Lucky Factors

Campbell R. Harvey

Duke University, Durham, NC 27708 USA
National Bureau of Economic Research, Cambridge, MA 02138 USA

Yan Liu[*]

Texas A&M University, College Station, TX 77843 USA

Current version: September 29, 2015

Abstract

We propose a new method to select amongst a large group of candidate factors — many of which might arise as a result of data mining — that purport to explain the cross-section of expected returns. The method is robust to general distributional characteristics of both factor and asset returns. We allow for the possibility of time-series as well as cross-sectional dependence. The technique accommodates a wide range of test statistics such as t-ratios. Our method can be applied to both asset pricing tests based on portfolio sorts as well as tests using individual asset returns. While our main application focuses on asset pricing, the method can be applied in any situation where regression analysis is used in the presence of multiple testing. This includes, for example, the evaluation of investment manager performance as well as time-series prediction of asset returns.

Keywords: Factors, Factor selection, Variable selection, Bootstrap, Data mining, Orthogonalization, Multiple testing, Predictive regressions, Fama-MacBeth, GRS.

[*] Current Version: September 29, 2015. First posted on SSRN: November 20, 2014. Previously circulated under the title "How Many Factors?" and "Incremental Factors". Send correspondence to: Campbell R. Harvey, Fuqua School of Business, Duke University, Durham, NC 27708. Phone: +1 919.660.7768, E-mail: cam.harvey@duke.edu . We appreciate the comments of Tony Cookson, Gene Fama, Thomas Flury, Ken French, Bryan Kelly, Hagen Kim, Marco Rossi, and Michael Stutzer as well as seminar participants at Texas A&M University, Baylor University, UC Boulder, Wharton Jacobs Levy Conference, New York, Society for Quantitative Analysis, New York, and Man Quant Conference, New York. We thank Yong Chen for supplying us with the mutual fund data. We thank Gene Fama, Ken French, Bryan Kelly and Lu Zhang for sharing their factor returns data.

1 Introduction

There is a common thread connecting some of the most economically important problems in finance. For example, how do we determine that a fund manager has "outperformed" given that there are thousands of managers and even those following random strategies might outperform? How do we assess whether a variable such as a dividend yield predicts stock returns given that so many other variables have been tried? Should we use a three-factor model for asset pricing or a new five factor model given that recent research documents that over 300 variables have been published as candidate factors? The common thread is multiple testing or data mining.

Our paper proposes a new method that enables us to better identify the flukes. The method is based on a bootstrap that allows for general distributional characteristics of the observables, a range of test statistics (e.g., R^2, t-ratios, etc.), and, importantly, preserves both the cross-sectional and time-series dependence in the data. Our method delivers specific recommendations. For example, for a p-value of 5%, our method delivers a marginal test statistic. In performance evaluation, this marginal test statistic identifies the funds that outperform or underperform. In our main application which is asset pricing, it will allow us to choose a specific group of factors, i.e., we answer the question: How many factors?

Consider the following example in predictive regressions to illustrate the problems we face. Suppose we have 100 candidate X variables to predict a variable Y. Our first question is whether any of the 100 X variables appear to be individually significant. This is not as straightforward as one thinks because what comes out as significant at the conventional level may be "significant" by luck. We also need to take the dependence among the X variables into account since large t-statistics may come in bundles if the X variables are highly correlated. Suppose these concerns have been addressed and we find a significant predictor, how do we proceed to find the next one? Presumably, the second one needs to predict Y in addition to what the first variable can predict. This additional predictability again needs to be put under scrutiny given that 99 variables can be tried. Suppose we establish the second variable is a significant predictor. When should we stop? Finally, suppose instead of predictive regressions, we are trying to determine how many factors are important in a cross-sectional regression. How should our method change in order to answer the same set of questions but accommodate the potentially time-varying risk loadings in a Fama-MacBeth type of regression?

We provide a new framework that answers the above questions. Several features distinguish our approach from existing studies.

First, we take data mining into account.[1] This is important given the collective effort in mining new factors by both academia and the finance industry. Data mining has a large impact on hypothesis testing. In a single test where a single predetermined variable X is used to explain the left-hand side variable Y, a t-statistic of 2.0 suffices to overcome the 5% p-value hurdle. When there are 100 candidate X variables and assuming independence, the 2.0 threshold for the maximal t-statistic corresponds to a p-value of 99%, rendering useless the 2.0 cutoff in single tests.[2] Our paper proposes appropriate statistical cutoffs that control for the search among the candidate variables.

While cross-sectional independence is a convenient assumption to illustrate the point of data snooping bias, it turns out to be a big assumption. First, it is unrealistic for most of our applications since almost all economic and financial variables are intrinsically linked in complicated ways. Second, a departure from independence may have a large impact on the results. For instance, in our previous example, if all 100 X variables are perfectly correlated, then there is no need for a multiple testing adjustment and the 99% p-value incorrectly inflates the original p-value by a factor of 20 ($= 0.99/0.05$). Recent work on mutual fund performance shows that taking cross-sectional dependence into account can materially change inference.[3]

Our paper provides a framework that is robust to the form and amount of cross-sectional dependence among the variables. In particular, our method maintains the dependence information in the data matrix, including higher moment and nonlinear dependence. Additionally, to the extent that higher moment dependence is difficult to measure in finite samples and this may bias standard inference, our method automatically takes sampling uncertainty (i.e., the observed sample may underrepresent the population from which it is drawn from) into account and provides inference that does not rely on asymptotic approximations.

Our method uses a bootstrap method. When the data are independent through time, we randomly sample the time periods with replacement. Importantly, when we bootstrap a particular time period, we draw the entire cross-section at that point in time. This allows us to preserve the contemporaneous cross-sectional dependence structure of the data. Additionally, by matching the size of the resampled data with the original data, we are able to capture the sampling uncertainty of the original sample. When the data are dependent through time, we sample with blocks to capture time-series dependence, similar in spirit to White (2000) and Politis and Romano (1994). In essence, our method reframes the multiple hypothesis testing problem in

[1] Different literature uses different terminologies. In physics, multiple testing is dubbed "looking elsewhere" effect. In medical science, "multiple comparison" is often used for simultaneous tests, particularly in genetic association studies. In finance, "data mining" "data snooping" and "multiple testing" are often used interchangeably. We also use these terms interchangeably and do not distinguish them in this paper.

[2] Suppose we have 100 tests and each test has a t-statistic of 2.0. Under independence, the chance to make at least one false discovery is $1 - 0.95^{100} = 1 - 0.006 = 0.994$.

[3] See Fama and French (2010) and Ferson and Yong (2014).

regression models in a way that permits the use of bootstrapping to make inferences that are both intuitive and distribution free.

Empirically, we show how to apply our method to both predictive regression and cross-sectional regression models — the two areas of research for which data snooping bias is likely to be the most severe. However, our method applies to other types of regression models as well. Essentially, what we are providing is a general approach to perform multiple testing and variable selection within a given regression model.

Our paper adds to the recent literature on the multidimensionality of the cross-section of expected returns. Harvey, Liu and Zhu (2015) document 316 factors discovered by academia and provide a multiple testing framework to adjust for data mining. Green, Hand and Zhang (2013) study more than 330 return predictive signals that are mainly accounting based and show the large diversification benefits by suitably combining these signals. McLean and Pontiff (2015) use an out-of-sample approach to study the post-publication bias of discovered anomalies. The overall finding of this literature is that many discovered factors are likely false. But how many factors are true factors? We provide a new testing framework that simultaneously addresses multiple testing, variable selection, and test dependence in the context of regression models.

Our method is inspired by and related to a number of influential papers, in particular, Foster, Smith and Whaley (FSW, 1997) and Fama and French (FF, 2010). In the application of time-series prediction, FSW simulate data under the null hypothesis of no predictability to help identify true predictors. Our method bootstraps the actual data, can be applied to a number of test statistics, and does not need to appeal to asymptotic approximations. More importantly, our method can be adapted to study cross-sectional regressions where the risk loadings can potentially be time-varying. In the application of manager evaluation, FF (2010) (see also, Kosowski et al., 2006, Barras et al., 2010, and Ferson and Yong, 2014) employ a bootstrap method that preserves cross-section dependence. Our method departs from theirs in that we are able to determine a specific cut-off whereby we can declare that a manager has significantly outperformed or that a factor is significant in the cross-section of expected returns.[4]

Our paper is organized as follows. In the second section, we present our testing framework. In the third section, we apply our method to the selection of risk factors. We offer insights on both tests based on traditional portfolio sorts as well as raw tests based on individual assets. Some concluding remarks are offered in the final section.

[4] See Harvey and Liu (2015) for the application of our method to investment fund performance evaluation.

2 Method

Our framework is best illustrated in the context of predictive regressions. We highlight the difference between our method and the current practice and relate to existing research. We then extend our method to accommodate cross-sectional regressions.

2.1 Predictive Regressions

Suppose we have a $T \times 1$ vector Y of returns that we want to predict and a $T \times M$ matrix X that includes the time-series of M right-hand side variables, i.e., column i of matrix X (X_i) gives the time-series of variable i. Our goal is to select a subset of the M regressors to form the "best" predictive regression model. Suppose we measure the goodness-of-fit of a regression model by the summary statistic Ψ. Our framework permits the use of an arbitrary performance measure Ψ, e.g., R^2, t-statistic or F-statistic. This feature stems from our use of the bootstrap method, which does not require any distributional assumptions on the summary statistics to construct the test. In contrast, Foster, Smith and Whaley (FSW, 1997) need the finite-sample distribution on R^2 to construct their test. To ease the presentation, we describe our approach with the usual regression R^2 in mind but will point out the difference when necessary.

Our bootstrap-based multiple testing adjusted incremental factor selection procedure consists of three major steps:

Step I. Orthogonalization Under the Null

Suppose we already selected k ($0 \leq k < M$) variables and want to test if there exists another significant predictor and, if there is, what it is. Without loss of generality, suppose the first k variables are the pre-selected ones and we are testing among the rest $M - k$ candidate variables, i.e., $\{X_{k+j}, j = 1, \ldots, M - k\}$. Our null hypothesis is that none of these candidate variables provides additional explanatory power of Y, following White (2000) and FSW (1997). The goal of this step is to modify the data matrix X such that this null hypothesis appears to be true in-sample.

To achieve this, we first project Y onto the group of pre-selected variables and obtain the projection residual vector $Y^{e,k}$. This residual vector contains information that cannot be explained by pre-selected variables. We then orthogonalize the $M - k$ candidate variables with respect to $Y^{e,k}$ such that the orthogonalized variables are uncorrelated with $Y^{e,k}$ for the entire sample. In particular, we individually project $X_{k+1}, X_{k+2}, \ldots, X_M$ onto $Y^{e,k}$ and obtain the projection residuals $X^e_{k+1}, X^e_{k+2}, \ldots, X^e_M$, i.e.,

$$X_{k+j} = c_j + d_j Y^{e,k} + X^e_{k+j}, \quad j = 1, \ldots, M - k, \tag{1}$$

where c_j is the intercept, d_j is the slope and X^e_{k+j} is the residual vector. By construction, these residuals have an in-sample correlation of zero with $Y^{e,k}$. Therefore, they appear to be independent of $Y^{e,k}$ if joint normality is assumed between X and $Y^{e,k}$.

This is similar to the simulation approach in FSW (1997), in which artificially generated independent regressors are used to quantify the effect of the multiple testing. Our approach is different from FSW because we use real data. In addition, we use bootstrap and block bootstrap to approximate the empirical distribution of test statistics.

We achieve the same goal as FSW while losing as little information as possible for the dependence structure among the regressors. In particular, our orthogonalization guarantees that the $M - k$ orthogonalized candidate variables are uncorrelated with $Y^{e,k}$ in-sample.[5] This resembles the independence requirement between the simulated regressors and the left-hand side variables in FSW (1997). Our approach is distributional free and maintains as much information as possible among the regressors. We simply purge $Y^{e,k}$ out of each of the candidate variables and therefore keep all the distributional information among the variables that is not linearly related to $Y^{e,k}$ intact. For instance, the tail dependency among all the variables — both pre-selected and candidate — is preserved. This is important because higher moment dependence may have a dramatic impact on the test statistics in finite samples.[6]

A similar idea has been applied to the recent literature on mutual fund performance. In particular, Kosowski et al. (2006) and Fama and French (2010) subtract the in-sample fitted alphas from fund returns, thereby creating "pseudo" funds that exactly generate a mean return of zero in-sample. Analogously, we orthogonalize candidate regressors such that they exactly have a correlation of zero with what is left to explain in the left-hand side variable, i.e., $Y^{e,k}$.

[5] In fact, the zero correlation between the candidate variables and $Y^{e,k}$ not only holds in-sample, but also in the bootstrapped population provided that each sample period has an equal chance of being sampled in the bootstrapping, which is true in an independent bootstrap. When we use a stationary bootstrap to take time dependency into account, this is no longer true as samples on the boundary time periods are sampled less frequently. But we should expect this correlation to be small for a long enough sample as the boundary periods are a small fraction of the total time periods.

[6] See Adler, Feldman and Taqqu (1998) for how distributions with heavy tails affect standard statistical inference.

Step II. Bootstrap

Let us arrange the pre-selected variables into $X^s = [X_1, X_2, \ldots, X_k]$ and the orthogonalized candidate variables into $X^e = [X^e_{k+1}, X^e_{k+2}, \ldots, X^e_M]$. Notice that for both the residual response vector $Y^{e,k}$ and the two regressor matrices X^s and X^e, rows denote time periods and columns denote variables. We bootstrap the time periods (i.e., rows) to generate the empirical distributions of the summary statistics for different regression models. In particular, for each draw of the time index $t^b = [t_1^b, t_2^b, \ldots, t_T^b]'$, let the corresponding left-hand side and right variables be Y^{eb}, X^{sb}, and X^{eb}.

The diagram below illustrates how we bootstrap. Suppose we have five periods, one pre-selected variable X^s, and one candidate variable X^e. The original time index is given by $[t_1 = 1, t_2 = 2, t_3 = 3, t_4 = 4, t_5 = 5]'$. By sampling with replacement, one possible realization of the time index for the bootstrapped sample is $t^b = [t_1^b = 3, t_2^b = 2, t_3^b = 4, t_4^b = 3, t_5^b = 1]'$. The diagram shows how we transform the original data matrix into the bootstrapped data matrix based on the new time index.

$$[Y^{e,k}, X^s, X^e] = \underbrace{\begin{bmatrix} y_1^e & x_1^s & x_1^e \\ y_2^e & x_2^s & x_2^e \\ y_3^e & x_3^s & x_3^e \\ y_4^e & x_4^s & x_4^e \\ y_5^e & x_5^s & x_5^e \end{bmatrix}}_{\text{Original data matrix}} \begin{pmatrix} t_1 = 1 \\ t_2 = 2 \\ t_3 = 3 \\ t_4 = 4 \\ t_5 = 5 \end{pmatrix} \Rightarrow \begin{pmatrix} t_1^b = 3 \\ t_2^b = 2 \\ t_3^b = 4 \\ t_4^b = 3 \\ t_5^b = 1 \end{pmatrix} \underbrace{\begin{bmatrix} y_3^e & x_3^s & x_3^e \\ y_2^e & x_2^s & x_2^e \\ y_4^e & x_4^s & x_4^e \\ y_3^e & x_3^s & x_3^e \\ y_1^e & x_1^s & x_1^e \end{bmatrix}}_{\text{Bootstrapped data matrix}} = [Y^{eb}, X^{sb}, X^{eb}]$$

Returning to the general case with k pre-selected variables and $M - k$ candidate variables, we bootstrap and then run $M - k$ regressions. Each of these regressions involves the projection of Y^{eb} onto a candidate variable from the data matrix X^{eb}. Let the associated summary statistics be $\Psi^{k+1,b}$, $\Psi^{k+2,b}$, $\ldots$, $\Psi^{M,b}$, and let the maximum among these summary statistics be Ψ_I^b, i.e.,

$$\Psi_I^b = \max_{j \in \{1,2,\ldots,M-k\}} \{\Psi^{k+j,b}\}. \tag{2}$$

Intuitively, Ψ_I^b measures the performance of the best fitting model that augments the pre-selected regression model with one variable from the list of orthogonalized candidate variables.

The max statistic models data snooping bias. With $M - k$ factors to choose from, the factor that is selected may appear to be significant through random chance. We adopt the max statistic as our test statistic to control for multiple hypothesis testing, similar to White (2000), Sullivan, Timmermann and White (1999) and FSW (1997). Our bootstrap approach allows us to obtain the empirical distribution of the max statistic under the joint null hypothesis that none of the $M-k$ variables is true. Due to multiple testing, this distribution is very different from the null distribution of the test statistic in a single test. By comparing the realized (in the data) max statistic to this distribution, our test takes multiple testing into account.

Which statistic should we use to summarize the additional contribution of a variable in the candidate list? Depending on the regression model, the choice varies. For instance, in predictive regressions, we typically use the R^2 or the adjusted R^2 as the summary statistic. In cross-sectional regressions, we use the t-statistic to test whether the average slope is significant.[7] One appealing feature of our method is that it does not require an explicit expression for the null distribution of the test statistic. It therefore can easily accommodate different types of summary statistics. In contrast, FSW (1997) only works with the R^2.

For the rest of the description of our method, we assume that the statistic that measures the incremental contribution of a variable from the candidate list is given and generically denote it as Ψ_I or Ψ_I^b for the b-th bootstrapped sample.

We bootstrap $B = 10,000$ times to obtain the collection $\{\Psi_I^b, b = 1, 2, \ldots, B\}$, denoted as $(\Psi_I)^B$, i.e.,

$$(\Psi_I)^B = \{\Psi_I^b, b = 1, 2, \ldots, B\}. \quad (3)$$

This is the empirical distribution of Ψ_I, which measures the maximal additional contribution to the regression model when one of the orthogonalized regressors is considered. Given that none of these orthogonalized regressors is a true predictor in population, $(\Psi_I)^B$ gives the distribution for this maximal additional contribution when the null hypothesis is true, i.e., null of the $M-k$ candidate variables is true. $(\Psi_I)^B$ is the bootstrapped analogue of the distribution for maximal R^2's in FSW (1997). Similar to White (2000) and advantageous over FSW (1997), our bootstrap method is essentially distribution-free and allows us to obtain the exact distribution of the test statistic through sample perturbations.[8]

Our bootstrapped sample has the same number of time periods as the original data. This allows us to take the sampling uncertainty of the original data into account. When there is little time dependence in the data, we simply treat each time period as the sampling unit and sample with replacement. When time dependence is an issue, we use a block bootstrap, as explained in detail in the appendix. In either case, we only resample the time periods. We keep the cross-section intact to preserve the contemporaneous dependence among the variables.

[7] In cross-sectional regressions, sometimes we use the average pricing errors (e.g., mean absolute pricing error) as the summary statistics. In this case, Ψ^{eb} should be understood as the minimum among the average pricing errors for the candidate variables.

[8] We are able to generalize FSW (1997) in two significant ways. First, our approach allows us to maintain the distributional information among the regressors, helping us avoid the Bonferroni type of approximation in Equation (3) of FSW (1997). Second, even in the case of independence, our use of bootstrap takes the sampling uncertainty into account, providing a finite sample version of what is given in Equation (2) of FSW (1997).

Step III: Hypothesis Testing and Variable Selection

Working on the original data matrix X, we can obtain a Ψ_I statistic that measures the maximal additional contribution of a candidate variable. We denote this statistic as Ψ_I^d. Hypothesis testing for the existence of the $(k+1)$-th significant predictor amounts to comparing Ψ_I^d with the distribution of Ψ_I under the null hypothesis, i.e., $(\Psi_I)^B$. With a pre-specified significance level of α, say 5%, we reject the null if Ψ_I^d exceeds the $(1-\alpha)$-th percentile of $(\Psi_I)^B$, that is,

$$\Psi_I^d > (\Psi_I)_{1-\alpha}^B, \tag{4}$$

where $(\Psi_I)_{1-\alpha}^B$ is the $(1-\alpha)$-th percentile of $(\Psi_I)^B$.

The result of the hypothesis test tells us whether there exists a significant predictor among the remaining $M-k$ candidate variables, after taking multiple testing into account. Had the decision been positive, we declare the variable with the largest test statistic (i.e., Ψ_I^d) as significant and include it in the list of pre-selected variables. We then start over from Step I to test for the next predictor, if not all predictors have been selected. Otherwise, we terminate the algorithm and arrive at the final conclusion that the pre-selected k variables are the only ones that are significant.

2.2 GRS and Panel Regression Models

Our method can be adapted to study panel regression models commonly used in asset pricing tests. The idea is to demean factor returns such that the demeaned factors have zero impact in explaining the cross-section of expected returns. However, their ability to explain variation in asset returns in time-series regressions is preserved. This way, we are able to disentangle the time-series vs. cross-sectional contribution of a candidate factor.

We start by writing down a time-series regression model,

$$R_{it} - R_{ft} = a_i + \sum_{j=1}^{K} b_{ij} f_{jt} + \epsilon_{it}, i = 1, \ldots, N, \tag{5}$$

in which the time-series of excess returns $R_{it} - R_{ft}$ are projected onto K contemporaneous factor returns f_{it}. Factor returns are the long-short strategy returns corresponding to zero cost investment strategies. If the set of factors are mean-variance efficient (or, equivalently, if the corresponding beta pricing model is true), the cross-section of regression intercepts should be indistinguishable from zero. This constitutes the testable hypothesis for the Gibbons, Ross and Shanken (GRS, 1989) test.

The GRS test is widely applied in empirical asset pricing. However, several issues hinder further applications of the test, or time-series tests in general. First, the GRS test almost always rejects. This means that almost no model can adequately explain the cross-section of expected returns. As a result, most researchers use the GRS test statistic as a heuristic measure for model performance (see, e.g., Fama and French, 2015a). For instance, if Model A generates a smaller GRS statistic than Model B, we would take Model A as the "better" Model, although neither model survives the GRS test. But does Model A "significantly" outperform B? The original GRS test cannot answer answer this question because the overall null of the test is that all intercepts are strictly at zero. When two competing models both generate intercepts that are not at zero, the GRS test is not designed to measure the relative performance of the two models. Our method provides a solution to this problem. In particular, for two models that are nested, it allows us to tell the incremental contribution of the bigger model relative to the smaller one, even if both models fail to meet the GRS null hypothesis.

Second, compared to cross-sectional regressions (e.g., the Fama-MacBeth regression), time-series regressions tend to generate a large time-series R^2. This makes them appear more attractive than cross-sectional regressions because the cross-sectional R^2 is usually much lower.[9] However, why would it be the case that a few factors that explain more than 90% of the time-series variation in returns are often not even significant in cross-sectional tests? Why would the market return explain a significant fraction of variation in individual stock and portfolio returns in time-series regressions but offer little help in explaining the cross-section? These questions point to a general inquiry into asset pricing tests: is there a way to disentangle the time-series vs. cross-sectional contribution of a candidate factor? Our method achieves this by demeaning factor returns. By construction, the demeaned factors have zero impact on the cross-section while having the same explanatory power in time-series regressions as the original factors. Through this, we test a factor's significance in explaining the cross-section of expected returns, holding its time-series predictability constant.

Third, the inference for the GRS test which is based on asymptotic approximations can be problematic. For instance, MacKinlay (1987) shows that the test tends to have low power when the sample size is small. Affleck-Graves and McDonald (1989) show that nonnormalities in asset returns can severely distort its size and/or power. Our method relies on bootstrapped simulations and is thus robust to small-sample or nonnormality distortions. In fact, bootstrap based resampling techniques are often recommended to mitigate these sources of bias.

Our method tries to overcome the aforementioned shortcomings in the GRS test by resorting to our bootstrap framework. The intuition behind our method is already given in our previous discussion on predictive regressions. In particular, we orthogonalize (or more precisely, demean) factor returns such that the orthogonalized factors

[9] See Lewellen, Nagel and Shanken (2010).

do not impact the cross-section of expected returns.[10] This absence of impact on the cross-section constitutes our null hypothesis. Under this null, we bootstrap to obtain the empirical distribution of the cross-section of pricing errors. We then compare the realized (i.e., based on the real data) cross-section of pricing errors generated under the original factor to this empirical distribution to provide inference on the factor's significance. We describe our panel regression method as follows.

Without loss of generality, suppose we only have one factor (e.g., the excess return on the market $f_{1t} = R_{mt} - R_{ft}$) on the right-hand side of Equation (5). By subtracting the mean from the time-series of f_{1t}, we rewrite Equation (5) as

$$R_{it} - R_{ft} = \underbrace{[a_i + b_{i1}E(f_{1t})]}_{\text{Mean excess return}=E(R_{it}-R_{ft})} + b_{i1} \underbrace{[f_{1t} - E(f_{1t})]}_{\text{Demeaned factor return}} + \epsilon_{it}. \tag{6}$$

The mean excess return of the asset can be decomposed into two parts. The first part is the time-series regression intercept (i.e., a_i), and the second part is the product of the time-series regression slope and the average factor return (i.e., $b_{i1}E(f_{1t})$).

In order for the one-factor model to work, we need $a_i = 0$ across all assets. Imposing this condition in Equation (6), we have $b_{i1}E(f_{1t}) = E(R_{it} - R_{ft})$. Intuitively, the cross-section of $b_{i1}E(f_{1t})$'s need to line up with the cross-section of expected asset returns (i.e., $E(R_{it} - R_{ft})$) in order to fully absorb the intercepts in time-series regressions. This condition is not easy to satisfy in time-series regressions because the cross-section of risk loadings (i.e., b_i) are determined by individual time-series regressions. The risk loadings may happen to line up with the cross-section of asset returns and thereby making the one-factor model work or they may not. This explains why it is possible for some factors (e.g., the market factor) to generate large time-series regression R^2's but contribute little in explaining the cross-section of asset returns.

Another important observation from Equation (6) is that by setting $E(f_{1t}) = 0$, factor f_{1t} exactly has zero impact on the cross-section of expected asset returns. Indeed, if $E(f_{1t}) = 0$, the cross-section of intercepts from time-series regressions (i.e., a_i) exactly equal the cross-section of average asset returns (i.e., $E(R_{it} - R_{ft})$) that the factor model is supposed to help explain in the first place. On the other hand, whether or not the factor mean is zero does not matter for time-series regressions. In particular, both the regression R^2 and the slope coefficient (i.e., b_{i1}) are kept intact when we alter the factor mean.

The above discussion motivates our test design. For the one-factor model, we define a "pseudo" factor $\tilde{f}_{1t}$ by subtracting the in-sample mean of f_{1t} from its time-series. This demeaned factor maintains all the time-series predictability of f_{1t} but has no role in explaining the cross-section of expected returns. With this pseudo

[10] More precisely, our method makes sure that the orthogonalized factors have a zero impact on the cross-section of expected returns *unconditionally*. This is because panel regression models with constant risk loadings focus on unconditional asset returns.

factor, we bootstrap to obtain the distribution of a statistic that summarizes the cross-section of pricing errors (i.e., regression intercepts). Candidate statistics include mean/median absolute pricing errors, mean squared pricing errors, and t-statistics. We then compare the realized statistic for the original factor (i.e., f_{1t}) to this bootstrapped distribution.

Our method generalizes straightforwardly to the situation when we have multiple factors. Suppose we have K pre-selected factors and we want to test the $(K+1)$-th factor. We first project the $(K+1)$-th factor onto the pre-selected factors through a time-series regression. We then define the new pseudo factor by subtracting the regression intercept from the $(K+1)$-th factor. This is analogous to the previous one-factor model example. In the one-factor model, demeaning is equivalent to projecting the factor onto a constant.

We use an example to illustrate how our method works when there are multiple factors. Suppose we have one pre-selected factor (f_{1t}) in the baseline model. The regression equation for asset i is:

$$R_{it} - R_{ft} = a_i + b_{i1} f_{1t} + e_{it}. \tag{7}$$

Now suppose we add another factor f_{2t} to the baseline model and denote the augmented model as:

$$R_{it} - R_{ft} = a_i^* + b_{i1}^* f_{1t} + b_{i2}^* f_{2t} + e_{it}^*. \tag{8}$$

In general, $a_i \neq a_i^*$. Our goal is to adjust f_{2t} such that the adjusted f_{2t} (denoted as f_{2t}^*) guarantees $a_i = a_i^*$, that is, the regression intercept under the augmented model is the same as the intercept under the baseline model. Our description in the previous paragraph achieves this. In particular, let the regression equation that projects f_{2t} onto f_{1t} be:

$$f_{2t} = \alpha + \beta f_{1t} + \varepsilon_t, \tag{9}$$

and define f_{2t}^* as

$$f_{2t}^* \equiv f_{2t} - \alpha = \beta f_{1t} + \varepsilon_t. \tag{10}$$

Thus defined f_{2t}^*, when substituting f_{2t} in (8), makes sure that $a_i^* = a_i$. To see this, we replace f_{2t}^* with f_{2t} in (8) and rewrite the regression equation as:

$$\begin{aligned} R_{it} - R_{ft} &= a_i^* + b_{i1}^* f_{1t} + b_{i2}^* (\beta f_{1t} + \varepsilon_t) + e_{it}^* && (11) \\ &= a_i^* + (b_{i1}^* + b_{i2}^* \beta) f_{1t} + \underbrace{(b_{i2}^* \varepsilon_t + e_{it}^*)}_{u_{it}^*}. && (12) \end{aligned}$$

By construction, both ε_t and e_{it}^* are orthogonal to f_{1t} and a vector of ones. Hence, by treating $u_{it}^* = b_{i2}^*\varepsilon_t + e_{it}^*$ as the new regression residual, the OLS assumptions hold. Comparing (12) with (7), we must have:

$$a_i^* = a_i, \quad b_{i1}^* + b_{i2}^*\beta = b_{i1}. \tag{13}$$

Our adjustment makes economic sense. Taking unconditional expectations on both sides of (10), we have

$$E(f_{2t}^*) = \beta E(f_{1t}). \tag{14}$$

Therefore, the adjusted factor f_{2t}^* is spanned by the pre-selected factor f_{1t} in the sense that its premium is completely explained by its exposure to the pre-selected factor. When this happens, the adjusted factor has zero incremental impact on the cross-section of expected returns. In the meantime, it has perfect time-series correlation with the original factor in-sample and has the same time-series correlation with the pre-selected variable as the original factor. Hence, the adjusted factor preserves the time-series properties of the original factor aside from the mean.

With this pseudo factor, we bootstrap to generate the distribution of pricing errors. In this step, the difference from the one-factor case is that, for both the original regression and the bootstrapped regressions based on the pseudo factor, we always keep the original K factors in the model. This way, our test captures the incremental contribution of the candidate factor. When multiple testing is the concern and we need to choose from a set of candidate variables, we can rely on the max statistic (in this case, the min statistic since minimizing the average pricing error is the objective) discussed in the previous section to provide inference.

2.3 Cross-sectional Regressions

Our method can also be adapted to test factor models in cross-sectional regressions. In particular, we show how an adjustment of our method applies to Fama-MacBeth type of regressions (FM, Fama and MacBeth, 1973) — one of the most important testing frameworks that allow time-varying risk loadings.

One hurdle in applying our method to FM regressions is the time-varying slopes in cross-sectional regressions. In particular, separate cross-sectional regressions are performed for each time period to obtain a collection of cross-sectional regression slopes. These slopes reflect the variability in the risk compensation for a given factor. We test the significance of a factor by looking at the time averaged cross-sectional slope coefficient. Therefore, in the FM framework, the null hypothesis is that the slope is zero in population. We adjust our method such that this condition exactly holds in-sample for the adjusted regressors.

First, we need to orthogonalize. Suppose we run a Fama-MacBeth regression on a baseline model and obtain the panel of residual excess returns. In particular, at time t, let the vector of residual excess returns be Y_t. We are testing the incremental contribution of a candidate factor in explaining the cross-section of expected returns. Let the vector of risk loadings (i.e., β's) for the candidate factor be X_t. Suppose there are n_t assets in the cross-section at time t so the dimension of both Y_t and X_t is $n_t \times 1$. Notice that n_t can be time-dependent as it is straightforward for our method to handle unbalanced panels. In a typical Fama-MacBeth regression, we would project Y_t onto X_t. For our orthogonalization to work, we reverse the process, similar to what we do in predictive regressions. More specifically, we stack the collection of Y_t's and X_t's into two column vectors that have a dimension of $\sum_{t=1}^{T} n_t \times 1$, and run the following constrained regression model:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} + \xi_{1\times 1} \cdot \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1}, \quad (15)$$

where ϕ_t is the constant vector of intercepts for time t, $\xi_{1\times 1}$ is a scalar, and $[\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_T']'$ is the vector of projected regressors that will be used in the follow-up bootstrap analysis. This is a constrained regression as we have a single regression slope (i.e., ξ) throughout the sample. Had we allowed different slopes across time, we would have the usual unconstrained regression model where X_t is projected onto Y_t period-by-period. Having a single slope coefficient is key for us to achieve the null hypothesis in-sample for the FM model.

Alternatively, we can view the above regression model as an adaptation of the orthogonalization procedure that we use in predictive regressions. It pools returns and factor loadings together to estimate a single slope coefficient. What is different, however, is the use of separate intercepts for different time periods. This is natural since the FM procedure allows time-varying intercepts and slopes. To purge the variation in Y_t's out of X_t's, we need to allow for time-varying intercepts as well. Mathematically, the time-dependent intercepts allow the regression residuals to sum up to zero within each period. This property proves very important in that it allows us to form the FM null hypothesis in-sample, as we shall see later.

Next, we scale each residual vector ε by its sum of squares $\varepsilon'\varepsilon$ and generate the orthogonalized regressor vectors:

$$X_t^e = \varepsilon_t / (\varepsilon_t' \varepsilon_t), \ t = 1, 2, \ldots, T. \quad (16)$$

These orthogonalized regressors are the FM counterparts of the orthogonalized regressors in predictive regressions. They satisfy the FM null hypothesis in cross-sectional regressions. In particular, suppose we run cross-sectional OLS with these orthogonalized regressor vectors for each period:

$$Y_t = \mu_t + \gamma_t X_t^e + \eta_t,\ t = 1, 2, \ldots, T, \tag{17}$$

where μ_t is the $n_t \times 1$ vector of intercepts, γ_t is the scalar slope for the t-th period, and η_t is the $n_t \times 1$ vector of residuals. We show in Appendix A that the following FM null hypothesis holds in-sample:

$$\sum_{t=1}^{T} \gamma_t = 0. \tag{18}$$

The above orthogonalization is the only step that we need to adapt to apply our method to the FM procedure. The rest of our method follows for factor selection in FM regressions. In particular, with a pre-selected set of right-hand side variables, we orthogonalize the rest of the right-hand side variables to form the joint null hypothesis that none of them is a true factor. We then bootstrap to test this null hypothesis. If we reject, we add the most significant one to the list of pre-selected variables and start over to test the next variable. Otherwise, we stop and end up with the set of pre-selected variables.

2.4 Discussion

Across the three different scenarios, our orthogonalization works by adjusting the right-hand side or forecasting variables so they appear irrelevant in-sample. That is, they achieve what are perceived as the null hypotheses in-sample. However, the null differs in different regression models. As a result, a particular orthogonalization method that works in one model may not work in another model. For instance, in the panel regression model the null is that a factor does not help reduce the cross-section of pricing errors. In contrast, in Fama-MacBeth type of cross-sectional regressions, the null is that the time averaged slope coefficients is zero. Following the same procedure as what we do in panel regressions will not achieve the desired null in the cross-sectional regressions.

Our method builds on the statistics literature on bootstrap. Jeong and Maddala (1993) suggest that there are two uses of bootstrap that can be justified both theoretically and empirically. First, bootstrap provides a tractable way to conduct statistical analysis (e.g., hypothesis tests, confidence intervals, etc.) when asymptotic theory is not tractable for certain models. Second, even when asymptotic theory is available, it may not be accurate in samples of the sizes of used in applications.[11]

Our application follows this advice. First, it is a daunting task to derive asymptotic distributions given the complicated structure of the cross-section of stocks returns, e.g., unbalanced panel, cross-sectional dependency, number of firms (N) is large relative to the number of time periods (T), etc. Second, as shown in Affleck-Graves and McDonald (1989), the GRS test is distorted when the returns for test portfolios are non-normally distributed. The problem is likely to be even worse given our use of individual stocks as test assets. Our bootstrap approach allows us to overcome these difficulties and conduct robust statistical inference.

More specially, our method falls into the category of nonparametric bootstrap that is routinely used for hypothesis testing. Hall and Wilson (1991) provide two valuable guidelines for nonparametric bootstrap hypothesis testing. The first guideline, which can have a large impact on test power, is that bootstrap resampling should be done in a way that reflects the null hypothesis, even if the true hypothesis is distant from the null.[12] The second guideline is to use pivotal statistics (that is, statistics whose distributions do not depend on unknown parameters).[13]

The design of our tests closely follows these principles. Take our panel regression model as an example. The first step orthogonalization, which is core to our method, ensures that the null hypothesis that a factor has zero explanatory power for the cross-section of expected returns is exactly achieved in-sample. Our method therefore abides by the first principle and can potentially have a higher test power compared to alternative designs of the hypothesis tests. When constructing the test statistics corresponding to the panel regression model, we make sure that pivotal statistics (e.g., t-statistics of the regression intercepts) are considered along with other test statistics. We therefore also take the second principle into account in the construction of the test statistics.

[11] For other references on bootstrap and its applications to financial time series, see Li and Maddala (1996), Veall (1992, 1998), Efron and Tibshirani (1993), and MacKinnon (2006).

[12] Young (1986), Beran (1988) and Hinkley (1989) discuss the first guideline in more detail.

[13] To give an example of the use of pivotal statistics in bootstrap hypothesis testing, suppose our sample is $\{x_1, x_2, \ldots, x_n\}$ and the hypothesis under test is that the population mean equals θ_0, i.e., $H_0 : \theta = \theta_0$. One test statistic one may want to use is $\hat{\theta}^* - \theta_0$, where $\hat{\theta}^* = \sum_{i=1}^{n} x_i/n$ is the sample mean. However, this statistic is not pivotal in that its distribution depends on the population standard deviation σ, which is an unknown parameter. According to Hall and Wilson (1991), a better statistic is to divide $\hat{\theta}^* - \theta_0$ by $\hat{\sigma}^*$, where $\hat{\sigma}^*$ is the standard deviation estimate. The new test statistic $(\hat{\theta}^* - \theta_0)/\hat{\sigma}^*$ is an example of a pivotal test statistic.

3 Identifying Factors

3.1 Candidate Risk Factors

We study risk factors that have been discovered by the literature. In principle, we can apply our method to the grand task of sorting out all the risk factors that have been proposed. One attractive feature of our method is that it allows the number of risk factors to be larger than the number of test portfolios, which is infeasible in conventional multiple regression models. However, we do not pursue this in the current paper but instead focus on a selected group of prominent risk factors. The choice of the test portfolios is a major confounding issue. Different test portfolios lead to different results. In contrast, individual stocks avoid the arbitrary portfolio construction. We apply our method to both popular test portfolios and individual stocks.

In particular, we apply our panel regression method to 14 risk factors that are proposed by Fama and French (2015a), Frazzini and Pedersen (2014), Novy-Marx (2013), Pastor and Stambaugh (2003), Carhart (1997), Asness, Frazzini and Pedersen (2013), Hou, Xue and Zhang (2015), Harvey and Siddique (2000), and Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014).[14]

We first provide acronyms for factors. Fama and French (2015a) add profitability (*rmw*) and investment (*cma*) to the three-factor model of Fama and French (1993), which has market (*mkt*), size (*smb*) and book-to-market (*hml*) as the pricing factors. Hou, Xue and Zhang (2015) propose similar profitability (*roe*) and investment (*ia*) factors. Other factors include betting againist beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic et al. (2014). We treat these 14 factors as candidate risk factors and incrementally select the group of "true" factors. True is in quotation marks because there are a number of other issues such as the original set of factors that we consider. Had we considered a larger set of factors, our results could have been different. We leave these extensions to future research. Similar to Fama and French (2015a), we focus on tests that rely on time-series regressions.

[14] The factors in Fama and French (2015a), Hou, Xue and Zhang (2015), Harvey and Siddique (2000) and Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014) are provided by the authors. The factors for the rest of the papers are obtained from the authors' webpages. Across the 14 factors, the liquidity factor in Pastor and Stambaugh (2003) has the shortest length (i.e., January 1968 - December 2012). We therefore focus on the January 1968 to December 2012 period to make sure that all factors have the same sampling period.

3.2 Test Statistics

We provide three types of test statistics that are economically sensible and statistically sound. Intuitively, a good test statistic in our context should be able to tell the difference in explaining the cross-section of expected returns between a baseline model and an augmented model that adds one additional variable to the baseline model. For the panel regression model, let $\{a_i^b\}_{i=1}^N$ be the cross-section of regression intercepts for the baseline model and $\{a_i^g\}_{i=1}^N$ be the cross-section of regression intercepts for the augmented model. Our first test statistic is given by

$$EW_I^m \equiv \sum_{i=1}^N |a_i^g|/N - \sum_{i=1}^N |a_i^b|/N,$$

where EW is equal weight, 'm'= mean, and 'I'= intercept. Intuitively, EW_I^m measures the difference in the mean absolute intercept between the augmented model and the baseline model. We would expect EW_I^m to be negative if the augmented model improves the baseline model. The significance of the improvement is evaluated against the bootstrapped empirical distribution that is generated under the null hypothesis that the additional variable in the augmented model has zero incremental contribution in explaining the cross-section of expected returns.

While EW_I^m calculates the difference in the mean absolute intercept, it may not be robust to extreme observations in the cross-section, especially when we use individual stocks as test assets. We therefore also consider a robust version of EW_I^m that calculates the difference in the median absolute intercept, that is,

$$EW_I^d \equiv median(\{|a_i^g|\}_{i=1}^N) - median(\{|a_i^b|\}_{i=1}^N),$$

where $median(\cdot)$ denotes the median of a group of variables and is denoted by a superscript 'd'.

To take the uncertainty in the estimation of the intercepts into account, we also consider the difference in the mean and median absolute t-statistic of the regression intercept, denoted with a subscript 'T'. Let $\{t_i^b\}_{i=1}^N$ and $\{t_i^g\}_{i=1}^N$ be the cross-section of the t-statistics for the regression intercepts for the baseline model and the augmented model, respectively. The test statistics are given by

$$\begin{aligned} EW_T^m &\equiv \sum_{i=1}^N |t_i^g|/N - \sum_{i=1}^N |t_i^b|/N, \\ EW_T^d &\equiv median(\{|t_i^g|\}_{i=1}^N) - median(\{|t_i^b|\}_{i=1}^N). \end{aligned}$$

There are many reasons for us to consider the t-statistic instead of the original intercept. First, in a time-series regression model, by thinking of the fitted combination of zero-cost portfolios (that is, factor proxies) as a benchmark index, the t-statistic of the intercept is essentially the *information ratio* of the strategy that takes a long position in the test asset and a short position in the benchmark index. When test assets are not diversified portfolios, information ratio is a better scaled metric to gauge the economic significance of the investment strategy. This is similar to the use of the t-statistic instead of the Jensen's alpha in performance evaluation. The t-statistic of alpha — not alpha itself — tells us how "abnormal" the returns are that are produced by a fund manager.

Second, the use of the t-statistic takes the heterogeneity in return volatility into account. Suppose two stocks generate the same regression intercept by fitting a factor model. Then the degree of mispricing by the factor model, as measured by the absolute value of the regression intercept, should be higher for the stock that is less noisy. In other words, we should assign less weight to stocks that are more noisy in our panel regression model. This is particularly important when we consider individual stocks as test assets as there is a large amount of heterogeneity in return volatility for individual stocks.

Finally, as mentioned previously, our use of the t-statistic is consistent with the second principle for bootstrap hypothesis testing in Hall and Wilson (1991). The use of pivotal statistics is recommended as it can potentially improve the accuracy of the test.[15]

Another way of weighting the cross-section of regression intercepts that is economically sensible is to use value weighting. Intuitively, for two stocks that generate the same regression intercept, the mispricing of the factor model should be more significant economically for the stock that has a higher market value. Our final two test statistics therefore use the market value to weight the cross-section of regression intercepts. In particular, let $\{me_{i,t}\}_{t=1}^{T}$ be the time-series of market equity for stock i, and let $ME_t = \sum_{i=1}^{N}\{me_{i,t}\}$ be the aggregate market equity at time t. The test statistics are given by

$$\begin{aligned} VW_I &\equiv (\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{me_{i,t}}{ME_t}\times|a_i^g|)/T - (\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{me_{i,t}}{ME_t}\times|a_i^b|)/T, \\ VW_T &\equiv (\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{me_{i,t}}{ME_t}\times|t_i^g|)/T - (\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{me_{i,t}}{ME_t}\times|t_i^b|)/T. \end{aligned}$$

[15] When the null hypothesis is true (i.e., the intercept equals zero), the t-statistic of the intercept for OLS is asymptotically pivotal as its asymptotic distribution is a normal distribution $\mathcal{N}(0,1)$, which is independent of the unknown parameters in OLS (e.g., slope coefficients, error variance).

To see how VW_I value weights the cross-section of absolute intercepts, we define $VW_t^g = \sum_{i=1}^N \frac{me_{i,t}}{ME_t} \times |a_i^g|$ and rewrite the first component in the definition of VW_I as

$$(\sum_{t=1}^T \sum_{i=1}^N \frac{me_{i,t}}{ME_t} \times |a_i^g|)/T = \sum_{t=1}^T VW_t^g/T.$$

We can think of $|a_i|$ as the average level of mispricing for stock i throughout the sample. VW_t^g therefore calculates the value-weighted level of mispricing for the cross-section of assets at time t. By taking the time averaged VW_t^g, the first component in the definition of VW_I (that is, $(\sum_{t=1}^T \sum_{i=1}^N \frac{me_{i,t}}{ME_t} \times |a_i^g|)/T$) calculates the time-series average of the value-weighted level of mispricing for the augmented model. VW_I therefore evaluates the difference in the time averaged value-weighted level of mispricing between the augmented model and the baseline model. A similar interpretation applies to VW_T. Our value-weighted test statistics take the time variation in market value into account.

While we focus on the above test statistics, other weighting schemes are possible. For example, we can use volatilities to weight the cross-section of absolute intercepts. The fact that our framework allows us to consider a variety of test statistics demonstrates the flexibility of our bootstrap approach. With a few caveats in mind for the construction of a well-behaved test statistic, our approach is able to provide statistical inference for a variety of test statistics, some of which are of great interest to us from an economic perspective. Notice that EW_I^m is essentially the heuristic test statistic used in Fama and French (2015a) to evaluate the performance of their investment and profitability factors.[16] Our framework allows us to make precise statements about the statistical significance of such test statistics.

We can also interpret these test statistics from an investment perspective. However, we postpone such interpretations to later sections, where we discuss the drawbacks the GRS test in more detail.

[16] More specifically, one of the test statistics used in Fama and French (2015a) is $(\sum_{i=1}^N |a_i^g|/N)/(\sum_{i=1}^N |a_i^b|/N)$, similar to EW_I^m. We do not use Fama and French (2015a)'s test statistic because it puts more weight on the reduction in absolute intercept (between the augmented model and the baseline model) for stocks that have a larger absolute intercept. In contrast, EW_I^m weights the reduction in absolute intercept equally across stocks. With the same reduction in absolute intercept between the augmented model and the baseline model, there is no particular reason for why we should weight more on stocks with a larger intercept. We therefore use EW_I^m. Nevertheless, our results are similar if we replace EW_I^m with the test statistic in Fama and French (2015a).

3.3 Results: Portfolios as Test Assets

We first apply our method to popular test portfolios. In particular, we use the standard 25 size and book-to-market sorted portfolios that are available from Ken French's on-line data library.

Table 1 presents the summary statistics on portfolios and factors. The 25 portfolios display the usual monotonic pattern in mean returns along the size and book-to-market dimension that we try to explain. The 14 risk factors generate sizable long-short strategy returns. Nine of the strategy returns generate t-ratios above 3.0 which is the level advocated by Harvey, Liu and Zhu (2015) that takes multiple testing into account. The correlation matrix shows the existence of two groups of factors. The first group consists of book-to-market (*hml*), Fama and French (2015a)'s investment factor (*cma*), and Hou, Xue and Zhang (2015)'s investment factor (*ia*). The second group consists of Fama and French (2015a)'s profitability factor (*rmw*), Hou, Xue and Zhang (2015)'s profitability factor (*roe*), and Asness, Frazzini and Pedersen (2013)'s quality minus junk factor (*qmj*). For example, *cma* and *ia* have a correlation of 0.90, and *rmw* and *qmj* have a correlation of 0.76. These high levels of correlations might make it difficult to distinguish the factors within each of the two groups, as we shall see later.

We use the aforementioned test statistics to capture the cross-sectional goodness-of-fit of a regression model. In addition, we also include the standard GRS test statistic. However, our othogonalization design does not guarantee that the GRS test statistic of the baseline model stays the same as the test statistic when we add an othogonalized factor to the model. The reason is that, while the othogonalized factor by construction has zero impact on the cross-section of expected returns, it may still affect the residual covariance matrix. Since the GRS statistic uses the residual covariance matrix to weight the regression intercepts, it changes as the estimate for the covariance matrix changes. We think the GRS statistic is not appropriate in our framework as the weighting function is no longer optimal and may distort the comparison between candidate models. Indeed, for two models that generate the same regression intercepts, the GRS test is biased towards the model that explains a smaller fraction of variance in returns in time-series regressions. To avoid this bias, we focus on the six metrics previously defined that do not rely on a model-based weighting matrix. Again, we postpone a more detailed discussion of the GRS test to later sections.

We start by testing whether any of the 14 factors is individually significant in explaining the cross-section of expected returns. Panel A in Table 2, 3, and 4 present the results. Across the six metrics, the market factor appears to be the best among the candidate factors. For instance, as shown in Panel A of Table 2, it reduces the mean absolute regression intercept by 0.372% per month, much higher than what the other factors generate.

Table 1: **Summary Statistics, January 1968 - December 2012**

Summary statistics on portfolios and factors. We report the mean annual returns for Fama-French size and book-to-market sorted 25 portfolios and the five risk factors in Fama and French (2015a) (i.e., excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*)), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). We also report the correlation matrix for factor returns. The sample period is from January 1968 to December 2012.

Panel A: Portfolio Returns

	Low	2	3	4	High
Small	0.009	0.078	0.085	0.106	0.120
2	0.039	0.074	0.095	0.101	0.108
3	0.047	0.082	0.082	0.093	0.119
4	0.062	0.061	0.077	0.087	0.090
Big	0.046	0.061	0.053	0.059	0.069

Panel B.1: Factor Returns

	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
Mean	0.052	0.022	0.048	0.081	0.024	0.055	0.068	0.057	0.048	0.105	0.039	0.047	0.033	0.060
t-stat	[2.17]	[1.32]	[3.08]	[3.54]	[1.84]	[2.99]	[5.09]	[5.76]	[3.74]	[5.98]	[3.24]	[4.44]	[2.92]	[3.48]

Panel B.2: Factor Correlation Matrix

	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
mkt	1.00													
smb	0.30	1.00												
hml	-0.32	-0.24	1.00											
mom	-0.14	-0.03	-0.15	1.00										
skew	-0.02	-0.05	0.23	0.03	1.00									
psl	-0.05	-0.04	0.03	-0.03	0.10	1.00								
roe	-0.19	-0.39	-0.11	0.51	0.19	-0.06	1.00							
ia	-0.39	-0.26	**0.69**	0.04	0.15	0.02	0.04	1.00						
qmj	-0.54	-0.54	0.02	0.26	0.13	0.03	**0.68**	0.15	1.00					
bab	-0.09	-0.07	0.40	0.18	0.24	0.06	0.25	0.35	0.19	1.00				
gp	0.08	0.06	-0.34	0.01	-0.01	-0.03	0.34	-0.26	0.45	-0.11	1.00			
cma	-0.41	-0.16	**0.71**	0.01	0.05	0.03	-0.10	**0.90**	0.07	0.32	-0.34	1.00		
rmw	-0.21	-0.42	0.11	0.10	0.27	0.03	**0.68**	0.05	**0.76**	0.26	0.49	-0.08	1.00	
civ	0.17	0.27	0.13	-0.18	0.04	0.05	-0.26	-0.00	-0.28	0.11	-0.00	0.04	-0.10	1.00

To evaluate the significance of the market factor, we follow our method and orthogonalize the 14 factors so they have a zero impact on the cross-section of expected returns in-sample. We bootstrap to obtain the empirical distributions of the individ-

Table 2: **Portfolios as Test Assets, Equally Weighted Intercepts**

Test results on 14 risk factors using portfolios. We use Fama-French size and book-to-market sorted portfolios to test 14 risk factors. They are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*) in Fama and French (2015a), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., EW_I^m and EW_I^d), which measure the difference in equally weighted mean/median absolute intercepts, are defined in Section 4.2. GRS reports the Gibbons, Ross and Shanken (1989) test statistic.

Panel A: Baseline = No factor								Panel B: Baseline = ***mkt***					
		single test			single test				single test			single test	
Factor	EW_I^m	5*th*-percentile	*p*-value	EW_I^d	5th-percentile	*p*-value	GRS	EW_I^m	5th-percentile	*p*-value	EW_I^d	5th-percentile	*p*-value
mkt	**-0.372**	[-0.305]	(0.031)	**-0.411**	[-0.320]	(0.019)	4.290						
smb	-0.137	[-0.184]	(0.108)	-0.143	[-0.188]	(0.114)	4.402	-0.018	[-0.049]	(0.258)	-0.022	[-0.075]	(0.251)
hml	0.137	[-0.071]	(1.000)	0.145	[-0.078]	(1.000)	4.050	-0.115	[-0.074]	(0.006)	**-0.126**	[-0.085]	(0.009)
mom	0.143	[-0.066]	(1.000)	0.166	[-0.072]	(0.997)	4.302	0.051	[-0.020]	(0.999)	0.047	[-0.027]	(0.990)
skew	-0.006	[-0.025]	(0.260)	0.006	[-0.030]	(0.801)	4.454	-0.026	[-0.023]	(0.038)	-0.024	[-0.028]	(0.067)
psl	0.030	[-0.023]	(0.951)	0.038	[-0.028]	(0.966)	4.286	-0.008	[-0.007]	(0.039)	-0.017	[-0.013]	(0.030)
roe	0.340	[-0.108]	(1.000)	0.311	[-0.111]	(1.000)	4.919	0.096	[-0.029]	(1.000)	0.080	[-0.039]	(0.996)
ia	0.414	[-0.095]	(1.000)	0.469	[-0.109]	[1.000]	4.553	-0.090	[-0.053]	(0.005)	-0.065	[-0.059]	(0.038)
qmj	0.543	[-0.206]	(1.000)	0.530	[-0.220]	(1.000)	5.594	0.138	[-0.035]	(1.000)	0.121	[-0.045]	(1.000)
bab	0.038	[-0.022]	(0.989)	0.029	[-0.036]	(0.944)	**3.718**	-0.113	[-0.040]	(0.000)	-0.118	[-0.049]	(0.000)
gp	-0.030	[-0.021]	(0.025)	-0.020	[-0.032]	(0.100)	4.096	0.043	[-0.021]	(0.999)	0.042	[-0.031]	(0.992)
cma	0.304	[-0.089]	(1.000)	0.353	[-0.101]	(1.000)	4.238	**-0.128**	[-0.060]	(0.001)	-0.116	[-0.065]	(0.006)
rmw	0.185	[-0.098]	(0.997)	0.186	[-0.104]	(0.996)	4.325	0.016	[-0.015]	(0.992)	0.003	[-0.032]	(0.651)
civ	-0.179	[-0.095]	(0.004)	-0.226	[-0.098]	(0.000)	4.132	-0.055	[-0.031]	(0.007)	-0.031	[-0.036]	(0.061)
		multiple test			multiple test				multiple test			multiple test	
	min	[-0.325]	(0.034)		[-0.328]	(0.020)		*min*	[-0.083]	(0.005)		[-0.102]	(0.016)

Panel C: Baseline = ***mkt*** + ***hml***						
		single test			single test	
Factor	EW_I^m	5*th*-percentile	*p*-value	EW_I^d	5th-percentile	*p*-value
mkt						
smb	**-0.044**	[-0.085]	(0.172)	**-0.046**	[-0.092]	(0.188)
hml						
mom	0.008	[-0.008]	(0.972)	0.031	[-0.015]	(0.997)
skew	-0.001	[-0.006]	(0.312)	0.002	[-0.012]	(0.716)
psl	-0.004	[-0.005]	(0.080)	0.001	[-0.010]	(0.671)
roe	0.131	[-0.027]	(1.000)	0.164	[-0.031]	(1.000)
ia	0.027	[-0.008]	(0.999)	0.057	[-0.014]	(1.000)
qmj	0.198	[-0.036]	(1.000)	0.253	[-0.041]	(1.000)
bab	-0.010	[-0.010]	(0.043)	0.013	[-0.016]	(0.959)
gp	-0.024	[-0.013]	(0.004)	-0.007	[-0.018]	(0.186)
cma	-0.013	[-0.010]	(0.030)	0.011	[-0.014]	(0.953)
rmw	0.048	[-0.029]	(1.000)	0.055	[-0.033]	(0.998)
civ	-0.024	[-0.021]	(0.036)	-0.022	[-0.026]	(0.076)
		multiple test			multiple test	
	min	[-0.085]	(0.177)		[-0.092]	(0.209)

Table 3: **Portfolios as Test Assets, Equally Weighted T-Statistics**

Test results on 14 risk factors using portfolios. We use Fama-French size and book-to-market sorted portfolios to test 14 risk factors. They are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*) in Fama and French (2015a), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., EW_T^m and EW_T^d), which measure the difference in equally weighted mean/median absolute t-statistics of intercepts, are defined in Section 4.2.

Panel A: Baseline = No factor							Panel B: Baseline = ***mkt***					
		single test			single test			single test			single test	
Factor	EW_T^m	5th-percentile	p-value	EW_T^d	5th-percentile	p-value	EW_T^m	5th-percentile	p-value	EW_T^d	5th-percentile	p-value
mkt	**-0.636**	[1.314]	(0.000)	**-0.739**	[1.458]	(0.000)						
smb	-0.185	[-0.258]	(0.064)	-0.226	[-0.352]	(0.072)	0.539	[0.333]	(0.169)	0.303	[0.052]	(0.146)
hml	0.578	[-0.204]	(0.996)	0.606	[-0.197]	(0.998)	-0.709	[-0.309]	(0.005)	-0.698	[-0.404]	(0.017)
mom	0.601	[-0.227]	(0.999)	0.692	[-0.260]	(0.999)	0.425	[-0.138]	(0.999)	0.388	[-0.203]	(0.982)
skew	-0.043	[-0.099]	(0.156)	0.007	[-0.138]	(0.620)	-0.220	[-0.159]	(0.024)	-0.240	[-0.196]	(0.026)
psl	0.090	[-0.086]	(0.913)	0.120	[-0.096]	(0.938)	-0.092	[-0.057]	(0.017)	-0.269	[-0.102]	(0.000)
roe	1.400	[-0.286]	(1.000)	1.317	[-0.313]	(1.000)	0.786	[-0.112]	(1.000)	0.811	[-0.177]	(1.000)
ia	1.696	[-0.287]	(1.000)	1.736	[-0.299]	[1.000]	-0.614	[-0.351]	(0.005)	-0.523	[-0.393]	(0.031)
qmj	3.230	[-0.297]	(1.000)	3.285	[-0.347]	(1.000)	1.249	[-0.148]	(1.000)	1.771	[-0.218]	(1.000)
bab	0.018	[-0.083]	(0.669)	0.005	[-0.129]	(0.552)	-0.890	[-0.287]	(0.001)	-0.905	[-0.346]	(0.001)
gp	-0.130	[-0.082]	(0.017)	0.047	[-0.111]	(0.825)	0.428	[-0.162]	(0.999)	0.523	[-0.195]	(0.999)
cma	1.296	[-0.280]	(1.000)	1.368	[-0.300]	(1.000)	**-0.914**	[-0.337]	(0.000)	**-0.976**	[-0.422]	(0.001)
rmw	0.796	[-0.259]	(0.998)	0.806	[-0.259]	(0.998)	0.140	[-0.053]	(0.965)	0.236	[-0.174]	(0.895)
civ	-0.601	[-0.264]	(0.002)	-0.716	[-0.285]	(0.001)	-0.414	[-0.155]	(0.002)	-0.203	[-0.205]	(0.050)
		multiple test			multiple test			multiple test			multiple test	
	min	[-0.528]	(0.018)		[-0.584]	(0.022)	*min*	[-0.467]	(0.002)		[-0.612]	(0.006)

Panel C: Baseline = ***mkt*** + ***cma***						
		single test			single test	
Factor	EW_T^m	5th-percentile	p-value	EW_T^d	5th-percentile	p-value
mkt						
smb	0.082	[-0.190]	(0.138)	0.038	[-0.434]	(0.198)
hml	0.097	[-0.090]	(0.532)	0.022	[-0.191]	(0.344)
mom	0.089	[-0.077]	(0.966)	0.113	[-0.141]	(0.897)
skew	0.017	[-0.059]	(0.557)	0.102	[-0.150]	(0.882)
psl	-0.037	[-0.040]	(0.057)	0.205	[-0.088]	(0.997)
roe	0.979	[-0.145]	(1.000)	1.275	[-0.245]	(1.000)
ia	0.415	[-0.121]	(1.000)	0.588	[-0.210]	(1.000)
qmj	1.568	[-0.163]	(1.000)	1.911	[-0.304]	(1.000)
bab	0.102	[-0.072]	(0.982)	0.065	[-0.164]	(0.789)
gp	**-0.296**	[-0.079]	(0.000)	-0.102	[-0.109]	(0.055)
cma						
rmw	0.647	[-0.105]	(1.000)	0.697	[-0.185]	(0.999)
civ	-0.165	[-0.121]	(0.025)	**-0.191**	[-0.177]	(0.045)
		multiple test			multiple test	
	min	[-0.290]	(0.054)		[-0.529]	(0.329)

Table 4: **Portfolios as Test Assets, Value Weighted Intercepts/T-Statistics**

Test results on 14 risk factors using portfolios. We use Fama-French size and book-to-market sorted portfolios to test 14 risk factors. They are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*) in Fama and French (2015a), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., VW_I and VW_T), which measure the difference in value weighted absolute intercepts and t-statistics, are defined in Section 4.2.

Panel A: Baseline = No factor							Panel B: Baseline = ***mkt***					
		single test			single test			single test			single test	
Factor	VW_I	5*th*-percentile	*p*-value	VW_T	5th-percentile	*p*-value	VW_I	5th-percentile	*p*-value	VW_T	5th-percentile	*p*-value
mkt	**-0.382**	[-0.305]	(0.019)	**-1.146**	[1.561]	(0.000)						
smb	-0.058	[-0.074]	(0.097)	-0.192	[-0.218]	(0.065)	0.007	[-0.018]	(0.869)	0.207	[-0.097]	(0.810)
hml	0.100	[-0.054]	(0.999)	0.572	[-0.178]	(0.996)	-0.016	[-0.078]	(0.375)	0.181	[-0.482]	(0.367)
mom	0.123	[-0.067]	(1.000)	0.588	[-0.222]	(0.117)	0.047	[-0.016]	(1.000)	0.441	[-0.121]	(1.000)
skew	-0.010	[-0.026]	(0.154)	-0.058	[-0.105]	(0.117)	-0.029	[-0.023]	(0.016)	-0.312	[-0.201]	(0.010)
psl	0.024	[-0.017]	(0.943)	0.091	[-0.087]	(0.898)	-0.004	[-0.005]	(0.090)	-0.052	[-0.054]	(0.052)
roe	0.173	[-0.062]	(1.000)	0.773	[-0.216]	(0.999)	0.046	[-0.015]	(1.000)	0.446	[-0.114]	(0.999)
ia	0.317	[-0.092]	(1.000)	1.568	[-0.334]	[1.000]	0.025	[-0.055]	(0.868)	0.419	[-0.409]	(0.899)
qmj	0.373	[-0.162]	(1.000)	2.269	[-0.415]	(0.999)	0.076	[-0.022]	(1.000)	0.926	[-0.178]	(1.000)
bab	0.018	[-0.017]	(0.914)	-0.001	[-0.075]	(0.321)	**-0.063**	[-0.032]	(0.002)	**-0.597**	[-0.295]	(0.002)
gp	-0.024	[-0.022]	(0.043)	-0.117	[-0.082]	(0.025)	0.080	[-0.041]	(1.000)	1.003	[-0.359]	(1.000)
cma	0.265	[-0.099]	(1.000)	1.414	[-0.351]	(1.000)	-0.014	[-0.060]	(0.335)	0.040	[-0.439]	(0.367)
rmw	0.085	[-0.043]	(0.988)	0.393	[-0.161]	(0.978)	-0.008	[-0.015]	(0.111)	-0.061	[-0.133]	(0.127)
civ	-0.124	[-0.068]	(0.006)	-0.569	[-0.242]	(0.001)	-0.026	[-0.013]	(0.003)	-0.246	[-0.094]	(0.001)
		multiple test			multiple test			multiple test			multiple test	
	min	[-0.305]	(0.019)		[-0.516]	(0.000)	*min*	[-0.082]	(0.097)		[-0.623]	(0.059)

Panel C: Baseline = ***mkt*** + ***bab***						
		single test			single test	
Factor	VW_I	5*th*-percentile	*p*-value	VW_T	5th-percentile	*p*-value
mkt						
smb	-0.002	[-0.037]	(0.529)	0.091	[-0.250]	(0.539)
hml	0.047	[-0.054]	(1.000)	0.758	[-0.337]	(0.999)
mom	-0.002	[-0.025]	(0.429)	**-0.069**	[-0.177]	(0.181)
skew	0.002	[-0.014]	(0.753)	0.029	[-0.137]	(0.699)
psl	**-0.003**	[-0.004]	(0.101)	-0.039	[-0.045]	(0.060)
roe	0.045	[-0.017]	(1.000)	0.459	[-0.160]	(0.998)
ia	0.128	[-0.037]	(1.000)	1.465	[-0.283]	(1.000)
qmj	0.076	[-0.021]	(1.000)	0.920	[-0.209]	(1.000)
bab						
gp	0.019	[-0.028]	(0.962)	0.231	[-0.235]	(0.911)
cma	0.093	[-0.034]	(1.000)	1.123	[-0.279]	(1.000)
rmw	0.010	[-0.016]	(0.963)	0.092	[-0.207]	(0.859)
civ	0.001	[-0.008]	(0.719)	0.065	[-0.077]	(0.909)
		multiple test			multiple test	
	min	[-0.062]	(0.927)		[-0.499]	(0.676)

ual test statistics. We then evaluate the realized test statistics against these empirical distributions to provide p-values. Take, again, the results in Panel A of Table 2 as an example. The bootstrapped 5th percentile of EW_I^m for the market factor is -0.305%. This means that bootstrapping under the null, i.e., the market factor has no ability to explain the cross-section, produces a distribution of increments to the intercept. At the 5th percentile, there is an intercept reduction of 0.305%. The actual factor reduces the intercept by more than the 5th percentile, 0.372%, and we declare it significant. More precisely, by evaluating the 0.372% reduction against the empirical distribution of EW_I^m for the market factor alone, the single-test p-value for the market factor is 3.1%.

We can also bootstrap to obtain the empirical distribution of the minimum statistic. In particular, following the bootstrap procedure in Section 2, we resample the time periods. For each bootstrapped sample, we first obtain the test statistic for each of the 14 orthogonalized factors and then record the minimum test statistic across all 14 statistics. The minimum statistic is the the largest intercept reduction among the 14 factors. Since all factors are orthogonalized and therefore have no impact on the cross-section of expected returns, the minimum statistic shows what the largest intercept reduction can be just by chance and therefore controls for multiple testing. It is important that all 14 test statistics are based on the same bootstrapped sample as this controls for test correlations, as emphasized by Fama and French (2010). Lastly, we compare the realized minimum statistic with the bootstrapped distribution of the minimum statistic to provide p-values.

Panel A of Table 2 shows the results on multiple testing as well. In particular, the bootstrapped 5th percentile of EW_I^m for the minimum statistic is -0.325%, which, as expected, is lower and thus more stringent than the 5th percentile under single test (i.e., -0.305%). By evaluating the 0.372% reduction against the empirical distribution of the minimum statistic, the p-value is 3.4%. Therefore, the multiple-test p-value is 3.4%, which is higher than the single-test p-value of 3.1% but still below the 5% cutoff. We therefore also declare the market factor significant from a multiple testing perspective. Across the six metrics we consider, the market factor is the dominating factor and is significant at 5% level, both from a single-test and a multiple-test perspective.

Notice that if our goal is just to obtain the single-test p-values, we can simply run standard panel regressions, obtain the t-statistics for intercepts, and then read the p-values off the significance table — there is not so much need for bootstrap. In our framework, bootstrap is necessary as it helps us obtain the empirical distribution of the minimum statistic, which is key to multiple testing adjustment.

One interesting observation based on Table 2, 3 and 4 is that the best factor that is selected may not be the one with the lowest single test p-value. For instance, in Panel A of Table 2 and for EW_I^m, the market factor is the first factor that we select despite a lower single test p-value for *civ*. On the surface, this happens because the minimum test statistic picks the factor that has the lowest EW_I^m (i.e., highest reduction in absolute intercept), not its p-value. As a result, the market factor, which has a lower EW_I^m, is favored over *civ*.

On a deeper level, should we use a minimum test statistic that depends on the p-values instead of the levels of the EW_I^m's? We think not. The use of EW_I^m allows us to focus on the economic significance rather than the statistical significance of a factor. This is especially important for our sequential selection procedure that incrementally identifies the group of true factors. We give a higher priority to a factor that has a large reduction in absolute intercept while passing a certain statistical hurdle than a factor that has a tiny reduction in absolute intercept but having a very small p-value.[17]

While different tests uniformly identify the market factor as the first and significant risk factor, they differ when it comes to the second risk factor. When we equally weight the regression intercepts, the mean test statistic (i.e., EW_I^m) picks up *cma* whereas the median test statistic (i.e., EW_I^d) picks up *hml*. This is not surprising given that *cma* and *hml* are highly correlated (correlation coefficient = 0.71). Given that *hml* has a longer history than *cma* and that median-based test statistic is typically more robust to outliers compared to mean-based test statistic, we take *hml* as the second factor identified. It has a single-test p-value of 0.9% and a multiple-test p-value of 1.6%, both significant under the 5% cutoff. After *hml* is identified and included in the baseline model, we continue to search for the third factor. This time both EW_I^m and EW_I^d favor *smb* among the candidate factors. However, *smb* is not significant. We therefore terminate the search and conclude with a two-factor model, i.e., $mkt + hml$.

Overall, our results using equally weighted regression intercepts confirm the idea that *hml* and *mkt* are helpful in explaining the cross-section of returns of Fama-French 25 portfolios. This is expected as *hml* and Fama-French 25 portfolios use the same characteristics to sort the cross-section of stocks. What is interesting in our results is that *hml* survives after *mkt* is included. *smb* does not. This is consistent with the critique of *smb* being a true risk factor (See, e.g., Berk, 1995, Harvey, Liu and Zhu, 2015).

When we equally weight the t-statistics of the regression intercepts (i.e., EW_T^m and EW_T^d), the results are similar to the previous results in that a value factor is identified as the second risk factor. However, *cma* is picked up instead of *hml*. However, both are close and correlated.

[17] Notice that a different scaling of a factor (i.e., long-short portfolio return) will not change the test statistics or their p-values. This is because we run time-series regressions on the factors. Factor loadings adjust for different scalings. For example, when *mkt* is used as the factor, suppose we have a beta estimate of 1.0 for a certain asset. When $2 \times mkt$ is used, the beta estimate will drop to 0.5, offsetting the scaling on *mkt*. Meanwhile, neither the regression intercept nor its significance will be affected by the scaling.

Lastly, when we value weight the regression intercepts or t-statistics using the monthly updated average firm size for each Fama-French 25 portfolio,[18] neither *hml* nor *cma* is able to incrementally explain portfolio returns after the market factor is included. Instead, *bab* is identified as the second risk factor, whose p-value is 9.7% under VW_I and 5.9% under VW_T. No third factor seems significant once *bab* is included in the baseline model.

Our results using value weighting have important implications for the current practice of using portfolios as test assets in asset pricing tests. Portfolio mean returns are dispersed in the cross-section, which is good news for asset pricing tests as it can potentially increase test power. However, the cross-section is small. Indeed, the anomalous returns of the Fama-French 25 portfolios are concentrated in a few portfolios that cover small stocks. Under equal weighting, current asset pricing assets are likely to identify factors that can explain these few extreme portfolios. This provides grounds for factor dredging as it is not hard to find some factors that "accidentally" correlate with the returns of these portfolios.[19] This also makes little economic sense as portfolios that cover small stocks are less important than portfolios that cover big stocks to an average investor that invests heavily in big stocks. Our approach provides a new way to take the market value of a portfolio into account when constructing an asset pricing test.

While our results based on Fama-French 25 portfolios are interesting, we are reluctant to offer any deeper interpretation given the main drawback of the portfolio approach: tests based on characteristics-sorted portfolios are likely to be biased towards factors that are constructed using the same characteristics. In the next section, we apply our method to individual stocks and hope to provide an unbiased assessment of the 14 risk factors.

3.4 Why We Abandon the GRS

The GRS test statistic is problematic in our context from a variety of perspectives. For instance, with *mkt* as the only factor in the baseline model and by adding the orthogonalized *smb* to the baseline model, the GRS is 6.039 (not shown in table), much larger than 4.290 in Panel A of Table 2, which is the GRS for the real data with *mkt* as the only factor. This means that by adding the orthogonalized *smb*, the GRS becomes much larger. By construction, the orthogonalized *smb* has no impact on the regression intercepts. The only way it can affect the GRS is through the error covariance matrix. Hence, the orthogonalized factor makes the GRS larger by reducing the error variance estimates. This insight also explains the discrepancy between EW_I^m and the GRS in Panel A of Table 2: *mkt*, which implies a much

[18] The average firm size data for Fama-French 25 portfolios are available from Ken French's online data library.

[19] See Lewellen et al. (2010) for a similar argument.

smaller mean absolute intercept in the cross-section, has a larger GRS than *bab* as *mkt* absorbs a larger fraction of variance in returns in time-series regressions and thereby putting more weight on regression intercepts compared to *bab*.

The weighting in the GRS does not seem appropriate for model comparison when none of the candidate models is expected to be the true model, i.e., the true underlying factor model that fully explains the cross-section of expected returns. Between two models that imply the same time-series regression intercepts, it favors the model that explains a smaller fraction of variance in returns. This does not make sense. We choose to focus on the six metrics that do not depend on the error covariance matrix estimate.

The way that the GRS test uses the residual covariance matrix to scale regression intercepts is likely to become even more problematic when we use individual stocks as test assets. Given a large cross-section and a limited time-series, the residual covariance matrix will be poorly measured. To make things worse, this covariance matrix needs to be inverted to obtain the weights for intercepts. As a result, the GRS test is likely to be very unstable and potentially distorted when applied to individual stocks.

Our findings about the GRS test resonate with a recent study by Fama and French (2015b). They find that the GRS test often implies unrealistically large short positions on certain assets, which does not make economic sense. To explain their findings, notice that the GRS test can be interpreted as the difference between the Sharpe ratio constructed using both the left-hand side assets and the right-hand side factors (call this Sharpe ratio SR_1) and the Sharpe ratio using only the right-hand side factors (call this Sharpe ratio SR_2). A rejection is found if SR_1 is significantly larger than SR_2. What Fama and French (2015b) find is that certain left-hand side assets need to take extreme short positions in order to achieve SR_1. By imposing short sale constraints, SR_1 is often much smaller, reducing the contribution of the left-hand side assets to the tangency portfolio formed using the right-hand side factors alone. This causes us to question the economic usefulness of the GRS test.

Our framework provides an economically meaningful approach to evaluate the incremental contribution of SR_1 over SR_2. In a panel regression model, the regression intercepts capture mispricing for the assets in the cross-section. An investor who is trying to exploit this mispricing will be long assets that have positive intercepts and short assets that have negative intercepts. By taking equally-weighted positions in the cross-section, the abnormal return for her portfolio (that is, returns with factor risks purged out) equals the equally weighted absolute intercepts plus a residual component that is the equally weighted average of the regression residuals. When we have a large cross-section — which will be the case when we use individual stocks as test assets — the residual component will be small. Therefore, the equally weighted absolute intercepts — key to our definition of EW_I^m — captures the abnormal return earned by an investor that tries to exploit the mispricing of the cross-section of assets relative to a factor model.

We have motivated our first test statistic, EW_I^m. An obvious extension is to take the heterogeneity of residual volatilities into account and use the residual volatilities weighted intercepts. This motivates our test statistics (e.g., EW_T^m) that are based on the t-statistics.[20] Finally, an average investor in the economy will invest in proportion to the market capitalizations of assets. Hence, a value-weighted metric may better reflect the economic significance of asset mispricing in the cross-section. This motivates our last two test statistics, e.g., VW_I and VW_T.

3.5 Results: Individual Stocks as Test Assets

- Challenge in using individual stocks (unbalanced panel, noise) and why our method is advantageous over existing methods in dealing with these issues.
- Describe some of the details in the implementation that are unique to individual stocks
- Present results
- Discuss findings; link to results using portfolios

3.6 Robustness

- Russell 1000&1500
- Industry portfolios
- Stationarity? Use block bootstrap
- Lagged factors

- Summarize findings
- Relate to Roll (2015) and Shanken et al.
- Discuss extensions (e.g., time-varying factor loadings)

[20] Technically, a t-statistics is not the same as the regression intercept divided by the residual volatility. It is also a function of the covariance matrix of the regressors. However, this difference is inconsequential for our application as we regress asset returns on the same set of regressors (that is, factors). We therefore use the t-statistic of the intercept for simplicity. Our choice is also consistent with the literature on performance evaluation.

4 Conclusions

We present a new method that allows researchers to meet the challenge of multiple testing in financial economics. Our method is based on a bootstrap and allows for general distributional characteristics, cross-sectional as well as time-series dependency, and a range of test statistics.

Our applications at this point are only illustrative. However, our method is general. It can be used for time-series prediction. The method applies to the evaluation of fund management. Finally, it allows us, in an asset pricing application, to address the problem of lucky factors. In the face of hundreds of candidate variables, some factors will appear significant by chance. Our method provides a new way to separate the factors that are lucky from the ones that explain the cross-section of expected returns.

Finally, while we focus on the asset pricing implications, our technique can be applied to any regression model that faces the problem of multiple testing. Our framework applies to many important areas of corporate finance such as the variables that explain the cross-section of capital structure. Indeed, there is a growing need for new tools to navigate the vast array of "big data". We offer a new compass.

References

Adler, R., R. Feldman and M. Taqqu, 1998, A practical guide to heavy tails: Statistial techniques and applications, *Birkhäuser.*

Affleck-Graves, J. and B. McDonald, 1989, Nonnormalities and tests of asset pricing theories, *Journal of Finance 44, 889-908.*

Ahn, D., J. Conrad and R. Dittmar, 2009, Basis assets, *Review of Financial Studies 22, 5133-5174.*

Asness, C., A. Frazzini and L.H. Pedersen, 2013, Quality minus junk, *Working Paper.*

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

Beran, R., 1988, Prepivoting test statistics: A bootstrap view of asymptotic refinements, *Journal of the American Statistical Association 83, 682-697.*

Berk, J.B., 1995, A critique of size-related anomalies, *Review of Financial Studies 8, 275-286.*

Bernard, H., B.T. Kelly, H.N. Lustig, and S. Van Nieuwerburgh, 2014, The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *Working Paper.*

Carhart,M.M., On persistence in mutual fund performance, *Journal of Finance 52, 57-82.*

Ecker, F., Asset pricing tests using random portfolios, *Working Paper, Duke University.*

Efron, 1987, Better bootstrap confidence intervals, *Journal of the American Statistical Associations 82, 171-185.*

Efron, B. and R.J. Tibshirani, 1993, *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Fama, E.F. and J.D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy 81, 607-636.*

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

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

Fama, E.F. and K.R. French, 2015a, A five-factor asset pricing model, *Journal of Financial Economics 116, 1-22.*

Fama, E.F. and K.R. French, 2015b, Incremental variables and the investment opportunity set, *Journal of Financial Economics 117, 470-488.*

Ferson, W.E. and Y. Chen, 2014, How many good and bad fund managers are there, really? *Working Paper, USC.*

Foster, F. D., T. Smith and R. E. Whaley, 1997, Assessing goodness-of-fit of asset pricing models: The distribution of the maximal R^2, *Journal of Finance 52, 591-607.*

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

Gibbons, M.R., S.A. Ross and J. Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica 57, 1121-1152.*

Green, J., J.R. Hand and X.F. Zhang, 2013, The remarkable multidimensionality in the cross section of expected US stock returns, *Working Paper, Pennsylvania State University.*

Hall, P., 1988, Theoretical comparison of bootstrap confidence intervals (with Discussion), *Annals of Statistics 16, 927-985.*

Hall, P. and S.R. Wilson, 1991, Two guidelines for bootstrap hypothesis testing, *Biometrics 47, 757-762.*

Harvey, C.R. and Akhtar Siddique, 2000, Conditional skewness in asset pricing tests, *Journal of Finance, 55, 1263-1295.*

Harvey, C.R., Y. Liu and H. Zhu, 2015, ... and the cross-section of expected returns, *Forthcoming, Review of Financial Studies.*
SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2249314

Harvey, C.R. and Y. Liu, 2014, Multiple testing in financial economics, *Working Paper, Duke University.*
SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2358214

Harvey, C.R. and Y. Liu, 2015, Dissecting luck vs. skill in investment manager performance, *Work In Progress, Duke University.*

Hou, Kewei, Chen Xue and Lu Zhang, 2014, Digesting anomalies: An investment approach, *Review of Financial Studies, Forthcoming.*

Hinkley, D.V., 1989, Bootstrap significance tests, In *Proceedings of the 47th Session of the International Statistical Institute*, Paris, 29 August - 6 September 1989, 3, 65-74.

Jeong, J. and G.S. Maddala, 1993, A perspective on application of bootstrap methods in econometrics, In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds), *Handbook of Statistics*, Vol. 11. Amsterdam: North Holland, 573-610.

Lewellen, J., S. Nagel and J. Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial Economics 96, 175-194.*

Li, Q. and G.S. Maddala, 1996, Bootstrapping time series models, *Econometric Reviews 15, 115-195.*

MacKinlay, A.C., 1987, On multivariate tests of the CAPM, *Journal of Financial Economics 18, 341-371.*

MacKinnon, J.G., 2006, Bootstrap methods in econometrics, *Economic Record 82, S2-18.*

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

McLean, R.D. and J. Pontiff, 2015, Does academic research destroy stock return predictability? *Journal of Finance, Forthcoming.*

Novy-Marx, R., 2013, The other side of value: The gross profitability premium, *Journal of Financial Economics 108, 1-28.*

Pástor, L. and R.F. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy 111(3).*

Politis, D. and J. Romano, 1994, The Stationary Bootstrap, *Journal of the American Statistical Association 89, 1303-1313.*

Pukthuanthong, K. and R. Roll, 2014, A protocol for factor identification, *Working Paper, University of Missouri.*

Sullivan, Ryan, Allan Timmermann and Halbert White, 1999, Data-snooping, technical trading rule performance, and the bootstrap, *Journal of Finance 54, 1647-1691.*

Veall, M.R., 1992, Bootstrapping the process of model selection: An econometric example, *Journal of Applied Econometrics 7, 93-99.*

Veall, M.R., 1998, Applications of the bootstrap in econometrics and economic statistics, In D.E.A. Giles and A. Ullah (eds), *Handbook of Applied Economic Statistics.* New York: Marcel Dekker, chapter 12.

White, Halbert, 2000, A reality check for data snooping, *Econometrica 68, 1097-1126.*

Young, A., 1986, Conditional data-based simulations: Some examples from geometrical statistics, *International Statistical Review 54, 1-13.*

A Proof for Fama-MacBeth Regressions

The corresponding objective function for the regression model in equation (15) is given by:

$$\mathcal{L} = \sum_{t=1}^{T} [X_t - (\phi_t + \xi Y_t)]'[X_t - (\phi_t + \xi Y_t)]. \tag{19}$$

Taking first order derivatives with respect to $\{\phi_t\}_{t=1}^T$ and ξ, respectively, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \phi_t} &= \sum_{t=1}^{T} \iota_t' \varepsilon_t = 0, \ t = 1, \ldots, T, \qquad (20) \\ \frac{\partial \mathcal{L}}{\partial \xi} &= \sum_{t=1}^{T} Y_t' \varepsilon_t = 0, \qquad (21) \end{aligned}$$

where ι_t is a $n_t \times 1$ vector of ones. Equation (20) says that the residuals within each time period sum up to zero, and equation (21) says that the Y_t's are on average orthogonal to the ε_t's across time. Importantly, Y_t is not necessarily orthogonal to ε_t within each time period. As explained in the main text, we next define the orthogonalized regressor X_t^e as the rescaled residuals, i.e.,

$$X_t^e = \varepsilon_t / (\varepsilon_t' \varepsilon_t), \ t = 1, \ldots, T. \tag{22}$$

Solving the OLS equation (17) for each time period, we have:

$$\begin{aligned} \gamma_t &= (X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime} (Y_t - \mu_t), \qquad (23) \\ &= (X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime} Y_t - (X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime} \mu_t, \ t = 1, \ldots, T. \qquad (24) \end{aligned}$$

We calculate the two components in equation (24) separately. First, notice X_t^e is a rescaled version of ε_t. By equation (20), the second component (i.e., $(X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime} \mu_t$) equals zero. The first component is calculated as:

$$\begin{aligned} (X_t^{e\prime} X_t^e)^{-1} X_t^{e\prime} Y_t &= [(\frac{\varepsilon_t}{\varepsilon_t' \varepsilon_t})'(\frac{\varepsilon_t}{\varepsilon_t' \varepsilon_t})]^{-1} (\frac{\varepsilon_t}{\varepsilon_t' \varepsilon_t})' Y_t, \qquad (25) \\ &= \varepsilon_t' Y_t, \ t = 1, \ldots, T, \qquad (26) \end{aligned}$$

where we again use the definition of X_t^e in equation (25). Hence, we have:

$$\gamma_t = \varepsilon_t' Y_t, \ t = 1, \ldots, T. \tag{27}$$

Finally, applying equation (21), we have:

$$\sum_{t=1}^{T} \gamma_t = \sum_{t=1}^{T} \varepsilon_t' Y_t = 0.$$

B A Simulation Study

C The Block Bootstrap

Our block bootstrap follows the so-called stationary bootstrap proposed by Politis and Romano (1994) and subsequently applied by White (2000) and Sullivan, Timmermann and White (1999). The stationary bootstrap applies to a strictly stationary and weakly dependent time-series to generate a pseudo time series that is stationary. The stationary bootstrap allows us to resample blocks of the original data, with the length of the block being random and following a geometric distribution with a mean of $1/q$. Therefore, the smoothing parameter q controls the average length of the blocks. A small q (i.e., on average long blocks) is needed for data with strong dependence and a large q (i.e., on average short blocks) is appropriate for data with little dependence. We describe the details of the algorithm in this section.

Suppose the set of time indices for the original data is $1, 2, \ldots, T$. For each bootstrapped sample, our goal is to generate a new set of time indices $\{\theta(t)\}_{t=1}^{T}$. Following Politis and Romano (1994), we first need to choose a smoothing parameter q that can be thought of as the reciprocal of the average block length. The conditions that $q = q_n$ needs to satisfies are:

$$0 < q_n \leq 1, q_n \to 0, nq_n \to \infty.$$

Given this smoothing parameter, we follow the following steps to generate the new set of time indices for each bootstrapped sample:

- Step I. Set $t = 1$ and draw $\theta(1)$ independently and uniformly from $1, 2, \ldots, T$.
- Step II. Move forward one period by setting $t = t + 1$. Stop if $t > T$. Otherwise, independently draw a uniformly distributed random variable U on the unit interval.
    1. If $U < q$, draw $\theta(t)$ independently and uniformly from $1, 2, \ldots, T$.
    2. Otherwise (i.e., $U \geq q$), set $\theta(t) = \theta(t-1) + 1$ if $\theta(t) \leq T$ and $\theta(t) = 1$ if $\theta(t) > T$.
- Step III. Repeat step II.

For most of our applications, we experiment with different levels of q and show how our results change with respect to the level of q.

D FAQ

D.1 General Questions

- *Can we "test down" for variable selection instead of "testing up" ? (Section 2)*

  Our method does not apply to the "test down" approach. To see why this is the case, imagine that we have 30 candidate variables. Based on our method, each time we single out one variable and measure how much it adds to the explanatory power of the other 29 variables. We do this 30 times. However, there is no baseline model across the 30 tests. Each model has a different null hypothesis and we do not have an overall null.

  Besides this technical difficulty, we think that "testing up" makes more sense for finance applications. For finance problems, as a prior, we usually do not believe that there should exist hundreds of variables explaining a certain phenomenon. "Testing up" is more consistent with this prior.