Slope Factors Outperform?

Evidence from a Large Comparative Study*

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

June 2023; November 2023

*We thank Gavin Feng, Kewei Hou, Yan Liu, Michael O'Doherty, Seth Pruitt, Shrihari Santosh, Lingxiao Zhao, Guofu Zhou and participants of the China International Conference in Finance (2022), the Annual Volatility Institute Conference at NYU Shanghai (2022), and the International Conference on Finance & Technology (2022) for helpful comments.

[†]Corresponding author: 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

ABSTRACT

We study slope factors (estimated OLS slopes from Fama-Macbeth regressions of firm returns on lagged standardized characteristics) in relation to sorted and ranked factors. We show that slope factors provide significant value if they are more pure play, i.e., purged of the effects of a broad set of characteristics. Starting from forty-seven characteristics, we show, by a new risk factor discovery method, that the best SDF varies by factor construction class. On a large collection of test assets, the best SDF from slope factors uniformly prices more portfolios, ETFs, and stocks. These findings support a greater use of slope factors in asset pricing.

JEL Classification: G11, G12, G14

Keywords: factor risk premia; firm level characteristics; marginal likelihood; pricing test; stochastic discount factor; risk factors

1 Introduction

An important question relevant to the vast literature on pricing in 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. Although the dominant approach to creating factors is by sorting methods, typically 3 by 2 sorts of firms by the characteristic of interest and size, a method we call the differential method, there are at least two alternatives, the rank factor method and the slope factor method.[1] In the slope factor method, factors are constructed by running Fama and Macbeth cross-sectional regressions of firm-level returns on firm-level lagged characteristics. The OLS estimates of the slopes in these cross-sectional regressions are long-short portfolios that give unit weighted exposure to each standardized lagged characteristic *and* zero weighted exposure to all other standardized lagged characteristics. Thus, these OLS estimates of the slopes are characteristic-specific long-short portfolios that load on that characteristic, removing the influence of other characteristics included in the regression.

There are fundamental conceptual differences between slope factors on the one hand and differential and rank factors on the other. Because the characteristics at the firm level are correlated, when we double-sort, or use ranks, and take long positions in firms with a high value of a characteristic, we also indirectly take long positions at high values of any characteristic that is positively correlated with that characteristic, and we take long positions at low values of any characteristic that is negatively correlated with that characteristic. Moreover, when we double-sort and take short positions on firms with a low value of a characteristic, we also indirectly take short positions at low values of any characteristic that is positively correlated with that characteristic

[1] Rank factors are constructed by grouping excess returns by market cap and then using the normalized rank of lagged characteristics as weights, for example, Asness, Frazzini, and Pedersen (2019); Kelly, Pruitt, and Su (2019); Chammaen, Pelger, and Zhu (2022); Freyberger, Neuhierl, and Weber (2020); Kozak, Nagel, and Santosh (2020).

and take short positions at high values of any characteristic that is negatively correlated with that characteristic. Therefore, factors constructed by double-sort and rank methods also incorporate returns to positions in characteristics that are correlated with the characteristic of interest.[2] Because it is based on the regression of returns on lagged standardized characteristics, the slope factor method potentially can elegantly overcome this problem.

Recently, Fama and French (2020) have compared differential and slope factors (what they call cross section factors) in the context of the FF5 and FF6 models. In their comparison, Fama and French (2020) replace the factors in the FF5 model by the OLS estimates of the four slopes in the FM regressions

$$r_{it} = \alpha_t + \beta_{t,\text{mve}}\text{mve}_{it-1} + \beta_{t,\text{bm}}\text{bm}_{it-1} + \beta_{t,\text{op}}\text{op}_{it-1} + \beta_{t,\text{inv}}\text{inv}_{it-1} + \varepsilon_{it} \tag{1}$$

where r_{it} is the excess return on equity of firm i, $i = 1,\ldots,n_t$ in month t, $t = 1,\ldots,T$, and the RHS variables are the month $(t-1)$ standardized values of size, the book-to-market ratio, operating profitability, and the rate of growth of assets. Then, for the FF6 model, Fama and French (2020) create a *new* set of slope factors, these being the OLS estimates of the five slopes in the FM regressions

$$r_{it} = \alpha_t + \beta_{t,\text{mve}}\text{mve}_{it-1} + \beta_{t,\text{bm}}\text{bm}_{it-1} + \beta_{t,\text{op}}\text{op}_{it-1} + \beta_{t,\text{inv}}\text{inv}_{it-1} + \beta_{t,\text{mom}}\text{mom}_{it-1} + \varepsilon_{it} \tag{2}$$

where mom_{it-1} is the last period standardized momentum characteristic. Then on 210 test portfolios, Fama and French (2020) show that the FF5 and FF6 models based on slope factors do better pricing than the models based on the original differential factors.

[2] One can attempt to correct this problem by sorting on more characteristics, something that is almost never done. However, it is impractical to sort by more than a few characteristics, so even this solution, if implemented, would best only partially correct the problem.

Although intriguing, the paper leaves several questions unanswered. If mom is correlated with one or more of the other lagged characteristics in the first of the above FM regression, then the OLS estimates (the slope factors) of the four common slopes from the two regressions would be different. There is no discussion in the paper on why one set should be adopted over the other. In fact, it is possible to create any number of different sets of slope factors for the FF5 (or FF6) model. Consider the following specification of the FM regression in the tth cross section:

$$r_{it} = \alpha_t + \beta_{t,\text{mve}}\text{mve}_{it-1} + \beta_{t,\text{bm}}\text{bm}_{it-1} + \beta_{t,\text{op}}\text{op}_{it-1} + \beta_{t,\text{inv}}\text{inv}_{it-1} + \beta'_{t,\text{otherc}}\text{otherc} + \varepsilon_{it} \tag{3}$$

where otherc denotes a vector of other lagged and standardized characteristics that can be entered as controls. The slope factors of mve, bm, op and inv are the OLS estimates of $\beta_{t,\text{mve}}$, $\beta_{t,\text{bm}}$, $\beta_{t,\text{op}}$ and $\beta_{t,\text{inv}}$, respectively. One obtains different slope factors by altering the composition of otherc. If the number of elements in otherc is k, one can construct 2^k different sets of slope factors for the FF5 model. Fama and French (2020) do not consider or comment on this issue. Another question left unanswered is whether the outperformance of slope factors, documented in a set of 210 test portfolios, would persist in a larger and more comprehensive collection of test assets.

In this paper, we extend the analysis in Fama and French (2020) in several crucial respects. Our first point is that to obtain value from slope factors, one should construct slope factors from Fama-Macbeth (FM) regressions that include a broad set of lagged (standardized) characteristics on the right-hand side (RHS).[3] The resulting slope factors are then more pure play (that is, more closely connected to the underlying characteristics) than slope factors constructed with fewer controls. Importantly, only then are the slope factors materially different from the differential and rank

[3] We resolve the ambiguity surrounding controls by including as many controls as possible, upper bounding this number to ensure that enough firms with complete data on that many controls are available and that multicollinearity remains within limits.

factors. Second, we develop generalized slope factors from linear and quadratic controls. These generalized slope factors offer better pricing than slope factors constructed from linear controls. Third, we show that to properly evaluate slope factors in relation to factors constructed by other methods, it is necessary to estimate and then compare SDFs based on the different sets of factors.[4] Fourth, we show that estimated SDFs based on more pure play slope factors do better pricing of the cross section than estimated SDFs based on less pure play slope factors. Finally, we document the evidence in favor of the estimated slope factor-based SDF on a large collection of test assets.

The basis of our analysis are factors constructed from 47 characteristics at the firm level. Data on these characteristics are taken from Green, Hand, and Zhang (2017) and Gu, Kelly, and Xiu (2020), and sourced from Compustat and I/B/E/S, for the period January 1989 to December 2020. We do not consider more characteristics because that tends to reduce the cross-sectional sample sizes and increases multicollinearity, causing instabilities in the FM least-squares regressions. Furthermore, this number of characteristics is already large enough to make a clear distinction between more pure play factors and less pure play factors.

To find risk factors from our pool of slope factors, we use a Bayesian model comparison approach. This approach is based on the model scan methodology of Chib and Zeng (2020) and comprises three steps. We refer to these steps as "pruning - augmentation - model scanning", or PAMS for short.[5] To briefly summarize, in Step 1, the pruning step, we apply a method to prune the set of factors to determine an initial set of risk factors. This pruning step is not based on a purely statistical method such as LASSO, but rather on a finance-driven test of the (incremental) value of each factor, under the assumption that every other factor is a risk factor. Factors that

[4] That the SDF, and the factors in the SDF, the risk factors, are the foundation for pricing is a point forcefully made in Cochrane (2009); see also Feng, Giglio, and Xiu (2020).

[5] Possible alternatives include Kozak et al. (2020), Hwang and Rubesam (2022) and Bryzgalova, Huang, and Julliard (2023).

affirmatively satisfy this test, at a particular level of Bayesian posterior probability confidence, are not pruned. We then have additional steps in the method in which the assessment of the likely risk factors made in Step 1 is revised and updated. In particular, in Step 2, the augmentation step, we infer which of the pruned factors in Step 1 cannot be priced by the factors that were not pruned. This step is a way to catch the false negatives from Step 1. Then in Step 3, the model scanning step, the factors from Step 1, augmented with the non-priced factors from Step 2, are subjected to a model scan to remove false positives. In model scanning, all possible models of the SDF with risk factors and non-risk factors are estimated.[6] These models are compared by Bayesian marginal likelihoods. The risk factors in the best model (the model with the highest marginal likelihood) are then considered the best risk factors.

In our data, PAMS produces twenty slope risk factors, 14 differential risk factors, and 19 rank risk factors. There are fewer differential and rank risk factors because each of these incorporates the returns to implicitly held positions in characteristics correlated with that characteristic. Therefore, fewer of these "jumbo" factors are needed in the SDF. Although the sets of risk factors are different, the best SDFs are directly comparable. Thus, it becomes possible to evaluate the performance of the different construction methods even though the factors themselves are not directly comparable and the number and composition of risk factor sets are different.

In order to fully evaluate the relative worth of the estimated SDFs for pricing the cross section, we consider a large number of test assets consisting of the excess returns on 1150 portfolios, 1480 ETFs and 6024 stocks. In addition, we also consider common (extant) risk factors as test assets. Our pricing comparison shows that the estimated SDF based on slope factors offers better pricing than the estimated SDFs based on differential and rank factors.

[6]This model enumeration and estimation is impossible at the outset because the number of models in the model space with forty-eight factors is $2^{48} - 1$, which is prohibitively large. However, typically, at the end of the pruning and augmentation steps, about 20-25 factors remain, and model scanning is feasible and effective.

The remainder of the paper is organized as follows. In Section 2, we review the three-factor construction methods with supplementary commentary on aspects of the resulting factors that are relevant to our discussion. In Section 3.2 we discuss the Bayesian methodology for the discovery of risk factors that can be applied to our large pool of factors to infer the risk factors most supported by the data. Section 4 presents the risk factors that we discover applying our Bayesian inference procedure. In Section 5 we first detail Bayesian pricing criteria to assess if a testing asset is priced, and we detail the application of these criteria to determine the pricing performance of the different risk factor collections in portfolios, ETFs, and stocks. Section 6 provides some information on why slope factors outperform. Section 7 concludes.

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 et al. (2020), and are sourced from Compustat and I/B/E/S. Our data contain information from 14,860 firms on 40 characteristics from January 1989 to December 2020.[7]

For our analysis, we began 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. Our aim is to be able to estimate cross-sectional regressions for firms that have a complete set of characteristics

[7]The initial data, which we source from Green et al. (2017), has ninety-four characteristics. We pared these down to 47 characteristics using the following reasonable filters. First, to mitigate multicollinearity, characteristics with variance inflation factors greater than seven are discarded. Second, the characteristics with high missingness (greater than 50% of the sample) are also removed. Finally, the characteristics that are indicator variables are removed.

in the preceding cross sections. One common approach is only analyzing firms with nonmissing values of all characteristics in each cross section. However, this approach sacrifices a large part of the data that may contain valuable information. Another approach is to replace the missing value with the mean of those characteristics between firms.

However, the latter approach is somewhat coarse in that it ignores the fact that different characteristics are likely to be correlated and that characteristics are likely, on average, to be different for different-sized firms and for firms in different industries. With this in mind, in our approach to imputing the missing characteristics, we first classify firms into two groups, small and large, by the median value of firm sizes in that cross section, and then within each size group, we further categorize each firm into ten industry groups based on its SIC4 code. In this way, each firm is uniquely assigned to one of the $2 \times 10 = 20$ groups. Then, if any firm has a missing value for a particular characteristic, we replace that missing value with the group mean of that characteristic from firms within the same group. This imputation procedure is based on the assumption that firms of similar size within the same industry would share similar characteristics. It is possible that this nonparametric imputation could be extended to involve other characteristics, but grouping/matching on too many characteristics reduces the group sample size and makes the imputation much more noisy. Our grouping on size and industry, on the other hand, brings in (plausibly) the most relevant information for doing effective imputations.

After imputation, we obtain a rich collection of cross-sectional data sets in which the minimum number of firms is 2051 and the maximum number of firms is 5184. We summarize the data, by the forty-seven characteristics in Table 1.

Table 1 The descriptive statistics of the 47 characteristics

This table presents acronym, full names, definition, and descriptive statistics of characteristics generated by the code from Green et al. (2017). The min, mean, max, 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	Mean	Max	Median	Std
acc	Working capital accruals	Annual income before extraordinary items (ib) minus operating cash flows (oancf) divided by average total assets (at); if oancf is missing then set to change in act - change in che - change in lct + change in dlc + change in txp - dp	-1.022	-0.045	0.500	-0.055	0.109
age	# years since first Compustat coverage	Number of years since first Compustat coverage	1.000	13.000	58.000	16.653	12.94
agr	Asset growth	Annual percent change in total assets (at)	-0.685	0.067	6.062	0.148	0.419
baspread	Bid-ask spread	Monthly average of daily bid-ask spread divided by average of daily spread	0.000	0.035	0.901	0.048	0.051
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	-0.742	0.992	3.937	1.068	0.666
bm	Book-to-market	Book value of equity (ceq) divided by end of fiscal-year-end market capitalization	-2.346	0.531	7.644	0.651	0.605
cash	Cash holdings	Cash and cash equivalents divided by average total assets	-0.079	0.077	0.978	0.161	0.204
cashdebt	Cash flow to debt	Earnings before depreciation and extraordinary items (ib+dp) divided by avg. total liabilities (lt)	-99.683	0.111	2.176	-0.015	1.245
cashpr	Cash productivity	Fiscal-year-end market capitalization plus long-term debt (dltt) minus total assets (at) divided by cash and equivalents (che)	-520.623	-0.072	600.277	-1.224	55.49
cfp	Cash flow to price ratio	Operating cash flows divided by fiscal-year-end market capitalization	-2.797	0.075	2.623	0.074	0.235
chatoia	Industry-adjusted change in asset turnover	2-digit SIC - fiscal-year mean-adjusted change in sales (sale) divided by average total assets (at)	-1.429	0.001	1.194	-0.003	0.216
chcsho	Change in shares outstanding	Annual percent change in shares outstanding (csho)	-0.891	0.008	2.576	0.100	0.298
chmpia	Industry-adjusted change in profit margin	Industry-adjusted change in number of employees	-24.162	-0.077	3.502	-0.151	0.796
depr	Depreciation/PP&E	Depreciation divided by PP&E	-0.984	0.188	6.703	0.307	0.432
dy	Dividend to price	Total dividends (dvt) divided by market capitalization at fiscal-year-end	-6.122	0.000	0.350	0.014	0.033

Table 1 The descriptive statistics of the 47 characteristics

Acronym	Firm characteristics	Definition	Min	Mean	Max	Median	Std
egr	Growth in common shareholder equity	Annual percent change in book value of equity (ceq)	-3.837	0.072	8.286	0.134	0.699
ep	Earnings to price	Annual income before extraordinary items (ib) divided by end of fiscal-year market cap	-7.523	0.043	0.437	-0.038	0.345
gma	Gross profitability	Revenues (revt) minus cost of goods sold (cogs) divided by lagged total assets (at)	-0.961	0.297	1.778	0.345	0.335
grcapx	Growth in capital expenditures	Percent change in capital expenditures from year t-2 to year t	-13.886	0.225	61.947	0.910	3.294
herf	Industry sales concentration	2-digit SIC - fiscal-year sales concentration (sum of squared percent of sales in industry for each company)	0.009	0.044	1.000	0.070	0.077
hire	Employee growth rate	Percent change in number of employees (emp)	-0.711	0.027	3.973	0.088	0.323
idiovol	Idiosyncratic return volatility	Standard deviation of residuals of weekly returns on weekly equal weighted market returns for 3 years prior to month end	0.000	0.055	0.279	0.064	0.037
ill	Illiquidity	Average of daily (absolute return / dollar volume)	0.000	0.000	0.001	0.000	0.000
indmom	Industry momentum	Equal weighted average industry 12-month returns	-0.761	0.112	3.641	0.138	0.284
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)	-0.507	0.032	1.385	0.061	0.153
lev	Leverage	Total liabilities (lt) divided by fiscal-year-end market capitalization	0.000	0.622	77.752	2.211	4.674
lgr	Growth in long-term debt	Annual percent change in total liabilities (lt)	-0.758	0.069	9.612	0.229	0.727
mom12m	12-month momentum	11-month cumulative returns ending one month before month end	-0.957	0.056	11.952	0.125	0.582
mom1m	1-month momentum	1-month cumulative return	-0.721	0.002	2.167	0.011	0.153
mve	Size	Natural log of market capitalization at end of month t-1	2.357	12.313	19.018	12.414	2.257
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	0.000	1.000	8.000	0.989	1.326
operprof	Operating profitability	Revenue minus cost of goods sold - SG&A expense - interest expense divided by lagged common shareholders' equity	-8.828	0.614	13.119	0.783	1.201
pchgm_pchsale	% change in gross margin - % change in sales	Percent change in gross margin (sale-cogs) minus percent change in sales (sale)	-12.26	-0.004	4.761	-0.070	0.843
pricedelay	Price delay	The proportion of variation in weekly returns for 36 months ending in month explained by 4 lags of weekly market returns incremental to contemporaneous market return	-15.849	0.068	15.597	0.153	1.076

Table 1 The descriptive statistics of the 47 characteristics

Acronym	Firm characteristics	Definition	Min	Mean	Max	Median	Std
ps	Financial statement score	Sum of 9 indicator variables to form fundamental health score	0.000	5.000	9.000	4.621	1.655
roaq	Return on assets	Income before extraordinary items (ibq) divided by one quarter lagged total assets (atq)	-0.590	0.005	0.159	-0.004	0.055
roeq	Quarterly return on equity	Earnings before extraordinary items divided by lagged common shareholders' equity	-2.280	0.021	1.766	-0.001	0.154
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)	-23.554	0.059	1.005	-0.143	1.267
salecash	Sales to cash	Annual sales divided by cash and cash equivalents	-300.275	7.889	2503.483	58.36	190.378
saleinv	Sales to inventory	Annual sales divided by total inventory	-35.442	12.160	1031.216	34.889	72.929
salerec	Sales to receivables	Annual sales divided by accounts receivable	-21796	5.949	210.006	11.472	70.916
sgr	Sales growth	Annual percent change in sales (sale)	-0.936	0.086	8.500	0.171	0.522
sp	Sales to price	Annual revenue (sale) divided by fiscal-year-end market capitalization	-4.131	0.870	37.551	1.778	2.882
std_dolvol	Volatility of liquidity (dollar trading volume)	Monthly standard deviation of daily dollar trading volume	0.000	0.708	2.783	0.813	0.427
std_turn	Volatility of liquidity (share turnover)	Monthly standard deviation of daily share turnover	0.000	2.342	736.352	4.833	11.773
tang	Debt capacity/firm tangibility	Cash holdings + 0.715 * receivables + 0.547 * inventory + 0.535 * PPE/ total assets	0.000	0.525	0.982	0.520	0.162
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	-27.344	-0.048	15.362	-0.096	1.685

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 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 other standardized lagged characteristics. Thus, these OLS estimates are characteristic-specific long-short portfolios that load on that characteristic.

To construct these factors, we estimate a sequence of cross-sectional regressions for $t = 1,2,...,T$ (in our sample t run from January 1989 to December 2020). 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 sample vector of excess returns and let the characteristic of j^{th} firm be c_j. Let the sample data on the c_j at the end of time $t-1$ be denoted by the $n_t \times 1$ vector, $\boldsymbol{c}_{j,t-1} = (c_{j,1t-1},...,c_{j,n_tt-1})$, $j = 1,2,....,47$. Let $\tilde{\boldsymbol{c}}_{j,t-1}$ denote the characteristic after standardization, i.e., after subtracting the sample mean and dividing by the sample standard deviation. In vector-matrix notation, the tth cross-sectional regression is given by

$$\boldsymbol{r}_t = \boldsymbol{X}_t\beta_t + \boldsymbol{\varepsilon}_t \tag{4}$$

where $\boldsymbol{X}_t = (\boldsymbol{i}_{n_t}, \tilde{\boldsymbol{c}}_{1,t-1}, \tilde{\boldsymbol{c}}_{2,t-1}, ..., \tilde{\boldsymbol{c}}_{47,t-1})$ is a $n_t \times 48$ matrix of consisting of $\boldsymbol{i}_{n_t}$ (a vector of ones) and sample data on the 47 characteristics. Then, the sequence of OLS estimates of β_t, namely $\hat{\beta}_t = (\hat{\alpha}_t, \hat{\beta}_{1,t}, ..., \hat{\beta}_{47,t}) = (\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\boldsymbol{r}_t$, $t = 1,...,T$, are a sequence of long-short portfolios that load purely on characteristic c_j, $j = 1,2,...,47$. The estimate $\hat{\alpha}_t$ is a long portfolio that can be thought of as the market portfolio.

It is important to understand that the RHS variable $\boldsymbol{X}_t$ in these cross-sectional regressions is standardized for each characteristic j within each cross section. Thus, the sample mean and standard deviation of $\boldsymbol{X}_t$ are zero and one, respectively, for all j and t. Therefore, the lagged variables on the RHS are unitless and the slope coefficients on the RHS are in the same units as the stock returns on the LHS. Only because of this standardization do the OLS slopes become long-short portfolios.

REMARK 1 *In applying this approach, to capture potential nonlinearities, for every characteristic, we also include on the RHS the square of the characteristic (standardized to have a mean of zero and the standard deviation of one). In particular, the matrix of covariates in these cross-sectional regressions is*

$$\boldsymbol{X}_t = (\boldsymbol{i}_{n_t}, \tilde{\boldsymbol{c}}_{1,t-1}, \tilde{\boldsymbol{c}}^2_{1,t-1}, ..., \tilde{\boldsymbol{c}}_{47,t-1}, \tilde{\boldsymbol{c}}^2_{47,t-1})$$

and the long-short portfolio of the j^{th} characteristic is computed by averaging the OLS estimates of the pair of linear and quadratic terms. Our experiments show that this approach produces better slope factors (in the sense that these slope factors provide better pricing of the cross section).

It is not difficult to show the long-short property of the OLS estimates of the slopes from these cross-sectional regressions. For simplicity, consider the case of two characteristics. Then, on letting $\boldsymbol{W}_t' = (\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'$,

$$\hat{\beta}_t = \boldsymbol{W}_t'\boldsymbol{r}_t \tag{5}$$

which can be written out as

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

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

$$\begin{pmatrix} \boldsymbol{w}_0' \\ \boldsymbol{w}_1' \\ \boldsymbol{w}_2' \end{pmatrix} (\boldsymbol{\imath}_{n_t}, \boldsymbol{c}_{1,t-1}, \boldsymbol{c}_{2,t-1}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{6}$$

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

$$\begin{aligned} &\boldsymbol{w}_0'\boldsymbol{\imath}_{n_t} = 1; \;\; \boldsymbol{w}_0'\boldsymbol{c}_{1,t-1} = 0; \;\; \boldsymbol{w}_0'\boldsymbol{c}_{2,t-1} = 0 \\ &\boldsymbol{w}_1'\boldsymbol{\imath}_{n_t} = 0; \;\; \boldsymbol{w}_1'\boldsymbol{c}_{1,t-1} = 1; \;\; \boldsymbol{w}_1'\boldsymbol{c}_{2,t-1} = 0 \\ &\boldsymbol{w}_2'\boldsymbol{\imath}_{n_t} = 0; \;\; \boldsymbol{w}_2'\boldsymbol{c}_{1,t-1} = 0; \;\; \boldsymbol{w}_2'\boldsymbol{c}_{2,t-1} = 1 \end{aligned} \tag{7}$$

Reading these restrictions row by row, we can now conclude that $\hat{\alpha}_t = \boldsymbol{w}_0'\boldsymbol{r}_t$ is a long portfolio (its weights $\boldsymbol{w}_0$ sum to one and it gives zero weighted exposure to the other two characteristics) and can be viewed as the market portfolio; that $\hat{\beta}_{1,t} = \boldsymbol{w}_1'\boldsymbol{r}_t$ is a 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 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.

We supplement these slope factors with the market portfolio from Kenneth French data library. We provide summary statistics of Mkt and the constructed slope factors in Table 2.

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	1.19	4.362	-13.61	11.445
s.acc	-0.081	-0.152	0.672	-1.574	1.769
s.age	-0.03	-0.007	0.225	-0.657	0.570
s.agr	-0.072	-0.071	0.44	-1.425	1.352
s.baspread	-0.111	-0.195	0.697	-1.968	2.453
s.beta	-0.039	-0.027	0.722	-2.242	2.122
s.bm	0.061	0.072	0.338	-0.970	1.020
s.cash	0.050	0.050	0.404	-1.177	1.378
s.cashdebt	0.072	0.046	1.023	-3.029	3.043
s.cashpr	-0.013	-0.012	0.337	-1.167	1.051
s.cfp	-0.023	0.025	0.645	-2.576	1.865
s.chatoia	0.034	0.025	0.345	-0.910	1.045
s.chcsho	-0.033	-0.035	0.234	-0.673	0.862
s.chempia	0.038	0.070	1.531	-4.340	3.910
s.depr	-0.002	-0.003	0.241	-0.640	0.679
s.dy	-0.068	-0.047	0.576	-2.427	2.185
s.egr	-0.017	-0.018	0.318	-0.773	0.872
s.ep	-0.034	0.025	1.666	-6.449	4.263
s.gma	0.042	0.029	0.410	-1.062	1.307
s.grcapx	-0.018	-0.029	0.219	-0.616	0.940
s.herf	-0.02	-0.025	0.211	-0.535	0.870
s.hire	-0.056	-0.056	0.797	-2.805	2.416
s.idiovol	0.018	-0.028	0.641	-1.471	2.331
s.ill	0.158	0.125	0.415	-0.815	1.438
s.indmom	0.032	0.076	1.247	-4.415	3.942
s.invest	-0.007	-0.019	0.327	-0.835	0.938
s.lev	-0.017	-0.022	0.503	-1.765	1.412
s.lgr	0.005	0.001	0.303	-0.864	1.161
s.mom12m	0.078	0.073	0.526	-1.511	1.955
s.mom1m	-0.162	-0.085	0.660	-2.948	1.904
s.mve	0.014	0.031	0.258	-0.786	0.566
s.nincr	0.047	0.040	0.157	-0.394	0.443
s.operprof	0.009	0.004	0.275	-0.807	0.817
s.pchgm_pchsale	0.004	-0.025	0.759	-2.043	2.492
s.pricedelay	-0.022	-0.032	0.271	-0.73	0.799
s.ps	0.020	0.016	0.226	-0.491	0.696
s.roaq	0.104	0.169	1.249	-4.018	3.650
s.roeq	0.023	-0.019	0.716	-2.041	2.277
s.roic	-0.216	-0.168	1.787	-6.459	4.773
s.salecash	-0.002	-0.011	0.254	-0.849	0.687
s.saleinv	0.004	0.000	0.155	-0.464	0.590
s.salerec	-0.012	-0.030	0.748	-2.637	3.395
s.sgr	-0.039	-0.026	0.322	-1.000	0.736
s.sp	0.024	0.018	0.389	-0.967	1.441

Table 2 continued: The descriptive statistics of the slope factors

	Mean	Median	Std	0.5% quantile	99.5% quantile
s.std_dolvol	-0.018	0.020	0.467	-1.693	1.127
s.std_turn	-0.028	-0.040	0.406	-0.890	1.138
s.tang	0.009	0.023	0.353	-0.962	0.922
s.tb	0.008	0.006	0.281	-1.016	1.176

2.3 Differential factors

In contrast to slope factors, differential factors are constructed from characteristics by the 3 by 2 double-sorting method of Fama and French (1993) and Fama and French (2015) only control for size. The goal is to make a (zero-cost) long-short (LS) portfolio that takes long positions on firms with a high value of a given characteristic and short positions on firms with a low value of that characteristic. The return on this portfolio in month t is the realized value of that factor in that month.

Let c_j denote the characteristic of interest other than mve. In each cross-section t, one divides the stocks 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, one further sorts the stocks in the small and large groups into an additional (say) 3 groups based on the 0.3 and 0.7 quantiles of the *lagged* values of that characteristic, $c_{j,i,t-1}$, $i \leq n_t$. Thus, with this double-sorting method, the stocks are allocated to six buckets or, equivalently, an array containing 3 rows and 2 columns. A 3 by 2 arrays such as this is calculated for each characteristic, excluding the size characteristic.

Next, the excess return of these six buckets is value-weighted, i.e. 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

(3,1) and (3,2) buckets. Similarly, a short portfolio is constructed as the sum of the value-weighted returns in the (1,1) and (1,2) buckets. Then, the differential factor for the time period t is given by the difference of these long and short portfolios.

Finally, for the size characteristic mve, the 3 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 create a long-short size portfolio that controls for bm by summing the three rows of the bm array down the first column and subtracting the sum of the three rows in the bm array down the second column. In the same way, we use the 3 by 2 arrays of operating profit and asset growth to create long-short-size portfolios that control operprof and agr. The size factor for this cross section t is then given by the average of these three long-short portfolios. The descriptive statistics of the 47 differential factors are in Table 3.

Table 3 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. Data are from January 1989 to December 2020.

	Mean	Median	Std	0.5% quantile	99.5% quantile
Mkt	0.739	1.190	4.362	-13.610	11.445
d.acc	-0.319	-0.272	1.947	-7.464	4.421
d.age	-0.073	-0.051	2.744	-10.236	9.712
d.agr	-0.374	-0.127	1.862	-7.236	4.150
d.baspread	-0.250	-0.425	7.146	-20.573	26.953
d.beta	0.154	0.209	6.174	-17.78	25.661
d.bm	0.242	0.132	3.162	-10.005	9.874
d.cash	0.416	0.570	3.786	-10.118	13.431
d.cashdebt	0.159	0.170	2.709	-8.537	7.741
d.cashpr	-0.122	-0.010	3.387	-10.024	11.336
d.cfp	0.422	0.371	3.937	-11.273	14.539
d.chatoia	0.177	0.197	1.147	-3.354	2.891
d.chcsho	-0.382	-0.205	2.473	-8.421	5.878
d.chempia	-0.142	-0.033	1.306	-3.354	4.046
d.depr	0.363	0.418	3.238	-10.809	11.270
d.dy	-0.207	-0.330	3.675	-12.629	10.381
d.egr	-0.265	-0.117	1.708	-6.548	4.110

Table 3 continued: The descriptive statistics of the differential factors

	Mean	Median	Std	0.5% quantile	99.5% quantile
d.ep	0.195	0.118	4.212	-14.038	13.255
d.gma	0.277	0.230	2.423	-5.976	6.906
d.grcapx	-0.196	-0.041	1.682	-5.326	4.605
d.herf	-0.135	-0.005	2.170	-7.437	6.352
d.hire	-0.158	-0.049	2.216	-7.063	6.685
d.idiovol	-0.014	0.020	6.726	-21.524	22.623
d.ill	-0.417	-0.024	5.467	-19.937	14.289
d.indmom	0.445	0.485	3.743	-12.049	14.641
d.invest	-0.274	-0.118	2.070	-6.683	6.014
d.lev	0.025	0.121	3.968	-13.492	10.981
d.lgr	-0.188	-0.230	1.522	-4.422	4.573
d.mom12m	0.645	0.745	4.921	-18.853	14.963
d.mom1m	-0.359	-0.194	3.877	-14.820	13.697
d.mve	0.450	-0.486	10.845	-25.309	42.551
d.nincr	0.357	0.388	1.256	-4.262	3.591
d.operprof	0.273	0.283	1.743	-5.091	5.349
d.pchgm_pchsale	0.231	0.242	1.448	-3.743	4.377
d.pricedelay	-0.010	-0.084	1.926	-5.486	6.111
d.ps	0.127	0.202	1.699	-6.298	4.246
d.roaq	0.456	0.556	3.287	-12.429	9.151
d.roeq	0.422	0.473	3.139	-12.559	8.870
d.roic	0.190	0.255	3.145	-10.062	9.460
d.salecash	-0.095	-0.029	2.611	-7.543	8.041
d.saleinv	0.018	0.058	1.741	-4.974	4.715
d.salerec	0.060	-0.072	1.767	-4.791	4.815
d.sgr	-0.246	-0.169	2.148	-5.735	4.788
d.sp	0.288	0.299	3.108	-9.850	10.202
d.std_dolvol	0.191	0.111	3.415	-10.722	10.495
d.std_turn	0.383	0.352	4.562	-12.922	18.009
d.tang	0.302	0.346	2.664	-6.967	8.454
d.tb	0.225	0.249	2.107	-7.066	7.584

2.4 Rank factors

Just as in the previous method, in each cross-section t, one divides the stocks at time t into two groups, small and large, based on the median of market-capitalization, $\text{mve}_{i,t-1}$, $i \leq n_t$. 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 $c_{j,0,t-1} = \{c_{j,i,t-1} : i \in I_{t0}\}$ be the vector of characteristics of length n_{t0} at time $(t-1)$ of all small firms, and similarly let $c_{j,1,t-1} = \{c_{j,i,t-1} : i \in I_{t1}\}$ 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 $rank_{j,0,t-1} = \frac{ra_{j,0,t-1}}{n_{t0}+1}$ 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 $rank_{j,1,t-1} = \frac{ra_{j,1,t-1}}{n_{t1}+1}$. Further, let the sample mean of the values in $rank_{j,0,t-1}$ be denoted by $\bar{rank}_{j,0,t-1}$ and similarly let $\bar{rank}_{j,1,t-1}$ denote the sample mean of the values in $rank_{j,1,t-1}$. Now define the vectors of weights

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

which each sums 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

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

where $\cdot$ is the dot-product operator for multiplying two vectors. The descriptive statistics of the 47 rank factors constructed from our sample are given in Table 4.

Table 4 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. Data are from January 1989 to December 2020.

	Mean	Median	Std	0.5% quantile	99.5% quantile
Mkt	0.739	1.190	4.362	-13.610	11.445
rank.acc	-0.378	-0.258	1.967	-7.135	3.716
rank.age	-0.105	0.055	2.726	-10.976	7.405

Table 4 continued: The descriptive statistics of the rank factors

	Mean	Median	Std	0.5% quantile	99.5% quantile
rank.agr	-0.714	-0.530	2.226	-9.669	4.200
rank.baspread	0.235	-0.047	6.626	-16.581	25.781
rank.beta	0.081	-0.118	5.838	-14.819	23.162
rank.bm	0.434	0.346	2.790	-8.458	8.502
rank.cash	0.422	0.430	3.735	-11.195	14.646
rank.cashdebt	-0.030	0.322	3.396	-15.876	8.055
rank.cashpr	-0.199	-0.085	2.918	-8.186	9.847
rank.cfp	0.290	0.392	4.090	-14.188	13.326
rank.chatoia	0.139	0.163	0.984	-2.371	2.650
rank.chcsho	-0.439	-0.376	2.537	-8.572	6.685
rank.chempia	-0.290	-0.208	1.565	-5.470	3.565
rank.depr	0.397	0.375	3.272	-8.936	13.614
rank.dy	-0.332	-0.540	3.313	-10.173	8.850
rank.egr	-0.439	-0.178	2.227	-10.055	5.071
rank.ep	-0.059	0.047	4.575	-16.842	12.162
rank.gma	0.179	0.291	2.285	-4.794	5.730
rank.grcapx	-0.356	-0.385	1.594	-4.838	3.982
rank.herf	-0.086	-0.087	2.216	-5.838	6.493
rank.hire	-0.378	-0.250	1.947	-5.925	5.140
rank.idiovol	0.298	-0.117	6.545	-16.38	27.745
rank.ill	0.342	0.039	3.744	-8.430	14.625
rank.indmom	0.560	0.639	3.869	-12.145	14.535
rank.invest	-0.458	-0.328	1.828	-5.182	4.682
rank.lev	0.085	0.011	3.970	-14.065	10.533
rank.lgr	-0.436	-0.459	1.358	-4.027	2.899
rank.mom12m	0.338	0.790	4.991	-20.868	11.333
rank.mom1m	-0.838	-0.378	4.321	-16.690	10.584
rank.mve	-0.536	-0.123	4.508	-19.221	9.868
rank.nincr	0.287	0.343	1.207	-4.145	3.306
rank.operprof	0.143	0.290	1.958	-7.497	4.360
rank.pchgm_pchsale	0.147	0.372	1.526	-4.959	3.455
rank.pricedelay	0.084	0.003	1.739	-5.117	5.071
rank.ps	0.057	0.325	2.612	-10.255	5.962
rank.roaq	0.202	0.616	4.020	-15.635	9.586
rank.roeq	0.177	0.356	3.999	-16.539	10.516
rank.roic	-0.046	0.057	3.636	-16.193	9.472
rank.salecash	-0.096	0.164	2.813	-8.062	8.034
rank.saleinv	-0.001	0.062	1.622	-5.766	4.311
rank.salerec	0.087	0.052	1.762	-4.471	6.431
rank.sgr	-0.444	-0.369	1.927	-5.787	4.566
rank.sp	0.413	0.498	3.182	-9.004	11.468
rank.std_dolvol	0.352	0.223	2.785	-8.089	8.834
rank.std_turn	0.365	0.128	4.411	-11.321	18.151
rank.tang	0.321	0.219	2.746	-6.846	11.271
rank.tb	0.037	0.185	2.426	-10.72	7.222

2.5 Are the factors the same?

Due to the conceptual differences between slope factors on the one hand and differential and rank factors on the other, as discussed in the Introduction, the correlation between the corresponding more pure play slope factors and the other two factors is expected to be weak. This is supported by the evidence.

First, consider Table 5 where we give the pairwise correlations between the slope and differential factors (labeled corrsd in the table), that between slope and rank factors (labeled corrsr) and finally between differential and rank factors (labeled corrdr) for each of the forty characteristic-based factors (the f followed by a dot in front of the characteristic name is our notation for a factor corresponding to that characteristic; as there are three different factors for each characteristic, f.acc, for example, stands for s.acc, d.acc and rank.acc). One can see from the entries in the first two columns of this table that the correlation between slope and differential and rank factors tends to be weak.[8]

Table 5 Pairwise correlations between slope, differential, and rank factors.
In the table, f. followed by the characteristic name stands for the factor corresponding to that characteristic, constructed by one of the s, d, and rank methods; corrsd is the pairwise correlation between the s and d factors; corrsr is the pairwise correlation between the s and rank factors; and corrdr is the correlation between the d and rank factors. The last two columns present pairwise correlations between differential and rank factors with less pure play slope factors (which control only for size) analogs.

factor	corrsd	corrsr	corrdr	corrdds	corrrrs
	more pure play			less pure play	
f.acc	0.248	0.266	0.831	0.733	0.936
f.age	0.452	0.501	0.826	0.865	0.981
f.agr	0.093	0.208	0.722	0.737	0.801
f.baspread	0.553	0.662	0.903	0.833	0.971
f.beta	0.809	0.819	0.978	0.974	0.996
f.bm	0.153	0.350	0.860	0.800	0.956

[8]Note that in Table 5, d.mve is negatively correlated with s.mve and r.mve. This is because d.mve is constructed like SMB, with positive weights on small firms and negative weights on large firms, while these signs are flipped on s.mve and r.mve.

f.cash	0.133	0.143	0.957	0.929	0.982
f.cashdebt	-0.007	0.020	0.758	0.568	0.83
f.cashpr	0.078	0.132	0.915	0.792	0.899
f.cfp	0.126	0.175	0.907	0.822	0.957
f.chatoia	0.131	0.337	0.630	0.460	0.845
f.chcsho	0.066	0.130	0.912	0.782	0.844
f.chempia	0.222	0.228	0.630	0.528	0.697
f.depr	0.078	0.170	0.925	0.838	0.948
f.dy	0.110	0.114	0.911	0.838	0.857
f.egr	0.081	0.084	0.714	0.676	0.779
f.ep	0.165	0.211	0.901	0.729	0.919
f.gma	0.318	0.458	0.821	0.714	0.967
f.grcapx	0.036	0.122	0.790	0.610	0.695
f.herf	0.239	0.319	0.819	0.771	0.811
f.hire	0.201	0.240	0.783	0.765	0.879
f.idiovol	0.377	0.559	0.904	0.869	0.987
f.ill	0.072	0.202	0.270	0.207	0.753
f.indmom	0.289	0.223	0.952	0.922	0.975
f.invest	0.303	0.359	0.733	0.647	0.888
f.lev	0.418	0.432	0.958	0.894	0.914
f.lgr	0.003	0.140	0.732	0.614	0.709
f.mom12m	0.291	0.364	0.911	0.903	0.969
f.mom1m	0.371	0.477	0.906	0.875	0.984
f.mve	-0.412	-0.024	-0.764	0.878	0.807
f.nincr	0.364	0.467	0.730	0.716	0.939
f.operprof	0.058	0.149	0.789	0.642	0.883
f.pchgm_pchsale	0.131	0.278	0.736	0.34	0.677
f.pricedelay	0.080	0.161	0.745	0.459	0.831
f.ps	0.240	0.259	0.758	0.763	0.995
f.roaq	0.345	0.344	0.855	0.786	0.937
f.roeq	-0.024	-0.003	0.863	0.789	0.943
f.roic	0.137	0.184	0.817	0.721	0.854
f.salecash	-0.047	-0.089	0.910	0.775	0.792
f.saleinv	0.329	0.394	0.822	0.747	0.833
f.salerec	0.014	0.059	0.786	0.541	0.653
f.sgr	0.135	0.207	0.822	0.697	0.751
f.sp	0.285	0.311	0.917	0.801	0.854
f.std_dolvol	0.478	0.514	0.476	0.563	0.891
f.std_turn	0.288	0.336	0.948	0.869	0.917
f.tang	0.183	0.223	0.901	0.894	0.986
f.tb	0.123	0.116	0.837	0.759	0.91

In keeping with the argument outline above, if the slope factors are less pure, the correlations between slope factors and the other factors will tend to increase. To see this, suppose that we were to construct slope factors that only control for size (we can call these less pure play slope factors). Now, for each characteristic in each cross section, we can sort stocks into three groups (top 30%, middle 40%, bottom 30%), and then include only stocks in the top and bottom groups

in the cross-sectional regression with excess returns on the LHS and that characteristic and size on the RHS. This would produce less pure play slope factor analogs of differential factors. Analogues of rank factors can be made similarly by not dropping the middle 40% of the sample. We give the pairwise correlations in the last two columns of Table 5. As expected, the correlations are much higher. This shows that if one is going to seek value from slope factors, it is essential to construct slope factors that are more pure play.

3 Methodology

3.1 Motivation

It seems reasonable to believe that the more pure play slope factors, due to properties isolated in the preceding discussion, would improve the pricing of the cross section. But to confirm such a conjecture it is not enough to take an existing factor model (say the FF6 model) and replace its five characteristic-based factors (SMB, HML, CMA, RMW, MOM) with the corresponding s.mve, s.bm, s.agr, s.operprof, s.mom12m, factors. This is because if we start from (say) the starting set of 47 slope factors, the best risk factors are not necessarily s.mve, s.bm, s.agr, s.operprof and s.mom12m. Thus, replacing FF6 with the corresponding slope factors is certainly possible, but this does not necessarily reveal the correct differences in pricing ability of the factors constructed by the different methods.

To show what happens if one prices with the slope, differential, and rank versions of the FF6 model, we give in Table 6, the pricing performance on a large cross-section of test assets consisting of 1150 portfolios, 1480 equity ETFs and 6024 stocks.[9] The table gives the number of test assets

[9]Test assets are 1150 portfolios from 5×5 sorts on size (mve) and 46 characteristics from our sample data; 1480

that are priced by the s.ff6, d.ff6 and r.ff6 for each of the three categories of test assets. Ignoring for the moment how we determine if a particular asset is priced (more on this methodology below), one can see that s.ff6 outperforms d.ff6 and r.ff6 in portfolios, that d.ff6 is best for the ETFs, and r.ff6 dominates on stocks.

Table 6 Slope, differential, and rank version of the FF6 factors: pricing performance on 1150 portfolios; 1480 ETFs and 6024 stocks. This table reports the number of assets that are priced at the 0.75 threshold, representing (odds of 3:1 in favor), 0.667 (odds of 2:1 in favor) and 0.80 (odds of 4:1). The factors consist of Mkt and those constructed from mve, bm, agr, operprof, and mom12m. In the case of the slope factors, the factors control 42 other characteristics. Pricing is based on log marginal likelihood differences of regressions with each test asset in the LHS and factors on the RHS, without and with an intercept, as explained in the text. The results show that replacing differential factors in the FF6 model with slope or rank factors does not uniformly improve or worsen performance. To understand which class of factors offers better pricing, it is necessary to find the best risk factors in each class and compare the pricing performance of the best risk factors.

factor set	# priced at 2:1	# priced at 3:1	# priced at 4:1
		1150 Portfolios	
s.ff6	842	732	629
d.ff6	689	517	288
rank.ff6	735	592	393
		1480 ETFs	
s.ff6	693	465	280
d.ff6	889	589	366
rank.ff6	821	546	342
		6024 Stocks	
s.ff6	4342	3009	1910
d.ff6	4363	3020	1930
rank.ff6	4503	3263	2175

On the basis of this sort of comparison, one would tend to reach the conclusion that factors behave differently depending on the asset class and that, since there is no clear winner, one could just continue using one's favored construction method. As we now show, it is a mistake to simply

ETFs obtained from CRSP (share code 73) that have at least 60 months of observation between January 1989 - December 2020; 6024 common stocks obtained from CRSP (share code 10 and 11) that have at least 60 months of observations within January 1989 - December 2020, financial firms, firms with negative book equity, and stocks with P/S lower than $5 are excluded.

replace one set of factors with another set of factors in an existing asset pricing model. This is because different factor construction methods carry different information about the underlying characteristics, and these differences result in different SDFs (equivalently, different risk factors). Therefore, it is important to infer the risk factors within each class of factors and then to compare the pricing performance of these different sets of risk factors.

3.2 PAMS: Pruning - Augmentation - Model Scanning

We now describe a methodology for determining the risk factors from a large starting group of factors. We denote this starting pool of factors by

$$f_t = (f_{1,t}, f_{2,t}, ..., f_{d,t})' \ , \ t \leq T$$

where d denotes the total number of factors. We always assume that the first factor in f_t is the market factor.

Step 1: Soft Pruning

The idea behind the soft pruning step is to remove factors that are unlikely to be risk factors. The factors that are pruned from this step are considered again in Step 2 so the pruning in Step 1 may be called soft pruning as opposed to hard pruning. To decide if this factor is a possible risk factor (or whether it should be pruned), we calculate

$$p_k = \Pr(f_k \text{ is a risk factor} | \boldsymbol{f}_{1:T}, \{f_l\}_{l \neq k} \text{ are risk factors}) \ , \ k \leq d$$

where $\boldsymbol{f}_{1:T}$ is the sample data on the factors and $\{f_l\}_{l\neq k}$ denotes the remaining set of factors. Before we show how this probability can be calculated, it is important to note that the idea behind this question is to interrogate the (incremental) value of each factor, under the assumption that every other factor is a risk factor. Intuitively, factors that have a high value of p_k remain risk factors even when all other factors are a risk factor. For this pruning step, we use a cutoff of.75. In other words, we do not prune f_k if

$$p_k > .75$$

The threshold probability is arbitrary to an extent, but a value higher than 0.75 would unnecessarily omit important factors (increasing the number of false negatives). Although there are additional steps in which the assessments made in Step 1 are revised and updated, omitting important factors in this step, by setting the bar too high, would lead to a misspecified pool of factors, which would jeopardize the effectiveness of the remaining steps. This step is called the pruning step (rather than a selection step) for a reason. Its goal is to eliminate factors that are less likely to be risk factors, not to isolate risk factors (the latter is done in Step 3).

REMARK 2 *To form cut-points, it is useful to think in terms of odds in favor of an event A vs. its complement A^C. The 0.75 threshold represents the odds of 3:1 in favor; other thresholds are 0.667 (odds of 2:1 in favor) and 0.80 (odds of 4:1). We also use the language of odds in our discussion below.*

We are now ready to show how we calculate p_k. Define the two models,

$$\mathbb{M}_{1,k} : f_k \text{ is a risk factor } \cap \{f_l\}_{l\neq k} \text{ are risk factors}$$

and

$$\mathbb{M}_{2,k} : f_k \text{ is not a risk factor } \cap \ \{f_l\}_{l \neq k} \text{ are risk factors}$$

Suppose that the prior probabilities, $\Pr(\mathbb{M}_{1,k})$ and $\Pr(\mathbb{M}_{2,k})$, are each 0.5. Then, by Bayes' Theorem,

$$p_k = \frac{\Pr(\mathbb{M}_{1,k}) m(\boldsymbol{f}_{1:T} | \mathbb{M}_{1,k})}{\Pr(\mathbb{M}_{1,k}) m(\boldsymbol{f}_{1:T} | \mathbb{M}_{1,k}) + \Pr(\mathbb{M}_{2,k}) m(\boldsymbol{f}_{1:T} | \mathbb{M}_{2,k})},$$

where $m_{s,k} = m(\boldsymbol{f}_{1:T} | \mathbb{M}_{s,k})$ $(s = 1,2)$ are the densities of the data under the two models. These are marginal likelihoods (i.e., the sampling densities of the data marginalized over the parameters). On cancelation of the prior probability terms we have

$$p_k = \frac{1}{1 + \exp(-(\log m_{1,k} - \log m_{2,k}))} \ , \ k \leq d$$

To complete the calculation we need to find $m_{1,k}$ and $m_{2,k}$ for all f_k. These can be computed easily from the general theory given below in the discussion of model scanning. To calculate $m_{1,k}$ we estimate the model

$$\boldsymbol{x}_t = \lambda_x + \varepsilon_t,$$

where $\boldsymbol{x}_t$ is the entire set of factors $\boldsymbol{f}_t$, λ_x is the vector of factor risk-premia and ε_t is a correlated Gaussian error. We fit this model under the prior parameters given below and calculate the marginal likelihood (available in closed form). This gives us $m_{1,k}$ (note that this does not depend on k, but we keep this dependence in the notation).

To calculate $m_{2,k}$ we remove f_k from $\boldsymbol{x}_t$ and estimate the model

$$\boldsymbol{x}_{t\backslash f_{k,t}} = \boldsymbol{\lambda}_{x\backslash k} + \boldsymbol{\varepsilon}_{t\backslash k}, \tag{8}$$

$$w_{k,t} = \Gamma \boldsymbol{x}_{t\backslash f_{k,t}} + \varepsilon_{w\cdot x,t} \tag{9}$$

where $\backslash$ is the element exclusion operator, $w_{k,t}$ is $f_{k,t}$, which by virtue of not being a risk factor is priced by the remaining assumed risk factors. The errors are block-correlated Gaussian. Again, under the priors given below, the marginal likelihood $m_{2,k}$ of this model is in closed form.

We repeat this comparison for every factor f_k and denote the factors not pruned by $\boldsymbol{x}^1$.

Step 2: Augmentation

Step 2 is designed to catch any false negatives (factors that are classified as non-risk factors in Step 1 that could be risk factors). Specifically, given the set of factors in $\boldsymbol{x}^1$ we look at the remaining factors in $\boldsymbol{f}$. We call these

$$\boldsymbol{w}^1 = \boldsymbol{f}\backslash\boldsymbol{x}^1$$

These are the factors that were judged to be non-risk factors at the end of the (soft) pruning Step 1. But it is possible that some of these factors are risk factors. Generally, the false negatives in this set tend to be factors whose posterior probability in Step 1 is close to 0.75 from below. The factors whose posterior probability in Step 1 is much smaller than the threshold, say, 0.3 or below, are generally not false negatives. Regardless, to find these false negatives, we ask a fresh question:

> Which factors in $\boldsymbol{w}^1$ are not priced by $\boldsymbol{x}^1$ with a posterior probability of at least 0.75 (ie, posterior odds of 3:1 of not priced vs. priced)?

We set the posterior probability bar at 0.75 to allow more factors to pass this threshold.[10] We calculate this posterior probability of not being priced for each factor in $\boldsymbol{w}^1$. Let a particular factor in $\boldsymbol{w}^1$ be denoted by w_j. Then, for every j, we estimate the following two Bayesian regression models, the first one without an intercept, and the second with an intercept:

$$\begin{aligned} &\text{Observation eq: } w_{j,t} = \beta'_{0,j}\boldsymbol{x}_t^1 + \varepsilon_{0,j,t}\ ,\ \varepsilon_{0,j,t} \sim \mathbb{N}(0, \sigma^2_{0,j})\ , t \leq T \\ &\text{Prior: } \beta_{0,j} \sim \mathbb{N}(\boldsymbol{b}_{0,j}, \boldsymbol{B}_{0,j})\ ,\ \sigma^2_{0,j} \sim \mathbb{IG}\left(\frac{\nu_{0,j}}{2}, \frac{\delta_{0,j}}{2}\right) \end{aligned} \tag{10}$$

and

$$\begin{aligned} &\text{Observation eq: } w_{j,t} = \alpha_j + \beta'_{1,j}\boldsymbol{x}_t^1 + \varepsilon_{1,t}\ ,\ \ \varepsilon_{1,j,t} \sim \mathbb{N}(0, \sigma^2_{1,j})\ , t \leq T \\ &\text{Prior: } \left(\alpha_j, \beta_{1,j}\right) \sim \mathbb{N}(\boldsymbol{b}_{1,j}, \boldsymbol{B}_{1,j})\ ,\ \sigma^2_{1,j} \sim \mathbb{IG}\left(\frac{\nu_{1,j}}{2}, \frac{\delta_{1,j}}{2}\right) \end{aligned} \tag{11}$$

where $\mathbb{IG}$ denotes the inverse gamma distribution. The hyperparameters of the prior distributions are determined from a training sample using the first 30% of the sample (see Greenberg (2012) for more on training sample priors). We estimate these models using MCMC methods and calculate the marginal likelihood of each model in this comparison using the method of Chib (1995), based on the output of the MCMC simulation.

If we let $m_{j,0}$ denote the marginal likelihood of the model without an intercept, and $m_{j,1}$ denote the marginal likelihood of the model with the intercept and then the posterior probability that w_j is not priced by $\boldsymbol{x}^1$ given the data on the factors is

$$\Pr(w_j \text{ is not priced by } \boldsymbol{x}^1 | \boldsymbol{f}_{1:T}) = \frac{1}{1 + \exp\left(-\left(\log m_{j,1} - \log m_{j,0}\right)\right)}$$

[10] In a different context, and with a different aim, Chib, Zhao, and Zhou (2022) also use a not priced test, but to select genuine anomalies from a large pool of anomalies, given a collection of risk factors.

If this posterior probability is at least 0.75 (i.e., posterior odds of not being priced is 3:1), we put w_j in the set $\boldsymbol{w}_{\text{np}}^2$ (where the superscript 2 stands for Step 2 and np for not-priced).

We assemble the factors $\boldsymbol{x}^1$ and $\boldsymbol{w}_{\text{np}}^2$ in the set $\boldsymbol{f}^2$.

Step 3: Model Scanning

The purpose of Step 3 is to find the best risk factors from the set $\boldsymbol{f}^2 = \left\{\boldsymbol{x}^1, \boldsymbol{w}_{\text{np}}^2\right\}$ by exhaustively estimating models with all possible combinations of risk factors and non-risk factors. This step can be seen as screening out any false positives that may be present in $\boldsymbol{f}^2$.

Specifically, the j^{th} model in this scan is defined as

$$\boldsymbol{x}_{j,t} = \boldsymbol{\lambda}_{x,j} + \boldsymbol{\varepsilon}_{x,j,t}, \tag{12}$$

$$\boldsymbol{w}_{j,t} = \Gamma_j \boldsymbol{x}_{j,t} + \boldsymbol{\varepsilon}_{w \cdot x,j,t}, \tag{13}$$

consisting of the risk factors $\boldsymbol{x}_{j,t} : k_{x,j} \times 1$, and the complementary set of factors (the non-risk factors) $\boldsymbol{w}_{j,t} : k_{w,j} \times 1$, and the errors are block independent Gaussian

$$\begin{pmatrix} \boldsymbol{\varepsilon}_{x,j,t} \\ \boldsymbol{\varepsilon}_{w \cdot x,j,t} \end{pmatrix} \sim \mathbb{N}_K \left(\mathbf{0}, \begin{pmatrix} \Omega_{x,j} & \mathbf{0} \\ \mathbf{0} & \Omega_{w \cdot x,j} \end{pmatrix} \right), \tag{14}$$

These splits are formed from the factors in $\boldsymbol{f}^2$. Since the number of factors in $\boldsymbol{f}^2$ is typically much smaller than d (the number of initial factors), the number of models estimated and compared in this scan is manageable. With abuse of notation, let d denote the dimension of $\boldsymbol{f}^2$. There are therefore $J = 2^d - 1$ such splits (assuming that the risk factor set cannot be empty). Each of these combinations defines a particular asset pricing model $\mathbb{M}_j$, $j = 1, ..., J$.

Let

$$\theta_j = (\lambda_{x,j}, \Omega_{x,j}, \Gamma_j, \Omega_{w\cdot x,j})$$

denote the parameters of $\mathbb{M}_j$. Priors, for every j, are given by

$$\pi(\theta_j|\mathbb{M}_j) = \pi(\Omega_{x,j}, \Gamma_j, \Omega_{w\cdot x,j}|\mathbb{M}_j)\pi(\lambda_{x,j}|\mathbb{M}_j, \Omega_{x,j}, \Gamma_j, \Omega_{w\cdot x,j}) \tag{15}$$

where

$$\pi(\Omega_{x,j}, \Gamma_j, \Omega_{w\cdot x,j}|\mathbb{M}_j) = c|\Omega_{x,j}|^{-\frac{2k_{x,j}-k+1}{2}}|\Omega_{w\cdot x,j}|^{-\frac{k+1}{2}},$$

$$\pi(\lambda_{x,j}|\mathbb{M}_j, \Omega_{x,j}, \Gamma_j, \Omega_{w\cdot x,j}) = \mathbb{N}_{k_{x,j}}(\lambda_{x,j}|\lambda_{x,j,0}, \kappa_j\Omega_{x,j}),$$

and $\mathbb{N}_d(\cdot|\mu, \Omega)$ is the d-dimensional multivariate normal density function with mean μ and covariance matrix Ω. These are the priors in Chib, Zeng, and Zhao (2020). They arise as a special case of the priors in Chib and Zeng (2020).

Note that in the model space, there is one model, the full model, in which all factors are in $\boldsymbol{x}$; thus $\boldsymbol{w}$ is empty. Letting this be the first model, we get from the above that its prior is $\pi(\Omega_{x,1}|\mathbb{M}_1) = c|\Omega_{x,1}|^{-\frac{K+1}{2}}$.

Under these priors and the sampling density of the factors given the parameters, the model marginal likelihoods, defined as the integral of the sampling density over the parameters, are available in closed form. In particular, we have

$$\log m_1(\boldsymbol{f}_{1:T}|\mathbb{M}_1) = -\frac{Tk}{2}\log\pi - \frac{k}{2}\log(T\kappa_1 + 1) - \frac{T}{2}\log|\Psi_1| + \log\Gamma_d\left(\frac{T}{2}\right), \tag{16}$$

and

$$\log m_j(\boldsymbol{y}_{1:T}|\mathbb{M}_j) = -\frac{Tk_{x,j}}{2}\log\pi - \frac{k_{x,j}}{2}\log(T\kappa_j+1) - \frac{\left(T+k_{x,j}-d\right)}{2}\log|\Psi_j| + \log\Gamma_{k_{x,j}}\left(\frac{T+k_{x,j}-d}{2}\right)$$
$$-\frac{(d-k_{x,j})(T-k_{x,j})}{2}\log\pi - \frac{\left(d-k_{x,j}\right)}{2}\log|W_j^*| - \frac{T}{2}\log|\Psi_j^*| + \log\Gamma_{d-k_{x,j}}\left(\frac{T}{2}\right),\ j>1, \tag{17}$$

and

$$\Psi_j = \sum_{t=1}^{T}(\boldsymbol{x}_{j,t}-\hat{\lambda}_{x,j})(\boldsymbol{x}_{j,t}-\hat{\lambda}_{x,j})' + \frac{T}{T\kappa_j+1}\left(\hat{\lambda}_{x,j}-\lambda_{xj0}\right)\left(\hat{\lambda}_{x,j}-\lambda_{xj0}\right)'$$
$$W_j^* = \sum_{t=1}^{T}\boldsymbol{x}_{j,t}\boldsymbol{x}_{j,t}'\ ,\ \ \Psi_j^* = \sum_{t=1}^{T}(\boldsymbol{w}_{j,t}-\hat{\Gamma}_j\boldsymbol{x}_{j,t})(\boldsymbol{w}_{j,t}-\hat{\Gamma}_j\boldsymbol{x}_{j,t})'.$$

Note that the variables in the hat in the above expressions are the least squares estimates calculated using the estimation sample, and $\Gamma_d(\cdot)$ denotes the d dimensional multivariate gamma function.

If we give each model in the model space the prior probability

$$\Pr(\mathbb{M}_j) = 1/J \tag{18}$$

then from Bayes theorem one gets that the posterior model probability of $\mathbb{M}_j$ is

$$\Pr(\mathbb{M}_j|\boldsymbol{f}_{1:T}) = \frac{1}{1+\sum\limits_{l=1,l\neq j}^{J}\exp\left(-\left(\log m_j(\boldsymbol{f}_{1:T}|\mathbb{M}_j)-\log m_l(\boldsymbol{f}_{1:T}|\mathbb{M}_l)\right)\right)} \tag{19}$$

We rank models according to these posterior probabilities. Let the model that has the largest log marginal likelihood be $\mathbb{M}_{j^*}$. Then the risk factors in this model, namely $\boldsymbol{x}_{j^*}$, are considered the best risk factors.

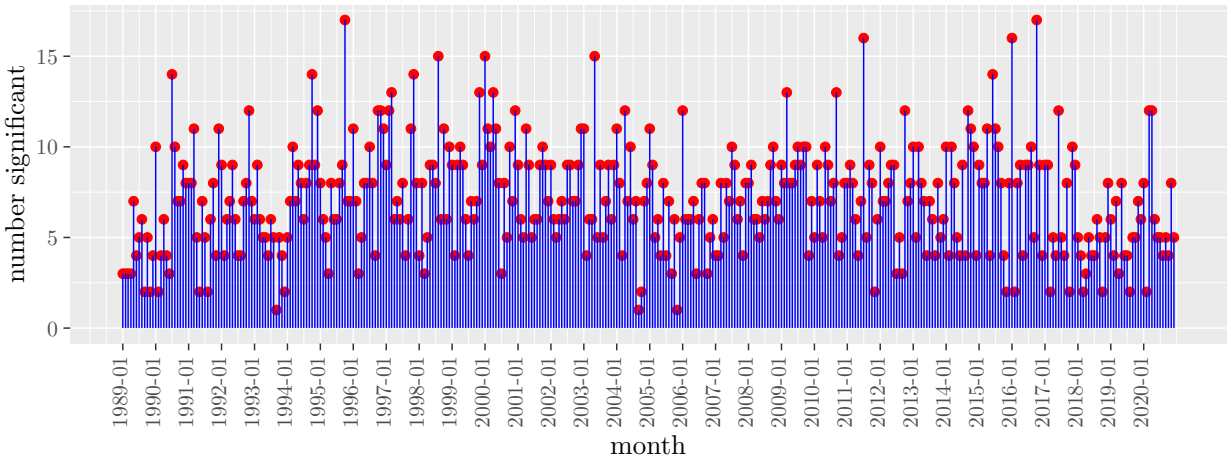

Figure 1 Number of significant characteristics ($|\text{t-ratio}| > 2.5$) in month-by-month cross-sectional regressions of firm-level excess returns on lagged standardized characteristics.

4 Evidence

For motivation, it can be insightful to run cross-sectional regressions of returns on lagged standardized characteristics to see which characteristics are significant in such regressions, and how that significance changes over time.

In Figure 1, we plot the number of significant characteristics (including the constant) in monthly regressions of excess firm returns on standardized lagged characteristics, where significance is measured by $|\text{t-ratio}| > 2.5$. What the plot shows is that the number of significant characteristics varies from a minimum of one to a maximum of seventeen and that there is considerable variation in this number across months. If there was no variation at all in this number, we might expect that, at least for slope factors, the risk factor discovery procedure would discover that many risk factors and that these would be the ones that correspond to the significant characteristics.

For the data at hand, given the variability over months, it is difficult to determine from such

a calculation which factors and how many would end up being risk factors. This connection is even more strained for differential and rank factors because these factors capture the influence of several characteristics at once, as discussed above. However, the somewhat long right tail of the distribution of significant characteristics suggests that 10-20 risk factors may be needed to price the cross section.

4.1 Step 1

To avoid repetition, we use the example of slope factors to detail the risk factor discovery process. The data on these factors are

$$\boldsymbol{f}_{\text{s},t} = (\text{Mkt}_t, \text{s.acc}_t, \text{s.age}_t, \text{s.agr}_t, ..., \text{s.tb}_t),\ t = 1, 2, ..., T$$

and the goal is to use these data to learn about the factors that are in the SDF.

As discussed above, the PAMS procedure involves three steps. Step 1 is a dimension reduction step. Factors that are unlikely to be risk factors are soft-pruned before the model scanning step. For easy replicability, we have packaged Step 1 in R code with the simple call `x1 = Step1(data = Sf,trainpct = .3,workers = 25,probcut = .75)`, where the first argument takes in the slope factor data.frame object, the second argument directs the function to use the first 30% of the sample to fix the prior hyperparameters, the third argument specifies the number of cores for parallel processing and the last argument sets the posterior probability cutoff to determine $\boldsymbol{x}_1$, the factors that are not pruned. The highly optimized computations coded in C++ take 14.23 seconds on a Macbook Pro computer with an M1 Max chip.

We give the results in Figure 2. It is clear from this plot that the probability cut-point of 0.75

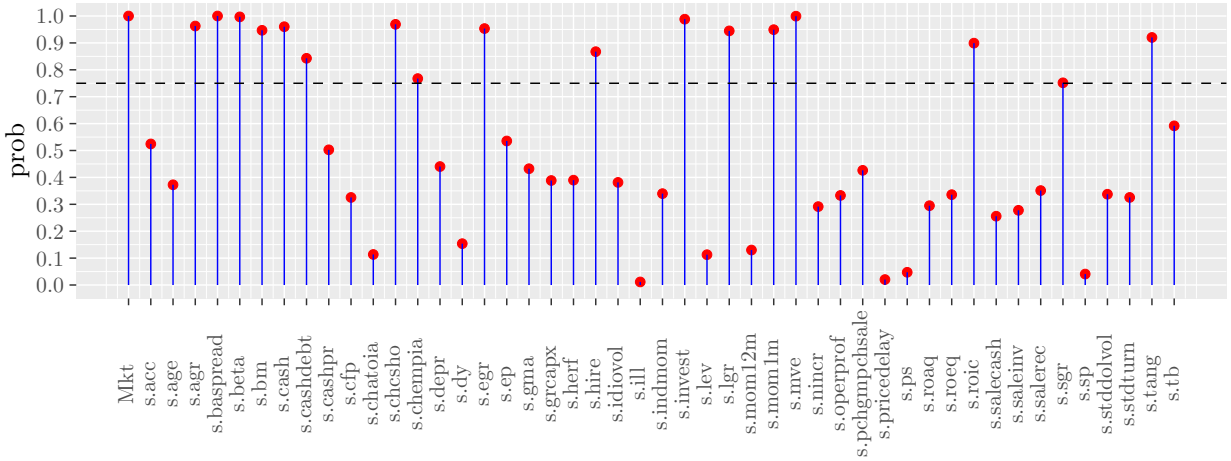

Figure 2 Step 1: The posterior probability that s.j (as j runs through the 48 factors) is a risk factor, given the data and all other factors are risk factors. The horizontal line has a probability cut-value of 0.75. Factors that have a posterior probability exceeding this threshold are not pruned.

effectively divides the set of 48 factors into two distinct groups. Eighteen factors, namely,

$$\begin{aligned}\boldsymbol{x}_{\mathrm{s}}^{1} =&(\mathrm{Mkt}, \mathrm{s.agr}, \mathrm{s.baspread}, \mathrm{s.beta}, \mathrm{s.bm}, \mathrm{s.cash}, \mathrm{s.cashdebt}, \mathrm{s.chcsho}, \mathrm{s.chempia}, \mathrm{s.egr},\\ &\mathrm{s.hire}, \mathrm{s.invest}, \mathrm{s.lgr}, \mathrm{s.mom1m}, \mathrm{s.mve}, \mathrm{s.roic}, \mathrm{s.sgr}, \mathrm{s.tang})\end{aligned}$$

have a posterior probability that exceeds the threshold, where the subscript s is the label for slope factors. These are the factors that are not pruned in Step 1. The figure shows that for each of the remaining 30 slope factors, the posterior probability of being a risk factor, given the rest are risk factors, is much lower than the threshold and are, therefore, *soft* pruned. Recall that soft pruning is not final. Factors that are soft-pruned are re-examined in Step 2.

REMARK 3 *As one can observe from Figure 2, the composition of $\boldsymbol{x}_s^1$ is robust to a reduction in the threshold to, say, 0.67 (for 2:1 odds). However, going in the other direction would reduce the cardinality of $\boldsymbol{x}_s^1$, and this would not be desirable since the potential risk factors would probably be*

trimmed. Even though it is possible that such pruned risk factors (false negatives) may be identified in Step 2, this may not happen because the pricing test conducted in Step 2 would be less effective if the RHS variables $\boldsymbol{x}_s^1$ used in that test came in misspecified. Our experiments conducted on many simulated data sets confirm this guidance. The probability threshold can be reduced from 0.75 but should generally not be adjusted upward.

4.2 Step 2

The next step in the discovery procedure is to examine the thirty pruned slope factors of Step 1 to determine which of these thirty can or cannot be priced by $\boldsymbol{x}_{\text{s}}^1$. We have implemented this pricing test in another R function called `wnp2 = pricing (x1 = x1, data = data, workers = 25)`, where the first argument inputs the factors that come out of Step 1, and the second and third arguments are as in Step 1. This function goes through the factors in the factor data set that are not in $\boldsymbol{x}_{\text{s}}^1$ and for each of those factors fits the two Bayesian regression models described above by MCMC methods, with priors determined from a training sample using 15% of the data, and the marginal likelihoods of every model computed by the method of Chib (1995). The function returns the names of the factors that are not priced by $\boldsymbol{x}_{\text{s}}^1$ at a minimum of 3:1 odds (posterior probability of at least 0.75).

On applying this function, with a run-time of 9.7 seconds on a Macbook Pro M1 Max computer, we get that

$$\boldsymbol{w}_{\text{s.np}}^2 = (\text{s.acc, s.ill, s.mom12m, s.nincr, s.std\_dolvol, s.std\_turn})$$

are the six out of thirty factors that are not priced by $\boldsymbol{x}_{\text{s}}^1$. Therefore, at the end of the two steps our set of potential risk factors consists of $\boldsymbol{x}_{\text{s}}^1$ augmented with the six factors in $\boldsymbol{w}_{\text{s.np}}^2$, for a total of twenty-four out of forty-eight factors.

4.3 Step 3

In the last step of the discovery procedure, we take the factors

$$\boldsymbol{f}_{\mathrm{s}}^{2} = \left\{\boldsymbol{x}_{\mathrm{s}}^{1}, \boldsymbol{w}_{\mathrm{s.np}}^{2}\right\}$$

and estimate and compare the 16.77722 million possible splits of $\boldsymbol{f}_{\mathrm{s}}^{2}$ into risk factors and non risk factors. Although, undoubtedly, this systematic estimation of all possible splits is computationally intensive, it represents a reliable way of removing any false positives and isolating the risk factors that are best supported by the data.

In practice, one can reduce the computational intensity of model scanning (without any change in the final answer) by including some factors from $\boldsymbol{f}_{\mathrm{s}}^{2}$ in all models. The idea is to look for factors that are almost certain to be among the best of 16.77722 million models. Such factors can be found by applying the calculation of Step 1, but this time to the factors in $\boldsymbol{f}_{\mathrm{s}}^{2}$. By setting a very high inclusion probability threshold of 0.995 one can ensure that one finds just those factors that are certain to be in the final model. In the current problem, this calculation shows that {Mkt, s.baspread, s.chcsho, s.invest, s.mve} each has a posterior probability greater than 0.995 of being risk factors, given that the rest of the factors in $\boldsymbol{f}_{\mathrm{s}}^{2}$ are risk factors. Therefore, the model scan can be performed on the subset of models that contain these five factors as risk factors. The dimension of this restricted model space is $2^{19} - 1 = 524,287$, which is smaller than the entire model space by a factor of 32.

We have coded an R function to perform these computations efficiently and quickly. It has the simple call `Step3(data=data,x1=x1,wnp2=wnp2,mustinclude=TRUE,probformustinclude=.995, trainpct=.3,workers=25)`, where the `mustinclude` argument signals that those factors with

`probformustinclude` greater than 0.995 should be held fixed in every model combination. For this problem, this scan takes about 31 minutes on a Macbook Pro M1 Max computer. In contrast, the scan with `mustinclude = FALSE` takes approximately 16 hours and produces exactly the same final best model.

These computations show that the SDF best supported by the evidence contains the slope factors

$$\begin{aligned}\boldsymbol{x}_{\mathrm{s}}^{*} =&(\text{Mkt, s.acc, s.agr, s.baspread, s.beta, s.bm, s.cash, s.chcsho, s.chempia, s.egr}\\ &\text{s.hire, s.ill, s.invest, s.lgr, s.mom1m, s.mve, s.nincr, s.roic, s.sgr, s.std\_turn})\end{aligned}$$

While twenty risk factors may seem non-standard, it is a consequence of the more pure play property. Since each factor is purged of the influence of a large number of other characteristics, more of these factors are needed in the SDF.

In Table 7 we report the first two centered moments of the marginal posterior distributions of the factor risk premia λ^* and the market price of factor risks (the SDF loadings) $\boldsymbol{b}^*$. The parameters are starred to emphasize that these are summaries of the risk factors in the best model. From the table one can see that 19 of the 20 twenty market prices of factor risks have an absolute value of the posterior mean more than twice bigger than the posterior sd. Fewer than 19 factor risk premia are significant in the same way, but this is less important than the significance of the market price parameters.

Table 7 Slope factor SDF: Posterior estimates of the factor risk premia λ^* and market prices of factor risks $\boldsymbol{b}^*$ of $\boldsymbol{x}_s^*$, the 20 slope risk factors.
This table shows the posterior estimates (mean and standard deviation) of the factor risk premia λ^* and the SDF coefficients $\boldsymbol{b}^*$ of the slope risk factors in the best model from Step 3 of the PAMS discovery method. SDF loadings with 95% posterior credibility intervals excluding zero are marked in bold. The data runs from January 1989 to December 2020.

	λ^*		$\boldsymbol{b}^*$	
$\boldsymbol{x}_s^*$	Mean	Std	Mean	Std
Mkt	0.682	0.274	**0.119**	0.025
s.acc	-0.079	0.041	-0.205	0.111
s.agr	-0.054	0.027	**-0.926**	0.269
s.baspread	-0.125	0.044	**-0.533**	0.122
s.beta	-0.049	0.047	**-0.389**	0.136
s.bm	0.055	0.020	**0.800**	0.243
s.cash	0.055	0.025	**0.435**	0.197
s.chcsho	-0.031	0.015	**-1.319**	0.329
s.chempia	0.038	0.098	**-0.144**	0.068
s.egr	-0.025	0.019	**-0.917**	0.269
s.hire	-0.057	0.044	**-0.389**	0.145
s.ill	0.131	0.023	**0.729**	0.209
s.invest	-0.012	0.020	**-0.935**	0.274
s.lgr	0.002	0.017	**-0.759**	0.354
s.mom1m	-0.133	0.041	**-0.258**	0.116
s.mve	0.001	0.016	**-0.754**	0.307
s.nincr	0.042	0.009	**1.523**	0.492
s.roic	-0.264	0.117	**-0.114**	0.042
s.sgr	-0.044	0.019	**-0.494**	0.250
s.std_turn	-0.038	0.025	**-0.402**	0.185

4.4 Risk factor discovery: Differential and Rank factors

Suppressing details to save space, PAMS applied to the differential and rank factors reveals 14 and 19 risk factors, respectively. These are

$$\begin{aligned}
\boldsymbol{x}^*_{\text{d}} =&(\text{Mkt, d.acc, d.agr, d.beta, d.grcapx, d.herf, d.hire, d.mve,}\\
&\text{d.operprof, d.pchgm_pchsale, d.roaq, d.sgr, d.std_turn, d.tb})\\
\boldsymbol{x}^*_{\text{rank}} =&(\text{Mkt, rank.agr, rank.bm, rank.cash, rank.cashdebt, rank.cashpr, rank.cfp, rank.depr,}\\
&\text{rank.herf, rank.indmom, rank.lgr, rank.mom1m, rank.mve, rank.operprof,}\\
&\text{rank.pricedelay, rank.ps, rank.roaq, rank.sgr, rank.tb})
\end{aligned}$$

Table 8 shows the posterior distribution of the factor risk premia λ^* and the SDF loadings $\boldsymbol{b}^*$ of these differential and rank risk factors. Most of the SDF loadings are significant in the Bayesian sense.

We collect the different sets of risk factors in Figure 3, with the characteristics labeled on the x-axis and the risk factors indicated by points, colored red for slope, green for differential and blue for rank risk factors. Though there is some overlap, the risk factor sets are broadly different.[11] Thus, the method of factor construction matters for which risks are captured. To our knowledge, this is the first time such a result has been documented in the literature. Thus, the factor construction method should not be an off-handed choice.

[11] It is not the aim of this paper to provide a theoretical justification for these risk factors. However, the empirical finding that the SDF varies depending on how the factors are constructed has practical implications. For example, to check any factor-based economic theory of asset pricing, it would not be enough to base that confirmation on factors constructed by one method. Because what may be true with differential factors may not be true with slope factors, and vice versa.

Table 8 Differential and rank factor SDFs: The posterior estimates of the factor risk premia λ^* and market prices of factor risks b^*. There are fourteen factors in the differential factor SDF and nineteen in the rank factor SDF.
This table shows the posterior estimates (mean and standard deviation) of the factor risk premia λ^* and SDF coefficients b^* of the differential and rank risk factors in the best model from Step 3 of the PAMS discovery method. Market prices of factor risks with 95% posterior credibility intervals excluding zero are marked in bold. The data runs from January 1989 to December 2020.

	λ^*		b^*			λ^*		b^*	
x_d^*	Mean	Std	Mean	Std	x_r^*	Mean	Std	Mean	Std
Mkt	0.682	0.270	**0.165**	0.029	Mkt	0.680	0.273	**0.222**	0.035
d.acc	-0.360	0.127	-0.070	0.045	rank.agr	-0.66	0.132	**-0.375**	0.106
d.agr	-0.365	0.112	**-0.256**	0.074	rank.bm	0.329	0.175	**0.426**	0.114
d.beta	0.189	0.402	**-0.088**	0.034	rank.cash	0.444	0.251	**0.410**	0.090
d.grcapx	-0.161	0.103	0.108	0.056	rank.cashdebt	-0.038	0.207	**-0.365**	0.121
d.herf	-0.091	0.142	**-0.269**	0.045	rank.cashpr	-0.094	0.184	**0.572**	0.128
d.hire	-0.150	0.137	0.125	0.074	rank.cfp	0.205	0.268	0.141	0.086
d.mve	0.782	0.654	**0.034**	0.010	rank.depr	0.395	0.212	**-0.241**	0.092
d.operprof	0.29	0.102	**0.202**	0.053	rank.herf	-0.085	0.137	**-0.356**	0.072
d.pchgm_pchsale	0.226	0.091	**0.118**	0.058	rank.indmom	0.485	0.257	**-0.050**	0.025
d.roaq	0.413	0.217	**0.184**	0.049	rank.lgr	-0.379	0.081	**0.364**	0.129
d.sgr	-0.281	0.129	**-0.185**	0.067	rank.mom1m	-0.767	0.282	**-0.060**	0.024
d.std_turn	0.443	0.294	**0.136**	0.043	rank.mve	-0.520	0.260	**-0.370**	0.065
d.tb	0.233	0.133	**0.117**	0.053	rank.operprof	0.163	0.123	**0.365**	0.088
					rank.pricedelay	0.072	0.107	**0.151**	0.071
					rank.ps	0.019	0.165	**0.539**	0.128
					rank.roaq	0.161	0.254	**0.287**	0.09
					rank.sgr	-0.415	0.113	**-0.237**	0.085
					rank.tb	0.064	0.150	**0.254**	0.075

5 Pricing of the cross section

To understand the pricing capabilities of the different sets of risk factors, we consider a large collection of assets consisting of portfolios, ETFs and stocks. These are the same test assets used in Table 6 above. We also discuss the pricing of existing risk factors in prevalent asset pricing models.

We primarily apply a Bayesian procedure to infer which assets are priced though we also

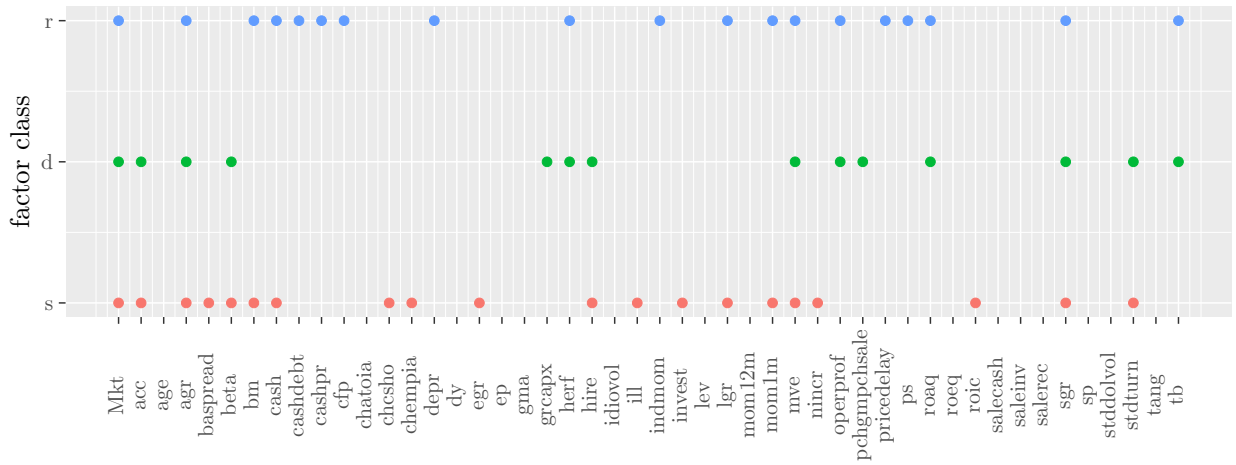

Figure 3 Risk factors within each factor class. There are 20 slope risk factors (colored red), 14 differential risk factors (colored green), and 19 rank risk factors (colored blue), as described in the text. The underlying characteristics are labeled on the horizontal axis.

consider the frequentist test for the null that the intercept is zero. The Bayesian pricing methodology is the same as in Step 2 of the discovery procedure, with the change that the variable w_j in equations (10) and (11) now stands for the excess return of the j^{th} test asset. If we let $m_{j,0}$ denote the marginal likelihood of the regression model (10) with w_j on the LHS and the risk factors on the RHS *without* an intercept, and $m_{j,1}$ denote the marginal likelihood of the model (11) with w_j on the LHS and the risk factors on the RHS *with* an intercept, then the posterior probability that w_j is priced by the risk factors $\boldsymbol{x}_{\mathrm{f}}^*$, $\mathrm{f} \in \{\mathrm{s,d,r}\}$, given the data on the risk factors is

$$\Pr(w_j \text{ is priced by } \boldsymbol{x}_{\mathrm{f}}^* | \boldsymbol{f}_{1:T}) = \frac{1}{1+\exp\left(-\left(\log m_{j,0} - \log m_{j,1}\right)\right)}$$

It can be checked that w_j is priced by $\boldsymbol{x}_{\mathrm{f}}^*$ with at least 2:1 odds, if

$$d_{j,01} = \left(\log m_{j,0} - \log m_{j,1}\right) > 0.69$$

It is priced at posterior odds of 3:1 if $d_{j,01} > 1.09$ and priced at posterior odds of 4:1 if $d_{j,01} > 1.38$. Once again, the prior hyperparameters are based on the initial 15% of the data, with the following qualifications. If the sample size of a certain asset is 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, the first 15form the training sample prior. Then, for each asset, we estimate two models (one without an intercept and one with) to obtain evidence about pricing versus nonpricing. This evidence is summarized by the number priced vs. not priced at 2:1, 3:1, and 4:1 posterior odds.

5.1 Portfolios, ETFs and Stocks

The pricing performance of the different sets of risk factors is given in Table 9. To benchmark the results, we also include the pricing performance of the FF6 factors of Fama and French (2018) (the FF6 factors are constructed using the differential / sorting method, downloaded from the Kenneth French data library).

Consider first the case of portfolios (double-sorted portfolios) that we have constructed from our sample. In particular, for each of our 46 characteristics (excluding size), we construct 5×5 sorts on size and characteristic, leading to 1150 ($= 46 \times 25$) value-weighted portfolios. Under the at least 2:1 posterior odds threshold criteria, the slope risk factors price 996 of these portfolios, while the differential and rank factor models price 778 and 763, respectively, a substantial difference in pricing ability. Under the even more demanding 4:1 posterior odds threshold, the slope risk factors price 697 of these portfolios, compared to 342 and 542 by the differential and slope risk factors, respectively. Each set of risk factors provides better pricing than the benchmark FF6 risk factors.

Consider next the sample of 1480 ETFs that we have assembled from CRSP (share code 73), spanning February 1993 (the earliest month data on ETFs is available) to December 2020. Within this period, each ETF has at least 60 months of data. ETFs are diversified portfolios that are liquid, transparent, and inexpensive 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. Again, under the least 2:1 posterior odds in favor of pricing criteria, one can see that slope risk factors outperform the other sets of risk factors.

Finally, we consider pricing a sample of 6024 stocks (CRSP sharecodes 10 and 11). The sample period is January 1989 to December 2020, and we ensure that there are at 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. In our view, it is useful to benchmark the pricing performance of stocks given that these assets tend to be more volatile than portfolios, which poses a greater hurdle.

Table 9 gives performance evidence. As can be seen from the table, under the (default) 2:1 criteria, the slope risk factors price 4883 stocks, and the differential and rank risk factors price 4738 and 4801 stocks, respectively, and the FF6 risk factors price 4331 stocks. These results provide evidence that one can price more of the cross-section of stocks (relative to the differential/sorted construction method) by adopting either the rank construction method (for some improved performance), or the slope construction method (for even greater improved performance). These gains can be achieved with effectively zero marginal effort.

Table 9 Slope, differential and rank risk factors: pricing performance on 1150 portfolios; 1480 ETFs and 6024 stocks. This table reports the number of assets that are priced at 0.75 threshold representing (odds of 3:1 in favor), 0.667 (odds of 2:1 in favor), and 0.80 (odds of 4:1). $\boldsymbol{x}_{\text{s}}^*$ denote the twenty slope risk factors; $\boldsymbol{x}_{\text{d}}^*$ denote the 14 differential risk factors, and $\boldsymbol{x}_{\text{rank}}^*$ denote the 19 rank risk factors discovered by the 3-step Bayesian methodology developed in the paper. Whether an asset is priced is determined by the log marginal likelihood differences of regressions with each test asset on the LHS and respective risk factors on the RHS, without and with an intercept, as explained in the text. We include the pricing results of the FF6 risk factors as a reference. The results show that slope risk factors uniformly price more of these test assets than differential, rank and FF6 risk factors.

factor set	# priced at 2:1	# priced at 3:1	# priced at 4:1
		1150 Portfolios	
$\boldsymbol{x}_{\text{s}}^*$	996	867	697
$\boldsymbol{x}_{\text{d}}^*$	778	566	342
$\boldsymbol{x}_{\text{rank}}^*$	763	640	542
FF6	460	328	202
		1480 ETFs	
$\boldsymbol{x}_{\text{s}}^*$	1190	1036	854
$\boldsymbol{x}_{\text{d}}^*$	1058	788	555
$\boldsymbol{x}_{\text{rank}}^*$	1012	806	597
FF6	870	561	343
		6024 Stocks	
$\boldsymbol{x}_{\text{s}}^*$	4883	3983	3071
$\boldsymbol{x}_{\text{d}}^*$	4738	3528	2378
$\boldsymbol{x}_{\text{rank}}^*$	4801	3821	2887
FF6	4331	2891	1789

5.2 Pricing of common risk factors

We conclude our evaluation of pricing power using the risk factors in the FF6 Fama and French (2018), Q5 Hou, Mo, Xue, and Zhang (2021), and DHS Daniel, Hirshleifer, and Sun (2020) models as test assets. The pricing methodology is the same: the FF6, Q5, and DHS factors are on the LHS of regressions, with the slope/differential/rank factors on the RHS. We then fit these models without an intercept and with an intercept and calculate $d_{01} = \log m_0 - \log m_1$, the difference in the respective log marginal likelihoods. Furthermore, we also estimate the model with an intercept by OLS and calculate $\hat{\alpha}$, the OLS estimate of the intercept and the associated absolute value of the

t-statistic.

Table 10 Slope/differential/rank risk factors: pricing of FF6, Q5 and DHS risk factors. The columns labeled $\hat{\alpha}$ have the intercept estimates in regressions with the factor in the LHS and the risk factors $\boldsymbol{x}_{\mathrm{f}}^{*}$ in the RHS; the columns labeled |t| has the absolute value of the OLS estimates of the intercept; and the columns labeled d_{01} have the difference in the log marginal likelihoods of the model (equation (10)) without an intercept, and the log marginal likelihood of the model (equation (11)) with an intercept: these log marginal likelihoods are computed by the method of Chib (1995) under priors based on the first 15% of the data. A factor is priced at posterior odds of 2:1 if $d_{01} > 0.69$. These are marked in bold.

	$\hat{\alpha}$	|t|	d_{01}	$\hat{\alpha}$	|t|	d_{01}	$\hat{\alpha}$	|t|	d_{01}
	s risk factors $\boldsymbol{x}_{\mathrm{s}}^{*}$			d risk factors $\boldsymbol{x}_{\mathrm{d}}^{*}$			rank risk factors $\boldsymbol{x}_{\mathrm{rank}}^{*}$		
SMB	0.14	1.12	**1.00**	-0.11	0.99	**0.72**	0.25	1.48	-0.56
HML	-0.11	0.58	**0.72**	0.12	1.05	**2.25**	-0.22	2.10	0.00
RMW	0.12	0.88	**1.04**	0.05	0.57	**0.95**	-0.01	0.11	**1.43**
CMA	0.20	1.73	**0.79**	0.02	0.27	**0.95**	-0.03	0.33	**1.38**
MOM	0.39	1.43	**1.79**	-0.10	0.43	**0.71**	-0.05	0.30	**1.31**
ME	0.07	0.49	**1.64**	-0.14	1.28	0.43	0.18	1.07	0.11
IA	0.20	1.63	**1.24**	0.00	0.03	**1.15**	-0.09	0.95	**0.99**
ROE	0.30	2.09	**4.72**	0.17	1.83	**1.74**	0.12	1.13	**2.09**
EG	0.50	4.55	-3.85	0.37	3.83	-1.88	0.29	2.45	-0.20
PEAD	0.47	3.75	-1.60	0.44	3.61	-1.44	0.40	3.13	-0.97
FIN	0.31	1.45	**1.34**	0.38	2.74	-0.42	0.23	1.49	**1.59**

The results are given in Table 10. The key columns are the third, sixth, and ninth. Focusing on the slope factors panel, one sees that d_{01} exceeds 0.69 for all risk factors except for EG in the Q5 model and PEAD in the DHS model. Therefore, slope risk factors can price nine of these common risk factors at posterior odds of 2:1. Reading through, one sees that whenever d_{01} is greater than 0.69, the $|t|$ statistic in that row is small, and when $d_{01} < 0.69$, the $|t|$ statistic in that row is large.

In contrast, differential risk factors cannot price four of the eleven common risk factors at posterior odds of 2:1. These are ME and EG in the Q5 model and PEAD and FIN in the DHS model. The rank risk factors perform even less well. These risk factors cannot price five of the 11 at posterior odds of 2:1, namely, SMB and HML in the FF6 model, ME and EG in the Q5 model, and PEAD in the DHS model. Thus, the evidence shows that slope risk factors outperform

differential and rank risk factors in pricing a large collection of test assets and pricing existing risk factors.

6 Importance of more pure play

In Section 3.1 we had shown that the short-cut approach of replacing risk factors in an existing model, say those in the FF6 model, is misleading about the relative worth of these three different factor construction methods. The correct way to see the performance of the slope factors is to compare the respective risk factors, as we did in the previous section.

Additionally, it is important to note that to realize the potential of slope factors, these factors must be constructed from cross-sectional regressions that control for a range of characteristics. If this is not done, the factors would be less pure play. For illustration, and to see how the SDF is based on less pure play slope factor prices, we construct less pure play slope factors in two different ways.

In the first way, for each standardized characteristic $c_j, j = 1,2,...,46$, other than mve, we run the cross-sectional regressions

$$r_{i,t} = \alpha_{j,t} + \beta_{1,j,t} c_{i,j,t} + \beta_{2,j,t} c_{i,j,t}^2 + \beta_{3,t} \text{mve}_{i,t} + \beta_{4,t} \text{mve}_{i,t}^2 + \varepsilon_{i,t}$$

and for the mve characteristic, we run the cross-sectional regression

$$r_{i,t} = \alpha_t + \beta_{1,t} \text{mve}_{i,t} + \beta_{2,t} \text{mve}_{i,t}^2 + \varepsilon_{i,t}$$

Then, the slope factor for c_j in month t is the average of the estimated $\beta_{1,j,t}$ and $\beta_{2,j,t}$, and the

slope factor for mve in month t is the average of the estimated $\beta_{1,t}$ and $\beta_{2,t}$. We denote the less pure play slope factors made this way as $\boldsymbol{x}_{\text{s.np1}}$, where the np label stands for non pure-play.

In a second way, we augment each of the above regressions with the five characteristics in the FF6 model: mve, bm, agr, operprof, and mom12m; each entered quadratically. Once again, the slope factor for a particular characteristic for a given month is the average of the estimated coefficients multiplying that characteristic's linear and quadratic terms. We denote the resulting less pure-play slope factors as $\boldsymbol{x}_{\text{s.np2}}$.

Note that, in general, less pure play slope factors can be constructed from a limited set of starting characteristics. However, less pure play slope factors trade off limited data requirements for performance.

Because more pure play and less pure play slope factors are different objects, the SDF based on less pure play risk factors will differ from the SDF we found above with more pure play risk factors. Specifically, after applying the PAMS method to the two sets of forty-seven less pure play slope factors as described above, $\boldsymbol{x}_{\text{s.np1}}$ and $\boldsymbol{x}_{\text{s.np2}}$, we find that the SDFs best supported by the data contain the factors

$$
\begin{aligned}
\boldsymbol{x}^{*}_{\text{s.np1}} =&(\text{Mkt, s.np1.acc, s.np1.cash, s.np1.chcsho, s.np1.lgr,}\\
&\text{s.np1.mom1m, s.np1.mve, s.np1.pricedelay, s.np1.sgr}),\\
\boldsymbol{x}^{*}_{\text{s.np2}} =&(\text{Mkt, s.np2.agr, s.np2.chcsho, s.np2.herf, s.np2.idiovol, s.np2.invest,}\\
&\text{s.np2.lev, s.np2.mom1m, s.np2.mve, s.np2.std\_dolvol, s.np2.tang}),
\end{aligned}
$$

respectively. Nine factors from the $\boldsymbol{x}_{\text{s.np1}}$ set are identified as risk factors while eleven factors from the $\boldsymbol{x}_{\text{s.np2}}$ set are identified as risk factors. It should be noted that this result does not imply that

these risk factors are capturing nine or eleven underlying risks. Because the characteristics are correlated, even after controlling for the FF6 characteristics, and these factors did not control for other characteristics, these selected characteristics each represent the risk emanating from other correlated characteristics, albeit coarsely. As a result, such factors cannot capture some risks, or others can only be captured incompletely. This weakness is manifested in degraded pricing performance.

Table 11 has the pricing results. One can see that at the default greater than 2:1 posterior odds threshold, the first set of less pure play slope risk factors $\boldsymbol{x}^*_{\text{s.np1}}$ price 682 out of 1150 portfolios, 921 out of 1480 ETFs, and 4551 out of 6024 individual stocks. Looking back at Table 9, these numbers are uniformly lower than the pricing numbers from the more pure play risk factors, $\boldsymbol{x}^*_{\text{s}}$, $\boldsymbol{x}^*_{\text{d}}$, and even the rank risk factors, $\boldsymbol{x}^*_{\text{r}}$. Meanwhile, the second set, $\boldsymbol{x}^*_{\text{s.np2}}$, prices 838 portfolios, 1096 ETFs, and 4715 stocks at the 2:1 posterior odds threshold. This improved performance shows that one can produce better slope factors even if those factors are constructed using a limited set of controls (here bm, agr, operprof and mom12m). The risk factors $\boldsymbol{x}^*_{\text{s.np2}}$ even outperform $\boldsymbol{x}^*_{\text{d}}$, but under perform the more pure play $\boldsymbol{x}^*_{\text{s}}$, as can be seen from Table 9. Thus, performance is closely connected to the more pure play property of slope factors.

7 Conclusion

In this paper, we have studied slope factors about differential/sorted and rank factors and reached several important conclusions for empirical asset pricing. First, we have shown that slope factors provide value when they are more pure play, i.e., when the slope factors are constructed from FM regressions that are broadly specified (in terms of having many RHS characteristics as controls). In this way, the resulting slope factors are purged of the effects of more characteristics.

Table 11 Less pure play slope risk factors: pricing performance on 1150 portfolios; 1480 ETFs and 6024 stocks. This table reports the number of assets that are priced at 0.75 threshold, representing (odds of 3:1 in favor), 0.667 (odds of 2:1 in favor), and 0.80 (odds of 4:1). There are nine non-pure-play slope risk factors $\boldsymbol{x}^*_{\text{s.np1}}$ and eleven non-pure-play slope risk factors $\boldsymbol{x}^*_{\text{s.np2}}$ discovered by the Bayesian PAMS methodology developed in the paper. Whether an asset is priced is determined by the log marginal likelihood differences of regressions with each test asset on the LHS and respective risk factors on the RHS, without and with an intercept, as explained in the text.

factor set	# priced at 2:1	# priced at 3:1	# priced at 4:1
		1150 Portfolios	
$\boldsymbol{x}^*_{\text{s.np1}}$	682	510	304
$\boldsymbol{x}^*_{\text{s.np2}}$	838	696	516
		1480 ETFs	
$\boldsymbol{x}^*_{\text{s.np1}}$	921	663	412
$\boldsymbol{x}^*_{\text{s.np2}}$	1096	845	589
		6024 Stocks	
$\boldsymbol{x}^*_{\text{s.np1}}$	4551	3225	2141
$\boldsymbol{x}^*_{\text{s.np2}}$	4715	3601	2564

Second, we show that it is not enough to take an existing asset pricing model, say the FF6, and replace its factors with slope factors to realize the value of slope factors. To realize the potential of slope factors, one should determine which slope factors from the starting pool of slope factors are in the SDF. To provide evidence of this, we have developed a new risk factor discovery methodology, PAMS, which is short for pruning, augmentation, and model scanning. The PAMS methodology is a general tool for risk factor discovery that can be used beyond the context of this paper.

Third, we show that the SDF best supported by the data varies by factor construction method. The PAMS methodology, applied to each set of factors, shows that the slope factor SDF has twenty risk factors, the differential factor SDF has fourteen risk factors, and the rank factor SDF has nineteen risk factors. Thus, the method of factor construction matters for which risks are captured. To our knowledge, this is the first time such a result has been documented in the literature. Thus, the factor construction method should not be an off-handed choice. It has far-reaching implications

for understanding the risks embedded in the cross-section of expected returns.

Fourth, on an extensive set of test assets consisting of 1150 portfolios, 1480 ETFs and (importantly) 6024 stocks, we have provided evidence that the slope factor SDF provides uniformly improved pricing of the cross section than the differential and rank factor SDFs. Since this finding is based on a broader and more representative collection of test assets than the norm, this result engenders confidence that the result is not an artifact of the test assets. We also document that slope risk factors outperform common (extant) risk factors in pricing. We relate this improved performance to the more pure play property and show that less pure play factors have degraded performance. These findings on pricing are a strong argument for adopting slope factors in empirical asset pricing.

Data, software, and code for reproducing the results in this paper are available on request.

Bibliography

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

Svetlana Bryzgalova, Jiantao Huang, and Christian Julliard. Bayesian solutions for the factor zoo: We just ran two quadrillion models. *The Journal of Finance*, 78:487–557, 2023.

Luyang Chammaen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science, in press*, 2022.

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

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.

Siddhartha Chib, Lingxiao Zhao, and Guofu Zhou. Winners from winners: A tale of risk factors. *Management Science, in press*, 2022.

John H Cochrane. *Asset pricing: Revised edition*. Princeton university press, 2009.

Kent Daniel, David Hirshleifer, and Lin Sun. Short- and long-horizon behavioral factors. *The Review of Financial Studies*, 33(4):1673–1736, 2020.

Eugene F Fama. *Foundations Of Finance*. 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. Choosing factors. *Journal of Financial Economics*, 128 (2):234–252, 2018.

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.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370, 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.

Edward Greenberg. *Introduction to Bayesian econometrics*. Cambridge University Press, 2012.

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, Haitao Mo, Chen Xue, and Lu Zhang. An augmented q-factor model with expected growth. *Review of Finance*, 25(1):1–41, 2021.

Soosung Hwang and Alexandre Rubesam. Bayesian selection of asset pricing factors using individual stocks. *Journal of Financial Econometrics*, 20(4):716–761, 2022.

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.