Slope Factors Outperform:

Evidence from a Large Comparative Study

Siddhartha Chib[1], **Yi Chun Lin**[2], **Kuntara Pukthuanthong**[3], **Xiaming Zeng**[4]

January 2023

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

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

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

[4]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 pure-play, ie., purged of the effects of a broad set of characteristics. Starting from data on forty seven characteristics, we show, by a new risk factor discovery method, that the SDF varies by factor construction class, and that this has implications for pricing of the cross-section. On a large collection of test assets, the slope factor based SDF uniformly prices more portfolios, ETFs and stocks. These findings support an increased 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, with relevance to the vast literature on pricing of the cross-section, is whether the method used to construct factors from firm-level characteristics has a bearing on the pricing performance of those factors. Recently, Fama and French (2020) show that *slope factors* constructed as the estimated OLS slopes from cross-sectional regressions of firm excess returns on standardized lagged characteristics provide better pricing of average returns on some test assets than the corresponding *differential factors* that are constructed from the difference in value-weighted returns of stocks in the extremes of 3 by 2 sorts, sorted by market capitalization and lagged characteristics. Our goal in this paper is to further study slope factors in order to understand the possible role that such factors can play in empirical asset pricing.

The first point revealed by our study is that it is important to construct slope factors from Fama-Macbeth (FM) regressions with a broad set of lagged (standardized) characteristics on the right hand side (RHS). In this way, the resulting slope factors are purged of the effects of more characteristics, and thus come closer to what might be called pure-play factors. In contrast, slope factors, which come from FM regressions with a limited set of lagged standardized characteristics on the RHS, are non pure-play. For example, the slope factors in Fama and French (2020), which were constructed from FM regressions with five characteristics, would be classified as non pure-play. We show that this is more than semantics. Pure-play slope factors provide improved pricing of the cross-section than non pure-play factors. In fact, it is possible to get even better pure-play slope factors (as measured by the metric of better pricing) by modeling each characteristic in FM regressions nonlinearly, for example, by a quadratic or cubic polynomial.

It is the pure-play property, and the flexibility one has in specifying the cross-sectional regressions, that makes slope factors such a useful device for compressing characteristics into factors. What our work shows is that if one intends to use slope factors, then it is necessary that those slope factors are constructed from FM regressions that are broadly specified (in terms of having many RHS characteristics as controls) and flexibly specified (in terms of nonlinearities).[1]

To reach these important conclusions, we focus on forty seven firm level characteristics 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. Although one could include more characteristics (to produce even purer slope factors), in practice it can lead to smaller sample sizes (because it becomes more difficult to find firms with complete data on all those characteristics) and multicollinearity, which can make least square unstable (the latter cannot be fixed by penalized least squares because the resulting slope estimates would not be long-short portfolios). We avoid these problems with our selection of forty seven broadly representative characteristics. Besides, this number of characteristics is already sufficiently large to make a clear distinction between pure-play and non pure-play factors, and to make the point about the importance of working with pure-play factors.

The second point revealed by our analysis is that to realize the value of slope factors it is not enough to take an existing asset pricing model, say the FF6, and replace its five characteristics based factors (SMB, HML, CMA, RMW, MOM) with the corresponding pure-play, or non pure-play slope factors. One could, of course, do this (as was done in Fama and French (2020)), but to realize the potential of slope factors one should determine which slope factors from the starting pool of slope factors are in the SDF.[2] To provide evidence on the latter point we develop

[1] An implication of the latter point is that even if one is interested in constructing a single slope factor, say a pollution factor, then it would be necessary to get this pollution slope factor from FM regressions that control for a broad set of other firm-level characteristics. It would not be enough to run FM regressions with the firm-level pollution characteristic on the RHS, along with few other firm level controls.

[2] 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).

a new Bayesian statistical method to determine which factors are in the SDF. This new method is underpinned by the model scanning methodology of Chib and Zeng (2020), with the priors of Chib, Zhao, and Zhou (2022). We refer to it as "pruning - augmentation - model scanning", or PAMS for short. To briefly summarize, in Step 1 of the method, 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 affirmatively satisfy this test, at a particular level of Bayesian posterior probability confidence, are not pruned. We then have additional steps in the method where the assessments of likely risk factors made in Step 1 are revised and updated. In particular, in Step 2, the augmentation step, we find which factors that were pruned in Step 1 cannot be priced by the factors that were not pruned in Step 1. 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 not priced factors from Step 2, are subjected to an complete model scan to remove false positives. This is done by estimating all possible models of the SDF with risk factors and non risk factors.[3] These models are compared by Bayesian marginal likelihoods. The risk factors in the best model (the model with the highest marginal likelihood) are then taken to be the best risk factors.

Any other risk factor discovery method, comparable to PAMS, that maintains the identity of the slope factors, could be used in place of PAMS. It would likely lead to the same, or roughly similar risk factors. We do not provide results for different methodologies in this paper because that would complicate the paper enormously. If the risk factors turn out to be different, then the methodologies themselves would have to be compared to determine which methodology does a

[3]This model enumeration and model estimation is not possible at the outset because the number of models in the model space with forty eight factors would be $2^{48} - 1$, which is prohibitively large. But, typically, at the end of the pruning and augmentation steps, about 20-25 factors are left, and model scanning is both feasible and effective.

better job in inferring the risk factors. This is rightfully a different project. Our focus here is narrower: to isolate the relative worth of slope factors, with the risk factors determined by one methodology. We leave the comparison of methodologies question for future work.[4]

Our third finding relates to the fact that the data evidence supports the conclusion that slope risk factors provide better pricing of the cross-section than risk factors from pools of factors constructed by the differential/sorted method, and by the *rank factor* method.[5] This simultaneous comparison of slope risk factors with differential and rank risk factors is the first of its kind. Factors in most studies are just constructed by one method. But, we show that the composition of the SDF changes depending on the factor construction method that is used, and that the pricing performance also depends on the factor construction method.

Specifically, starting from each set of three sets of forty eight factors (the Mkt factor plus the forty characteristics based factors constructed by the slope, differential and rank factor methods) we show that the slope factor set produces twenty risk factors; the differential factor set, fourteen; and the rank factor set, nineteen.[6] Thus, the SDF changes depending on the factor construction method. The evidence is also compelling that slope risk factors price more of the cross-section than differential and rank risk factors. We reach this conclusion by using each risk factor set to price a large set of test assets consisting of the excess returns of 1150 portfolios, 1480 ETFs and 6024 stocks. This is a larger set of test assets than is usually used in practice. On these test assets,

[4]Possible alternatives include Kozak, Nagel, and Santosh (2020), Hwang and Rubesam (2022) and Bryzgalova, Huang, and Julliard (2023). The goals of Feng et al. (2020) and Chen, Roussanov, and Wang (2021) are different. The former does not determine risk factors from a pool of factors; it uses a given set of risk factors to screen which of an additional set of factors could potentially be in the SDF. In the latter, FM slopes are used as response variables in semiparametric regressions. Thus, the estimated factors from this methodology are functions of the slope factors, and the identity of the slope factors is obscured in the estimated factors, unlike in the output from the PAMS approach.

[5]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 et al. (2020)

[6]The large number of risk factors that emerge in each case is consistent with the main thesis of Kozak and Nagel (2022) that the SDF should depend on a large number of factors.

we show that the slope risk factors outperform the differential and rank risk factors.

The rest 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 is relevant for our discussion. In Section 3 we discuss the Bayesian methodology for risk factor discovery that can be applied to our large pool of factors to infer the risk factors most supported by the data. Section 4 presents the risk factors we discover from applying our Bayesian inference procedure. In Section 5 we first detail a Bayesian pricing criteria to assess if a testing asset is priced, and detail the application of this criteria to determine pricing performance of the different risk factor collections on portfolios, ETFs and stocks. Section 6 provides some insights about 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 forty characteristics for the period January 1989 to December 2020.

For our analysis, we begin the sample from January 1989, which is the earliest month for which complete data on our selected characteristics are available in the I/B/E/S data set. Our aim is to be able to estimate cross-sectional regressions for firms that have a complete set of characteristics in the preceding cross-sections. One common approach is to only analyze firms with non-missing values of all characteristics in each cross-section. This approach, however, sacrifices a large portion of the data that may contain valuable information. Another approach is to replace the missing value with the mean of that characteristics across firms.

The latter approach is, however, 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 big, 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 $2 \times 10 = 20$ groups. Then, if any firm has 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 non-parametric 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 *pure-play* long-short portfolios. Specifically, these OLS coefficients give unit weighted exposure to each standardized lagged characteristic in the cross-section regression *and* zero weighted exposure to all the other standardized lagged characteristics. Thus, these OLS estimates are characteristic specific long-short portfolios that load entirely on that characteristic.

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 jth firm characteristic 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, $\boldsymbol{c}_{j,t-1} = (c_{j,1t-1},...,c_{j,n_t t-1})$, $j = 1,2,....,47$. Let $\tilde{\boldsymbol{c}}_{j,t-1}$ denote the characteristic after standardization, ie., after subtracting the sample mean and dividing by the sample standard deviation. In vector-matrix notation, the tth cross-sectional regression is given by

$$\boldsymbol{r}_t = \boldsymbol{X}_t \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \tag{1}$$

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 $\boldsymbol{\beta}_t$, namely $\hat{\boldsymbol{\beta}}_t = (\hat{\alpha}_t, \hat{\beta}_{1,t},...,\hat{\beta}_{47,t}) = (\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\boldsymbol{r}_t$, $t = 1,...,T$, 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.

Remark 1 In applying this approach, to capture potential non-linearities, for every characteristic, we also include on the RHS the square of the characteristic (standardized to have a mean of zero and 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 jth characteristic is computed by averaging the OLS estimates of the pair of linear and quadratic terms. Our experiments show that this approach, which, to our knowledge, has not be used before, 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 $\boldsymbol{W}_t' = (\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'$,

$$\hat{\beta}_t = \boldsymbol{W}_t'\boldsymbol{r}_t \tag{2}$$

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

$$\begin{pmatrix} \boldsymbol{w}_0' \\ \boldsymbol{w}_1' \\ \boldsymbol{w}_2' \end{pmatrix} (\boldsymbol{i}_{n_t}, \boldsymbol{c}_{1,t-1}, \boldsymbol{c}_{2,t-1}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3}$$

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

$$\begin{aligned}
&\boldsymbol{w}_0'\boldsymbol{i}_{n_t} = 1;\ \ \boldsymbol{w}_0'\boldsymbol{c}_{1,t-1} = 0;\ \ \boldsymbol{w}_0'\boldsymbol{c}_{2,t-1} = 0 \\
&\boldsymbol{w}_1'\boldsymbol{i}_{n_t} = 0;\ \ \boldsymbol{w}_1'\boldsymbol{c}_{1,t-1} = 1;\ \ \boldsymbol{w}_1'\boldsymbol{c}_{2,t-1} = 0 \\
&\boldsymbol{w}_2'\boldsymbol{i}_{n_t} = 0;\ \ \boldsymbol{w}_2'\boldsymbol{c}_{1,t-1} = 0;\ \ \boldsymbol{w}_2'\boldsymbol{c}_{2,t-1} = 1
\end{aligned} \tag{4}$$

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

We supplement these slope factors with the market portfolio from the 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

Table 2 continued: The descriptive statistics of the slope factors

	Mean	Median	Std	0.5% quantile	99.5% quantile
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
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 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. Let c_j denote the characteristic of interest other than *mve*. In each cross-section t, one divides the stock-premiums at time t into two groups, small and large, based on the median of market-capitalization, $mve_{i,t-1}$, $i \leq n_t$. Then, one further sort the stock premiums 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, $c_{j,it-1}$, $i \leq n_t$. Thus, with this double-sorting method, the stock-premiums are allocated to six buckets or, equivalently, an array containing 3 rows and 2 columns. A 3 by 2 array such as this is calculated for each characteristic, excluding the size characteristic.

Next, the excess return of these six buckets is value-weighted, ie., multiplied by its stock market cap divided by the total market cap in that bucket. Then, a long portfolio (a portfolio that goes long on that characteristic) is constructed as the sum of the value-weighted stock excess returns in the (3,1) and (3,2) buckets. Similarly, a short portfolio is constructed as the sum of the value-weighted stock-returns in the (1,1) and (1,2) buckets. Then, the differential factor for time period t is given by the difference 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 make 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 for $operprof$ and agr. The size factor for that cross-section t is then given by the average of these three long-short portfolios. The descriptive statistics of the forty seven 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. The 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
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

Table 3 continued: The descriptive statistics of the differential factors

	Mean	Median	Std	0.5% quantile	99.5% quantile
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 stock-premiums at time t into two groups, small and large, based on the median of market-capitalization, $mve_{i,t-1}$, $i \leq n_t$. Let $I_{t0} = \{i : \text{firm } i \text{ is a small firm}\}$ and let $I_{t1} = \{i : \text{firm } i \text{ is a large firm}\}$ denote the indices of small firms and large firms at time t. Let the number of firms in each group be n_{t0} and n_{t1}, respectively. Now for each characteristic c_j, let $c_{j,0,t-1} = \{c_{j,it-1} : i \in I_{t0}\}$ be the vector of characteristics of length n_{t0} at time $(t-1)$ of all small firms, and similarly let $c_{j,1,t-1} = \{c_{j,it-1} : i \in I_{t1}\}$ be the vector of characteristics of length n_{t1} at time $(t-1)$ of all large firms. Now let $ra_{j,0,t-1}$ denote the vector of ranks of the values in $c_{j,0,t-1}$, and let $r_{j,0,t-1} = \frac{ra_{j,0,t-1}}{n_{t0}+1}$ denote the normalized ranks. Likewise, let $ra_{j,1,t-1}$ denote the vector of ranks of the values in $c_{j,1,t-1}$, and let $r_{j,1,t-1} = \frac{ra_{j,1,t-1}}{n_{t1}+1}$. Further, let the sample mean of the values in $r_{j,0,t-1}$ be denoted by $\bar{r}_{j,0,t-1}$ and similarly let $\bar{r}_{j,1,t-1}$ denote the sample mean of the values in $r_{j,1,t-1}$. Now define the vectors of weights

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

which each sum to zero.

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

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

where $\cdot$ is the dot-product operator for multiplying two vectors. The descriptive statistics of the forty seven 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. The 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
r.acc	-0.378	-0.258	1.967	-7.135	3.716
r.age	-0.105	0.055	2.726	-10.976	7.405
r.agr	-0.714	-0.530	2.226	-9.669	4.200
r.baspread	0.235	-0.047	6.626	-16.581	25.781
r.beta	0.081	-0.118	5.838	-14.819	23.162
r.bm	0.434	0.346	2.790	-8.458	8.502
r.cash	0.422	0.430	3.735	-11.195	14.646
r.cashdebt	-0.030	0.322	3.396	-15.876	8.055
r.cashpr	-0.199	-0.085	2.918	-8.186	9.847
r.cfp	0.290	0.392	4.090	-14.188	13.326
r.chatoia	0.139	0.163	0.984	-2.371	2.650
r.chcsho	-0.439	-0.376	2.537	-8.572	6.685
r.chempia	-0.290	-0.208	1.565	-5.470	3.565
r.depr	0.397	0.375	3.272	-8.936	13.614
r.dy	-0.332	-0.540	3.313	-10.173	8.850
r.egr	-0.439	-0.178	2.227	-10.055	5.071
r.ep	-0.059	0.047	4.575	-16.842	12.162
r.gma	0.179	0.291	2.285	-4.794	5.730
r.grcapx	-0.356	-0.385	1.594	-4.838	3.982
r.herf	-0.086	-0.087	2.216	-5.838	6.493
r.hire	-0.378	-0.250	1.947	-5.925	5.140
r.idiovol	0.298	-0.117	6.545	-16.38	27.745
r.ill	0.342	0.039	3.744	-8.430	14.625
r.indmom	0.560	0.639	3.869	-12.145	14.535
r.invest	-0.458	-0.328	1.828	-5.182	4.682
r.lev	0.085	0.011	3.970	-14.065	10.533
r.lgr	-0.436	-0.459	1.358	-4.027	2.899
r.mom12m	0.338	0.790	4.991	-20.868	11.333

Table 4 continued: The descriptive statistics of the rank factors

	Mean	Median	Std	0.5% quantile	99.5% quantile
r.mom1m	-0.838	-0.378	4.321	-16.690	10.584
r.mve	-0.536	-0.123	4.508	-19.221	9.868
r.nincr	0.287	0.343	1.207	-4.145	3.306
r.operprof	0.143	0.290	1.958	-7.497	4.360
r.pchgm_pchsale	0.147	0.372	1.526	-4.959	3.455
r.pricedelay	0.084	0.003	1.739	-5.117	5.071
r.ps	0.057	0.325	2.612	-10.255	5.962
r.roaq	0.202	0.616	4.020	-15.635	9.586
r.roeq	0.177	0.356	3.999	-16.539	10.516
r.roic	-0.046	0.057	3.636	-16.193	9.472
r.salecash	-0.096	0.164	2.813	-8.062	8.034
r.saleinv	-0.001	0.062	1.622	-5.766	4.311
r.salerec	0.087	0.052	1.762	-4.471	6.431
r.sgr	-0.444	-0.369	1.927	-5.787	4.566
r.sp	0.413	0.498	3.182	-9.004	11.468
r.std_dolvol	0.352	0.223	2.785	-8.089	8.834
r.std_turn	0.365	0.128	4.411	-11.321	18.151
r.tang	0.321	0.219	2.746	-6.846	11.271
r.tb	0.037	0.185	2.426	-10.72	7.222

2.5 Are the factors the same?

It is well enough that we have these three different ways of computing factors from characteristics, but, what are the properties of these factors? We will get into the pricing capabilities below. For now, we make the point that the differential and rank factors produce factors that are not that similar to slope factors. In Table 5, we give the pair-wise 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 r.acc). One can see from the entries in the first two columns of this table that the correlation between the slope and differential and rank factors tends

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 r methods; corrsd is the pairwise correlation between the s and d factors; corrsr is the pairwise correlation between the s and r factors; and corrdr is the correlation between the d and r factors.

factor	corrsd	corrsr	corrdr
f.acc	0.248	0.266	0.831
f.age	0.452	0.501	0.826
f.agr	0.093	0.208	0.722
f.baspread	0.553	0.662	0.903
f.beta	0.809	0.819	0.978
f.bm	0.153	0.350	0.860
f.cash	0.133	0.143	0.957
f.cashdebt	-0.007	0.020	0.758
f.cashpr	0.078	0.132	0.915
f.cfp	0.126	0.175	0.907
f.chatoia	0.131	0.337	0.630
f.chcsho	0.066	0.130	0.912
f.chempia	0.222	0.228	0.630
f.depr	0.078	0.170	0.925
f.dy	0.110	0.114	0.911
f.egr	0.081	0.084	0.714
f.ep	0.165	0.211	0.901
f.gma	0.318	0.458	0.821
f.grcapx	0.036	0.122	0.790
f.herf	0.239	0.319	0.819
f.hire	0.201	0.240	0.783
f.idiovol	0.377	0.559	0.904
f.ill	0.072	0.202	0.270
f.indmom	0.289	0.223	0.952
f.invest	0.303	0.359	0.733
f.lev	0.418	0.432	0.958
f.lgr	0.003	0.140	0.732
f.mom12m	0.291	0.364	0.911
f.mom1m	0.371	0.477	0.906
f.mve	-0.412	-0.024	-0.764
f.nincr	0.364	0.467	0.730
f.operprof	0.058	0.149	0.789
f.pchgm_pchsale	0.131	0.278	0.736
f.pricedelay	0.080	0.161	0.745
f.ps	0.240	0.259	0.758
f.roaq	0.345	0.344	0.855
f.roeq	-0.024	-0.003	0.863
f.roic	0.137	0.184	0.817
f.salecash	-0.047	-0.089	0.910
f.saleinv	0.329	0.394	0.822
f.salerec	0.014	0.059	0.786
f.sgr	0.135	0.207	0.822
f.sp	0.285	0.311	0.917
f.std_dolvol	0.478	0.514	0.476
f.std_turn	0.288	0.336	0.948
f.tang	0.183	0.223	0.901
f.tb	0.123	0.116	0.837

to be weak.

One can ask why the differential and rank factors are this different from slope factors. After all, the former two construction methods also start from the same data and each differential and rank factor is supposed to connect to the underlying characteristic. The connection is, however, not direct because the characteristics themselves tend to be correlated. Stated another way, slope factors are pure-play (ie., purged of the effects of other characteristics) whereas the differential and rank factors only control for size. There would be essentially little difference between the factors from the three construction methods if the characteristics were not correlated after controlling for size. But, this is not the case. In Figure 1 we plot the pairwise (partial) correlations between every characteristic (other than *mve*), controlling for size. These partial correlations are computed from the residuals of regression of each characteristic on *mve*. These residuals are calculated cross-section by cross-section and then the correlations are calculated after pooling the residuals across time. This plot shows that even after controlling for size (*mve*), the characteristics tend to be correlated with several other characteristics. Thus, the differential and rank factors that control for just size incorporate the influence of these other characteristics, not just the risk emanating from that characteristic.

3 Methodology

3.1 Motivation

Intuitively, it seems reasonable to believe that slope factors, due to their pure-play property, should provide superior rendering of the underlying risk emanating from the characteristics, and, therefore, improved pricing of the cross-section. But to confirm such a conjecture it is not enough

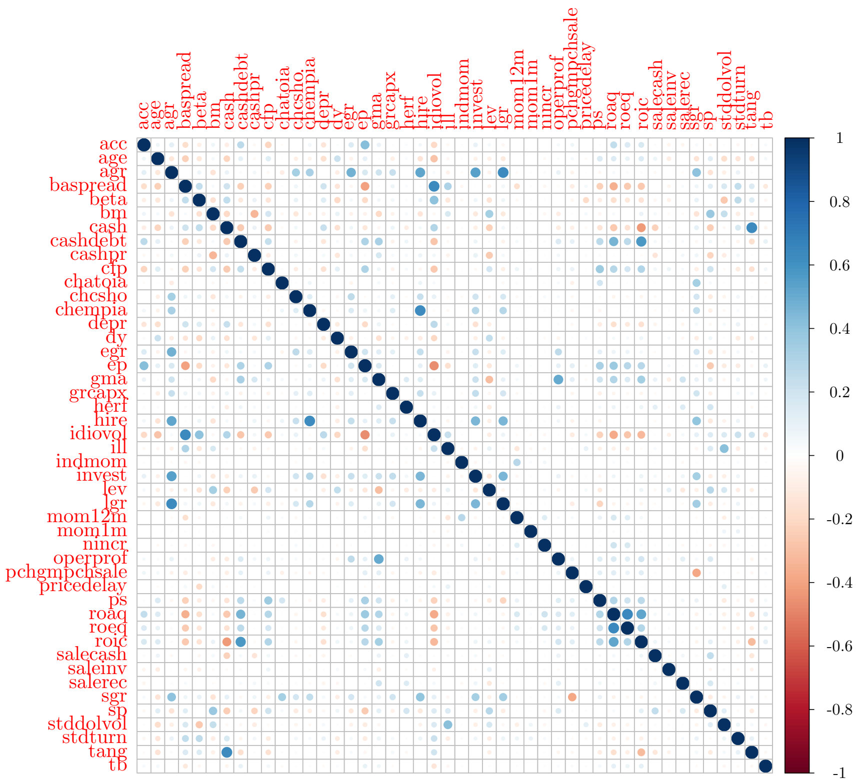

Figure 1 Pairwise partial correlations between the 46 characteristics in Table 1, controlling for size (*mve*).

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 f.mve, f,bm, f.agr, f.operprof, f.mom12m, factors, where f $\in$ {s,r}. This is because if we start from (say) the starting set of forty seven slope factors, the best risk factors are not necessarily s.mve, s.bm, s.agr, s.operprof and s.mom12m. Thus, replacing the FF6 by the corresponding s and r factors is certainly possible, but this need not give 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.[7] The table gives the number of test assets 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 on portfolios, that d.ff6 is best for the ETF's, and r.ff6 dominates on stocks.

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 ones favored construction method. As we now show, it is a mistake to simply replace one set of factors with another set of factors in an existing asset pricing model. This is because the different factor construction methods carry different information about the underlying characteristics and that those differences result in different SDFs (equivalently, different risk factors). Thus, it is important to infer the risk factors within each class of factors and then to

[7] Test assets are 1150 portfolios from 5×5 sorts on size (*mve*) and 46 characteristics from our sample data; 1480 ETFs obtained from CRSP (sharecode 73) that have at least 60 months of observation between January 1989 - December 2020; 6024 common stocks obtained from CRSP (sharecode 10 and 11) that have 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.

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 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). Factors consist of the Mkt and those constructed from *mve*, *bm*, *agr*, *operprof*, and *mom*12*m*. In the case of the slope factors, the factors control for for 42 other characteristics. The pricing is based on log marginal likelihood differences of regressions with each test asset on 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 does better pricing it is necessary to find the best risk factors in each class, and to compare the pricing performance of those 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
r.ff6	735	592	393
		1480 ETFs	
s.ff6	693	465	280
d.ff6	889	589	366
r.ff6	821	546	342
		6024 Stocks	
s.ff6	4342	3009	1910
d.ff6	4363	3020	1930
r.ff6	4503	3263	2175

compare the pricing performances of those different sets of risk factors.

3.2 Risk factor discovery

The problem we now discuss is how to determine which factors in each set of factors are in the SDF, ie., which factors are risk factors within each set of factors,

$$\boldsymbol{f}_{\mathrm{f},t} = \left(\mathrm{Mkt}_t, \mathrm{f.acc}_t, \mathrm{f.age}_t, \mathrm{f.agr}_t, \ldots, \mathrm{f.tb}_t\right) \ , \ t = 1, 2, \ldots, T$$

where $\mathrm{f} \in \{\mathrm{s}, \mathrm{d}, \mathrm{r}\}$ denotes the different factor construction methods (we suppress this in the notation below for convenience), t is the month and T the size of the time-series sample. Following Chib and Zeng (2020), one way to find the risk factors is to allow each factor to play the role of a risk factor i.e., an element of the SDF, or a non-risk factor, to produce different groupings of risk factors in a combinatorial fashion. Starting with the 48 initial possible risk factors, there are, therefore, $J = 2^{48} - 1$ possible risk factor combinations (assuming that the risk factor set cannot be empty). Each of these risk factor combinations defines a particular asset pricing model $\mathbb{M}_j$, $j = 1, \ldots, J$.

Consider now a specific model $\mathbb{M}_j$, $j = 1, \ldots, J$ consisting of the risk factors $\boldsymbol{x}_{j,t} : k_{x,j} \times 1$, and the complementary set of factors (the non-risk factors) $\boldsymbol{w}_{j,t} : k_{w,j} \times 1$. By definition, factors are risk factors if they are in the stochastic discount factor $M_{j,t}$. Following Hansen and Jagannathan (1991), let the SDF be

$$M_{j,t} = 1 - \boldsymbol{\lambda}'_{x,j} \Omega_{x,j}^{-1} \left(\boldsymbol{x}_{j,t} - \mathbb{E}[\boldsymbol{x}_{j,t}] \right), \tag{5}$$

where $\boldsymbol{b}_{x,j} \triangleq \Omega_{x,j}^{-1} \boldsymbol{\lambda}_{x,j} : k_{x,j} \times 1$ are the unknown risk factor loadings (the market prices of factor risks), and $\Omega_{x,j} : k_{x,j} \times k_{x,j}$ is the covariance matrix of $\boldsymbol{x}_j$. Enforcing the pricing restrictions implied by the no-arbitrage condition

$$\mathbb{E}[M_{j,t} \boldsymbol{x}'_{j,t}] = \mathbf{0} \quad \text{and} \quad \mathbb{E}[M_{j,t} \boldsymbol{w}'_{j,t}] = \mathbf{0}$$

for all t, we get that

$$\mathbb{E}[\boldsymbol{x}_{j,t}] = \boldsymbol{\lambda}_{x,j} \quad \text{and} \quad \mathbb{E}[\boldsymbol{w}_{j,t}] = \Gamma_j \boldsymbol{\lambda}_{x,j},$$

for some matrix $\Gamma_j : k_{w,j} \times k_{x,j}$.

If we assume that the joint distribution of $(\boldsymbol{x}_j, \boldsymbol{w}_j)$ is Gaussian (we suppress the t subscript for convenience), then the latter pricing restrictions imply that under a marginal-conditional decomposition of the factors, $\mathbb{M}_j$ has the restricted reduced form given by

$$\boldsymbol{x}_j = \boldsymbol{\lambda}_{x,j} + \boldsymbol{\varepsilon}_{x,j}, \tag{6}$$

$$\boldsymbol{w}_j = \Gamma_j \boldsymbol{x}_j + \boldsymbol{\varepsilon}_{w\cdot x,j}, \tag{7}$$

where the errors are block independent Gaussian

$$\begin{pmatrix} \boldsymbol{\varepsilon}_{x,j} \\ \boldsymbol{\varepsilon}_{w\cdot x,j} \end{pmatrix} \sim \mathbb{N}_K \left(\mathbf{0}, \begin{pmatrix} \Omega_{x,j} & \mathbf{0} \\ \mathbf{0} & \Omega_{w\cdot x,j} \end{pmatrix} \right), \tag{8}$$

and $\Omega_{w\cdot x,j} : k_{w,j} \times k_{w,j}$ is the covariance matrix of the conditional residuals $\boldsymbol{\varepsilon}_{w\cdot x,j}$.

We take a Bayesian approach to estimating each of these models. Let

$$\boldsymbol{\theta}_j = (\boldsymbol{\lambda}_{x,j}, \Omega_{x,j}, \Gamma_j, \Omega_{w\cdot x,j})$$

denote the parameters of $\mathbb{M}_j$. For doing Bayesian inference across models, we adopt the priors given in Chib, Zeng, and Zhao (2020), which are a special case of the priors of Chib and Zeng (2020). Suppose that K is the total number of factors. Then, the priors across models take the form

$$\pi(\boldsymbol{\theta}_j|\mathbb{M}_j) = \pi(\Omega_{x,j}, \Gamma_j, \Omega_{w\cdot x,j}|\mathbb{M}_j)\pi(\boldsymbol{\lambda}_{x,j}|\mathbb{M}_j, \Omega_{x,j}, \Gamma_j, \Omega_{w\cdot x,j}) \tag{9}$$

where

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

and

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

where $\mathbb{N}_d(\cdot|\mu,\Omega)$ is the d-dimensional multivariate normal density function with mean μ and covariance matrix Ω. The constant c in these priors is irrelevant because it is the same constant in all the priors and cancels in taking log marginal likelihood differences across models.

Also note that there is one model, let us call it the full model and give it the label $j=1$, in which all factors are in $\boldsymbol{x}$; thus $\boldsymbol{w}$ is empty. In that case the marginal prior $\pi(\Omega_{x,1},\Gamma_1,\Omega_{w\cdot x,1}|\mathbb{M}_1)$ reduces to $\pi(\Omega_{x,1}|\mathbb{M}_1) = c|\Omega_{x,1}|^{-\frac{K+1}{2}}$.

The key quantities for comparing these J models are the marginal likelihoods, defined as

$$m_j(\boldsymbol{f}_{1:T}|\mathbb{M}_j) = \int_{\Theta_j} p(\boldsymbol{y}_{1:T}|\mathbb{M}_j,\boldsymbol{\theta}_j)\pi(\lambda_{x,j}|\mathbb{M}_j,\boldsymbol{\eta}_j)\psi(\boldsymbol{\eta}_j|\mathbb{M}_j)\mathrm{d}\boldsymbol{\theta}_j, \tag{10}$$

where Θ_j is the domain of the parameters $\boldsymbol{\theta}_j$ and $\boldsymbol{f}_{1:T}$ is the data on the factors. Therefore, under the assumption of independence across time, we have that

$$p(\boldsymbol{f}_{1:T}|\mathbb{M}_j,\boldsymbol{\theta}_j) = \prod_{t=1}^{T}\mathbb{N}_{k_{x,j}}(\boldsymbol{x}_{j,t}|\lambda_{x,j},\Omega_{x,j})\mathbb{N}_{k_{w,j}}(\boldsymbol{w}_{j,t}|\Gamma_j\boldsymbol{x}_{j,t},\Omega_{w\cdot x,j})$$

is the density of the data.

We make extensive use of the log of these marginal likelihoods. Based on the method of Chib (1995), Chib et al. (2020) derive the following closed-form expressions of the *log* marginal likelihoods under the above priors. Specifically, they show that

$$\log m_1(\boldsymbol{f}_{1:T}|\mathbb{M}_1) = -\frac{TK}{2}\log\pi - \frac{K}{2}\log(T\kappa_1+1) - \frac{T}{2}\log|\Psi_1| + \log\Gamma_K\left(\frac{T}{2}\right), \tag{11}$$

and

$$\log m_j(\boldsymbol{y}_{1:T}|\mathbb{M}_j) = -\frac{Tk_{x,j}}{2}\log\pi - \frac{k_{x,j}}{2}\log(T\kappa_j+1) - \frac{(T+k_{x,j}-K)}{2}\log|\Psi_j| + \log\Gamma_{k_{x,j}}\left(\frac{T+k_{x,j}-K}{2}\right)$$
$$-\frac{(K-k_{x,j})(T-k_{x,j})}{2}\log\pi - \frac{(K-k_{x,j})}{2}\log|W_j^*| - \frac{T}{2}\log|\Psi_j^*| + \log\Gamma_{K-L_j}\left(\frac{T}{2}\right),\ j>1, \tag{12}$$

where we have set $c = 1$ (as this choice is irrelevant, by construction), and

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

Note that the variables in 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.

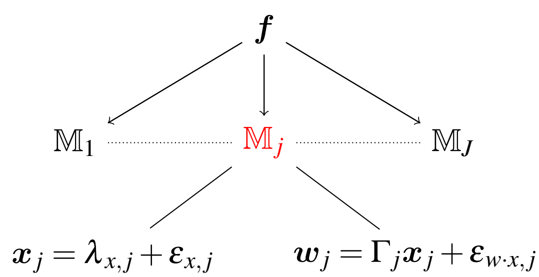

Figure 2 Model scanning (summary): The starting set of K factors imply $J = 2^K - 1$ different asset pricing models $\mathbb{M}_1,...,\mathbb{M}_J$, defined by different splits of the factors into risk factors $\boldsymbol{x}_j$ that enter the SDF and non risk factors $\boldsymbol{w}_j$, which are priced by $\boldsymbol{x}_j$. Under the assumption that the initial distribution of the factors is Gaussian, the errors in $\mathbb{M}_j$ are Gaussian and block independent. In model scanning each of these models is estimated and compared by Bayesian marginal likelihoods. The risk factors in the model with the largest marginal likelihood are the risk factors best supported by the evidence.

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

$$\Pr(\mathbb{M}_j) = 1/J \tag{13}$$

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

$$\Pr(\mathbb{M}_j|\boldsymbol{f}_{1:T}) = \frac{1}{1+\sum\limits_{l=1,l\neq j}^{J} \exp\left(-\left(\log m_j(\boldsymbol{f}_{1:T}|\mathbb{M}_j) - \log m_l(\boldsymbol{f}_{1:T}|\mathbb{M}_l)\right)\right)} \tag{14}$$

In the present context, however, we cannot apply this model scanning approach directly because J is too large. But, if J is say 25 or less, then there are at most 33.55443 million models, which can be estimated relatively quickly. For the present case, where $J = 48$, we discover our new asset pricing model by a new 3-step methodology in which model scanning plays a key role. We refer to this 3-step approach as "pruning - augmentation - model scanning", or PAMS for short.

3.3 PAMS: Pruning - augmentation - model scanning

To briefly summarize (more details are given in the following subsections), in the pruning step we apply a method to prune the set of factors to determine an initial set of risk factors that we call $\boldsymbol{x}^1$ (where the superscript 1 refers to Step 1). Then in the augmentation step we consider the remaining set of factors, ie., $\boldsymbol{f}\backslash\boldsymbol{x}^1$ (where the backslash is the set-difference operator) and see which of these factors cannot be priced by $\boldsymbol{x}^1$. We denote this set of non-priced remaining factors by $\boldsymbol{w}_{\text{np}}^2$ (where the superscript 2 refers to an operation carried out in Step 2). We let $\boldsymbol{f}^2 = (\boldsymbol{x}^1, \boldsymbol{w}_{\text{np}}^2)$ denote this set of factors. These constitute the set of likely potential risk factors. If the number of factors in $\boldsymbol{f}^2$ is greater than 25, we apply another pruning step to these factors to reduce the number of factors to 25 or below. We then take these factors $\boldsymbol{f}^2$ to Step 3 where we apply the model scanning method described above to estimate all possible models of risk factors and non risk factors. The risk factors in the best model (as measured by model marginal likelihoods) from model scanning are then taken to be the best risk factors.

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 describe this soft pruning step, consider for specificity, the factor Mkt. The way we decide if this factor is a possible risk factor (or whether it should be pruned) is to ask the following question:

> Is the posterior probability that Mkt is a risk factor, given that the remaining factors are risk factors, at least 0.75?

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. Factors that affirmatively satisfy this test are, therefore, not pruned. 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). Even though there are additional steps where 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 the factors that are less likely to be risk factors, not to isolate the risk factors (the latter is done in Step 3). For this reasons, the threshold probability of 0.75 serves its purpose admirably.

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 odds of 3:1 in favor; other threshold are 0.667 (odds of 2:1 in favor), and 0.80 (odds of 4:1). We use the language of odds as well in our discussion below.

We can obtain the answer to our question by simply comparing two models, each of which is a special case of the $\boldsymbol{x}_j = \lambda_{x,j} + \varepsilon_{x,t}$ and $\boldsymbol{w}_j = \Gamma_j \boldsymbol{x}_j + \varepsilon_{w \cdot x,j}$ formulation given above. In the first model Mkt is a risk factor and all other factors are risk factors (this is the full model mentioned in the previous section). In the second model, the risk factors are all the factors in $\boldsymbol{f}_t$ other than Mkt, and Mkt is the single element in $\boldsymbol{w}$. Specifically, the first model is

$$\boldsymbol{f}_t = \lambda + \varepsilon_t, \tag{15}$$

since $\boldsymbol{x}_t = \boldsymbol{f}_t$. To define the second model let $\boldsymbol{x}_2$ denote the factors $\boldsymbol{f}_t \backslash$Mkt and let w_2 denote the single factor Mkt. Then, the second model is

$$\boldsymbol{x}_{2,t} = \lambda_{x,2} + \varepsilon_{x,2,t}, \tag{16}$$

$$w_{2,t} = \Gamma_2 \boldsymbol{x}_{2,t} + \varepsilon_{w \cdot x,2,t} \tag{17}$$

Let the marginal likelihood of the first model (computed under the priors and formula given above) be m_1 and the marginal likelihood of the second model (again computed under the priors and formula given above) be m_2. Then, by Bayes theorem,

$$\Pr(\text{Mkt is a risk factor} | \boldsymbol{f}_{1:T}, \text{all other factors are risk factors}) = \frac{1}{1 + \exp(-(\log m_1 - \log m_2))}$$

If this posterior probability is at least 0.75, we put this factor in $\boldsymbol{x}^1$. Otherwise, we omit this factor from $\boldsymbol{x}^1$. We then move to the next factor and ask the same question, regardless of the answer to the questions regarding other factors. For example, to judge if f.age should be pruned in Step 1, we compute

$$\Pr(\text{f.age is a risk factor} | \boldsymbol{f}_{1:T}, \text{all other factors are risk factors})$$

Once again, the first model is the same as the full model (all factors are risk factors). In the second model, $\boldsymbol{x}_2$ now denotes the factors $\boldsymbol{f}_t \backslash$f.age, and $\boldsymbol{w}_2$ denote the single factor f.age. The posterior probability is again given by the same expression with m_2 now the marginal likelihood of the new second model.

This (soft) pruning step is new and from our experiments is very effective in pruning out factors that are unlikely to be risk factors. But, this step is not final. Our 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).

Step 2: Augmentation

The idea behind Step 2 is to catch any false negatives from Step 1. Specifically, given the set of factors in $\boldsymbol{x}^1$ we look at the remaining factors in $\boldsymbol{f}$. 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, in order to find these false negatives, we ask a fresh question:

> Which factors in $\boldsymbol{w}^1$ are not priced by $\boldsymbol{x}^1$ with 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.[8] 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:

$$\begin{gathered}\text{Observation eq: } w_{j,t} = \beta_{0,j}' \boldsymbol{x}_t^1 + \varepsilon_{0,j,t} \ , \ \varepsilon_{0,j,t} \sim \mathbb{N}(0, \sigma_{0,j}^2) \ , t \leq T \\ \text{Prior: } \beta_{0,j} \sim \mathbb{N}(\boldsymbol{b}_{0,j}, \boldsymbol{B}_{0,j}) \ , \ \sigma_{0,j}^2 \sim \mathbb{IG}\left(\frac{\nu_{0,j}}{2}, \frac{\delta_{0,j}}{2}\right)\end{gathered} \tag{18}$$

and

$$\begin{gathered}\text{Observation eq: } w_{j,t} = \alpha_j + \beta_{1,j}' \boldsymbol{x}_t^1 + \varepsilon_{1,t} \ , \ \ \varepsilon_{1,j,t} \sim \mathbb{N}(0, \sigma_{1,j}^2) \ , t \leq T \\ \text{Prior: } \left(\alpha_j, \beta_{1,j}\right) \sim \mathbb{N}(\boldsymbol{b}_{1,j}, \boldsymbol{B}_{1,j}) \ , \ \sigma_{1,j}^2 \sim \mathbb{IG}\left(\frac{\nu_{1,j}}{2}, \frac{\delta_{1,j}}{2}\right)\end{gathered} \tag{19}$$

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 by MCMC methods and calculate the marginal likelihood of each model in this comparison by 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

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

[8] In a different context, and with a different aim, Chib et al. (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 (ie. posterior odds of not being priced is 3:1), we put w_j in the set $\boldsymbol{w}^2_{\text{np}}$ (where the superscript 2 stands for Step 2 and np for not-priced).

We assemble the factors $\boldsymbol{x}^1$ and $\boldsymbol{w}^2_{\text{np}}$ in the set $\boldsymbol{f}^2$. The idea now is to find the risk factors from within this augmented set by full model scanning. This is done in Step 3.

Step 3: Model scanning

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

This exhaustive examination is conducted by the model scan method applied to the factors $\boldsymbol{f}^2$. Specifically, the jth model in this scan is defined as above by

$$\boldsymbol{x}_{j,t} = \lambda_{x,j} + \varepsilon_{x,j,t}, \tag{20}$$

$$\boldsymbol{w}_{j,t} = \Gamma_j \boldsymbol{x}_{j,t} + \varepsilon_{w \cdot x,j,t}, \tag{21}$$

where the errors are block independent Gaussian

$$\begin{pmatrix} \varepsilon_{x,j,t} \\ \varepsilon_{w \cdot x,j,t} \end{pmatrix} \sim \mathbb{N}_K \left(\mathbf{0}, \begin{pmatrix} \Omega_{x,j} & \mathbf{0} \\ \mathbf{0} & \Omega_{w \cdot x,j} \end{pmatrix} \right), \tag{22}$$

with the difference that $\boldsymbol{x}_{j,t}$ and $\boldsymbol{w}_{j,t}$ are splits formed from the factors in $\boldsymbol{f}^2$. Since the number of factors in $\boldsymbol{f}^2$ is typically much smaller than K (the number of initial factors), the number of models that are estimated and compared in this scan is substantially smaller than J.

We rank the models in this scan by log marginal likelihoods. 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 taken to be best risk factors.

4 Risk factor discovery: Evidence

We now take our PAMS risk factor discovery method to the data. It should be kept in mind that we are trying to determine the composition of the SDF, and how that composition varies across the three different sets of factors.

Before we apply the procedure, however, it is useful to think about what the data are likely to say. Since the factors are cross-section summaries of how returns depend on characteristics it is 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 3, we plot the number of significant characteristics (including the constant) in month-by-month regressions of firm level excess returns on lagged standardized 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 about that many risk factors, and that these would be the ones that correspond to the significant characteristics.

For the data at hand, however, given the variability across months, it is difficult to ascertain from such a calculation which factors, and how many, would end up being risk factors. That

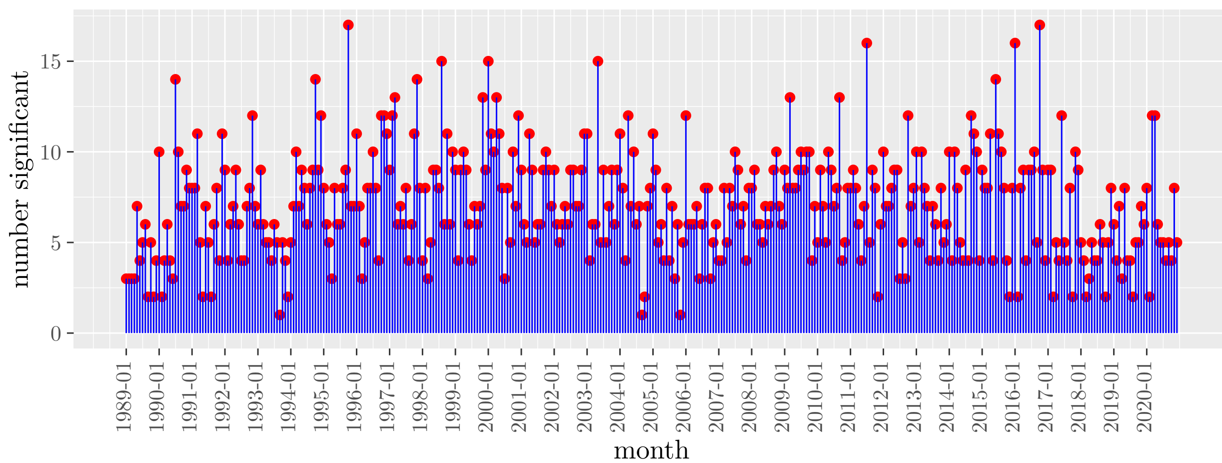

Figure 3 Number of significant characteristics (|t-ratio| $>$ 2.5) in month-by-month cross-sectional regressions of firm level excess returns on lagged standardized characteristics.

connection is even more strained for differential and rank factors because those factors capture the influence of several characteristics at once, as discussed above. Nonetheless, 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

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

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 cut-off for determining $\boldsymbol{x}_1$, the factors that are not pruned. The highly optimized computations coded in C++ take 14.23 seconds on a Macbook Pro computer with a M1 Max chip.

We give the results in Figure 4. It is clear from this plot that the probability cut-point of 0.75

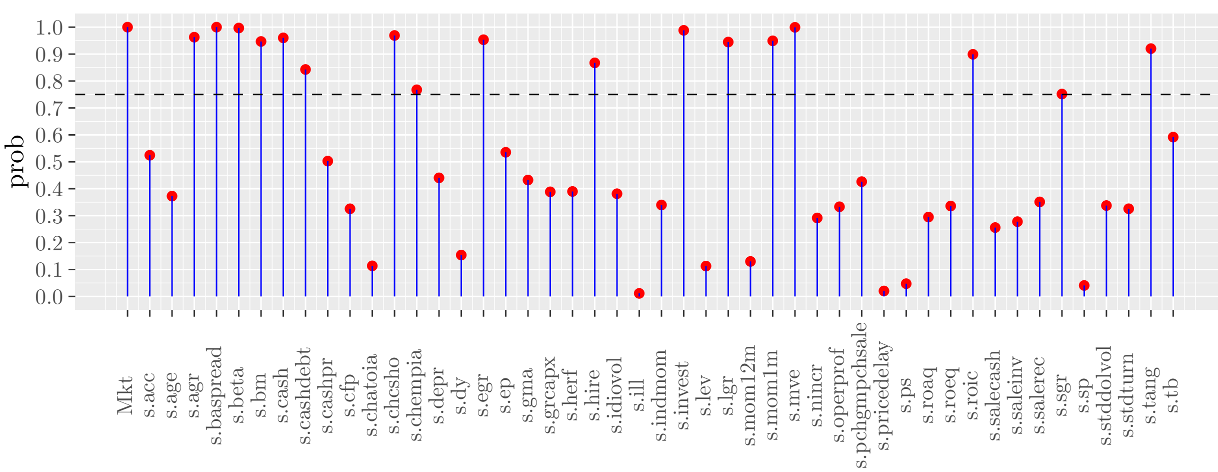

Figure 4 Step 1: Posterior probability that s.j (as j runs through the 48 factors) is a risk factor, given the data and all other factors are risk factors. The horizontal line is at the probability cut-value of 0.75. Factors that have a posterior probability exceeding this threshold are not pruned.

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

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

have 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 thirty slope factors, the posterior probability of being a risk factor, given the rest are risk factors, is far below 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 4, the composition of $\boldsymbol{x}_{\mathrm{s}}^{1}$ is robust to a reduction in the threshold to, say, 0.67 (for 2:1 odds). Going in the other direction, however, would reduce the cardinality of $\boldsymbol{x}_{\mathrm{s}}^{1}$, and this would not be desirable since potential risk factors would likely be pruned. 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 $\boldsymbol{x}_{\mathrm{s}}^{1}$ used in that test came in misspecified. Our experiments conducted over many simulated data sets confirm this guidance. The probability threshold can be reduced from 0.75, but should not generally be adjusted upwards.

4.2 Step 2

The next step in the discovery procedure is to examine the thirty pruned slope factors from Step 1 to determine which of these thirty can, or cannot, be priced by $\boldsymbol{x}_{\mathrm{s}}^{1}$. We have implemented this pricing test in another R function that is called as `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 $\boldsymbol{x}_{\mathrm{s}}^{1}$ 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 $\boldsymbol{x}_{\mathrm{s}}^1$ 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

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

are the six out of thirty factors that are not priced by $\boldsymbol{x}_{\mathrm{s}}^1$. Therefore, at the end of the two steps our set of potential risk factors consists of $\boldsymbol{x}_{\mathrm{s}}^1$ augmented with the six factors in $\boldsymbol{w}_{\mathrm{s.np}}^2$, 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. While 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 in the best of 16.77722 million models. One can find such factors 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 have 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. Thus, the model scan can be conducted over 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 full model space by a factor of 32.

We have coded a R function to do 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 likely a by-product of starting from a relatively large pool of initial characteristics.

It is interesting to examine the posterior distribution of the factor risk premia $\boldsymbol{\lambda}^*$ and the market price of factor risks (the SDF loadings) $\boldsymbol{b}^*$. We have a put a star to emphasize that these are the parameters associated with the risk factors in the best model. The first two centered moments of the marginal posterior distributions of these parameters are given in Table 7. It is noteworthy

Table 7 Slope factor SDF: The posterior estimates of the factor risk premia $\boldsymbol{\lambda}^*$ and market prices of factor risks $\boldsymbol{b}^*$ of $\boldsymbol{x}_{\mathrm{s}}^*$, the twenty slope risk factors.
This table shows the posterior estimates (mean and standard deviation) of the factor risk premia $\boldsymbol{\lambda}^*$ and 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 that exclude zero are marked in bold. The data runs from January 1989 to December 2020.

	λ^*		b^*	
$\boldsymbol{x}_{\mathrm{s}}^*$	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**	0.25
s.std_turn	-0.038	0.025	**-0.402**	0.185

that nineteen of the twenty market prices of factor risks have absolute value of the posterior mean more than twice bigger than the posterior sd. Thus, the PAMS method, which relies on marginal likelihoods (which, in turn, are objects marginalized over the parameters), and does not consider the posterior distribution of the market prices in the ranking of models, finds a model with posterior distributions of those market prices that have most of their posterior mass away from zero. It is less important that fewer than nineteen factor risk premia are significant in the same way. It is the market prices of factor risks that are relevant for pricing, rather than the factor risk premia.

4.4 Risk factor discovery: Differential and Rank factors

We next implement precisely the same PAMS discovery procedure to the set of differential and rank factors. We suppress the details because they run parallel to what we have described above for slope factors. The computations are reproducible using the R functions, data and R packages on the website associated with this paper. Once again, we get a large number of risk factors. For differential factors, we find fourteen risk factors and nineteen for rank factors:

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

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

For easy comparison, the different sets of risk factors are collected together in Figure 5, with the associated characteristics labeled on the x-axis and the risk factors indicated by points, colored in red for slope, green for differential and blue for rank risk factors. One can see that though there is some overlap, broadly the risk factor sets are different.[9] Thus, the method of factor construction matters for which risk are captured. To the best of our knowledge, this is the first time that such a

[9]It is not the aim of this paper to provide a theoretical justification for these risk factors. Nonetheless, the empirical finding that the SDF varies depending on how the factors are constructed has many practical ramifications. For example, to check any factor-based economic theory of asset pricing, it would not be enough to base that confirmation on factors that are 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 $\boldsymbol{\lambda}^*$ and market prices of factor risks $\boldsymbol{b}^*$. There are fourteen factors in the differential factor SDF and nineteen factors in the rank factor SDF.
This table shows the posterior estimates (mean and standard deviation) of the factor risk premia $\boldsymbol{\lambda}^*$ and SDF coefficients $\boldsymbol{b}^*$ 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 that exclude zero are marked in bold. The data runs from January 1989 to December 2020.

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

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.

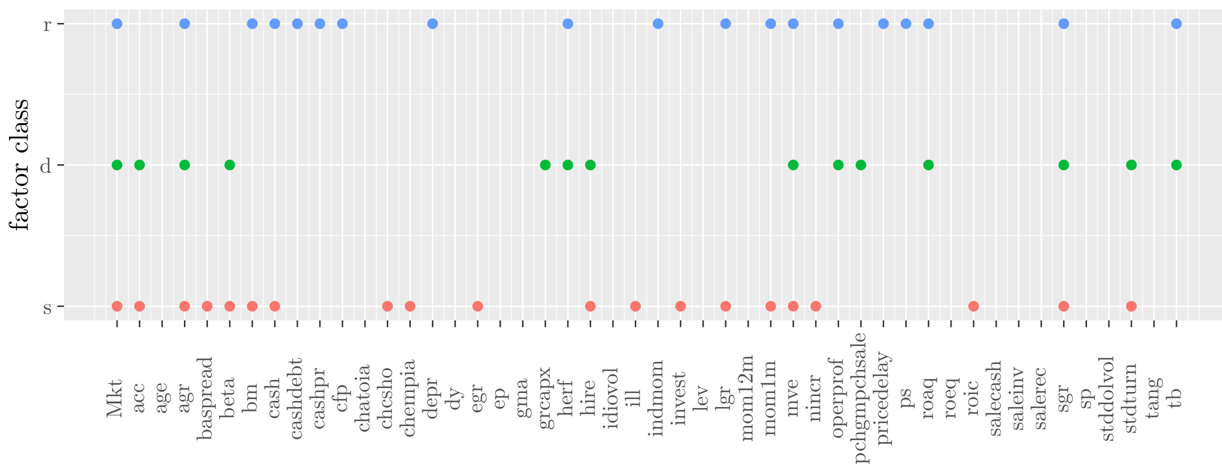

Figure 5 Risk factors within each factor class. There are 20 slope risk factors (colored in red), 14 differential risk factors (colored in green), and 19 rank risk factors (colored in blue) as described in text. The underlying characteristics are labeled on the horizontal axis.

5 Pricing of the cross section

It is one thing that the SDF's are different depending on the class of factors, but how much does this matter for pricing expected returns? To answer this question with confidence, we consider a large collection of test assets, consisting of portfolios, ETF's 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.

To judge pricing, we primarily apply a Bayesian approach. The Bayesian approach can give an answer to whether a particular asset is priced, whereas a frequentist approach can deliver an answer to whether an asset is not priced (since the null of pricing cannot be accepted). The pricing methodology is precisely the same as that discussed in Step 2 of the discovery procedure, with the change that the variable w_j in equations (18) and (19) now stands for the excess return of jth test asset. Specifically, if we let $m_{j,0}$ denote the marginal likelihood of the regression model (18) 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 (19) 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 hyperparameters of the prior 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 15% of the data is used to form 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 non-pricing. 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 from Fama and French (2018) (the FF6 factors are constructed by 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 of these portfolios, 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 of these risk factor sets provide better pricing than the benchmark FF6 risk factors.

Consider next the sample of 1480 ETFs. We assembled this sample from CRSP (sharecode 73) spanning the period February 1993 (the earliest month for which data on ETFs is available) to December 2020. Within this time-span, each ETF has at least 60 months of data. ETFs are diversified portfolios that are liquid, transparent, and cheap to trade. As a result, the premium of these assets is likely to be determined by exposure to the common non-diversifiable sources of risk manifested in the risk factors. Again, under the least 2:1 posterior odds in favor of pricing criteria, one can see that the slope risk factors outperform the other sets of risk factors.

Finally, we consider pricing a sample of 6024 stocks (CRSP sharecode 10 and 11). The sample period is from 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 pricing performance on stocks given that these assets tend to be more volatile than portfolios and, hence, pose a bigger hurdle.

Evidence on performance is given in Table 9. As can be seen from the table, under the (default) 2:1 criteria, the slope risk factors price 4883 stocks, 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

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). $\boldsymbol{x}_{\mathrm{s}}^{*}$ denote the twenty slope risk factors; $\boldsymbol{x}_{\mathrm{d}}^{*}$ denote the 14 differential risk factors, and $\boldsymbol{x}_{\mathrm{r}}^{*}$ 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 benchmark. The results show that slope risk factors uniformly price more of these test assets than the differential, rank and FF6 risk factors.

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

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 realized with effectively zero marginal effort.

5.2 Pricing of common risk factors

We conclude our evaluation of pricing power by 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, of course, 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 respective log marginal likelihoods. In addition, 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 $\hat{\alpha}$ has the estimates of the intercept in regressions with the factor on the LHS and the risk factors $\boldsymbol{x}_{\mathrm{f}}^*$ on the RHS; the columns labeled |t| has the absolute-value of the OLS estimates of the intercept; and the columns labeled d_{01} has the difference in the log marginal likelihoods of the model (equation (18)) without an intercept, and the log marginal likelihood of the model (equation (19)) 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 $\boldsymbol{x}_{\mathrm{d}}^*$			r risk factors $\boldsymbol{x}_{\mathrm{r}}^*$		
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 column in the last. Now 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. Thus, the slope risk factors can price nine of these common risk factors at posterior odds of 2:1. Reading across, 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, the differential risk factors cannot price four out 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 eleven 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 the slope risk factors outperform differential and rank risk factors not only in pricing a large collection of test assets but also in pricing existing risk factors.

6 Importance of 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 outperformance of slope factors is to compare the respective risk factors, as we did in the foregoing section.

Additionally, it is important to note that to realize the potential of slope factors it is necessary that these factors are constructed from cross-sectional regressions that control for a range of characteristics. If this is not done, the factors would not be pure-play. To be concrete, and to see what happens with non pure-play factors, let us construct the slope factors from linear, quadratic or cubic cross-sectional regressions (it does not matter which), but with only one control, mve. It also does not matter how mve is entered as a control. What is important is that we are not controlling for the effect for other relevant characteristics that may be available. Since the characteristics are correlated, even after controlling for size, as shown in Figure 1, the resulting slope factors will be non pure-play (and different from the corresponding pure play slope factors).

For the sake of illustration, and to see how the SDF based on non pure play slope factors prices, suppose that each characteristic is entered quadratically in the cross-sectional regression and mve is also entered quadratically. Thus, in each cross-section, forty six such regressions are run for characteristics other than mve. For the mve characteristic, the cross-sectional regression is (say) a quadratic in mve with no controls. Then, from each such regression in the cross-section, the slope factor for that characteristic, for that month, is the average of the estimated coefficients multiplying the linear and quadratic terms of that characteristic.

Non pure-play slope factors are simple to construct. If one is interested in the slope factor for a single characteristic, one can construct a non pure-play slope factor with data only on that characteristic and mve. As we now show, however, non pure-play slope factors trade-off limited data requirements for performance.

Because pure play and non pure play slope factors are different objects, the SDF based on non pure play risk factors will be different from the SDF we found above with play risk factors. Specifically, after applying the PAMS method to the set of forty seven non pure-play slope factors (we suppress the details for convenience), we find that the SDF best supported by the data contains the factors

$$\begin{aligned} \boldsymbol{x}^*_{\text{s.np}} = & (\text{Mkt, s.np.acc, s.np.cash, s.np.chcsho, s.np.lgr, s.np.mom1m,} \\ & \text{s.np.mve, s.np.pricedelay, s.np.sgr}), \end{aligned}$$

where the np label now stands for non pure-play. It should be noted that this result does not imply that nine underlying risks are being captured by these risk factors. Because the characteristics are correlated, even after controlling for size, and these factors did not control for other characteristics, these nine, in fact, each represents the risk emanating from other correlated characteristics, albeit coarsely. This coarsity of non pure play factors means that some risks cannot be captured by such factors, 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 risk factors $\boldsymbol{x}^*_{\mathrm{s.np}}$ 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 pure-play risk factors, $\boldsymbol{x}^*_{\mathrm{s}}$, $\boldsymbol{x}^*_{\mathrm{d}}$, and even the rank risk factors, $\boldsymbol{x}^*_{\mathrm{r}}$. Thus, outperformance is closely connected to the pure-play property of slope factors.

Table 11 Non 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 $\boldsymbol{x}^*_{\mathrm{s.np}}$ 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.

Test assets	# priced at 2:1	# priced at 3:1	# priced at 4:1
1150 Portfolios	682	510	304
1480 ETFs	921	663	412
6024 Stocks	4551	3225	2141

7 Conclusion

In this paper we have done a deep study of slope factors in relation to differential/sorted and rank factors and reached several conclusions that are important for empirical asset pricing. First, we have shown that slope factors provide value when they are pure-play, ie., 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, and thus come closer to what might be called pure-play factors.

Second, to realize the value of slope factors, it is not enough to take an existing asset pricing model, say the FF6, and replace its factors with 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, 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. We used PAMS to show that for our sample of forty seven characteristics, there are twenty slope factors, and that this SDF provides improved pricing of the cross-section relative to the FF6 model in which its factors are replaced by the corresponding slope factors.

Third, we show that the SDF best supported by the data varies by factor construction method. The PAMS methodology shows that the SDF has fourteen differential factors and nineteen rank factors. Thus, the method of factor construction matters for which risk are captured. To the best of our knowledge, this is the first time that 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.

Third, on an extensive set of test assets, consisting of 1150 portfolios, 1480 ETFs and (importantly) 6024 stocks, we have provided evidence that the SDF based on slope risk factors provides uniformly improved pricing of the cross section than the SDFs based on sorted and rank risk factors. Since this finding is based on a broader and more representative collection of test assets than the norm, this result engenders confidence that the outperformance is not an artifact of the test assets. We also document that the slope risk factors outperform in pricing the common (extant) risk factors. These findings on pricing outperformance could be used to justify an increased interest and use of slope factors in asset pricing.

Fourth, we show that the pricing outperformance of slope risk factors is due to the pure-play property. Non pure-play factors have degraded performance.

We believe that our paper should attract the attention of researchers in empirical finance since the evidence is compelling that the method of constructing factors leads to different risk factors (and hence different rendering of the risks that are priced in the cross-section), which, in turn, impacts pricing. It is possible that the evidence in this paper will hasten the greater use of slope factors in empirical finance.

Data, software and code to reproduce 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.

Qihui Chen, Nikolai Roussanov, and Xiaoliang Wang. Semiparametric conditional factor models: Estimation and inference. *arXiv preprint arXiv:2112.07121*, 2021.

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.

Lars Peter Hansen and Ravi Jagannathan. Implications of security market data for models of dynamic economies. *Journal of Political Economy*, 99(2):225–262, 1991.

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 and Stefan Nagel. When do cross-sectional asset pricing factors span the stochastic discount factor? *University of Chicago, Working Paper*, 2022.

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