PHBS WORKING PAPER SERIES

Winners from Winners: A Tale of Risk Factors

Siddhartha Chib
Washington University in St. Louis

Lingxiao Zhao
Peking University

Guofu Zhou
Washington University in St. Louis

November 2021

Working Paper 20211105

Abstract

Starting from the twelve distinct risk factors from Fama and French (1993, 2015, 2018), Hou, Xue, and Zhang (2015), Stambaugh and Yuan (2017), and Daniel, Hirshleifer, and Sun (2020), we construct and compare all possible asset pricing models by the Bayesian method of Chib, Zeng and Zhao (2020), and find that six risk factors, Mkt, SMB, MOM, ROE, MGMT, and PEAD, perform the best in terms of Bayesian posterior probability. A more extensive model comparison of 8,388,607 models, constructed from the twelve winners plus eleven principal components (PCs) of anomalies unexplained by the winners, shows that the best model consists of {Mkt, RMW, MOM, IA, ROE, MGT, PEAD, FIN} and the non-consecutive {PC1, PC3, PC4, PC5, and PC7}. The pricing performance suggests that these risk factors should be used for computing expected returns and for assessing investment strategies, instead of the risk factors in one of the other four collections.

Keywords: Model comparison, factor models, anomaly, discount factor, portfolio analysis.
JEL Classification: G12, C11, C12, C52, C58

Peking University HSBC Business School
University Town, Nanshan District
Shenzhen 518055, China

Winners from Winners: A Tale of Risk Factors*

Siddhartha Chib†, **Lingxiao Zhao**‡, **Guofu Zhou**§

January 2021

*We thank Doron Avramov, Amit Goyal, Sahn-Wook Huh, Hong Liu, Werner Ploberger, John Nachbar, Gaetano Antinolfi, and seminar participants at Washington University in St. Louis, Tongji University, Wuhan University, Hunan University, Antai College of Economics and Management (ACEM) of Shanghai Jiao Tong University, and Peking University HSBC Business School for insightful comments and suggestions. We are also grateful for Dashan Huang for his earlier contributions to this project, and thank Kyle Wang for research assistance.

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

‡Peking University HSBC Business School (PHBS), University Town, Nanshan District, Shenzhen, P.R.China 518055. e-mail lingxiao@phbs.pku.edu.cn.

§Olin School of Business, Washington University in St. Louis, 1 Bookings Drive, St. Louis, MO 63130. e-mail: zhou@wustl.edu

Winners from Winners: A Tale of Risk Factors

Abstract

Starting from the twelve distinct risk factors from Fama and French (1993, 2015, 2018), Hou, Xue, and Zhang (2015), Stambaugh and Yuan (2017), and Daniel, Hirshleifer, and Sun (2020), we construct and compare all possible asset pricing models by the Bayesian method of Chib, Zeng and Zhao (2020), and find that six risk factors, Mkt, SMB, MOM, ROE, MGMT, and PEAD, perform the best in terms of Bayesian posterior probability. A more extensive model comparison of 8,388,607 models, constructed from the twelve winners *plus* eleven principal components (PCs) of anomalies *unexplained by the winners*, shows that the best model consists of {Mkt, RMW, MOM, IA, ROE, MGT, PEAD, FIN} and the non-consecutive {PC1, PC3, PC4, PC5, and PC7}. The pricing performance suggests that these risk factors should be used for computing expected returns and for assessing investment strategies, instead of the risk factors in one of the other four collections.

JEL Classification: G12, C11, C12, C52, C58

Keywords: Model comparison, Factor models, Anomaly, Discount factor, Portfolio analysis.

1 Introduction

The question of which risk factors best explain the cross-section of expected equity returns continues to draw attention on account of the large importance of this topic for theoretical and empirical finance (Cochrane, 2011). Along with the market factor, hundreds of additional risk factors have emerged (Harvey, Liu, and Zhu, 2016; Hou, Xue, and Zhang, 2017), and the set of possible such factors continues to grow. Rather than add to this list, we ask two questions: (1) could we start with the risk factor collections that have generated support in the recent literature, take the union of the factors in those collections as the pool of winners, and then find a new set of risk factors (winners from winners) that gather even more support from the data, on both statistical and financial grounds? (2) can such a benchmark analysis be improved (and by how much) by considering not just the winners, but the winners plus (say) the 125 anomalies in Green, Hand, and Zhang (2017) and Hou, Xue, and Zhang (2017)?

To answer these questions, in what we call the tale of risk factors, we start with the four leading risk factor collections that we believe have spawned consensus support within the profession, namely those in the papers by Fama and French (1993, 2015, 2018), Hou, Xue, and Zhang (2015), Stambaugh and Yuan (2017), and Daniel, Hirshleifer, and Sun (2020). These collections, which cover the market, fundamental and behavioral factors, listed by author initials and risk factor abbreviations, are as follows[1]:

- FF6 collection: {Mkt, SMB, HML, RMW, CMA, MOM};
- HXZ collection: {Mkt, SMB, IA, ROE};
- SY collection: {Mkt, SMB, MGMT, PERF};
- DHS collection: {Mkt, PEAD, FIN}

[1]There are slight differences in these collections in relation to the size factor, which we ignore for simplicity.

In the first part of the analysis, the *benchmark model scan*, we focus on the set of these winners, and ask what collection of winners from winners emerge when each factor is allowed to play the role of a risk factor, ie., an element of the stochastic discount factor (SDF), or a non-risk factor, to produce different groupings of risk factors in a combinatorial fashion. Each grouping consists of a collection of factors that are risk factors in that grouping, and a complementary collection (the remaining factors) that are not risk factors in that grouping. We compare the resulting possible asset pricing models from a Bayesian model comparison perspective (Avramov and Chao, 2006; Barillas and Shanken, 2018; Chib and Zeng, 2020; Chib, Zeng, and Zhao, 2020). Specifically, we use the priors and marginal likelihoods given in Chib et al. (2020) to compare these models.

Our main result, from the benchmark model scan, is that a six-factor model, consisting of {Mkt, SMB, MOM, ROE, MGMT, and PEAD} from the twelve factors, gets the most support from the data. This model is closely followed by a second model, a seven-factor model that has PERF as an additional risk factor, and a third model, a five-factor model, that does not have MOM as a risk factor. In terms of probabilities, these models have posterior model probabilities of around 0.13, 0.11, 0.10, respectively. Though one would have liked to witness even more decisive posterior support for the best model, this is unrealistic on a large model space composed of models constructed from factors that are a prior winners. Nevertheless, the data evidence is clear in isolating the winners from winners as the posterior probability distribution beyond the top three slumps sharply. For example, the posterior probabilities of the fourth and fifth-best models are around 0.03, and the sixth is about 0.025. The posterior probabilities of the remaining models in the model space barely register, being roughly of the same size as the prior probability of 1/4,095 = 0.00025, and even below. On the basis of this evidence, the top three models isolated by our scan are worthy contenders for use in calculating risk-adjusted returns of funds and equity assets.

Some remarks about the models that are not in the top model set. For example, the model with all 12 factors as risk-factors, the full model, has a log-marginal likelihood of 14174.01 and is ranked 853. Thus, the default strategy of including all factors as risk-factors is dominated by many other models in the model space. Interestingly, the FF6, HXZ, SY, and DHS models also do not appear in the top model set. As we demonstrate later, this is due to fact that the risk factors in these 4 collections are not able to price the non-risk factors for that collection. Specifically, the single Mkt factor cannot price the remaining 11 non risk factors. The risk factors of the Fama and French family of models can price at most IA. The risk factors of HXZ can price all the Fama and French factors, but cannot price MGMT and PEAD. The risk factors of the SY and DHS models cannot price one non risk factor, PEAD and MGMT, respectively. In contrast, our six risk factors, {Mkt, SMB, MOM, ROE, MGMT, and PEAD}, can price all the remaining six factors.

We also document the performance of top models on other important statistical and economic dimensions. For one, we examine the performance of the top models in forecasting out-of-sample. Additionally, we find that the Sharpe ratio of the optimal mean-variance portfolio constructed from our six factors {Mkt, SMB, MOM, ROE, MGMT, and PEAD} is higher than the Sharpe ratios of the portfolios from the risk factors in the FF6, HXZ, SY, and DHS collections.

The next question we take up is about whether this benchmark scan can be improved by considering also the 125 anomalies in Green et al. (2017) and Hou, Xue, and Zhang (2017). Of course, the resulting model space of 2^{137} models is unfathomably large, and cannot be directly scanned, but it is possible to make the problem tractable. Our idea is to combine a finance motivated dimension reduction strategy (to reduce the number of factors) followed by our usual model scan. It works as follows. We first use the initial set of 12 factors (and a Bayesian model comparison method) to find the set of *genuine anomalies*, ie., anomalies that cannot be priced by the winners. Our empirical analysis reveals that this set just consists of twenty anomalies. This set of genuine anomalies is small enough that the model space of $2^{32}-1$ models can actually be scanned. Nonetheless, we reduce the dimension even further by converting these genuine anomalies to the space of principal components (PCs). Our model scan is then applied to the twelve initial factors and the first eleven principal components of the genuine anomalies. In other words, each of the twelve winners, and each of the eleven PC's, is allowed to play the role of a risk-factor or a non-risk factor. Each model in this space of 8,388,607 models is then compared by our Bayesian model comparison technique. We refer to this scan as the *benchmark plus genuine anomalies model scan*.

It should be noted that in this extended scan, we keep the key factors (the winners) as is, but only covert the less important (the genuine anomalies) into principal components. In this way, we can understand whether the PC factors provide incremental information for pricing assets over the factors from the benchmark scan, and which factors from the benchmark scan continue to be risk factors in this broader space of models.

Our extended model scan shows the benefit of incorporating information in genuine anomalies in explaining the cross-section of expected equity returns. For example, the Sharp ratios increases by about 30%. In addition, we find that the best model consists of {Mkt, RMW, MOM, IA, ROE, MGT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7}, and that the common factors in the top model set (the set consisting of the best model plus the models whose log-marginal distance to the best model is less than 1.15), are the nine factors, {Mkt, MOM, ROE, PEAD, MGMT} and {PC1, PC4, and PC5}.

This evidence about the risk factors {Mkt, RMW, MOM, IA, ROE, MGT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7} has strong implications for empirical work in finance. For one, the pricing evidence suggests that we should use this collection for computing expected returns, or for assessing performance of various investment strategies, instead of the risk factors in one of the other four collections. Admittedly, these thirteen risk factors arose from a statistical approach, but the statistical approach is grounded in the SDF, and is not ad-hoc. Therefore, given the pricing performance of these risk factors, and performance on other metrics, one can be confident about the empirical usefulness of these risk factors.

Our paper is part of a new wave of Bayesian approaches to risk factor selection. For instance, Kozak, Nagel, and Santosh (2020) focus on PC factors that are constructed from 50 anomalies, and utilize economic priors to estimate a single model with those 50 contending factors in a classical penalized regression estimation. Bryzgalova, Huang, and Julliard (2019) on the other hand, focus on the Fama-McBeth regression, and select from amongst 50 factors by a Bayesian variable selection method. More broadly, Bayesian contributions in finance include, for example, Pástor (2000), Pástor and Stambaugh (2000), Pástor and Stambaugh (2001), Avramov (2002), Ang and Timmermann (2012), Goyal, He, and Huh (2018) and Smith and Timmermann (2021). These papers consider a variety of problems, economic priors, predictability, model uncertainty, regime changes, and testing, from the Bayesian viewpoint, but do not directly intersect with the work in this paper.

We see our paper as contributing to the further understanding of risk-factor determination from a Bayesian model uncertainty perspective. A key contribution is the procedure that allows us to consider a large initial pool of factors by splitting that pool into winners, and using the winners to whittle down the genuine (or non-priced factors) into a set of principal components, and then applying our model scan strategy to the winners plus principal components. By only converting the genuine anomalies to PCs, we enhance interpretability of the final set of risk-factors. We would like to emphasize that the dimension reduction method we outline broadens the scope of the model scanning method of Chib and Zeng (2020) to a much larger pools of factors than has been considered to date.

The rest of the paper is organized as follows. In Section 2, we briefly outline the methodology that we use to conduct our model comparisons. In Section 3, we consider the benchmark model scan, and in Section 4 the winners plus genuine anomalies model scan. Section 5 and Section 6 contain results and Section 7 concludes.

2 Methodology

Suppose that the potential risk factor set is denoted by $\boldsymbol{f}_t : K \times 1$, where t denotes the t-th month. We now 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 K initial possible risk factors, there are, therefore, $J = 2^K - 1$ 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 $\mathbb{M}_j$, $j = 1, ..., J$.

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

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

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

$$\boldsymbol{x}_{j,t} = \boldsymbol{\lambda}_{x,j} + \varepsilon_{x,j,t}, \tag{2}$$

$$\boldsymbol{w}_{j,t} = \Gamma_j \boldsymbol{x}_{j,t} + \varepsilon_{w\cdot x,j,t}, \tag{3}$$

where the errors are block independent Gaussian

$$\begin{pmatrix} \varepsilon_{x,j,t} \\ \varepsilon_{w\cdot x,j,t} \end{pmatrix} \sim \mathbb{N}_K \left(\mathbf{0}, \begin{pmatrix} \Omega_{x,j} & \mathbf{0} \\ \mathbf{0} & \Omega_{w\cdot x,j} \end{pmatrix} \right), \tag{4}$$

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

The goal of the analysis is to calculate the support for these models, $\mathbb{M}_j,\ j = 1,\ldots,J$, given the sample data on the factors. To explain how this is done, we start with the prior distributions of the parameters across models. It is important that these priors are comparable across models to ensure that the ranking of the models depends on the data evidence, not on differences in the prior. Chib and Zeng (2020) introduce an approach for doing this with proper prior distributions. Subsequently, Chib, Zeng, and Zhao (2020) specialize their approach with improper priors.

The starting point of the Chib and Zeng (2020) priors is a prior on the model in which all factors are pricing factors (and therefore $k_{x,1} = K$). Call this model $\mathbb{M}_1$. Consider now

$$\boldsymbol{\eta}_1 = \Omega_{x,1}\,,$$

the variance parameter (the covariance matrix) in this model. Give $\Omega_{x,1}$ an inverse Wishart prior

$$\pi(\boldsymbol{\eta}_1|\mathbb{M}_1) = \mathbb{IW}_K(\nu, Q) \propto |\Omega_{x,1}|^{-\frac{\nu+K+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(Q\Omega_{x,1}^{-1}) \right), \tag{5}$$

where ν, the degree of freedom, and Q, the scale matrix, are the two parameters of this prior.

The next step is to notice that for any other model $\mathbb{M}_j$, the set of parameters

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

are simply 1-1 functions of $\boldsymbol{\eta}_1$. To see this, first note that the dimension of $\boldsymbol{\eta}_j$, which equals

$$\begin{aligned}
&\{k_{w,j}k_{x,j} + k_{x,j}(k_{x,j}+1)/2 + k_{w,j}(k_{w,j}+1)/2\} \\
=&\{k_{x,j}^2 + k_{w,j}^2 + 2k_{x,j}k_{w,j} + (k_{x,j}+k_{w,j})\}/2 \\
=&(K^2+K)/2,
\end{aligned}$$

is the same as the dimension of $\boldsymbol{\eta}_1$. To show that the mapping is 1-1, calculate the Jacobian,

$$\left|\frac{\mathrm{d}\boldsymbol{\eta}_1}{\mathrm{d}\boldsymbol{\eta}_j}\right| = |\Omega_{x,j}|^{k_{w,j}}$$

which is positive. The derivation of this Jacobian is in Chib, Zeng, and Zhao (2020).

Based on these facts, Chib and Zeng (2020) derive the prior of $\boldsymbol{\eta}_j$ (for $j > 1$) from the prior of $\boldsymbol{\eta}_1$ by the change of variable formula

$$\pi(\boldsymbol{\eta}_j|\mathbb{M}_j) = \pi(\boldsymbol{\eta}_1|\mathbb{M}_1)\left|\frac{\mathrm{d}\boldsymbol{\eta}_1}{\mathrm{d}\boldsymbol{\eta}_j}\right| \tag{6}$$

where the Jacobian is given above. By this tactic, all the priors are derived from a *single prior distribution*. In other words, all the different priors are fixed by just specifying ν and Q. These different priors can be viewed as identical. Substituting (5) on the right hand side, using properties of the determinant of a partitioned matrix, that of the inverse of a partitioned matrix, and those of the trace operator, we get

$$\pi(\boldsymbol{\eta}_j|\mathbb{M}_j) = \mathbb{IW}_{k_{x,j}}(\Omega_{x,j}|\nu - k_{w,j}, Q_{x,j})\mathbb{IW}_{k_{w,j}}(\Omega_{w\cdot x,j}|\nu, Q_{w\cdot x,j}) \tag{7}$$

$$\times \mathbb{N}_{k_{xj}\times k_{wj}}\left(\gamma_j|\text{vec}(Q_{wx,j}Q_{x,j}^{-1}), Q_{x,j}^{-1}\otimes\Omega_{w\cdot x,j}\right), \tag{8}$$

where $\gamma_j = vec(\Gamma_j)$, and the different Q matrices are partitioned according to Ω.

In this paper we use the priors of Chib et al. (2020) instead. These can be derived from (5), (7) and (8) by setting $\nu = 0$ and $Q = 0$. Then for the full model we get Jeffreys' improper prior

$$\pi(\boldsymbol{\eta}_1|\mathbb{M}_1) = c|\Omega_{x,1}|^{-\frac{K+1}{2}}, \tag{9}$$

and for all the remaining models we get

$$\pi(\boldsymbol{\eta}_j|\mathbb{M}_j) = c|\Omega_{x,j}|^{-\frac{-k_{w,j}+k_{x,j}+1}{2}} \times |\Omega_{w\cdot x,j}|^{-\frac{k_{w,j}+1}{2}} \times |\Omega_{w\cdot x,j}|^{-\frac{k_{x,j}}{2}} \tag{10}$$

$$= c|\Omega_{x,j}|^{-\frac{2k_{x,j}-K+1}{2}}|\Omega_{w\cdot x,j}|^{-\frac{K+1}{2}},\ j > 1. \tag{11}$$

One key benefit of the priors in (9) and (11) is that the marginal likelihoods, discussed next, are all in closed form. The constant c is irrelevant because it is the same constant in all the priors and cancels in taking log marginal likelihood differences across models.

Finally, we complete the prior distributions by supposing that, conditional on $\boldsymbol{\eta}_j$,

$$\boldsymbol{\lambda}_{x,j}|\mathbb{M}_j, \boldsymbol{\eta}_j \sim \mathbb{N}_{k_{x,j}}(\boldsymbol{\lambda}_{x,j,0}, \kappa_j\Omega_{x,j}),$$

where $\boldsymbol{\lambda}_{x,j,0}$ and κ_j are model-specific hyperparameters that are determined from a training sample.

2.1 Marginal Likelihoods

Assume that each model $\mathbb{M}_j \in \mathscr{M}$ has a prior model probability of $\Pr(\mathbb{M}_j)$ of being the correct model. The objective is to calculate the posterior model probability $\Pr(\mathbb{M}_j|\boldsymbol{y}_{1:T})$, where $\boldsymbol{y}_{1:T} = (\boldsymbol{f}_1, ..., \boldsymbol{f}_T)$ is the estimation sample of the potential risk factors.

The key quantities for performing this prior-posterior update for the models in $\mathscr{M}$ are the marginal likelihoods, defined as

$$m_j(\boldsymbol{y}_{1:T}|\mathbb{M}_j) = \int_{\Theta_j} p(\boldsymbol{y}_{1:T}|\mathbb{M}_j, \boldsymbol{\theta}_j)\pi(\boldsymbol{\lambda}_{x,j}|\mathbb{M}_j, \boldsymbol{\eta}_j)\psi(\boldsymbol{\eta}_j|\mathbb{M}_j)\mathrm{d}\boldsymbol{\theta}_j, \tag{12}$$

where Θ_j is the domain of $\boldsymbol{\theta}_j$,

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

is the density of the data and $\mathbb{N}_d(\cdot|\mu, \Omega)$ is the d-dimensional multivariate normal density function with mean μ and covariance matrix Ω.

Notice that the phrase marginal likelihood encapsulates two concepts: one that it is a function that is marginalized over the parameters of model j, hence the word marginal; and second that it is the likelihood of the model parameter $\mathbb{M}_j$, hence the word likelihood.

Under our assumptions, the *log* marginal likelihoods are in closed form. Specifically,

$$\log m_1(\boldsymbol{y}_{1:T}|\mathbb{M}_1) = -\frac{TK}{2}\log\pi - \frac{K}{2}\log(T\kappa_1 + 1) - \frac{T}{2}\log|\Psi_1| + \log\Gamma_K\left(\frac{T}{2}\right), \tag{13}$$

and

$$\log m_j(\boldsymbol{y}_{1:T}|\mathbb{M}_j) = -\frac{Tk_{x,j}}{2}\log\pi - \frac{k_{x,j}}{2}\log(T\kappa_j+1) - \frac{\left(T+k_{x,j}-K\right)}{2}\log|\Psi_j| + \log\Gamma_{k_{x,j}}\left(\frac{T+k_{x,j}-K}{2}\right) \tag{14}$$

$$-\frac{(K-k_{x,j})(T-k_{x,j})}{2}\log\pi - \frac{\left(K-k_{x,j}\right)}{2}\log|W_j^*| - \frac{T}{2}\log|\Psi_j^*| + \log\Gamma_{K-L_j}\left(\frac{T}{2}\right),\ j>1\ , \tag{15}$$

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

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

Note that the variables in 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.

2.2 Model Scan Approach

We conduct a prior-posterior analysis on the model space denoted by $\mathscr{M} = \{\mathbb{M}_1, \mathbb{M}_2, \ldots, \mathbb{M}_J\}$. Assume that each model in the model space is given an uninformative and equalized prior model probability, that is, for any j

$$\Pr(\mathbb{M}_j) = 1/J. \tag{16}$$

Applying Bayes theorem to the unknown model parameter $\mathbb{M}_j$, the posterior model probability is given by

$$\Pr(\mathbb{M}_j|\boldsymbol{y}_{1:T}) = \frac{m_j(\boldsymbol{y}_{1:T}|\mathbb{M}_j)}{\sum_{l=1}^{J} m_l(\boldsymbol{y}_{1:T}|\mathbb{M}_l)}\ , \tag{17}$$

as the model prior probabilities in the numerator and the denominator cancel out.

Both the prior and posterior probability distributions on model space acknowledge the notion of *model uncertainty*. The prior distribution on model space represents our beliefs about the models before we see the data. A discrete uniform prior is our default, but, of course, it is possible to consider other prior distributions. The posterior distribution retains model uncertainty unless the sample size is large in relation to the dimension of the model space. By this we mean that the posterior distribution on model space will not concentrate on a single model. As T becomes large, however, the asymptotic theory of the marginal likelihood (see, e.g., Chib, Shin, and Simoni (2018)), implies that the posterior model probabilities will concentrate on the true model (if it is in the set of models), or on the model that is closest to the true model in the Kullback-Leibler distance.

Regardless of the sample size, however, the end-product of our analysis is a ranking of models by posterior model probabilities (equivalently, by marginal likelihoods given that the denominator in the posterior probability calculation is just a normalization constant). We indicate these ranked models by

$$\{\mathbb{M}_{1*}, \mathbb{M}_{2*}, \ldots, \mathbb{M}_{J*}\}$$

such that

$$m_{1*}(\boldsymbol{y}_{1:T}|\mathbb{M}_{1*}) > m_{2*}(\boldsymbol{y}_{1:T}|\mathbb{M}_{2*}) > \cdots > m_{J*}(\boldsymbol{y}_{1:T}|\mathbb{M}_{J*}).$$

This ranking provides the basis for our empirical Bayesian model comparison.

3 Benchmark Model Scan

As mentioned in the introduction, we conduct an initial analysis with the 12 winners to benchmark what is possible by working with just well known factors. This sets the stage for our extended analysis on a larger space of models that are based on the standard factors plus anomaly factors.

Details of these standard factors, {Mkt, SMB, HML, RMW, CMA, MOM, IA, ROE, MGMT, PERF, PEAD, FIN}, are given in Table 1. While Mkt captures the overall market risk, SMB, HML, RMW, CMA, MOM, IA, and ROE are well-known characteristic-based factors and are constructed by sorting stocks in a relatively simple way. For those remaining novel four mispricing factors, MGMT and PERF are constructed based on average rankings of eleven anomalies of stocks, and PEAD and FIN are related to investors' psychology. Although the construction and motivation of those factors are different, those twelve factors are named "winners" as they are believed and proved to have strong power in explaining the cross-section of expected equity returns.

The data on these standard factors are monthly, spanning the period from January 1974 to December 2018, for a total of 540 starting observations. Of these the last 12 months of data are held-out for out-of-sample prediction validation purpose. For the other 528 in-sample monthly observations, the first 10 percent is used as a training sample to construct the prior distributions of the risk premia parameters, leaving a sample size of $T = 475$ as estimation sample.

3.1 Two special cases

To understand how the framework is applied, consider first the model in which all twelve winners are risk factors. In this case, model $\mathbb{M}_1$ (say), the general model reduces to

$$\boldsymbol{x}_{1,t} = \underbrace{\boldsymbol{\lambda}_{x,1}}_{12\times 1} + \varepsilon_{x,1,t}, \;\; \varepsilon_{x,1,t} \sim \mathbb{N}_{12}(\mathbf{0}, \Omega_{x,1}), \tag{18}$$

where $\boldsymbol{x}_{1,t} = (\text{Mkt}, \text{SMB}, \text{SML}, \text{RMW}, \text{CMA}, \text{MOM}, \text{IA}, \text{ROE}, \text{MGMT}, \text{PERF}, \text{PEAD}, \text{FIN})'_t$ and, since the non-risk factor collection $\boldsymbol{w}_{1,t}$ is empty, $k_{x,1} = 12$ and $\Omega_{x,1} : 12 \times 12$.

Now consider $\mathbb{M}_2$ (say) with the FF6 risk factors $\boldsymbol{x}_{2,t} = (\text{Mkt}, \text{SMB}, \text{SML}, \text{RMW}, \text{CMA}, \text{MOM})'_t$.

Then, we have

$$\boldsymbol{x}_{2,t} = \underbrace{\boldsymbol{\lambda}_{x,2}}_{6\times 1} + \boldsymbol{\varepsilon}_{x,2,t}, \ \ \boldsymbol{\varepsilon}_{x,2,t} \sim \mathbb{N}_6(\mathbf{0}, \Omega_{x,2}),$$

while for $\boldsymbol{w}_{2,t} = (\text{IA}, \text{ROE}, \text{MGMT}, \text{PERF}, \text{PEAD}, \text{FIN})'_t$ we have

$$\boldsymbol{w}_{2,t} = \underbrace{\Gamma_2}_{6\times 6} \boldsymbol{x}_{2,t} + \boldsymbol{\varepsilon}_{w\cdot x,2,t}, \ \ \boldsymbol{\varepsilon}_{w\cdot x,2,t} \sim \mathbb{N}_6(\mathbf{0}, \Omega_{w\cdot x,2}),$$

where $k_{x,2} = 6$, $k_{w,2} = 6$, $\Omega_{x,2} : 6 \times 6$, and $\Omega_{w\cdot x,2} : 6 \times 6$. The latter model embodies the pricing restrictions that the assumed risk factors of this model price the non-risk factors $\boldsymbol{w}_{2,t}$.

There are $J = 4,095$ such models in the model space $\mathscr{M}$. Our goal is to compare these J models using the model scan approach described in Section 2.

3.2 Benchmark Model Scan Results

3.2.1 Top Model Set

To get a clear picture of the prior-posterior update on the model space $\mathscr{M}$, we view each model as a point in that space. The prior distribution of models on that space is uniform. The posterior probabilities of the models are proportional to the product of the uniform prior and the marginal likelihoods. We can use these posterior probabilities to plot these points (or models) in that space in decreasing order. From Figure 1, which plots the posterior model probability of the top 220 models. We can see from the figure that the posterior model probabilities drop sharply beyond the top three models. For example, the posterior probabilities of the fourth and fifth-best models are around 0.03, and the sixth is about 0.025. The posterior probabilities of the remaining models in the model space barely register, being roughly of the same size as the prior probability of 1/4,095 = 0.00025, and even below.

In Figure 2 we plot these posterior model probabilities but, this time, only for the top 5 models. We see that the top three models have a joint posterior probability of 0.3407. In a sense, we can think of these models as being indistinguishable, or equivalent. To make this more precise, in the notation introduction above, let $\mathbb{M}_{1*}$ denote the highest posterior probability model. If we now let the Bayes factor of the best model against any other model be denoted by

$$B_{1j} = \frac{m_{1*}(\boldsymbol{y}_{1:T}|\mathbb{M}_{1*})}{m_{j*}(\boldsymbol{y}_{1:T}|\mathbb{M}_{j*})},$$

then, according to Jeffreys' scale, if $B_{1j} \leq 10^{\frac{1}{2}}$, the evidence supporting $\mathbb{M}_{1*}$ over $\mathbb{M}_{j*}$ is barely worth mentioning. Equivalently, in terms of the log Bayes factor, the indistinguishably condition above can be expressed as

$$\log B_{1j} = \log m_{1*}(\boldsymbol{y}_{1:T}|\mathbb{M}_{1*}) - \log m_{j*}(\boldsymbol{y}_{1:T}|\mathbb{M}_{j*}) \leq 1.15.$$

We can, therefore, refer to $\mathbb{M}_{j*}$ that is in the radius of the best model in this way as being indistinguishable from the best model.

Applying this criterion, we conclude that $\mathbb{M}_{1*}$, $\mathbb{M}_{2*}$, and $\mathbb{M}_{3*}$ constitute the top model set $\mathscr{M}_*$ in the winners scan, while $\mathbb{M}_{4*}$ and $\mathbb{M}_{5*}$ also given in Figure 2 are not in the top model set.

Table 2 shows that the six-factor model $\mathbb{M}_{1*}$ consisting of Mkt, SMB, MOM, ROE, MGMT, and PEAD as risk factors gets the most support from the data. This model is closely followed by the seven-factor model $\mathbb{M}_{2*}$ that has PERF as an additional risk factor, and the five-factor model $\mathbb{M}_{3*}$ that does not have MOM as a risk factor.

Interestingly, Mkt, SMB, ROE, MGMT, and PEAD, are present in each of the three top groupings. It appears that both fundamental and behavioral factors play an important role in pricing the cross-section of expected equity returns. It should also be noted that the top groupings feature between five and seven-factors, similar to the number of factors in most of the literature.

Besides, the ratio of the posterior model probability and the prior model probability of any given model $\mathbb{M}_j$, denoted by $\frac{\Pr(\boldsymbol{y}_{1:T}|\mathbb{M}_{j*})}{\Pr(\mathbb{M}_{j*})}$, is provided in Table 2. That ratio reflects the information improvement of posterior over the same prior for $\mathbb{M}_j$ when data are observed. Therefore it is a good measure for evaluating the joint superiority of candidate models. For comparison, Panel B of Table 2 reports the marginal likelihoods and that ratios for $\mathbb{M}_1$ and models with the same risk factor sets as CAPM, Fama and French family, HXZ, SY, and DHS. It can seen that the information improvement of each of those models is substantially smaller than that of top models. Because $\mathbb{M}_1$ is also not supported by the data, we can conclude that the twelve factors together contain information redundancies.

3.2.2 Parameter Updating for the Best Model of the Benchmark Model Scan

We now provide more details about the best model in the benchmark model scan $\mathbb{M}_{1*}$, which takes the form

$$\mathbb{M}_{1*}: \begin{pmatrix} \text{Mkt}_t \\ \text{SMB}_t \\ \text{MOM}_t \\ \text{ROE}_t \\ \text{MGMT}_t \\ \text{PEAD}_t \end{pmatrix} = \underbrace{\boldsymbol{\lambda}_{x,1*}}_{6\times 1} + \varepsilon_{x,1*,t}, \quad \varepsilon_{x,1*,t} \sim \mathbb{N}_6(\mathbf{0}, \boldsymbol{\Omega}_{x,1*}),$$

$$
\begin{pmatrix} \text{HML}_t \\ \text{RMW}_t \\ \text{CMA}_t \\ \text{IA}_t \\ \text{PERF}_t \\ \text{FIN}_t \end{pmatrix} = \underbrace{\Gamma_{1*}}_{6\times 6} \begin{pmatrix} \text{Mkt}_t \\ \text{SMB}_t \\ \text{MOM}_t \\ \text{ROE}_t \\ \text{MGMT}_t \\ \text{PEAD}_t \end{pmatrix} + \varepsilon_{w\cdot x,1*,t}, \;\; \varepsilon_{w\cdot x,1*,t} \sim \mathbb{N}_6(\mathbf{0}, \Omega_{w\cdot x,1*}).
$$

In calculating the marginal likelihood of this model, which equals 14186.43, as reported earlier in Table 2, we used the prior on $\boldsymbol{\eta}_{1*} = (\Gamma_{1*}, \Omega_{x,1*}, \Omega_{w\cdot x,1*})$ from (11) equal to $|\Omega_{x,1*}|^{-\frac{1}{2}} |\Omega_{w\cdot x,1*}|^{-\frac{13}{2}}$, and the proper prior of the risk premia parameters $\boldsymbol{\lambda}_{x,1*}$ from the training sample equal to

$$
\pi(\boldsymbol{\lambda}_{x,1*}|\mathbb{M}_{1*}, \boldsymbol{\eta}_{1*}) = \mathbb{N}_6(\boldsymbol{\lambda}_{x,1*}|\boldsymbol{\lambda}_{x,1*,0}, 0.1915 \times \Omega_{x,1*}),
$$

where $\boldsymbol{\lambda}_{x,1*,0} = (0.0017, 0.0130, 0.0044, 0.0041, 0.0084, 0.0085)'$.

Under our prior distributions, it is easy to confirm that the posterior distributions $\pi(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T})$ of parameters $\boldsymbol{\theta}_j$ of any given model $\mathbb{M}_j$ have the marginal-conditional forms given by

$$
\pi(\Omega_{x,j}|\mathbb{M}_j, \boldsymbol{y}_{1:T}) = \mathbb{W}^{-1}_{k_{x,j}}(\Omega_{x,j}|\Psi_j, T + k_{x,j} - K), \tag{19}
$$

$$
\pi(\boldsymbol{\lambda}_{x,j}|\mathbb{M}_j, \boldsymbol{y}_{1:T}, \Omega_{x,j}) = \mathbb{N}_{k_{x,j}}(\boldsymbol{\lambda}_{x,j}|\boldsymbol{\lambda}_{xj1}, \left(1/\kappa_j + T\right)^{-1} \Omega_{x,j}), \tag{20}
$$

and

$$
\pi(\Omega_{w\cdot x,j}|\mathbb{M}_j, \boldsymbol{y}_{1:T}) = \mathbb{W}^{-1}_{k_{w,j}}(\Omega_{w\cdot x,j}|\Psi^*_j, T), \tag{21}
$$

$$
\pi(\text{vec}(\Gamma_j)|\mathbb{M}_j, \boldsymbol{y}_{1:T}, \Omega_{w\cdot x,j}) = \mathbb{N}_{k_{w,j}\times k_{x,j}}(\text{vec}(\Gamma_j)|\text{vec}(\hat{\Gamma}_j), W_j^{*-1} \otimes \Omega_{w\cdot x,j}), \tag{22}
$$

where $\boldsymbol{\lambda}_{x,j,1} = \frac{1}{T\kappa_j+1}\boldsymbol{\lambda}_{x,j,0} + \frac{T\kappa_j}{T\kappa_j+1}\hat{\boldsymbol{\lambda}}_{x,j}$ and $\mathbb{W}^{-1}(\Psi, \nu)$ denotes the inverse Wishart distribution with ν degrees of freedom and scale matrix Ψ. Thus, the posterior distribution $\pi(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T})$ is given by the product of equations (19), (20), (21), and (22). We can apply this result to generate a large number of simulated draws, first by sampling the marginal distribution, and then by the conditional distribution given the draws from the marginal distributions. These sampled draws can be used to make marginal posterior distributions of relevant parameters, and other summaries.

Applying the above sampling procedure to $\mathbb{M}_{1*}$, we obtain the marginal posterior distributions of the risk premia parameters $\boldsymbol{\lambda}_{x,1^*}$, given in Figure 3, and the posterior means, standard deviations and quantiles, given in Table 3. It is interesting that the posterior means of the risk premia parameters are similar, except for that of SMB, while the posterior standard deviations of the risk premia of Mkt and MOM are almost twice as large as the rest.

3.3 Pricing Performance

We now discuss why the FF6, HXZ, SY, and DHS risk factor models do not appear in the top model set. The reason for this is interesting. Essentially, these models are not able to price the left out factors. For example, consider the case of the {Mkt, SMB, HML, RMW, CMA, and MOM} risk factors in the FF6 collection. The *w* factors (the remaining factors) for this collection are {IA, ROE, MGMT, PERF, PEAD, and FIN}. The asset pricing model for FF6 assumes that these six *w* factors can be priced - hence the intercepts are all zero. But what if the FF6 factors cannot price these *w* factors? In that case, the data would prefer that we did not impose the restrictions, but the model requires that we do - the pricing condition is, therefore, binding. The *x* factors in that model can be said to fail an internal consistency condition, and the model suffers a marginal likelihood penalty.

Hence, this is our definition. We say that a particular model satisfies the internal consistency condition if its risk factors can price the set of non risk factors in that model without incurring a penalty. In other words, a penalty is incurred, and the marginal likelihood suffers if the constraint that the intercepts in w model, which by assumption must be all zero, is binding. If the constraints are not binding, then its marginal likelihood is significantly higher.

Consider model $\mathbb{M}_j$

$$
\begin{aligned}
\boldsymbol{x}_{j,t} &= \boldsymbol{\lambda}_{x,j} + \varepsilon_{x,j,t}, \\
\boldsymbol{w}_{j,t} &= \Gamma_j \boldsymbol{x}_{j,t} + \varepsilon_{w\cdot x,j,t},
\end{aligned}
$$

with risk factors $\boldsymbol{x}_{j,t}$ and non risk factors $w_{j,t} = (w_{j,1,t}, \ldots, w_{j,k_{w,j},t})'$ with dimension $k_{w,j} \times 1$. Now for each non risk factor $w_{j,i,t}$, $i = 1, 2, \ldots, k_{w,j}$, we compare the two models,

$$\mathbb{M}^i_{j,0}: \; w_{j,i,t} = \gamma'_{j,i} \boldsymbol{x}_{j,t} + \varepsilon_{j,i,t} \tag{23}$$

and

$$\mathbb{M}^i_{j,1}: \; w_{j,i,t} = \alpha_{j,i} + \gamma'_{j,i} \boldsymbol{x}_{j,t} + \varepsilon_{j,i,t} \tag{24}$$

using marginal likelihoods. If the log marginal likelihood of the second model does not exceed that of the first model by more than 1.15, then, by an application of Jeffreys' rule, we can conclude that imposing the zero $\alpha_{j,i}$ pricing restriction does not result in a marginal likelihood penalty. Stated yet another way, this means that the non risk factor $w_{j,i}$ can be priced by the risk factor set $\boldsymbol{x}_j$ of that model $\mathbb{M}_j$. If this condition holds for each of the factors in w_j, we conclude that the risk factor set of that model satisfies the ICC condition.

Our analysis shows that the CAPM, Fama and French family, HXZ, SY, and DHS models do not satisfy ICC. Specifically, the single Mkt factor cannot price the remaining 11 non risk factors. The risk factor sets of the Fama and French family of models can price at most one non risk factor (IA). The risk factors of HXZ can price all of the Fama and French factors, but cannot price MGMT and PEAD. The risk factors of SY and DHS models cannot price one non risk factor, PEAD and MGMT, respectively. In contrast, the top models in $\mathscr{M}_*$ satisfy the ICC condition fully, which helps to explain why those models rank high in the benchmark model scan.

3.4 Prediction

It is worthwhile to consider how well the top models perform out-of-sample. From the Bayesian perspective, an elegant way to examine this question is by calculating the *predictive likelihood* for a set of future observations. This predictive likelihood, which like the marginal likelihood, is a number when evaluated at a particular sample of future observations, can be used to rank the predictive performance of each model in the model space.

To define the predictive likelihood, let $\pi(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T})$ denote the posterior distributions of the parameters $\boldsymbol{\theta}_j$ of $\mathbb{M}_j$, and let $\boldsymbol{y}_{(T+1):(T+12)} = (\boldsymbol{f}_{T+1}, ..., \boldsymbol{f}_{T+12})$ denote 12 months of held-out out-of-sample data on those winners. Then, for any given model $\mathbb{M}_j$, the predictive likelihood is defined as

$$m_j\left(\boldsymbol{y}_{(T+1):(T+12)}|\mathbb{M}_j, \boldsymbol{y}_{1:T}\right) = \int_{\Theta_j} p(\boldsymbol{y}_{(T+1):(T+12)}|\mathbb{M}_j, \boldsymbol{\theta}_j)\pi(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T})\mathrm{d}\boldsymbol{\theta}_j,$$

where

$$p(\boldsymbol{y}_{(T+1):(T+12)}|\mathbb{M}_j, \boldsymbol{\theta}_j) = \prod_{s=1}^{12} \mathbb{N}_{k_{x,j}}(\boldsymbol{x}_{j,T+s}|\boldsymbol{\lambda}_{x,j}, \Omega_{x,j})\mathbb{N}_{k_{w,j}}(\boldsymbol{w}_{j,T+s}|\Gamma_j \boldsymbol{x}_{j,T+s}, \Omega_{w\cdot x,j})$$

is the out-of-sample density of the factors given the parameters. We can compute this integral by Monte Carlo. Taking draws $\{\boldsymbol{\theta}_j^{(1)}, ..., \boldsymbol{\theta}_j^{(G)}\}$ from $\pi(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T})$, with G being a large integer, we calculate the predictive likelihood as the Monte Carlo average

$$m_j\left(\boldsymbol{y}_{(T+1):(T+12)}|\mathbb{M}_j, \boldsymbol{y}_{1:T}\right) = \frac{1}{G}\sum_{g=1}^{G} p(\boldsymbol{y}_{(T+1):(T+12)}|\mathbb{M}_j, \boldsymbol{\theta}_j^{(g)}).$$

Table 4 reports the log predictive likelihoods of the top three models as well as those competing models. We can see that the top three models also have larger predictive likelihoods, which means that they outperform the competing models on the predictive dimension.

4 Benchmark Plus Genuine Anomalies Model Scan

We now show that there are some benefits to including additional potential risk factors along with the winners. There are many additional risk factors to draw upon and the approach we describe can be used with any set of additional potential risk factors. For our analysis here we focus on the 125 anomalies in **?** and **?**. These anomalies, which are re-balanced at an annual or quarterly frequency, exclude anomalies that are re-balanced at a monthly frequency, because the latter cease to be anomalies once transaction costs are taken into account (Novy-Marx and Velikov, 2016; Patton and Weller, 2020). All these portfolios are value-weighted and held for one month. Just as in the case of the winners, we have monthly observations on these anomalies spanning the period from January 1974 to December 2018, for a total of 540 observations. We partition these observations into out-of-sample and in-sample, which consists of the training sample and the estimation sample, just as in Section 3.

What we aim to show is that there is information in these anomalies that can be captured to produce better statistical performance (in terms of marginal likelihoods) and higher Sharpe-ratios of portfolios built from the best fitting risk-factors. In order to show this, we recognize that the winners are already carefully vetted factors and, therefore, the anomalies that are allowed to enter the pool of augmented potential risk factors must be those that cannot be priced by these winners. This point helps to limit the dimension of the model space and allows us to design a full model scan approach, as we now detail.

4.1 Genuine Anomalies

The model space with all 125 anomalies is 2^{137}, which is astronomically large. However, it is unnecessary to consider such a large model space because many of the anomalies can actually be priced by the winners. In fact, Harvey, Liu, and Zhu (2016) and **?** have cast doubt on the credibility of these anomalies and Cochrane (2011) has raised similar concerns.

The first step, therefore, is to eliminate anomalies that are not *genuine anomalies*. An anomaly is a genuine anomaly if it cannot be priced by the winners. Here is how we sort this issue out. Let z_i, $i = 1, 2, \ldots, 125$, denote the anomalies. Let $\boldsymbol{x} =$ (Mkt, SMB, HML, RMW, CMA, MOM, IA, ROE, MGMT, PERF, PEAD, FIN) denote the twelve winners. Now for each anomaly z_i as the response, and x as the covariates, we estimate two models, one without an intercept and one with:

$$\mathbb{M}_0^i: \; z_{i,t} = \gamma_i' \boldsymbol{x}_t + \varepsilon_{i,t}, \; \varepsilon_{i,t} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, \sigma_i^2) \tag{25}$$

and

$$\mathbb{M}_1^i: \; z_{i,t} = \alpha_i + \gamma_i' \boldsymbol{x}_t + \varepsilon_{i,t} \; \varepsilon_{i,t} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, \sigma_i^2). \tag{26}$$

In estimating these models, the first 10 percent of the data are used as a training sample to pin down the hyperparameters of the proper prior distributions. We then use standard Bayesian Markov chain Monte Carlo methods to estimate model on the remaining 90 percent of the data, $\boldsymbol{y}_{1:T}$. The log marginal likelihood of each model is computed by the Chib (1995) method. Denote these marginal likelihoods by $\log m_0^i(\boldsymbol{y}_{1:T}|\mathbb{M}_0^i)$ and $\log m_1^i(\boldsymbol{y}_{1:T}|\mathbb{M}_0^i)$. Then, based on the Jeffreys

(1961) scale, if the following condition holds

$$\log m_1^i(\boldsymbol{y}_{1:T}|\mathbb{M}_1^i) - \log m_0^i(\boldsymbol{y}_{1:T}|\mathbb{M}_0^i) > 1.15, \tag{27}$$

then x is not able to price z_i. In this case, we assert that z_i is a genuine anomaly; otherwise z_i is not a genuine anomaly.

Applying the procedure described above, twenty genuine anomalies emerge, namely, acc, age, currat, hire, lev, quick, salecash, sgr, Em, Lbp, dFin, Cop, Lfe, SA, sue, cash, OLAQ, CLAQ, TBIQ, and BLQ. Details about these anomalies are given in Table 5.

4.2 The Potential Risk Factor Set

Now instead of conducting our model scan on the original twelve winners and these twenty genuine anomalies, which leads to a model space of around four billion models ($2^{32} - 1 = 4,294,967,295$), we apply a second dimension reduction step by converting the genuine anomalies to principal components (with the rotated mean added back in), of which we then consider the first eleven that explain in total approximating 91% of the variation in the genuine anomalies, as can be seen from Table 6. This set, of the twelve winners and the first eleven PCs of the genuine anomalies, constitutes our potential risk factor set which we use to launch our extended risk factor analysis.

We note that this strategy of blending of winners and the PCs in this way appears to be new to the literature. By this strategy, we are able to limit the model space to a reasonable dimension (of around eight million models), while simultaneously avoiding the problem of working with twenty correlated PCs that are also quite correlated with the winners. For instance, some anomalies are related to leverage (currat, lev, quick, Lbp, and BLQ) and some are linked to sales status (salecash and sgr). Considering two groups of risk factors in this way, where some are in their original form and some are PCs, appears to be novel. It allows us to show the value of including anomalies as potential risk factors.

We also note that the idea of transforming our genuine anomalies into their corresponding principal components is similar to Kozak, Nagel, and Santosh (2020) who argued that "a relatively small set of principal component from the universe of potential characteristics-based factors can approximate the SDF quite well." Our idea is related, but distinct, as we keep the key factors (the winners) as is, but only covert the less important (the genuine anomalies) into principal components.

We emphasize again that our approach of reducing the 125 anomalies to the set of twenty genuine anomalies is a *dimension-reduction strategy*. Furthermore, our idea of converting these anomalies to the space of principal components, is another element of that same strategy. Of course, whether as anomalies, or as PC's, these factors are portfolios of assets. Our purpose here is to understand whether the PC factors provide incremental information for pricing assets.

5 Benchmark Plus Genuine Anomalies Model Scan Results

5.1 Top Model Set

Starting with the potential risk factor set of dimension $\tilde{K} = 23$, twelve winners plus eleven PCs, and applying the methodology given in Section 2, we calculate the marginal likelihood of each of the $\tilde{J} = 8,388,607$ models in $\tilde{\mathscr{M}}$. Converting these marginal likelihoods into posterior model probabilities, the ratios of these posterior model probabilities and the prior model probability (assumed equal to $1/\tilde{J}$) can be calculated. The ratio, $\frac{\Pr(\tilde{\boldsymbol{y}}_{1:T}|\tilde{\mathbb{M}}_{j*})}{\Pr(\tilde{\mathbb{M}}_{j*})}$, defined earlier in Section 3.2.1, makes it easier to see which models receive more support from the data. We report these ratios for the top 220 models in Figure 4, in which the dashed blue vertical line represents the cutoff of the top model set.

The top model set, denoted by $\tilde{\mathscr{M}}_*$ as in Section 3.2.1, can be defined in relation to the best model $\tilde{\mathbb{M}}_{1*}$. A model is in the top model set if its distance to the best model on the log marginal likelihood scale is less than 1.15. These 29 models, along with the including associated risk factor sets, log marginal likelihoods, and the ratios of posterior model probability and prior model probability are provided in Panel A of Table 7. The risk factors common to all these top models are Mkt, MOM, ROE, PEAD, MGMT, PC1, PC4, and PC5. We also note that the risk factors that are common to the top 3 models in the benchmark model scan, {Mkt, SMB, ROE, PEAD, MGMT}, are also the risk factors that are common to the top 29 models in the extended model scan except that SMB is replaced by MOM, which is also risk factors of the top 2 models of the winners scan.

5.1.1 Parameter Updating for the Best Model

We now provide more details about the best model in the winners plus genuine model scan $\tilde{\mathbb{M}}_{1*}$, in which the risk factor set is given by $\tilde{\boldsymbol{x}}_{1*,t} = (\text{Mkt}, \text{RMW}, \text{MOM}, \text{IA}, \text{ROE}, \text{MGMT}, \text{PEAD}, \text{FIN}, \text{PC1}, \text{PC3}, \text{PC4}, \text{PC5}, \text{PC7})'_t$ and the non risk factor set is given by $\tilde{\boldsymbol{w}}_{1*,t} = (\text{SMB}, \text{HML}, \text{CMA}, \text{PERF}, \text{PC2}, \text{PC6}, \text{PC8}, \text{PC9}, \text{PC10}, \text{PC11})'_t$:

$$\begin{aligned}
\tilde{\mathbb{M}}_{1*}:\ \tilde{\boldsymbol{x}}_{1*,t} &= \underbrace{\tilde{\boldsymbol{\lambda}}_{x,1*}}_{13\times 1} + \tilde{\boldsymbol{\varepsilon}}_{x,1*,t},\ \ \tilde{\boldsymbol{\varepsilon}}_{x,1*,t} \sim \mathbb{N}_{13}(\mathbf{0}, \tilde{\boldsymbol{\Omega}}_{x,1*}),\\
\tilde{\boldsymbol{w}}_{1*,t} &= \underbrace{\tilde{\boldsymbol{\Gamma}}_{1*}}_{10\times 10} \tilde{\boldsymbol{x}}_{1*,t} + \tilde{\boldsymbol{\varepsilon}}_{w\cdot x,1*,t},\ \ \tilde{\boldsymbol{\varepsilon}}_{w\cdot x,1*,t} \sim \mathbb{N}_{10}(\mathbf{0}, \tilde{\boldsymbol{\Omega}}_{w\cdot x,1*}).
\end{aligned}$$

Similar to Section 3.2, the prior and posterior statistics, i.e. prior mean, posterior mean, posterior standard deviation, posterior median, 2.5% quantile, and 97.5% quantile for the risk premia $\tilde{\boldsymbol{\lambda}}_{1*}$ are provided in Table 8.

5.2 Pricing performance: Internal consistency check

Just as in Section 3.3, we can see that 27 out of 29 models in the top model set $\tilde{\mathscr{M}}_*$ satisfy the ICC completely. The two exceptions occur for $\mathbb{M}_{16*}$ which is unable to explain IA and $\mathbb{M}_{16*}$ which is unable to explain RMW. In constrast, models with the same risk factor sets as CAPM, Fama and French family, HXZ, SY, and DHS deviate from ICC further, leaving out 13, 12, 10, 9, 5, 5, and 4 non-risk factors unexplained. Moreover, none of models with the same risk factor sets as the top three models in the winners scan can explain PC1, PC2, and PC3.

5.3 Prediction

Similar to Section 3.4, it is important to consider how well the winning model performs out-of-sample and we compute the *predictive likelihood* for a set of future observations $\tilde{\boldsymbol{y}}_{(T+1):(T+12)} = (\tilde{\boldsymbol{f}}_{T+1}, ..., \tilde{\boldsymbol{f}}_{T+12})$ denote 12 months of out-of-sample data on the winners and principal components. Table 9 reports the log predictive likelihoods for top models in $\tilde{\mathscr{M}}_*$ as well as models with the same risk factor sets as CAPM, Fama and French family models, SY and DHS. We can tell that those top models do not fail out of sample.

6 Economic Performance: Sharpe Ratios

Now suppose that based on the identity of the risk factors in (say) the best model $\tilde{\mathbb{M}}_{1*}$ of the winners plus genuine anomalies model scan, namely, Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, and PC7, we construct an optimal mean-variance portfolio of these risk factors together with a risk-free asset. Similarly, we construct the optimal mean-variance portfolios from the risk factors in $\tilde{\mathbb{M}}_1$, $\mathbb{M}_1$ CAPM, Fama and French family, HXZ, SY, and DHS collections, as well as the risk factors of top models of the benchmark model scan. This leads to the important question: how do the Sharpe ratios of those different portfolios compare?

We construct these different portfolios in the following manner. Consider model $\mathbb{M}_j$ with associated risk factors $\boldsymbol{x}_j$. Given the data $\boldsymbol{y}_{1:T}$, consider calculating the *predictive* mean of $\boldsymbol{x}_{j,T+1}$

$$\mathbb{E}[\boldsymbol{x}_{j,T+1}|\mathbb{M}_j, \boldsymbol{y}_{1:T}] \triangleq \boldsymbol{x}_{j,T+1|T} = \int \boldsymbol{x}_{j,T+1} \mathbb{N}_{k_{x,j}}(\boldsymbol{x}_{j,T+1}|\boldsymbol{\lambda}_{x,j}, \Omega_{x,j}) \boldsymbol{\pi}(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T}) \mathrm{d}\boldsymbol{\theta}_j \mathrm{d}\boldsymbol{x}_{j,T+1},$$

which by changing the order of the integration can be seen to just equal the posterior mean of $\boldsymbol{\lambda}_{x,j}$

$$\boldsymbol{x}_{j,T+1|T} = \hat{\boldsymbol{\lambda}}_{x,j} \triangleq \int_{\Theta_j} \boldsymbol{\lambda}_{x,j} \boldsymbol{\pi}(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T}) \mathrm{d}\boldsymbol{\theta}_j,$$

and the *predictive* covariance of $\boldsymbol{x}_{j,T+1}$

$$\Omega_{x,j,T+1|T} \triangleq \int (\boldsymbol{x}_{j,T+1} - \boldsymbol{x}_{j,T+1|T})(\boldsymbol{x}_{j,T+1} - \boldsymbol{x}_{j,T+1|T})' \mathbb{N}_{k_{x,j}}(\boldsymbol{x}_{j,T+1}|\boldsymbol{\lambda}_{x,j}, \Omega_{x,j}) \boldsymbol{\pi}(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T}) \mathrm{d}\boldsymbol{\theta}_j \mathrm{d}\boldsymbol{x}_{j,T+1},$$

which by the law of iterated expectations for covariances simplifies to the sum of the posterior mean of $\Omega_{x,j}$ and the posterior variance of $\boldsymbol{\lambda}_{x,j}$:

$$\Omega_{x,j,T+1|T} = \int_{\Theta_j} \Omega_{x,j} \boldsymbol{\pi}(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T}) \mathrm{d}\boldsymbol{\theta}_j + \int_{\Theta_j} (\boldsymbol{\lambda}_{x,j} - \hat{\boldsymbol{\lambda}}_{x,j})(\boldsymbol{\lambda}_{x,j} - \hat{\boldsymbol{\lambda}}_{x,j})' \boldsymbol{\pi}(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T}) \mathrm{d}\boldsymbol{\theta}_j.$$

Of course, both these quantities are straightforwardly estimated from the output of the simulation of the posterior density $\boldsymbol{\pi}(\boldsymbol{\theta}_j|\mathbb{M}_j, \boldsymbol{y}_{1:T})$. Given these predictive moments, with certain calculations, the Sharpe ratio of the optimal mean-variance portfolio of the factors in $\boldsymbol{x}_j$ plus a risk-free asset is available as

$$\mathrm{Sh}_j = \left(\hat{\boldsymbol{\lambda}}_{x,j}' \Omega_{x,j,T+1|T}^{-1} \hat{\boldsymbol{\lambda}}_{x,j}\right)^{\frac{1}{2}}.$$

In Table 10 we report the Sharpe ratios of risk-factor portfolios based on several asset pricing models. In Panel A we consider the risk factor sets of the top models in $\tilde{\mathscr{M}}_*$, in Panel B for the risk factor sets of the top models in $\mathscr{M}_*$, and in Panel C for the $\tilde{\mathbb{M}}_1$, $\mathbb{M}_1$, CAPM, Fama and French family, HXZ, SY, and DHS models.

Looking at the results in Panel B and C, we can see that the Sharpe ratios are higher for the top models from the winners scan than those from some of the existing asset pricing models. Comparing the results in Panel A and Panel B, we see that the top models from the winners plus genuine anomalies model scan provide even higher Sharpe ratios. Taken together, if we consider the best performing DHS model as the benchmark from Panel C, we can see that the models in $\mathscr{M}_*$ have about 17% higher Sharpe ratios, and the models in $\tilde{\mathscr{M}}_*$ have about 49% higher Sharpe ratios.

Finally, in the benchmark model scan, all twelve winners can achieve a Sharpe ratio of 0.56, whereas investing in the seven winners of winners in $\mathbb{M}_{2*}$ gives a Sharpe ratio of 0.55. And in the winners plus genuine anomalies model scan, investing in those top risk factor sets produces a Sharpe ratio as high as 0.70 while investing in all twelve winners plus eleven PCs gives 0.71.

In summary, based on the Sharpe ratios, we have three observations. First, the portfolios of risk factors from the top models perform as well as those from the complete set of risk factors. In other words, there is some information redundancy in the potential risk factor set. Second, the marginal likelihood ranking and Sharpe ratio rankings are aligned. Third, the PCs have incremental explanatory power, but further research is needed to identify if some individual factors from amongst the selected PCs are equivalent to those PCs.

7 Pricing the cross-section

We conclude by providing a comparison of our two risk factor collections

- Mkt, SMB, MOM, ROE, MGMT, and PEAD and

- {Mkt, RMW, MOM, IA, ROE, MGT, PEAD, FIN} and {PC1, PC3, PC4, PC5, and PC7}

in pricing the cross-section of stocks. In particular, the testing assets are excess returns of individual stocks over the period 1974:01 to 2018:12, which is the period of our estimation sample. For this sample period, we have 282 such stock returns. Following the method used to extract the genuine anomalies, for each individual stock return, we estimate two models for each risk factor set, one without an intercept and one with, and use the 1.15 log-marginal likelihood criterion to decide whether that testing asset is priced.

The results show that the first risk factor collection fails to price three testing assets. On the other hand, the second risk factor set (from the benchmark plus genuine anomalies scan) is able to price all 282 testing assets. We view this as further compelling evidence of the empirical usefulness of this collection for computing expected returns and assessing investment strategies.

8 Conclusion

We have studied the implications of the Bayesian model scan framework of Chib and Zeng (2020) and Chib et al. (2020) for risk factor selection. Starting from the twelve distinct risk factors in Fama and French (1993, 2015, 2018), Hou, Xue, and Zhang (2015), Stambaugh and Yuan (2017), and Daniel, Hirshleifer, and Sun (2020), a comparison of all possible implied asset pricing models shows that the six-factor model consisting of {Mkt, SMB, MOM, ROE, MGMT, and PEAD} gets the most support from the data in terms of posterior model probability, out-of-sample predictive performance, and Sharpe ratio.

We also devise a way for conducting a model scan when the initial set of potential risk factors leads to a model space that is astronomically large. For example, we consider the case where one is interested in selecting the best risk factors from the initial set of 12 risk factors plus 125 anomalies.

The resulting model space of 2^{137} models is intractable. It is not known which of the existing methods (if any) could work in such a case. The idea we develop employs a finance motivated dimension reduction strategy to reduce the model space to a manageable size.

The results from the benchmark plus genuine anomalies scan show that the best model consists of {Mkt, RMW, MOM, IA, ROE, MGT, PEAD, FIN} plus the non-consecutive principal components, {PC1, PC3, PC4, PC5, PC7}. The pricing comparison suggests that we should use this collection for computing expected returns and assessing investment strategies, instead of the risk factors in one of the standard collections.

The general approach that we have described in this paper has applications beyond this paper. The idea of combining well vetted factors to come up the best few, and the idea of using a finance-motivated dimension reduction to make scanning of otherwise insurmountably large model spaces possible, are likely to prove useful for other asset categories such as bonds, currencies and commodities, and is likely to open up other avenues for further research.

Table 1 Winners Definitions

Factors	Definitions
Mkt	the excess return of the market portfolio
SMB	the return spread between diversified portfolios of small size and big size stocks
HML	the return spread between diversified portfolios of high and low B/M stocks
RMW	the return spread between diversified portfolios of stocks with robust and weak profitability
CMA	the return spread between diversified portfolios of the stocks of low (conservative) and high (aggressive) investment firms
MOM	the momentum factor based on two prior returns
IA	the investment factor based on annual changes in total assets divided by lagged total assets
ROE	the profitability factor based on income before extraordinary items divided by one-quarter-lagged book equity
MGMT	the mispricing factor controlled by management
PERF	the mispricing factor related to performance
PEAD	the short-horizon behavioral factor motivated by investor inattention and evidence of short-horizon under reaction
FIN	the long-horizon behavioral factor exploiting the information in managers' decisions to issue or repurchase equity

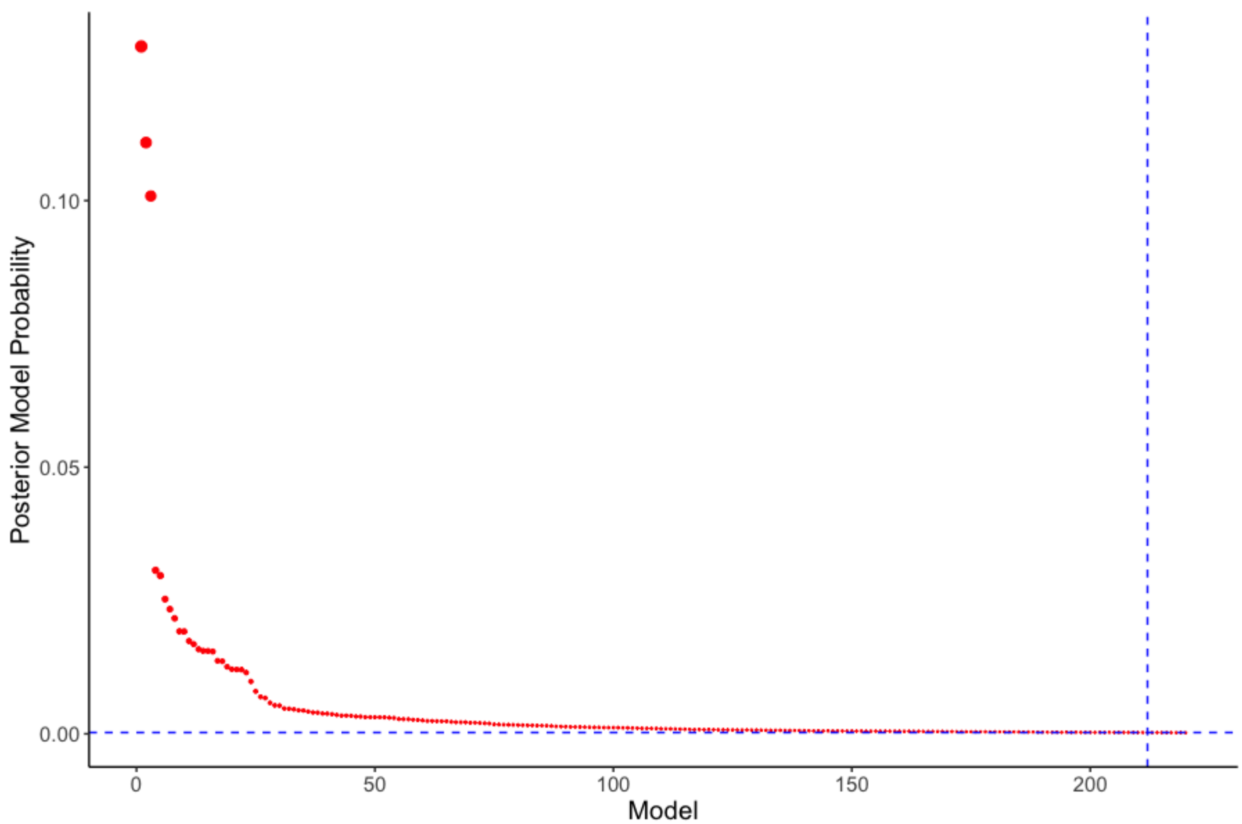

Figure 1 Top 220 Models of the Winners Model Scan

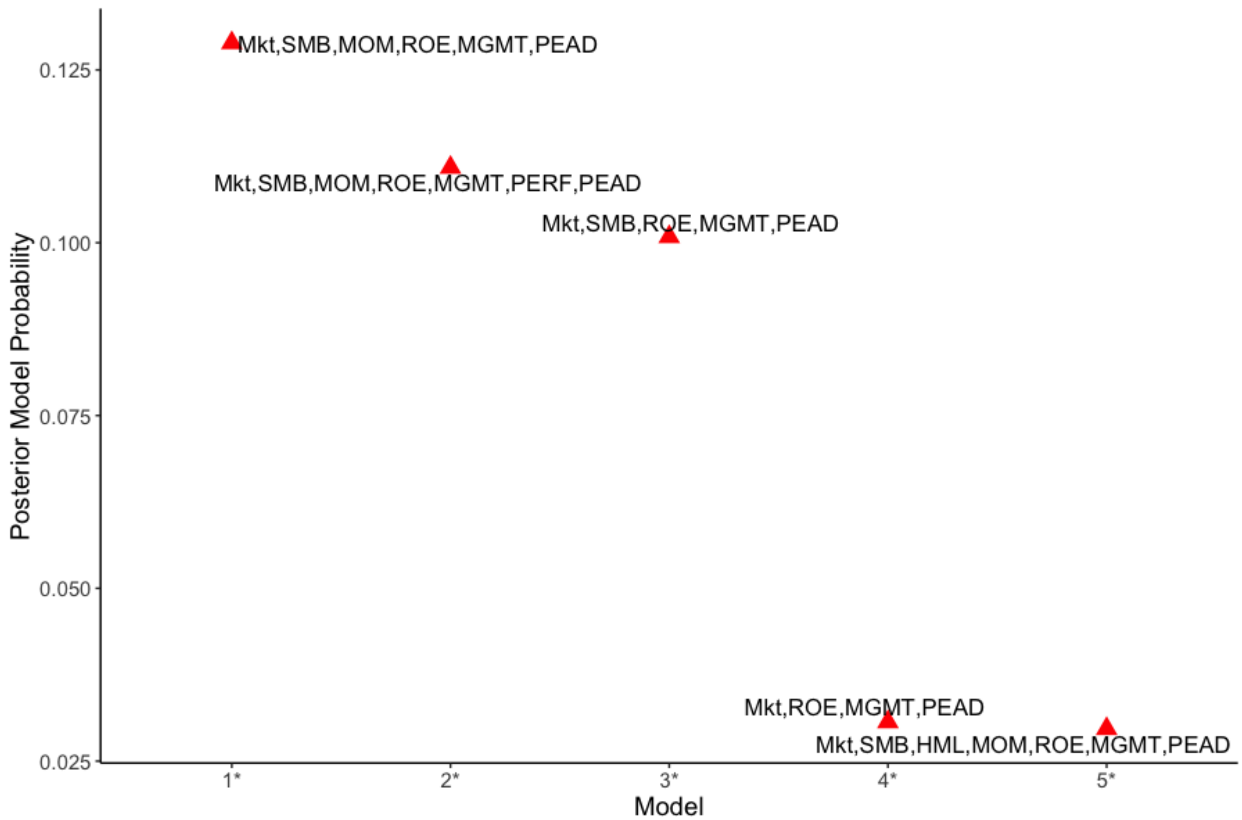

Figure 2 Top 5 Models of the Winners Model Scan

Table 2 Marginal Likelihoods and the Ratios of Posterior Model Probability and Prior Model Probability of Selected Models in the Winners Model Space $\mathscr{M}$

Results from the comparison of the $\tilde{J} = 4,095$ models. Panel A has the results for the top three models, and Panel B for $\tilde{\mathbb{M}}_1$ and models with the same risk factor sets as CAPM, FF3, FF5, FF6, HXZ, SY, and DHS.

Risk factors	$\log m_j(\boldsymbol{y}_{1:T}\|\mathbb{M}_j)$	$\frac{\Pr(\mathbb{M}_j\|\boldsymbol{y}_{1:T})}{\Pr(\mathbb{M}_j)}$
Panel A: Top three models		
Mkt, SMB, MOM, ROE, MGMT, PEAD	14186.43	527.89
Mkt, SMB, MOM, ROE, MGMT, PERF, PEAD	14186.28	454.09
Mkt, SMB, ROE, MGMT, PEAD	14186.18	413.01
Panel B: $\tilde{\mathbb{M}}_1$ and models with the same risk factor sets as CAPM, FF3, FF5, FF6, HXZ, SY, and DHS		
12 winners	1.66×10^{-1}	
Mkt	14140.85	8.44×10^{-18}
Mkt, SMB, HML	14140.32	4.98×10^{-18}
Mkt, SMB, HML, RMW, CMA	14152.79	1.30×10^{-12}
Mkt, SMB, HML, RMW, CMA, MOM	14154.45	6.87×10^{-12}
Mkt, SMB, IA, ROE	14164.47	1.54×10^{-7}
Mkt, SMB, MGMT, PERF	14173.32	1.07×10^{-3}
Mkt, PEAD, FIN	14178.86	2.73×10^{-1}

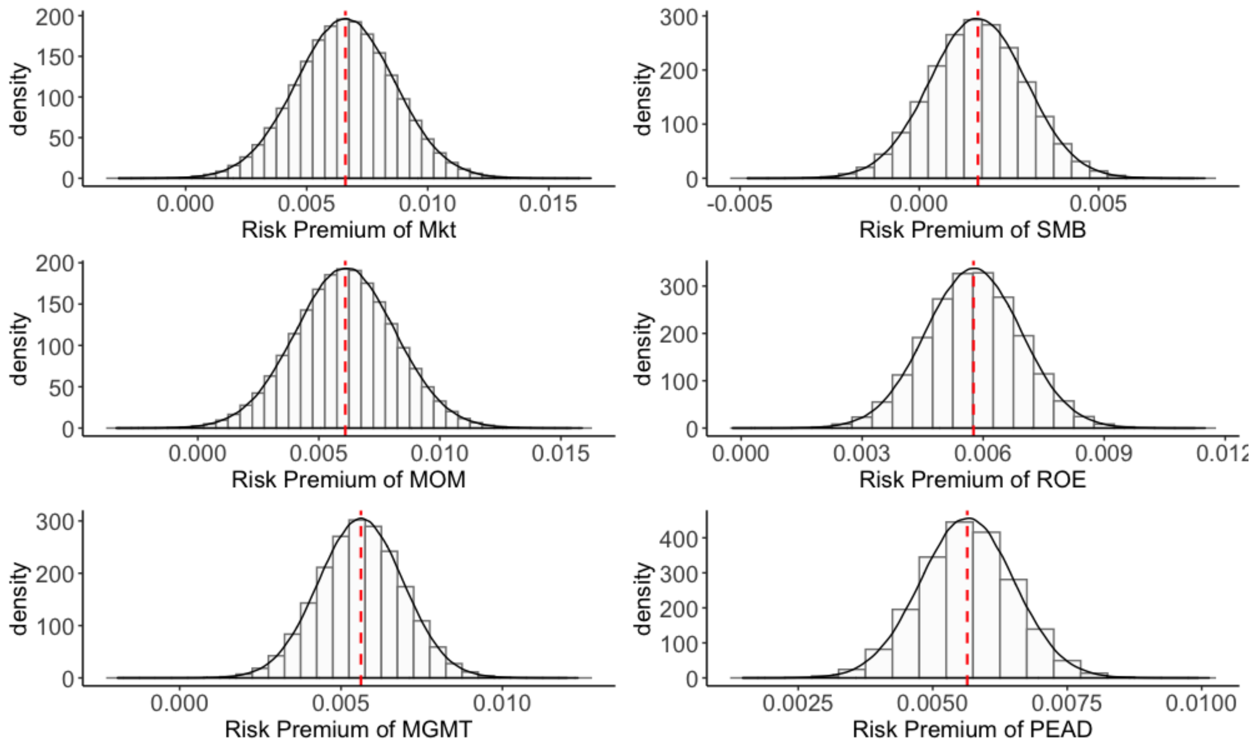

Figure 3 Posterior Distributions of the Risk Premia Parameters of the Best of the Winners Model Scan $\mathbb{M}_{1*}$

Table 3 Prior and Posterior Statistics of the Risk Premia Parameters of the Best of the Winners Model Scan $\mathbb{M}_{1*}$

Prior and posterior statistics, i.e. prior mean, posterior mean, posterior standard deviation, posterior median, 2.5% quantile, and 97.5% quantile for the risk premia $\boldsymbol{\lambda}_{x,1*}$ of the best model $\mathbb{M}_1^*$, which has Mkt, SMB, MOM, ROE, MGMT, and PEAD as risk factors.

	Prior mean	Posterior mean	Posterior sd	Posterior median	2.5% Quantile	97.5% Quantile
Mkt	0.0017	0.0066	0.0020	0.0066	0.0026	0.0106
SMB	0.0130	0.0016	0.0013	0.0016	-0.0010	0.0043
MOM	0.0044	0.0061	0.0021	0.0061	0.0020	0.0101
ROE	0.0041	0.0058	0.0012	0.0058	0.0034	0.0081
MGMT	0.0084	0.0056	0.0013	0.0056	0.0030	0.0082
PEAD	0.0085	0.0056	0.0009	0.0056	0.0039	0.0074

Table 4 Predictive Likelihoods for the Winners Model Scan

Predictive likelihoods of selected asset pricing models in winners model scan.

Risk factors	$\log m_j(\boldsymbol{y}_{(T+1):(T+12)} \mid \mathbb{M}_j)$
Panel A: Top three models	
Mkt, SMB, MOM, ROE, MGMT, PEAD	383.48
Mkt, SMB, MOM, ROE, MGMT, PERF, PEAD	383.65
Mkt, SMB, ROE, MGMT, PEAD	383.55
Panel B: Models with the same risk factor sets as CAPM, FF3, FF5, FF6, HXZ, SY, and DHS	
Mkt	382.46
Mkt, SMB, HML	381.76
Mkt, SMB, HML, RMW, CMA	381.67
Mkt, SMB, HML, CMA, RMW, MOM	381.83
Mkt, SMB, IA, ROE	381.87
Mkt, SMB, MGMT, PERF	382.89
Mkt, FIN, PEAD	382.76

Table 5 Surviving Anomalies Explanations

Anomalies	Explanations
acc	annual income before extraordinary items minus operating cash flows divided by average total assets
age	number of years since first Compustat coverage
currat	current assets / current liabilities
hire	percent change in number of employees
lev	total liabilities divided by fiscal year-end market capitalization
quick	(current assets - inventory) / current liabilities
salecash	annual sales divided by cash and cash equivalents
sgr	annual percent change in sales (sale)
Em	enterprise value divided by operating income before depreciation (Compustat annual item OIBDP)
Lbp	leverage component of book to price
dFin	the change in net financial assets
Cop	cash-based operating profitability
Lfe	labor force efficiency
SA	SA index measuring financial constraint
sue	the high-minus-low earnings surprise
cash	cash and cash equivalents divided by average total assets
OLAQ	quarterly operating profits-to-lagged assets
CLAQ	quarterly cash-based operating profits-to-lagged assets
TBIQ	quarterly taxable income-to-book income
BLQ	quarterly book leverage

Table 6 Importance of Principal Components

	PC1	PC2	PC3	PC4	PC5
Standard deviation	0.1143	0.0785	0.0441	0.0415	0.0408
Proportion of Variance	0.3953	0.1866	0.0590	0.0521	0.0503
Cumulative Proportion	0.3953	0.5819	0.6409	0.6930	0.7433
	PC6	**PC7**	**PC8**	**PC9**	**PC10**
Standard deviation	0.0366	0.0331	0.0298	0.0282	0.0274
Proportion of Variance	0.0405	0.0332	0.0269	0.0241	0.0227
Cumulative Proportion	0.7837	0.8170	0.8438	0.8680	0.8907
	PC11	**PC12**	**PC13**	**PC14**	**PC15**
Standard deviation	0.0257	0.0231	0.0218	0.0201	0.0194
Proportion of Variance	0.0200	0.0162	0.0144	0.0122	0.0113
Cumulative Proportion	0.9106	0.9268	0.9412	0.9535	0.9648
	PC16	**PC17**	**PC18**	**PC19**	**PC20**
Standard deviation	0.0172	0.0162	0.0158	0.0142	0.0125
Proportion of Variance	0.0090	0.0079	0.0075	0.0061	0.0047
Cumulative Proportion	0.9737	0.9817	0.9892	0.9953	1.0000

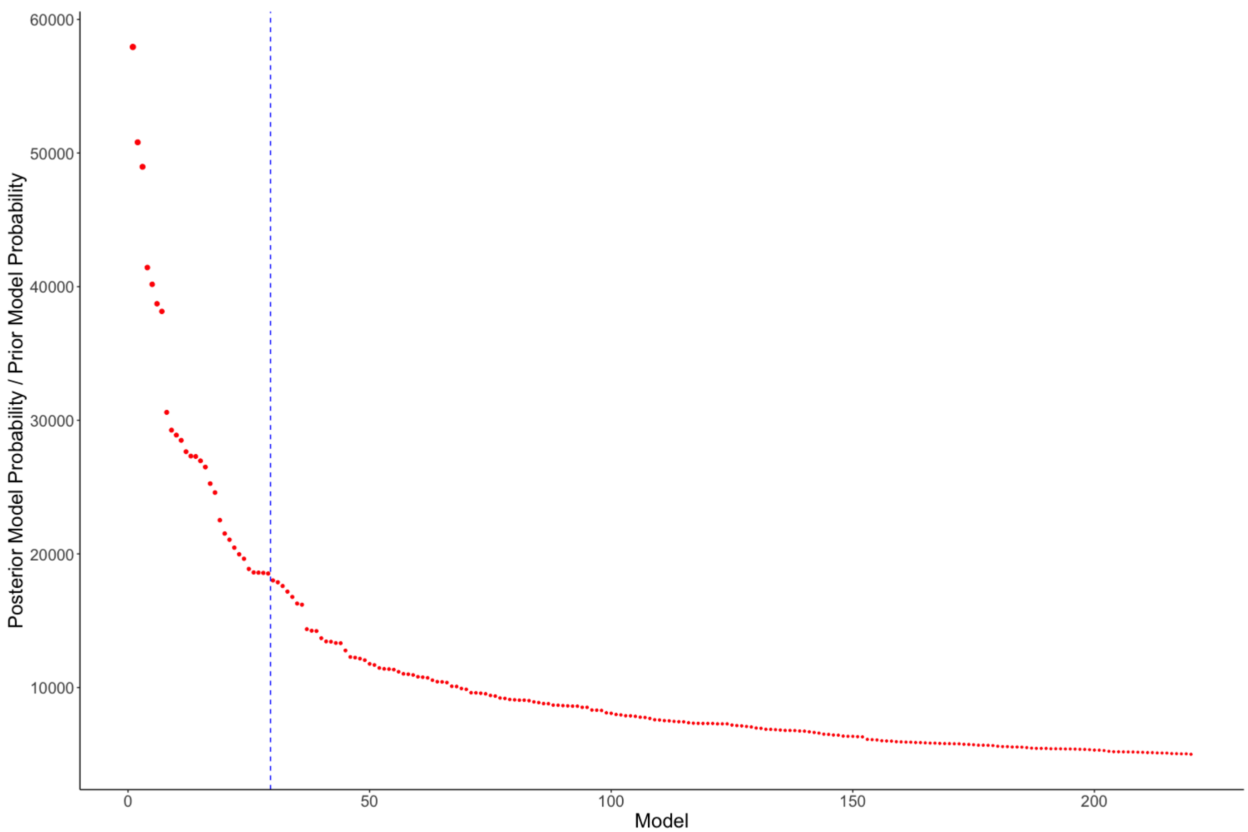

Figure 4 Top 220 Models of the Winners Plus Genuine Anomalies Model Scan

Table 7 Marginal Likelihoods and Ratios of Posterior Model Probability and Prior Model Probability of Selected Models in the Winners Plus Genuine Anomalies Model Space $\tilde{\mathscr{M}}$

Risk Factors	$\log m_j(\tilde{\boldsymbol{y}}_{1:T}\|\tilde{\mathbb{M}}_j)$	$\frac{\Pr(\tilde{\boldsymbol{y}}_{1:T}\|\tilde{\mathbb{M}}_j)}{\Pr(\tilde{\mathbb{M}}_j)}$
Panel A: Top models in $\tilde{\mathscr{M}}_*$		
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	24621.85	57939.28
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC7	24621.72	50803.12
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	24621.68	48969.29
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5	24621.51	41429.51
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC3, PC4, PC5, PC7	24621.48	40171.69
Mkt, RMW, CMA, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	24621.44	38720.10
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5	24621.43	38145.58
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC3, PC4, PC5	24621.21	30594.32
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC4, PC5	24621.16	29262.14
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC7, PC9	24621.15	28892.28
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC4, PC5, PC7	24621.14	28496.78
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC4, PC5, PC7	24621.11	27648.27
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC4, PC5	24621.10	27321.30
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	24621.09	27292.03
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5	24621.08	26966.60
Mkt, MOM, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	24621.07	26499.81
Mkt, RMW, CMA, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5	24621.02	25260.55
Mkt, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	24620.99	24590.60
Mkt, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	24620.90	22526.00
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7, PC9	24620.86	21524.20
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC9	24620.84	21061.93
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC7	24620.81	20467.26
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, PC1, PC4, PC5, PC7	24620.78	19969.41
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7, PC9	24620.77	19636.23
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, PC1, PC4, PC5	24620.73	18874.75
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5	24620.71	18612.73
Mkt, CMA, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	24620.71	18597.04
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, PC1, PC3, PC4, PC5, PC7	24620.71	18577.15
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5	24620.71	18531.14
Panel B: $\tilde{\mathbb{M}}_1$ and models with the same risk factor sets as CAPM, FF3, FF5, FF6, HXZ, SY, and DHS		
12 winners and 11 PCs	24606.63	1.71×10^{-9}
Mkt	24560.93	2.42×10^{-29}
Mkt, SMB, HML	24560.17	1.13×10^{-29}
Mkt, SMB, HML, RMW, CMA	24572.01	1.57×10^{-24}
Mkt, SMB, HML, RMW, CMA, MOM	24573.47	6.74×10^{-24}
Mkt, SMB, IA, ROE	24583.58	1.66×10^{-19}
Mkt, SMB, MGMT, PERF	24592.21	9.25×10^{-16}
Mkt, PEAD, FIN	24597.68	2.20×10^{-13}
Panel C: Models with the same risk factor sets as the top three models in the benchmark model scan		
Mkt, SMB, MOM, ROE, MGMT, PEAD	24604.69	2.44×10^{-10}
Mkt, SMB, MOM, ROE, MGMT, PERF, PEAD	24604.41	1.84×10^{-10}
Mkt, SMB, ROE, MGMT, PEAD	24604.60	2.22×10^{-10}

Table 8 Prior and Posterior Statistics of the Risk Premia Parameters of the Best of the Winners Plus Genuine Model Scan $\tilde{\mathbb{M}}_{1*}$

Prior and posterior statistics, i.e. prior mean, posterior mean, posterior standard deviation, posterior median, 2.5% quantile, and 97.5% quantile for the risk premia $\tilde{\boldsymbol{\lambda}}_{x,1*}$ of the best model $\mathbb{M}_1^*$, which has Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, and PC7 as risk factors.

	Prior mean	Posterior mean	Posterior sd	Posterior median	2.5% Quantile	97.5% Quantile
Mkt	0.0017	0.0066	0.0021	0.0066	0.0025	0.0106
RMW	-0.0015	0.0036	0.0011	0.0036	0.0014	0.0057
MOM	0.0044	0.0061	0.0021	0.0061	0.0020	0.0102
IA	0.0075	0.0033	0.0009	0.0033	0.0016	0.0050
ROE	0.0041	0.0058	0.0012	0.0058	0.0034	0.0081
MGMT	0.0084	0.0056	0.0013	0.0056	0.0030	0.0082
PEAD	0.0085	0.0056	0.0009	0.0056	0.0039	0.0074
FIN	0.0077	0.0070	0.0018	0.0070	0.0034	0.0106
PC1	0.0044	-0.0006	0.0056	-0.0006	-0.0117	0.0103
PC3	0.0191	0.0082	0.0020	0.0082	0.0043	0.0121
PC4	0.0221	0.0112	0.0018	0.0112	0.0076	0.0147
PC5	0.0009	0.0013	0.0019	0.0013	-0.0024	0.0051
PC7	-0.0005	-0.0011	0.0015	-0.0011	-0.0040	0.0018

Table 9 Predictive Likelihoods for the Winners Plus Genuine Anomalies Model Scan

Predictive likelihoods of selected models in the winners plus genuine anomalies model scan.

Risk factors	$\log m_j(\tilde{\boldsymbol{y}}_{(T+1):(T+12)}\|\tilde{\mathbb{M}}_j)$
Panel A: Top models in $\tilde{\mathscr{M}}_*$	
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	639.27
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC7	639.36
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	639.46
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5	639.01
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC3, PC4, PC5, PC7	639.86
Mkt, RMW, CMA, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	639.09
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5	639.08
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC3, PC4, PC5	639.59
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC4, PC5	639.36
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC7, PC9	639.32
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC4, PC5, PC7	639.65
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC4, PC5, PC7	639.16
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC4, PC5	638.86
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	639.35
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5	639.23
Mkt, MOM, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	639.53
Mkt, RMW, CMA, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5	638.82
Mkt, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	639.48
Mkt, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	639.41
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7, PC9	639.43
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC9	639.08
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC7	639.25
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, PC1, PC4, PC5, PC7	639.71
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7, PC9	639.19
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, PC1, PC4, PC5	639.42
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5	638.95
Mkt, CMA, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	639.26
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, PC1, PC3, PC4, PC5, PC7	639.90
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5	639.09
Panel B: $\tilde{\mathbb{M}}_1$ and models with risk factor sets same as CAPM, FF3, FF5, FF6, HXZ, SY, and DHS	
12 winners and 11 PCs	638.61
Mkt	640.03
Mkt, SMB, HML	639.35
Mkt, SMB, HML, RMW, CMA	639.27
Mkt, SMB, HML, CMA, RMW, MOM	639.43
Mkt, SMB, IA, ROE	639.48
Mkt, SMB, MGMT, PERF	640.48
Mkt, FIN, PEAD	640.35

Table 10 Sharpe Ratios

Sharpe ratios for the risk factor sets of selected asset pricing models based on $G = 100,000$.

Risk factors	Sharpe ratios
Panel A: Risk factor sets of the top models in $\tilde{\mathscr{M}}_*$	
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	0.69
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC7	0.68
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	0.69
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5	0.68
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC3, PC4, PC5, PC7	0.68
Mkt, RMW, CMA, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	0.69
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5	0.67
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC3, PC4, PC5	0.67
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC4, PC5	0.66
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC7, PC9	0.69
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, PC1, PC4, PC5, PC7	0.67
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC4, PC5, PC7	0.68
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC4, PC5	0.67
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	0.70
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5	0.68
Mkt, MOM, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	0.68
Mkt, RMW, CMA, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5	0.68
Mkt, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	0.67
Mkt, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	0.68
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5, PC7, PC9	0.70
Mkt, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC9	0.68
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5, PC7	0.69
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, PC1, PC4, PC5, PC7	0.68
Mkt, RMW, MOM, IA, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7, PC9	0.70
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, PC1, PC4, PC5	0.67
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC4, PC5	0.68
Mkt, CMA, MOM, ROE, MGMT, PEAD, FIN, PC1, PC3, PC4, PC5, PC7	0.68
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, PC1, PC3, PC4, PC5, PC7	0.68
Mkt, RMW, MOM, IA, ROE, MGMT, PERF, PEAD, FIN, PC1, PC3, PC4, PC5	0.69
Panel B: Risk factor sets of the top three models in the benchmark model scan	
Mkt, SMB, MOM, ROE, MGMT, PEAD	0.54
Mkt, SMB, MOM, ROE, MGMT, PERF, PEAD	0.55
Mkt, SMB, ROE, MGMT, PEAD	0.53
Panel C: Risk factor sets of $\tilde{\mathbb{M}}_1$, $\mathbb{M}_1$, CAPM, FF3, FF5, FF6, HXZ, SY, and DHS models	
12 winners and 11 PCs	0.71
12 winners	0.56
Mkt	0.15
Mkt, SMB, HML	0.20
Mkt, SMB, HML, RMW, CMA	0.34
Mkt, SMB, HML, CMA, RMW, MOM	0.36
Mkt, SMB, IA, ROE	0.40
Mkt, SMB, MGMT, PERF	0.45
Mkt, FIN, PEAD	0.47

References

Ang, A., and A. Timmermann. 2012. Regime Changes and Financial Markets. *Annu. Rev. Financ. Econ.* 4:313–337.

Avramov, D. 2002. Stock return predictability and model uncertainty. *Journal of Financial Economics* 64:423–458.

Avramov, D., and J. C. Chao. 2006. An Exact Bayes Test of Asset Pricing Models with Application to International Markets. *Journal of Business* 79:293–324.

Barillas, F., and J. Shanken. 2018. Comparing Asset Pricing Models. *Journal of Finance* 73:715–754.

Bryzgalova, S., J. Huang, and C. Julliard. 2019. Bayesian Solutions for the Factor Zoo: We Just Ran Two Quadrillion Models. *Available at* SSRN.

Chib, S. 1995. Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association* 90:1313–1321.

Chib, S., M. Shin, and A. Simoni. 2018. Bayesian Estimation and Comparison of Moment Condition Models. *Journal of the American Statistical Association* 113:1656–1668.

Chib, S., and X. Zeng. 2020. Which Factors are Risk Factors in Asset Pricing? A Model Scan Framework. *Journal of Business & Economic Statistics* 38:771–783.

Chib, S., X. Zeng, and L. Zhao. 2020. On Comparing Asset Pricing Models. *Journal of Finance* 75:551–577.

Cochrane, J. H. 2011. Presidential Address: Discount Rates. *Journal of Finance* 66:1047–1108.

Daniel, K., D. Hirshleifer, and L. Sun. 2020. Short-and Long-Horizon Behavioral Factors. *Review of Financial Studies* 33:1673–1736.

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

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

Fama, E. F., and K. R. French. 2018. Choosing factors. *Journal of Financial Economics* 128:234–252.

Goyal, A., Z. L. He, and S.-W. Huh. 2018. Distance-Based Metrics: A Bayesian Solution to the Power and Extreme-Error Problems in Asset-Pricing Tests. *Swiss Finance Institute Research Paper* .

Green, J., J. R. M. Hand, and X. F. Zhang. 2017. The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *Review of Financial Studies* 30:4389–4436.

Hansen, L. P., and R. Jagannathan. 1991. Implications of Security Market Data for Models of Dynamic Economies. *Journal of Political Economy* 99:225–262.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. ... and the Cross-Section of Expected Returns. *Review of Financial Studies* 29:5–68.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting Anomalies: An Investment Approach. *Review of Financial Studies* 28:650–705.

Hou, K., C. Xue, and L. Zhang. 2017. Replicating Anomalies. *Review of Financial Studies* forthcoming.

Jeffreys, H. 1961. Theory of Probability. 3rd Edn Oxford: Oxford University Press.

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

Novy-Marx, R., and M. Velikov. 2016. A Taxonomy of Anomalies and Their Trading Costs. *Review of Financial Studies* 29:104–147.

Pástor, L. 2000. Portfolio Selection and Asset Pricing Models. *Journal of Finance* 55:179–223.

Pástor, L., and R. F. Stambaugh. 2000. Comparing asset pricing models: an investment perspective. *Journal of Financial Economics* 56:335–381.

Pástor, L., and R. F. Stambaugh. 2001. The Equity Premium and Structural Breaks. *Journal of Finance* 56:1207–1239.

Patton, A. J., and B. M. Weller. 2020. What you see is not what you get: The costs of trading market anomalies. *Journal of Financial Economics* forthcoming.

Smith, S. C., and A. Timmermann. 2021. Break risk. *The Review of Financial Studies* 34:2045–2100.

Stambaugh, R. F., and Y. Yuan. 2017. Mispricing Factors. *Review of Financial Studies* 30:1270–1315.