# Sparse Structure of Stochastic Discount Factor in the Chinese Stock Market: A Bayesian Interpretable Machine-learning Approach *

Yaohan Chen

School of Economics, Singapore Management University

**Abstract**

This paper reviews a Bayesian interpretable machine-learning method proposed by Kozak, Nagel, and Santosh (2020). We show how the method can link two strands of literature, namely the literature on empirical asset pricing and the literature on statistical learning. Based on a recently developed data-cleaning technique, we obtain 123 financial and accounting cross-sectional equity characteristics in the Chinese stock market. When applying the method of Kozak, Nagel, and Santosh (2020) to the Chinese stock market, we find that it is futile to summarize the stochastic discount factor (SDF) in the Chinese stock market as the exposure of several dominant cross-sectional equity characteristics in-sample. A cross-validated out-of-sample analysis further supports this finding.

**Keywords:** Asset pricing; Bayesian; Stochastic discount factor; Machine-learning; Chinese stock market

**JEL Classification:** C02, C55, C80, G12

* I am grateful for the discussion with Tao Zeng and Xiaobin Liu. All errors are my own. You may contact me at yaohan.chen.2017@phdecons.smu.edu.sg.

# 1 Introduction

> *One of our central themes is that if assets are priced rationally, variables that are related to average returns, such as size and book-to-market equity, must proxy for sensitivity to common (shared and thus and undiversifiable) risk factors in returns.*
>
> Fama and French (1993)

> *We have a lot of questions to answer: First, which characteristics really provide independent information about average returns? Which are subsumed by others? Second, does each new anomaly variable also correspond to a new factor formed on those same anomalies? ... Third, how many of these new factors are really important?*
>
> Cochrane (2011)

As the two quotes cited above suggest, a formidable challenge faced within the community of financial researchers is how to handle the high-dimensionality in the potential predictors for the expected return. There are at least two difficulties associated with high-dimensionality in the potential predictors. First, whether or not there exists a sparse exposure structure of stochastic discount factor (SDF) is difficult to know. Second, what should be a reasonable functional relationship between the expected return and intrinsically useful predictors.

The first difficulty has attracted a great deal of attentions in recent years, given that a huge number of firm-level characteristics have been proposed to be the predictors in the literature. Many studies rely on the $p$-value of the standard test statistics (such as the t statistic) as the evidence to support or be against the use of firm-level characteristics. However, Harvey, Liu, and Zhu (2016) points out the so-called $p$-hacking issue in the conventional statistical test. They further propose an adjusted $p$-value to check the statistical evidence of the usefulness of firm-level characteristics.

To deal with the second difficulty, one way is to allow for nonlinear relationships between the expected return and predictors in the model specification. This is the exact reason why nonparametric methods have becomes increasingly popular in this literature. With the development of modern computational power and statistical algorithms, some advanced nonparametric methods have been proposed (see Freyberger, Neuhierl, and Weber, 2020). Machine-learning methods are one of the popular nonparametric techniques.

Studies that employ machine-learning methods to study return predictability can be divided into three groups. The first group of studies aims to use and design machine-learning methods to generate good out-of-sample performance. These methods usually are flexible given the generic nonparametric feature in the methods. Gu, Kelly, and Xiu (2020) compare many machine-learning methods in terms of their predictive power of the U.S. equity returns. It is found that neural network and regression trees perform relatively well. Other studies that use machine learning method to analyse cross-sectional returns include but not restricted to (Freyberger, Neuhierl, and Weber, 2020;

Chinco, Clark-Joseph, and Ye, 2019; Han, He, Rapach, and Zhou, 2019; Chen, Pelger, and Zhu, 2019).

The second group of studies assume that there exists a factor structure in the potentially useful predictors. The number of factors is usually much lower than the number of available characteristics. This approach has been an important part of the literature ever since the seminar works of Fama and French (1992, 1993, 1996). Generally there are two alternative ways to introduce a factor structure in this rather extensive literature. The first one uses pre-specified and observed factors based on the prior knowledge about the cross-sectional accounting information. Many factors have been established for explaining the cross-sectional variations associated with asset returns; see, for example, Fama and French (1993), Fama and French (2015), Hou, Xue, and Zhang (2015). More references can be found in two recent excellent surveys, that is, Hou, Xue, and Zhang (2018) and Chen and Zimmermann (2020). The second one assumes that factors are latent variables. In this case, statistical factor analysis techniques, such as principal component analysis (PCA), are used to extract factors and factor loadings simultaneously. Studies of this kind can be traced back at least to Connor and Korajczyk (1986) and Chamberlain and Rothschild (1983). Recently, the latent factor approach has been employed in Fan, Liao, and Wang (2016); Kozak, Nagel, and Santosh (2018); Kelly, Pruitt, and Su (2019); Kozak (2020); Lettu and Pelger (2020a,b) to study stock returns.

The third group of studies focuses directly on addressing the high-dimension problem. These studies use model selection and variable selection techniques to select useful firm-level characteristics. Since many machine-learning methodologies inherit ideas from statistical theory (Vapnik, 1998; Hastie, Tibshirani, and Friedman, 2001; Catoni, 2004, 2007), some machine-learning methodologies designed for handling the high-dimension problem are essentially statistical learning methods. Among all the methods of this type, the most representative ones are LASSO, ridge regression and elastic net. These are also the major statistical learning methods mostly applied in economics and finance literature: Rapach, Strauss, and Zhou (2010), Messmer and Audrino (2017), Giannone, Lenza, and Primiceri (2021) and Bakalli, Guerrier, and Scaillet (2021) discuss how LASSO is applied in selecting useful predictors for making predictions in economics and finance when there are a slew of predictors; Gabauer, Gupta, Marfatia, and Miller (2020) establish a high-dimensional vector autoregressive model with $L^2$-penalty (i.e. it essentially belongs to ridge regression) for estimating price network connectedness in the U.S. housing market; Kim and Swanson (2014) provide empirical evidence of how elastic net is applied in the high-dimension problem setting for forecasting financial and macroeconomic variables. Huang, Li, and Wang (2021) apply elastic net as one intermediate step for aggregating cross-sectional information of equities to construct the disagreement index in the U.S market. Bali, Goyal, Huang, Jiang, and Wen (2021) discuss the application of elastic net in addressing bond return predictability in the setting where the individual bond is cross-sectionally exposed to the high-dimensional information vector. It is known in literature (Zou and Hastie, 2005) that in comparison to LASSO as the dimension reduction method for variable selection, elastic net combines LASSO and ridge penalties and produces a model with more flexibilities and good out-of-sample prediction accuracy. A useful addition to this group of studies is Linero (2018) where a

Bayesian additive regression trees (BART) method is found to perform well. In the BART method of Linero (2018), a sparsity-inducing Dirichlet hyper-prior is used to solve the high-dimension problem. This method is empirically successful in selecting relevant variables that can yield good out-of-sample prediction accuracy, and is later theoretically justified by Ročková (2019), Ročková and Saha (2019), and Ročková and van der Pas (2020). However, it naturally inherits the major disadvantage of Chipman, George, and McCulloch (2010), which is computationally heavy in comparison to LASSO, ridge regression, or elastic net, mainly due to its underlying MCMC sampling scheme. Besides, for all the existing machine-learning methodologies, rarely is there any discussion on whether or not these methodologies can be interpreted through the lens of existing economic theories.

Another Bayesian method to address the high-dimension problem is the Bayesian interpretable machine-learning method proposed in Kozak, Nagel, and Santosh (2020). There are a number of good features in this methods. First, the modelling framework is parsimonious but still powerful in characterizing the key asset-pricing structure. Second, it reconciles well with economic theory through the Bayesian lens (that is the reason why it is referred to as a Bayesian interpretable method) as well as with the statistical learning theory (which facilitates implementation and computation). Basically, by imposing an economically motivated prior on SDF, it is possible to show how the machine-learning methods (specifically, the penalized regression such as the ridge regression with the objective function being the Hansen-Jagannathan distance or the elastic net method with dual penalty) are related to the SDF-based asset pricing theory. Because of these attractive features, in this paper, we apply it to analyze the returns of the Chinese stock market. In particular, we use the method to check whether or not there exists a sparse exposure structure of SDF to several dominant cross-sectional equity characteristics in the Chinese stock market.

The rest of this paper is structured as follows: In Section 2, we review the theoretical modelling framework for the SDF-based linear asset pricing theory and explain how the economic theory is related to some of the machine-learning methods so that the machine-learning methods are interpretable through the Bayesian lens. In Section 3, we discuss how cross-sectional anomaly variables (or equivalently firm-level characteristics) are constructed. In Section 4, we report the main empirical findings. Finally Section 5 concludes this paper.

## 2 Basic Modelling Framework

### 2.1 SDF and cross-sectional asset pricing

In much of the finance literature, the central goal is to explain the differences in returns in the cross-sectional dimension. Specifically for individual stock, denote $R_{t+1,i}$ as the return of asset $i$ at $t+1$. The fundamental no-arbitrage condition is closely related to the existence of SDF derived from the first-order condition of the Euler equation. That is, for any return in excess of the risk-free rate $R^e_{t+1,i} = R_{t+1,i} - R^f_{t+1}$, the following key pricing formula (conditional) holds

$$\mathbb{E}_t \left[ M_{t+1} R^e_{t+1,i} \right] = 0. \tag{1}$$

Following the convention in the literature (see Hansen and Jagannanthan, 1991; Haugen and Baker, 1996) and without loss of generality, we can assume that the SDF is of a linear functional form as

$$M_{t+1} = 1 - \omega_t^\top \left(R_{t+1}^e - \mathbb{E}_t R_{t+1}^e\right),$$

where $\omega_t$ is a $N \times 1$ vector of SDF coefficients with $N$ being the number of firms cross-sectionally.[1] This specification implies that we normalize the excess return by the corresponding conditional mean, $\mathbb{E}_t R_{t+1}^e$.

To see how it is connected with the factor-modeling framework, considering the following construction,

$$\omega_t = Z_t \omega, \tag{2}$$

where $Z_t$ is an $N \times L$ matrix of asset characteristics and $\omega$ is an $L \times 1$ vector of time-invariant coefficients. Usually the entries of matrix $Z_t$ in (2) collects the information of firm-level characteristics (specifically, each row $i$ of $Z_t$ collects the characteristic information of firm $i$ at time $t$). As documented in empirical asset pricing literature, usually researchers search for new measurable asset characteristics that approximately span $\omega_t$. For example, Fama and French (1993) use two characteristics, market capitalization and the book-to-market equity ratio. Similarly, by plugging this equation into the fundamental pricing equation (1), we have

$$\begin{aligned} M_{t+1} &= 1 - \omega_t^\top \left(R_{t+1}^e - \mathbb{E}_t R_{t+1}^e\right) \\ &= 1 - \omega^\top Z_t^\top \left(R_{t+1}^e - \mathbb{E}_t R_{t+1}^e\right). \end{aligned}$$

We can then define $L$ multi-factors as

$$F_{t+1} = Z_t^\top R_{t+1}^e \tag{3}$$

which simply leads to the normalized representation of SDF as following

$$\begin{aligned} M_{t+1} &= 1 - \omega^\top \left(F_{t+1} - Z_t^\top \mathbb{E}_t R_{t+1}^e\right) \\ &= 1 - \omega^\top \left(F_{t+1} - \mathbb{E}_t F_{t+1}\right). \end{aligned} \tag{4}$$

Note that $F_{t+1}$ is essentially assets in a portfolio form. Hence, it is possible to plug it into the key pricing formula as in (1). Without loss of generality we can replace the conditional mean of factors, $\mathbb{E}_t F_{t+1}$ with the unconditional mean $\mathbb{E} F_{t+1}$ (i.e. $M_{t+1} = 1 - \omega^\top \left(F_{t+1} - \mathbb{E} F_{t+1}\right)$). We can then have following unconditional pricing formula for the managed portfolios,

$$\mathbb{E}_t \left[M_{t+1} F_{t+1}^\top\right] = 0 \quad \Rightarrow \quad \mathbb{E}\left[M_{t+1} F_{t+1}^\top\right] = 0,$$

[1] As pointed in Kozak, Nagel, and Santosh (2018), the ground for a linear factor-based representation of SDF is essentially the law of one prices (LOP). As long as LOP holds, the factors used to represent SDF are a linear combination of asset payoffs.

which implies that

$$\begin{aligned} & \mathbb{E}F_{t+1}^{\top} - \omega^{\top}\mathbb{E}\left[\left(F_{t+1} - \mathbb{E}F_{t+1}\right)F_{t+1}^{\top}\right] \\ = \; & \mathbb{E}F_{t+1}^{\top} - \omega^{\top}\mathbb{E}\left[\left(F_{t+1} - \mathbb{E}F_{t+1}\right)\left(F_{t+1} - \mathbb{E}F_{t+1}\right)^{\top}\right] \; = \; 0. \end{aligned}$$

Hence we have

$$\mathbb{E}F_{t+1} \; = \; \mathbb{E}\left[\left(F_{t+1} - \mathbb{E}F_{t+1}\right)\left(F_{t+1} - \mathbb{E}F_{t+1}\right)^{\top}\right]\omega. \tag{5}$$

This constant specification imposed on the managed portfolio processes implicitly suggests that we focus on unconditional asset pricing. It brings convenience using the corresponding sample moment over the time-series dimension to estimate $\mathbb{E}F_{t+1}$ and $\mathbb{E}\left[\left(F_{t+1} - \mathbb{E}F_{t+1}\right)\left(F_{t+1} - \mathbb{E}F_{t+1}\right)^{\top}\right]$, denoted by $\mu$ (managed portfolio's time-series mean) and $\Sigma$ (variance-covariance matrix), respectively, for the following discussion. It will be seen in the following discussion that, the time-series analogue of the managed portfolios $\bar{\mu}$ and $\bar{\Sigma}$ can be regarded as the data used for constructing the posterior to update the prior information. This constant specification reconciles well with the empirical Bayes logic. See the corresponding discussion in **Remark** 2.3.

## 2.2 Interpretation from a Bayesian perspective

In this section, we discuss how SDF is connected with penalized cross-sectional regression from a Bayesian perspective. The discussions follow the main ideas from Kozak, Nagel, and Santosh (2020) but are more detailed than those in Kozak, Nagel, and Santosh (2020). Essentially the Bayesian prior structure is imposed on $\mu$ as follows (assuming $\Sigma$ is known, and we will discuss how to obtain $\Sigma$ in **Remark** 2.3).

$$\mu \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau}\Sigma^{\eta}\right), \tag{6}$$

where

$$\tau = \mathrm{Tr}\left[\Sigma\right],$$

and $\kappa$, $\eta$ are tuning parameters to be discussed later.

**Remark 2.1** *$\mu$ is $L \times 1$ vector that collects the expected return of each managed portfolio over the time-series dimension. The cross-sectional heterogeneity is captured by the prior (6). Thus, the prior captures investors' ex-ante belief about the expected return of individual managed portfolio. Integrating $\mu$ out of $\mu^{\top}\Sigma^{-1}\mu$ (the squared Sharpe ratio) (i.e., integrating out the ex-ante uncertainty*

*associated with $\mu$) yields the root expected Sharpe ratio under the prior distribution,*

$$\begin{aligned}
&\mathbb{E}\left[\mu^\top \Sigma^{-1} \mu\right]^{1/2} \\
=\ &\mathbb{E}\left[\Sigma^{-1} \operatorname{Tr}\left(\mu\mu^\top\right)\right]^{1/2} \\
=\ &\left\{\Sigma^{-1} \operatorname{Tr}\left(\mathbb{E}\left[\mu\mu^\top\right]\right)\right\}^{1/2} \\
=\ &\operatorname{Tr}\left(\Sigma^{-1} \frac{\kappa^2}{\tau} \Sigma^\eta\right)^{1/2} \\
=\ &\left\{\frac{\kappa^2}{\tau} \operatorname{Tr}\left(\Sigma^{\eta-1}\right)\right\}^{1/2}.
\end{aligned}$$

*It is $\kappa$ if $\eta = 2$ so that we can use $\kappa$ to capture investors' belief about the root expected Sharpe ratio of the managed portfolios.*

Given the previous discussion, the prior imposed on $\mu$ as in (6) also implies that the prior information for $\omega$ should be

$$\omega = \Sigma^{-1}\mu \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau} \mathrm{I}_L\right), \qquad \text{with } \eta = 2,$$

where $\mathrm{I}_L$ refers to an identity matrix of dimension $L$. The matrix representation is

$$\underset{L\times 1}{F_t} = \underset{L\times 1}{\mu} + \underset{L\times 1}{\varepsilon}, \qquad \varepsilon \sim (0, \Sigma). \tag{7}$$

or equivalently in the stacked matrix form

$$\underset{LT\times 1}{f} = \begin{pmatrix} F_1 \\ \vdots \\ F_T \end{pmatrix} = \underbrace{(\mathbf{1}_T \otimes \mathrm{I}_L)}_{X} \underset{L\times 1}{\mu} + \underset{LT\times 1}{\tilde{\varepsilon}}, \qquad \tilde{\varepsilon} \sim (0, \Xi), \tag{8}$$

$$\Xi = \mathrm{I}_T \otimes \Sigma.$$

The structure of $\tilde{\Sigma}$ implies that there is no time-series correlation. Recall the usual conjugate posterior for $\mu$ under the linear model framework, denoted by $\hat{\mu}$, is

$$\hat{\mu} = \left(\Xi_0^{-1} + X^\top \Xi^{-1} X\right)^{-1} \left(\Xi_0^{-1}\mu_0 + X^\top \Xi^{-1} f\right).$$

In the case for (8), by construction we have

$$\mu_0 = 0, \quad \Xi_0 = \frac{\kappa^2}{\tau}\Sigma^\eta, \quad X = \mathbf{1}_T \otimes \mathrm{I}_L.$$

Hence

$$\hat{\mu} \;=\; \left(\Xi_0^{-1} + X^\top \Xi^{-1} X\right)^{-1} X^\top \Xi^{-1} f.$$

Note that

$$\begin{aligned}
X^\top \Xi^{-1} X &= (\mathbf{1}_T \otimes \mathrm{I}_L)^\top \tilde{\Sigma}^{-1} (\mathbf{1}_T \otimes \mathrm{I}_L) \\
&= (\mathbf{1}_T \otimes \mathrm{I}_L)^\top (\mathrm{I}_T \otimes \Sigma)^{-1} (\mathbf{1}_T \otimes \mathrm{I}_L) \\
&= \left(\mathbf{1}_T^\top \otimes \mathrm{I}_L\right) \left(\mathrm{I}_T \otimes \Sigma^{-1}\right) (\mathbf{1}_T \otimes \mathrm{I}_L) \\
&= \left(\mathbf{1}_T^\top \otimes \Sigma^{-1}\right) (\mathbf{1}_T \otimes \mathrm{I}_L) \\
&= \mathbf{1}_T^\top \mathbf{1}_T \otimes \Sigma^{-1} = T\Sigma^{-1}
\end{aligned}$$

$$\begin{aligned}
X^\top \Xi^{-1} f &= (\mathbf{1}_T \otimes \mathrm{I}_L)^\top (\mathrm{I}_T \otimes \Sigma)^{-1} f \\
&= \left(\mathbf{1}_T^\top \otimes \mathrm{I}_L\right) \left(\mathrm{I}_T \otimes \Sigma^{-1}\right) f \\
&= \left(\mathbf{1}_T^\top \otimes \Sigma^{-1}\right) f \\
&= \mathrm{vec}\left(\Sigma^{-1} \tilde{f} \mathbf{1}_T\right),
\end{aligned}$$

where

$$f = \mathrm{vec}\left(\tilde{f}\right), \qquad \tilde{f} = \begin{pmatrix} F_1 & \cdots & F_T \end{pmatrix}_{L \times T}, \qquad \tilde{f}\mathbf{1}_T = T\bar{\mu}.$$

Thus,

$$X^\top \Xi^{-1} f = \mathrm{vec}\left(\Sigma^{-1} T \bar{\mu}\right) = \Sigma^{-1} T \bar{\mu}.$$

Finally

$$\hat{\mu} = \left(\Xi_0^{-1} + T\Sigma^{-1}\right)^{-1} T\Sigma^{-1} \bar{\mu}.$$

Consequently,

$$\begin{aligned}
\hat{\omega} &= \Sigma^{-1}\hat{\mu} \\
&= \Sigma^{-1}\left(\Xi_0^{-1} + T\Sigma^{-1}\right)^{-1} T\Sigma^{-1}\bar{\mu} \\
&= \left[T^{-1}\Sigma\left(\Xi_0^{-1} + T\Sigma^{-1}\right)\Sigma\right]^{-1}\bar{\mu} \\
&= \left[T^{-1}\Sigma\frac{\tau}{\kappa^2}\Sigma^{-\eta}\Sigma + \Sigma\right]^{-1}\bar{\mu} \\
&= \left[T^{-1}\frac{\tau}{\kappa^2}\Sigma^{2-\eta} + \Sigma\right]^{-1}\bar{\mu} \\
&= \left[\frac{\tau}{T\kappa^2}\Sigma^{2-\eta} + \Sigma\right]^{-1}\bar{\mu}. \qquad (9)
\end{aligned}$$

If $\eta = 2$, we have

$$\hat{\omega} = \left(\gamma \mathrm{I}_L + \Sigma\right)^{-1}\bar{\mu} \qquad \gamma = \frac{\tau}{T\kappa^2}. \qquad (10)$$

Similarly, the posterior covariance of $\hat{\mu}$ is

$$\mathrm{Var}\left(\hat{\mu}\right) = \left(\Xi_0^{-1} + X^\top\Xi^{-1}X\right)^{-1} = \left(\frac{\kappa}{\tau^2}\Sigma^{-\eta} + T\Sigma^{-1}\right)^{-1}.$$

The posterior covariance matrix can expressed as

$$\begin{aligned}
\mathrm{Var}\left(\hat{\omega}\right) &= \Sigma^{-1}\left(\frac{\tau}{\kappa^2}\Sigma^{-\eta} + T\Sigma^{-1}\right)^{-1}\Sigma^{-1} \\
&= \left[\Sigma\left(\frac{\tau}{\kappa^2}\Sigma^{-\eta} + T\Sigma^{-1}\right)\Sigma\right]^{-1} \\
&= \left[\left(\frac{\tau}{\kappa^2}\Sigma^{2-\eta} + T\Sigma\right)\right]^{-1} = \frac{1}{T}\left[\frac{\tau}{T\kappa^2}\Sigma^{2-\eta} + \Sigma\right]^{-1}.
\end{aligned}$$

Since $\eta = 2$, we have

$$\mathrm{Var}\left(\hat{\omega}\right) = \frac{1}{T}\left(\gamma \mathrm{I}_L + \Sigma\right)^{-1}, \qquad (11)$$

where $\mathrm{Var}\left(\hat{\omega}\right)$ can be used to construct the confidence interval or $t$-statistic.

**Remark 2.2 (Connection with penalized estimator)** *The proposed Bayesian estimator is closely related to the penalized estimator. Consider the following cases where each penalized estimator is constructed based on different objective function*

*(i) The objective function is constructed to maximize the cross-sectional $R^2$ with the penalty imposed on the model implied Sharpe ratio,*

$$\mathbb{E}F = \Sigma\omega, \ \textit{and} \ \left(\Sigma\omega\right)^\top\Sigma^{-1}\left(\Sigma\omega\right) = \omega^\top\Sigma\omega$$

*then*

$$\hat{\omega} = \arg\min_{\omega} \left\{ (\bar{\mu} - \Sigma\omega)^{\top} (\bar{\mu} - \Sigma\omega) + \gamma \omega^{\top} \Sigma \omega \right\} \tag{12}$$

*(ii) The objective function is constructed to minimize the HJ distance,*

$$\hat{\omega} = \arg\min_{\omega} \left\{ (\bar{\mu} - \Sigma\omega)^{\top} \Sigma^{-1} (\bar{\mu} - \Sigma\omega) + \gamma \omega^{\top} \omega \right\} \tag{13}$$

*(iii) The objective function is constructed as that in the ridge regression,*

$$\hat{\omega} = \arg\min_{\omega} \left\{ (\bar{\mu} - \Sigma\omega)^{\top} (\bar{\mu} - \Sigma\omega) + \gamma \omega^{\top} \omega \right\} \tag{14}$$

*(i) and (ii) share the same solution and the solution is the same as the case when $\eta = 2$. (iii) is the same as the case $\eta = 3$. This is because the first order condition with respect to $\omega$ in (14) yields*

$$-\Sigma \left( \bar{\mu} - \Sigma\omega \right) + \gamma\omega = \mathbf{0}.$$

*Solving this equation, we have*

$$\hat{\omega} = \left( \Sigma + \gamma \Sigma^{-1} \right)^{-1} \bar{\mu}$$

*which is the case implied by (9) when $\eta = 3$. For (ii), when $\eta = 2$, we can regard (13) as the $L^2$-norm penalized cross-sectional regression with the HJ distance as the objective function (alternatively, it can be understood as the extension of the ridge regression with the objective function being the HJ distance). Consequently, the tuning parameter associated with the $L^2$-norm penalized cross-sectional regression, $\gamma$ is closely related with the root expected Sharpe ratio (under the prior), $\kappa$ (implied from (10)). In this regard, the imposed Bayesian prior structure (6) brings the corresponding economic theory to the tuning procedure.*

**Remark 2.3** *The justification of the Bayesian interpretation of SDF is essentially given by the prior imposed on $\mu$ conditional on the fact that investors update their knowledge about the cross-sectional variance-covariance structure via the observed returns. This maps well to the robust estimator for a relatively large variance-covariance matrix in the literature (Ledoit and Wolf, 2004a,b). This connection can be easily seen from the Wishart prior imposed on the precision matrix (in general, the inverse of variance-covariance matrix, i.e., $P = \Sigma^{-1}$) commonly used for Bayesian analysis. Suppose we have the following prior for the precision matrix*

$$\Sigma^{-1} \sim \mathcal{W}\left( U_0, \varpi_0 \right),$$

*where $U_0$ a $L \times L$ positive definite matrix with $\varpi_0$ degrees of freedoms such that $\varpi_0 > L - 1$. Let $\mathbf{x}$ follow a multivariate normal distribution with mean zero. The conditional density function is given by*

$$p(\mathbf{x} \mid P) = (2\pi)^{-L/2} |P|^{1/2} \exp\left( -\frac{1}{2} \mathbf{x}^{\top} P \mathbf{x} \right).$$

*Since the probability density function of the Wishart distribution is*

$$p(P) = \frac{|P|^{(\varpi_0 - L - 1)/2} \exp\left[-\operatorname{Tr}\left(U_0^{-1} P\right)/2\right]}{2^{\frac{\varpi_0 L}{2}} \Gamma\left(\varpi_0/2\right) |U_0|^{\varpi_0/2}},$$

*the posterior distribution given* $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ *is*

$$\begin{aligned} p(P \mid \mathbf{X}) &\propto p(P)(\mathbf{X} \mid P) \\ &\propto \prod_{i=1}^{T} \left[ |P|^{1/2} \exp\left(-\frac{1}{2} \mathbf{x}_i^\top P \mathbf{x}_i\right)\right] |P|^{(\varpi_0 - L - 1)/2} \exp\left[-\operatorname{tr}\left(U_0^{-1} P\right)/2\right] \\ &= |P|^{(T + \varpi_0 - L - 1)/2} \exp\left\{-\frac{1}{2} \operatorname{Tr}\left[\left(T\mathbf{S} + U_0^{-1}\right) P\right]\right\} \end{aligned}$$

*, where*

$$\mathbf{S} = \frac{1}{T} \sum_{i=1}^{T} \mathbf{x}_i \mathbf{x}_i^\top,$$

*is the sample counterpart of the variance-covariance matrix. This suggests that the posterior distribution of* $P$ *is also a Wishart distribution such that*

$$P \mid \mathbf{X} \sim \mathcal{W}\left(\left(T\mathbf{S} + U_0^{-1}\right)^{-1}, T + \varpi_0\right).$$

*To make it connected to our discussion in the main context, replacing* $U_0$ *and* $\varpi_0$ *with* $\frac{1}{L}\Sigma_0^{-1}$ *and* $L$*, we have*

$$\Sigma^{-1} \sim \mathcal{W}\left(\frac{1}{L}\Sigma_0^{-1}, L\right).$$

*Replacing* $\mathbf{X}$ *with the demeaned return over the time-series dimension and the variance-covariance matrix with* $\mathbf{S} = \Sigma_T$*, we have*

$$\Sigma^{-1} \mid \mathbf{X} \sim \mathcal{W}\left(\left(T\Sigma_T + L\Sigma_0\right)^{-1}, T + L\right).$$

*The expected value (posterior) is*

$$\mathbb{E}\left[\Sigma^{-1} \mid \mathbf{X}\right] = (T + L)\left(T\Sigma_T + L\Sigma_0\right)^{-1} = \left[\left(\frac{L}{T + L}\right)\Sigma_0 + \left(\frac{T}{T + L}\right)\Sigma_T\right]^{-1}.$$

*The typical choice for* $\Sigma_0$ *is* $\Sigma_0 = \frac{1}{L}\operatorname{Tr}\left(\Sigma_T\right)\mathbf{I}_L$ *where* $\mathbf{I}_L$ *is the* $L \times L$ *identity matrix. Consequently, we use*

$$\bar{\Sigma} = \left(\frac{L}{T + L}\right)\Sigma_0 + \left(\frac{T}{T + L}\right)\Sigma_T. \tag{15}$$

*to replace* $\Sigma$ *in all relevant formulas in this paper.*

### 2.3 Dual-penalty in combination of two norms

We discussed a key insight of Kozak, Nagel, and Santosh (2020) in detail, that is, the $L^2$-norm penalty imposed on the cross-sectional regression has a nice Bayesian interpretation, which is grounded on economics theory. However, a more strict shrinkage penalty can also be used. In our empirical analysis, for example, we also consider the following dual $L^1$-$L^2$ penalized cross-sectional regression by adding the following $L^1$-norm penalty term

$$\hat{\omega} = \arg\min_{\omega} (\bar{\mu} - \Sigma\omega)^\top \Sigma^{-1} (\bar{\mu} - \Sigma\omega) + \gamma_2 \omega^\top \omega + \gamma_1 \sum_{i=1}^{L} |\omega_i|. \tag{16}$$

This choice is related to the elastic-net method proposed in Zou and Hastie (2005), with the objective function slightly modified to be the HJ distance. The objective function for cross-validation is the cross-sectional $R^2$ defined by

$$R^2_{\text{oos}} = 1 - \frac{\left(\bar{\mu}_{\text{O}} - \bar{\Sigma}_{\text{O}}\hat{\omega}\right)^\top \left(\bar{\mu}_{\text{O}} - \bar{\Sigma}_{\text{O}}\hat{\omega}\right)}{\bar{\mu}_{\text{O}}^\top \bar{\mu}_{\text{O}}} \tag{17}$$

This is similar to the standard routine in the statistical learning literature where the whole sample is divided into $K$ sub-samples. In each fold of cross-validation, $K-1$ sub-samples are used as training samples to calculate the sample mean and variance-covariance matrix (over time-series dimension), denoted by $\bar{\mu}_{\text{I}}$ and $\bar{\Sigma}_{\text{I}}$, while the remained samples are used as the testing samples to calculate the sample mean and variance-covariance matrix (over time-series dimension), denoted by $\bar{\mu}_{\text{O}}$ and $\bar{\Sigma}_{\text{O}}$. For the penalized cross-sectional regression with the $L^2$-norm penalty, $\omega$ is estimated using (12) or (13). For the penalized cross-sectional regression with the dual $L^1$-$L^2$-norm penalty, $\omega$ is estimated using (16).

## 3 Data

In cross-sectional asset pricing studies, it is important for researchers to carefully construct cross-sectional equity characteristics. In this section, we first briefly discuss the recent literature on constructing cross-sectional equity characteristics for asset pricing studies and explain how we use the existing methods to construct equity characteristics in the Chinese stock market. Then we discuss how characteristic-managed portfolios are constructed based on daily returns of individual assets in the Chinese stock market.

### 3.1 Individual equity characteristic data

Following Harvey and Liu (2014, 2015); Harvey, Liu, and Zhu (2016); Mclean and Pontiff (2016); Green, Hand, and Zhang (2017); Hou, Xue, and Zhang (2018); Gu, Kelly, and Xiu (2020); Demiguel, Martín, Nogales, and Uppal (2020); Freyberger, Neuhierl, and Weber (2020); Kozak, Nagel, and Santosh (2020); Kozak (2020), we obtain firm-level equity characteristic data. Several standard

data-cleaning routines are available in the literature. The method of Chen and Zimmermann (2020) is a successful response to the call for transparency and cooperation (Welch, 2019). Besides, Jensen, Kelly, and Pedersen (2022) provide a more comprehensive analysis by constructing a global dataset in response to the recent discussions on the replication crisis in empirical asset pricing studies.[2] We combine both the data cleaning routines in Chen and Zimmermann (2020) and Jensen, Kelly, and Pedersen (2022) to replicate 123 finance and accounting anomaly variables in the Chinese stock market from 1995 to 2020. All the data (including returns and accounting data) are obtained from the Center for Research in Security Prices (`CRSP`), `Compustat`, and the China Stock Market & Accounting Research (`CSMAR`) database, all of which can be downloaded from the Wharton Research Data Service (WRDS). These anomaly variables are normalized as in Freyberger, Neuhierl, and Weber (2020) so that each characteristic is normalized over the cross-sectional dimension to take a value between 0 and 1. More precisely,

$$rc_{i,t}^{s} = \frac{\operatorname{rank}\left(c_{i,t}^{s}\right)}{n_t + 1}, \tag{18}$$

where $c_{i,t}^{s}$ denotes the originally unscaled firm-level equity characteristic (indexed by superscript $s$) associated with stock $i$ at time $t$ and $n_t$ denotes the total number of individual assets available for observations at time $t$. $\operatorname{rank}(\cdot)$ denotes the cross-sectional ranking order of specific variable. Then, for each rank-transformed characteristic $rc_{s,t}^{i}$, we center it around the cross-sectional mean and divide it by the sum of average deviations from the cross-sectional mean for available stocks. Hence, we have,

$$z_{i,t}^{s} = \frac{\left(rc_{i,t}^{s} - \overline{rc}_{t}^{s}\right)}{\sum_{i=1}^{n_t}\left|rc_{i,t}^{s} - \overline{rc}_{t}^{s}\right|}, \tag{19}$$

where

$$\overline{rc}_{t}^{s} = \frac{1}{n_t}\sum_{i=1}^{n_t} rc_{i,t}^{s}.$$

Each column of $Z_t$ is $\left(z_{1,t}^{s}, \ldots, z_{n_{t,1}}^{s}\right)^{\top}$. It is known in practice that individual characteristic data is imbalanced panel dat. For this reason, we exploit $n_t$ rather than $N$ to emphasize the time-varying cross-sectional dimension.[3]

### 3.2 Characteristic-managed portfolios

Annual accounting data is realigned with monthly return data based on the following annual rebalancing rule. Returns at the monthly frequency from July of year $t$ to June of year $t+1$ are matched to the annual accounting variables in December of $t-1$. This is also the mechanism in which

[2] Jensen, Kelly, and Pedersen (2022) also makes their replication procedures and data publicly available at `https://github.com/bkelly-lab/ReplicationCrisis`.

[3] This also implicitly suggests that for each cross-section we only use those individual assets available as observations both for the corresponding returns and specific characteristics (indexed by $s$).

we realign data to construct cross-sectional equity characteristic data. For monthly rebalancing to construct the daily characteristic-managed portfolios, a similar scheme applies. That is, to construct the daily characteristic-managed portfolios in month $t+1$ based on equity $s$, returns at the daily frequency are matched with the normalized characteristics $z^s_{i,t}$ in month $t$ and $z^s_{i,t}$ are used as the weights for constructing the daily characteristic-managed portfolios. Characteristics normalized as in (19) ensure the managed portfolios, to some extent, mimic the long-short trading strategies so that we can use the normalized characteristics as the weights for constructing portfolios. These normalized variables are then used to construct 123 characteristic-managed portfolios (either in the monthly or daily frequency, and portfolios constructed at the daily frequency will be mainly used for the empirical analysis). More comprehensive descriptions of these anomaly variables are listed in the appendix along with acronyms used in our replication procedure. The corresponding papers in which these anomaly variables were initially proposed are listed in the appendix as well.

# 4 Empirical Findings

We now apply this Bayesian interpretable machine-learning method in analyzing the sparse structure of cross-sectional exposure of SDF using the data for the Chinese stock market constructed above. Our main empirical finding is that, in general, it is a futile effort to summarize the SDF as the exposure to several dominant cross-sectional characteristics in the Chinese stock market.

We first demonstrate results generated from imposing the $L^2$-norm penalty on cross-sectional regression (i.e., ridge regression with H-J distance as the objective function, which is also nicely interpretable from the Bayesian perspective). As we have discussed in the previous section that the tuning parameter ($\gamma_2$) associated with the $L^2$-penalized cross-sectional regression is closely related to the expected Sharpe ratio under the Bayesian prior ($\kappa$), we plot both IS (in-sample)/OOS (out-of-sample) $R^2$ across different $\kappa$ values in the following figure

**[Place Figure 1 about here]**

In the following figure, we plot the coefficient path associated with the $L^2$-norm-penalized regression across different root expected Sharpe ratios under the Bayesian prior ($\kappa$), that is, different strengths of the $L^2$-penalty imposed on the cross-sectional regression.

**[Place Figure 2 about here]**

Next, we summarize both the estimated coefficients $\hat{\omega}$ and the associated absolute value of the $t$-statistic calculated using (10) and (11).

**[Place Table 1 about here]**

The main implication from Table 1 is that although there is rarely any redundancy of the cross-sectional equity characteristics to summarize SDF in the Chinese stock market since absolute

values of the SDF coefficients associated with the leading SDF factors (among 123 SDF factors) listed in Table 1(a) are not close to zero. However, according to Table 1(b), there are 2 to 3 leading latent factors (i.e., principal components) that are statistically significant with relatively large estimated SDF coefficients. Based on the classification in Jensen, Kelly, and Pedersen (2022), it is not surprising to see that **Size**, **Value** and **Investment** related equity characteristics matter for SDF in the Chinese stock market. This empirical result is not far away from that in the U.S stock market. The $t$-statistics reported in Table 1(a) are calculated using (11). These reported $t$-statistics are for reference since, in general, the joint selection matters more for the penalized regression with the $L^2$-norm penalty than the single selection.

Finally, let us discuss how this Bayesian interpretable machine-learning algorithm (i.e., single $L^2$-norm penalized cross-sectional regression with objective function adjusted as HJ-distance) can be extended to the cross-sectional regression with a dual-penalty by adding additional $L^1$-norm penalty for accommodating shrinkage purpose.

**[Place Figure 3 about here]**

Figure 3 essentially provide the main conclusion of this paper. It is obvious from this figure that the optimal tuning parameter pair $(\gamma_1, \gamma_2)$ (or equivalent $(\gamma_1, \kappa)$) leading to the highest OOS $R^2$ resides in the area with relatively smaller $\gamma_1$ and $\gamma_2$ (or equivalently reflected in the number of variables retained in the SDF, over $y$-axis and the root expected Sharpe ratio, over $x$-axis) in Figure 3. This cross-validated out-of-sample analysis implies that it is futile to summarize SDF in the Chinese stock market as the exposure to several dominant cross-sectional equity characteristics.

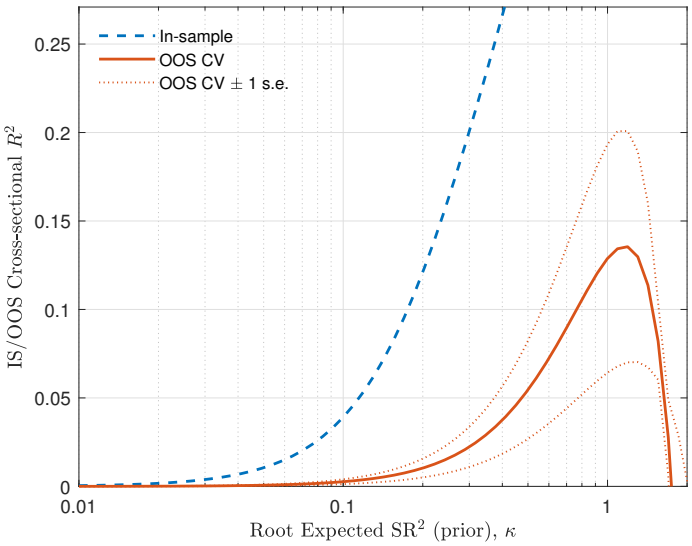
## 5 Conclusion

In this paper, we review an interpretable machine-learning method that features an economic-theory-based foundation from a Bayesian perspective. The cross-sectional regression with the $L^2$-norm penalty (the ridge regression with H-J distance as the objective function) has interpretation with the economic grounds from the Bayesian perspective. Given the attractive property of the methodology proposed in Kozak, Nagel, and Santosh (2020), we apply this methodology to analyze whether there exists a sparse structure of the SDF in the Chinese stock market. From the empirical perspective, we follow the cutting-edge data cleaning routine that is in response to recent discussions about the replication crisis in the empirical cross-sectional asset pricing literature to successfully replicate and construct 123 finance and accounting characteristics of individual assets in the Chinese stock market and hence construct the corresponding characteristics (anomalies) managed portfolios. Based on these constructed characteristics (anomalies)-managed portfolios, we apply both the pure $L^2$-penalized cross-sectional regression (the ridge regression with the H-J distance as the objective function) and the extended $L^1$-$L^2$ penalized cross-sectional regression (elastic-net regularization) to check whether there is a sparse exposure of SDF in the Chinese stock market. Our empirical study suggests that staying within the 123 cross-sectional equity characteristic universe, it is still hard to

characterize the SDF in the Chinese stock market using a few dominant characteristics, although our empirical analysis may suggest that there exist several dominant latent factors (principal components) to summarize the SDF in the Chinese stock market.

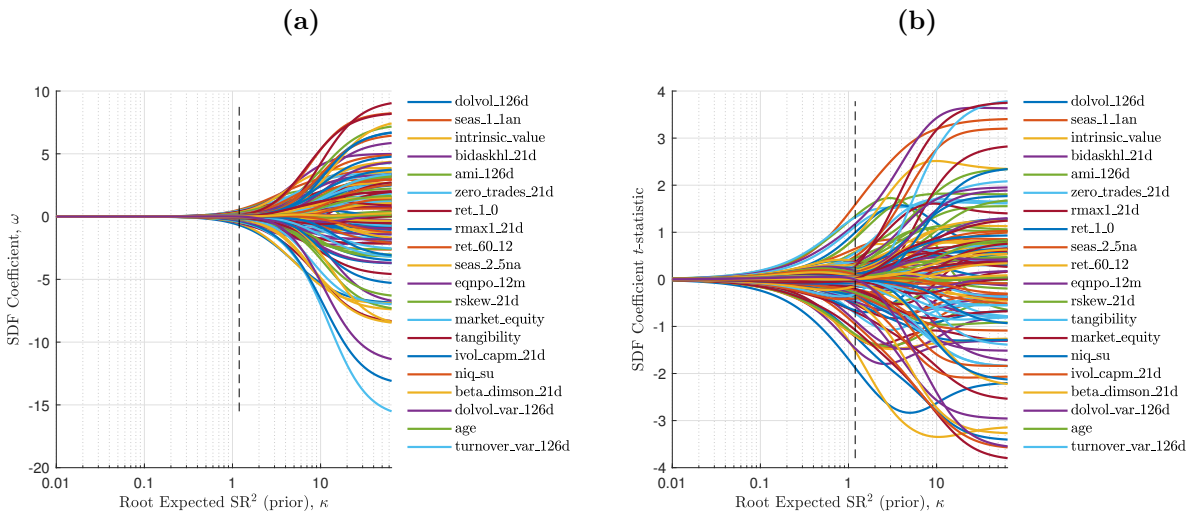
# Figures and Tables

## Figure 1




**Note:** In the figure above, we apply the Bayesian interpretable machine-learning method to the Chinese stock market by constructing characteristic-managed portfolios based on 123 anomaly variables. All characteristic-managed portfolio returns are constructed at the daily frequency. As we have discussed in the main context about the relationship between the root maximum squared Sharpe Ratio ($\kappa$) and the penalty parameter $\gamma$, in this figure we demonstrate cross-sectional $R^2$ and $\kappa$ (both in-sample (dashed blue line) and out-of-sample (solid red line)). The standard-error (dotted red line) is calculated based on splitted sample for cross-validation (3-folds cross-validation is used for this implementation).

## Figure 2




**Note:** In the figure above, we apply the Bayesian interpretable machine-learning method in the Chinese stock market by constructing characteristic-managed portfolios based on 123 anomaly variables. This is a plot demonstrating coefficient paths associated with the penalized cross-sectional regression with the $L^2$-norm penalty: $\hat{\omega}$ as estimated SDF coefficients across different $\kappa$ in (a) and corresponding $t$-statistics (using equation (10) and (11)) in (b). All the corresponding variables are sorted according to the absolute values. Vertical dashed lines both in (a) and (b) indicate the optimal tuning parameter based on cross-validation, i.e. $\kappa$ associated with the highest OOS $R^2$.

**Table 1.** Largest SDF factors in the Chinese Stock Market

**(a)**

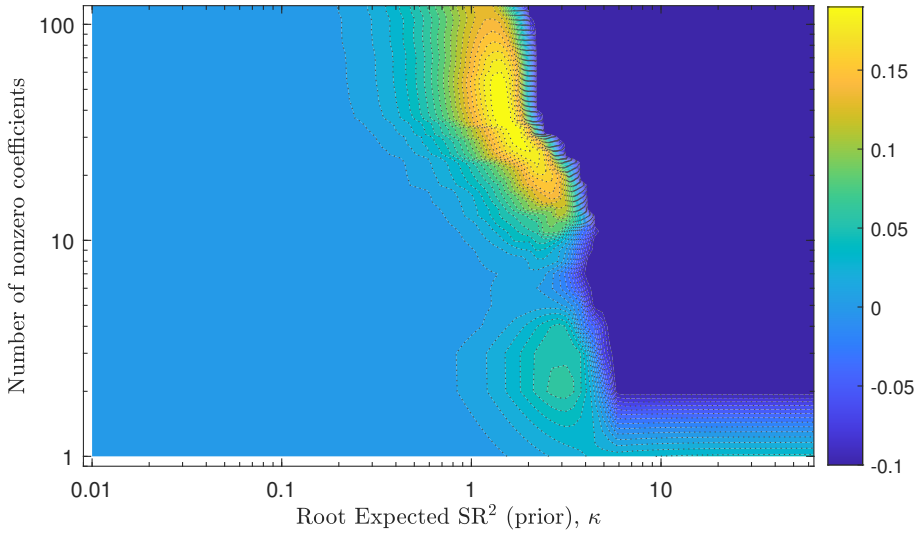
| | $\omega$ | $t$-stat |
|---|---|---|
| Dollar trading volume [**Size**] | -0.6815 | 1.8792 |
| Year 1-lagged return, annual [**Profit Growth**] | 0.5474 | 1.5867 |
| Intrinsic-value [**Value**] | -0.5456 | 1.5178 |
| 21 Day high-low bid-ask spread [**Low Leverage**] | -0.5164 | 1.4670 |
| Amihud measure [**Size**] | 0.4863 | 1.3422 |
| Number of zero trades (1 month) [**Low Risk**] | 0.4722 | 1.3180 |
| Maximum daily return [**Low Risk**] | -0.4390 | 1.2135 |
| Short-term reversal [**Size**] | -0.4446 | 1.2103 |
| Years 2-5 lagged returns, nonannual [**Investment**] | -0.4271 | 1.2017 |
| Long-term reversal [**Investment**] | -0.4283 | 1.2013 |

**(b)**

| | $\omega$ | $t$-stat |
|---|---|---|
| PC 7 | **1.0265** | **3.3041** |
| PC 8 | **1.0365** | **3.2092** |
| PC 3 | **0.4018** | **2.0319** |
| PC 6 | 0.4594 | 1.6777 |
| PC 11 | 0.5407 | 1.6288 |
| PC 24 | 0.5201 | 1.4496 |
| PC 28 | -0.4708 | 1.3003 |
| PC 17 | 0.4283 | 1.2330 |
| PC 1 | -0.1212 | 1.1643 |
| PC 13 | -0.3400 | 1.0087 |

**Note:** In the table above, we summarize corresponding results obtained from applying the Bayesian interpretable-machine learning method to the Chinese stock market by constructing characteristic-managed portfolios based 123 anomaly variables. In (a) we summarize estimated coefficients $\hat{\omega}$ at the optimal $L^2$-norm penalty tuning parameter $\gamma_2$ (or equivalently the root expected Sharpe ratio $\kappa$ under the prior distribution (based on cross-validation)). There 123 anomaly portfolios in all. In (b), anomaly portfolios returns are rotated into principal component (PC) space and corresponding estimated coefficients are demonstrated there. Coefficients are sorted descending on the absolute $t$-statistic values.

## Figure 3




**Note:** In the figure above, we apply the Bayesian interpretable machine-learning method to the Chinese stock market by constructing characteristic-managed portfolios based 123 anomaly variables. This is a plot demonstrating OOS $R^2$ associated with the $L^1$-$L^2$-penalized cross-sectional regression discussed in the main context. $L^2$-penalty is tuned via $\gamma_2$, which is closely related to the root expected Sharpe ratio $\kappa$ (over $x$-axis) under prior distribution; $L^1$-penalty is tuned via $\gamma_1$ and is in general proportional to the reciprocal of the number nonzero coefficients in cross-sectional regression. Hence we use the number of nonzero coefficients (i.e. number of variables retained in SDF) to characterize the strength associated with $L^1$-penalty (over $y$-axis). Both axes are plotted on logarithmic scale. Yellow color depicts the higher OOS $R^2$ while the dark blue area depicts $(\gamma_1, \gamma_2)$ pair for which the corresponding OOS $R^2$ is low.

## References


Abarbanell, J. S., and B. J. Bushee (1998): "Abnormal Returns to a Fundamental Analysis Strategy," *The Accounting Review*, 73(1), 19–45. [Cited on pages A-2 and A-4.]

Ali, A., L.-S. Hwang, and M. A. Trombley (2003): "Arbitrage risk and the book-to-market anomaly," *Journal of Financial Economics*, 69(2), 355–373. [Cited on page A-3.]

Amihud, Y. (2002): "Illiquidity and stock returns: cross-section and time-series effects," *Journal of Financial Markets*, 5(1), 31–56. [Cited on page A-1.]

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006): "The Cross-Section of Volatility and Expected Returns," *The Journal of Finance*, 61(1), 259–299. [Cited on pages A-1 and A-6.]

Asness, C., A. Frazzini, N. J. Gormsen, and L. H. Pedersen (2020): "Betting against correlation: Testing theories of the low-risk effect," *Journal of Financial Economics*, 135(3), 629–652. [Cited on pages A-2 and A-6.]

Asness, C. S., A. Frazzini, and L. H. Pedersen (2019): "Quality minus junk," *Review of Accounting Studies*, 24(1), 34–112. [Cited on page A-5.]

Bakalli, G., S. Guerrier, and O. Scaillet (2021): "A penalized two-pass regression to predict stock returns with time-varying risk premia," Swiss Finance Institute Research Paper Series 21-09, Swiss Finance Institute. [Cited on page 2.]

Balakrishnan, K., E. Bartov, and L. Faurel (2010): "Post loss/profit announcement drift," *Journal of Accounting and Economics*, 50(1), 20–41. [Cited on page A-4.]

Bali, T., A. Goyal, D. Huang, F. Jiang, and Q. Wen (2021): "Different Strokes: Return Predictability Across Stocks and Bonds with Machine Learning and Big Data," Working paper. [Cited on page 2.]

Bali, T. G., S. J. Brown, and Y. Tang (2017): "Is economic uncertainty priced in the cross-section of stock returns?," *Journal of Financial Economics*, 126(3), 471–489. [Cited on page A-6.]

Bali, T. G., N. Cakici, and R. F. Whitelaw (2011): "Maxing out: Stocks as lotteries and the cross-section of expected returns," *Journal of Financial Economics*, 99(2), 427–446. [Cited on page A-6.]

Bali, T. G., R. F. Engle, and S. Murray (2016): *Empirical Asset Pricing: The Cross Section of Stock Returns*. John Wiley & Sons, Hoboken, NJ. [Cited on pages A-3 and A-6.]

Ball, R., J. Gerakos, J. T. Linnainmaa, and V. Nikolaev (2016): "Accruals, cash flows, and operating profitability in the cross section of stock returns," *Journal of Financial Economics*, 121(1), 28–45. [Cited on pages A-2 and A-5.]

Banz, R. W. (1981): "The relationship between return and market value of common stocks," *Journal of Financial Economics*, 9(1), 3–18. [Cited on page A-3.]

Barth, M. E., J. A. Elliott, and M. W. Finn (1999): "Market Rewards Associated with Patterns of Increasing Earnings," *Journal of Accounting Research*, 37(2), 387–413. [Cited on page A-4.]

Basu, S. (1983): "The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence," *Journal of Financial Economics*, 12(1), 129–156. [Cited on page A-4.]

Belo, F., and X. Lin (2012): "The Inventory Growth Spread," *The Review of Financial Studies*, 25(1), 278–313. [Cited on page A-3.]

Bhandari, L. C. (1988): "Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence," *The Journal of Finance*, 43(2), 507–528. [Cited on page A-2.]

Bouchaud, J.-P., P. Krüger, A. Landier, and D. Thesmar (2019): "Sticky Expectations and the Profitability Anomaly," *The Journal of Finance*, 74(2), 639–674. [Cited on page A-5.]

Bradshaw, M. T., S. A. Richardson, and R. G. Sloan (2006): "The relation between corporate financing activities, analysts' forecasts and stock returns," *Journal of Accounting and Economics*, 42(1), 53–85, Conference Issue on Implications of Changing Financial Reporting Standards. [Cited on page A-2.]

Brennan, M. J., T. Chordia, and A. Subrahmanyam (1998): "Alternative factor specifications, security characteristics, and the cross-section of expected stock returns," *Journal of Financial Economics*, 49(3), 345–373. [Cited on page A-2.]

Catoni, O. (2004): *Statistical Learning Theory and Stochastic Optimization*. Saint- Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer. [Cited on page 2.]

——— (2007): *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes— Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH. [Cited on page 2.]

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51(5), 1281–1304. [Cited on page 2.]

Chen, A. Y., and T. Zimmermann (2020): "Open Source Cross-Sectional Asset Pricing," Working paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3604626. [Cited on pages 2 and 12.]

Chen, L., M. Pelger, and J. Zhu (2019): "Deep Learning in Asset Pricing," Working paper. [Cited on page 2.]

Chinco, A., A. D. Clark-Joseph, and M. Ye (2019): "Sparse Signals in the Cross-Section of Returns," Manuscript, forthcoming for *Journal of Finance*. [Cited on page 2.]

Chipman, H. A., E. I. George, and R. E. McCulloch (2010): "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4(1), 266–298. [Cited on page 3.]

Chordia, T., A. Subrahmanyam, and V. Anshuman (2001): "Trading activity and expected stock returns," *Journal of Financial Economics*, 59(1), 3–32. [Cited on pages A-2 and A-6.]

Connor, G., and R. A. Korajczyk (1986): "Performance measurement with the arbitrage pricing theory: A new framework for analysis," *Journal of Financial Economics*, 15(3), 373–394. [Cited on page 2.]

Cooper, M., H. Gulen, and M. J. Schill (2008): "Asset Growth and the Cross-Section of Stock Returns," *The Journal of Finance*, 63(4), 1609–1651. [Cited on page A-1.]

Corwin, S. A., and P. Schultz (2012): "A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices," *The Journal of Finance*, 67(2), 719–760. [Cited on page A-1.]

Daniel, K., and S. Titman (2006): "Market Reactions to Tangible and Intangible Information," *The Journal of Finance*, 61(4), 1605–1643. [Cited on page A-3.]

Datar, V. T., N. Y. Naik, and R. Radcliffe (1998): "Liquidity and stock returns: An alternative test," *Journal of Financial Markets*, 1(2), 203–219. [Cited on page A-6.]

De Bondt, W. F. M., and R. Thaler (1985): "Does the Stock Market Overreact?," *The Journal of Finance*, 40(3), 793–805. [Cited on page A-6.]

Dechow, P. M., R. G. Sloan, and M. T. Soliman (2004): "Implied Equity Duration: A New Measure of Equity Risk," *Review of Accounting Studies*, 9, 197–228. [Cited on page A-3.]

Demiguel, V., A. Martín, F. J. Nogales, and R. Uppal (2020): "A Transaction-Cost Perspective on the Multitude of Firm Characteristics," *The Review of Financial Studies*, 33(5), 2180–2122. [Cited on page 11.]

Dichev, I. D. (1998): "Is the Risk of Bankruptcy a Systematic Risk?," *The Journal of Finance*, 53(3), 1131–1147. [Cited on page A-6.]

Fairfield, P. M., J. S. Whisenant, and T. L. Yohn (2003): "Accrued Earnings and Growth: Implications for Future Profitability and Market Mispricing," *The Accounting Review*, 78(1), 353–371. [Cited on page A-3.]

Fama, E. F., and K. R. French (1992): "The Cross-Section of Expected Stock Returns," *The Journal of Finance*, 47(2), 427–465. [Cited on pages 2 and A-1.]

——— (1993): "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33(1), 3–56. [Cited on pages 2 and 4.]

——— (1996): "Multifactor Explanations of Asset Pricing Anomalies," *The Journal of Finance*, 51(1), 55–84. [Cited on page 2.]

——— (2015): "A Five-factor Asset Pricing Model," *Journal of Financial Economics*, 116(1), 1 – 22. [Cited on pages 2 and A-5.]

Fama, E. F., and J. D. Macbeth (1973): "Risk, Return and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81(3), 607–636. [Cited on page A-1.]

Fan, J., Y. Liao, and W. Wang (2016): "Projected Principal Component Analysis in Factor Models," *The Annals of Statistics*, 44(1), 219–254. [Cited on page 2.]

Foster, G., C. Olsen, and T. Shevlin (1984): "Earnings Releases, Anomalies, and the Behavior of Security Returns," *The Accounting Review*, 59(4), 574–603. [Cited on page A-4.]

Fowler, D. J., and C. Rorke (1983): "Risk measurement when shares are subject to infrequent trading: Comment," *Journal of Financial Economics*, 12(2), 279–283. [Cited on page A-1.]

Francis, J., R. LaFond, P. M. Olsson, and K. Schipper (2004): "Costs of Equity and Earnings Attributes," *The Accounting Review*, 79(4), 967–1010. [Cited on pages A-3 and A-4.]

Frankel, R., and C. M. Lee (1998): "Accounting valuation, market expectation, and cross-sectional stock returns," *Journal of Accounting and Economics*, 25(3), 283–319. [Cited on page A-3.]

Frazzini, A., and L. H. Pedersen (2014): "Betting against beta," *Journal of Financial Economics*, 111(1), 1–25. [Cited on page A-1.]

Freyberger, J., A. Neuhierl, and M. Weber (2020): "Dissecting Characteristics Nonparametrically," *The Review of Financial Studies*, 33(5), 2326–2377. [Cited on pages 1, 11, and 12.]

Gabauer, D., R. Gupta, H. A. Marfatia, and S. M. Miller (2020): "Estimating U.S. Housing Price Network Connectedness: Evidence from Dynamic Elastic Net, Lasso, and Ridge Vector Autoregressive Models," Working Papers 202065, University of Pretoria, Department of Economics. [Cited on page 2.]

George, T. J., and C.-Y. Hwang (2004): "The 52-Week High and Momentum Investing," *The Journal of Finance*, 59(5), 2145–2176. [Cited on page A-5.]

Giannone, D., M. Lenza, and G. E. Primiceri (2021): "Economic predictions with big data: The illusion of sparsity," ECB Working Paper 2542. [Cited on page 2.]

Green, J., J. R. M. Hand, and X. F. Zhang (2017): "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns," *The Review of Financial Studies*, 30(12), 4389–4436. [Cited on page 11.]

Gu, S., B. Kelly, and D. Xiu (2020): "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, 33(5), 2223–2273. [Cited on pages 1 and 11.]

Hafzalla, N., R. Lundholm, and E. M. V. Winkle (2011): "Percent Accruals," *The Accounting Review*, 86(1), 209–236. [Cited on pages A-4 and A-6.]

Hahn, J., and H. Lee (2009): "Financial Constraints, Debt Capacity, and the Cross-Section of Stock Returns," *The Journal of Finance*, 64(2), 891–921. [Cited on page A-6.]

Han, Y., A. He, D. E. Rapach, and G. Zhou (2019): "What Firm Characteristics Drive US Stock Returns," Working paper. [Cited on page 2.]

Hansen, L. P., and R. Jagannanthan (1991): "Implications of Security Market Data for Models of Dynamic Economies," *Journal of Political Economy*, 99(2), 225–262. [Cited on page 4.]

Harvey, C. R., and Y. Liu (2014): "Evaluationg Trading Strategies," *Journal of Portfolio Management*, 40(5), 108–118. [Cited on page 11.]

——— (2015): "Backtesting," *Journal of Portfolio Management*, 42(1), 13–28. [Cited on page 11.]

Harvey, C. R., Y. Liu, and H. Zhu (2016): "...and the Cross-Section of Expected Returns," *The Review of Financial Studies*, 29(1), 5–68. [Cited on pages 1 and 11.]

Harvey, C. R., and A. Siddique (2000): "Conditional Skewness in Asset Pricing Tests," *The Journal of Finance*, 55(3), 1263–1295. [Cited on page A-2.]

Hastie, T., R. Tibshirani, and J. Friedman (2001): *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA. [Cited on page 2.]

Haugen, R. A., and N. L. Baker (1996): "Commonality in the determinants of expected stock returns," *Journal of Financial Economics*, 41(3), 401–439. [Cited on pages 4, A-1, A-2, and A-4.]

Heston, S. L., and R. Sadka (2008): "Seasonality in the cross-section of stock returns," *Journal of Financial Economics*, 87(2), 418–445. [Cited on page A-6.]

Hirshleifer, D., K. Hou, S. Teoh, and Y. Zhang (2004): "Do Investors Overvalue Firms with Bloated Balance Sheets," *Journal of Accounting and Economics*, 38, 297–331. [Cited on page A-4.]

Hou, K., C. Xue, and L. Zhang (2015): "Digesting Anomalies: An Investment Approach," *The Review of Financial Studies*, 28(3), 650–705. [Cited on pages 2 and A-4.]

——— (2018): "Replicating Anomalies," *The Review of Financial Studies*, 33(5), 2019–2133. [Cited on pages 2 and 11.]

Huang, D., J. Li, and L. Wang (2021): "Are disagreements agreeable? Evidence from information aggregation," *Journal of Financial Economics*, 141(1), 83–101. [Cited on page 2.]

Jegadeesh, N. (1990): "Evidence of Predictable Behavior of Security Returns," *The Journal of Finance*, 45(3), 881–898. [Cited on page A-5.]

Jegadeesh, N., and S. Titman (1993): "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *The Journal of Finance*, 48(1), 65–91. [Cited on pages A-5 and A-6.]

Jensen, T. I., B. Kelly, and L. H. Pedersen (2022): "Is There a Replication Crisis in Finance," Working paper, conditionally accepted in *The Journal of Finance*. [Cited on pages 12, 14, and A-1.]

Jiang, G., C. M. Lee, and Y. Zhang (2005): "Information Uncertainty and Expected Returns," *Review of Accounting Studies*, 10, 185–221. [Cited on page A-1.]

Kelly, B. T., S. Pruitt, and Y. Su (2019): "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, 134(3), 501–524. [Cited on page 2.]

Kim, H. H., and N. R. Swanson (2014): "Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence," *Journal of Econometrics*, 178, 352–367, Recent Advances in Time Series Econometrics. [Cited on page 2.]

Kozak, S. (2020): "Kernel Trick for the Cross Section," Working paper. [Cited on pages 2 and 11.]

Kozak, S., S. Nagel, and S. Santosh (2018): "Interpreting Factor Models," *The Journal of Finance*, 73(3), 1183–1223. [Cited on pages 2 and 4.]

——— (2020): "Shrinking the Cross Section," *Journal of Financial Economics*, 135(2), 271–292. [Cited on pages 3, 5, 11, and 14.]

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994): "Contrarian Investment, Extrapolation, and Risk," *The Journal of Finance*, 49(5), 1541–1578. [Cited on page A-6.]

Ledoit, O., and M. Wolf (2004a): "Honey, I Shrunk the Sample Covariance Matrix," *The Journal of Portfolio Management*, 30(4), 110–119. [Cited on page 9.]

——— (2004b): "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, 88(2), 365–411. [Cited on page 9.]

Lettu, M., and M. Pelger (2020a): "Estimating Latent Asset-Pricing Factors," forthcoming in *The Journal of Econometrics*. [Cited on page 2.]

——— (2020b): "Factors that Fit the Time-Series and Cross-Section of Stock Returns," *Review of Financial Studies*, 33(5), 2274–2325. [Cited on page 2.]

Lev, B., and D. Nissim (2004): "Taxable Income, Future Earnings, and Equity Values.," *Accounting Review*, 79(4), 1039–1074. [Cited on page A-5.]

Linero, A. R. (2018): "Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection," *Journal of the American Statistical Association*, 113(522), 626–636. [Cited on pages 2 and 3.]

Litzenberger, R. H., and K. Ramaswamy (1979): "The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence," *Journal of Financial Economics*, 7(2), 163–195. [Cited on page A-2.]

Liu, W. (2006): "A liquidity-augmented capital asset pricing model," *Journal of Financial Economics*, 82(3), 631–671. [Cited on page A-7.]

Loughran, T., and J. W. Wellman (2011): "New Evidence on the Relation between the Enterprise Multiple and Average Stock Returns," *Journal of Financial and Quantitative Analysis*, 46(6), 1629–1650. [Cited on page A-3.]

Lyandres, E., L. Sun, and L. Zhang (2008): "The New Issues Puzzle: Testing the Investment-Based Explanation," *Review of Financial Studies*, 21(6), 2825–2855. [Cited on pages A-2 and A-5.]

Mclean, R. D., and J. Pontiff (2016): "Does Academic Research Destroy Stock Return Predictability," *Journal of Finance*, 71(1). [Cited on page 11.]

Messmer, M., and F. Audrino (2017): "The (adaptive) Lasso in the Zoo-Firm Characteristic Selection in the Cross-Section of Expected Returns," . [Cited on page 2.]

Miller, M. H., and M. S. Scholes (1982): "Dividends and Taxes: Some Empirical Evidence," *Journal of Political Economy*, 90(6), 1118–1141. [Cited on page A-5.]

Novy-Marx, R. (2010): "Operating Leverage," *Review of Finance*, 15(1), 103–134. [Cited on page A-5.]

——— (2012): "Is momentum really momentum?," *Journal of Financial Economics*, 103(3), 429–453. [Cited on page A-5.]

——— (2013): "The other side of value: The gross profitability premium," *Journal of Financial Economics*, 108(1), 1–28. [Cited on page A-3.]

Ortiz-Molina, H., and G. M. Phillips (2014): "Real Asset Illiquidity and the Cost of Capital," *Journal of Financial and Quantitative Analysis*, 49(1), 1–32. [Cited on page A-1.]

Palazzo, B. (2012): "Cash holdings, risk, and expected returns," *Journal of Financial Economics*, 104(1), 162–185. [Cited on page A-1.]

Penman, S., S. A. Richardeson, and I. Tuna (2007): "The Book-to-Price Effect in Stock Returns: Accounting for Leverage," *Journal of Accounting Research*, 45(2), 427–467. [Cited on pages A-1 and A-4.]

Piotroski, J. D. (2000): "Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers," *Journal of Accounting Research*, 38, 1–41. [Cited on page A-3.]

Pontiff, J., and A. Woodgate (2008): "Share Issuance and Cross-sectional Returns," *The Journal of Finance*, 63(2), 921–945. [Cited on page A-2.]

Rapach, D. E., J. K. Strauss, and G. Zhou (2010): "Out-of-sample Equity Premium Prediction: Combination Forecasts and links to the Real Economy," *The Review of Financial Studies*, 23(2), 821–862. [Cited on page 2.]

Richardson, S. A., R. G. Sloan, M. T. Soliman, and İrem Tuna (2005): "Accrual reliability, earnings persistence and stock prices," *Journal of Accounting and Economics*, 39(3), 437–485. [Cited on pages A-1, A-2, A-3, A-4, and A-6.]

Rosenberg, B., K. Reid, and R. Lanstein (1985): "Persuasive evidence of market inefficiency," *The Journal of Portfolio Management*, 11(3), 9–16. [Cited on page A-1.]

Ročková, V. (2019): "On Semi-parametric Bernstein-von Mises Theorems for BART," Working paper. [Cited on page 3.]

Ročková, V., and E. Saha (2019): "On Theory for BART," Working paper, $22^{nd}$ International Conference on Artificial Intelligence and Statistics. [Cited on page 3.]

Ročková, V., and S. van der Pas (2020): "Posterior concentration for Bayesian regression trees and forests," *The Annals of Statistics*, 48(4), 2108–2131. [Cited on page 3.]

Sloan, R. G. (1996): "Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings?," *The Accounting Review*, 71(3), 289–315. [Cited on page A-4.]

Soliman, M. T. (2008): "The Use of DuPont Analysis by Market Participants," *The Accounting Review*, 83(3), 823–853. [Cited on pages A-3 and A-6.]

Stambaugh, R. F., and Y. Yuan (2016): "Mispricing Factors," *The Review of Financial Studies*, 30(4), 1270–1315. [Cited on page A-3.]

Thomas, J. K., and H. Zhang (2002): "Inventory Changes and Future Returns," *Review of Accounting Studies*, 7, 163–187. [Cited on pages A-3 and A-6.]

Vapnik, V. (1998): *Statistical Learning Theory*. Wiley, New York. [Cited on page 2.]

Welch, I. (2019): "Reproducing, Extending, Updating, Replicationg, Reexamining, and Reconciling," *Critical Finance Review*, 8(1-2), 301–304. [Cited on page 12.]

William C. Barbee, J., S. Mukherji, and G. A. Raines (1996): "Do Sales-Price and Debt-Equity Explain Stock Returns Better than Book-Market and Firm Size?," *Financial Analysts Journal*, 52(2), 56–60. [Cited on page A-6.]

Zou, H., and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320. [Cited on pages 2 and 11.]

# Appendix

## A Anomaly variables used in Chinese stock market

We summarize the main cross-sectional equity characteristics (firm-level characteristics) used in empirical analysis of this paper. We follow the cutting-edge data-cleaning routine proposed in Jensen, Kelly, and Pedersen (2022) to replicate following 123 equity characteristics in Chinese stock market. In each item, we list the brief descriptions of corresponding anomaly variables with the acronym (in typewriter format collected in parenthesis) and in general the category (in bold collected square brackets) it belongs to in finance and accounting literature. We also list the corresponding literature that initially proposes equity characteristics. The corresponding information and description inherit directly from Jensen, Kelly, and Pedersen (2022) and readers should refer to documentation released along with Jensen, Kelly, and Pedersen (2022) for more about construction details.

1. Firm age (`age`) [**Low Leverage**], Jiang, Lee, and Zhang (2005).
2. Liquidity of book assets (`aliq_at`) [**Investment**], Ortiz-Molina and Phillips (2014).
3. Liquidity of market assets (`aliq_mat`) [**Low leverage**], Ortiz-Molina and Phillips (2014).
4. Amihud measure (`ami_126d`) [**Size**], Amihud (2002).
5. Book leverage (`at_be`) [**Low leverage**], Fama and French (1992).
6. Asset growth (`at_gr1`) [**Investment**], Cooper, Gulen, and Schill (2008).
7. Assets-to-market (`at_me`) [**Value**], Fama and French (1992).
8. Capital turnover (`at_turnover`) [**Quality**], Haugen and Baker (1996).
9. Change in common equity (`be_gr1a`) [**Investment**], Richardson, Sloan, Soliman, and İrem Tuna (2005).
10. Book-to-market equity `be_me` [**Value**], Rosenberg, Reid, and Lanstein (1985).
11. Market beta (`beta_60m`) [**Low Risk**], Fama and Macbeth (1973).
12. Dimson beta (`beta_dimson_21d`) [**Low Risk**], Fowler and Rorke (1983).
13. Frazzini-Pedersen market beta (`betabab_1260d`) [**Low Risk**], Frazzini and Pedersen (2014).
14. Downside beta (`betadown_252d`) [**Low Risk**], Ang, Hodrick, Xing, and Zhang (2006).
15. Book-to-market enterprise value (`bev_mev`) [**Value**], Penman, Richardeson, and Tuna (2007).
16. 21 Day high-low bid-ask spread (`bidaskhl_21d`) [**Low Leverage**], Corwin and Schultz (2012).
17. Cash-to-assets (`cash_at`) [**Low Leverage**], Palazzo (2012).

18. Net stock issues (`chcsho_12m`) [**Value**], Pontiff and Woodgate (2008).

19. Change in current operating assets (`coa_gr1a`) [**Investment**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

20. Change in current liabilities (`col_gr1a`) [**Investment**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

21. Cash-based operating profits-to-book assets (`cop_at`) [**Quality**], Haugen and Baker (1996).

22. Cash-based operating profits-to lagged book assets (`cop_atl1`) [**Quality**], Ball, Gerakos, Linnainmaa, and Nikolaev (2016).

23. Market correlation (`corr_1260d`) [**Seasonality**], Asness, Frazzini, Gormsen, and Pedersen (2020).

24. Coskewness (`coskew_21d`) [**Seasonality**], Harvey and Siddique (2000).

25. Change in current operating working capital (`cowc_gr1a`) [**Accruals**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

26. Net debt issuance (`dbnetis_at`) [**Net debt issuance**], Bradshaw, Richardson, and Sloan (2006).

27. Growth in book debt (3 years) (`debt_gr3`) [**Debt Issuance**], Lyandres, Sun, and Zhang (2008).

28. Debt-to-market (`debt_me`) [**Value**], Bhandari (1988).

29. Change gross margin minus change sales (`dgp_dsale`) [**Quality**], Abarbanell and Bushee (1998).

30. Dividend yield (`div12m_me`) [**Value**], Litzenberger and Ramaswamy (1979).

31. Dollar trading volume (`dolvol_126d`) [**Size**], Brennan, Chordia, and Subrahmanyam (1998).

32. Coefficient of variation for dollar trading volume (`dolvol_var_126d`) [**Profitability**], Chordia, Subrahmanyam, and Anshuman (2001).

33. Change sales minus change inventory (`dsale_dinv`) [**Profit Growth**], Abarbanell and Bushee (1998).

34. Change sales minus change receivables (`dsale_drec`) [**Profit Growth**], Abarbanell and Bushee (1998).

35. Change sales minus change SG&A (`dsale_dsga`) [**Profit Growth**], Abarbanell and Bushee (1998).

36. Earnings variability (`earnings_variability`) [**Low Risk**], Francis, LaFond, Olsson, and Schipper (2004).

37. Return on net operating assets (`ebit_bev`) [**Profitability**], Soliman (2008).

38. Profit margin (`ebit_sale`) [**Profit Growth**], Soliman (2008).

39. Ebitda-to-market enterprise value (`ebitda_mev`) [**Value**], Loughran and Wellman (2011).

40. Equity duration (`eq_dur`) [**Value**], Dechow, Sloan, and Soliman (2004).

41. Equity net payout (`eqnpo_12m`) [**Value**], Daniel and Titman (2006).

42. Pitroski F-score (`f_score`) [**Profitability**], Piotroski (2000).

43. Change in financial liabilities (`fnl_gr1a`) [**Debt Issuance**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

44. Gross profits-to-assets (`gp_at`) [**Quality**], Novy-Marx (2013).

45. Gross profits-to-lagged assets (`gp_atl1`) [**Quality**], Novy-Marx (2013).

46. Intrinsic-value (`intrinsic_value`) [**Value**], Frankel and Lee (1998).

47. Inventory growth (`inv_gr1`) [**Investment**], Belo and Lin (2012).

48. Inventory change (`inv_gr1a`) [**Investment**], Thomas and Zhang (2002).

49. Idiosyncratic skewness from the CAPM (`iskew_capm_21d`) [**Skewness**], Bali, Engle, and Murray (2016).

50. Idiosyncratic volatility from the CAPM (21 days) (`ivol_capm_21d`) [**Low Risk**], Ali, Hwang, and Trombley (2003).

51. Idiosyncratic volatility from the CAPM (252 days) (`ivol_capm_252d`) [**Low Risk**] Ali, Hwang, and Trombley (2003).

52. Change in long-term net operating assetsn (`lnoa_gr1a`) [**Investment**], Fairfield, Whisenant, and Yohn (2003).

53. Change in long-term investments (`lti_gr1a`) [**Seasonality**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

54. Market equity (`market_equity`) [**Size**], Banz (1981).

55. Mispricing factor: Management (`mispricing_mgmt`) [**Investment**], Stambaugh and Yuan (2016).

56. Mispricing factor: Performance (`mispricing_perf`) [**Quality**], Stambaugh and Yuan (2016).

57. Change in noncurrent operating assets (`nroa_gr1a`) [**Investment**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

58. Change in noncurrent operating liabilities (`ncol_gr1a`)) [**Debt Issuance**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

59. Net debt-to-price (`netdebt_me`) [**Low Leverage**], Penman, Richardeson, and Tuna (2007).

60. Change in net financial assets (`nfna_gr1a`) [**Debt Issuance**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

61. Earnings persistence (`ni_ar1`) [**Debt Issuance**], Francis, LaFond, Olsson, and Schipper (2004).

62. Return on equity (`ni_be`) [**Profitability**], Haugen and Baker (1996).

63. Number of consecutive quarters with earnings increases (`ni_inc8q`) [**Quality**], Barth, Elliott, and Finn (1999).

64. Earnings volatility (`ni_ivol`) [**Low Leverage**], Francis, LaFond, Olsson, and Schipper (2004).

65. Earnings-to-price (`ni_me`) [**Value**], Basu (1983).

66. Quarterly return on assets (`niq_at`) [**Quality**], Balakrishnan, Bartov, and Faurel (2010).

67. Change in quarterly return on assets (`niq_at_chg1`) [**Profit Growth**], Abarbanell and Bushee (1998).

68. Quarterly return on equity (`niq_be`) [**Profitability**], Hou, Xue, and Zhang (2015).

69. Change in quarterly return on equity (`niq_be_chg1`) [**Profit Growth**], Abarbanell and Bushee (1998).

70. Standardized earnings surprise (`niq_su`) [**Profit Growth**], Foster, Olsen, and Shevlin (1984).

71. Change in net noncurrent operating assets (`nncoa_gr1a`) [**Investment**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

72. Net operating assets (`noa_at`) [**Debt Issuance**], Hirshleifer, Hou, Teoh, and Zhang (2004).

73. Change in net operating assets (`noa_gr1a`) [**Investment**], Hirshleifer, Hou, Teoh, and Zhang (2004).

74. Operating accruals (`oaccruals_at`) [**Accruals**], Sloan (1996).

75. Percent operating accruals (`oaccruals_ni`) [**Accruals**], Hafzalla, Lundholm, and Winkle (2011).

76. Operating cash flow to assets (`ocf_at`) [**Profitability**], Bouchaud, Krüger, Landier, and Thesmar (2019).

77. Change in operating cash flow to assets (`ocf_at_chg1`) [**Profit Growth**], Bouchaud, Krüger, Landier, and Thesmar (2019).

78. Operating cash flow to market (`ocf_me`) [**Value**], Bouchaud, Krüger, Landier, and Thesmar (2019).

79. Operating cash flow to assets (`ocf_at`) [**Profitability**], Bouchaud, Krüger, Landier, and Thesmar (2019).

80. Operating profits-to-lagged book assets (`op_atl1`) [**Quality**], Ball, Gerakos, Linnainmaa, and Nikolaev (2016).

81. Operating profits to book equity (`ope_be`) [**Profitability**], Fama and French (2015).

82. Operating profits to lagged book equity (`ope_bel1`) [**Profitability**], Fama and French (2015).

83. Operating leverage (`opex_at`) [**Quality**], Novy-Marx (2010).

84. Taxable income-to-book income (`pi_nix`) [**Seasonality**], Lev and Nissim (2004).

85. Change PPE and Inventory (`ppeinv_gr1a`) [**Investment**], Lyandres, Sun, and Zhang (2008).

86. Price and share (`prc`) [**Size**], Miller and Scholes (1982).

87. Current price to high price over last year (`prc_highprc_252d`) [**Momentum**], George and Hwang (2004).

88. Quality minus Junk: Composite (`qmj`) [**Quality**], Asness, Frazzini, and Pedersen (2019).

89. Quality minus Junk: Growth (`qmj_growth`) [**Quality**], Asness, Frazzini, and Pedersen (2019).

90. Quality minus Junk: Profitability (`qmj_prof`) [**Quality**], Asness, Frazzini, and Pedersen (2019).

91. Quality minus Junk: Safety (`qmj_safety`) [**Quality**], Asness, Frazzini, and Pedersen (2019).

92. Short-term reversal (`ret_1_0`) [**Size**], Jegadeesh (1990).

93. Price momentum $t-12$ to $t-1$ (`ret_12_1`) [**Momentum**], Jegadeesh and Titman (1993).

94. Price momentum $t-12$ to $t-7$ (`ret_12_7`) [**Profit Growth**], Novy-Marx (2012).

95. Price momentum $t-3$ to $t-1$ (`ret_3_1`) [**Momentum**], Jegadeesh and Titman (1993).

96. Price momentum $t-6$ to $t-1$ (`ret_6_1`) [**Momentum**], Jegadeesh and Titman (1993).

97. Long-term reversal (`ret_60_12`) [**Investment**], De Bondt and Thaler (1985).

98. Price momentum $t-9$ to $t-1$ (`ret_9_1`) [**Momentum**], Jegadeesh and Titman (1993).

99. Maximum daily return (`rmax1_21d`) [**Low Risk**], Bali, Cakici, and Whitelaw (2011).

100. Highest 5 days of return (`rmax5_21d`) [**Low Risk**], Bali, Brown, and Tang (2017).

101. Highest 5 days of return scaled by volatility (`rmax5_rvol_21d`) [**Skewness**], Asness, Frazzini, Gormsen, and Pedersen (2020).

102. Total skewness (`rskew_21d`) [**Skewness**], Bali, Engle, and Murray (2016).

103. Return volatility (`rvol_21d`) [**Low Risk**], Ang, Hodrick, Xing, and Zhang (2006).

104. Asset turnover (`sale_bev`) [**Quality**], Soliman (2008).

105. Sale growth (1 year) (`sale_gr1`) [**Investment**], Lakonishok, Shleifer, and Vishny (1994).

106. Sale growth (3 years) (`sale_gr3`) [**Investment**], Lakonishok, Shleifer, and Vishny (1994).

107. Sale to market (`sale_me`) [**Value**], William C. Barbee, Mukherji, and Raines (1996).

108. Sale growth (1 quarter) (`saleq_gr3`) [**Investment**], Lakonishok, Shleifer, and Vishny (1994).

109. Year 1-lagged return, annual (`seas_1_1an`) [**Profit Growth**], Heston and Sadka (2008).

110. Year 1-lagged return, nonannual (`seas_1_1na`) [**Momentum**], Heston and Sadka (2008).

111. Years 2-5 lagged returns, annual (`seas_2_5an`) [**Seasonality**], Heston and Sadka (2008).

112. Years 2-5 lagged returns, nonannual (`seas_2_5na`) [**Investment**], Heston and Sadka (2008).

113. Change in short-term investments (`sti_gr1a`) [**Seasonality**], Heston and Sadka (2008).

114. Total accruals (`taccruals_at`) [**Accruals**], Richardson, Sloan, Soliman, and İrem Tuna (2005).

115. Percent total accruals (`taccruals_ni`) [**Accruals**], Hafzalla, Lundholm, and Winkle (2011).

116. Asset tangibility (`tangibility`) [**Low Leverage**], Hahn and Lee (2009).

117. Tax expense surprise (`tax_gr1a`) [**Profit Growth**], Thomas and Zhang (2002).

118. Share turnover (`turnover_126d`) [**Low Risk**], Datar, Y. Naik, and Radcliffe (1998).

119. Coefficient of variation for share turnover (`turnover_var_126d`) [**Profitability**], Chordia, Subrahmanyam, and Anshuman (2001).

120. Altman Z-score (`z_score`) [**Low Leverage**], Dichev (1998).

121. Number of zero trades with turnover as tiebreaker (6 months) (`zero_trades_126d`) [**Low Risk**], Liu (2006).

122. Number of zero trades with turnover as tiebreaker (1 month) (`zero_trades_21d`) [**Low Risk**], Liu (2006).

123. Number of zero trades with turnover as tiebreaker (12 months) (`zero_trades_252d`) [**Low Risk**], Liu (2006).