial and Quantitative Analysis* 55, 1840–1874.

Barillas, Francisco and Jay Shanken, 2017. Which alpha? *Review of Financial Studies* 30, 1316–1338.

Barillas, Francisco and Jay Shanken, 2018. Comparing asset pricing models. *Journal of Finance* 73, 715–754.

Cochrane, John, 2011. Presidential address: Discount rates. *Journal of Finance* 66, 1047–1108.

Fama, Eugene and Kenneth French, 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 3–56.

Fama, Eugene and Kenneth French, 2015. A five-factor asset-pricing model. *Journal of Financial Economics* 116, 1–22.

Fama, Eugene and Kenneth French, 2018. Choosing factors. *Journal of Financial Economics* 128, 234–252.

Gibbons, Michael, Stephen Ross, and Jay Shanken, 1989. A test of the efficiency of a given portfolio. *Econometrica* 57, 1121–1152.

Hansen, Lars and Ravi Jagannathan, 1991. Implications of security market data for models of dynamic economies. *Journal of Political Economy* 99, 225–262.

Harvey, Campbell and Yan Liu, 2021. Lucky factors. *Journal of Financial Economics* 141, 413–435.

Harvey, Campbell, Yan Liu, and Heqing Zhu, 2016. … and the cross-section of expected returns. *Review of Financial Studies* 29, 5–68.

Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang, 2019. Which factors? *Review of Finance* 23, 1–35.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015. Digesting anomalies: An investment approach. *Review of Financial Studies* 28, 650–705.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020. Shrinking the cross-section. *Journal of Financial Economics* 135, 271–292.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010. A skeptical appraisal of asset pricing tests. *Journal of Financial Economics* 96, 175–194.

MacKinlay, A. Craig, 1995. Multifactor models do not explain deviations from the CAPM. *Journal of Financial Economics* 38, 3–28.

Mood, Alexander, Franklin Graybill, and Duane Boes, 1974. *Introduction to the Theory of Statistics*, 3rd edition. McGraw-Hill, New York.

Morrison, Donald, 1990. *Multivariate Statistical Methods*, 3rd edition. McGraw-Hill, New York.

Stambaugh, Robert and Yu Yuan, 2016. Mispricing factors. *Review of Financial Studies* 30, 1270–1315.

Table 1: Alphas

This table reports alpha estimates and standard errors for value-weighted momentum, profitability, and asset growth deciles using several factor models: (i) the CAPM; (ii) the Fama-French (1993) three-factor model ('FF3'); (iii) an expanded version of the FF model that uses all six of the underlying size-B/M portfolios ('FF6'); (iv) an expanded version of the FF model that uses all 25 of Fama and French's size-B/M portfolios ('FF25'); (v) an augmented version of the FF model that adds returns on Fama and French's 30 industry portfolios ('FF3+30 ind'); and (vi) a 40-factor model based on the anomaly portfolios of Kozak, Nagel, and Santosh (2020) ('KNS40'). Models (i)–(v) use data for July 1963 through June 2021 from Ken French's website, while model (vi) uses data for July 1963 through December 2019 from Serhiy Kozak's website.

	CAPM		FF3		FF6		FF25		FF3+30 ind		KNS40	
Portfolio	a	se	a	se	a	se	a	se	a	se	a	se
Panel A: Momentum portfolios												
Low	-0.94	0.20	-1.10	0.19	-0.91	0.19	-0.75	0.19	-0.98	0.17	-0.31	0.14
2	-0.36	0.13	-0.50	0.13	-0.42	0.13	-0.32	0.13	-0.41	0.11	-0.08	0.09
3	-0.13	0.10	-0.24	0.10	-0.23	0.10	-0.13	0.10	-0.19	0.09	0.05	0.09
4	-0.03	0.08	-0.11	0.08	-0.10	0.08	-0.03	0.08	-0.08	0.07	0.02	0.07
5	-0.03	0.07	-0.10	0.06	-0.11	0.06	-0.08	0.06	-0.05	0.06	0.06	0.07
6	0.03	0.06	-0.02	0.06	-0.03	0.06	-0.01	0.06	-0.01	0.05	0.03	0.07
7	0.04	0.06	0.01	0.06	-0.05	0.06	-0.01	0.06	-0.02	0.06	-0.05	0.07
8	0.15	0.06	0.14	0.06	0.09	0.06	0.09	0.07	0.13	0.06	0.11	0.07
9	0.24	0.08	0.23	0.08	0.19	0.08	0.17	0.08	0.21	0.07	0.08	0.07
High	0.48	0.13	0.56	0.11	0.56	0.12	0.46	0.12	0.55	0.11	0.40	0.09
High–Low	1.42	0.27	1.66	0.26	1.47	0.27	1.21	0.27	1.53	0.25	0.71	0.17
Panel B: Profitability portfolios												
Low	-0.26	0.10	-0.42	0.09	-0.33	0.09	-0.32	0.09	-0.27	0.06	-0.16	0.07
2	-0.05	0.09	-0.23	0.07	-0.19	0.07	-0.18	0.07	-0.09	0.05	-0.03	0.06
3	-0.10	0.07	-0.18	0.07	-0.15	0.07	-0.14	0.07	-0.11	0.06	0.03	0.07
4	0.04	0.07	-0.05	0.06	0.01	0.06	-0.02	0.06	0.00	0.05	-0.05	0.06
5	0.01	0.07	-0.07	0.06	-0.04	0.06	-0.01	0.06	0.00	0.05	0.00	0.06
6	0.06	0.06	0.00	0.06	0.03	0.06	0.04	0.06	0.03	0.05	0.06	0.06
7	0.06	0.06	0.01	0.06	0.02	0.06	0.03	0.06	0.04	0.05	-0.07	0.06
8	0.03	0.05	0.02	0.05	-0.02	0.05	-0.02	0.05	0.00	0.05	-0.07	0.05
9	0.01	0.05	0.04	0.05	-0.01	0.05	-0.02	0.05	0.02	0.05	-0.02	0.05
High	0.11	0.07	0.26	0.05	0.15	0.04	0.12	0.05	0.12	0.04	0.22	0.05
High–Low	0.37	0.14	0.68	0.11	0.49	0.11	0.44	0.11	0.39	0.08	0.38	0.09
Panel C: Asset growth portfolios												
Low	0.22	0.09	0.09	0.08	0.12	0.08	0.15	0.08	0.08	0.08	-0.01	0.08
2	0.21	0.07	0.10	0.07	0.09	0.07	0.14	0.07	0.14	0.07	0.04	0.07
3	0.18	0.06	0.09	0.06	0.11	0.05	0.14	0.06	0.06	0.05	-0.01	0.06
4	0.11	0.06	0.03	0.05	0.03	0.05	0.00	0.05	0.01	0.05	-0.04	0.06
5	0.09	0.05	0.02	0.05	0.00	0.05	0.02	0.05	0.02	0.05	0.07	0.06
6	0.06	0.05	0.02	0.05	-0.02	0.05	-0.06	0.05	-0.02	0.05	0.01	0.06
7	0.06	0.05	0.05	0.05	-0.02	0.05	-0.05	0.05	0.02	0.05	0.05	0.05
8	-0.03	0.05	-0.02	0.05	-0.09	0.05	-0.11	0.05	0.00	0.05	-0.04	0.06
9	0.05	0.07	0.14	0.06	0.09	0.06	0.11	0.06	0.10	0.06	0.18	0.07
High	-0.39	0.08	-0.27	0.07	-0.20	0.07	-0.19	0.07	-0.22	0.06	-0.16	0.07
High–Low	-0.61	0.13	-0.36	0.11	-0.32	0.11	-0.34	0.11	-0.30	0.11	-0.16	0.11

Table 2: Power of asset-pricing tests

This table reports the probability of rejecting the null hypothesis at a 5% significance level that a five-factor model F_1 or an extended model F_2, that adds up to 32 redundant factors (in addition to the factors in F_1), explains the cross-section of expected returns. Results are shown for various combinations of T (months of data), q (number of redundant factors), and $SH^2(R_{all}) - SH^2(F)$ (true pricing errors). The null hypothesis, $SH^2(R_{all}) - SH^2(F) = 0$, is tested using the Gibbons, Ross, and Shanken (1989) F-statistic. Panels labeled '$g(F_1)$' show rejection probabilities for model F_1 and panels labeled '$g(F_2)$' show rejection probabilities for model F_2. The 'left-hand side' assets include all of the factors in F_1 and F_2 along with 25 additional test assets. Returns are assumed to be normally distributed.

			$SH^2(R_{all}) - SH^2(F)$										
		q	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
T=240	$g(F_1)$	0	0.050	0.092	0.147	0.214	0.290	0.371	0.454	0.535	0.611	0.680	0.741
		4	0.050	0.087	0.137	0.197	0.265	0.340	0.417	0.493	0.567	0.636	0.699
		8	0.050	0.084	0.129	0.183	0.245	0.313	0.385	0.458	0.529	0.597	0.660
		12	0.050	0.081	0.122	0.171	0.228	0.291	0.358	0.426	0.494	0.561	0.623
		16	0.050	0.079	0.116	0.161	0.214	0.272	0.334	0.399	0.464	0.528	0.589
		20	0.050	0.077	0.111	0.153	0.201	0.255	0.313	0.374	0.436	0.497	0.557
		24	0.050	0.075	0.107	0.146	0.190	0.240	0.295	0.352	0.411	0.470	0.528
		28	0.050	0.074	0.104	0.139	0.181	0.228	0.278	0.332	0.388	0.445	0.501
		32	0.050	0.072	0.100	0.134	0.173	0.216	0.264	0.314	0.367	0.421	0.475
	$g(F_2)$	0	0.050	0.092	0.147	0.214	0.290	0.371	0.454	0.535	0.611	0.680	0.741
		4	0.050	0.091	0.144	0.210	0.284	0.364	0.445	0.524	0.600	0.669	0.730
		8	0.050	0.090	0.142	0.206	0.278	0.356	0.436	0.514	0.589	0.657	0.719
		12	0.050	0.089	0.140	0.202	0.272	0.348	0.426	0.504	0.577	0.646	0.708
		16	0.050	0.088	0.138	0.198	0.267	0.341	0.417	0.493	0.566	0.634	0.696
		20	0.050	0.087	0.136	0.194	0.261	0.333	0.408	0.483	0.555	0.622	0.684
		24	0.050	0.086	0.133	0.190	0.255	0.326	0.399	0.472	0.543	0.610	0.672
		28	0.050	0.085	0.131	0.187	0.250	0.318	0.389	0.461	0.531	0.598	0.659
		32	0.050	0.084	0.129	0.183	0.244	0.311	0.380	0.451	0.520	0.585	0.647
T=480	$g(F_1)$	0	0.050	0.154	0.308	0.482	0.643	0.772	0.864	0.924	0.960	0.980	0.990
		4	0.050	0.144	0.284	0.446	0.603	0.735	0.835	0.903	0.946	0.971	0.985
		8	0.050	0.136	0.264	0.416	0.568	0.701	0.806	0.881	0.930	0.961	0.979
		12	0.050	0.129	0.247	0.390	0.536	0.669	0.777	0.858	0.914	0.950	0.972
		16	0.050	0.123	0.233	0.367	0.508	0.639	0.749	0.835	0.896	0.938	0.964
		20	0.050	0.119	0.221	0.347	0.482	0.611	0.723	0.812	0.878	0.925	0.955
		24	0.050	0.114	0.210	0.329	0.459	0.585	0.697	0.789	0.860	0.910	0.945
		28	0.050	0.111	0.201	0.313	0.438	0.561	0.673	0.767	0.841	0.896	0.934
		32	0.050	0.108	0.192	0.299	0.418	0.538	0.650	0.745	0.822	0.881	0.923
	$g(F_2)$	0	0.050	0.154	0.308	0.482	0.643	0.772	0.864	0.924	0.960	0.980	0.990
		4	0.050	0.153	0.305	0.477	0.638	0.767	0.860	0.921	0.958	0.979	0.990
		8	0.050	0.151	0.302	0.473	0.633	0.762	0.856	0.918	0.956	0.977	0.989
		12	0.050	0.150	0.299	0.468	0.627	0.757	0.852	0.915	0.954	0.976	0.988
		16	0.050	0.149	0.296	0.464	0.622	0.752	0.848	0.912	0.952	0.975	0.987
		20	0.050	0.148	0.293	0.459	0.617	0.747	0.844	0.909	0.949	0.973	0.987
		24	0.050	0.147	0.290	0.454	0.611	0.742	0.839	0.905	0.947	0.972	0.986
		28	0.050	0.146	0.287	0.450	0.606	0.736	0.834	0.902	0.945	0.970	0.985
		32	0.050	0.145	0.284	0.445	0.600	0.731	0.830	0.898	0.942	0.969	0.984

Figure 1: Sample vs. true Sharpe ratios

This figure shows the 5th, 50th, and 95th percentiles of the sampling distribution of $sh^2(F)$, the sample maximum squared Sharpe ratio attainable from a given set of factors, plotted against $SH^2(F)$, the population maximum squared Sharpe ratio attainable from the factors. The graph on the left uses four factors and graph on the right uses 16 factors, all normally distributed. The points indicated in the graphs represent the upper and lower bounds of a 90% confidence interval for the true $SH^2(F)$ if the sample $sh^2(F)$ equals 0.10. T = 480.

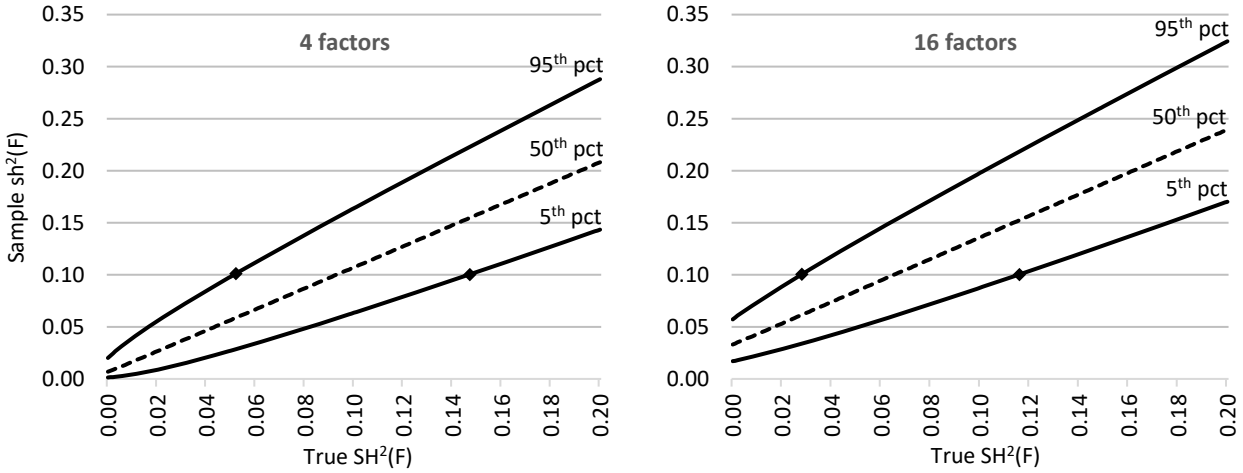


Table 3: Confidence intervals for $SH^2(F)$

The top panel reports the 50th percentile of the sampling distribution for $sh^2(F)$, the sample maximum squared Sharpe ratio attainable from a given set of factors, for different combinations of the number of factors (K) and true squared Sharpe ratio $SH^2(F)$. The second and third panels report the lower and upper bounds of a 90% confidence interval for $SH^2(F)$ if the sample $sh^2(F)$ equals the number in the top panel. The bottom panel reports the width of the confidence interval (upper bound minus lower bound). Returns are assumed to be normally distributed. T = 480.

		$SH^2(F)$										
	K	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Median	4	0.007	0.017	0.027	0.037	0.047	0.057	0.067	0.077	0.087	0.097	0.107
$sh^2(F)$	8	0.016	0.025	0.035	0.046	0.056	0.066	0.076	0.086	0.096	0.107	0.117
	12	0.024	0.034	0.044	0.055	0.065	0.075	0.085	0.096	0.106	0.116	0.126
	16	0.033	0.043	0.053	0.064	0.074	0.084	0.095	0.105	0.115	0.126	0.136
	20	0.042	0.052	0.063	0.073	0.083	0.094	0.104	0.115	0.125	0.136	0.146
	24	0.051	0.062	0.072	0.082	0.093	0.103	0.114	0.124	0.135	0.146	0.156
	28	0.061	0.071	0.082	0.092	0.103	0.113	0.124	0.134	0.145	0.156	0.166
	32	0.070	0.081	0.091	0.102	0.113	0.123	0.134	0.145	0.155	0.166	0.177
Lower	4	0.000	0.000	0.003	0.009	0.015	0.021	0.028	0.035	0.042	0.050	0.057
bound	8	0.000	0.000	0.002	0.007	0.013	0.020	0.027	0.034	0.041	0.048	0.056
	12	0.000	0.000	0.000	0.006	0.012	0.019	0.026	0.033	0.040	0.047	0.055
	16	0.000	0.000	0.000	0.004	0.011	0.017	0.024	0.032	0.039	0.046	0.054
	20	0.000	0.000	0.000	0.003	0.009	0.016	0.023	0.030	0.038	0.045	0.053
	24	0.000	0.000	0.000	0.001	0.008	0.015	0.022	0.029	0.037	0.044	0.052
	28	0.000	0.000	0.000	0.000	0.007	0.014	0.021	0.028	0.036	0.043	0.051
	32	0.000	0.000	0.000	0.000	0.005	0.012	0.020	0.027	0.034	0.042	0.050
Upper	4	0.016	0.033	0.048	0.063	0.077	0.091	0.104	0.117	0.130	0.142	0.155
Bound	8	0.020	0.036	0.051	0.065	0.079	0.092	0.105	0.118	0.131	0.144	0.156
	12	0.023	0.039	0.053	0.067	0.081	0.094	0.107	0.120	0.132	0.145	0.158
	16	0.026	0.041	0.055	0.069	0.082	0.095	0.108	0.121	0.134	0.146	0.159
	20	0.029	0.043	0.057	0.071	0.084	0.097	0.110	0.122	0.135	0.148	0.160
	24	0.031	0.045	0.059	0.072	0.085	0.098	0.111	0.124	0.136	0.149	0.161
	28	0.033	0.047	0.061	0.074	0.087	0.100	0.113	0.125	0.138	0.150	0.163
	32	0.035	0.049	0.062	0.075	0.088	0.101	0.114	0.126	0.139	0.151	0.164
Width	4	0.016	0.033	0.045	0.054	0.062	0.070	0.076	0.082	0.088	0.092	0.098
	8	0.020	0.036	0.049	0.058	0.066	0.072	0.078	0.084	0.090	0.096	0.100
	12	0.023	0.039	0.053	0.061	0.069	0.075	0.081	0.087	0.092	0.098	0.103
	16	0.026	0.041	0.055	0.065	0.071	0.078	0.084	0.089	0.095	0.100	0.105
	20	0.029	0.043	0.057	0.068	0.075	0.081	0.087	0.092	0.097	0.103	0.107
	24	0.031	0.045	0.059	0.071	0.077	0.083	0.089	0.095	0.099	0.105	0.109
	28	0.033	0.047	0.061	0.074	0.080	0.086	0.092	0.097	0.102	0.107	0.112
	32	0.035	0.049	0.062	0.075	0.083	0.089	0.094	0.099	0.105	0.109	0.114

Table 4: Confidence intervals for $SH^2(F)$ for empirical factor models

This table reports the sample maximum squared Sharpe ratio $sh^2(F)$ and a 90% confidence interval for the true $SH^2(F)$ for several factor models: (i) the CAPM; (ii) the Fama-French (1993) three-factor model ('FF3'); (iii) an expanded version of the FF model that uses all six of the underlying size-B/M portfolios ('FF6'); (iv) an expanded version of the FF model that uses all 25 of Fama and French's size-B/M portfolios ('FF25'); (v) an augmented version of the FF model that adds returns on Fama and French's 30 industry portfolios ('FF3+30 ind'); and (vi) a 40-factor model based on the anomaly portfolios of Kozak, Nagel, and Santosh (2020) ('KNS40'). Models (i)–(v) use data for July 1963 through June 2021 from Ken French's website, while model (vi) uses data for July 1963 through December 2019 from Serhiy Kozak's website. Returns are monthly. The bottom panel shows results for maximum Sharpe ratios (not squared), equal to the square root of the corresponding number in the top panel.

	CAPM	FF3	FF6	FF25	FF3+30 ind	KNS40
$sh^2(F)$	0.017	0.035	0.088	0.175	0.135	0.627
Lower bound for $SH^2(F)$	0.005	0.013	0.047	0.089	0.046	0.429
Upper bound for $SH^2(F)$	0.037	0.058	0.120	0.189	0.129	0.646
sh(F)	0.131	0.186	0.296	0.419	0.368	0.792
Lower bound for SH(F)	0.071	0.114	0.217	0.298	0.214	0.655
Upper bound for SH(F)	0.192	0.241	0.346	0.435	0.359	0.804