Slope Factors Outperform:

Evidence from a Large Comparative Study

Siddhartha Chib[*], **Yi Chun Lin**[†], **Kuntara Pukthuanthong**[‡], **Xiaming Zeng**[§]

December 2021

[*]Olin School of Business, Washington University in St. Louis, 1 Bookings Drive, St. Louis, MO 63130. E-mail: chib@wustl.edu

[†]Department of Economics, Washington University in St. Louis, 1 Bookings Drive, St. Louis, MO 63130. E-mail: l.yichun@wustl.edu

[‡]Department of Finance, Trulaske College of Business, University of Missouri, Columbia MO 65203. Email: pukthuantthongk@missouri.edu

[§]Investment professional. E-mail: zengxiaming@wustl.edu

Slope Factors Outperform:
Evidence from a Large Comparative Study

Abstract

Does the method used to construct long-short factors from firm-level characteristics change the factor risks that are incorporated in the cross-section, and the pricing performance of those risk-factors? Starting from 62 firm-level characteristics, we construct and compare pure-play slope factors, differential factors, and rank factors. For each set of 62 factors, plus the market factor, we examine the posterior distributions of the SDF coefficients, and apply a Bayesian pricing criteria to determine the evidence in favor of pricing for a large collection of assets. The evidence shows that the slope risk-factors strongly dominate, with important implications for empirical finance.

JEL Classification: G11, G12, G14

Keywords: asset pricing, Bayesian, characteristics, factors, risk factors

1 Introduction

An important question, with relevance to the vast literature on pricing of the cross-section, is whether the method used to construct factors from firm-level characteristics has a bearing on the pricing performance of those factors. Recently, Fama and French (2020) show that *slope factors* constructed as the estimated OLS slopes from cross-sectional regressions of firm excess returns on standardized lagged characteristics provide better pricing of average returns on some test assets than the corresponding *differential factors* that are constructed from the difference in value-weighted returns of stocks in the extremes of 2 by K sorts, sorted by market capitalization and lagged characteristics. Our goal in this paper is to further study this important question to determine if their finding holds more generally.

We proceed by expanding on Fama and French (2020) in three dimensions. First, we broaden the scope of the comparison to include *rank factors* constructed according to the method described in Asness, Frazzini, and Pedersen (2019), Kelly, Pruitt, and Su (2019), Chen, Pelger, and Zhu (2020), Freyberger, Neuhierl, and Weber (2020), and Kozak, Nagel, and Santosh (2020), where after grouping excess returns by market cap, factors are constructed using the normalized rank of lagged characteristics as weights.

Second, we go beyond the five factors in Fama and French (2020) and construct, for each method, a collection of 62 factors from firm-level characteristics based on Green, Hand, and Zhang (2017). It turns out that this extension is important because slope factors constructed from a limited set of characteristics, what we refer to as *limited-pure-play* slope factors, perform worse in pricing than the pure-play factors that we construct from our broad set of 62 characteristics. Fama and French (2020) dealt with only limited-pure-play slope factors and did not mention the importance of starting from a larger pool of characteristics.

Third, to precisely quantify the relative worth of the different factors in pricing the cross-section, we consider the pricing performance of only those factors that are common sources of factor risks (namely, the risk-factors). Working with risk-factors turns out to be important. Pricing

comparisons done with the factors (not the risk-factors), as in Fama and French (2020), can provide less precise information about the pricing capabilities of the different factor construction methods.

Our analysis is underpinned by methodologies for determining the risk-factors and for determining if a given test asset is priced. Both methodologies are based on Bayesian thinking. For risk-factor discovery, we use Bayesian MCMC methods to derive the marginal posterior distributions of the SDF coefficients. We use these posterior distributions to determine which factor risks are strongly supported by the data. The latter are the risk-factors. For pricing, we use Bayesian model comparison principles to develop a criteria that gives the evidence *in favor of pricing*. In this approach, we classify a test asset as priced if the posterior odds in favor of pricing exceed at least 2:1.

We obtain several noteworthy results. To begin, our analysis shows the importance of constructing slope factors from a broad set of initial characteristics. We also show that the factor risks that are incorporated in the cross-section vary depending on how the factors are constructed. This interesting, and surprising finding, appears to be new. In particular, we document that along with the market factor, nineteen of the slope factors are risk-factors, and two differential and four rank factors are risk-factors. Only one underlying characteristic, namely, volatility of liquidity (std_turn), is represented in each of these method-specific risk-factor collections. Thus, it appears that the factors from the different methods do not represent the same risks, even though each is a long-short portfolio designed to capture the (potential) premium associated with the same characteristics.

The most direct link to the risk in the underlying characteristics is offered by the slope factors. By their very construction, as the OLS slopes from cross-sectional regressions, slope factors are *pure-play* long-short portfolios.[1] Each slope factor has unit weighted exposure to the corresponding standardized characteristic and zero weighted exposure to the other standardized characteristics in the regression. Thus, these pure-play portfolios load on the risk of that

[1]For further discussion, one can see Fama (1976), Back, Kapadia, and Ostdiek (2013), Back, Kapadia, and Ostdiek (2015) and Appendix A of this paper.

characteristic alone, which is exactly what we want when we try to represent characteristics in terms of factors. On the other hand, the differential and rank factors are not pure-play factors. Thus, EAR_D constructed by the second method, or EAR_R constructed by the third method, do not *purely* represent the risk associated with the earnings announcement return characteristic.[2] Because *ear* is likely correlated with many other firm characteristics, sorting can only provide partial control for this dependence on other characteristics. Unsatisfactorily, therefore, a factor made by the differential or rank construction methods captures the risk of that characteristic, along with the risks of those other characteristics it is correlated with. Because of this correlation, it can be difficult to infer which risks are precisely represented by the differential and rank risk-factors.

The central objective of this study is to document the pricing performance of these different method-specific risk-factors. From time series regression models with various test assets, including excess returns on approximately 6000 individual stocks and 1480 ETFs, and the inferred risk-factors of each method, we find that the pricing performance of the slope risk-factors dominates that of the differential and rank risk-factors. Since the pure-play property is the main differentiating feature of slope risk-factors, one can conclude that this property is valuable. Thus, our extensive study of pricing performance across a large number of test assets corroborates and extends the finding of Fama and French (2020) about the relative superiority of slope factors. This conclusion has significant implications for empirical asset pricing.

The rest of the paper is organized as follows. In Section 2, for pedagogical completeness, we briefly review the three factor construction methods. In Section 3 we discuss the Bayesian methodology for risk-factor discovery that can be applied to our large pool of factors to infer the risk-factors most supported by the data. In Section 4 we first detail a Bayesian pricing criteria to assess if a testing asset is priced, and detail the application of this criteria to determine pricing performance of the different risk-factor collections on portfolios, ETFs and stocks. Section 5 concludes.

[2]**Notation**: Moving forward, we let C_j denote the factor corresponding to characteristic c_j. The method used to construct the factor is indicated by a subscript M, where $\text{M} \in \{\text{S}, \text{D}, \text{R}\}$, where S stands for slope factor, D for differential factor and R for rank factor.

2 Factor Construction Methods

2.1 Data

We collect monthly stock returns data from CRSP. The set of characteristics are those considered in Green et al. (2017) and Gu, Kelly, and Xiu (2020), and are sourced from Compustat and I/B/E/S. Our data contain information from 14,860 firms on 102 characteristics for the period January 1972 to December 2020. On average, there are 114 months of data on a firm and 29 firms exist for the entire time span.

For our analysis, we begin the sample from January 1989, which is the earliest month for which complete data on our selected characteristics are available in the I/B/E/S data set. Since our aim is to be able to estimate cross-sectional regressions for firms that have a complete set of characteristics in the preceding cross-sections, it is necessary to remove firms and characteristics from the sourced data that do not meet this criteria. This means that we have to drop firms with missing characteristics and, in addition, we have to drop characteristics that are highly collinear to avoid multicollinearity in the OLS cross-sectional estimations. With these general aims in mind, we proceed as follow to create a clean, workable sequence of sample cross-sectional data-sets.

First, we drop a characteristic that is missing for a large number of firms in the pooled sample. We use the rule that missingness is excessive if that characteristic is unavailable for more than 5500 firms in the pooled data. Second, we drop characteristics that are highly collinear with other characteristics. In particular, characteristics that have variance inflation factors greater than 7 are omitted. These VIFs, following Green et al. (2017), are calculated for each characteristic, as $\frac{1}{1-R^2}$, where R^2 is the multiple correlation coefficient in the pooled regression of that characteristic against all other characteristics. These two steps produce 62 characteristics that we use in our subsequent analysis. Third, by scanning the cross-sections, month-by-month, we remove firms that do not have complete data on the selected 62 characteristics in the previous month. Equivalently, this amounts to including all firms that have non-missing values for the 62 characteristics in that

month plus non-missing returns in the following month. These steps produce a rich collection of cross-sectional data sets in which the minimum number of firms is 994 and the maximum number of firms is 1970. We summarize these data, by characteristics, in Table 1.

Note that the cross-section data sets we create allow us to estimate the cross-sectional regressions without facing sample size or multicollinearity problems. If the intention was only to construct the differential and rank factors, a larger data set could be used since the construction of those factors does not require that a firm has data on the complete set of 62 characteristics. Nevertheless, to conduct a fair comparison, we apply the different factor construction methods to common cross-sectional data sets.

Finally, after we construct the three different sets of 62 factors, we augment each set with the market excess return factor, Mkt. This factor is obtained from Professor Kenneth French's data library.

2.2 Slope factors

As first stated in Fama (1976), the OLS estimates of the coefficients in cross-sectional regressions of excess returns on standardized lagged characteristics are *pure-play* long-short portfolios. Specifically, these OLS coefficients give unit weighted exposure to each standardized lagged characteristic in the cross-section regression *and* zero weighted exposure to all the other standardized lagged characteristics. Thus, these OLS estimates are characteristic specific long-short portfolios that load entirely on that characteristic. Since this fundamental property is not that well known, we provide a simple explanation of it in Appendix A.

In our study, we construct these slope factors from cross-sectional regressions with stock premiums on the LHS and 62 lagged characteristics on the RHS, where each of these lagged characteristics is standardized within each cross-section (see Coqueret (2021) for a theoretical model in which characteristics are a source of return heterogeneity). In particular, we estimate the

cross-sectional regressions

$$r_{it} = \alpha_{it} + \sum_{j=1}^{62} \beta_{j,t} c_{j,it-1} + \varepsilon_{it},\ i = 1, \ldots n_t \tag{1}$$

where r_{it} is the excess return of firm i in month t, with t running from January 1989 to December 2020, α_{it} is the intercept, $\beta_{j,t}$ is the slope of the characteristic $c_{j,it-1}$, and ε_{it} is the error term which is assumed to be iid across firms in the cross-section. The total number of firms in the cross-section, n_t, varies across time.

It is important to understand that the RHS variable $c_{j,it-1}$ in these cross-sectional regressions are standardized for each characteristic j within each cross-section. Thus, the sample mean, and standard-deviation, of $c_{j,t-1} = (c_{j,1t-1}, c_{j,2t-1}, \ldots, c_{j,n_t t-1})$ are zero and one, respectively, for every j and t. Therefore, the lagged variables on the RHS are unit-less, and the slope coefficients on the RHS are in the same units as the stock returns on the LHS. It is only because of this standardization (see Appendix A equation (A.3)) that the OLS slopes become pure-play long-short portfolios.

Let the cross-sectional OLS estimates of $\beta_{j,t}$ be denoted by $\hat{\beta}_{j,t}$, for $j = 1, 2, \ldots 62$. These OLS estimates are the slope factors corresponding to the 62 characteristics at time t. We denote these time t slope factors as $\mathrm{C}_{j,\mathrm{S},t}$, where C_j emphasizes that this is the factor corresponding to characteristic c_j, and S emphasizes that this is the factor constructed by the slope factor method. A sequence of these slope factors is obtained by estimating the cross-sectional regressions for each month t. We provide summary statistics of these constructed slope factors in Table 2.

Remark 1: It is important to note that each slope factor, by the OLS property, captures the effect of $c_{j,t-1}$, purged of the influence of all other lagged characteristics on the RHS. This highlights the importance of including a broad set of relevant lagged characteristics on the RHS. While one could construct slope factors with a limited set of characteristics on the RHS, say just five, as in Fama and French (2020), we would get five slope factors whose pure-play property would, by definition, be limited to that set of characteristics. For clarity, we refer to such slope factors as *limited-pure-play* slope factors. Our results show that such limited-pure-play factors

perform worse in pricing than the pure-play factors we have constructed from our broad set of 62 characteristics.

Remark 2: Another point to note is that the pure-play slope factors, again due to the way they are constructed, tend to be less mutually correlated than limited-pure-play slope factors, and the differential and rank factors (which are constructed one characteristic at a time). As an illustration, we provide in Table 3 the correlation matrix of five pure-play slope factors and five limited-pure-play slope factors corresponding to the Fama and French (2020) characteristics. From this table, one sees that the highest correlation within the set of the pure-play slope factors is 0.207, that between the slope factor representing the size characteristic, MVE_S, and the slope factor representing the book-to-market ratio characteristic, BM_S. The corresponding limited-pure-play factors have a correlation of 0.419. As can be seen from the table, the correlation among the limited-pure-play slope factors is, in general, higher than that of the pure-play slope factors. This indicates the importance of constructing slope factors from a reasonable, broad set of characteristics. It is infeasible, however, to include all possible characteristics because of missingness problems, multicollinearity and sample size issues, as discussed above.

2.3 Differential factors

A popular (and common) way of constructing long-short factors from characteristics is by the double-sorting method of Fama and French (1993) and Fama and French (2015). In this method, long-short factors are constructed one characteristic at a time.

In each cross-section t, one divides the stock-premiums at time t into two groups, small and large, based on the median of market-capitalization, $mve_{i,t-1}$, $i \leq n_t$. Then, for the characteristic of interest, c_j, which for the moment is not *mve*, the stock premiums in each group are further sorted into (say) 10 decile groups constructed from the *lagged* values of that characteristic, $c_{j,it-1}$, $i \leq n_t$. Thus, with this double-sorting method, the stock-premiums are allocated to twenty buckets or, equivalently, an array containing 10 rows and 2 columns. A 10 by 2 array such as this is calculated

for each characteristic, excluding the size characteristic.

Next, the excess return of these 20 buckets is value-weighted, ie., multiplied by its stock market cap divided by the total market cap in that bucket. Then, a long portfolio (a portfolio that goes long on that characteristic) is constructed as the sum of the value-weighted stock excess returns in the (10,1) and (10,2) buckets. Similarly, a short portfolio is constructed as the sum of the value-weighted stock-returns in the (1,1) and (1,2) buckets. Then, the differential factor for time period t is given by the difference (differential) of these long and short portfolios. We denote the resulting differential factor by $\mathrm{C}_{j,\mathrm{D},t}$, where C_j emphasizes that this is the factor corresponding to the characteristic c_j, and D emphasizes that this is the factor constructed by the differential factor method.

Finally, for the size characteristic *mve*, the 10 by 2 sorted and value-weighted arrays made in the preceding step for the book-to market (*bm*), operating profitability ($operprof$) and asset growth (*agr*) characteristics, are used to form long-short portfolios that represent the size factor. In particular, we make a long-short size portfolio that controls for *bm* by summing the ten rows of the *bm* array down the first column and subtracting the sum of the ten rows in the *bm* array down the second column. In the same way, we use the 10 by 2 arrays of operating profit and asset growth to create long-short size portfolios that control for $operprof$ and agr. The size factor for that cross-section t is then given by the average of these three long-short portfolios. We refer to this size factor as $\mathrm{MVE}_{\mathrm{D},t}$. The descriptive statistics of the 62 differential factors are in Table 4.

Remark 3: Instead of the 10 by 2 sorts (or arrays) used above, we could have used the more common 3 by 2 or 5 by 2 sorts. We do not do this, however, because the pricing performance of the resulting differential factors is worse than that of the factors created by 10 by 2 sorts. In the interest of space, we only report the results below for differential factors created by 10 by 2 sorting. Pricing results for the other sets of differential factors are available from us.

2.4 Rank factors

Our final method for constructing long-short factors from characteristics is the rank factor method. To explain this method, for a given characteristic c_j, the long-short rank factor for time t is constructed as follows.

Just as in the previous method, in each cross-section t, one divides the stock-premiums at time t into two groups, small and large, based on the median of market-capitalization, $mve_{i,t-1}$, $i \leq n_t$. Let $I_{t0} = \{i : \text{firm } i \text{ is a small firm}\}$ and let $I_{t1} = \{i : \text{firm } i \text{ is a large firm}\}$ denote the indices of small firms and large firms at time t. Let the number of firms in each group be n_{t0} and n_{t1}, respectively. Now for each characteristic c_j, let $c_{j,0,t-1} = \{c_{j,it-1} : i \in I_{t0}\}$ be the vector of characteristics of length n_{t0} at time $(t-1)$ of all small firms, and similarly let $c_{j,1,t-1} = \{c_{j,it-1} : i \in I_{t1}\}$ be the vector of characteristics of length n_{t1} at time $(t-1)$ of all large firms. Now let $ra_{j,0,t-1}$ denote the vector of ranks of the values in $c_{j,0,t-1}$, and let $r_{j,0,t-1} = \frac{ra_{j,0,t-1}}{n_{t0}+1}$ denote the normalized ranks. Likewise, let $ra_{j,1,t-1}$ denote the vector of ranks of the values in $c_{j,1,t-1}$, and let $r_{j,1,t-1} = \frac{ra_{j,1,t-1}}{n_{t1}+1}$. Further, let the sample mean of the values in $r_{j,0,t-1}$ be denoted by $\bar{r}_{j,0,t-1}$ and similarly let $\bar{r}_{j,1,t-1}$ denote the sample mean of the values in $r_{j,1,t-1}$. Now define the vectors of weights

$$w_{j,0,t-1} = \frac{r_{j,0,t-1} - \bar{r}_{j,0,t-1}}{\text{sum}|r_{j,0,t-1} - \bar{r}_{j,0,t-1}|} \text{ and } w_{j,1,t-1} = \frac{r_{j,1,t-1} - \bar{r}_{j,1,t-1}}{\text{sum}|r_{j,1,t-1} - \bar{r}_{j,1,t-1}|}$$

which each sum to zero.

Finally, let $prm_{0,t}$ and $prm_{1,t}$ be the vectors of excess returns at time t of small and large firms, respectively. The rank factor corresponding to the characteristic c_j is now defined as

$$\text{C}_{j,\text{R},t} = \text{sum}(w_{j,0,t-1} \cdot prm_{0,t}) + \text{sum}(w_{j,1,t-1} \cdot prm_{1,t})$$

where $\cdot$ is the dot-product operator for multiplying two vectors, and R emphasizes that this is the factor constructed by the rank factor method. The descriptive statistics of the 62 rank factors constructed from our sample are given in Table 5.

Remark 2 (continued): Since the differential and rank-factors are essentially constructed one characteristic at a time, controlling only for size, the pool of differential and rank factors tends to be more mutually correlated than slope factors. To exemplify this, consider once again the differential and rank factors corresponding to the Fama and French five factors. As can be seen from Table 6, the absolute correlations between AGR_D and three other differential factors are around 0.4; and the absolute correlations between AGR_R and the same three other rank factors are also around 0.4. In contrast, as discussed above in Remark 2, the correlations between AGR_S and the other four slope factors are insignificant. Besides complicating interpretation of the factors, the higher mutual correlations amongst the pool of differential and rank factors have implications for risk-factor discovery, as continued further in Remark 2 below.

3 Risk Factor Discovery

We have developed a two-part Bayesian methodology that helps to shed light on whether the factor construction method has a bearing on pricing performance. Let

$$\boldsymbol{f}_\text{M} = (Mkt, C_{1,\text{M}}, ..., C_{62,\text{M}}),\ \text{M} \in \{\text{S,D,R}\},$$

denote the three collections of factors, each inclusive of the market premium factor. To compare the relative performance of these factors in pricing test assets, in the first part of the methodology, discussed in this section, we determine which factors in $\boldsymbol{f}_\text{M}$, M$\in \{$S,D,R$\}$, are actually risk-factors for asset premiums. In the second part, discussed in Section 4, we construct a method for inferring if a given test asset is priced by the risk-factors.

3.1 Finding the risk factors

For each set of factors $\boldsymbol{f}_{\mathrm{M}}$, constructed by method $\mathrm{M} \in \{\mathrm{S,D,R}\}$, suppose that the stochastic discount factor (SDF) is given by

$$m_{\mathrm{M}} = 1 - \boldsymbol{\lambda}_{\mathrm{M}}' \Omega_{\mathrm{M}}^{-1} (\boldsymbol{f}_{\mathrm{M}} - \boldsymbol{\mu}_{\mathrm{M}}),$$

where $E[\boldsymbol{f}_{\mathrm{M}}] = \boldsymbol{\mu}_{\mathrm{M}}$ and $V[\boldsymbol{f}_{\mathrm{M}}] = \Omega_{\mathrm{M}}$ are, respectively, the mean vector and covariance matrix of $\boldsymbol{f}_{\mathrm{M}}$. This is the familiar form of the SDF except that we indicate its dependence on the method M used to construct the factors.

Let $\boldsymbol{b}_{\mathrm{M}} = \Omega_{\mathrm{M}}^{-1} \boldsymbol{\lambda}_{\mathrm{M}}$ denote the SDF coefficients. Following conventional practice, we use the non-zero elements of these coefficients to determine which factors are risk-factors. Under the SDF pricing restrictions, $E[m_{\mathrm{M}} \boldsymbol{f}_{\mathrm{M}}'] = 0$, we have that

$$\boldsymbol{\mu}_{\mathrm{M}} = \boldsymbol{\lambda}_{\mathrm{M}}, \quad \mathrm{M} \in \{\mathrm{S,D,R}\},$$

and thus $\boldsymbol{\lambda}_{\mathrm{M}}$ is identified and can be estimated from a joint model of the factors

$$\boldsymbol{f}_{\mathrm{M}} = \boldsymbol{\lambda}_{\mathrm{M}} + \varepsilon_{\mathrm{M}}, \quad \mathrm{M} \in \{\mathrm{S,D,R}\},$$

for a vector of idiosyncratic errors ε_{M} with mean zero and covariance Ω_{M}. We take the latter model to the data and calculate the posterior distribution of $(\boldsymbol{\lambda}_{\mathrm{M}}, \Omega_{\mathrm{M}})$. The marginal posterior distributions of the elements of $\boldsymbol{b}_{\mathrm{M}} = \Omega_{\mathrm{M}}^{-1} \boldsymbol{\lambda}_{\mathrm{M}}$ are obtained from this posterior distribution by transformation. We form 99% credibility intervals and use these intervals to infer which elements of $\boldsymbol{b}_{\mathrm{M}}$ are non-zero, given the data. The factors for which these SDF coefficients are inferred to be non-zero are the risk-factors. The 99% choice is designed to control the false-positives. Our experiments with simulated data show that the 99% level strikes a balance between the unavoidable

trade-offs between the false positives and false negatives.[3]

Now let $\boldsymbol{f}_{\mathrm{M},t}$, for t running from January 1989 to December 2020 denote the sample data on the factors that we have constructed above. We suppose that these data are an iid sample from the population $\boldsymbol{f}_{\mathrm{M}} = \boldsymbol{\lambda}_{\mathrm{M}} + \varepsilon_{\mathrm{M}}$, where, importantly, we suppose that the errors are distributed as multivariate-t (MVT) with ν degrees of freedom, ie.,

$$\varepsilon_{\mathrm{M}} \sim MVt(0, \boldsymbol{\Delta}_{\mathrm{M}}, \nu), \quad \mathrm{M} \in \{\mathrm{S,D,R}\},$$

parameterized in the usual way in terms of a dispersion matrix $\boldsymbol{\Delta}_{\mathrm{M}}$. The covariance matrix of this distribution is $\boldsymbol{\Omega}_{\mathrm{M}} = \frac{\nu}{\nu-2}\boldsymbol{\Delta}_{\mathrm{M}}$. We use the multivariate student-t distribution to accommodate the potentially thick tails of the factor distribution. This assumption is better supported by the data than the Gaussian.[4]

To estimate our models, we introduce the Gamma distributed iid random variables

$$\tau_t \sim Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$$

for a given value of ν. Later we repeat the analysis for different values of ν and use marginal likelihoods to determine the best value. Then, conditioned on these Gamma variables, the likelihood function, conditioned on $\{\tau_t\}$, is a product of multivariate normal densities

$$p(\boldsymbol{f}_{\mathrm{M},1:T}|\boldsymbol{\lambda}_{\mathrm{M}}, \boldsymbol{\Delta}_{\mathrm{M}}, \{\tau_t\}) = \prod_{t=1}^{T} N(\boldsymbol{f}_{\mathrm{M},t}|\boldsymbol{\lambda}_{\mathrm{M}}, \tau_t^{-1}\boldsymbol{\Delta}_{\mathrm{M}}), \quad \mathrm{M} \in \{\mathrm{S,D,R}\},$$

where we have used the standard scale mixture representation of the multivariate student-t distribution in terms of normal random variables and gamma scale variables. For the parameters

[3] One could impose an even stronger 99.9% threshold for risk-factor discovery, as we do later, but the results are qualitatively similar.

[4] Kozak et al. (2020) also estimate this joint model of the factors. However, they adopt the parameterization $\boldsymbol{f} = \boldsymbol{\Omega}\boldsymbol{b} + \varepsilon$. Moreover, they assume Gaussian errors and do not estimate $\boldsymbol{\Omega}$, using instead a plug-in estimate. For us, this parameterization is not convenient because, when $\boldsymbol{\Omega}$ is unknown, there is no easy way of estimating it as it appears in both the mean and in the distribution of the noise.

we inject some weak but informative prior information based on the priors in Chib and Zeng (2020) and Chib, Zeng, and Zhao (2020). Specifically, we suppose that

$$\boldsymbol{\lambda}_{\mathrm{M}}|\boldsymbol{\Delta}_{\mathrm{M}} \sim MVN(\boldsymbol{\lambda}_{\mathrm{M},0}, \kappa_{\mathrm{M}}\boldsymbol{\Delta}_{\mathrm{M}})$$

and

$$\boldsymbol{\Delta}_{\mathrm{M}} \sim IW(\rho_0, \boldsymbol{Q}_{\mathrm{M},0})$$

where IW is the inverse Wishart distribution. The hyperparameters

$$(\boldsymbol{\lambda}_{\mathrm{M},0}, \kappa_{\mathrm{M}}, \boldsymbol{Q}_{\mathrm{M},0}), \quad \mathrm{M} \in \{\mathrm{S,D,R}\},$$

of these priors are determined from a training sample (a sample of data prior to the estimation sample). In particular, we estimate the model on the first 30% of the data (from January 1989 to July 1997) and use the estimates from this estimation to fix $\boldsymbol{\lambda}_{\mathrm{M},0}$, a 63×1 vector and $\boldsymbol{Q}_{\mathrm{M},0}$, the 63×63 scale matrix. With 30% of the data we have enough observations to estimate these high-dimensional hyperparameters of the prior distribution. In addition, the use of a training sample leads to a prior that is more objective than what one could specify by a-priori reasoning. Of course, the training sample data is then omitted from the subsequent analysis. We use the same training sample to fix the value of κ_{M}, following the method given in Chib et al. (2020). Finally, we fix ρ_0 at $63+6$, where the choice of six ensures that we have a weakly informative, but proper (integrating to one) inverse Wishart density with finite mean and variance. By injecting reasonable prior information in this way we improve the learning of the high-dimensional covariance matrix $\boldsymbol{\Omega}_{\mathrm{M}}$ and, consequently, the learning of the SDF coefficients.

We sample the resulting posterior distribution $\pi(\boldsymbol{\lambda}_{\mathrm{M}}, \boldsymbol{\Delta}_{\mathrm{M}}, \{\tau_t\}|\boldsymbol{f}_{\mathrm{M},1:T})$ by Markov Chain Monte Carlo (MCMC) methods. Details about the MCMC steps can be found in Chib and Zeng (2020). This algorithm is available is a user-friendly R-package that can be downloaded from the home-page of the first author.

Denote the output of the MCMC simulations by

$$\left\{\boldsymbol{\lambda}_{\mathrm{M}}^{(g)}, \boldsymbol{\Omega}_{\mathrm{M}}^{(g)}\right\},\ g \leq G, \quad \mathrm{M} \in \{\mathrm{S,D,R}\},$$

where $\boldsymbol{\Omega}_{\mathrm{M}}^{(g)} = \frac{\nu}{\nu-2}\boldsymbol{\Delta}_{\mathrm{M}}^{(g)}$ and G is the MCMC sample size. Then, by the usual MCMC theory, see, for example, Chib (2001), the transformed values

$$\boldsymbol{b}_{\mathrm{M}}^{(g)} = \boldsymbol{\Omega}_{\mathrm{M}}^{(g)-1}\boldsymbol{\lambda}_{\mathrm{M}}^{(g)},\ g \leq G, \quad \mathrm{M} \in \{\mathrm{S,D,R}\},$$

are a sample from the posterior distribution of $\boldsymbol{b}_{\mathrm{M}}$.[5] We use this sample to make inferences about the SDF coefficients. For example, we use the draws $\{\boldsymbol{b}_{\mathrm{M}}^{(g)}\}_{g=1}^{m}$ to infer which components of $\boldsymbol{b}_{\mathrm{M}}$ are a posteriori non-zero.

Consider, for example, the j component, $b_{\mathrm{M},j}$. Let (l_j, u_j) denote (say) the 99% *posterior credibility interval* of this component. This is the interval such that $\int_{l_j}^{u_j} \pi(b_{\mathrm{M},j}|\ \boldsymbol{f}_{\mathrm{M},1:T})db_{\mathrm{M},j} = 0.99$, where $\pi(b_{\mathrm{M},j}|\boldsymbol{f}_{\mathrm{M},1:T})$ is the marginal posterior distribution of $b_{\mathrm{M},j}$. Note that this interval refers to an interval for $b_{\mathrm{M},j}$ and is conditioned on the given data. We obtain (l_j, u_j) from the quantiles of the simulated draws.

Now if any (l_j, u_j) interval with 99% posterior content is entirely on one, or the other side of zero, then we infer that the corresponding factor is a risk-factor. Henceforth, we denote the risk-factors determined from this computation by $\boldsymbol{f}_{\mathrm{M}}^{*}$. In the next subsection we detail which risk-factors emerge for each value of M.

Remark 4: It is important to realize that the marginal posterior distributions $\pi(b_{\mathrm{M},j}|\boldsymbol{f}_{\mathrm{M},1:T})$, $j = 1,2,...,63$, are derived from the joint distribution $\pi(\boldsymbol{b}_{\mathrm{M}}|\boldsymbol{f}_{\mathrm{M},1:T})$ of all the SDF coefficients, $\boldsymbol{b}_{\mathrm{M}} = \boldsymbol{\Omega}_{\mathrm{M}}^{-1}\boldsymbol{\lambda}_{\mathrm{M}}$. In particular, the posterior distribution of $b_{\mathrm{M},j}$ depends on the entire precision matrix $\boldsymbol{\Omega}_{\mathrm{M}}^{-1}$ and the entire factor premium vector $\boldsymbol{\lambda}_{\mathrm{M}}$. Moreover, the marginal distribution of $b_{\mathrm{M},j}$ depends on the data on all the factors $\boldsymbol{f}_{\mathrm{M},1:T}$ (as noted in the conditioning variable of the marginal

[5]Note that from a non-Bayesian perspective one could multiply together the estimates of $\boldsymbol{\Omega}_{\mathrm{M}}^{-1}$ and $\boldsymbol{\lambda}_{\mathrm{M}}$, but the sampling distribution of this estimate would be difficult to calculate.

density). In other words, these marginal densities account for the correlation amongst the factors. Put another way, these marginal densities depend on the information contained in the full joint model and full data on the entire set of factors.

3.2 Results

We now report our results on risk-factor discovery. One might expect that the factor risks that are incorporated in the cross-section of expected returns would be similar, regardless of the way the factors are constructed, because, after all, these sets of factors are constructed from the same set of characteristics on the same common data. However, this is not the case. The method-specific risk-factors are quite different. Before we explore the ramifications of this finding for pricing and for the interpretation of these risk-factors, we describe the implementation of the methodology developed in the preceding section.

As described there, the starting point of the analysis is the joint multivariate-t based model

$$\boldsymbol{f}_{\mathrm{M}} = \boldsymbol{\lambda}_{\mathrm{M}} + \varepsilon_{\mathrm{M}},$$

for each method-specific factors $\boldsymbol{f}_{\mathrm{M}}$, $\mathrm{M} \in \{\mathrm{S},\mathrm{D},\mathrm{R}\}$, which we estimate on our constructed factor data, $\boldsymbol{f}_{\mathrm{M},1:T}$, that runs from January 1989 to December 2020, for a total $T = 384$ observations. Here $\boldsymbol{\lambda}_{\mathrm{M}}$ is the 63×1 vector of factor means and ε_{M} is a vector of idiosyncratic errors distributed as multivariate-t with ν degrees of freedom, mean zero and 63×63 covariance matrix $\boldsymbol{\Omega}_{\mathrm{M}}$. In fitting each of these three models, we use a training sample prior that is based on the first 30% of the sample data. As mentioned above, the training sample approach is particularly useful for arriving at a reasonable prior (and a suitable regularizer of the likelihood) given the large sizes of $\boldsymbol{\lambda}_{\mathrm{M}}$ and $\boldsymbol{\Omega}_{\mathrm{M}}$. The resulting posterior distributions are sampled by MCMC methods, coded in a user-friendly R-package. We get 20,000 MCMC draws from the posterior distribution of $\boldsymbol{\lambda}_{\mathrm{M}}$ and $\boldsymbol{\Omega}_{\mathrm{M}}$, following a burn-in of 1000 draws, and use these draws of the parameters to calculate

marginal likelihoods by the method of Chib (1995) (to determine which ν is best supported by the data), and to get posterior draws of the SDF coefficients (to determine the risk-factors).

3.2.1 Importance of the MVT assumption

It is important to note that the multivariate-t model improves considerably on the Gaussian equivalent, as measured by the standard Bayesian marginal likelihood criteria. In the case of the slope factors, $\boldsymbol{f}_{\mathrm{S}}$, for example, the highest marginal likelihood MVT model, selected over ν on the grid $\{4, 4.5, 5, 5.5, 6, 6.5, 7\}$ along with the Gaussian distribution (when $\nu =$ Infinity), is a multivariate-t model with $\nu = 6$ degrees of freedom and a log-marginal value of -20421.30. In contrast, the log-marginal likelihood of the Gaussian equivalent has the much inferior value of -21587.91, showing the importance of adopting the MVT assumption. Similar large improvements occur for $\boldsymbol{f}_{\mathrm{D}}$ and $\boldsymbol{f}_{\mathrm{R}}$, as can be seen from the results in Table 7.

3.2.2 Posterior of SDF coefficients

We summarize the pertinent posterior estimation results in Tables 8, 9, 10, and Figures 1, 2, and 3 where we plot the 99% credibility intervals of each SDF coefficients, for each set of 63 factors. We can see from these posterior credibility intervals, that the risk-factors (identified in the plots by 99% credibility intervals that exclude zero and colored in red) are different even though each set of factors is supposed to represent the same set of underlying characteristics. This shows that the construction method has concrete implications for the implied factor mimicking portfolios for understanding which sources of factor risk are incorporated in stock equity premiums and, as we shall see below, performance in pricing excess returns in the cross-section.

From inspecting the credibility intervals in Figure 1 we can see that the data suggest that SDF

coefficients of the following factors,

$$\begin{aligned} \boldsymbol{f}_{\mathrm{S}}^{*} = \{&Mkt, \mathrm{AGR_S}, \mathrm{BETA_S}, \mathrm{CASHDEBT_S}, \mathrm{CHINV_S}, \mathrm{EP_S}, \mathrm{GRLTNOA_S}, \mathrm{INDMOM_S}, \mathrm{LEV_S}, \mathrm{LGR_S}, \\ &\mathrm{MOM1M_S}, \mathrm{MVE_S}, \mathrm{NINCR_S}, \mathrm{PS_S}, \mathrm{ROAQ_S}, \mathrm{ROEQ_S}, \mathrm{ROIC_S}, \mathrm{STD\_TURN_S}, \mathrm{TB_S}, \mathrm{TURN_S}\} \end{aligned} \tag{2}$$

are non-zero and, hence, can be taken as the slope risk-factors. There is some question about the SDF coefficients of $\{\mathrm{CASH_S}, \mathrm{CFP_S}, \mathrm{GRCAPX_S}, \mathrm{INVEST_S}, \mathrm{PRICEDELAY_S}, \mathrm{TANG_S}\}$ as the 99% credibility intervals are mostly on one side of zero. Arguably, these six could be added to the preceding set of risk-factors, but, under our 99% rule, fail to make the cut.

Continuing to the differential factors, from Figure 2, we see that the following three factors:

$$\boldsymbol{f}_{\mathrm{D}}^{*} = \{Mkt, \mathrm{CONVIND_D}, \mathrm{STD\_TURN_D}\} \tag{3}$$

have SDF coefficients that are non-zero. These are the differential risk-factors. Finally, from Figure 3, we infer that the following five factors,

$$\boldsymbol{f}_{\mathrm{R}}^{*} = \{Mkt, \mathrm{CONVIND_R}, \mathrm{STD\_TURN_R}, \mathrm{TB_R}, \mathrm{TURN_R}\} \tag{4}$$

are rank risk-factors.

Remark 2 (continued): Our results show that the SDF is sparse in the differential and rank factors. This is mainly due to the fact that the starting pool of these factors is more correlated than the starting pool of slope factors. In fact, scree plots (see Appendix B for details) show that about 15 principal components (PCs) explain 85% of the total variance of the differential and rank factors. Therefore, while there are fewer factors in $\boldsymbol{f}_{\mathrm{D}}^{*}$ and $\boldsymbol{f}_{\mathrm{R}}^{*}$, these risk-factor sets likely represent the risk of several additional characteristics.[6] Of course, one would obtain more differential and

[6] As mentioned in Footnote 3, Kozak et al. (2020) also estimate a joint model (for *only* rank factors) under the parameterization, $\boldsymbol{f} = \boldsymbol{\Omega} \boldsymbol{b} + \boldsymbol{\varepsilon}$. Applied to 50 rank factors, assuming Gaussian errors, a plug-in estimate of $\boldsymbol{\Omega}$, and a different implied prior on $\boldsymbol{\lambda}$, they find that the SDF is sparse in the rank factor PCs, but not in the rank factors

rank risk-factors under a 95% credibility interval rule, but, because of the likely increased number of false-positives, the credibility of these would be questionable. Given the limited number of differential and rank risk-factors discovered, one need not be concerned about false-positives. As for the slope factors, we discover twenty risk-factors mainly because the slope factor set is less correlated. This also shows up in the fact that 37 PCs are needed to explain 85% of the total variance of the slope factors. Even under a 99.9% rule, the number of slope risk-factors only drops to twelve, as shown in Table 11. Informally, more slope factors are involved in pricing because each represents the risk of one characteristic, not the risk of an amalgamation of characteristics and, therefore, if several risks are priced, several slope risk-factors emerge, one for each of those risks. We are, thus, confident in what the data and our analysis are telling us: that there are 3-5 times more slope risk-factors than differential and rank risk-factors.

In Table 11, we follow Hou, Xue, and Zhang (2020) and categorize the risk-factors into six characteristic-specific categories: Momentum, Value-versus-Growth, Investment, Profitability, Intangibles, and Trading frictions. One can see that the slope factors represent characteristics in all six of the categories. This is likely one reason behind the superior performance of the slope risk-factors that we document next. It is noteworthy that most risk-factors under the 99% and 99.9% rules fall under the trading frictions category, which is consistent with the finding of Hou et al. (2020). Under the 99% rule, volatility of liquidity (std_turn) is the only characteristic that shows up in all three final sets of risk-factors.

(while we find sparsity in the rank factors as well). The latter difference could be due to the difference in distributional assumptions, the fact that uncertainty of an unknown $\mathbf{\Omega}$ is not incorporated, and/or because the decision about sparsity is not based on Bayesian posterior distributions, or frequentist sampling distributions, but on whether the LASSO estimates of the SDF coefficients have shrunk to zero. The latter criteria may allow some factors to appear as risk-factors when, in fact, a decision that takes into account sampling uncertainty of the estimates, or Bayesian uncertainty of the parameters, as we do in our approach, might suggest otherwise.

4 Pricing of the cross-section

The results in the preceding section show that the risk-factors $\boldsymbol{f}_{\mathrm{S}}^{*}$, $\boldsymbol{f}_{\mathrm{D}}^{*}$ and $\boldsymbol{f}_{\mathrm{R}}^{*}$ are quite different. Clearly, the way the factors are constructed alters the make-up of the risk-factors. To assess which of these risk-factors better represent the common sources of factor risks that affect the cross-section of expected excess returns, we examine the pricing performance of these risk-factors, part of the required risk-factor protocol described in Pukthuanthong, Roll, and Subrahmanyam (2019).

4.1 Pricing test

We begin by outlining our Bayesian methodology for determining if a given traded equity asset is priced by a set of risk-factors. Suppose that the premium (excess return) of a test asset is denoted by prm. Under the SDF m_{M}, composed of the risk factors $\boldsymbol{f}_{\mathrm{M}}^{*}$ constructed by method M, this premium is priced if $E[m_{\mathrm{M}}\mathrm{prm}] = 0$, or equivalently, if

$$\mathrm{prm} = \boldsymbol{\beta}_0{}'\boldsymbol{f}_{\mathrm{M}}^{*} + \varepsilon_0, \quad \mathrm{M} \in \{\mathrm{S,D,R}\}, \tag{5}$$

and mispriced if

$$\mathrm{prm} = \alpha + \boldsymbol{\beta}_1{}'\boldsymbol{f}_{\mathrm{M}}^{*} + \varepsilon_1, \quad \mathrm{M} \in \{\mathrm{S,D,R}\}. \tag{6}$$

In the traditional, non-Bayesian approach, one estimates the latter model and tests the null of correct pricing, i.e., $\alpha = 0$. In this approach, one can reject the null, but not accept the null that the asset is priced. For example, in a t-test, if the value of the t-statistic is less than the threshold, one cannot conclude that the premium is priced. Instead, we take a Bayesian approach and treat the two possibilities symmetrically to obtain evidence in favor of the null of correct pricing. Let $\mathscr{M}_0$ denote the regression model of premium under correct pricing given in (5), and $\mathscr{M}_1$ denote the regression model of premium under mispricing given in (6). We then calculate $\Pr(\mathscr{M}_0|\mathrm{Data})$ and use this probability to infer if the premium of the test asset is priced.

We assume that the noise terms ε_0 and ε_1 are student-t $t(0,\delta^2,\nu)$ distributed with unknown ν degrees of freedom. Let $\pi(\beta_0,\delta^2|\mathscr{M}_0)$ denote the prior of the parameters in $\mathscr{M}_0$ and let $\pi(\alpha,\beta_1,\delta^2|\mathscr{M}_1)$ denote the prior of the parameters in $\mathscr{M}_1$. We take these priors to be a product of multivariate normal and inverse-gamma distributions and fix the hyperparameters of these prior distributions by a training sample approach. This training sample consists of the first 15% of the sample.[7] Also let

$$p(\text{prm}_{1:T}|\mathscr{M}_j,\boldsymbol{f}^*_{\text{M},1:T},\boldsymbol{\theta}_j)$$

denote the joint density of the sample premium (a product of student-t densities) under model $\mathscr{M}_j$ where $\boldsymbol{\theta}_0=(\beta_0,\delta^2)$ and $\boldsymbol{\theta}_1=(\alpha,\beta_1,\delta^2)$. Now denote the respective marginal likelihoods, the integral of the latter density over the prior of the parameters, by

$$\text{marglik}_j(\text{prm}_{1:T}|\mathscr{M}_j,\boldsymbol{f}^*_{\text{M},1:T})=\int p(\text{prm}_{1:T}|\mathscr{M}_j,\boldsymbol{f}^*_{\text{M},1:T},\boldsymbol{\theta}_j)d\pi(\boldsymbol{\theta}_j|\mathscr{M}_j)$$

which we each compute by the method of Chib (1995). Then, by Bayes theorem, under the assumption that the prior probabilities of the models are equal, the posterior probability of $\mathscr{M}_0$ is given by

$$\Pr(\mathscr{M}_0|\text{Data})=\frac{1}{1+e^{-d}}$$

where

$$d=\log(\text{marglik}_0(\text{prm}_{1:T}|\mathscr{M}_0,\boldsymbol{f}^*_{\text{M},1:T})-\log(\text{marglik}_1(\text{prm}_{1:T}|\mathscr{M}_1,\boldsymbol{f}^*_{\text{M},1:T}))$$

is the difference in log-marginal likelihoods.

We classify the premium as priced if $\Pr(\mathscr{M}_0|\text{Data})$ is at least 0.667 (which means the posterior odds in favor of correct pricing are 2:1). An easy calculation shows that then d should be at least

[7] Actually, if the sample size of a certain stock is small, less than 100 (as in the case of some small stocks), we use the first 40% of the data as a training sample. If the available sample size is between 100 and 150, we use the first 30% of the sample as a training sample. When the sample size exceeds 150, the most common case, we deploy the first 15% of the data to form the training sample prior.

0.693. In other words, if the log-marginal likelihood of $\mathscr{M}_0$ exceeds that of $\mathscr{M}_1$ by more than 0.693, the odds in favor of pricing is at least 2:1. We also report results on pricing for the case where $\Pr(\mathscr{M}_0|\text{Data})$ is at least 0.75 (meaning the posterior odds in favor of pricing are 3:1). In this case, d has to be at least 1.099.[8] Finally, we report results on whether an asset is priced under the rule that d is at least as big as 1.386, equivalently, the posterior odds are at least 4:1.

Remark 5: In the case of the slope risk-factors $\boldsymbol{f}_{\text{S}}^*$ *and* stocks as the test assets, another related pricing test is possible, following Fama and French (2020). To describe this alternative, now let prm denote the excess return of the test stock and let $\boldsymbol{c}_{-1}^*$ denote the 19×1 vector of the lagged standardized stock characteristics that correspond to the risk-factors in $\boldsymbol{f}_{\text{S}}^*$. Then, since the $\boldsymbol{f}_{\text{S}}^*$ are the estimated slopes from cross-sectional regressions, one can say that the test stock premium is priced if

$$\text{prm} = \boldsymbol{c}_{-1}^{*\prime}\boldsymbol{f}_{\text{S}}^* + \varepsilon_0 \tag{7}$$

and mispriced if

$$\text{prm} = \alpha + \boldsymbol{c}_{-1}^{*\prime}\boldsymbol{f}_{\text{S}}^* + \varepsilon_1 \tag{8}$$

and compare these two models by marginal likelihoods. The main advantage of setting up the comparison in this way is that one does not have to estimate the regression parameters - these are already estimated as the slope factors. Thus, in order to calculate the marginal likelihoods, there is just one parameter to marginalize out in the former model and two parameters in the latter model. A downside of this approach, however, is that one would have to collect time-series data on the characteristics for each of our 6024 testing stocks. Fama and French (2020) did not face this problem because they applied this approach to the same stocks that produced their five slope-factors. In our pricing test, we are considering stocks that are not necessarily the same as those we used to construct the slope-factors. We just have returns data on our testing stocks, but not the data on characteristics. Another downside of this approach is that it cannot be applied to other testing

[8] In Jeffreys' (1961) scale, the evidence in favor of the null is substantial if d exceeds 1.15, equivalently, the posterior probability of the null model is at least 0.759, implying a posterior odds ratio of 3.15 to 1. We use 3:1 to maintain integer odds.

assets, portfolios and ETFs, since, in each of those cases, it is not possible to define characteristics that correspond to our slope risk-factors. These difficulties are by-passed in the pricing approach we use.

4.2 Pricing results

We now apply our pricing methodology to each of our test assets in turn, and calculate the evidence in favor of pricing for posterior odds of at least 2:1, at least 3:1 and at least 4:1. To begin, we consider testing assets that are portfolios constructed by us from our sample data. In particular, for each of our 61 characteristics (excluding size), we construct 5×5 sorts on size and characteristic, leading to 1525 ($= 25 \times 61$) value-weighted portfolios. We then apply our pricing test to each of these LHS testing assets with, in turn, each set of risk-factors on the RHS. The results given in Table 12 show that the slope risk-factors price more of these portfolios than the differential and rank risk-factors. In particular, under the at least 2:1 posterior odds threshold in favor of pricing criteria, the slope risk-factors price 1405 of these portfolios, while the differential and rank risk-factors price 1151 and 470 of these portfolios, respectively. Under the even more demanding, at least 4:1 posterior odds threshold in favor of pricing criteria, the slope risk-factors maintain their decisive edge in pricing these portfolios, managing to price 1176 of the 1525 portfolios, while the differential and rank risk-factors manage to only price 814 and 267, respectively.

In the second part of our pricing evaluation, we also consider as test assets a large collection of equity ETFs. These assets are diversified portfolios that are traded, liquid, transparent, and cheap to trade. As a result, the premium of these assets is likely to be determined by exposure to the common non-diversifiable sources of risk manifested in the risk-factors. With this in mind, we collect data on 1480 ETFs from CRSP (sharecode 73) spanning the period February 1993 (the earliest month for which data on ETFs is available) to December 2020. Within this time-span, each of 1480 ETFs has at least least 60 months of data. Our pricing methodology applied to these test assets shows that under the at least 2:1 posterior odds in favor of pricing criteria, the slope

risk-factors are able to price 1153 ETFs, beating the other sets of risk-factors. Under the more demanding at least 4:1 posterior odds in favor of pricing criteria, the slope risk-factors outperform the other sets of risk-factors as well. Full details are given in Table 13.

Finally, as a more rigorous examination of pricing capabilities, we consider a large collection of excess returns on common stocks as test assets. While, in general, stocks can be difficult to price because of the excess noise relative to portfolios, stocks as test assets can be viewed as more objective test assets than portfolios. This is because portfolios by diversifying away risk may mask relevant risk-factors embedded in portfolio premiums (Roll, 1977). Second, portfolios constructed by sorting on characteristics can introduce a spurious factor structure across testing portfolios; see for example Lewellen, Nagel, and Shanken (2010) and Lo and MacKinlay (1990). Third, forming portfolios might mask cross-sectional relation between average returns and beta exposure (see Roll (1977) and Ang, Liu, and Schwarz (2020) for example). Lastly, the statistical significance and economic significance of risk premium might be distorted depending on the choice of testing portfolios. It is known Fama and French (1993)'s HML and SMB factors demand risk-premiums when testing portfolios are sorted by book-to-market and size, respectively, but they do not demand risk premiums when testing portfolios are sorted by momentum. Brennan, Chordia, and Subrahmanyam (1998) show different results for different sets of portfolios depending on characteristics used to form such portfolios.

Our sample consists of monthly excess returns of 6024 common stocks from CRSP (sharecode 10 and 11). The sample period is from January 1989 to December 2020. Once again, we ensure that we have least 60 months of observations within this time frame on any given stock. Financial firms and firms with negative book equity are excluded. Stocks with prices per share lower than \$5 are also excluded. The results on pricing of these data are summarized in Table 14. As can be seen from the table, the pricing performance of the slope risk-factors strictly dominates that of the other two sets of risk-factors. Specifically, under the 2:1 criteria, the slope risk-factors are able to price 5062 of these stocks, and 3480 of these stocks under the more stringent 4:1 criteria. On the other hand, the differential and rank factors manage to price 4034 and 4312, respectively, under

the 2:1 criteria and 1487 and 1823, respectively under the 4:1 rule. This represents strong evidence in support of the pure-play slope factors.

Remark 1 (continued). We would like to emphasize again the importance of slope factors constructed from a broad set of initial characteristics. Limited-pure-play slope factors have weaker pricing performance. There are many ways to construct the limited-pure-play slope factors. For instance, one can make the limited-pure-play factors only using the nineteen characteristics that are represented by $\boldsymbol{f}_{\mathrm{S}}^{*}$. That is, one can construct slope factors by controlling only for the nineteen characteristics associated with risk (see Table 8) rather than a full swath of 62 characteristics in the cross-sectional regressions. One can view these as the most competitive limited-pure-play factors relative to $\boldsymbol{f}_{\mathrm{S}}^{*}$. Nonetheless, the pricing performance of these nineteen (plus the market factor), which is given in Table 15, is clearly worse than that of the $\boldsymbol{f}_{\mathrm{S}}^{*}$ risk-factors derived from our broad set of initial characteristics, uniformly across all three sets of test assets.

Remark 6: It is also important to emphasize that one can only get an accurate assessment of pricing performance if one has first determined which factors in $\boldsymbol{f}_{\mathrm{M}}$, $\mathrm{M} \in \{\mathrm{S,D,R}\}$, are actually risk-factors. After all, by definition, the asset premium is compensation for the exposure to the risk-factors. To demonstrate the importance of this point, suppose we were to randomly select twenty slope factors from Table 2, making sure that the market factor is always in this set, and then proceed to price the three sets of test assets with these twenty factors. We repeat this experiment (say) 100 times. In Table 16 we report the average number of assets priced across these 100 trials. The table shows that the pricing is uniformly worse than the pricing by the risk-factors in $\boldsymbol{f}_{\mathrm{S}}^{*}$.

5 Conclusion

In this paper we have extended the work of Fama and French (2020) on slope factors in several directions. One extension has been to include rank factors as competitors to slope factors and differential factors. A second extension has been to consider slope factors representing a larger

collection of characteristics. This extension is important because slope factors constructed from a limited set of characteristics, what we refer to as *limited-pure-play* slope factors, perform worse in pricing than the pure-play factors we have constructed from our broad set of 62 characteristics. That this would be the case was not obvious from Fama and French (2020) where only limited-pure-play slope factors were constructed and the likely benefits of considering more characteristics were not mentioned. A third extension has been to consider the pricing performance of only those factors that are common sources of factor risks. Again, this extension is important for properly quantifying the relative strengths of the factors constructed by the different methods. By doing this we improve, and refine, the comparison on pricing performance reported in Fama and French (2020) between the limited-pure-play factors and the differential factors.

Finally, we develop an approach for determining if a given test asset is priced by the risk-factors, departing from conventional approaches that falsify the null of pricing. In our approach we classify a test asset as priced if the posterior odds in favor of pricing exceed the strong 2:1, or, stronger, 3:1, or, strongest, 4:1 thresholds. These thresholds quantify the strength of evidence in support of pricing. On applying our methodology to a large collection of portfolios, ETFs and common stocks, we have shown that the slope risk-factors outperform the other risk-factors. This is due to the pure-play property of the slope factors, the feature that differentiates the slope factors from the differential and rank factors. We view our study as both extending and refining the findings of Fama and French (2020) about the relative superiority of slope factors. We believe that their conclusion, now further supported by our extensive analysis, has significant implications for empirical asset pricing.

Data, software and code to reproduce the results in this paper are available on request.

Figures and Tables

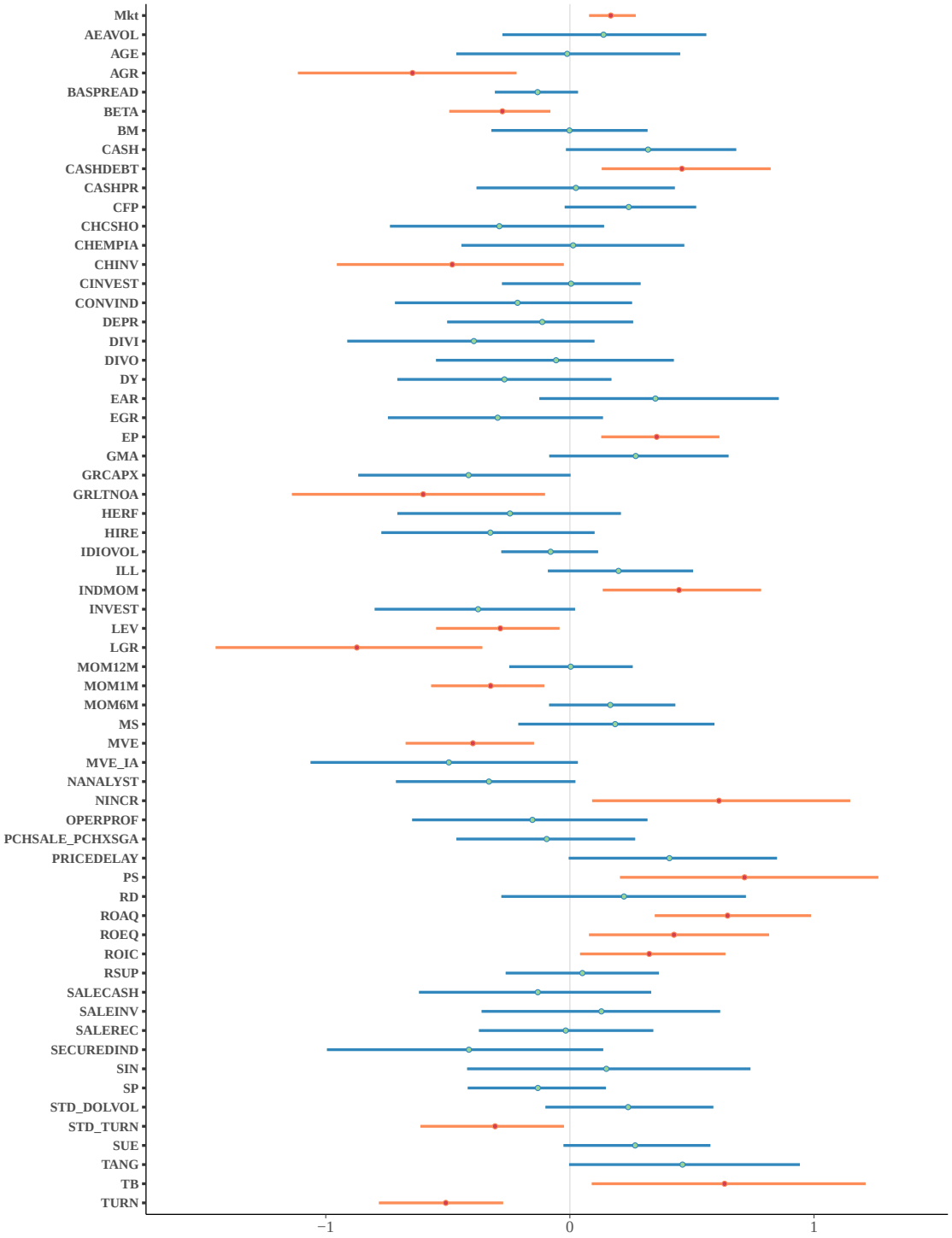

Figure 1 Slope factors: Posterior 99% credibility intervals of the SDF coefficients
This figure presents posterior 99% credibility intervals of the SDF coefficients of the *Mkt* and 62 slope factors. Intervals colored in red exclude zero and are associated with risk-factors. See Table 1 for acronyms.

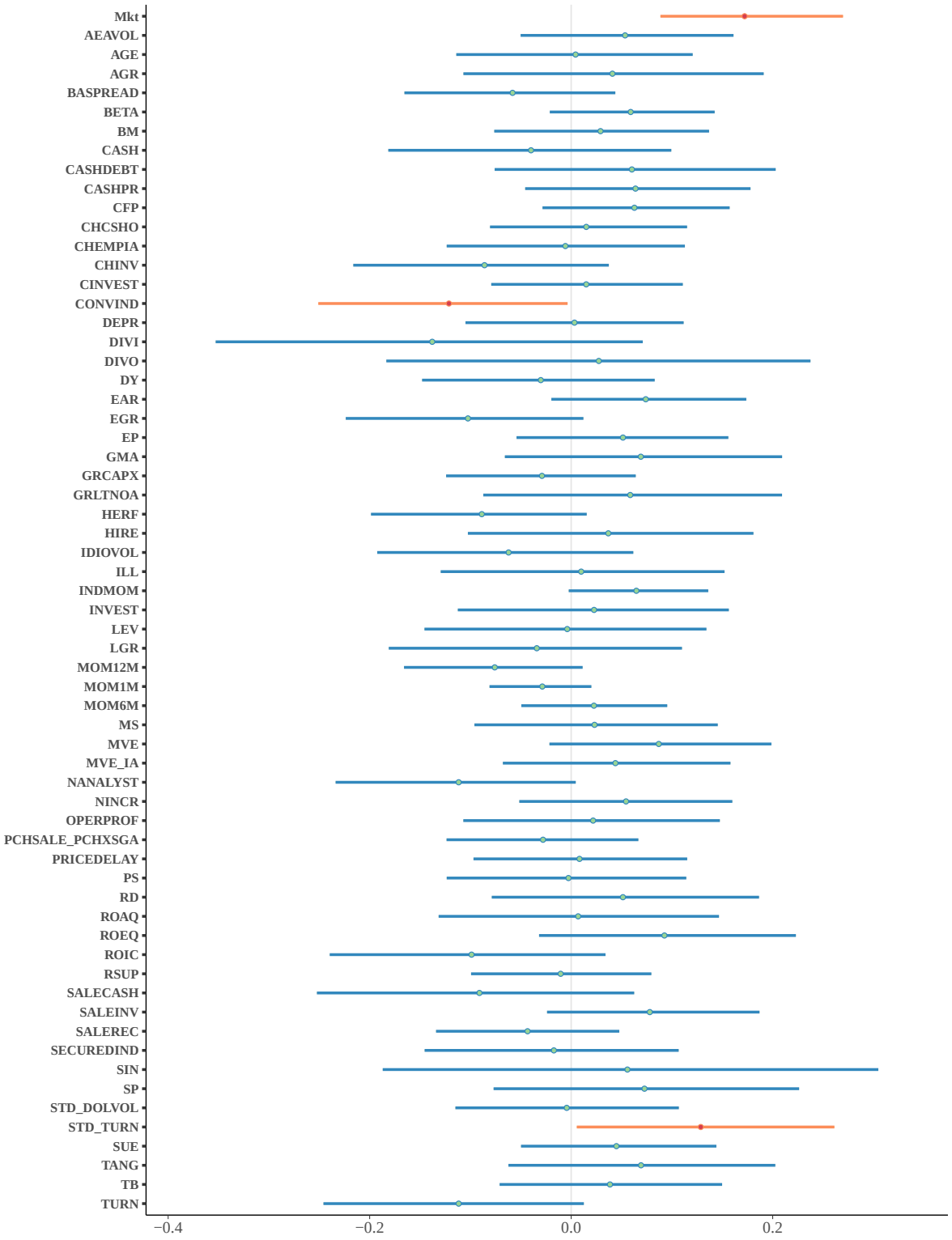

Figure 2 Differential factors: Posterior 99% credibility intervals of the SDF coefficients
This figure presents posterior 99% credibility intervals of the SDF coefficients of the *Mkt* and 62 differential factors. Intervals colored in red exclude zero and are associated with risk-factors. See Table 1 for acronyms.

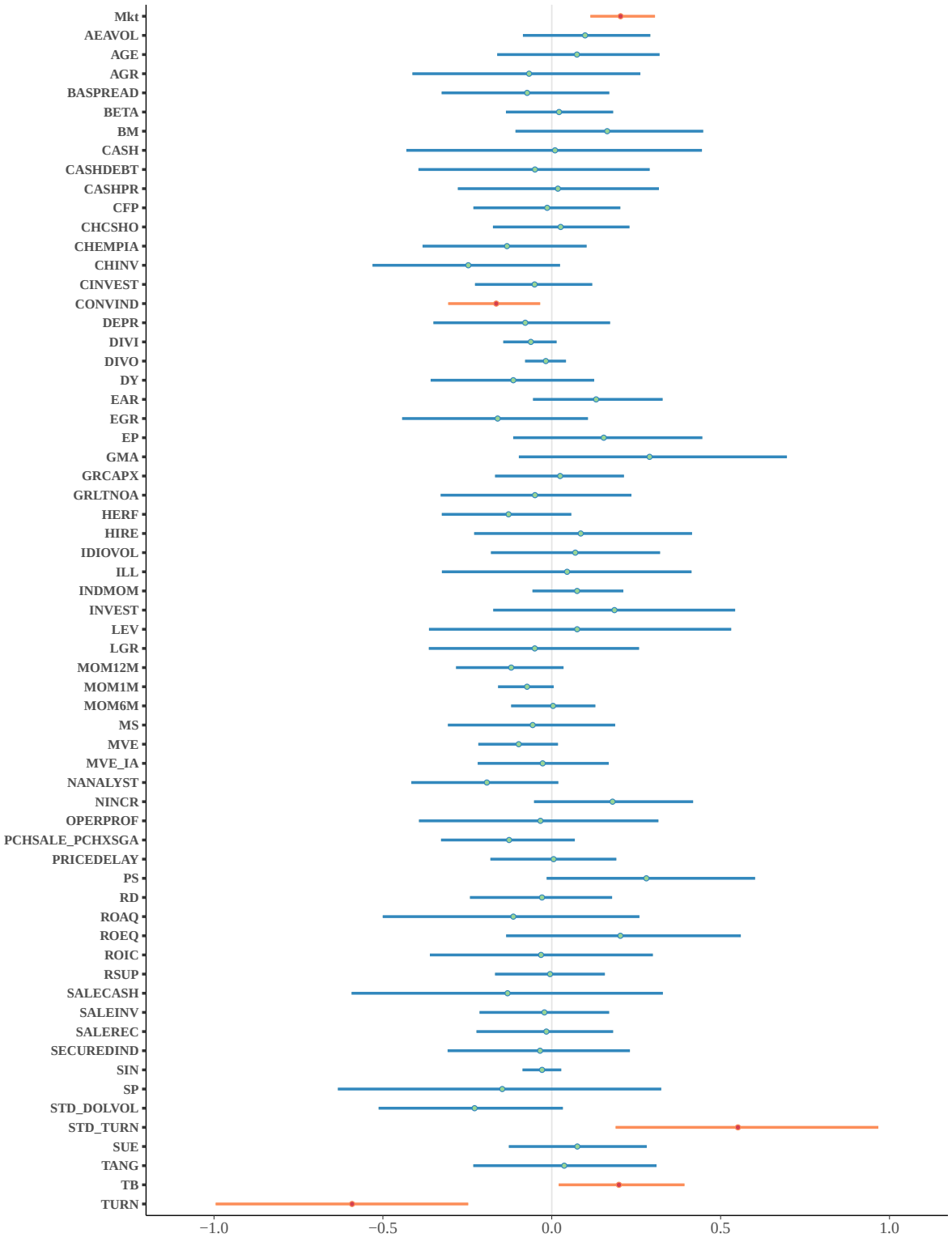

Figure 3 Rank factors: Posterior 99% credibility intervals of the SDF coefficients

This figure presents posterior 99% credibility intervals of the SDF coefficients of the *Mkt* and 62 rank factors. Intervals colored in red exclude zero and are associated with risk-factors. See Table 1 for acronyms.

Table 1 The descriptive statistics of the 62 characteristics

This table presents acronym, full names, definition, and descriptive statistics of characteristics generated by the code from Green et al. (2017). The min, max, mean, median, and standard deviation are for the characteristics across firms and months. The data are from January 1989 to December 2020.

Acronym	Firm characteristics	Definition	Min	Max	Mean	Median	Std
aeavol	Abnormal earnings announcement volume	Average daily trading volume (vol) for 3 days around earnings announcement minus average daily volume for 1-month ending 2 weeks before earnings announcement divided by 1-month average daily volume. Earnings announcement day from Compustat quarterly (rdq)	-1.00	21.06	1.08	0.52	2.15
age	# years since first Compustat coverage	# years since first Compustat coverage	3.00	58.00	20.36	18.00	13.29
agr	Asset growth	Annual percent change in total assets (at)	-0.68	6.06	0.12	0.05	0.36
baspread	Bid-ask spread	Monthly average of daily bid-ask spread divided by average of daily spread	0.00	0.90	0.04	0.03	0.04
beta	Beta	Estimated market beta from weekly returns and equal weighted market returns for 3 years ending month t-1 with at least 52 weeks of returns	-0.74	3.94	1.12	1.05	0.62
bm	Book-to-market	Book value of equity (ceq) divided by end of fiscal year-end market capitalization	-2.35	7.64	0.64	0.50	0.60
cash	Cash holdings	Cash and cash equivalents divided by average total assets	-0.08	0.97	0.14	0.08	0.17
cashdebt	Cash flow to debt	Earnings before depreciation and extraordinary items (ib+dp) divided by avg. total liabilities (lt)	-32.12	2.18	0.15	0.17	0.69
cashpr	Cash productivity	Fiscal year-end market capitalization plus long-term debt (dltt) minus total assets (at) divided by cash and equivalents (che)	-520.62	600.28	2.85	0.96	54.15
cfp	Cash-flow-to-price ratio	Operating cash flows divided by fiscal-year-end market capitalization	-2.80	2.62	0.08	0.08	0.21
chcsho	Changes in shares outstanding	Annual percent change in shares outstanding (csho)	-0.89	2.58	0.08	0.01	0.27
chempia	Industry-adjusted change in number of employees	Industry-adjusted change in number of employees	-24.16	3.50	-0.17	-0.08	0.88
chinv	Change in inventory	Change in inventory (inv) scaled by average total assets (at)	-0.29	0.36	0.01	0.00	0.05
cinvest	Corporate investment	Change over one quarter in net PP&E (ppentq) divided by sales (saleq) - average of this variable for prior 3 quarters; if saleq= 0, then scale by 0.0	-160.50	473.77	0.00	0.00	1.35
convind	Convertible debt obligations	An indicator equal to 1 if company has convertible debt obligations	0.00	1.00	0.13	0.00	0.34

Table 1 The descriptive statistics of the 62 characteristics

Acronym	Firm characteristics	Definition	Min	Max	Mean	Median	Std
depr	Depreciation/PP&E	Depreciation divided by PP&E	-0.12	6.70	0.28	0.18	0.37
divi	Dividend initiation	An indicator variable equal to 1 if company pays dividends but did not in prior year	0.00	1.00	0.03	0.00	0.17
divo	Dividend omission	An indicator variable equal to 1 if company does not pay dividend but did in prior year	0.00	1.00	0.02	0.00	0.15
dy	Dividend to price	Total dividends (dvt) divided by market capitalization at fiscal year-end	-0.20	0.35	0.01	0.00	0.02
ear	Earnings announcement return	Sum of daily returns in three days around earnings announcement. Earnings announcement from Compustat quarterly file (rdq)	-0.46	0.51	0.00	0.00	0.09
egr	Growth in common shareholder equity	Annual percent change in book value of equity (ceq)	-3.84	8.29	0.11	0.07	0.63
ep	Earnings to price	Annual income before extraordinary items (ib) divided by end of fiscal year market cap	-7.52	0.44	-0.03	0.04	0.33
gma	Gross profitability	Revenues (revt) minus cost of goods sold (cogs) divided by lagged total assets (at)	-0.90	1.78	0.45	0.40	0.28
grcapx	Growth in capital expenditures	Percent change in capital expenditures from yeart-2 to year t	-9.33	61.95	0.75	0.13	2.81
grltnoa	Growth in long term net operating assets	Growth in long-term net operating assets	-0.62	1.14	0.08	0.05	0.14
herf	Industry sales concentration	2-digit SIC - fiscal-year sales concentration (sum of squared percent of sales in industry for each company).	0.01	1.00	0.08	0.05	0.08
hire	Employee growth rate	Percent change in number of employees (emp)	-0.71	3.97	0.07	0.02	0.29
idiovol	Idiosyncratic return volatility	Standard deviation of residuals of weekly returns on weekly equal weighted market returns for 3 years prior to month end	0.01	0.28	0.06	0.06	0.03
ill	Illiquidity	Average of daily (absolute return / dollar volume).	0.00	0.00	0.00	0.00	0.00
indmom	Industry momentum	Equal weighted average industry 12-month returns	-0.76	3.64	0.14	0.11	0.30
invest	Capital expenditures and inventory	Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at)	-0.51	1.38	0.06	0.04	0.15
lev	Leverage or total liabilities (lt) divided by fiscal year-end market capitalization	Total liabilities (lt) divided by fiscal year-end market capitalization	0.00	75.46	1.04	0.48	2.14
lgr	Annual percent change in total liabilities (lt)	Annual percent change in total liabilities (lt)	-0.76	9.61	0.18	0.04	0.64

Table 1 The descriptive statistics of the 62 characteristics

Acronym	Firm characteristics	Definition	Min	Max	Mean	Median	Std
mom12m	12-month momentum	11-month cumulative returns ending one month before month end	-0.96	11.95	0.13	0.05	0.59
mom1m	1-month momentum	1-month cumulative return	-0.72	2.17	0.01	0.00	0.15
mom6m	6-month momentum	5-month cumulative returns ending one month before month end	-0.92	7.84	0.06	0.02	0.37
ms	Financial statement score	Sum of 8 indicator variables for fundamental performance	0.00	8.00	4.20	4.00	1.70
mve	Size	Natural log of market capitalization at end of month t-1	4.46	19.02	12.62	12.55	2.29
mve_ia	Industry-adjusted size	2-digit SIC industry-adjusted fiscal year-end market capitalization	-17506.35	146752.21	726.88	-711.47	10586.59
nanalyst	Number of analysts covering stock	Number of analyst forecasts from most recently available I/B/E/S summary files in month prior to month of portfolio formation. nanalyst set to zero if not covered in I/B/E/S summary file	0.00	56.00	5.98	3.00	7.48
nincr	Number of earnings increases	Number of consecutive quarters (up to eight quarters) with an increase in earnings (ibq) over same quarter in the prior year	0.00	8.00	0.94	1.00	1.24
operprof	Operating profit	Revenue minus cost of goods sold - SG&A expense - interest expense divided by lagged common shareholders' equity	-8.83	13.12	0.95	0.75	1.23
pchsale_pchxsga	% change in sales - % change in SG&A	Annual percent change in sales (sale) minus annual percent change in SG&A (xsga)	-1.25	4.10	0.02	0.00	0.25
pricedelay	Price delay	The proportion of variation in weekly returns for 36 months ending in montht explained by 4 lags of weekly market returns incremental to contemporaneous market return	-15.85	15.60	0.11	0.05	0.96
ps	Financial statement score	Sum of 9 indicator variables to form fundamental health score	0.00	8.00	4.81	5.00	1.65
rd	R&D expense	An indicator variable equal to 1 if R&D expense as a percentage of total assets has an increase greater than 5%.	0.00	1.00	0.18	0.00	0.38
roaq	Return on assets	Income before extraordinary items (ibq) divided by one quarter lagged total assets (atq)	-0.59	0.16	0.00	0.01	0.05
roeq	Quarterly return on equity	Earnings before extraordinary items divided by lagged common shareholders' equity	-2.28	1.77	0.01	0.02	0.14
roic	Return on invested capital	Annual earnings before interest and taxes (ebit) minus nonoperating income (nopi) divided by non-cash enterprise value (ceq+lt-che)	-23.55	1.01	0.03	0.08	0.43
rsup	Revenue surprise	Sales from quarter t minus sales from quarter t-4 (saleq) divided by fiscal-quarter-end market capitalization (cshoq * prccq)	-4.58	1.10	0.00	0.01	0.19

Table 1 The descriptive statistics of the 62 characteristics

Acronym	Firm characteristics	Definition	Min	Max	Mean	Median	Std
salecash	Sales to cash	Annual sales divided by cash and cash equivalents	-118.32	2503.48	73.81	13.84	209.68
saleinv	Sales to inventory	Annual sales divided by total inventory	0.00	1031.22	23.98	7.99	60.70
salerec	Sales to receivables	Annual sales divided by accounts receivable	0.00	210.01	14.79	6.66	27.99
securedind	Secured debt indicator	An indicator equal to 1 if company has secured debt obligations	0.00	1.00	0.56	1.00	0.50
sin	Sin stocks	An indicator variable equal to 1 if a company's primary industry classification is in smoke or tobacco, beer or alcohol, or gaming	0.00	1.00	0.01	0.00	0.11
sp	Sales to price	Annual revenue (sale) divided by fiscal year-end market capitalization	0.00	37.55	2.16	1.15	3.16
std_dolvol	Volatility of liquidity (dollar trading volume)	Monthly standard deviation of daily dollar trading volume	0.00	2.78	0.76	0.64	0.41
std_turn	Volatility of liquidity (share turnover)	Monthly standard deviation of daily share turnover	0.00	736.35	4.97	2.58	10.60
sue	Unexpected quarterly earnings	Unexpected quarterly earnings divided by fiscal-quarter-end market cap. Unexpected earnings is I/B/E/S actual earnings minus median forecasted earnings if available, else it is the seasonally differenced quarterly earnings before extraordinary items from Compustat quarterly file	-7.05	1.69	0.00	0.00	0.11
tang	Debt capacity/firm tangibility	Cash holdings + 0.715 _ receivables +0.547 _ inventory + 0.535 _ PPE/ totl assets	0.00	0.97	0.51	0.52	0.14
tb	Tax income to book income	Tax income, calculated from current tax expense divided by maximum federal tax rate, divided by income before extraordinary items	-27.34	15.36	-0.04	0.02	1.85
turn	Share turnover	Average monthly trading volume for most recent 3 months scaled by number of shares outstanding in current month	0.00	68.09	1.44	0.90	1.90

Table 2 The descriptive statistics of the slope factors

This table presents the descriptive statistics of the slope factors in units of monthly returns (%). The acronym is described in Table 1. Descriptive statistics include the mean, median, standard deviation, and the 0.5%, 99.5% quantiles across the time-series of each slope factor. The data are from January 1989 to December 2020.

	Mean	Median	Std	0.5% quantile	99.5% quantile
Mkt	0.739	1.190	4.357	-13.609	11.445
$AEAVOL_S$	-0.015	-0.057	0.515	-1.418	1.509
AGE_S	-0.019	-0.012	0.593	-1.605	1.697
AGR_S	-0.109	-0.089	1.326	-4.324	3.996
$BASPREAD_S$	-0.007	-0.120	1.658	-3.433	6.474
$BETA_S$	0.089	-0.011	1.614	-3.833	6.503
BM_S	0.043	0.025	0.854	-2.438	2.817
$CASH_S$	0.172	0.138	0.918	-2.171	3.026
$CASHDEBT_S$	0.038	0.092	0.932	-2.885	2.903
$CASHPR_S$	0.023	0.016	0.579	-1.581	1.964
CFP_S	0.095	0.145	0.967	-3.07	2.626
$CHCSHO_S$	-0.027	-0.058	0.671	-2.162	1.800
$CHEMPIA_S$	0.025	0.045	0.722	-2.004	1.914
$CHINV_S$	-0.058	-0.006	0.828	-2.283	2.401
$CINVEST_S$	-0.037	0.012	1.143	-2.875	2.899
$CONVIND_S$	0.003	-0.018	0.487	-1.489	1.473
$DEPR_S$	0.003	-0.032	0.708	-1.853	2.121
$DIVI_S$	-0.044	-0.051	0.549	-1.407	1.700
$DIVO_S$	-0.028	-0.040	0.473	-1.283	1.590
DY_S	-0.06	-0.098	0.595	-1.6	1.906
EAR_S	0.085	0.051	0.574	-1.462	1.520
EGR_S	-0.104	-0.119	0.827	-2.926	2.667
EP_S	0.054	0.088	1.113	-2.786	3.160
GMA_S	0.123	0.120	0.839	-2.21	2.303
$GRCAPX_S$	-0.078	-0.111	0.615	-1.926	1.875
$GRLTNOA_S$	-0.009	-0.046	0.823	-2.261	2.203
$HERF_S$	-0.028	-0.046	0.523	-1.466	1.875
$HIRE_S$	-0.022	-0.037	0.784	-2.417	1.970
$IDOLVOL_S$	0.027	-0.077	1.459	-3.271	5.183
ILL_S	0.151	0.106	0.988	-2.318	3.152
$INDMOM_S$	0.209	0.161	0.773	-1.691	2.877
$INVEST_S$	0.028	0.046	1.072	-2.468	3.140
LEV_S	-0.151	-0.174	1.243	-3.224	4.677
LGR_S	-0.019	-0.074	0.791	-2.118	1.989
$MOM12M_S$	0.059	0.035	1.095	-3.742	3.681
$MOM1M_S$	-0.538	-0.326	1.29	-5.122	2.379
$MOM6M_S$	0.059	0.098	1.17	-4.661	2.361
MS_S	0.054	0.041	0.708	-1.919	2.156
MVE_S	-0.466	-0.258	2.006	-8.293	4.993
MVE_IA_S	0.002	-0.014	0.652	-1.901	1.928
$NANALYST_S$	0.251	0.215	1.039	-2.004	4.154
$NINCR_S$	0.110	0.087	0.481	-1.002	1.582
$OPERPROF_S$	-0.006	0.047	0.649	-1.968	1.474
$PCHSALE\_PCHXSGA_S$	-0.022	-0.043	0.670	-2.379	1.920

Table 2 continued: The descriptive statistics of the slope factors

	Mean	Median	Std	0.5% quantile	99.5% quantile
$PRICEDELAY_S$	-0.005	-0.011	0.589	-2.062	1.718
PS_S	0.008	0.048	0.59	-1.604	1.518
RD_S	0.081	0.07	0.484	-0.974	1.932
$ROAQ_S$	0.048	0.099	1.137	-3.315	2.928
$ROEQ_S$	0.019	0.06	0.884	-2.795	2.573
$ROIC_S$	0.014	0.028	1.136	-2.788	3.616
$RSUP_S$	0.109	0.131	0.801	-2.414	2.381
$SALECASH_S$	-0.093	-0.071	0.506	-1.54	1.267
$SALEINV_S$	-0.008	-0.009	0.535	-1.556	1.581
$SALEREC_S$	-0.062	-0.07	0.674	-1.727	1.886
$SECUREDIND_S$	0.009	0.004	0.435	-1.443	1.129
SIN_S	0.033	0.009	0.434	-1.425	1.443
SP_S	0.13	0.055	1.095	-2.77	4.167
STD_DOLVOL_S	-0.188	-0.165	1.123	-4.067	1.993
STD_TURN_S	0.318	0.317	1.081	-2.607	3.176
SUE_S	0.152	0.113	0.859	-2.742	2.516
$TANG_S$	-0.016	-0.041	0.763	-2.149	1.900
TB_S	0.042	0.041	0.42	-1.123	1.081
$TURN_S$	-0.497	-0.514	1.298	-3.529	3.677

Table 3 The correlation matrix of the five pure-play slope factors and five limited-pure-play slope factors corresponding to the Fama-French five characteristics

Pure-play	MVE_S	BM_S	$OPERPROF_S$	AGR_S	$MOM12M_S$
MVE_S	1.000	0.207	0.022	0.123	0.047
BM_S	0.207	1.000	0.096	-0.006	-0.054
$OPERPROF_S$	0.022	0.096	1.000	0.175	0.037
AGR_S	0.123	-0.006	0.175	1.000	0.010
$MOM12M_S$	0.047	-0.054	0.037	0.010	1.000
Limited-pure-play					
MVE_S	1.000	0.419	0.210	0.074	-0.011
BM_S	0.419	1.000	0.319	-0.140	-0.188
$OPERPROF_S$	0.210	0.319	1.000	-0.070	-0.125
AGR_S	0.074	-0.140	-0.070	1.000	0.050
$MOM12M_S$	-0.011	-0.188	-0.125	0.050	1.000

The pure-play set of slope factors are constructed from cross-sectional regressions that involve the full set of 62 characteristics in Table 1. The limited-pure-play set of slope factors are constructed from cross-sectional regressions that contain the limited set of characteristics, size (*mve*), book-to-market ratio (*bm*), operating profitability ($operprof$), asset growth (*agr*), and momentum ($mom12m$), on the RHS. Sample data are monthly, spanning the period January 1989 - December 2020.

Table 4 The descriptive statistics of the differential factors

This table presents the descriptive statistics of the differential factors in units of monthly returns (%). The acronym is described in Table 1. Descriptive statistics include the mean, median, standard deviation, and the 0.5%, 99.5% quantiles across the time-series of each differential factor. The data are from January 1989 to December 2020.

	Mean	Median	Std	0.5% quantile	99.5% quantile
Mkt	0.739	1.190	4.357	-13.609	11.445
$AEAVOL_D$	0.238	0.101	2.345	-5.840	7.668
AGE_D	-0.148	0.107	3.871	-12.651	9.021
AGR_D	-0.615	-0.457	3.470	-11.612	8.832
$BASPREAD_D$	0.150	-0.517	8.643	-24.629	40.628
$BETA_D$	0.273	-0.193	8.810	-21.666	34.884
BM_D	0.431	0.461	3.757	-12.318	11.207
$CASH_D$	0.571	0.425	4.505	-13.932	14.612
$CASHDEBT_D$	0.114	0.644	4.628	-17.766	11.131
$CASHPR_D$	-0.065	-0.267	3.235	-10.116	7.943
CFP_D	0.571	0.739	4.581	-16.061	12.599
$CHCSHO_D$	-0.591	-0.533	3.472	-12.192	10.651
$CHEMPIA_D$	-0.281	-0.188	2.740	-7.446	9.660
$CHINV_D$	-0.604	-0.580	2.465	-8.505	4.878
$CINVEST_D$	0.169	0.169	2.485	-6.857	6.746
$CONVIND_D$	-1.070	-0.770	3.350	-10.644	7.546
$DEPR_D$	0.429	0.209	4.344	-11.753	16.618
$DIVI_D$	-1.038	-0.434	4.034	-13.506	6.755
$DIVO_D$	-0.972	-0.447	4.035	-16.203	7.564
DY_D	-0.926	-0.625	3.928	-15.208	8.605
EAR_D	0.565	0.723	2.586	-9.273	7.479
EGR_D	-0.633	-0.400	3.111	-11.643	7.257
EP_D	0.217	0.657	4.408	-17.656	12.058
GMA_D	0.360	0.636	3.567	-11.066	9.031
$GRCAPX_D$	-0.603	-0.583	2.863	-8.928	6.930
$GRLTOA_D$	-0.508	-0.610	2.712	-7.880	7.006
$HERF_D$	-0.360	-0.222	2.701	-10.061	6.365
$HIRE_D$	-0.442	-0.393	2.923	-9.684	7.662
$IDIOVOL_D$	0.132	-0.112	7.655	-24.472	30.232
ILL_D	0.154	0.080	3.375	-10.712	10.404
$INDMOM_D$	0.749	0.860	5.045	-18.629	16.076
$INVEST_D$	-0.580	-0.613	2.693	-7.226	7.733
LEV_D	0.192	0.061	5.078	-15.501	18.686
LGR_D	-0.328	-0.273	2.517	-8.846	8.262
$MOM12M_D$	0.617	1.093	7.166	-35.693	18.073
$MOM1M_D$	-1.301	-0.681	6.128	-23.231	15.148
$MOM6M_D$	0.540	0.935	6.885	-36.823	19.135
MS_D	0.341	0.574	3.852	-15.369	11.336
MVE_D	0.314	0.083	3.271	-7.347	14.798
MVE_IA_D	-0.058	-0.104	2.335	-10.094	5.063
$NANALYST_D$	-0.851	-0.672	4.137	-16.447	11.420
$NINCR_D$	-0.132	0.203	3.309	-11.510	8.213
$OPERPROF_D$	0.283	0.545	3.472	-13.410	7.609
$PCHSALE\_PCHXSGA_D$	-0.249	-0.232	2.719	-10.655	7.223

Table 4 continued: The descriptive statistics of the differential factors

	Mean	Median	Std	0.5% quantile	99.5% quantile
$PRICEDELAY_D$	0.091	0.171	2.406	-10.722	5.723
PS_D	0.194	0.372	3.781	-15.778	10.339
RD_D	-0.447	-0.292	3.268	-10.931	8.006
$ROAQ_D$	0.484	1.120	4.907	-21.121	11.610
ROQ_D	0.517	0.904	4.514	-20.472	10.936
$ROIC_D$	0.150	0.378	4.865	-17.551	12.295
$RSUP_D$	0.091	0.387	3.168	-11.868	6.823
$SALECASH_D$	-0.261	-0.218	4.585	-14.889	13.672
$SALEINV_D$	0.342	0.527	2.621	-10.097	7.464
$SALEREC_D$	0.124	0.056	3.806	-13.610	12.965
$SECUREDIND_D$	-0.494	-0.214	3.215	-13.468	7.128
SIN_D	-0.920	-0.274	4.350	-16.933	8.480
SP_D	0.344	0.252	5.090	-15.420	14.711
STD_DOLVOL_D	0.182	0.092	3.518	-11.434	10.447
STD_TURN_D	0.617	-0.026	6.226	-14.175	27.441
SUE_D	0.799	0.965	2.782	-10.633	8.242
$TANG_D$	0.404	0.080	3.855	-9.356	14.347
TB_D	0.093	0.230	2.401	-7.417	5.986
$TURN_D$	-0.023	-0.323	7.121	-20.305	29.132

Table 5 The descriptive statistics of the rank factors

This table presents the descriptive statistics of the rank factors in units of monthly returns (%). The acronym is described in Table 1. Descriptive statistics include the mean, median, standard deviation, and the 0.5%, 99.5% quantiles across the time-series of each rank factor. The data are from January 1989 to December 2020.

	Mean	Median	Std	0.5% quantile	99.5% quantile
Mkt	0.739	1.190	4.357	-13.609	11.445
AEAVOL_R	0.174	0.091	1.457	-4.181	4.776
AGE_R	-0.090	-0.014	2.601	-8.807	6.745
AGR_R	-0.486	-0.347	2.032	-6.626	4.706
BASPREAD_R	0.192	-0.023	5.578	-15.516	25.649
BETA_R	0.141	-0.121	5.542	-13.516	20.824
BM_R	0.287	0.193	2.429	-7.646	7.881
CASH_R	0.350	0.284	3.040	-8.850	10.988
CASHDEBT_R	-0.005	0.318	2.845	-10.402	6.894
CASHPR_R	-0.177	-0.106	2.398	-8.180	7.292
CFP_R	0.331	0.411	3.251	-12.927	11.415
CHCSHO_R	-0.304	-0.303	2.542	-8.239	7.507
CHEMPIA_R	-0.214	-0.117	1.428	-4.518	3.802
CHINV_R	-0.328	-0.278	1.460	-3.919	3.125
CINVEST_R	0.044	-0.029	1.379	-3.334	4.130
CONVIND_R	-0.221	-0.268	2.331	-5.234	8.561
DEPR_R	0.332	0.068	2.921	-7.889	10.751
DIVI_R	-0.402	-0.499	3.542	-10.035	11.691
DIVO_R	-0.271	-0.399	3.362	-10.979	9.174
DY_R	-0.249	-0.287	2.347	-7.231	6.153
EAR_R	0.296	0.393	1.396	-4.686	3.467
EGR_R	-0.341	-0.166	1.975	-6.853	4.733
EP_R	-0.001	0.110	3.221	-13.432	10.487
GMA_R	0.233	0.373	2.194	-6.449	5.763
GRCAPX_R	-0.332	-0.325	1.604	-4.600	3.663
GRLTNOA_R	-0.338	-0.350	1.540	-4.879	4.141
HERF_R	-0.096	-0.124	2.135	-7.735	6.434
HIRE_R	-0.320	-0.180	1.804	-5.721	5.082
IDOLVOL_R	0.246	0.146	5.205	-15.307	17.372
ILL_R	0.120	0.128	2.119	-6.739	6.061
INDMOM_R	0.544	0.552	3.063	-9.132	12.150
INVEST_R	-0.377	-0.335	1.625	-4.497	5.318
LEV_R	0.116	0.094	3.213	-9.758	11.691
LGR_R	-0.257	-0.273	1.322	-3.732	3.890
MOM12M_R	0.422	0.831	4.479	-20.239	10.709
MOM1M_R	-0.818	-0.433	3.664	-13.060	7.190
MOM6M_R	0.301	0.557	4.247	-21.828	11.068
MS_R	0.220	0.162	2.380	-7.736	7.287
MVE_R	-0.578	-0.149	4.527	-21.551	9.323
MVE_IA_R	-0.156	-0.086	1.869	-7.508	5.655
NANALYST_R	-0.094	-0.216	2.277	-6.835	9.349
NINCR_R	0.362	0.314	1.218	-2.804	4.047
OPERPROF_R	0.145	0.367	1.966	-7.021	4.189
$\text{PCHSALE\_PCHXSGA}_\text{R}$	-0.094	-0.003	1.499	-4.557	3.692

Table 5 continued: The descriptive statistics of the rank factors

	Mean	Median	Std	0.5% quantile	99.5% quantile
$PRICEDELAY_R$	0.114	0.081	1.520	-4.473	3.877
PS_R	0.182	0.362	2.186	-9.748	6.511
RD_R	0.357	0.173	2.168	-5.202	9.028
$ROAQ_R$	0.313	0.648	3.119	-13.296	8.384
$ROEQ_R$	0.264	0.497	2.992	-15.582	7.162
$ROIC_R$	0.064	0.257	2.903	-11.933	7.371
$RSUP_R$	0.114	0.389	1.988	-7.376	4.753
$SALECASH_R$	-0.208	-0.142	3.056	-11.567	8.689
$SALEINV_R$	0.143	0.131	1.495	-4.372	4.717
$SALEREC_R$	-0.022	-0.029	2.308	-7.678	6.838
$SECUREDIND_R$	-0.016	-0.001	1.190	-3.124	3.615
SIN_R	0.187	0.090	5.034	-15.277	16.084
SP_R	0.243	0.103	3.153	-9.266	10.283
STD_DOLVOL_R	0.144	0.275	2.354	-8.331	5.481
STD_TURN_R	0.443	0.105	4.130	-10.380	18.624
SUE_R	0.451	0.544	1.597	-5.828	3.866
$TANG_R$	0.232	0.097	2.257	-5.885	8.622
TB_R	0.101	0.148	1.935	-7.869	5.364
$TURN_R$	0.064	-0.211	4.676	-13.201	18.262

Table 6 The correlation matrix of the five differential factors and five rank factors corresponding to the Fama-French five characteristics

Differential factors	MVE_D	BM_D	$OPERPROF_D$	AGR_D	$MOM12M_D$
MVE_D	1.000	0.030	-0.349	-0.393	-0.385
BM_D	0.030	1.000	0.093	-0.225	-0.031
$OPERPROF_D$	-0.349	0.093	1.000	0.427	0.231
AGR_D	-0.393	-0.225	0.427	1.000	0.015
$MOM12M_D$	-0.385	-0.031	0.231	0.015	1.000
Rank factors	MVE_R	BM_R	$OPERPROF_R$	AGR_R	$MOM12M_R$
MVE_R	1.000	-0.044	0.267	0.439	0.410
BM_R	-0.044	1.000	-0.014	-0.447	-0.053
$OPERPROF_R$	0.267	-0.014	1.000	0.422	0.073
AGR_R	0.439	-0.447	0.422	1.000	0.112
$MOM12M_R$	0.410	-0.053	0.073	0.112	1.000

Correlation matrix of differential and rank factors representing size (*mve*), book-to-market ratio (*bm*), operating profitability ($operprof$), asset growth (*agr*), and momentum ($mom12m$). Sample data are monthly, spanning the period January 1989 - December 2020.

Table 7 Log marginal likelihoods for the multivariate-t model $f_M = \lambda_M + \varepsilon_M$, for different degrees of freedom ν

ν	Slope factors	Differential factors	Rank factors
4	-20428.83	-39028.23	-28785.78
4.5	-20424.63	-39023.90	-28783.02
5	-20423.08	-39020.45	-28780.65
5.5	-20422.38	-39019.48	**-28780.14**
6	**-20421.30**	**-39019.42**	-28781.08
6.5	-20422.56	-39019.78	-28782.80
7	-20424.08	-39021.12	-28784.77
Inf	-21587.91	-40089.09	-29932.48

Estimation based on training sample priors that use the first 30% of the sample data (see text for further details). For each model, MCMC sampling sampling of the posterior is used to produce 20,000 draws of the parameters, beyond a burn-in of 1000 cycles. This output is used to calculate the log-marginal likelihoods using the method of Chib (1995). The best model, equivalently, the best ν corresponds to the model with the largest log marginal likelihood, indicated in bold. Inf stands for infinity.

Table 8 Slope factors: selected marginal posterior distributions of the SDF coefficients, $\boldsymbol{b}_{\mathrm{S}}$

Factor	Mean	Sd	Median	0.5% quantile	99.5% quantile
Mkt	0.168	0.037	0.168	0.079	0.270
$\mathrm{AGR_S}$	-0.649	0.171	-0.645	-1.114	-0.218
$\mathrm{BETA_S}$	-0.278	0.082	-0.277	-0.494	-0.080
$\mathrm{CASHDEBT_S}$	0.460	0.134	0.459	0.130	0.823
$\mathrm{CHINV_S}$	-0.484	0.177	-0.482	-0.955	-0.025
$\mathrm{EP_S}$	0.358	0.094	0.356	0.129	0.613
$\mathrm{GRLTNOA_S}$	-0.604	0.203	-0.601	-1.139	-0.102
$\mathrm{INDMOM_S}$	0.449	0.126	0.447	0.135	0.783
$\mathrm{LEV_S}$	-0.287	0.098	-0.285	-0.548	-0.042
$\mathrm{LGR_S}$	-0.879	0.212	-0.873	-1.451	-0.358
$\mathrm{MOM1M_S}$	-0.326	0.089	-0.325	-0.568	-0.104
$\mathrm{MVE_S}$	-0.399	0.103	-0.397	-0.673	-0.147
$\mathrm{NINCR_S}$	0.613	0.207	0.611	0.091	1.148
$\mathrm{PS_S}$	0.717	0.205	0.715	0.205	1.264
$\mathrm{ROAQ_S}$	0.649	0.124	0.646	0.348	0.988
$\mathrm{ROEQ_S}$	0.428	0.141	0.426	0.078	0.816
$\mathrm{ROIC_S}$	0.327	0.115	0.325	0.042	0.637
$\mathrm{STD\_TURN_S}$	-0.309	0.114	-0.306	-0.612	-0.024
$\mathrm{TB_S}$	0.636	0.218	0.634	0.090	1.212
$\mathrm{TURN_S}$	-0.511	0.099	-0.508	-0.782	-0.273

This table shows twenty of the 63 marginal posterior distributions of the SDF coefficients, $\boldsymbol{b}_{\mathrm{S}}$, from the estimation of the model $\boldsymbol{f}_{\mathrm{S},t} = \boldsymbol{\lambda}_{\mathrm{S},t} + \varepsilon_{\mathrm{S},t}$, for t running from January 1989 to December 2020, with the error distributed as multivariate-t with $\nu = 6$ degrees of freedom, whose 99% credibility intervals exclude zero. Thus, these are the posterior distributions of the inferred risk-factors. The remaining 43 marginal posterior distributions of the SDF coefficients are not shown.

Table 9 Differential factors: selected marginal posterior distributions of SDF coefficients, $\boldsymbol{b}_{\mathrm{D}}$

Factor	Mean	Sd	Median	0.5% quantile	99.5% quantile
Mkt	0.173	0.035	0.172	0.089	0.270
$\mathrm{CONVIND}_{\mathrm{D}}$	-0.122	0.048	-0.121	-0.251	-0.004
$\mathrm{STD\_TURN}_{\mathrm{D}}$	0.129	0.049	0.129	0.006	0.261

This table shows three of the 63 marginal posterior distributions of the SDF coefficients, $\boldsymbol{b}_{\mathrm{D}}$ from the estimation of the model $\boldsymbol{f}_{\mathrm{D},t} = \boldsymbol{\lambda}_{\mathrm{D},t} + \varepsilon_{\mathrm{D},t}$, for t running from January 1989 to December 2020, with the error distributed as multivariate-t with $\nu = 6$ degrees of freedom, whose 99% credibility intervals exclude zero. Thus, these are the posterior distributions of the inferred risk-factors. The remaining 60 marginal posterior distributions of the SDF coefficients are not shown.

Table 10 Rank factors: selected marginal posterior distributions of the SDF coefficients, $\boldsymbol{b}_{\mathrm{R}}$

Factor	Mean	Sd	Median	0.5% quantile	99.5% quantile
Mkt	0.205	0.038	0.204	0.114	0.306
$\mathrm{CONVIND}_{\mathrm{R}}$	-0.166	0.053	-0.164	-0.307	-0.034
$\mathrm{STD\_TURN}_{\mathrm{R}}$	0.555	0.150	0.551	0.189	0.967
TB_{R}	0.200	0.072	0.199	0.020	0.393
$\mathrm{TURN}_{\mathrm{R}}$	-0.596	0.146	-0.591	-0.995	-0.247

This table shows five of the 63 marginal posterior distributions of the SDF coefficients, $\boldsymbol{b}_{\mathrm{R}}$, from the estimation of the model $\boldsymbol{f}_{\mathrm{R},t} = \boldsymbol{\lambda}_{\mathrm{R},t} + \varepsilon_{\mathrm{R},t}$, for t running from January 1989 to December 2020, with the error distributed as multivariate-t with $\nu = 5.5$ degrees of freedom, whose 99% credibility intervals exclude zero. Thus, these are the posterior distributions of the inferred risk-factors. The remaining 58 marginal posterior distributions of the SDF coefficients are not shown.

Table 11 The category classification of the underlying characteristics of the identified risk-factors from slope factors, differential factors, and rank factors, January 1989 - December 2020.

Category of characteristics	Slope factors	Differential factors	Rank factors	Characteristics
Momentum	*indmom**			Industry momentum
	*mom*1*m**			1-month momentum
	nincr			Number of earnings increases
Value-versus-Growth	*cashdebt**			Cash flow to debt
	chinv			Change in inventory
	*ep**			Earnings to price
Investment	*agr**			Asset growth
	grltnoa			Growth in long-term net operating assets
Profitability	*lev*			Leverage
	*ps**			Financial statements score
	*roaq**			Return on assets
	roeq			Return on equity
	roic			Return on invested capital
	tb		*tb*	Tax income to book income
Intangibles		*convind*	*convind*	Convertible debt indicator
	*lgr**			Growth in long-term debt
Trading frictions	*beta**			Beta
	*mve**			Size
	*turn**		*turn*	Share turnover
	std_turn	*std_turn*	*std_turn*	Volatility of liquidity (share turnover)
The market factor	*Mkt**	*Mkt*	*Mkt*	
Total number of factors	20	3	5	

This table presents the discovered slope, differential and rank factors, and the characteristics corresponding to each risk-factor (excluding, of course, the *Mkt* factor). The starred slope risk-factors are those discovered under a 99.9% credibility interval rule. The data are from January 1989 to December 2020. See Table 1 for acronyms.

Table 12 Pricing results: portfolios as test assets

Posterior probability of pricing at least equal to	0.667	0.750	0.800
Posterior odds of pricing vs mispricing at least equal to	2	3	4
Slope risk-factors $\boldsymbol{f}_{\mathrm{S}}^{*}$	1405	1310	1176
Differential risk-factors $\boldsymbol{f}_{\mathrm{D}}^{*}$	1151	990	814
Rank risk-factors $\boldsymbol{f}_{\mathrm{R}}^{*}$	470	368	267

Number of priced portfolios. LHS assets are 1525 portfolios from 5×5 sorts on size and 61 characteristics (see Table 1 for the full list of 61 characteristics) from our sample data. Right hand side variables are the discovered risk-factors given in Tables 8, 9 and 10. The number of priced portfolios is reported for different minimum posterior odds ratios of pricing vs mispricing: 2:1, 3:1 and 4:1 (see Section 4 for further details).

Table 13 Pricing results: ETFs as test assets

Posterior probability of pricing at least equal to	0.667	0.750	0.800
Posterior odds of pricing vs mispricing at least equal to	2	3	4
Slope risk-factors $\boldsymbol{f}_{\mathrm{S}}^{*}$	1153	966	756
Differential risk-factors $\boldsymbol{f}_{\mathrm{D}}^{*}$	734	470	316
Rank risk-factors $\boldsymbol{f}_{\mathrm{R}}^{*}$	1027	697	430

Number of priced ETFs. LHS assets are 1480 ETFs obtained from CRSP (sharecode 73). We select ETFs that have at least 60 months of observation within January 1989 - December 2020. Right hand side variables are the three sets of discovered risk-factors given in Tables 8, 9 and 10. The number of priced ETFs is reported for different minimum posterior odds ratios of pricing vs mispricing: 2:1, 3:1 and 4:1 (see Section 4 for further details).

Table 14 Pricing results: stocks as test assets

Posterior probability of pricing at least equal to	0.667	0.750	0.800
Posterior odds of pricing vs mispricing at least equal to	2	3	4
Slope risk-factors $\boldsymbol{f}_{\mathrm{S}}^{*}$	5062	4360	3480
Differential risk-factors $\boldsymbol{f}_{\mathrm{D}}^{*}$	4034	2553	1487
Rank risk-factors $\boldsymbol{f}_{\mathrm{R}}^{*}$	4312	2931	1823

Number of priced stocks. LHS assets are 6024 common stocks obtained from CRSP (sharecode 10 and 11). We select firms that have at least 60 months of observation within January 1989 - December 2020. Financial firms and firms with negative book equity are excluded. Stocks with prices per share lower than \$5 are also excluded. Right hand side variables are the three sets of discovered risk-factors given in Tables 8, 9 and 10. The number of priced stocks is reported for different minimum posterior odds ratios of pricing vs mispricing: 2:1, 3:1 and 4:1 (see Section 4 for further details).

Table 15 Pricing results: limited-pure-play slope factors as RHS variables

Posterior probability of pricing at least equal to	0.667	0.750	0.800
Posterior odds of pricing vs mispricing at least equal to	2	3	4
(1) 1525 portfolios	1021	865	661
(2) 1480 ETFs	1124	959	767
(3) 6024 common stocks	4940	4107	3272

Number of priced test assets. LHS assets are in the first column: (1) 1525 portfolios from 5×5 sorts on size and 61 characteristics (see Table 1 for the full list of 61 characteristics) from our sample data, (2) 1480 ETFs obtained from CRSP (sharecode 73), (3) 6024 common stocks obtained from CRSP (sharecode 10 and 11). RHS variables are the market factor plus 19 limited-pure-play slope factors constructed by only controlling for the 19 characteristics (as in Table 8) rather than 62 characteristics in the cross-sectional regressions. The number of priced assets is reported for different minimum posterior odds ratios of pricing vs mispricing: 2:1, 3:1 and 4:1 (see Section 4 for further details).

Table 16 Pricing results: the market factor plus 19 randomly selected slope factors as RHS variables (repeated for 100 trials, average number shown)

Posterior probability of pricing at least equal to Posterior odds of pricing vs mispricing at least equal to	0.667 2	0.750 3	0.800 4
(1) 1525 portfolios	1023.89	845.76	652.42
(2) 1480 ETFs	999.61	816.06	642.95
(3) 6024 common stocks	4860.36	4000.63	3086.89

This table shows the average number of priced test assets across 100 trials. LHS assets are in the first column: (1) 1525 portfolios from 5×5 sorts on size and 61 characteristics (see Table 1 for the full list of 61 characteristics) from our sample data, (2) 1480 ETFs obtained from CRSP (sharecode 73), (3) 6024 common stocks obtained from CRSP (sharecode 10 and 11). RHS variables are the market factor plus 19 randomly selected slope factors from Table 2. The number of priced assets is reported for different minimum posterior odds ratios of pricing vs mispricing: 2:1, 3:1 and 4:1 (see Section 4 for further details).

Appendices

A OLS estimates are factors

In this appendix we provide a simple explanation of the fact that OLS estimates in cross-sectional regressions on standardized lagged characteristics are pure-play long-short portfolios. For expositional and notational clarity, we provide this derivation (without loss of generality) for two characteristics.

Suppose that the tth cross-section consists of n_t firms that are independently sampled from the population of firms at time t. Let $\boldsymbol{r}_t = (r_{1t}, ..., r_{n_t,t})$ denote the the sample vector of excess returns and suppose that two firm characteristics, c_1 and c_2 are measured. Let the sample data on these characteristics at time t be denoted by the $n_t \times 1$ vectors, $\boldsymbol{c}_{j,t} = (c_{j,1}, ..., c_{j,n_t})$, $j = 1, 2$. Assume that $\boldsymbol{c}_{j,t}$ are each standardized by subtracting the respective sample means and dividing by the respective sample standard deviations. In vector-matrix notation, the tth cross-sectional regression is given by

$$\boldsymbol{r}_t = \boldsymbol{X}_t \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t$$

where $\boldsymbol{X}_t = (\boldsymbol{i}_{n_t}, \boldsymbol{c}_{1,t-1}, \boldsymbol{c}_{2,t-1})$ is a $n_t \times 3$ matrix of sample data on the intercept and the two characteristics and $\boldsymbol{i}_{n_t}$ is a vector of ones. The coefficient vector $\boldsymbol{\beta}_t$ is a 3×1 vector of cross-section specific coefficients and $\boldsymbol{\varepsilon}_t$ is a vector of iid homoskedastic cross-sectional errors. Now consider the OLS estimate of $\boldsymbol{\beta}_t$, namely $\hat{\boldsymbol{\beta}}_t = (\boldsymbol{X}_t' \boldsymbol{X}_t)^{-1} \boldsymbol{X}_t' \boldsymbol{r}_t$. This can be expressed as a linear combination of the excess returns as

$$\hat{\boldsymbol{\beta}}_t = \boldsymbol{W}_t' \boldsymbol{r}_t \tag{A.1}$$

or as

$$\begin{pmatrix} \hat{\beta}_{0,t} \\ \hat{\beta}_{1,t} \\ \hat{\beta}_{2,t} \end{pmatrix} = \begin{pmatrix} \boldsymbol{w}_0' \boldsymbol{r}_t \\ \boldsymbol{w}_1' \boldsymbol{r}_t \\ \boldsymbol{w}_2' \boldsymbol{r}_t \end{pmatrix}$$

where $\boldsymbol{w}_j'$ is the j^{th} row of $\boldsymbol{W}_t'$. Now from the trivial identity $\boldsymbol{W}_t'\boldsymbol{X}_t = \boldsymbol{I}_3$, where $\boldsymbol{I}_3$ is the 3×3 identity matrix, written out in full as

$$\begin{pmatrix} \boldsymbol{w}_0' \\ \boldsymbol{w}_1' \\ \boldsymbol{w}_2' \end{pmatrix} (\boldsymbol{i}_{n_t}, \boldsymbol{c}_{1,t-1}, \boldsymbol{c}_{2,t-1}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{A.2}$$

we can see, by row-column multiplication, that $\boldsymbol{w}_j$ are proper weights that satisfy the following restrictions,

$$\begin{aligned} &\boldsymbol{w}_0'\boldsymbol{i}_{n_t} = 1; \;\; \boldsymbol{w}_0'\boldsymbol{c}_{1,t-1} = 0; \;\; \boldsymbol{w}_0'\boldsymbol{c}_{2,t-1} = 0 \\ &\boldsymbol{w}_1'\boldsymbol{i}_{n_t} = 0; \;\; \boldsymbol{w}_1'\boldsymbol{c}_{1,t-1} = 1; \;\; \boldsymbol{w}_1'\boldsymbol{c}_{2,t-1} = 0 \\ &\boldsymbol{w}_2'\boldsymbol{i}_{n_t} = 0; \;\; \boldsymbol{w}_2'\boldsymbol{c}_{1,t-1} = 0; \;\; \boldsymbol{w}_2'\boldsymbol{c}_{2,t-1} = 1 \end{aligned} \tag{A.3}$$

Reading these restrictions row by row, we can now conclude that $\hat{\beta}_{0,t} = \boldsymbol{w}_0'\boldsymbol{r}_t$ is a pure-play long portfolio (its weights $\boldsymbol{w}_0$ sum to one and it gives zero weighted exposure to the other two characteristics); that $\hat{\beta}_{1,t} = \boldsymbol{w}_1'\boldsymbol{r}_t$ is a pure-play long-short portfolio (its weights $\boldsymbol{w}_1$ sum to zero, it gives unit weighted exposure to the first lagged characteristic and zero weighted exposure to the second lagged characteristic); and that $\hat{\beta}_{2,t} = \boldsymbol{w}_2'\boldsymbol{r}_t$ is a pure-play long-short portfolio (its weights $\boldsymbol{w}_2$ sum to zero, it gives zero weighted exposure to the first lagged characteristic and unit weighted exposure to the second lagged characteristic). Thus, $\hat{\beta}_{1,t}$ and $\hat{\beta}_{2,t}$ are characteristic specific long-short portfolios.

It should also be clear from this derivation that regularized estimates, such as the LASSO, or the Bayesian posterior mean with a proper prior, would not satisfy this property.

If we run the preceding cross-sectional regression separately in sequence for $t = 1, 2, ..., T$, then the sequence of OLS estimates $\hat{\beta}_{j,t}$, $t = 1, ..., T$, are a sequence of long-short portfolios that load purely on characteristic c_j, $j = 1, 2$.

B Scree plots

This appendix shows the scree plots for the slope, differential, and rank factors.

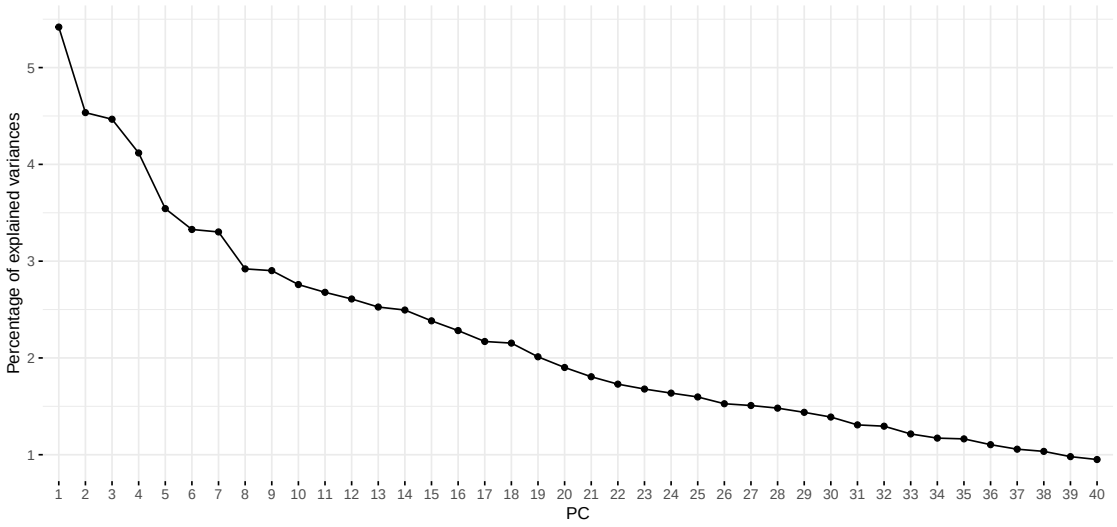

Figure B.1 Slope factors: Scree plot
Scree plot from the PCA of the slope factors data. This shows that 37 PC's explain 85% of the total variance.

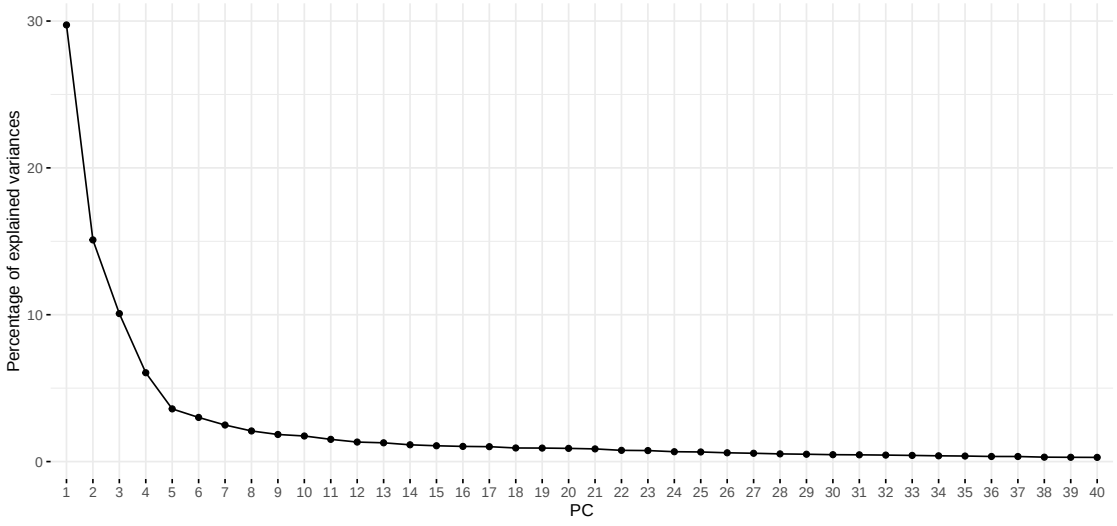

Figure B.2 Differential factors: Scree plot
Scree plot from the PCA of the differential factors data. This shows that 14 PC's explain 85% of the total variance.

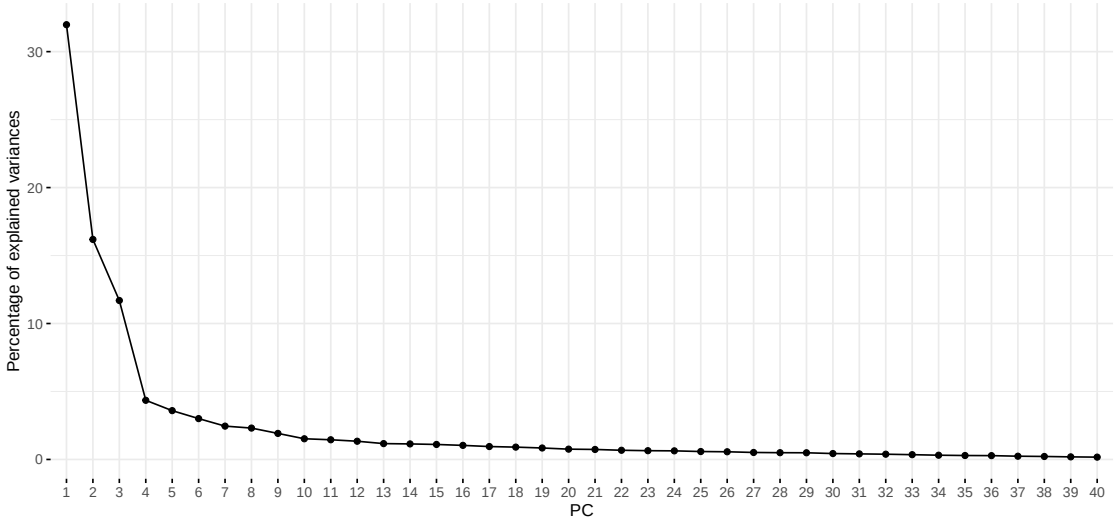

Figure B.3 Rank factors: Scree plot
Scree plot from the PCA of the rank factors data. This shows that 17 PC's explain 85% of the total variance.

Bibliography

Andrew Ang, Jun Liu, and Krista Schwarz. Using stocks or portfolios in tests of factor models. *Journal of Financial and Quantitative Analysis*, 55(3):709–750, 2020.

Clifford S Asness, Andrea Frazzini, and Lasse Heje Pedersen. Quality minus junk. *Review of Accounting Studies*, 24(1):34–112, 2019.

Kerry Back, Nishad Kapadia, and Barbara Ostdiek. Slopes as factors: Characteristic pure plays. *Available at SSRN 2295993*, 2013.

Kerry Back, Nishad Kapadia, and Barbara Ostdiek. Testing factor models on characteristic and covariance pure plays. *Available at SSRN 2621696*, 2015.

Michael J Brennan, Tarun Chordia, and Avanidhar Subrahmanyam. Alternative factor specifications, security characteristics, and the cross-section of expected stock returns. *Journal of Financial Economics*, 49(3):345–373, 1998.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Available at SSRN 3350138*, 2020.

Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

Siddhartha Chib. Markov chain monte carlo methods: Computation and inference. *Handbook of econometrics*, 5:3569–3649, 2001.

Siddhartha Chib and Xiaming Zeng. Which factors are risk factors in asset pricing? A model scan framework. *Journal of Business & Economic Statistics*, 38(4):771–783, 2020.

Siddhartha Chib, Xiaming Zeng, and Lingxiao Zhao. On comparing asset pricing models. *The Journal of Finance*, 75(1):551–577, 2020.

Guillaume Coqueret. Characteristics-driven returns in equilibrium. *Available at SSRN 3941195*, 2021.

Eugene F Fama. *Foundations Of Finan*. Basic books, 1976.

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

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

Eugene F Fama and Kenneth R French. Comparing cross-section and time-series factor models. *The Review of Financial Studies*, 33(5):1891–1926, 2020.

Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 2020.

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

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133, 2020.

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

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

Jonathan Lewellen, Stefan Nagel, and Jay Shanken. A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, 96(2):175–194, 2010.

Andrew W Lo and A Craig MacKinlay. When are contrarian profits due to stock market overreaction? *The Review of Financial Studies*, 3(2):175–205, 1990.

Kuntara Pukthuanthong, Richard Roll, and Avanidhar Subrahmanyam. A protocol for factor identification. *The Review of Financial Studies*, 32(4):1573–1607, 2019.

Richard Roll. A critique of the asset pricing theory's tests part I: On past and potential testability of the theory. *Journal of Financial Economics*, 4(2):129–176, 1977.