Cross-Sectional Learning and Inference for the Stochastic Discount Factor[*†]

Zhanhui Chen[‡] Yi Ding[§] Yingying Li[¶] Xinghua Zheng[‖]

Abstract

We develop a statistical learning framework for constructing the stochastic discount factor (SDF) portfolio. To address the dimensionality challenge, we extend the MAXSER method (Ao, Li, and Zheng, 2019) to allow for $N \gg T$; prove that it surely screens for useful characteristics; and establish asymptotic normality for the SDF loading estimates. Using 153 characteristics returns from cross-sectional regressions (Fama and French, 2020), our framework not only constructs an SDF with a high out-of-sample Sharpe ratio that successfully prices the cross-section of expected returns, but also allows us to identify key characteristic themes and test the significance of their contributions.

JEL classification: C55, C58, G11, G12

Keywords: Stochastic discount factor; Factor models; High dimensions; Cross-sectional factors; Maximum Sharpe ratio regression

*Zhanhui Chen acknowledges the financial support from the Hong Kong Research Grants Council (GRF 16504522 and GRF 16503423). Yi Ding acknowledges the financial support from NSFC of China (72101226), SRG2023-00001-FBA, MYRG-GRG2023-00119-FBA and MYRG-GRG2024-00072-FBA of the University of Macau, and FDCT grants 0057/2024/ITP2. Yingying Li acknowledges the financial support from RGC grants (GRF 16504223, SRFS 2324-6H02 and RIF 6010-23). Xinghua Zheng acknowledges the financial support from RGC grants (GRF 16304521).

†We are grateful to comments from Yachine Ait-Sahalia, Ravi Jagannathan, Torben Anderson, Lin William Cong, Victor DeMiguel, Robert Engle, Jianqing Fan, Stefano Giglio, William Goetzmann, Raymond Kan, Bryan Kelly, Jia Li, Andrew Patton, Guanhao Feng, Viktor Todorov, Xiao Qiao, Dacheng Xiu, as well as seminar and conference participants at Princeton University, Northwestern University, the Imperial College London, London Business School, the Market Microstructure, Quantitative Trading, High Frequency, and Large Data Conference in Chicago, Barcelona Workshop in Financial Econometrics, the 16th Annual Society for Financial Econometrics (SoFiE 2024) Conference, FinEML Conference 2024 at USI Lugano, HKUST IAS-SBM Joint Workshop on Financial Econometrics in the Big Data Era, the 36th Asian Finance Association Annual Conference, the first INFORMS Conference on Financial Engineering and FinTech, the 19th International Symposium on Econometric Theory and Applications (Macau), the 4th Hong Kong Conference on FinTech, AI & Big Data in Business, and the Tsinghua SIEM Machine Learning Application in Financial Economics Conference. This manuscript has been previously circulated under the title "Learning the Stochastic Discount Factor" since 2023.

‡Corresponding Author. Department of Finance, Hong Kong University of Science and Technology, E-mail: chenzhanhui@ust.hk.

§Faculty of Business Administration, University of Macau, E-mail: yiding@um.edu.mo.

¶Department of ISOM and Department of Finance, Hong Kong University of Science and Technology, E-mail: yyli@ust.hk.

‖Department of ISOM, Hong Kong University of Science and Technology, E-mail: xhzheng@ust.hk.

1 Introduction

How can we construct an efficient stochastic discount factor (SDF) portfolio from the vast universe of potentially predictive firm characteristics while simultaneously detecting which of them are statistically significant drivers of factor premia? That is, as Cochrane (2011) points out, "which characteristics provide truly independent useful information explaining the returns, and hence contribute to the SDF?" This paper develops a statistical learning framework to solve this estimation and inference problem, including characteristic selection, portfolio construction, and inference tests for the contribution of individual characteristics or a theme to the SDF.

Learning the SDF is challenging due to the high-dimensional nature of the problem. There are hundreds of potentially useful factors identified in the literature, while the time series are limited, e.g., typically only a few hundred monthly observations are available. When the number of assets is about the same order of magnitude or even larger than the number of periods, estimating the high-dimensional covariance matrix is non-trivial and creates substantial statistical challenges. In addition, obtaining an accurate estimation of the expected returns is difficult because we need a long period to estimate the means but factors might have limited number of observations or experience structural changes over time.

Traditionally, factors and test assets are constructed through portfolio sorting (long-short portfolios based on firm characteristics, e.g., significant anomalies), and factor models are then tested against certain assets. This approach suffers from two major limitations. First, it can only handle a few (usually two or three) characteristics and cannot fully address the interaction among many different characteristics (see, e.g., Daniel, Mota, Rottke, and Santos, 2020; Giglio, Xiu, and Zhang, 2025; Cong, Feng, He, and He, 2025). Second, the choice of test assets is critical but less explored (Barillas and Shanken, 2018). Because different assets could have heterogeneous risk prices of factors (Patton and Weller, 2022), the choice of test assets affects the empirical evaluation. Moreover, there can be weak factors or even irrelevant factors which test assets do not have strong exposure to, and the empirical tests may be biased (see, e.g., Kan and Zhang, 1999; Gospodinov, Kan, and Robotti, 2017; Giglio, Xiu, and Zhang, 2025).[1]

To address these difficulties, our approach rests on three pillars: constructing SDF as a high-dimensional portfolio optimization problem; using cross sectional regression-based characteristic returns as basis assets; developing statistical methodologies and theories for factor selection and

[1]For example, many macroeconomic factors appear to be weak factors and test assets may be unable to identify risk premia of these weak factors.

inference tests.

First, for our objective, we follow the right-hand side approach suggested in Barillas and Shanken (2017) and Fama and French (2018) to construct the SDF. Barillas and Shanken (2017) and Fama and French (2018) show that for models with tradable factors, it is sufficient to examine the model's maximum squared Sharpe ratio (which is equivalent to the Hansen and Jagannathan (1997) distance), while the choice of test assets is irrelevant. That is, the model with "less mispricing" should have a higher squared Sharpe ratio.[2] Following the right-hand side approach, we set our objective to be identifying the portfolio with the maximum Sharpe ratio on the mean-variance efficient frontier. By construction, such an SDF portfolio is mean-variance efficient and prices asset returns. This is also consistent with the no arbitrage condition that if there exists a factor model that prices all assets, then the factor portfolio is mean-variance efficient.

Second, for input assets, we use characteristic-based returns estimated from cross-sectional regressions as the building blocks, instead of long-short characteristic-sorted portfolio returns (e.g., portfolios constructed by sorting on characteristics). The reason is that long-short characteristic-sorted portfolios may not be mean-variance efficient and could capture both priced and unpriced risks (Daniel, Mota, Rottke, and Santos, 2020). Empirically, Fama and French (2020) show that cross-sectional factors constructed from the cross-sectional regression of Fama and MacBeth (1973) outperform time-series factors constructed from portfolio sorts. Therefore, we follow Fama and French (2020) and estimate characteristic-based returns directly from the cross-sectional regressions and then use them to construct the SDF.

Third, our statistical methodology and theory provide rigorous guarantees. We develop a comprehensive statistical theory for constructing the SDF in a high-dimensional setting, by advancing the maximum-Sharpe ratio estimated & sparse regression (MAXSER) approach proposed by Ao, Li, and Zheng (2019) in three important directions: (a) We extend it to the setting where the number of assets can be much larger than the sample size ($N \gg T$). (b) We establish its sure screening property, proving that it retains the true signals with a high probability while effectively reducing irrelevant variables. (c) We develop a novel asymptotic theory for the estimated SDF loadings, which is crucial for addressing the over-selection in LASSO-type methods and allows us to formally test the significance of individual characteristics and thematic groups

[2] Barillas and Shanken (2018) apply this observation and suggest a Bayesian procedure which computes the probabilities for all models created by a given set of factors. Barillas, Kan, Robotti, and Shanken (2020) apply this metric to empirically compare several prevailing factor models.

in the SDF portfolio. The only existing inference result in the literature of asset pricing with machine learning techniques, to the best of our knowledge, is in Feng, Giglio, and Xiu (2020). They use a set of test assets and double-selection LASSO to select useful factors, and then conduct post-selection estimation and inference. But we use the right-hand side approach, which does not need to specify test assets.

Our empirical analysis uses a large set of firm characteristics from Jensen, Kelly, and Pedersen (2023). We test our estimated SDF against benchmark factor models such as CAPM, the Fama-French three-factor model (Fama and French, 1992, hereafter, FF3), the Fama-French five-factor model (Fama and French, 2015, hereafter, FF5), the Fama-French five-factor model augmented with momentum factor (Fama and French, 2018, labeled as FF6 hereafter), Q/Q5 models (Hou, Xue, and Zhang, 2015; Hou, Mo, Xue, and Zhang, 2021), the six-factor model in Barillas and Shanken (2018, hereafter, BS6), and behavioral models such as Stambaugh and Yuan (2017, hereafter, SY4) and Daniel, Hirshleifer, and Sun (2020, hereafter, DHS3). We also include the SDF estimator of Kozak, Nagel, and Santosh (2020, hereafter, KNS) as a benchmark case.

The test results show that our SDF significantly outperforms these benchmark factor models both statistically and economically, providing a more comprehensive description of the cross-section of expected returns than existing models. Our inference framework reveals that themes like "Low Risk" and "Short-Term Reversal" are persistent and statistically significant contributors to the SDF premium, while "Debt Issuance" and "Profitability" are not. Moreover, we find that using characteristic returns from cross-sectional regressions leads to a monthly Sharpe ratio of 0.78 for our estimated SDF, while the Sharpe ratio from using the long-short characteristic-sorted portfolio returns is only 0.38, suggesting the importance of using cross-sectional regression-based characteristic returns. We also show that our method does not simply select the most significant anomalies. Instead, it includes characteristics with insignificant alphas relative to some benchmark models but crucial for portfolio diversification.

Trading strategies based on machine learning techniques often face implementation issues (Jensen, Kelly, Malamud, and Pedersen, 2025). As pointed out in Patton and Weller (2020), "what you see is not what you get," implementation cost can significantly undermine the practical relevance of trading strategies based on anomalies, especially for small stocks. We formally address this tradability issue in three ways. First, we show that our estimated SDF portfolio neither concentrates on small-cap stocks nor has extreme weights on individual stocks. Second, we truncate the stock-level weights at the 1% level and find the weight-truncated portfolio performs similarly to the original one, suggesting that the superior performance of our SDF portfolio is not driven by extreme weights. Last, we show that explicitly accounting for the transaction costs does not significantly affect our SDF portfolio's performance. In fact, combining lots of characteristics could provide trading diversification effects (DeMiguel, Martin-Utrera, Nogales, and Uppal, 2020; DeMiguel, Martín-utrera, and Uppal, 2024). Overall, the results suggest that our proposed SDF portfolio is practically feasible.

Our paper is closely related to the emerging literature on applying various machine learning tools to identify factors and construct factor models. First, some papers apply principal component analysis (PCA) or its variants. For example, Kelly, Pruitt, and Su (2019) develop an instrument PCA method to estimate loadings of latent factor models as linear functions of characteristics. Lettau and Pelger (2020a,b) propose a risk-premium PCA to estimate the SDF at the presence of weak factors. Bryzgalova, DeMiguel, Li, and Pelger (2023) generalize PCA to consider economic targets of cross-sectional and time-series moments which helps to detect weak factors. Kelly, Malamud, and Pedersen (2023) use singular value decomposition to extract cross-asset signals. Giglio, Xiu, and Zhang (2025) suggest using a supervised PCA to select test assets and estimate risk premia for potentially weak factors. Second, some papers apply tree-based methods. For example, Bryzgalova, Pelger, and Zhu (2025) emphasized the importance of considering the dependence of characteristics and apply decision trees to build portfolios, e.g., by pruning or regularizing a decision tree. Cong, Feng, He, and He (2025) propose a panel tree model to construct test assets and latent factors which are mean-variance efficient, e.g., specifying a global split criterion and growing a tree to maximize the Sharpe ratio of the SDF. We contribute to this strand of literature in two ways. First, earlier studies mostly focus on individual factors while we consider the mean-variance efficient SDF portfolio, which utilizes the characteristics information and covariance structure of asset returns. Second, we develop a solid statistical theory to test and support our SDF estimator.

Our study also relates to the empirical literature on identifying and interpreting characteristic-based factors. Numerous studies examine whether these factors are significant, robust, and replicable. In particular, the factor zoo raises concerns about data mining, significance, transaction costs, and replication.[3] Unlike these papers, our paper uses these characteristics to construct the SDF portfolio, following the right-hand side approach (Barillas and Shanken, 2017; Fama

[3]See, e.g., McLean and Pontiff (2016); Harvey, Liu, and Zhu (2016); Green, Hand, and Zhang (2017); Hou, Xue, and Zhang (2020); Jacobs and Müller (2020); Chen (2021); Harvey and Liu (2021); Jensen, Kelly, and Pedersen (2023); Goyal, Welch, and Zafirov (2024).

and French, 2018). The characteristics we use may not be truly anomaly portfolios. In fact, we find many important characteristics in our SDF do not have significant alphas relative to some factor models. More importantly, our asymptotic theory allows us to detect weak factors or groups and provides economic interpretations.

Last, our paper contributes to the literature of high-dimensional mean-variance efficient portfolio optimization. It is well known that using sample plug-in portfolio performs poorly in constructing mean-variance optimal portfolios when the number of assets is large due to the sampling error in high-dimensions (see, e.g., Michaud, 1989; Britten-Jones, 1999). Various approaches have been proposed to improve the performance of large portfolios. One way is to adopt better mean and covariance matrix estimators, e.g., the factor-model-based covariance matrix estimators (Fan, Fan, and Lv, 2008; Fan, Liao, and Mincheva, 2011, 2013), and shrinkage estimators (Ledoit and Wolf, 2017; Kozak, Nagel, and Santosh, 2020; Kelly, Malamud, Pourmohammadi, and Trojani, 2024). Another approach is by adding constraints, e.g., gross exposure constraint (DeMiguel, Garlappi, Nogales, and Uppal, 2009). These approaches lead to improved numerical performance compared with plug-in portfolio but are suboptimal in achieving mean-variance targets. Ao, Li, and Zheng (2019) propose the MAXSER approach, which finds the mean-variance efficient portfolios by maximizing the Sharpe ratio directly, instead of relying on the estimates of mean and covariance matrix separately. This paper extends Ao, Li, and Zheng (2019) to allow for $N \gg T$ and more importantly, establishes sure screening property of MAXSER and asymptotic normality to test the estimated SDF loadings.

A contemporaneous work by Luo, Zhou, and Zhou (2025) also builds on the MAXSER approach (a LASSO regression) to construct the pricing kernel. However, their approach suffers from several issues. First, their theory relies on a crucial assumption that the optimal weights have a bounded ℓ_1 norm (Assumption (A4) therein). This assumption is invalid when the returns of the base assets exhibit a factor structure, which is typical in the data. Second, the characteristic returns they construct present significant multi-collinearity. When there is a factor structure in asset returns or strong multi-collinearity in base assets, the LASSO method fails, resulting in false selection of variables (see, e.g., Fan, Ke, and Wang, 2020). Third, they attempt to identify important characteristics by sequentially removing characteristics from asset pools and comparing the differences in the Sharpe ratios. However, due to the complex interactions among characteristics, this procedure does not guarantee the successful selection of the most important factors. As a result, their approach fails to answer the question of which factors are truly important in the pricing kernel. In contrast, our paper differs in three important ways. First, our theory explicitly allows for factor structure and proves the sure screening property of the MAXSER approach. This crucial property guarantees that MAXSER can be used for an initial screening of factors without the risk of missing important ones. Second, we establish an inference theory and apply "sample-splitting" technique to further select factors, avoiding the over-selection issue of LASSO or false factors in their paper. Lastly, our inference theory allows us to test the significance of characteristics at both the individual or theme level.

The rest of this paper is organized as follows. We present our methodology and statistical theories in Section 2. Section 3 describes the empirical procedures. Section 4 presents main empirical results. Finally, we conclude in Section 5. The proofs and additional empirical results are collected in Appendices A and B.

The following notations are used throughout the paper. For any matrix $\mathbf{A} = (a_{ij})$, its spectral norm is defined as $\|\mathbf{A}\| = \max_{\|\mathbf{x}\|\leq 1} \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{A}^\top \mathbf{x}}$, where $\|\mathbf{x}\| = \sqrt{\sum x_i^2}$ for any vector $\mathbf{x} = (x_i)$; the max norm is defined as $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$; and the ℓ_1 norm is defined as $\|\mathbf{A}\|_1 = \max_i \sum_j |a_{ij}|$. For two sequences of positive real numbers (a_n) and (b_n), we write $a_n \gg b_n$ if $a_n/b_n \to \infty$, and write $a_n \asymp b_n$ if $b_n/c < a_n < cb_n$ for some constant $c > 1$ and for all n. For any vector $\mathbf{v}$, $\mathbf{v}(i)$ or v_i stands for its ith entry, and for any matrix $\mathbf{A}$, $\mathbf{A}(i, j)$ or A_{ij} stands for its (i, j)th entry. Finally, "$\xrightarrow{p}$" stands for convergence in probability, and "$\xrightarrow{\mathcal{L}}$" stands for convergence in distribution.

2 Methodology and statistical theories

In this section,we develop comprehensive statistical methodologies and theories for factor selections and inferences. We build upon the MAXSER approach of Ao, Li, and Zheng (2019), advancing it in three key directions. First, we generalize the framework to the ultra high-dimensional setting where the number of assets N can be much larger than the time series length T. Second, we establish its sure screening property, meaning that true signals are retained with high probability while discarding most irrelevant variables. Finally, we establish asymptotic normality for the estimated SDF loadings, which enables further factor selection and inference on premia associated with various characteristics or themes.

2.1 SDF and the mean-variance efficient portfolio

A beta-pricing model uses some pricing factors with excess returns $\mathbf{R}$, which can be some characteristic-based stock portfolios. The stochastic discount factor (SDF) in this beta-pricing model can be written as

$$\begin{aligned} M &= \frac{1 - \mathbf{b}^\top(\mathbf{R} - E(\mathbf{R}))}{R_f}, \\ E(M\mathbf{R}) &= 0, \end{aligned} \tag{2.1}$$

where $\mathbf{b}$ is the loading of the SDF on the pricing factors, and R_f is the risk-free rate. Solving (2.1), we get

$$\mathbf{b} = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \tag{2.2}$$

where $\mathbf{\Sigma}$ is the covariance matrix and $\boldsymbol{\mu}$ is the mean of factor returns $\mathbf{R}$.

Note that SDF loading $\mathbf{b}$ is proportional to the weights of the optimal mean-variance portfolio. Specifically, the mean-variance optimal portfolio solves

$$\max_{\boldsymbol{\omega}} \boldsymbol{\mu}^\top\boldsymbol{\omega}, \text{ subject to } \boldsymbol{\omega}^\top\mathbf{\Sigma}\boldsymbol{\omega} \leq \sigma^2,$$

where σ is a target risk constraint.[4] The optimal weights, $\boldsymbol{\omega}^*$, are

$$\boldsymbol{\omega}^* = \frac{\sigma}{\sqrt{\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}. \tag{2.3}$$

Comparing (2.3) with (2.2), we see that the SDF loadings $\mathbf{b}$ are proportional to the optimal portfolio weights $\boldsymbol{\omega}^*$. Motivated by such an observation, we will consider the SDF construction as a mean-variance portfolio optimization problem.

2.2 Extending MAXSER to allow for $N \gg T$

We assume that there are N characteristic-based portfolios, and their (excess) returns $\mathbf{R}$ are potentially driven by a small set of factors such that

$$\mathbf{R} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{f} + \boldsymbol{\varepsilon} =: \boldsymbol{\beta}\mathbf{f} + \mathbf{U}, \tag{2.4}$$

[4] Note that the optimization problem considers the investment on risky assets so the weights do not sum up to one. The investment weight on the risk-free asset is $1 - \sum_i w_i$.

where $\mathbf{f}$ is a $K \times 1$ vector of factors, $\boldsymbol{\beta}$ are factor loadings, and $\mathbf{U}$ are the idiosyncratic returns. We will use $(\mathbf{f}, \mathbf{R})$ to construct the SDF.

We first recall the maximum-Sharpe-ratio estimated and sparse regression (MAXSER) approach in Ao, Li, and Zheng (2019). For any risk constraint σ, Proposition 3 of Ao, Li, and Zheng (2019) shows that the optimal allocation on $(\mathbf{f}, \mathbf{R})$, denoted by $\boldsymbol{\omega}^*_{f,R}$, is given by $\boldsymbol{\omega}^*_{f,R} = ((\boldsymbol{\omega}^*_f - \frac{\sigma}{\sqrt{\theta_{all}}}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha})^\top, (\boldsymbol{\omega}^*_R)^\top)^\top$, where

$$\boldsymbol{\omega}^*_f = \frac{\sigma}{\sqrt{\theta_{all}}}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\mu}_f, \quad \boldsymbol{\omega}^*_R = \frac{\sigma}{\sqrt{\theta_{all}}}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha}, \tag{2.5}$$

$\boldsymbol{\mu}_f$ and $\boldsymbol{\Sigma}_f$ are the mean and the covariance matrix of $\mathbf{f}$; $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}_u$ are the mean and the covariance matrix of $\mathbf{U}$; and $\theta_{all} = \boldsymbol{\mu}_f^\top \boldsymbol{\Sigma}_f^{-1}\boldsymbol{\mu}_f + \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha} := \theta_f + \theta_u$ is the squared Sharpe ratio of the optimal portfolio on $(\mathbf{f}, \mathbf{R})$, θ_f and θ_u are squared Sharpe ratio of the optimal portfolio on $\mathbf{f}$ and $\mathbf{U}$, respectively. Note that any portfolio on $(\mathbf{f}, \mathbf{R})$ can be regarded as a portfolio on $(\mathbf{f}, \mathbf{U})$. In particular, the optimal portfolio $\boldsymbol{\omega}^*_{f,R}$ on $(\mathbf{f}, \mathbf{R})$ corresponds to an optimal allocation on $(\mathbf{f}, \mathbf{U})$ as follows:

$$\boldsymbol{\omega}^*_{f,u} = \left(\frac{\sigma}{\sqrt{\theta_{all}}}\Big(\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\mu}_f\Big)^\top, \frac{\sigma}{\sqrt{\theta_{all}}}\Big(\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha}\Big)^\top\right)^\top =: \Big((\boldsymbol{\omega}^*_f)^\top, (\boldsymbol{\omega}^*_u)^\top\Big)^\top. \tag{2.6}$$

Moreover, according to Proposition 1 of Ao, Li, and Zheng (2019), $\boldsymbol{\omega}^*_{f,u}$ solves

$$\min_{\boldsymbol{\omega}} E\Big((r_{\sigma,f,u} - (\mathbf{f}^\top, \mathbf{U}^\top)\boldsymbol{\omega})^2\Big), \tag{2.7}$$

where

$$r_{\sigma,f,u} = \sigma(\sqrt{\theta_{all}} + 1/\sqrt{\theta_{all}}). \tag{2.8}$$

We study the composition of SDF. Under such a setting, there is no risk constraint. This frees us from estimating the maximum squared Sharpe ratio as in Ao, Li, and Zheng (2019). An important implication is that we can extend MAXSER to allow for $N \gg T$, where T stands for the sample size. Specifically, suppose that we have T i.i.d. sample from model (2.4), $(\mathbf{f}_t, \mathbf{R}_t)$, $t = 1, ..., T$. We first fit model (2.4) with ordinary least of squares (OLS) to get $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$, and

$\widehat{\mathbf{U}}_t = \mathbf{R}_t - \widehat{\boldsymbol{\beta}}\mathbf{f}_t$. Then we estimate $\boldsymbol{\omega}^*_{f,u}$ by

$$\widehat{\boldsymbol{\omega}}^*_{f,u} = ((\widehat{\boldsymbol{\omega}}^*_f)^\top, (\widehat{\boldsymbol{\omega}}^*_u)^\top)^\top := \underset{\boldsymbol{\omega}=(\boldsymbol{\omega}_f^\top, \boldsymbol{\omega}_u^\top)^\top}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^T (1 - \boldsymbol{\omega}^\top(\mathbf{f}_t^\top, \widehat{\mathbf{U}}_t^\top)^\top)^2, \quad \text{subject to } \|\boldsymbol{\omega}_u\|_1 \le L. \tag{2.9}$$

Let us emphasize that the ℓ_1 penalty is only imposed on $\boldsymbol{\omega}_u$. The economic justification is that there is a relatively small number of idiosyncratic components that contribute to the optimal allocation. The portfolio $\widehat{\boldsymbol{\omega}}^*_{f,u} = ((\widehat{\boldsymbol{\omega}}^*_f)^\top, (\widehat{\boldsymbol{\omega}}^*_u)^\top)^\top$ on $(\mathbf{f}, \mathbf{U})$ corresponds to the following portfolio on $(\mathbf{f}, \mathbf{R})$:

$$\widehat{\boldsymbol{\omega}}^*_{f,R} = ((\widehat{\boldsymbol{\omega}}^*_f - \widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\omega}}^*_u)^\top, (\widehat{\boldsymbol{\omega}}^*_u)^\top)^\top. \tag{2.10}$$

We now state the assumptions that ensure $\widehat{\boldsymbol{\omega}}^*_{f,R}$ is asymptotically optimal.

Assumption 1 *The number of assets N and the sample size T satisfy that $T \to \infty$ and $\log(N) = o(T)$.*

Assumption 1 allows N to be much larger than T, a condition that is much weaker than that of Ao, Li, and Zheng (2019).

We also recall the following three assumptions from Ao, Li, and Zheng (2019).

Assumption 2 $\mathbf{f}_t \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$, *$\mathbf{U}_t$ is independent of $\mathbf{f}_t$ with* $\mathbf{U}_t \overset{i.i.d.}{\sim} N(\boldsymbol{\alpha}, \boldsymbol{\Sigma}_u)$.

Assumption 3 *There exists a constant M such that*

$$\max\left\{\boldsymbol{\alpha}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha}, \max_{1\le i\le N, 1\le k\le K} |\alpha_i|, |\beta_{ik}|, \boldsymbol{\Sigma}_u(i,i)\right\} < M.$$

Assumption 4 *There exists a constant L such that* $\|\boldsymbol{\omega}^*_u\|_1 \le L$.

Denote the mean and the covariance matrix of $(\mathbf{f}, \mathbf{R})$ by $\boldsymbol{\mu}_{f,R}$, and $\boldsymbol{\Sigma}_{f,R}$, respectively. The following theorem shows that the estimated portfolio $\widehat{\boldsymbol{\omega}}^*_{f,R}$ in (2.10) asymptotically achieves the maximized Sharpe ratio.

Theorem 1 *Under Assumptions 1–4, the estimator $\widehat{\boldsymbol{\omega}}^*_{f,R}$ in (2.10) satisfies*

$$\left| \frac{\boldsymbol{\mu}_{f,R}^\top \widehat{\boldsymbol{\omega}}^*_{f,R}}{\sqrt{(\widehat{\boldsymbol{\omega}}^*_{f,R})^\top \boldsymbol{\Sigma}_{f,R} \widehat{\boldsymbol{\omega}}^*_{f,R}}} - \sqrt{\theta_{all}} \right| \xrightarrow{p} 0. \tag{2.11}$$

Proof: See Appendix A.1.

2.3 Sure screening property of MAXSER

In this paper, besides finding a portfolio that can deliver a high Sharpe ratio, we are equally concerned about the portfolio weights. In particular, we do not want to miss important characteristics nor do we want to include many unnecessary characteristics. That is, we want to make sure that our method can surely screen for true signals while reducing irrelevant variables. This is missing in Ao, Li, and Zheng (2019).

We investigate the weights of MAXSER, starting with the following assumptions. For any subset $S \subset \{1, ..., N+K\}$, let $|S|$ denote its cardinality and S^c denote its complement set. Denote $S_{signal} = \{i : \boldsymbol{\omega}^*_{f,u}(i) \neq 0, 1 \leq i \leq N+K\}$.

Note that the LASSO regression (2.9) is equivalent to the following format with a one-to-one correspondence between L in (2.9) and λ in (2.12) below (see Section 3.4.2 of Hastie, Tibshirani, Friedman et al. (2009)):

$$\widehat{\boldsymbol{\omega}}^*_{f,u} = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^{T}(1-\boldsymbol{\omega}^\top(\mathbf{f}_t^\top, \widehat{\mathbf{U}}_t^\top)^\top)^2 + \lambda\|\boldsymbol{\omega}_u\|_1. \tag{2.12}$$

Assumption 5 *There exists a set S_1 that satisfies $S_{signal} \subseteq S_1 \subset \{1, ..., N, ..., N+K\}$, $|S_1|^{3/2}\sqrt{\log N/T} = o(1)$, and $\min_{i\in S_{signal}} |\boldsymbol{\omega}^*_{f,u}(i)| \gg \lambda\sqrt{|S_1|}$ with $\lambda \gg |S_1|\sqrt{(\log N)/T}$ for λ in (2.12).*

Denote the mean and the covariance matrix of $(\mathbf{f}, \mathbf{U})$ by $\boldsymbol{\mu}_{f,u}$, and $\boldsymbol{\Sigma}_{f,u}$, respectively. Denote $\boldsymbol{\Sigma}_{f,u,11} = (\boldsymbol{\Sigma}_{f,u}(i,j))_{i,j\in S_1}$, $\boldsymbol{\Sigma}_{f,u,21} = (\boldsymbol{\Sigma}_{f,u}(i,j))_{i\in S_1^c, j\in S_1}$, $\boldsymbol{\Sigma}_{f,u,22} = (\boldsymbol{\Sigma}_{f,u}(i,j))_{i\in S_1^c, j\in S_1^c}$, $\boldsymbol{\mu}_{f,u,1} = (\boldsymbol{\mu}_{f,u}(i))_{i\in S_1}$, $\boldsymbol{\mu}_{f,u,2} = (\boldsymbol{\mu}_{f,u}(i))_{i\in S_1^c}$, and $\boldsymbol{\omega}^*_{S_1} = (\boldsymbol{\omega}^*_{f,u}(i))_{i\in S_1}$. Further define the following set of vectors with elements being either 1 or -1:

$$V = \{(\delta_1, ..., \delta_{|S_1|})^\top : \delta_i = \operatorname{sign}(\boldsymbol{\omega}^*_{S_1}(i)) \text{ if } \boldsymbol{\omega}^*_{S_1}(i) \neq 0, \text{ and } \delta_i \in \{-1, 1\} \text{ otherwise}\}.$$

Assumption 6 *There exists $0 < \eta < 1$ such that for all $\mathbf{v} \in V$, $\|(\boldsymbol{\Sigma}_{f,u,21} + \boldsymbol{\mu}_{f,u,2}\boldsymbol{\mu}_{f,u,1}^\top)(\boldsymbol{\Sigma}_{f,u,11} + \boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top)^{-1}\mathbf{v}\|_{\max} < 1-\eta$. Moreover, $\|\boldsymbol{\Sigma}_{f,u,11}^{-1}\| < M$ for some constant $M < \infty$.*

Assumption 5, the sparsity assumption, is slightly different from Ao, Li, and Zheng (2019). We require the number of nonzero weights to be not too big. Moreover, the minimum nonzero weight is not too small.

Assumption 6 includes the irrepresentable condition and the regularity condition on the precision matrix that are widely imposed in the LASSO literature. Different from usual assumptions for LASSO regression (e.g., Zhao and Yu (2006)) in which the irrepresentable condition is imposed on the design matrix, here we state it as a property in the population. In addition, we relax the condition such that it holds for a set that includes the true signals. Specifically, in Zhao and Yu (2006), the irrepresentable condition is imposed on all non-signal variables, S^c_{signal}. Here, we only impose the condition on the non-signal variables in the set S_1^c. Our assumptions on the non-signal variables are hence more general under the relaxation.

Zhao and Yu (2006) establish the sign consistency of the estimated coefficients from the LASSO regression. Wainwright (2009) gives sharp threshold conditions on sample sizes for sign recovery of LASSO regression. It is important to note that our setting is intrinsically different from the standard LASSO regression. Specifically, in the standard LASSO regression, the inference theory is obtained by conditioning on the x-variables and utilizing the i.i.d. assumption imposed on the noise term. In our setting, if one conditions on the "x-variables" $(\mathbf{f}_t^\top, \widehat{\mathbf{U}}_t^\top)^\top$ in (2.12), then there is no randomness left in the noise term. Because of this reason, we have to adopt a different approach and utilize the randomness of the x-variables to develop asymptotic theory. Our result is also fundamentally different from the sure independence screening proposed in Fan and Lv (2008), which rely on sorting correlations between the response variable and individual x-variables for variable selection. In our setting, the response variable is a constant, so the correlations are constantly zero and provide no information about the importance of individual variables.

The next theorem establishes the sure screening property of $\widehat{\boldsymbol{\omega}}^*_{f,u}$ in our setting.

Theorem 2 *Under Assumptions 1–3 and 5–6, the estimator $\widehat{\boldsymbol{\omega}}^*_{f,u}$ in (2.12) satisfies*

$$P\Big(\text{sign}(\widehat{\boldsymbol{\omega}}^*_{f,u}(i)) = \text{sign}(\boldsymbol{\omega}^*_{f,u}(i)) \quad \textit{for all } i \in S_{signal} \cup S_1^c\Big) \to 1. \tag{2.13}$$

Proof: See Appendix A.2.

Theorem 2 states that MAXSER achieves the sure screening property, namely, with a probability tending to one, $S_{signal} \subset \hat{S} = \{i : \widehat{\boldsymbol{\omega}}^*_{f,u}(i) \neq 0, i \leq N + K\}$. In addition, $\hat{S} \subset S_1$, hence $|\hat{S}|$ is small. In summary, Theorem 2 guarantees that MAXSER can screen for true signals and effectively reduce the irrelevant variables.

2.4 Statistical inference of SDF loadings

The sparsity of the estimated parameters from the LASSO regression depends on the tuning parameter λ. In a finite sample, it is well known that LASSO estimation tends to overselect non-signals with a tuning parameter λ chosen from cross-validation using the criterion of minimizing prediction error (see, e.g., Leng, Lin, and Wahba (2006)). In addition, Leng, Lin, and Wahba (2006) suggest using statistical tests to perform post variable selection. Inference of sparse regressions has been explored in the literature. For example, a widely used approach is the debiased low-dimensional linear projection method developed by Zhang and Zhang (2014). However, for the same reason mentioned in the previous subsection, i.e., no randomness left in the noise term after conditioning on the x-variables, these approaches are not applicable to our setting. Another approach is to perform post inference based on variable selection from sparse regressions, e.g., Belloni, Chernozhukov, and Hansen (2014) and Lee, Sun, Sun, and Taylor (2016). In finance applications, Feng, Giglio, and Xiu (2020) develop a factor testing method based on OLS after two-pass variable selection. Again, such inference results do not apply to our setting.

In this subsection, we develop an alternative approach for statistical inference of the SDF loadings. We will use the screening results from MAXSER and conduct a post screening plug-in estimation of the SDF loadings. We will then make statistical inference based on the plug-in estimator. To ensure the validity of post-screening inference, we adopt the widely-used "sample-splitting" technique (Wasserman and Roeder, 2009). Specifically, we split the observations into two subsamples of comparable sizes, apply MAXSER on one subsample of data, then conduct further estimation and inference using only the factors selected in the previous step and the remaining subsample of data.

We now explain the procedures in detail. For any set $S \subset \{1, ..., N+K\}$, the SDF weights based on variables in S are

$$\mathbf{b}_S = \mathbf{\Sigma}_S^{-1} \boldsymbol{\mu}_S,$$

where $\mathbf{\Sigma}_S$ and $\boldsymbol{\mu}_S$ are the covariance matrix and the mean of $((\mathbf{f}, \mathbf{R})(i))_{i \in S}$. We aim to make inference about $\mathbf{b}_S$. To do so, we compute the plug-in estimator of $\mathbf{b}_S$:

$$\widehat{\mathbf{b}}_S = \widehat{\mathbf{\Sigma}}_S^{-1} \widehat{\boldsymbol{\mu}}_S, \tag{2.14}$$

where $\widehat{\boldsymbol{\Sigma}}_S$ and $\widehat{\boldsymbol{\mu}}_S$ are the sample covariance matrix and the sample mean of the returns in S.

Theorem 3 gives the central limit theorem of $\widehat{\mathbf{b}}_S$.

Theorem 3 *Under Assumptions 1–3, if in addition,* $\|\boldsymbol{\Sigma}_{f,R}^{-1}\| = O(1)$*, then for any set* S *with* $|S|^2(\log|S|) = o(T)$*, any fixed* k *and any deterministic* $k \times |S|$ *matrix* $\mathbf{A}$ *with* $\|\mathbf{A}\| = O(1)$*, we have*

$$\sqrt{T}\mathbf{A}(\widehat{\mathbf{b}}_S - \mathbf{b}_S) \xrightarrow{\mathcal{L}} N(0, \boldsymbol{\Sigma}_A),$$

where $\boldsymbol{\Sigma}_A = \lim_{N\to\infty} \mathbf{A}((\boldsymbol{\mu}_S^\top \boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S + 1)\boldsymbol{\Sigma}_S^{-1} + \boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1})\mathbf{A}^\top$.

Proof: See Appendix A.3.

Theorem 3 states that the SDF estimator $\widehat{\mathbf{b}}_S$ enjoys asymptotic normality, which allows us to perform inference on the SDF loadings.

Theorems 1–3 lay down the statistics foundation of our methodology. We describe the empirical procedures in the next section.

3 Empirical procedures

3.1 Characteristics data

We use the firm characteristics data from Jensen, Kelly, and Pedersen (2023). The dataset includes 153 firm characteristics, which could correspond to various anomalies. Jensen, Kelly, and Pedersen (2023) categorize these anomalies into 13 themes: accruals, debt issuance, investment, low leverage, low risk, momentum, profit growth, profitability, quality, seasonality, size, short-term reversal and value. As we include the market portfolio as a base factor in constructing the SDF, we didn't include the characteristic of CAPM beta (*beta_60m*) from Jensen, Kelly, and Pedersen (2023). That is, we use 152 characteristics from Jensen, Kelly, and Pedersen (2023). The sample is from January 1975 to December 2023. To avoid the bias caused by small stocks, we exclude the stocks with market capitalization that is lower than 10% quantile in the cross section.

3.2 Using cross-sectional regressions to construct characteristic-based returns

We use characteristic-based returns as our inputs, as firm characteristics appear to be important for stock returns. For example, Daniel and Titman (1997) and Jegadeesh, Noh, Pukthuanthong, Roll, and Wang (2019) show that firm characteristics significantly drive the cross-sectional return variations. Prior literature often uses long-short characteristic-sorted portfolio returns, e.g., portfolios constructed by sorting on characteristics. However, such portfolios may not be mean-variance efficient and could capture both priced and unpriced risks, as suggested in Daniel, Mota, Rottke, and Santos (2020). Empirically, Fama and French (2020) show that cross-sectional factors constructed from the cross-sectional regression approach of Fama and MacBeth (1973) perform better than time-series factors. Therefore, we follow Fama and French (2020) and use firm characteristics directly in the cross-sectional regressions to estimate characteristic-based returns.

Specifically, we perform cross-sectional regressions of stock returns against firm characteristics for each period, then compute characteristic-based returns ($\mathbf{R}_t$) over time:

$$\mathbf{R}_t = (\mathbf{X}_{t-1}^\top \mathbf{X}_{t-1})^{-1} \mathbf{X}_{t-1}^\top \tilde{\mathbf{R}}_t, \tag{3.1}$$

where $\tilde{\mathbf{R}}_t$ are stock returns; and $\mathbf{X}_{t-1}$ are the firm characteristics at time $t-1$. Following Fama and French (2020), we standardize each characteristic to have zero mean and unit standard deviation in the cross-section.

Because many characteristics are highly multi-collinear, to alleviate the multi-collinearity, at each time t, we pre-screen the characteristics and exclude the ones that have high multi-collinearity with other characteristics in the cross-section. Specifically, we sequentially remove characteristics, one at a time, that has a variance inflation factor (VIF) higher than 10 and is the highest among all remaining characteristics. After the pre-screening, the VIFs of all characteristics in the cross-section are lower than 10. We then use these characteristics to perform the cross-sectional regression (3.1) and obtain the characteristics-based factors.

3.3 Building SDF with cross-sectional characteristic returns

We construct the SDF using the market portfolio as the base factor (given the theoretical foundation of CAPM) and other regression-based characteristic returns.[5] The portfolio optimization will be conducted every year using a rolling window of past 25 years, that is $T = 300$. We choose such a long training window because we intend to adopt the data splitting technique for further factor selection. The out-of-sample period is between 2000 and 2023.

[5]Other factors can be easily added if necessary.

We note that some highly-collinear characteristics have been removed when we perform the cross-sectional regression, which helps alleviate the multi-collinearity in the resulting factors in the time-series. To further mitigate the multi-collinearity in factors, we pre-screen the factors in each time period and sequentially exclude the ones that exhibit high multi-collinearity with others in the time series, based on the VIF as done previously.

We use MAXSER for a first step screening. Specifically, we split the sample into two subsamples according to even/odd years and apply MAXSER to one subsample. We will then only focus on the subset of factors with nonzero weights in the estimated MAXSER portfolio. This step reduces the large zoo of factors into a much smaller pool. However, due to the over-selection issue of LASSO, this reduced pool may still contain many redundant factors. To solve this issue, we will conduct further factor selection using the inference theory that we develop in Section 2.4.

To conduct further factor selection, we note that similar to LASSO, MAXSER actually yields a collection of sub-models indexed by the penalty parameter λ in (2.12), and the sub-models are nested and become smaller as λ increases (Efron, Hastie, Johnstone, and Tibshirani, 2004). Let $S_0 = \{i_{(1)}, i_{(2)}, \ldots, i_{(|S_0|)}\}$ denote the set of factors included in MAXSER, where the variables $(i_{(j)})_{j\leq|S_0|}$ are arranged in decreasing order of inclusion frequency in the collection of sub-models. For the last factor $i_{(|S_0|)}$, we test whether it is statistically significant using the withheld subsample and Theorem 3. Specifically, we compute the p-value

$$2\left(1-\Phi\left(\frac{|\widehat{b}^{(S_0)}_{i_{(|S_0|)}}|}{se(\widehat{b}^{(S_0)}_{i_{(|S_0|)}})}\right)\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $\widehat{b}^{(S_0)}_{i_{(|S_0|)}}$ is the plug-in estimator of the loading of $i_{(|S_0|)}$th variable from (2.14) for $i_{(|S_0|)} \in S_0$, and $se(\widehat{b}^{(S_0)}_{i_{(|S_0|)}})$ is the standard error from Theorem 3. If the p-value is greater than a pre-specified significance level τ, we repeat the procedure on the smaller sub-model $S_1 = \{i_{(1)}, i_{(2)}, \ldots, i_{(|S_0|-1)}\}$. We continue this process until the p-value of the last variable in the sub-model is smaller than τ. By doing so, we ensure that the last factor in the final sub-model is statistically significant at the chosen level τ. The reason for only testing the last variable rather than all variables is because the models in the LASSO solution path are nested, consequently, the other variables are included more frequently than the last variable, so they are likely more useful than the last variable.

Finally, the resulting SDF portfolio is estimated using the plug-in method based on the whole sample using the factors identified in the last sub-model. To facilitate comparison among SDF from different models, we impose an additional risk constraint σ, set to be the standard deviation of the market returns during the training period. We denote this portfolio as MAXSER-S(τ). For robustness check, we consider $\tau = 0.01, 0.05$ and 0.1. We summarize our proposed MAXSER-S algorithm for building SDF as follows:

Algorithm

Input: $(\mathbf{f}_t, \mathbf{R}_t)_{1\le t\le T}$, σ, τ

Output: $\widehat{\boldsymbol{\omega}}$, $\widehat{\mathbf{b}}$

Step I. Split sample into two subsets of comparable sizes $T_1 \cup T_2 = \{1, ..., T\}$, according to odd/even years.

Step II. Get weight estimator $\widehat{\boldsymbol{w}}^*_{f,u}$ via (2.12) based on $(\mathbf{f}_t, \mathbf{R}_t)_{t\in T_1}$, with λ chosen by cross-validation using the criterion of maximizing the Sharpe ratio, and get $S_0 = \{i_{(j)}; \widehat{\boldsymbol{\omega}}^*_{f,u}(i_{(j)}) \neq 0, 1 \le i \in N + K\}$ with $i_{(1)}, ..., i_{(|S_0|)}$ arranged in decreasing order of inclusion of the solution path.

Step III. Initialize $S = S_0$. While $|S| > 0$, do:

Compute $p_{i_{(|S|)}}$ for variable $i_{(|S|)}$ based on $((\mathbf{f}_t, \mathbf{R}_t)(i))_{i\in S, t\in T_2}$ and Theorem 3.

If $p_{i_{(|S|)}} < \tau$, then go to Step IV.

Else: $S \leftarrow \{i_{(1)}, ..., i_{(|S|-1)}\}$.

Step IV. Compute plugin SDF weight estimator $\widehat{\mathbf{b}}$ via (2.14) and $\widehat{\boldsymbol{\omega}} = \big((1 - |S|/T)\sigma/\sqrt{\widehat{\boldsymbol{\mu}}_S^\top \widehat{\boldsymbol{\Sigma}}_S^{-1} \widehat{\boldsymbol{\mu}}_S}\big)\ \widehat{\mathbf{b}}$, where $\widehat{\boldsymbol{\mu}}_S$ and $\widehat{\boldsymbol{\Sigma}}_S$ are the sample mean and sample covariance matrix of $((\mathbf{f}_t, \mathbf{R}_t)(i))_{i\in S, t\le T}$, respectively.[6]

3.4 Benchmark models

For comparison, we consider the following models proposed in the literature and use them as benchmark models.

[6]Note that the plug-in portfolio $\widehat{\boldsymbol{\omega}}_s$ is given by $\widehat{\boldsymbol{\omega}}_s = (\sigma/\sqrt{\widehat{\boldsymbol{\mu}}_S^\top \widehat{\boldsymbol{\Sigma}}_S^{-1} \widehat{\boldsymbol{\mu}}_S})\ \widehat{\mathbf{b}}$. Here we define $\widehat{\boldsymbol{\omega}} = (1 - |S|/T)\widehat{\boldsymbol{\omega}}_s$. The scalar $(1 - |S|/T)$ works as a finite-sample correction for out-of-sample risk of the plug-in estimator $\widehat{\boldsymbol{\omega}}_s$. Specifically, one can derive from Proposition 2 and (A.14) of Ao, Li, and Zheng (2019) that the out-of-sample risk of $\widehat{\boldsymbol{\omega}}_s$ satisfies $\widehat{\boldsymbol{\omega}}_s^\top \boldsymbol{\Sigma}_{f,R} \widehat{\boldsymbol{\omega}}_s \xrightarrow{p} \sigma^2/(1 - |S|/T)^2$. Hence, after the correction, $\widehat{\boldsymbol{\omega}}^\top \boldsymbol{\Sigma}_{f,R} \widehat{\boldsymbol{\omega}} \xrightarrow{p} \sigma^2$.

- CAPM: the market portfolio.

- FF3/FF5/FF6: plug-in optimal portfolio using Mkt-Rf, SMB and HML factors from the Fama-French three-factor model (Fama and French, 1992), additional CMA and RMW portfolios from the Fama-French five-factor model (Fama and French, 2015), and the momentum factor MOM (Fama and French, 2018).[7]

- Q/Q5: plug-in optimal portfolio using Mkt-Rf, R_ME, R_IA and R_ROE factors from the Q factor model (Hou, Xue, and Zhang, 2015) and additional R_EG factor for the Q5 factor model (Hou, Mo, Xue, and Zhang, 2021).[8]

- BS6: plug-in optimal portfolio using Mkt-Rf, SMB, IA, ROE, MOM, HML(m) factors from Barillas and Shanken (2018).[9]

- SY4: plug-in optimal portfolio using Mkt-Rf, SMB, PERF and MGMT factors from Stambaugh and Yuan (2017).[10]

- DHS3: plug-in optimal portfolio using Mkt-Rf, PEAD and FIN factors from Daniel, Hirshleifer, and Sun (2020).[11]

- KNS: the shrinkage SDF estimator in Kozak, Nagel, and Santosh (2020). The estimator is given by:

$$\widehat{\mathbf{b}}^{KNS} = \underset{\mathbf{b}}{\operatorname{argmin}} (\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\Sigma}}\mathbf{b})^\top \widehat{\boldsymbol{\Sigma}}^{-1} (\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\Sigma}}\mathbf{b}) + \gamma_1 \sum_{i=1}^{N} |\mathbf{b}_i| + \gamma_2 \mathbf{b}^\top \mathbf{b},$$

  where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are sample mean and sample covariance matrix of the training data, γ_1 and γ_2 are tuning parameters. Following Kozak, Nagel, and Santosh (2020), the tuning parameters are chosen by cross-validation with the criterion of maximizing the out-of-sample (oos) R^2:

$$R^2_{oos} = 1 - \frac{(\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\Sigma}}_2\mathbf{b})^\top (\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\Sigma}}_2\mathbf{b})}{\widehat{\boldsymbol{\mu}}_2^\top \widehat{\boldsymbol{\mu}}_2},$$

  where $\widehat{\boldsymbol{\mu}}_2$ and $\widehat{\boldsymbol{\Sigma}}_2$ are computed using the withheld sample. The weight $\widehat{\mathbf{b}}^{KNS}$ is then normalized to make the portfolio risk the same as the target risk.

[7]Monthly data of these factors are from French's data library.https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[8]Monthly data of the q-factors are from https://global-q.org/factors.html.

[9]HML(m) data are from https://pages.stern.nyu.edu/~afrazzin/data_library.htm.

[10]Monthly data of SY4 factors are from https://finance.wharton.upenn.edu/~stambaug/.

[11]Monthly data of DHS factors are from https://sites.google.com/view/linsunhome.

4 Empirical performance of SDF portfolios

4.1 Summary statistics of SDF portfolios

We summarize the out-of-sample performance of the SDF portfolios computed from our proposed approach and other benchmark models in Table 1. We report the monthly mean, standard deviation, Sharpe ratio, maximum drawdown, together with different distribution percentiles. In addition, we use the method in Ledoit and Wolf (2008) for testing the difference in Sharpe ratios between our baseline model MAXSER-S(0.05) and other models:

$$H_0 : SR_{MAXSER-S(0.05)} \leq SR_0 \qquad \text{vs.} \qquad H_1 : SR_{MAXSER-S(0.05)} > SR_0, \tag{4.1}$$

where $SR_{MAXSER-S(0.05)}$ is the Sharpe ratio of MAXSER-S(0.05) and SR_0 denotes the Sharpe ratio of another model.

Table 1: **Out-of-sample performances of SDF.** This table reports the out-of-sample performances of the optimal portfolios constructed from the proposed approach using cross-sectional regression-based characteristic returns and other benchmark models, including the MAXSER-S(τ) with various post screening significance levels (τ =0.01, 0.05, 0.1), the shrinkage SDF portfolio (KNS) by Kozak, Nagel, and Santosh (2020), and other factor models. The summary statistics include monthly mean return (Mean), standard deviation (SD), Sharpe ratio (SR), the maximum drawdown (MDD), the average number of factors in the portfolio (NUM) and the distribution percentiles. All are reported in percentage except the Sharpe ratio. The column "SR-pvalue" reports the *p*-value of testing the difference in the Sharpe ratios between the proposed MAXSER-S(0.05) and other models, while *, ** and *** indicate statistical significance at the level of 10%, 5% and 1%, respectively. The optimal portfolios are estimated every year, using a rolling window of past 25 years, and portfolio returns are calculated every month. The out-of-sample period is between 2000 and 2023.

Model	Mean	SD	SR	SR-pvalue	MDD	NUM	P1	P25	P50	P75	P99
Panel A: MAXSER-based SDF											
MAXSER-S(0.01)	4.23	5.61	0.75	0.17	29.58	28	-10.87	0.88	4.15	7.00	20.48
MAXSER-S(0.05)	4.79	6.13	0.78	–	29.58	37	-10.87	1.14	4.27	8.16	22.13
MAXSER-S(0.1)	4.87	6.16	0.79	0.16	30.44	39	-10.90	1.14	4.54	8.07	22.13
Panel B: SDF from benchmark models											
KNS	2.97	5.03	0.59	0.00 ***	22.03	17	-6.33	0.41	2.00	4.99	19.30
CAPM	0.55	4.63	0.12	0.00 ***	49.39	1	-10.40	-2.03	1.14	3.26	10.33
FF3	0.81	5.85	0.14	0.00 ***	66.32	3	-19.45	-1.89	1.08	4.26	13.78
FF5	1.82	6.64	0.27	0.00 ***	34.85	5	-11.82	-1.81	1.24	4.91	20.88
FF6	1.72	6.51	0.26	0.00 ***	37.24	6	-12.29	-1.56	1.36	4.37	18.10
Q	1.50	5.72	0.26	0.00 ***	32.46	4	-10.22	-1.47	1.58	4.02	20.49
Q5	1.91	5.39	0.36	0.00 ***	25.23	5	-11.75	-0.92	1.53	4.84	16.61
SY4	2.01	6.34	0.32	0.00 ***	27.03	4	-12.59	-1.31	2.02	4.68	19.95
DHS3	1.53	5.13	0.30	0.00 ***	37.00	3	-11.23	-1.49	1.39	4.42	13.79
BS6	1.57	6.61	0.24	0.00 ***	35.06	6	-14.84	-1.66	1.57	3.89	23.84

We see from Panel A of Table 1 that among the MAXSER-based portfolios with further selection, as the significance level increases from 0.01 to 0.05, the monthly Sharpe ratio increases from 0.75 to 0.78. When the significance level reaches 0.1, the performance is similar and the Sharpe Ratio is 0.79. The statistical tests of (4.1) show that the differences in Sharpe ratios from various MAXSER-based portfolios are statistically insignificant. For example, the difference between MAXSER-S(0.1) and MAXSER-S(0.05) is not statistically significant at the 10% level. This shows the robustness of our methodology. Therefore, balancing between the performance and the parsimony of the SDF, we will focus on MAXSER-S(0.05) as our baseline model. Hereafter, we refer to the MAXSER-S(0.05) SDF simply as MAXSER-S.

Panel B of Table 1 shows that MAXSER-S significantly outperforms all other benchmark methods. Among all benchmark portfolios, the KNS portfolio has the highest monthly Sharpe Ratio of 0.59, which is still significantly lower than that of MAXSER-S. Figure 1 plots the cumulative returns of various SDF portfolios, which illustrates the superior performance of MAXSER-S.

The out-of-sample Sharpe ratio of MAXSER-S (an annualized value of 2.7) is similar to or better than those reported in the literature using machine learning techniques. For example, using the panel-tree, Cong, Feng, He, and He (2025) report an annual Sharpe ratio of 3.12 over 2001–2020. Feng, He, Polson, and Xu (2024) show an annual Sharpe ratio of 2.95 to 3 over 2002–2021. Using deep reinforcement learning, Cong, Tang, Wang, and Zhang (2021) show an annual Sharpe ratio of 2.31 over 1990–2016, after excluding small stocks. Bryzgalova, Pelger, and Zhu (2025) find an annual Sharpe ratio of 2.39 after pruning to 40 portfolios, over 1994 to 2016.

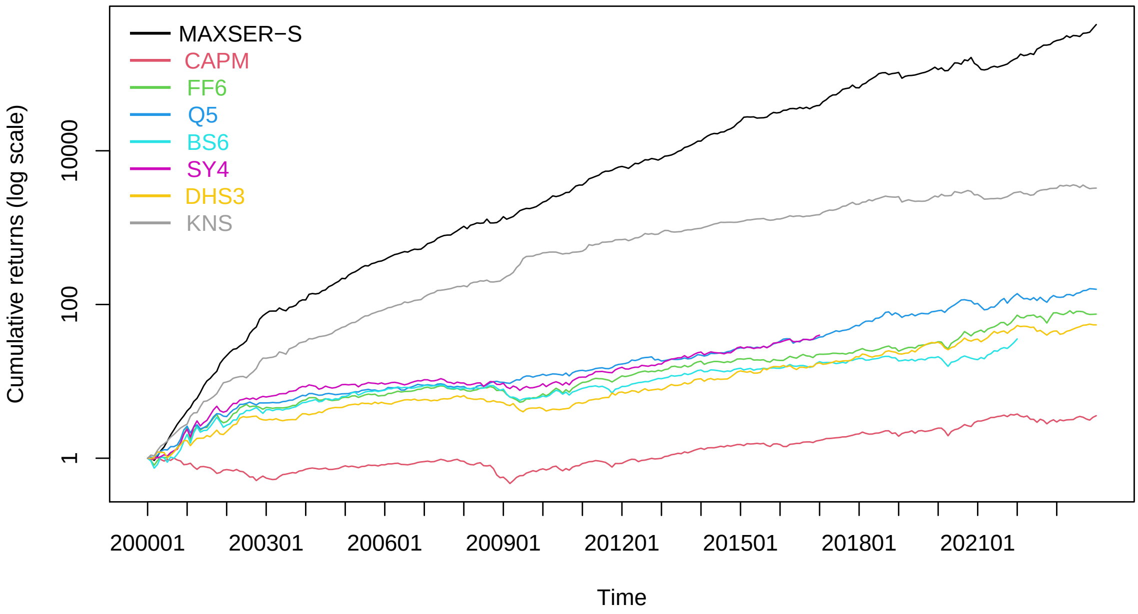


Figure 1: **Cumulative dollar returns of the SDF portfolios.** This figure plots the cumulative returns of various SDF portfolios from 2000 to 2023, starting with $1. We include the SDF portfolios constructed from MAXSER-S, CAPM, FF6, Q5, BS6, SY4, DHS3, and KNS. The portfolio risk is set to be the market risk.

4.2 Spanning tests

Table 1 suggests that MAXSER-S delivers a significantly higher Sharpe ratio than other benchmark models. One might still wonder to what extent our proposed SDF provides genuinely new pricing information or it is simply a repackaging of risks already captured by existing models. To investigate this, we perform a spanning regression, i.e., regressing the out-of-sample return of MAXSER-S against various benchmark models:

$$R_{MAXSER-S,t} = \alpha + \sum_{k=1}^{K} \beta_{ik} R_{k,t} + \varepsilon_{it}, \tag{4.2}$$

where $R_{MAXSER-S,t}$ is the return of MAXSER-S and $R_{k,t}$ is the factor return from a benchmark model.

Before presenting the results, it is important to clarify the economic hypothesis being tested. The intercept from this regression, the alpha, measures the average return of our SDF that is not accounted for by its exposures to the benchmark factors. A statistically significant, positive alpha would constitute strong evidence that our SDF contains information missed by the benchmark model, thereby offering a tangible improvement to the investor's mean-variance frontier.

Table 2 presents the alpha and its t-statistic from the regression (4.2). Newey and West (1987) robust standard error is used in calculating the t-statistic. We see from Table 2 that all benchmark models fail to fully price our MAXSER-S portfolio. The alpha of MAXSER-S is both economically large and statistically significant under all benchmark models. The monthly alpha is greater than 2% with a t-statistic higher than 6 across all models. These results suggest that the benchmark models can not capture the returns of our MAXSER-S portfolio. In other words, MAXSER-S contains new information.

Table 2: **Testing the pricing of MAXSER-S with benchmark models.** This table reports alphas from regressions of MAXSER-S against various benchmark models, using monthly returns from 2000 to 2023. Alpha in percentage and its corresponding Newey-West t-statistic are reported.

	Alpha (%)	t-statistic
CAPM	4.80	9.68
FF3	4.79	9.99
FF5	4.61	8.31
FF6	4.59	8.66
Q	4.61	7.62
Q5	4.29	7.67
SY4	4.96	6.55
BS6	4.55	8.55
DHS3	4.61	7.41
KNS	2.32	6.47

4.3 Encompassing tests

Next, we switch the roles between benchmark models and our estimated SDF and examine whether the commonly used pricing factors have significant alpha under MAXSER-S. This reverse test addresses the critical question of encompassing. A superior SDF should not only offer novel pricing information but also subsume, or encompass, the information contained in existing, well-established factors.

Specifically, we regress various pricing factors against MAXSER-S:

$$R_{k,t} = \alpha_k + \beta_k R_{MAXSER-S,t} + \varepsilon_{kt}. \tag{4.3}$$

Here, the null hypothesis is that the alpha is zero. A statistically insignificant alpha provides evidence that our SDF encompasses the benchmark factor, i.e., the factor's premium is captured by its exposure to our SDF. As a result, a zero alpha from such an encompassing test would suggest that the benchmark factor does not offer marginal improvement to the Sharpe ratio beyond what our SDF already provides.

Table 3 reports the alpha and its t-statistic from regression (4.3). We see from Table 3 that all alphas are insignificant at the t-statistic threshold level of 1.96. These results suggest that our SDF encompasses those commonly used pricing factors.

Table 3: **Testing the pricing of commonly used factors with MAXSER-S.** This table reports alphas from regressions of commonly used pricing factors against MAXSER-S, using monthly returns from 2000 to 2023. Alpha in percentage and its corresponding Newey-West t-statistic are reported.

Factor	Alpha (%)	t-statistic
Mkt.RF	0.61	1.59
SMB	0.42	1.36
HML	-0.07	-0.22
RMW	0.06	0.25
CMA	0.21	1.26
Mom	-0.05	-0.16
R_ME	0.46	1.41
R_IA	0.15	0.78
R_ROE	0.01	0.03
R_EG	0.08	0.41
HmLm	0.02	0.05
MGMT	-0.18	-0.76
PERF	0.34	0.68
PEAD	0.32	1.65
FIN	0.07	0.20
KNS	0.28	0.94

4.4 Asset pricing tests

Finally, we perform some standard asset pricing tests, following Hou, Mo, Xue, and Zhang (2021). That is, we test the pricing power of various factor models over a set of anomaly portfolios. Specifically, we use the 153 value-weighted long-short characteristic-sorted portfolios from Jensen, Kelly, and Pedersen (2023) as the test assets and regress those anomaly portfolios against various factor models to examine whether alpha is significant:

$$R_{i,t} = \alpha_i + \sum_{k=1}^{K} \beta_{ik} R_{k,t} + \varepsilon_{it}, \tag{4.4}$$

where $R_{i,t}$ is the return of anomaly portfolio i and $R_{k,t}$ is the return of factor k in a factor model. We use the SDF directly for MAXSER-S or KNS (see, e.g., Ghosh, Julliard, and Taylor, 2017, 2025; Cong et al., 2025), while using pricing factors for other factor models.

Table 4 summarizes the number of significant alphas for a given t-statistic level. We see that the proposed MAXSER-S portfolio has the lowest number of significant alphas. Under MAXSER-S, only 5 anomalies have significant alphas when using a threshold of t=1.96, while no significant alphas when using a threshold of t=3. On the other hand, all other benchmark models have more rejections. The next best models are Q5 and KNS. For example, given a threshold of t=1.96, Q5 has 21 significant alphas while KNS has 14 significant alphas; given a threshold of t=3, Q5 has 1 significant alpha. Overall, we see that MAXSER-S performs better in the low risk and profit growth themes than Q5 and it is better than KNS in the debt issuance, seasonality, and short-term reversal themes. The superior performance of MAXSER-S in minimizing pricing errors across a wide range of anomalies suggests that it provides a more comprehensive description of the cross-section return variations than existing models.

Table 4: **The number of significant alphas from testing 153 anomalies against various factor models and SDFs.** We examine the pricing power of various models over 153 anomalies. This table summarizes the total number of significant alphas and the number of significant alphas in each characteristic theme. The number in parenthesis is the total number of portfolios in each characteristic theme. The number of significant alphas under the threshold $t > 1.96$ or $t > 3$ is reported, where t-statistic uses Newey and West (1987) robust standard error. The evaluation period is between 2000 and 2023.

	All (153)		Accruals (6)		Debt Issuance (7)		Investment (22)	
Threshold of t	1.96	3	1.96	3	1.96	3	1.96	3
MAXSER-S	5	0	1	0	0	0	0	0
KNS	14	0	0	0	2	0	0	0
CAPM	59	21	0	0	1	1	4	0
FF3	75	45	2	0	2	1	6	1
FF5	38	13	2	0	1	1	2	0
FF6	39	11	2	0	1	1	3	0
Q	37	12	2	1	1	1	2	0
Q5	21	1	2	0	0	0	0	0
BS6	50	21	2	1	1	1	4	0
SY4	24	3	0	0	0	0	1	0
DHS3	28	4	1	0	1	0	1	0

	Low Leverage (11)		Low Risk (18)		Momentum (8)		Profit Growth (12)	
Threshold of t	1.96	3	1.96	3	1.96	3	1.96	3
MAXSER-S	0	0	0	0	0	0	0	0
KNS	0	0	1	0	1	0	1	0
CAPM	1	0	16	5	2	0	4	1
FF3	3	2	17	13	2	0	4	1
FF5	5	2	3	0	0	0	3	1
FF6	5	1	3	0	0	0	3	1
Q	1	0	3	0	0	0	2	2
Q5	1	0	3	0	0	0	3	1
BS6	7	2	3	0	0	0	2	2
SY4	0	0	1	0	1	0	1	1
DHS3	3	0	1	0	0	0	2	0

Table 4: **The number of significant alphas from testing 153 anomalies against various SDFs, continued.**

	Profitability (11)		Quality (17)		Seasonality (12)		Short-Term Reversal (5)	
Threshold of t	1.96	3	1.96	3	1.96	3	1.96	3
MAXSER-S	0	0	3	0	0	0	1	0
KNS	0	0	3	0	2	0	3	0
CAPM	8	6	12	7	3	0	2	0
FF3	9	7	15	13	2	0	2	2
FF5	4	1	12	5	1	1	1	1
FF6	4	0	11	5	1	0	1	1
Q	5	2	14	5	1	0	2	0
Q5	1	0	5	0	2	0	1	0
BS6	6	2	14	11	3	1	1	0
SY4	3	0	5	0	2	0	0	0
DHS3	3	0	9	2	2	0	1	1
	Size (5)		Value (18)					
Threshold of t	1.96	3	1.96	3				
MAXSER-S	0	0	0	0				
KNS	0	0	1	0				
CAPM	1	0	5	1				
FF3	1	0	10	5				
FF5	1	1	2	1				
FF6	1	1	3	1				
Q	1	1	3	0				
Q5	2	0	1	0				
BS6	1	1	6	0				
SY4	3	0	7	2				
DHS3	1	1	3	0				

4.5 Investigating the tradability issue

Trading strategies based on anomalies or statistical learning often suffer from tradability issues, especially for small stocks (Patton and Weller, 2020; Jensen, Kelly, Malamud, and Pedersen, 2025). Although we exclude the bottom decile of stocks, one might still concern about the tradability issue of our proposed SDF. We formally evaluate this issue in this subsection.

4.5.1 Examining the stock-level portfolio weights

We first investigate whether our SDF portfolio has extreme allocations to certain stocks or heavily invests in small stocks. From the cross-sectional regression, we can compute the investment weights on individual stocks as

$$\widehat{\boldsymbol{\omega}}_t^s = \widehat{\boldsymbol{\omega}}_t^\top (\mathbf{X}_{t-1}^\top \mathbf{X}_{t-1})^{-1} \mathbf{X}_{t-1}^\top,$$

where $\widehat{\boldsymbol{\omega}}_t$ are investment weights of the SDF portfolio on various factors from the cross-section regressions. In Table 5, we summarize the distribution of the stock-level investment weights. We see that MAXSER-S does not have extreme weights in stocks. Only 1.3% of stocks have an absolute weight greater than 1%. We also examine the maximum weight observed in each training window and find that the maximum weight is about 3% only. Overall, we see that our SDF portfolio does not assign extreme weights to any single stock.

Second, we examine how investment weights are distributed across stocks with different size. We sort all stocks into 10 groups, using NYSE size deciles. Then we compute the average of the absolute weights. The results are reported in Table 6. We see that the average weight of a stock is about 0.2% across all 10 size-sorted groups. Although the smallest size group has a slightly higher absolute weight of 0.28%, we don't see material difference in investment weights for stocks with different size.

In fact, as our estimator extends L1 regularization on the idiosyncratic component, together with using sureâĂŠscreening and postâĂŠselection inference to keep only economically meaningful characteristics, such regularization acts as turnover and concentration controls. That is, the MAXSER-S portfolio does not concentrate on small stocks.

Table 5: **Distribution of stock-level investment weights.** This table reports the fraction of the absolute weights in the SDF portfolio which are greater than various levels.

|Weight|	>3%	>2%	>1%	>0.5%
Fraction%	0.024%	0.13%	1.3%	8.7%

Table 6: **The average investment weights across 10 size-sorted groups.** This table reports the average of the absolute weights for 10 size-sorted groups, using NYSE size deciles.

NYSE Decile	Q1	Q2	Q3	Q4	Q5
Average |Weight|	0.0028	0.0022	0.0020	0.0019	0.0019
NYSE Decile	Q6	Q7	Q8	Q9	Q10
Average |Weight|	0.0018	0.0018	0.0018	0.0017	0.0018

4.5.2 Evaluating the impacts of transaction costs

In this subsection, we explicitly evaluate the impacts of transaction costs on the portfolio. We will add transaction costs to SDF constructed from MAXSER-S and KNS only, while leaving all benchmark models cost free, because we only have readily available data of portfolio weights for MAXSER-S and KNS. We compute stock-level transactions as follows. Based on the stock-level portfolio weights, the net returns with transaction costs subtracted are

$$r_{net,t} = \left(1 - c\sum_{t=1}^{T_{oos}} \|\widehat{\boldsymbol{\omega}}_{t+}^{s} - \widehat{\boldsymbol{\omega}}_{t+1}^{s}\|_1\right)(1 + r_t) - 1,$$

where r_t and $r_{net,t}$ represent the return of the SDF portfolio at month t before and after subtracting the transaction cost, $\widehat{\omega}_{it+}^{s}$ is the weight of asset i at the end of month t, $\widehat{\omega}_{it+1}^{s}$ is the weight of asset i at the beginning of month $t+1$, c is the transaction cost per dollar, and T_{oos} is the length of the portfolio's evaluation period.

Kan, Wang, and Zhou (2022) suggest using 20 bps per dollar of transaction when evaluating the post-transaction cost performance of investment in anomaly portfolios. We check the performance of the SDF portfolio after deducting the transaction costs, assuming a cost of c =10 bps or 20 bps per dollar of transaction. The transaction cost of the market portfolio is set to be 20 bps.

Note that transaction costs are closely related to the portfolio's turnover. We compute the monthly turnover at the stock level: $\sum_{t=1}^{T_{oos}} \sum_i |\widehat{\omega}_{it+}^{s} - \widehat{\omega}_{it+1}^{s}|/T_{oos}$. We find that the MAXSER-S has a monthly turnover of 9.10. Then with the absolute monthly weight exposure defined as $\sum_{t=1}^{T_{oos}} \sum_i |\omega_{it}^{s}|/T_{oos}$, we find that MAXSER-S has an absolute monthly weight exposure of 10.03. Therefore our SDF portfolio turnover per dollar exposure is 90.45%.

The performance of our SDF estimator after adjusting for the transaction costs is summarized in Table 7. We see from Table 7 that, the MAXSER-S maintains a high Sharpe ratio after adjusting for the transaction costs. This could be due to the trading diversification caused by combining characteristics into a portfolio. For example, DeMiguel, Martin-Utrera, Nogales, and Uppal (2020) and DeMiguel, Martín-utrera, and Uppal (2024) argue that combining characteristics could cancel trades over individual stocks during portfolio rebalancing. We see that MAXSER-S has a Sharpe ratio of 0.64 after removing 10 bps of transaction cost per dollar, and 0.48 when the cost is 20 bps per dollar. Meanwhile, KNS still has a much lower Sharpe ratio than that of MAXSER-S after adjusting for transaction costs.

Table 7: **Performance of MAXSER-S and KNS after adjusting for transaction costs.** The transaction cost is assumed to be 10 bps (20 bps) per dollar of transaction at stock level. The portfolios are re-estimated every year from 2000 to 2023. The portfolios are learned with cross-sectional characteristic projected portfolios. The summary statistics include the monthly mean return (Mean), standard deviation (SD), Sharpe ratio (SR), and the maximum drawdown (MDD) of the SDF portfolio. All are reported in percentage except the Sharpe ratio.

	Transaction cost	Mean	SD	SR	MDD
MAXSER-S	10 bps	3.84	6.04	0.64	31.70
	20 bps	2.88	5.95	0.48	33.71
KNS	10 bps	2.16	4.91	0.44	26.29
	20 bps	1.35	4.81	0.28	43.17

Next, we repeat Table 2, adjusting for transaction costs. That is, we regress the MAXSER-S SDF portfolio against various factor models (Eq. (4.2)) in Table 8. We subtract transaction costs from MAXSER-S and KNS. For the variables from other benchmark models, we use the original data without removing the transaction cost. We see that after removing the transaction costs, our MAXSER-S SDF portfolio still has statistically significant alphas against benchmark models, with a t-statistic above 3 in all cases, though the alphas are smaller than those reported in Table 2.

Table 8: **Testing the pricing of MAXSER-S with benchmark models, adjusting for transaction costs.** This table presents alphas from regressions of MAXSER-S against various benchmark models, using the monthly returns from 2000 to 2023. The portfolios are learned with cross-sectional characteristic projected portfolios. The transaction costs of MAXSER-S and KNS are assumed to be 10 bps or 20 bps per dollar of transaction for stocks. Transaction cost for the Market portfolio is 20 bps. The pricing factors from various benchmark models are the original data without removing the transaction costs. The portfolios are re-estimated every year. We report the alpha in percentage and Newey-West t-statistic.

Transaction cost	10 bps		20 bps	
Model	Alpha (%)	t-statistic	Alpha (%)	t-statistic
CAPM	3.84	10.72	2.90	8.19
FF3	3.83	10.70	2.88	8.17
FF5	3.66	9.83	2.71	7.40
FF6	3.64	9.78	2.69	7.36
Q	3.65	9.93	2.70	7.46
Q5	3.34	8.91	2.40	6.49
SY4	3.98	9.39	3.00	7.20
BS6	3.59	9.41	2.64	7.03
DHS3	3.67	9.87	2.70	7.40
KNS	2.03	7.11	1.76	6.51

Then, we repeat Table 3 but take into account of transaction costs. We regress various pricing factors against our SDF after adjusting for transaction costs (Eq. (4.3)). For the pricing factors from other benchmark models except KNS, we use their original returns without removing the transaction cost. The results are summarized in Table 9. We see that alphas of pricing factors from other benchmark models remains insignificant at a t threshold of 1.96. That is, our SDF captures the factor premia of these prevailing factors.

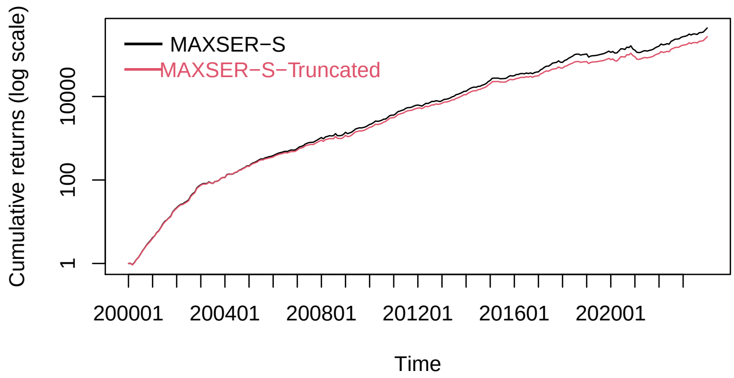
Table 9: **Testing the pricing of commonly used factors with MAXSER-S, adjusting for transaction costs.** This table reports alphas from regressions of various pricing factors against MAXSER-S after adjusting for the transaction costs, using the monthly returns from 2000 to 2023. The transaction cost of MAXSER-S and KNS is assumed to be 10 bps or 20 bps per dollar of transaction for stock level. The pricing factors from other models are the original returns without removing the transaction costs. Alpha in percentage and its Newey-West t-statistic are reported.

Transaction cost	10 bps		20 bps	
Factor	Alpha (%)	t-statistic	Alpha (%)	t-statistic
Mkt.RF	0.61	1.81	0.60	1.87
SMB	0.39	1.45	0.36	1.53
HML	-0.01	-0.05	0.04	0.17
RMW	0.13	0.58	0.21	1.03
CMA	0.23	1.48	0.24	1.66
Mom	-0.03	-0.14	-0.01	-0.05
R_ME	0.44	1.53	0.41	1.68
R_IA	0.16	0.95	0.18	1.13
R_ROE	0.06	0.25	0.12	0.55
R_EG	0.15	0.79	0.23	1.24
HmLm	0.08	0.22	0.14	0.41
MGMT	-0.07	-0.32	0.05	0.25
PERF	0.38	0.84	0.44	1.04
PEAD	0.31	1.79	0.30	1.92
FIN	0.16	0.54	0.25	0.95
KNS	0.04	0.14	-0.22	-0.81

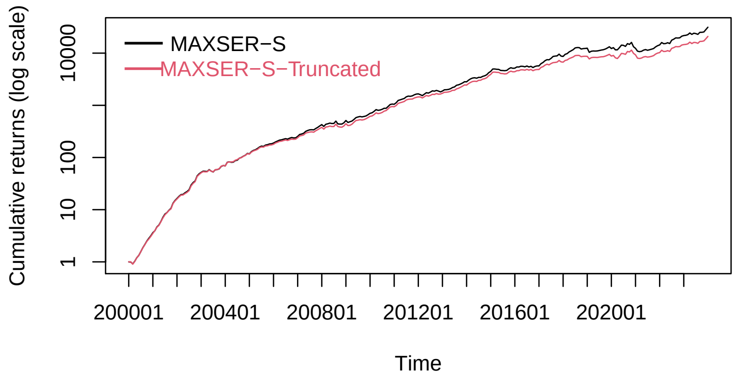
Last, we examine whether the superior performance of MAXSER-S is driven by extreme portfolio weights. To address this, we construct a truncated portfolio by capping the absolute weight at 1% for each stock. In Table 10, we summarize the performance of the truncated portfolios with and without transaction cost. Because the MAXSER-S differs from the truncated portfolio only in their extreme weights, a substantial performance gap would suggest that the extreme weights play a significant role. However, on the contrary, as shown in Table 10 and illustrated in Figure 2, we see that the truncated portfolio performs similarly to MAXSER-S both before and after adjusting for the transaction cost. For example, MAXSER-S still delivers a monthly Sharpe ratio of 0.78 after truncating the portfolio weights. This clears the doubt that the superior performance of MAXSER is driven by extreme weights.

Table 10: **Performance of MAXSER-S with and without weight truncation.** We summarize the performance of the truncated portfolio (denoted by MAXER-S-Truncated) before and after adjusting for transaction costs, compared with that of MAXSER-S without the truncation. The truncated portfolio caps the absolute portfolio weight at 1%. The transaction cost is assumed to be 10 bps (20 bps) per dollar of transaction at stock level. The summary statistics include the monthly mean return (Mean), standard deviation (SD), Sharpe ratio (SR), and the maximum drawdown (MDD) of the SDF portfolio. All are reported in percentage except the Sharpe ratio.

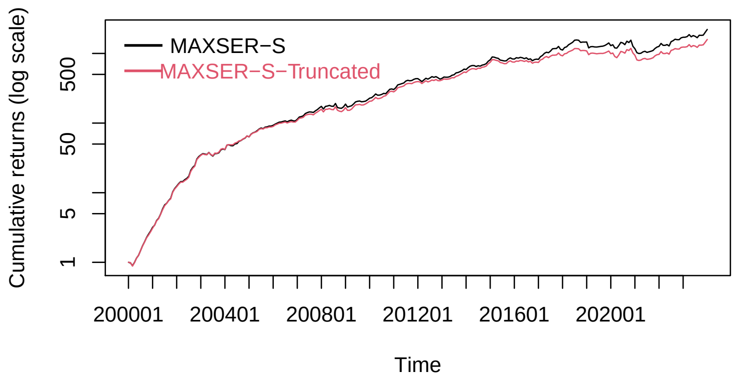
Transaction cost	Mean	SD	SR	MDD	Mean	SD	SR	MDD
	MAXER-S-Truncated				MAXER-S			
No costs	4.60	5.89	0.78	27.24	4.79	6.13	0.78	29.58
10 bps	3.67	5.79	0.63	29.36	3.84	6.04	0.64	31.70
20 bps	2.75	5.70	0.48	31.37	2.88	5.95	0.48	33.71



(a) No transaction cost



(b) Transaction costs of 10 bps



(c) Transaction costs of 20 bps

Figure 2: **Cumulative dollar returns of the MAXSER-S with and without truncating portfolio weights.** This figure plots the cumulative returns of the MAXSER-S with and without truncating portfolio weights, starting with $1. To avoid the extreme portfolio weights, the truncated portfolio caps the absolute portfolio weights at 1%. Panels (a)–(c) present the results with zero transaction costs, transaction costs of 10 bps or 20 bps, respectively.

4.6 Investigating the economic sources of MAXSER-S

The superior out-of-sample performance of MAXSER-S prompts a deeper investigation into its composition. In this section, we dissect the MAXSER-S to understand which characteristics are the primary drivers of its success, and we leverage our novel inference framework to assess the statistical and economic significance of their contributions.

4.6.1 Identifying key characteristics

We begin by identifying the characteristics that are most persistently selected by our methodology. Since we re-estimate the SDF portfolio annually using a rolling window from 2000 to 2023, we have 24 distinct portfolios. We can therefore measure the importance of a characteristic by its inclusion frequency across these periods.

Figure 3 presents the characteristics that were included in the estimated MAXSER-S portfolio more than 50% of the time (i.e., at least 13 out of 24 years). Aside from the market portfolio (Mkt-RF), which is always included as a base factor, a clear set of important characteristics emerges. The six most dominant characteristics, included in 100% of the trained portfolios, are share turnover over 126 days (turnover_126d), 1-month short-term reversal (ret_1_0), R&D scaled by market capitalization (rd_me), price relative to its 52-week high (prc_highprc_252d), performance-based mispricing (mispricing_perf), idiosyncratic volatility from CAPM in the last one year (ivol_capm_252d). Five characteristics are included in 96% of time, including the number of zero trades day in the last one year (zero_trades_252d), the highest 5 days of return scaled by volatility (rmax5_rvol_21d), price momentum (ret_6_1), quality minus junk - profitability (qmj_prof), liquidity scaled by lagged market assets (aliq_mat). These characteristics are primarily associated with the low risk, short-term reversal, quality, and momentum anomaly themes, as defined by Jensen, Kelly, and Pedersen (2023). Some of these important characteristics are related to liquidity and transaction costs (see, e.g., Chen and Velikov, 2023). Or, from the behavioral perspective, these factors might be attributed to investor attention, sentiment, or other behavioral bias.

As Figure 3 illustrates, the composition of our estimated SDF dynamically changes over time, a feature that is both intentional and economically intuitive. First, our rolling-window approach ensures the SDF responds to the dynamic economic environment; it naturally generates a time-varying SDF as it re-estimates. This is crucial for capturing the effects of shocks to investment opportunities or even structural changes the U.S. economy might experience over time, such as shifts in industry structure or the fiscal and monetary environment. Second, the characteristic portfolios might contain behavioral traits which could be time-varying, for instance, market sentiment or extrapolative beliefs (see, e.g., Baker and Wurgler, 2006; Cassella and Gulen, 2018). This leads to a time-varying SDF. Last, a time-varying SDF could be driven by variations in economic primitives, as discussed in conditional asset pricing models (see, e.g., Nagel and Singleton, 2011; Kelly, Pruitt, and Su, 2019).

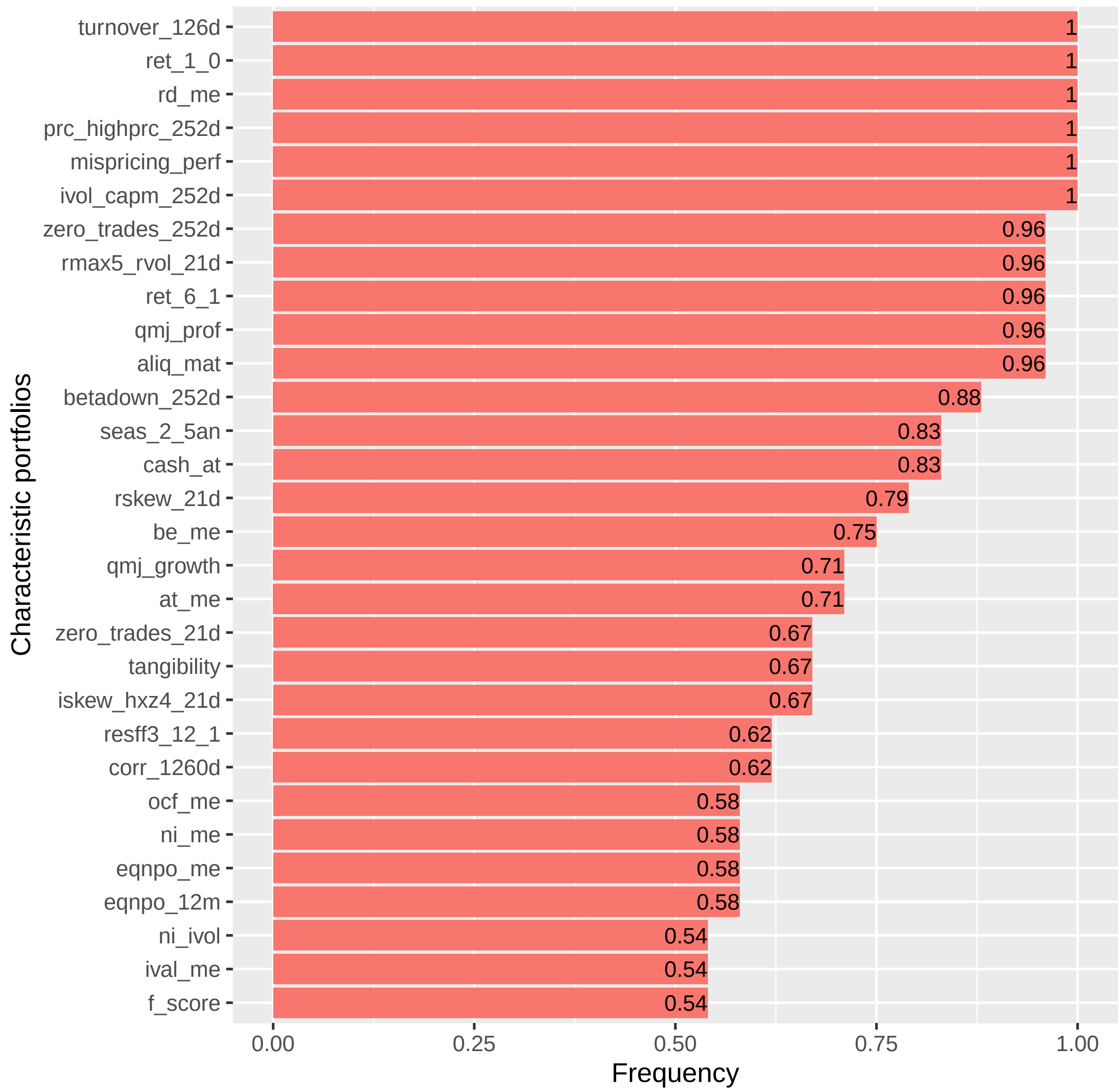


Figure 3: **Frequencies of characteristics included in MAXSER-S.** This figure reports the inclusion frequency of characteristics in the MAXSER-S from all training periods which have an inclusion frequency higher than 50%. We re-estimate the SDF every year from 2000 to 2023 using a rolling window of past 25 years. The market portfolio is always included hence omitted in the graph.

4.6.2 Properties of selected characteristics

Next, we investigate the selected characteristics. One might wonder whether our approach merely picks up the most significant anomalies. For example, linear factor models often add significant anomalies as additional pricing factors. We examine the significance of selected characteristics in Table 11. We provide summary statistics for the most frequently selected characteristics in MAXSER-S, including the t-statistics of their alphas against various benchmark models.

The results reveal a key insight: many of the selected characteristics do not exhibit statistically significant alphas. Taking the CAPM as an example, only 14 of the top 30 characteristics show a t-statistic greater than 1.96. The rest 16 characteristics do not have a significant alpha under CAPM and hence might be neglected by conventional factor models. In contrast, our approach retains them because they reduce overall portfolio risk through diversification, thereby improving the Sharpe ratio. This highlights the primary advantage of our methodology: rather than merely chasing significant anomalies, it accounts for cross-sectional dependence to select characteristics that work together to maximize the SDF's Sharpe ratio.

4.6.3 Statistical inference on theme-level factor premia

We further explore premia provided by characteristics. Instead of presenting results over individual characteristics, we consider the contribution of each characteristic theme to the SDF premium. We follow the classifications of Jensen, Kelly, and Pedersen (2023). The contribution of theme G to the SDF premium can be calculated as $\widehat{\boldsymbol{\omega}}_G^\top \boldsymbol{\mu}_G$, where $\widehat{\boldsymbol{\omega}}_G$ is the learned SDF weight of characteristics in theme G in the testing year and $\boldsymbol{\mu}_G$ is the average monthly return in that year.

Table 11: **Summary statistics of the most frequently selected characteristics in MAXSER-S.** This table reports the summary statistics of the most frequently selected characteristics (excluding the market portfolio) in constructing the MAXSER-S, including the frequency, theme name, and the Newey-West t-statistics of alphas when regressing a characteristic against some benchmark factor models, using the monthly returns from 2000 to 2023.

Variable	Freq	Theme	CAPM	FF3	FF5	FF6	Q	Q5	SY4	DHS3	BS6
turnover_126d	1	Low Risk	-3.94	-4.61	-3.94	-3.87	-3.85	-2.59	-3.68	-2.25	-3.13
ret_1_0	1	Short-Term Reversal	-2.99	-3.08	-3.61	-3.01	-2.61	-2.74	-2.18	-2.90	-2.95
rd_me	1	Size	5.22	4.82	5.61	5.52	5.10	4.62	3.05	4.90	5.25
prc_highprc_252d	1	Momentum	2.14	2.21	1.79	1.49	0.86	0.96	-0.18	0.30	1.42
mispricing_perf	1	Quality	3.82	3.97	3.52	3.22	3.28	3.15	3.57	3.20	3.18
ivol_capm_252d	1	Low Risk	-4.71	-4.60	-3.71	-3.63	-3.80	-3.32	-2.78	-4.07	-3.54
zero_trades_252d	0.96	Low Risk	1.38	1.41	1.21	1.35	1.50	0.75	2.10	1.50	1.84
rmax5_rvol_21d	0.96	Short-Term Reversal	-3.72	-3.60	-3.06	-3.18	-3.52	-3.50	-4.93	-3.15	-4.19
ret_6_1	0.96	Momentum	1.94	1.87	1.76	1.67	1.04	1.12	0.61	1.50	0.94
qmj_prof	0.96	Quality	2.34	2.61	1.43	1.51	1.98	1.44	1.90	1.82	1.80
aliq_mat	0.96	Low Leverage	1.96	1.92	1.85	1.78	1.28	1.71	1.15	1.57	0.67
betadown_252d	0.88	Low Risk	-2.22	-2.50	-1.70	-1.70	-1.76	-1.09	-1.88	-1.79	-1.69
seas_2_5an	0.83	Seasonality	1.59	1.63	1.75	1.73	1.55	0.83	2.24	1.71	1.53
cash_at	0.83	Low Leverage	1.16	2.10	2.96	2.83	2.26	1.34	0.37	1.99	2.79
rskew_21d	0.79	Short-Term Reversal	4.17	4.19	5.16	4.60	4.86	4.58	4.68	3.80	4.31
be_me	0.75	Value	1.58	1.47	0.91	1.22	2.19	1.81	1.90	1.77	1.62
qmj_growth	0.71	Quality	3.32	3.43	3.25	3.28	3.37	3.43	2.82	3.57	3.52
at_me	0.71	Value	1.11	0.34	1.70	1.59	0.11	0.66	0.05	0.22	-0.05
zero_trades_21d	0.67	Low Risk	-2.56	-3.34	-2.28	-1.69	-2.10	-1.41	-3.58	-1.60	-2.55
tangibility	0.67	Low Leverage	1.53	1.32	1.08	1.14	1.05	2.06	2.85	1.54	0.73
iskew_hxz4_21d	0.67	Short-Term Reversal	2.15	2.27	2.45	2.53	2.44	2.50	7.15	2.45	2.39
resff3_12_1	0.63	Momentum	0.98	1.02	1.02	1.00	1.19	0.58	1.44	1.17	0.58
corr_1260d	0.63	Seasonality	-4.30	-4.45	-5.25	-4.93	-4.96	-4.64	-3.24	-5.02	-4.69
ocf_me	0.58	Value	-0.15	-0.37	-0.42	-0.33	-0.18	-0.20	0.18	0.10	-0.26
ni_me	0.58	Value	1.10	1.08	0.72	0.67	0.69	0.24	0.19	0.44	0.55
eqnpo_12m	0.58	Value	2.16	2.06	2.30	2.12	1.85	1.76	1.70	1.89	1.19
eqnpo_me	0.58	Value	0.63	0.62	1.01	0.89	0.16	0.24	-0.56	0.35	0.00
ni_ivol	0.54	Low Leverage	-0.11	-0.25	-0.38	-0.54	-0.48	-0.68	-1.21	-0.44	-0.52
ival_me	0.54	Value	-1.57	-1.55	-1.84	-1.91	-2.38	-2.11	-1.72	-1.90	-2.23
f_score	0.54	Profitability	-0.59	-0.60	-0.79	-0.99	-1.65	-1.31	0.33	-0.73	-1.70

Figure 4 plots a stacked bar chart to compare the contribution of each characteristic theme to the SDF premium. Overall, we see that the "Low Risk" theme stands out as a persistently important contributor to the SDF premium. Other themes exhibit more dynamic behavior. For example, "Short-Term Reversal" theme becomes notably more pronounced in recent years, while "Quality" theme is more sizable in earlier years. Figure 4 also demonstrates the SDF's ability to adapt to changing market regimes. This is most evident in the performance of some specific themes during periods of market stress. The "Momentum" theme, for instance, delivered a significant negative premium during the market crash in 2009 and again during the COVID-19 turmoil in 2020. This dynamic exposure underscores the necessity of a time-varying SDF that can capture how different factors contribute to risk and return as economic conditions shift.

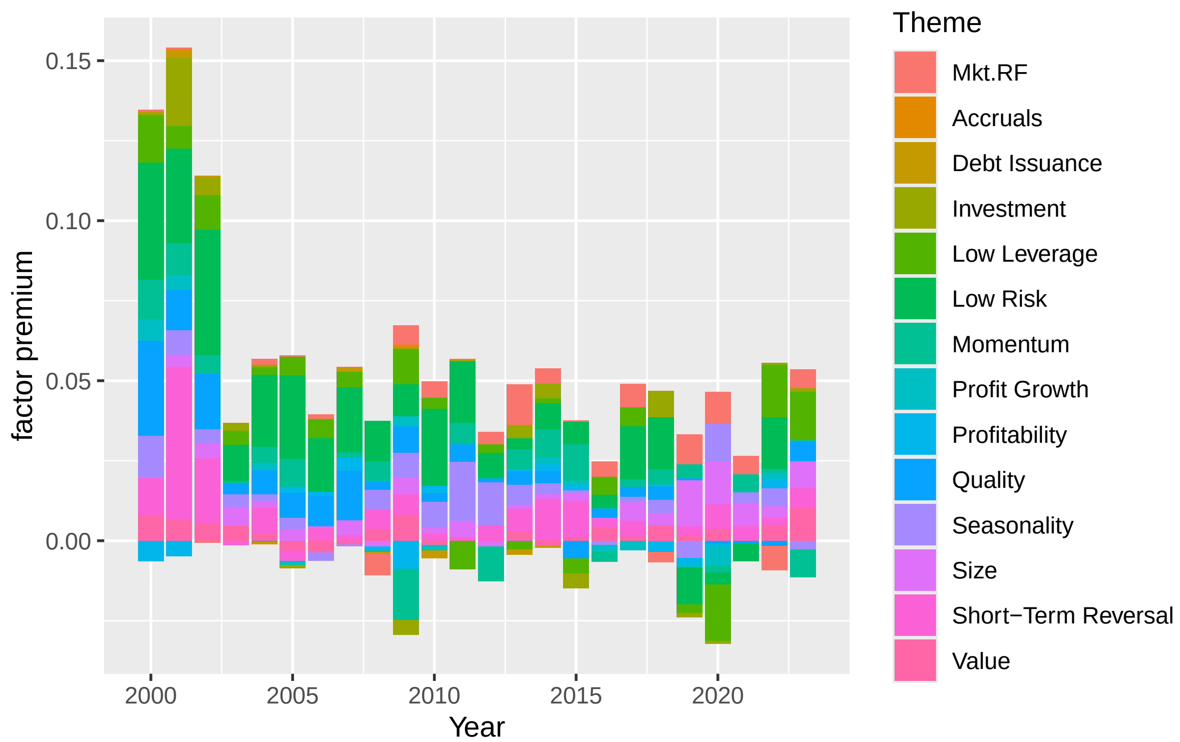


Figure 4: **The contribution of each characteristic theme to the MAXSER-S premium.** This barchart plots the average portfolio returns of each characteristic theme of MAXSER-S in each year from 2000 to 2023.

While Figure 4 illustrates the economic magnitude of each theme's contribution, our novel inference framework allows us to assess their statistical significance. This marks a key contribution of our paper: the development of a statistical inference theory (Theorem 3) that allows us to move beyond point estimates and formally assess the statistical significance of the economic contributions from different characteristic themes, crucially, in the presence of other characteristics.

Specifically, for each theme G, the estimated contribution to SDF is $\widehat{\boldsymbol{\omega}}_G^\top \boldsymbol{\mu}_G$, where recall that $\widehat{\boldsymbol{\omega}}_G$ is the estimated weights on theme G, and $\boldsymbol{\mu}_G$ is the premium vector of factors in theme G. To build a 95% confidence interval of the estimated theme contribution, we apply Theorem 3 and set $\mathbf{A} = \boldsymbol{\mu}_G^\top$. In the analysis below, for each year of evaluation, we set $\boldsymbol{\mu}_G$ to be the average monthly returns in that year. Figure 5 reports these contributions and their confidence intervals over time, including the factor premium of the whole SDF and the contribution of each theme, sorted by their overall importance.

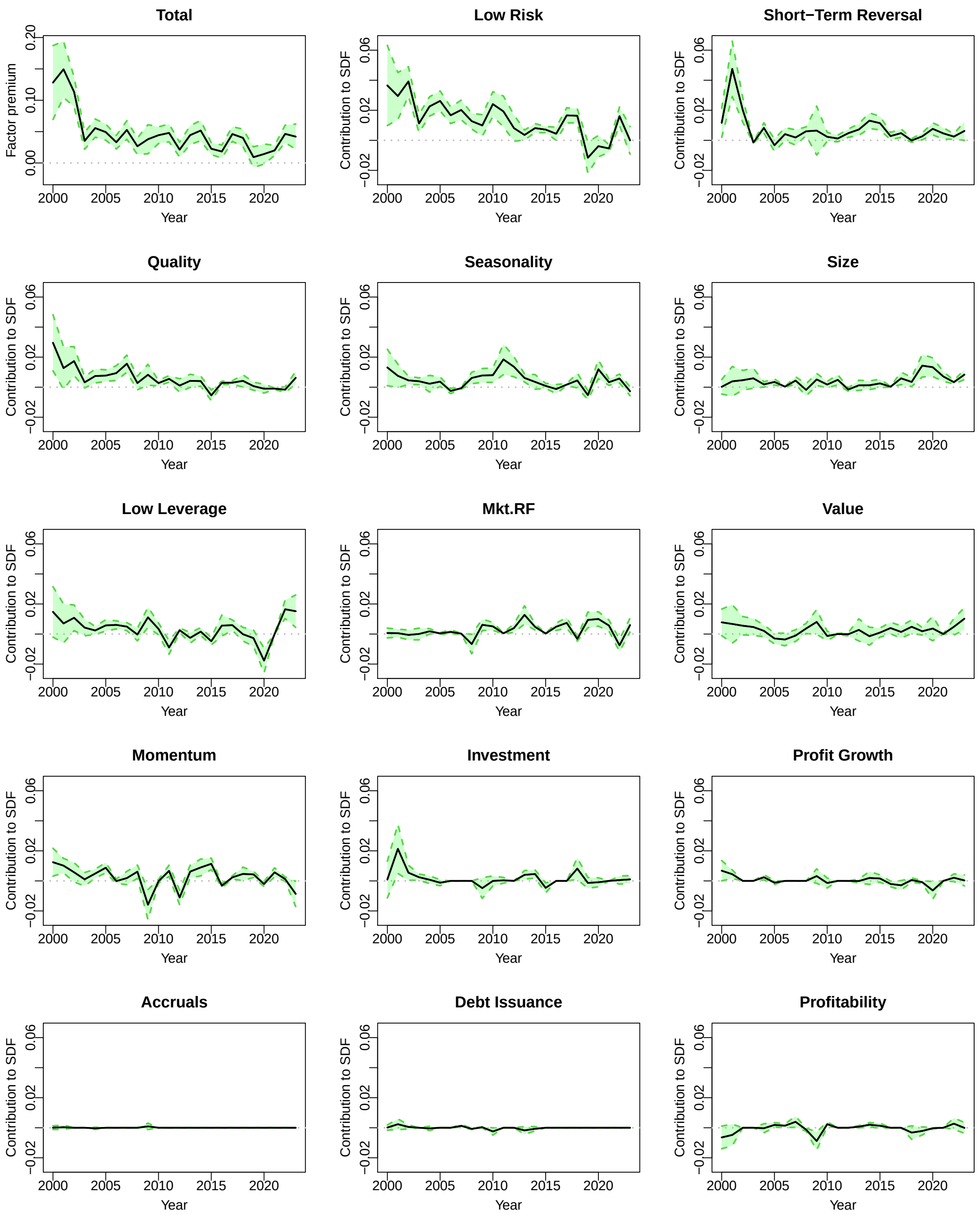


Figure 5: **Time series of theme-level MAXSER-S premia and 95% confidence intervals.** The solid line presents the evolution of the estimated monthly SDF premium as well as the contribution of each characteristic theme in each year from 2000 to 2023. The shaded areas represent the 95% confidence intervals of the premia calculated using the inference framework from Theorem 3.

The ability to construct these confidence intervals provides a powerful diagnostic tool. For example, Figure 5 reveals that the "Low Risk" theme provides a persistently positive and statistically significant premium to the SDF for most of the sample period. Low risk anomaly has been well documented in the literature (see, e.g., Frazzini and Pedersen, 2014; Schneider, Wagner, and Zechner, 2020). The "Short-Term Reversal" and "Quality" themes also show periods of significant positive contributions, consistent with previous studies (see, e.g., Jegadeesh, 1990; Asness, Frazzini, and Pedersen, 2019). In contrast, themes like "Accruals", "Debt Issuance", and "Profitability" are not only economically small but also statistically insignificant, suggesting there are unimportant. The reason could be due to cross-characteristics overlap (Green, Hand, and Zhang, 2017; Linnainmaa and Roberts, 2018; Jensen, Kelly, and Pedersen, 2023; Cakici, Fieberg, Metko, and Zaremba, 2024; Goyal, Welch, and Zafirov, 2024). For example, accruals anomaly may be subsumed by firm investment (Wu, Zhang, and Zhang, 2010); debt issuances and profitability correlate with earnings momentum, quality, investment, size, and growth options (Bouchaud, Krueger, Landier, and Thesmar, 2019; Bali, Del Viva, Lambertides, and Trigeorgis, 2020; Davydiuk, Marchuk, and Shaliastovich, 2024). Such formal inference allows us to rigorously differentiate important drivers of the SDF premium from those whose contributions are likely due to noise.

4.7 Alternative MAXSER-S SDF: Using long-short characteristic-sorted portfolios

Previously, we constructed the MAXSER-S SDF using characteristic-based returns from cross-sectional regressions. A natural question is how the MAXSER-S would perform if we instead use long-short characteristic-sorted portfolios (which are often used in the literature) as the input assets. To investigate this, we use the value-weighted long-short portfolios from Jensen, Kelly, and Pedersen (2023). Following our main analysis, we include the market portfolio as a base asset and remove the CAPM beta portfolio (*beta_60m*) from the set, resulting in 152 characteristic portfolios plus the market, each with 588 monthly returns between 1975 and 2023. We apply the same training procedures as before, learning the SDF portfolio annually using a rolling window of the past 25 years, with an out-of-sample period from 2000 to 2023. We also use the same long-short characteristic-sorted portfolios as inputs for KNS.

Table 12 reports the summary statistics for the SDF constructed from long-short characteristic-sorted portfolios. First, comparing the results reported in Tables 1 and 12, we see the Sharpe ratio of our main MAXSER-based SDF, MAXSER-S(0.05), decreases from 0.78 in Table 1 to 0.38 in Table 12. This deterioration is not unique to our method; the KNS model's Sharpe ratio also drops from 0.59 to 0.31.

Table 12: **Out-of-sample performances of SDFs.** This table reports the out-of-sample performances of the optimal portfolios constructed from the proposed approach and other benchmark models, including the MAXSER-S(τ) with various post screening significance levels (τ =0.01, 0.05, 0.1), the shrinkage SDF portfolio (KNS) by Kozak, Nagel, and Santosh (2020), and other factor models. The MAXSER-based SDFs and the KNS SDF are constructed by long-short characteristic-sorted portfolios. The summary statistics include monthly mean return (Mean), standard deviation (SD), Sharpe ratio (SR), the maximum drawdown (MDD), the average number of factors in the portfolio (NUM) and the distribution percentiles. All are reported in percentage except the Sharpe ratio. The column "SR-pvalue" reports the p-value of testing the difference in the Sharpe ratios between the proposed MAXSER-S(0.05) and other models, while *, ** and *** indicate statistical significance at the level of 10%, 5% and 1%, respectively. The optimal portfolios are estimated every year, using a rolling window of past 25 years, and portfolio returns are calculated every month. The out-of-sample period is between 2000 and 2023.

Model	Mean	SD	SR	SR-pvalue	MDD	NUM	P1	P25	P50	P75	P99
Panel A: MAXSER-based SDF											
MAXSER-S(0.01)	1.84	5.15	0.36	0.27	23.57	22	-9.08	-1.11	1.86	4.81	14.37
MAXSER-S(0.05)	1.95	5.13	0.38	–	30.59	30	-9.95	-1.11	1.65	5.35	15.35
MAXSER-S(0.1)	1.93	5.08	0.38	0.34	25.90	32	-10.21	-1.04	1.67	5.22	15.35
Panel B: SDF from benchmark models											
KNS	1.46	4.67	0.31	0.093 *	20.88	25	-8.76	-1.28	1.06	4.14	14.38
CAPM	0.55	4.63	0.12	0.000 ***	49.39	1	-10.40	-2.03	1.14	3.26	10.33
FF3	0.81	5.85	0.14	0.001 ***	66.32	3	-19.45	-1.89	1.08	4.26	13.78
FF5	1.82	6.64	0.27	0.080 **	34.85	5	-11.82	-1.81	1.24	4.91	20.88
FF6	1.72	6.51	0.26	0.056 **	37.24	6	-12.29	-1.56	1.36	4.37	18.10
Q	1.50	5.72	0.26	0.053 **	32.46	4	-10.22	-1.47	1.58	4.02	20.49
Q5	1.91	5.39	0.36	0.370	25.23	5	-11.75	-0.92	1.53	4.84	16.61
SY4	2.01	6.34	0.32	0.087 *	27.03	4	-12.59	-1.31	2.02	4.68	19.95
DHS3	1.53	5.13	0.30	0.140	37.00	3	-11.23	-1.49	1.39	4.42	13.79
BS6	1.57	6.61	0.24	0.014 **	35.06	6	-14.84	-1.66	1.57	3.89	23.84

Second, when compared against the benchmark models using the same portfolios as inputs (Panel B), MAXSER-S remains the top performer. It is statistically superior to several models (e.g., KNS, CAPM, FF3, FF5, FF6, and Q), though its performance is statistically indistinguishable from other models such as Q5, SY4, DHS3, and BS6. The cumulative returns for these various SDFs are plotted in Figure 6. We also perform spanning tests, encompassing tests, asset pricing tests, and consider the impacts of transaction costs over this MAXSER-S. We find messages similar to those reported in Tables 2–4 and 7–9 (see details in Appendix B).

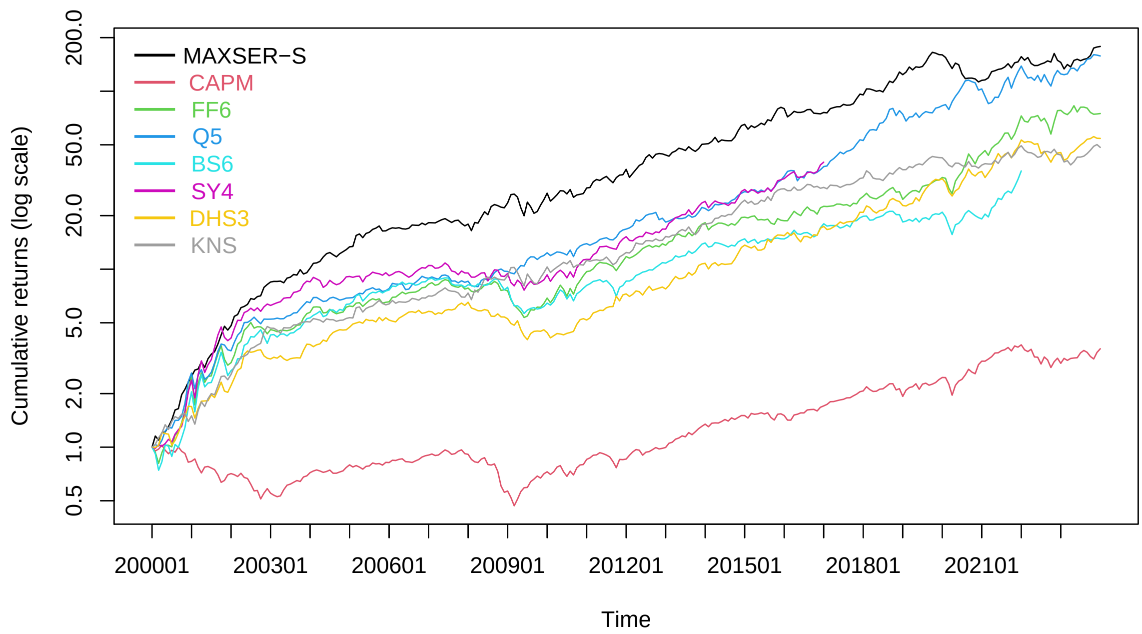


Figure 6: **Cumulative dollar returns of the SDF portfolios.** This figure plots the cumulative returns of various SDF portfolios from 2000 to 2023, starting with $1. We include the SDF portfolios constructed from the MAXSER-S, CAPM, FF6, Q5, BS6, SY4, DHS3, and KNS. The MAXSER-S SDF and the KNS SDF are constructed by long-short characteristic-sorted portfolios. The portfolio risk is set to be the market risk.

Overall, these findings lead to two clear conclusions. First, the MAXSER-S methodology is robust, as it still builds a superior SDF even when starting with less efficient input assets. Second, this highlights the significant impact of input asset choice on SDF performance. The dramatic drop in the Sharpe ratio occurs because long-short characteristic-sorted portfolios are not mean-variance efficient, as emphasized by Daniel, Mota, Rottke, and Santos (2020). This experiment confirms that using characteristic-based returns from cross-sectional regressions is the superior approach for constructing a high-performing SDF.

5 Conclusions

This paper develops a rigorous statistical framework to construct and interpret the stochastic discount factor (SDF) from the high-dimensional "zoo" of firm characteristics. We follow the right-hand side approach to asset pricing, framing the problem as one of maximizing the Sharpe ratio of the SDF portfolio. Theoretically, we extend the MAXSER approach in Ao, Li, and Zheng (2019) to handle the $N \gg T$ setting that is common in finance, establish its sure screening property to ensure no important factors are missed, and develop a novel statistical inference theory to test the estimated SDF loadings.

Our constructed SDF delivers a high and stable out-of-sample monthly Sharpe ratio of 0.78, significantly outperforming a comprehensive suite of established benchmark models. The superior performance is not an artifact of data mining or repackaged risks; spanning and encompassing tests confirm that our SDF contains new pricing information while subsuming the premia of existing factors. Our SDF portfolio doesn't have extreme allocation on certain stocks or concentrate on small stocks. These results hold after accounting for realistic transaction costs. Furthermore, we show that the choice of input assets is crucial: an SDF built from cross-sectional regression-based returns dramatically outperforms the one built from traditional long-short characteristic-sorted portfolios, highlighting the inefficiency of the latter.

The inference framework we develop provides a powerful diagnostic tool that allows researchers to go beyond point estimates and formally test the statistical significance of economic themes driving the SDF. By constructing confidence intervals for the contributions of characteristic themes, we can rigorously differentiate between economically meaningful drivers of factor premia and those whose contributions are statistically insignificant.

Our paper provides a more disciplined and interpretable approach to understanding the true composition of the SDF. Future research could apply this framework to other asset classes, such as corporate bonds or international equities, or use the time-varying SDF loadings to better understand the relationship between macroeconomic conditions and market-wide risk premia.

References

Ao, M., Y. Li, and X. Zheng. 2019. Approaching mean-variance efficiency for large portfolios. *Review of Financial Studies* 32:2890–919.

Asness, C. S., A. Frazzini, and L. H. Pedersen. 2019. Quality minus junk. *Review of Accounting studies* 24:34–112.

Bai, Z., and J. W. Silverstein. 1998. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Annals of Probability* 26:316–45.

Baker, M., and J. Wurgler. 2006. Investor sentiment and the cross-section of stock returns. *Journal of Finance* 61:1645–80.

Bali, T. G., L. Del Viva, N. Lambertides, and L. Trigeorgis. 2020. Growth options and related stock market anomalies: Profitability, distress, lotteryness, and volatility. *Journal of Financial and Quantitative Analysis* 55:2150–80.

Barillas, F., R. Kan, C. Robotti, and J. Shanken. 2020. Model comparison with Sharpe ratios. *Journal of Financial and Quantitative Analysis* 55:1840–74.

Barillas, F., and J. Shanken. 2017. Which alpha? *Review of Financial Studies* 30:1316–38.

———. 2018. Comparing asset pricing models. *Journal of Finance* 73:715–54.

Belloni, A., V. Chernozhukov, and C. Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81:608–50.

Bouchaud, J.-P., P. Krueger, A. Landier, and D. Thesmar. 2019. Sticky expectations and the profitability anomaly. *Journal of Finance* 74:639–74.

Britten-Jones, M. 1999. The sampling error in estimates of mean-variance efficient portfolio weights. *Journal of Finance* 54:655–71.

Bryzgalova, S., V. DeMiguel, S. Li, and M. Pelger. 2023. Asset-pricing factors with economic targets. *Available at SSRN 4344837* .

Bryzgalova, S., M. Pelger, and J. Zhu. 2025. Forest through the trees: Building cross-sections of asset returns. *Journal of Finance* 80:2447–506.

Cakici, N., C. Fieberg, D. Metko, and A. Zaremba. 2024. Do anomalies really predict market returns? new data and new evidence. *Review of Finance* 28:1–44.

Cassella, S., and H. Gulen. 2018. Extrapolation bias and the predictability of stock returns by price-scaled variables. *Review of Financial Studies* 31:4345–97. ISSN 0893-9454.

Chen, A. Y. 2021. The limits of p-hacking: Some thought experiments. *Journal of Finance* 76:2447–80.

Chen, A. Y., and M. Velikov. 2023. Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis* 58:968–1004.

Cochrane, J. H. 2011. Presidential address: Discount rates. *Journal of Finance* 66:1047–108.

Cong, L. W., G. Feng, J. He, and X. He. 2025. Growing the efficient frontier on panel trees. *Journal of Financial Economics* 167:104024–.

Cong, L. W., K. Tang, J. Wang, and Y. Zhang. 2021. AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI. Working Paper, Cornell University.

Daniel, K., D. Hirshleifer, and L. Sun. 2020. Short-and long-horizon behavioral factors. *Review of Financial Studies* 33:1673–736.

Daniel, K., L. Mota, S. Rottke, and T. Santos. 2020. The cross-section of risk and returns. *Review of Financial Studies* 33:1927–79.

Daniel, K., and S. Titman. 1997. Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance* 52:1–33.

Davydiuk, T., T. Marchuk, and I. Shaliastovich. 2024. Corporate bond issuance by financial institutions. *Available at SSRN 4811333* .

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal. 2009. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science* 55:798–812.

DeMiguel, V., A. Martin-Utrera, F. J. Nogales, and R. Uppal. 2020. A transaction-cost perspective on the multitude of firm characteristics. *Review of Financial Studies* 33:2180–222.

DeMiguel, V., A. Martín-utrera, and R. Uppal. 2024. A multifactor perspective on volatility-managed portfolios. *Journal of Finance* 79:3859–91.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. *Annals of Statististics* 32:407–99.

El Karoui, N. 2003. On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n $\to \infty$. *arXiv preprint math/0309355* .

Fama, E. F., and K. R. French. 1992. The cross-section of expected stock returns. *Journal of Finance* 47:427–65. ISSN 1540-6261.

———. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116:1–22.

———. 2018. Choosing factors. *Journal of Financial Economics* 128:234–52.

———. 2020. Comparing cross-section and time-series factor models. *Review of Financial Studies* 33:1891–926. ISSN 0893-9454.

Fama, E. F., and J. D. MacBeth. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy* 81:607–36.

Fan, J., Y. Fan, and J. Lv. 2008. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* 147:186–97.

Fan, J., Y. Ke, and K. Wang. 2020. Factor-adjusted regularized model selection. *Journal of econometrics* 216:71–85.

Fan, J., Y. Liao, and M. Mincheva. 2011. High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics* 39:3320–.

———. 2013. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75:603–80.

Fan, J., and J. Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70:849–911.

Feldheim, O. N., and S. Sodin. 2010. A universality result for the smallest eigenvalues of certain sample covariance matrices. *Geometric And Functional Analysis* 20:88–123.

Feng, G., S. Giglio, and D. Xiu. 2020. Taming the factor zoo: A test of new factors. *Journal of Finance* 75:1327–70.

Feng, G., J. He, N. G. Polson, and J. Xu. 2024. Deep learning in characteristics-sorted factor models. *Journal of Financial and Quantitative Analysis* 59:3001–36.

Frazzini, A., and L. H. Pedersen. 2014. Betting against beta. *Journal of financial economics* 111:1–25.

Ghosh, A., C. Julliard, and A. P. Taylor. 2017. What is the consumption-CAPM missing? An information-theoretic framework for the analysis of asset pricing models. *Review of Financial Studies* 30:442–504.

———. 2025. An information-theoretic asset pricing model. *Journal of Financial Econometrics* 23:nbae033–.

Giglio, S., D. Xiu, and D. Zhang. 2025. Test assets and weak factors. *Journal of Finance* 80:259–319.

Gospodinov, N., R. Kan, and C. Robotti. 2017. Spurious inference in reduced-rank asset-pricing models. *Econometrica* 85:1613–28.

Goyal, A., I. Welch, and A. Zafirov. 2024. A comprehensive 2022 look at the empirical performance of equity premium prediction. *Review of Financial Studies* 37:3490–557.

Green, J., J. R. Hand, and X. F. Zhang. 2017. The characteristics that provide independent information about average us monthly stock returns. *Review of Financial Studies* 30:4389–436.

Hansen, L. P., and R. Jagannathan. 1997. Assessing specification errors in stochastic discount factor models. *Journal of Finance* 52:557–90.

Harvey, C. R., and Y. Liu. 2021. Lucky factors. *Journal of Financial Economics* 141:413–35.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. ... and the cross-section of expected returns. *Review of Financial Studies* 29:5–68.

Hastie, T., R. Tibshirani, J. Friedman, et al. 2009. The elements of statistical learning.

Hou, K., H. Mo, C. Xue, and L. Zhang. 2021. An augmented q-factor model with expected growth. *Review of Finance* 25:1–41.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting anomalies: An investment approach. *Review of Financial Studies* 28:650–705.

———. 2020. Replicating anomalies. *Review of Financial Studies* 33:2019–133.

Jacobs, H., and S. Müller. 2020. Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics* 135:213–30.

Jegadeesh, N. 1990. Evidence of predictable behavior of security returns. *Journal of finance* 45:881–98.

Jegadeesh, N., J. Noh, K. Pukthuanthong, R. Roll, and J. Wang. 2019. Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation. *Journal of Financial Economics* 133:273–98.

Jensen, T. I., B. Kelly, and L. H. Pedersen. 2023. Is there a replication crisis in finance? *Journal of Finance* 78:2465–518.

Jensen, T. I., B. T. Kelly, S. Malamud, and L. H. Pedersen. 2025. Machine learning and the implementable efficient frontier. *Review of Financial Studies* forthcoming.

Kan, R., X. Wang, and G. Zhou. 2022. Optimal portfolio choice with estimation risk: No risk-free asset case. *Management Science* 68:2047–68.

Kan, R., and C. Zhang. 1999. Two-pass tests of asset pricing models with useless factors. *Journal of Finance* 54:203–35.

Kelly, B., S. Malamud, and L. H. Pedersen. 2023. Principal portfolios. *Journal of Finance* 78:347–87.

Kelly, B. T., S. Malamud, M. Pourmohammadi, and F. Trojani. 2024. Universal portfolio shrinkage. NBER Working Paper 32004.

Kelly, B. T., S. Pruitt, and Y. Su. 2019. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics* 134:501–24.

Kozak, S., S. Nagel, and S. Santosh. 2020. Shrinking the cross-section. *Journal of Financial Economics* 135:271–92.

Ledoit, O., and M. Wolf. 2008. Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance* 15:850–9.

———. 2017. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Review of Financial Studies* 30:4349–88.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor. 2016. Exact post-selection inference, with application to the Lasso. *Annals of Statistics* 44:907–27.

Leng, C., Y. Lin, and G. Wahba. 2006. A note on the Lasso and related procedures in model selection. *Statistica Sinica* 16:1273–84.

Lettau, M., and M. Pelger. 2020a. Estimating latent asset-pricing factors. *Journal of Econometrics* 218:1–31.

———. 2020b. Factors that fit the time series and cross-section of stock returns. *Review of Financial Studies* 33:2274–325.

Linnainmaa, J. T., and M. R. Roberts. 2018. The history of the cross-section of stock returns. *Review of Financial Studies* 31:2606–49.

Luo, B., G. Zhou, and T. Zhou. 2025. Which factors matter in the pricing kernel? *Working Paper* .

McLean, R. D., and J. Pontiff. 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71:5–32.

Merlevède, F., M. Peligrad, and E. Rio. 2011. A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields* 151:435–74.

Michaud, R. O. 1989. The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal* 45:31–42.

Nagel, S., and K. J. Singleton. 2011. Estimation and evaluation of conditional asset pricing models. *Journal of Finance* 66:873–909.

Newey, W., and K. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55:703–8.

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* 137:515–49.

———. 2022. Risk price variation: The missing half of empirical asset pricing. *The Review of Financial Studies* 35:5127–84.

Schneider, P., C. Wagner, and J. Zechner. 2020. Low-risk anomalies? *Journal of Finance* 75:2673–718.

Stambaugh, R. F., and Y. Yuan. 2017. Mispricing factors. *Review of Financial Studies* 30:1270–315.

Wainwright, M. J. 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE transactions on information theory* 55:2183–202.

Wasserman, L., and K. Roeder. 2009. High dimensional variable selection. *Annals of statistics* 37:2178–201.

Wu, J., L. Zhang, and X. F. Zhang. 2010. The q-theory approach to understanding the accrual anomaly. *Journal of Accounting Research* 48:177–223.

Zhang, C.-H., and S. S. Zhang. 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* 76:217–42.

Zhao, P., and B. Yu. 2006. On model selection consistency of Lasso. *Journal of Machine Learning Research* 7:2541–63.

Internet Appendix to "cross-sectional Learning and Inference for the Stochastic Discount Factor"

Appendix A Proofs

A.1 Proof of Theorem 1

Using $(\mathbf{f}, \mathbf{R})$ is equivalent to using the (infeasible) variable $(\mathbf{f}, \mathbf{U})$ to construct the optimal portfolio. The optimal weights invested on $(\mathbf{f}, \mathbf{U})$, $\boldsymbol{\omega}^*_{f,u}$, is given by (2.6). Recall that the estimated allocation on $(\mathbf{f}, \widehat{\mathbf{U}})$ is $\left((\widehat{\boldsymbol{\omega}}^*_f)^\top, (\widehat{\boldsymbol{\omega}}^*_u)^\top\right)^\top$. The corresponding allocation on the infeasible $(\mathbf{f}, \mathbf{U})$ is

$$\mathring{\boldsymbol{\omega}}^*_{f,u} = \left((\widehat{\boldsymbol{\omega}}^*_f)^\top + (\widehat{\boldsymbol{\omega}}^*_u)^\top(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}), (\widehat{\boldsymbol{\omega}}^*_u)^\top\right)^\top =: \left((\mathring{\boldsymbol{\omega}}^*_f)^\top, (\widehat{\boldsymbol{\omega}}^*_u)^\top\right)^\top.$$

Under Assumption 4 and by the definition (2.9), we have $\widehat{\boldsymbol{\omega}}^*_u = O_p(1)$. Moreover, it is easy to see that $\|\boldsymbol{\omega}^*_f\| < C$ and $\|\widehat{\boldsymbol{\omega}}^*_f\| < C$ with probability tending one. By a similar proof to that of Proposition 8 of Ao, Li, and Zheng (2019) with a slight modification by replacing $\widehat{\mathbf{U}}$ with $(\mathbf{f}, \widehat{\mathbf{U}})$, we have

$$\boldsymbol{\mu}^\top_{f,u}\widehat{\boldsymbol{\omega}}^*_{f,u} - \frac{1}{r_{\sigma,f,u}}\boldsymbol{\mu}^\top_{f,u}\boldsymbol{\omega}^*_{f,u} = o_p(1),$$

where recall that $r_{\sigma,f,u}$ is defined in (2.8), and

$$(\widehat{\boldsymbol{\omega}}^*_{f,u})^\top\boldsymbol{\Sigma}_{f,u}\widehat{\boldsymbol{\omega}}^*_{f,u} - \frac{1}{r^2_{\sigma,f,u}}(\boldsymbol{\omega}^*_{f,u})^\top\boldsymbol{\Sigma}_{f,u}\boldsymbol{\omega}^*_{f,u} = o_p(1).$$

Moreover, similarly to the proof of Theorem 4 of Ao, Li, and Zheng (2019) with a slight modification by replacing $\widehat{\mathbf{U}}$ with $(\mathbf{f}, \widehat{\mathbf{U}})$, we have $\|\widehat{\boldsymbol{\omega}}^*_{f,u} - \boldsymbol{\omega}^*_{f,u}/r_{\sigma,f,u}\| = o_p(1)$. Note further that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\max} = o_p(1)$; see page 25 of Appendix I of Ao, Li, and Zheng (2019). It follows that

$$\|(\widehat{\boldsymbol{\omega}}^*_u)^\top(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\| \leq \|(\widehat{\boldsymbol{\omega}}^*_u)\|_1\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_{\max} = o_p(1).$$

We then get,

$$\|\mathring{\boldsymbol{\omega}}^*_f - \frac{1}{r_{\sigma,f,u}}\boldsymbol{\omega}^*_f\| = o_p(1),$$

and

$$\boldsymbol{\mu}^\top_f\mathring{\boldsymbol{\omega}}^*_f - \frac{1}{r_{\sigma,f,u}}\boldsymbol{\mu}^\top_f\boldsymbol{\omega}^*_f = o_p(1),$$

$$(\mathring{\boldsymbol{\omega}}_f^*)^\top \boldsymbol{\Sigma}_f \mathring{\boldsymbol{\omega}}_f^* - \frac{1}{r_{\sigma,f,u}^2} (\boldsymbol{\omega}_f^*)^\top \boldsymbol{\Sigma}_f \boldsymbol{\omega}_f^* = o_p(1).$$

In addition, because $r_{\sigma,f,u} \asymp 1$, by the decompositions

$$\begin{aligned}
&\boldsymbol{\mu}_{f,R}^\top \widehat{\boldsymbol{\omega}}_{f,R}^* = \boldsymbol{\mu}_f^\top \mathring{\boldsymbol{\omega}}_f^* + \boldsymbol{\alpha}^\top \widehat{\boldsymbol{\omega}}_u^*, \\
&\boldsymbol{\mu}_{f,R}^\top \boldsymbol{\omega}_{f,R}^* = \boldsymbol{\mu}_f^\top \boldsymbol{\omega}_f^* + \boldsymbol{\alpha}^\top \boldsymbol{\omega}_u^*, \\
&(\widehat{\boldsymbol{\omega}}_{f,R}^*)^\top \boldsymbol{\Sigma}_{f,R} \widehat{\boldsymbol{\omega}}_{f,R}^* = (\mathring{\boldsymbol{\omega}}_f^*)^\top \boldsymbol{\Sigma}_f \mathring{\boldsymbol{\omega}}_f^* + (\widehat{\boldsymbol{\omega}}_u^*)^\top \boldsymbol{\Sigma}_u \widehat{\boldsymbol{\omega}}_u^*, \\
&(\boldsymbol{\omega}_{f,R}^*)^\top \boldsymbol{\Sigma}_{f,R} (\boldsymbol{\omega}_{f,R}^*) = (\boldsymbol{\omega}_f^*)^\top \boldsymbol{\Sigma}_f \boldsymbol{\omega}_f^* + (\boldsymbol{\omega}_u^*)^\top \boldsymbol{\Sigma}_u \boldsymbol{\omega}_u^*,
\end{aligned}$$

we get the desired convergence. □

A.2 Proof of Theorem 2

Without lost of generality, suppose that $S_1 = \{1, ..., |S_{signal}|, ..., |S_1|\}$ with $q_u = |S_1|$, and $S_{signal} = \{1, ..., |S_{signal}|\}$. Denote $\mathbf{r} = ((\mathbf{f}_1^\top, \mathbf{U}_1^\top)^\top, ..., (\mathbf{f}_T^\top, \mathbf{U}_T^\top)^\top) = (\mathbf{r}_1, ..., \mathbf{r}_T) =: (r_{it})$, and $\widehat{\mathbf{r}} = ((\mathbf{f}_1^\top, \widehat{\mathbf{U}}_1^\top)^\top, ..., (\mathbf{f}_T^\top, \widehat{\mathbf{U}}_T^\top)^\top) =: (\widehat{r}_{it})$. Denote by $\mathbf{r}_{S_1}$ the first q_u rows, and $\mathbf{r}_{S_1^c}$ the remaining rows of $\mathbf{r}$. Denote by $\widehat{\mathbf{r}}_{S_1}$ the first $|S_1|$ rows, and $\widehat{\mathbf{r}}_{S_1^c}$ the last remaining rows of $\widehat{\mathbf{r}}$. Recall that $r_{\sigma,f,u} = \sigma(1/\sqrt{\theta_{all}} + \sqrt{\theta_{all}})$. Because $\boldsymbol{\omega}_{f,u}^*$ are proportional to $\widetilde{\boldsymbol{\omega}}_c^* := \boldsymbol{\omega}_{f,u}^*/r_{\sigma,f,u}$, to show (2.13), it is equivalent to show

$$P\Big(\text{sign}(\widehat{\boldsymbol{\omega}}_{f,u}^*(i)) = \text{sign}(\widetilde{\boldsymbol{\omega}}_c^*(i)) \quad \text{for all } i \in S_{signal} \cup S_1^c\Big) \to 1. \tag{A.1}$$

Recall that η is defined in Assumption 6, and

$$\begin{aligned}
&C_{11} = \frac{1}{T}\widehat{\mathbf{r}}_{S_1}\widehat{\mathbf{r}}_{S_1}^\top, \quad C_{21} = \frac{1}{T}\widehat{\mathbf{r}}_{S_1^c}\widehat{\mathbf{r}}_{S_1}^\top, \\
&W_1 = \sqrt{\frac{1}{T}}\widehat{\mathbf{r}}_{S_1}\Big(\mathbf{1} - \widehat{\mathbf{r}}^\top \widetilde{\boldsymbol{\omega}}_c^*\Big), \quad \text{and} \quad W_2 = \sqrt{\frac{1}{T}}\widehat{\mathbf{r}}_{S_1^c}\Big(\mathbf{1} - \widehat{\mathbf{r}}^\top \widetilde{\boldsymbol{\omega}}_c^*\Big).
\end{aligned}$$

Using the proof of Proposition 1 of Zhao and Yu (2006), in particular, the KKT (Karush-Kuhn-Tucker) conditions (Lemma 1 of Zhao and Yu (2006)), if

$$\begin{aligned}
&\sqrt{T}\left(C_{11}\Big(\widehat{\boldsymbol{\omega}}_{f,u}^* - \widetilde{\boldsymbol{\omega}}_c^*\Big)_{S_1}\right)(i) - W_1(i) = -\frac{\lambda\sqrt{T}}{2}\text{sign}(\widetilde{\boldsymbol{\omega}}_c^*(i)) \text{ for all } i \in S_{signal}, \\
&\left|\Big(\widehat{\boldsymbol{\omega}}_{f,u}^* - \widetilde{\boldsymbol{\omega}}_c^*\Big)(i)\right| < |\widetilde{\boldsymbol{\omega}}_c^*(i)| \text{ for all } i \in S_{signal}, \text{ and} \\
&\left\|\sqrt{T}C_{21}\Big(\widehat{\boldsymbol{\omega}}_{f,u}^* - \widetilde{\boldsymbol{\omega}}_c^*\Big)_{S_1} - W_2\right\|_{\max} < \frac{\lambda\sqrt{T}}{2},
\end{aligned}$$

then $\text{sign}(\widehat{\boldsymbol{\omega}}^*_{f,u}(i)) = \text{sign}(\widetilde{\boldsymbol{\omega}}^*_c(i))$ for all $i \in S_{signal} \cup S_1^c$. Hence, one can show that

$$P\Big(\text{sign}(\widehat{\boldsymbol{\omega}}^*_{f,u}(i)) = \text{sign}(\widetilde{\boldsymbol{\omega}}^*_c(i)) \quad \text{for all } i \in S_{signal} \cup S_1^c\Big) \geq P\Big(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}\Big), \tag{A.2}$$

where

$$\begin{aligned}
\mathcal{A} =& \left\{ \big|(C_{11}^{-1}W_1)(i)\big| < \sqrt{T}\left(|\widetilde{\boldsymbol{\omega}}^*_c(i)| - \frac{\lambda}{2}\big|(C_{11}^{-1}\text{sign}(\widetilde{\boldsymbol{\omega}}^*_c))(i)\big|\right) \text{ for all } i = 1, ..., |S_{signal}| \right\},\\
\mathcal{B} =& \left\{ \|(C_{21}C_{11}^{-1}W_1 - W_2\|_{\max} < \frac{\lambda\sqrt{T}}{4}\eta \right\}, \quad \text{and}\\
\mathcal{C} =& \left\{ \max_{\tilde{\mathbf{1}} \in V} \|C_{21}C_{11}^{-1}\tilde{\mathbf{1}}\|_{\max} < 1 - \frac{\eta}{2} \right\},
\end{aligned}$$

where recall that $V = \{(\delta_1, ..., \delta_{|S_1|})^\top : \delta_i = \text{sign}(\boldsymbol{\omega}^*_{S_1}(i))$ when $(\boldsymbol{\omega}^*_{S_1}(i)) \neq 0,$ and $\delta_i \in \{-1, 1\}$ otherwise$\}$. Define for some $C > 0$, the following event

$$\begin{aligned}
\mathcal{G} =& \left\{ \|(W_1^\top, W_2^\top)\|_{\max} < C\sqrt{\log N} \right\}\\
& \bigcap \left\{ \|C_{11}^{-1} - (\boldsymbol{\Sigma}_{f,u,11} + \boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top)^{-1}\| < C\sqrt{\frac{q_u \log N}{T}} \right\}\\
& \quad \bigcap \left\{ \|C_{21} - \boldsymbol{\Sigma}_{f,u,21} - \boldsymbol{\mu}_{f,u,2}\boldsymbol{\mu}_{f,u,1}^\top\|_{\max} < Cq_u\sqrt{\frac{\log N}{T}} \right\}.
\end{aligned}$$

We will show that as $N, T \to \infty$,

$$P(\mathcal{G}) \to 1. \tag{A.3}$$

We claim that (A.3) implies the desired convergence (A.2). In fact, about event $\mathcal{A}$, under Assumptions 1 and 5, we have $q_u^2 \log N/T \to 0$. Assumption 6 implies that

$$\|(\boldsymbol{\Sigma}_{f,u,11} + \boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top)^{-1}\| = O(1).$$

Hence, under event $\mathcal{G}$,

$$\|C_{11}^{-1}\| < C. \tag{A.4}$$

Therefore,

$$\max_{1 \leq i \leq |S_{signal}|} |C_{11}^{-1}W_1|_i \leq \|C_{11}^{-1}\|\sqrt{q_u}\|W_1\|_{\max} = \quad O(\sqrt{q_u \log N}). \tag{A.5}$$

Similarly

$$\max_{1\le i\le |S_{signal}|}\left|\frac{\lambda}{2}\Big(C_{11}^{-1}\mathrm{sign}(\widetilde{\boldsymbol{\omega}}_c^*)\Big)_i\right| \le \frac{\lambda}{2}\|C_{11}^{-1}\|\sqrt{q_u} \le C\lambda\sqrt{q_u}. \tag{A.6}$$

By Assumption 5 that $\min_{1\le i\le |S_{signal}|}|\omega_i| \gg \lambda\sqrt{q_u}$ and $\lambda \gg \sqrt{(\log N)/T}$, we get that

$$P(\mathcal{A}) \to 1. \tag{A.7}$$

About event $\mathcal{B}$, by Assumption 3, under event $\mathcal{G}$, we have

$$\|C_{21}\|_{\max} < C. \tag{A.8}$$

Therefore,

$$\begin{aligned}\|C_{21}C_{11}^{-1}W_1\|_{\max} \le& \|C_{21}\|_{\max}\|(C_{11}^{-1}W_1)\|_1\\ \le&\sqrt{q_u}\|C_{21}\|_{\max}\cdot\|C_{11}^{-1}W_1\|\\ \le& q_u\|C_{21}\|_{\max}\cdot\|C_{11}^{-1}\|\cdot\|W_1\|_{\max}\\ =&O(q_u\sqrt{\log N}),\end{aligned}$$

where the second line holds by the Cauchy-Schwarz inequality, and the last equation holds by (A.5) and that under event $\mathcal{G}$, $\|W_1\|_{\max} = O(\sqrt{\log N})$. By Assumption 5 that $\lambda \gg q_u\sqrt{(\log N)/T}$, we get

$$\lambda\sqrt{T} \gg \max(\|C_{21}C_{11}^{-1}W_1\|_{\max}, \|W_2\|_{\max}).$$

Therefore, (A.3) implies that

$$P(\mathcal{B}) \to 1. \tag{A.9}$$

About event $\mathcal{C}$, we have

$$\begin{aligned}&\max_{\tilde{\mathbf{1}}\in V}\|C_{21}C_{11}^{-1}\tilde{\mathbf{1}}\|_{\max}\\ \le&\max_{\tilde{\mathbf{1}}\in V}\|(\boldsymbol{\Sigma}_{f,u,21}+\boldsymbol{\mu}_{f,u,2}\boldsymbol{\mu}_{f,u,1}^\top)(\boldsymbol{\Sigma}_{f,u,11}+\boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top)^{-1}\tilde{\mathbf{1}}\|_{\max}\\ &+\max_{\tilde{\mathbf{1}}\in V}\|(C_{21}-\boldsymbol{\Sigma}_{u,21}-\boldsymbol{\alpha}_2\boldsymbol{\mu}_{f,u,1}^\top)C_{11}^{-1}\tilde{\mathbf{1}}\|_{\max}\\ &+\max_{\tilde{\mathbf{1}}\in V}\|(\boldsymbol{\Sigma}_{u,21}+\boldsymbol{\mu}_{f,u,2}\boldsymbol{\mu}_{f,u,1}^\top)(C_{11}^{-1}-(\boldsymbol{\Sigma}_{f,u,11}+\boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top)^{-1})\tilde{\mathbf{1}}\|_{\max}\\ =:&I+II+III.\end{aligned}$$

By Assumption 6, $I \leq 1-\eta$. In addition, under event G, by (A.4) and (A.8), we have

$$
\begin{aligned}
II \leq & \max_{\tilde{\mathbf{1}}\in V} \|C_{21} - \boldsymbol{\Sigma}_{f,u,21} - \boldsymbol{\mu}_{f,u,2}\boldsymbol{\mu}_{f,u,1}^\top\|_{\max}\|C_{11}^{-1}\tilde{\mathbf{1}}\|_1 \\
\leq & \max_{\tilde{\mathbf{1}}\in V} \sqrt{q_u}\|C_{21} - \boldsymbol{\Sigma}_{f,u,21} - \boldsymbol{\mu}_{f,u,2}\boldsymbol{\mu}_{f,u,1}^\top\|_{\max} \cdot \|C_{11}^{-1}\tilde{\mathbf{1}}\| \\
\leq & q_u\|C_{21} - \boldsymbol{\Sigma}_{u,21} - \boldsymbol{\mu}_{f,u,2}\boldsymbol{\mu}_{f,u,1}^\top\|_{\max} \cdot \|C_{11}^{-1}\| \\
= & O\left(q_u\sqrt{\frac{\log N}{T}}\right).
\end{aligned}
$$

Similarly, under event $\mathcal{G}$, by Assumption 3, we have

$$
\begin{aligned}
III \leq & q_u\|\boldsymbol{\Sigma}_{u,21} + \boldsymbol{\mu}_{f,u,2}\boldsymbol{\mu}_{f,u,1}^\top\|_{\max} \cdot \|C_{11}^{-1} - (\boldsymbol{\Sigma}_{f,u,11} + \boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top)^{-1}\| \\
= & O\left(q_u^{3/2}\sqrt{\frac{\log N}{T}}\right).
\end{aligned}
$$

Under Assumptions 1 and 5, $q_u^{3/2}\sqrt{(\log N)/T} = o(1)$, hence (A.3) implies that

$$
P(\mathcal{C}) \to 1. \tag{A.10}
$$

Combining (A.2), (A.3), (A.7), (A.9) and (A.10) yields (A.1).

It remains to show (A.3). Define $\bar{\hat{\mathbf{r}}} = \hat{\mathbf{r}}\mathbf{1}/T$, $\bar{\mathbf{r}} = \mathbf{r}\mathbf{1}/T$, and $W = \hat{\mathbf{r}}\Big(\mathbf{1} - \hat{\mathbf{r}}^\top\widetilde{\boldsymbol{\omega}}_c^*\Big)/\sqrt{T}$. We have

$$
\begin{aligned}
W = & \sqrt{T}(\bar{\hat{\mathbf{r}}} - \bar{\mathbf{r}}) \\
& - \sqrt{\frac{1}{T}}(\hat{\mathbf{r}} - \mathbf{r})\mathbf{r}^\top\widetilde{\boldsymbol{\omega}}_c^* \\
& - \sqrt{\frac{1}{T}}\mathbf{r}(\hat{\mathbf{r}} - \mathbf{r})^\top\widetilde{\boldsymbol{\omega}}_c^* \\
& - \sqrt{\frac{1}{T}}(\hat{\mathbf{r}} - \mathbf{r})(\hat{\mathbf{r}} - \mathbf{r})^\top\widetilde{\boldsymbol{\omega}}_c \\
& + \sqrt{\frac{1}{T}}\mathbf{r}\left(\mathbf{1} - \mathbf{r}^\top\widetilde{\boldsymbol{\omega}}_c^*\right) \\
=: & I + II + III + IV + V.
\end{aligned}
$$

About term I, under Assumptions 1–3, by Bernstein's inequality (Theorem 1 of Merlevède, Peligrad, and Rio (2011)),

$$
P\left(\|\bar{\mathbf{r}} - \boldsymbol{\mu}_{f,u}\|_{\max} > \sqrt{\frac{\log N}{T}}\right) = O\left(\frac{1}{T^2}\right).
$$

Moreover, by the Cauchy-Schwarz inequality,

$$\|\widehat{\overline{\mathbf{r}}} - \overline{\mathbf{r}}\|_{\max} \leq \max_{1 \leq i \leq N} \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{U}_{it} - U_{it}\right)^2}. \tag{A.11}$$

We have

$$\sum_{t=1}^{T}(\widehat{U}_{it} - U_{it})^2 = \sum_{t=1}^{T}\left((\boldsymbol{\beta}_{i\cdot} - \widehat{\boldsymbol{\beta}}_{i\cdot})\mathbf{f}_t\right)^2 \leq K\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_{\max}^2 \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbf{f}_{t,k}^2,$$

where $\mathbf{f}_{t,k} = (\mathbf{f}_t)_k$, $\widehat{\boldsymbol{\beta}}_{i\cdot}$ and $\boldsymbol{\beta}_{i\cdot}$ are the ith row of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, respectively. Note that

$$\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} = \sum_{t=1}^{T}\mathbf{U}_t\check{\mathbf{f}}_t^\top(\check{\mathbf{F}}^\top\check{\mathbf{F}})^{-1},$$

where $\check{\mathbf{F}} = (\check{\mathbf{f}}_{t,k})_{1\leq t\leq T, 1\leq k\leq K}$, $\check{\mathbf{f}}_t = (\mathbf{f}_{t,k} - \bar{f}_k)_{1\leq k\leq K}$, and $\bar{f}_k = \sum_{t=1}^{T}\mathbf{f}_{t,k}/T$. By Assumptions 1 and 2, the independence between $\mathbf{U}$ and $\check{\mathbf{F}}$ and Bernstein's inequality, one can show that for some constant $C > 0$,

$$\begin{aligned} &P\left(\sum_{t=1}^{T}\|\mathbf{f}_t\|^2 < CT\right) > 1 - \frac{1}{T^2}, \\ &P\left(\left\|(\check{\mathbf{F}}^\top\check{\mathbf{F}})^{-1}\right\| < \frac{C}{T}\right) > 1 - \frac{1}{T^2}, \end{aligned} \tag{A.12}$$

and

$$P\left(\left\|\sum_{t=1}^{T}\mathbf{U}_t\check{\mathbf{f}}_t^\top\right\|_{\max} < C\sqrt{T\log N}\right) > 1 - \frac{1}{T}.$$

Hence

$$P\left(\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_{\max} > \sqrt{\frac{\log N}{T}}\right) < \frac{1}{T}.$$

By Bernstein's inequality again,

$$P\left(\max_{1\leq k\leq K}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_{t,k}^2 - E(\mathbf{f}_{t,k}^2)\right| < C\right) > 1 - \frac{1}{T},$$

Hence

$$P\left(\frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbf{f}_{t,k}^2 < C\right) > 1 - \frac{1}{T}.$$

Therefore,

$$P\left(\max_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}(\widehat{U}_{it} - U_{it})^2 > \frac{\log N}{T}\right) = O\left(\frac{1}{T}\right). \tag{A.13}$$

By (A.11) and (A.13), we get

$$P\left(\|\widehat{\bar{\mathbf{r}}}-\bar{\mathbf{r}}\|_{\max}>\sqrt{\frac{\log N}{T}}\right)=O\left(\frac{1}{T}\right).$$

Combining the results above yields

$$P\left(\|I\|_{\max}>\sqrt{\log N}\right)=O\left(\frac{1}{T}\right).$$

About term II, by the Cauchy-Schwarz inequality,

$$\|(\widehat{\mathbf{r}}-\mathbf{r})\mathbf{r}^{\top}\widetilde{\boldsymbol{\omega}}_c^*\|_{\max}\leq\left(\max_{1\leq i\leq N}\sqrt{\sum_{t=1}^{T}\left(\widehat{U}_{it}-U_{it}\right)^2}\right)\sqrt{\sum_{t=1}^{T}\left((\widetilde{\boldsymbol{\omega}}_c^*)^{\top}\mathbf{r}_t\right)^2}.$$

Note that $(\widetilde{\boldsymbol{\omega}}_c^*)^{\top}\mathbf{r}_t\overset{\text{i.i.d.}}{\sim}N(\sqrt{\theta_{all}}/r_{\sigma,f,u},1/r_{\sigma,f,u}^2)$, we get that for some $C>0$,

$$P\left(\left(\sum_{t=1}^{T}\left((\widetilde{\boldsymbol{\omega}}_c^*)^{\top}\mathbf{r}_t\right)^2\right)<CT\right)>1-\frac{1}{T^2}. \tag{A.14}$$

Combining (A.13) and (A.14) yields

$$P\left(\|II\|_{\max}>\sqrt{\log N}\right)=O\left(\frac{1}{T^2}\right).$$

About term III, by the Cauchy-Schwarz inequality, we get that

$$\|III\|_{\max}\leq\sqrt{\max_{1\leq i\leq N+K}\frac{1}{T}\sum_{t=1}^{T}r_{it}^2}\sqrt{\sum_{t=1}^{T}\left((\widetilde{\boldsymbol{\omega}}_c^*)^{\top}(\widehat{\mathbf{r}}_t-\mathbf{r}_t)\right)^2}. \tag{A.15}$$

Under Assumptions 2 and 3, there exists $C>0$,

$$P\left(\max_{1\leq i\leq N+K}\frac{1}{T}\sum_{t=1}^{T}(r_{it})^2<C\right)\geq1-\frac{1}{T^2}. \tag{A.16}$$

Note that $\widehat{\mathbf{U}}_t-\mathbf{U}_t=(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\mathbf{f}_t=\big(\sum_{t=1}^{T}\mathbf{U}_t\check{\mathbf{f}}_t^{\top}(\check{\mathbf{F}}^{\top}\check{\mathbf{F}})^{-1}\big)\mathbf{f}_t$. We have that

$$(\widetilde{\boldsymbol{\omega}}_c^*)^{T}(\widehat{\mathbf{r}}_t-\mathbf{r}_t)=\left(\sum_{t=1}^{T}(\boldsymbol{\omega}_u^*/r_{\sigma,f,u})^{\top}\mathbf{U}_t\check{\mathbf{f}}_t^{\top}(\check{\mathbf{F}}^{\top}\check{\mathbf{F}})^{-1}\right)\mathbf{f}_t.$$

Therefore,

$$\sum_{t=1}^{T}\Big((\widetilde{\boldsymbol{\omega}}_c^*)^\top(\widehat{\mathbf{r}}_t-\mathbf{r}_t)\Big)^2 \leq \Bigg\|\sum_{t=1}^{T}(\widetilde{\boldsymbol{\omega}}_U^*/r_{\sigma,f,u})^\top(\mathbf{U}_t\check{\mathbf{f}}_t^\top(\check{\mathbf{F}}^\top\check{\mathbf{F}})^{-1}\Bigg\|^2\Bigg(\sum_{t=1}^{T}\|\mathbf{f}_t\|^2\Bigg)$$
$$\leq \Bigg\|\sum_{t=1}^{T}(\widetilde{\boldsymbol{\omega}}_U^*/r_{\sigma,f,u})^\top\mathbf{U}_t\check{\mathbf{f}}_t^\top\Bigg\|^2\cdot\Big\|(\check{\mathbf{F}}^\top\check{\mathbf{F}})^{-1}\Big\|^2\Bigg(\sum_{t=1}^{T}\|\mathbf{f}_t\|^2\Bigg).$$

Because $(\widetilde{\boldsymbol{\omega}}_U^*/r_{\sigma,f,u})^\top\mathbf{U}_t \sim N(\sqrt{\theta_u}/r_{\sigma,f,u},\theta_u/(r_{\sigma,f,u}^2\theta_{all}))$, and independent with $\mathbf{f}_t$, we have

$$P\Bigg(\Bigg\|\sum_{t=1}^{T}(\widetilde{\boldsymbol{\omega}}_U^*/r_{\sigma,f,u})^\top\mathbf{U}_t\check{\mathbf{f}}_t^\top\Bigg\| > \sqrt{(\log N)T}\Bigg) = O\Bigg(\frac{1}{T^2}\Bigg).$$

It follows from (A.12) that

$$P\Bigg(\sum_{t=1}^{T}\Big((\widetilde{\boldsymbol{\omega}}_c^*)^\top(\widehat{\mathbf{r}}_t-\mathbf{r}_t)\Big)^2 > C\log N\Bigg) = O\Bigg(\frac{1}{T^2}\Bigg). \tag{A.17}$$

Combining (A.15), (A.16) and (A.17) yields

$$P\Big(III > \sqrt{\log N}\Big) = O\Bigg(\frac{1}{T^2}\Bigg).$$

About term IV, we have

$$\|(\widehat{\mathbf{r}}-\mathbf{r})(\widehat{\mathbf{r}}-\mathbf{r})^\top\widetilde{\boldsymbol{\omega}}_c^*\|_{\max}$$
$$\leq\sqrt{\max_{1\leq i\leq N}\Bigg(\sum_{t=1}^{T}(\widehat{U}_{it}-U_{it})^2\Bigg)}\sqrt{\sum_{t=1}^{T}\Big((\widetilde{\boldsymbol{\omega}}_c^*)^\top(\widehat{\mathbf{r}}_t-\mathbf{r}_t)\Big)^2}.$$

By (A.13) and (A.17), we then get

$$P\Bigg(\|IV\|_{\max} > \frac{\log N}{\sqrt{T}}\Bigg) = O\Bigg(\frac{1}{T^2}\Bigg).$$

Finally, about term V, we have

$$\|V\|_{\max} = \frac{1}{\sqrt{T}}\Bigg(\max_{1\leq i\leq N+K}\Bigg|\sum_{t=1}^{T}r_{it}\Big(1-\mathbf{r}_t^\top\widetilde{\boldsymbol{\omega}}_c^*\Big)\Bigg|\Bigg).$$

Note that

$$E\Big(\mathbf{r}_t(1-\mathbf{r}_t^\top\widetilde{\boldsymbol{\omega}}_c^*)\Big) = \mathbf{0}.$$

In addition, under Assumption 3, for some constant $C > 0$,

$$\max_{1\leq i\leq N+K} E\left(\left(r_{it}(1-\mathbf{r}_t^\top\widetilde{\boldsymbol{\omega}}_c^*)\right)^2\right) \leq \max_{1\leq i\leq N+K}\sqrt{E\left(r_{it}^4\right)}\sqrt{E\left((1-\mathbf{r}_t^\top\widetilde{\boldsymbol{\omega}}_c^*)^4\right)} < C.$$

Hence, under Assumption 2, by Bernstein's inequality, we get

$$P\left(\|V\|_{\max} \geq \sqrt{\log N}\right) = O\left(\frac{1}{T^2}\right).$$

Combining the results above yields

$$P\left(\|W\|_{\max} > C\sqrt{\log N}\right) = O\left(\frac{1}{T}\right) \to 0. \tag{A.18}$$

About term C_{11}, we have

$$\begin{aligned} C_{11} =& \frac{1}{T}(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})^\top \\ &+ \frac{1}{T}\mathbf{r}_{S_1}\mathbf{r}_{S_1}^\top \\ &+ \frac{1}{T}(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})\mathbf{r}_{S_1}^\top \\ &+ \frac{1}{T}\mathbf{r}_{S_1}(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})^\top. \end{aligned} \tag{A.19}$$

By (A.13), with probability approaching one,

$$\left\|\frac{1}{T}(\widehat{\mathbf{r}}-\mathbf{r})(\widehat{\mathbf{r}}-\mathbf{r})^\top\right\|_{\max} < C\frac{\log N}{T}. \tag{A.20}$$

By (A.20) and Assumption 5 that $\|\boldsymbol{\Sigma}_{f,u,11}^{-1}\| = O(1)$, we have

$$\left\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\right\| = O\left(\frac{q_u\log N}{T}\right). \tag{A.21}$$

Assumption 2 implies that $\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\mathbf{r}_{S_1,t} \sim N(\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}, \mathbf{I})$. By Bernstein's inequality, we have, with probability tending one,

$$\|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\bar{\mathbf{r}}_{S_1} - \boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\|_{\max} = O\left(\sqrt{\frac{\log N}{T}}\right). \tag{A.22}$$

Assumption 3 implies that $\|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\| = \sqrt{\boldsymbol{\mu}_{f,u,1}^\top \boldsymbol{\Sigma}_{f,u,11}^{-1}\boldsymbol{\mu}_{f,u,1}} = O(1)$. Therefore,

$$
\begin{aligned}
&\|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\bar{\mathbf{r}}_{S_1}\bar{\mathbf{r}}_{S_{11}}\boldsymbol{\Sigma}_{f,u,11}^{-1/2} - \boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\| \\
\leq& 2\|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\bar{\mathbf{r}}_{S_1} - \boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\| \cdot \|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\| + \|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\bar{\mathbf{r}}_{S_1} - \boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\|^2 \\
=& O\left(\sqrt{\frac{q_u \log N}{T}}\right).
\end{aligned}
\tag{A.23}
$$

Note that under Assumption 2, $\boldsymbol{\Sigma}_{f,u,11}^{-1/2}(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}$ follows Wishart distribution with df $T-1$ and covariance matrix $\mathbf{I}$, and can be written as

$$
\boldsymbol{\Sigma}_{f,u,11}^{-1/2}(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2} = \sum_{t=1}^{T-1}\mathbf{z}_t\mathbf{z}_t^\top,
$$

where $\mathbf{z}_t$ are i.i.d. multivariate standard normal. By Theorem 2 of El Karoui (2003) and Theorem I.1.1 of Feldheim and Sodin (2010), we have, with probability tending one,

$$
\left\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2} - \mathbf{I}\right\| = O\left(\sqrt{\frac{q_u}{T}}\right). \tag{A.24}
$$

Note also that $\frac{1}{T}\mathbf{r}_{S_1}\mathbf{r}_{S_1}^\top = \frac{1}{T}(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)^\top + \bar{\mathbf{r}}_{S_1}\bar{\mathbf{r}}_{S_1}^\top$. We have

$$
\begin{aligned}
&\left\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\mathbf{r}_{S_1}\mathbf{r}_{S_1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2} - \mathbf{I} - \boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\right\| \\
\leq&\left\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)(\mathbf{r}_{S_1} - \bar{\mathbf{r}}_{S_1}\mathbf{1}^\top)^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2} - \mathbf{I}\right\| \\
&+ \|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\bar{\mathbf{r}}_{S_1}\bar{\mathbf{r}}_{S_1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2} - \boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\|.
\end{aligned}
\tag{A.25}
$$

Combining (A.23), (A.24) and (A.25) yields

$$
\left\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\mathbf{r}_{S_1}\mathbf{r}_{S_1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2} - \mathbf{I} - \boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\right\| = O\left(\sqrt{\frac{q_u \log N}{T}}\right). \tag{A.26}
$$

By $\|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\| = O(1)$, $q_u^2(\log N) = o(T)$ and (A.26), we have

$$
\left\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\mathbf{r}_{S_1}\mathbf{r}_{S_1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\right\| = O(1). \tag{A.27}
$$

Therefore, by (A.24) and (A.27), we get that

$$\begin{aligned}
&\left\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})\mathbf{r}_{S_1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\right\| \\
\leq&\sqrt{\left\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\right\|}\sqrt{\left\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\mathbf{r}_{S_1}\mathbf{r}_{S_1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\right\|} \\
=&O\left(\sqrt{\frac{q_u\log N}{T}}\right),
\end{aligned} \tag{A.28}$$

and similarly,

$$\|\frac{1}{T}\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\mathbf{r}_{S_1}(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\| = O\left(\sqrt{\frac{q_u\log N}{T}}\right). \tag{A.29}$$

Combining (A.19), (A.21), (A.26), (A.28) and (A.29) yields that

$$\|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}C_{11}\boldsymbol{\Sigma}_{u,11}^{-1/2}-\mathbf{I}-\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\| = O\left(\sqrt{\frac{q_u\log N}{T}}\right).$$

Because $q_u^2(\log N)/T = o(1)$, we then get

$$\|\boldsymbol{\Sigma}_{f,u,11}^{1/2}C_{11}^{-1}\boldsymbol{\Sigma}_{f,u,11}^{1/2}-(\mathbf{I}+\boldsymbol{\Sigma}_{u,11}^{-1/2}\boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top\boldsymbol{\Sigma}_{f,u,11}^{-1/2})^{-1}\| = O\left(\sqrt{\frac{q_u\log N}{T}}\right).$$

In addition, because $\|\boldsymbol{\Sigma}_{f,u,11}^{-1/2}\| = O(1)$, we get that

$$\|C_{11}^{-1}-(\boldsymbol{\Sigma}_{f,u,11}+\boldsymbol{\mu}_{f,u,1}\boldsymbol{\mu}_{f,u,1}^\top)^{-1}\| = O\left(\sqrt{\frac{q_u\log N}{T}}\right).$$

Finally, about term C_{21}, we have

$$\begin{aligned}
C_{21} =&\frac{1}{T}(\widehat{\mathbf{r}}_{S_1^c}-\mathbf{r}_{S_1^c})(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})^\top \\
&+\frac{1}{T}(\widehat{\mathbf{r}}_{S_1^c}-\mathbf{r}_{S_1^c})\mathbf{r}_{S_1}^\top \\
&+\frac{1}{T}\mathbf{r}_{S_1^c}(\widehat{\mathbf{r}}_{S_1}-\mathbf{r}_{S_1})^\top \\
&+\frac{1}{T}\mathbf{r}_{S_1^c}\mathbf{r}_{S_1}^\top.
\end{aligned}$$

By Assumptions 2 and Bernstein's inequality, with probability approaching one,

$$\left\|\frac{1}{T}\mathbf{r}\mathbf{r}^\top-\boldsymbol{\Sigma}_{f,u}-\boldsymbol{\mu}_{f,u}\boldsymbol{\mu}_{f,u}^\top\right\|_{\max} < C\sqrt{\frac{\log N}{T}},$$

and hence

$$\begin{aligned}
&\left\|\frac{1}{T}(\widehat{\mathbf{r}}-\mathbf{r})\mathbf{r}^\top\right\|_{\max}\\
<&\sqrt{\left\|\frac{1}{T}(\widehat{\mathbf{r}}-\mathbf{r})(\widehat{\mathbf{r}}-\mathbf{r})^\top\right\|_{\max}}\sqrt{\left\|\frac{1}{T}\mathbf{r}\mathbf{r}^\top\right\|_{\max}}\\
<&C\sqrt{\frac{\log N}{T}}.
\end{aligned}$$

Similarly,

$$\left\|\frac{1}{T}\mathbf{r}(\widehat{\mathbf{r}}-\mathbf{r})^\top\right\|_{\max} < C\sqrt{\frac{\log N}{T}}.$$

We then get

$$\|C_{21}-\boldsymbol{\Sigma}_{f,u,21}-\boldsymbol{\mu}_{f,u,2}\boldsymbol{\mu}_{f,u,1}^\top\|_{\max} \le C\sqrt{\frac{\log N}{T}}.$$

The desired result (A.3) follows. □

A.3 Proof of Theorem 3

We have

$$\begin{aligned}
&\widehat{\boldsymbol{\Sigma}}_S^{-1}\widehat{\boldsymbol{\mu}}_S-\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\\
=&(\widehat{\boldsymbol{\Sigma}}_S^{-1}-\boldsymbol{\Sigma}_S^{-1})(\widehat{\boldsymbol{\mu}}_S-\boldsymbol{\mu}_S)\\
&+\boldsymbol{\Sigma}_S^{-1}(\widehat{\boldsymbol{\mu}}_S-\boldsymbol{\mu}_S)\\
&+\boldsymbol{\Sigma}_S^{-1/2}(\boldsymbol{\Sigma}_S^{1/2}\widehat{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S^{1/2}-\mathbf{I})(\mathbf{I}-\boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2})\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\\
&+(\boldsymbol{\Sigma}_S^{-1}-\boldsymbol{\Sigma}_S^{-1}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1})\boldsymbol{\mu}_S\\
=:&I+II+III+IV.
\end{aligned}\tag{A.30}$$

Assumption 3 implies that $\|\boldsymbol{\mu}_S\|_{\max}=O(1)$, and $\|\operatorname{diag}(\boldsymbol{\Sigma}_S)\|_{\max}=O(1)$. By Bernstein's inequality, we have

$$\|\widehat{\boldsymbol{\mu}}_S-\boldsymbol{\mu}_S\|=O_p\left(\sqrt{\frac{|S|(\log|S|)}{T}}\right).\tag{A.31}$$

Note that under Assumption 2, $T\boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2}$ follows Wishart distribution with df $T-1$ and covariance matrix $\mathbf{I}$, and can be written as

$$T\boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2}=\sum_{t=1}^{T-1}\mathbf{z}_t\mathbf{z}_t^\top,\tag{A.32}$$

where $\mathbf{z}_t$ are i.i.d. multivariate standard normal. By Theorem 2 of El Karoui (2003) and Theorem I.1.1 of Feldheim and Sodin (2010), we have,

$$\|\boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2}-\mathbf{I}\| = O_p\left(\sqrt{\frac{|S|}{T}}\right). \tag{A.33}$$

By the assumption that $|S|^2(\log|S|)=o(T)$, we get that, with probability tending one, for all large T, $\widehat{\boldsymbol{\Sigma}}_S^{-1}$ exists, $\|\widehat{\boldsymbol{\Sigma}}_S^{-1}\| = O_p(1)$, and

$$\|\boldsymbol{\Sigma}_S^{1/2}\widehat{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S^{1/2}-\mathbf{I}\| = o_p(1). \tag{A.34}$$

In addition, because $\widehat{\boldsymbol{\Sigma}}_S^{-1}-\boldsymbol{\Sigma}_S^{-1}=\boldsymbol{\Sigma}_S^{-1/2}(\boldsymbol{\Sigma}_S^{1/2}\widehat{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S^{1/2}-\mathbf{I})\boldsymbol{\Sigma}_S^{-1/2}$, we get that

$$\|\widehat{\boldsymbol{\Sigma}}_S^{-1}-\boldsymbol{\Sigma}_S^{-1}\| = O_p\left(\sqrt{\frac{|S|}{T}}\right). \tag{A.35}$$

It follows that

$$\|I\| = O_p\left(|S|\frac{\sqrt{\log|S|}}{T}\right) = o_p\left(\frac{1}{\sqrt{T}}\right), \tag{A.36}$$

where the last equality holds by the assumption that $|S|^2(\log|S|)=o(T)$.

About term II, by Assumption 2, we have

$$II \sim N\left(0, \frac{1}{T}\boldsymbol{\Sigma}_S^{-1}\right). \tag{A.37}$$

About term III, we have

$$\begin{aligned}\|III\| =& \|\boldsymbol{\Sigma}_S^{-1/2}(\boldsymbol{\Sigma}_S^{1/2}\widehat{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S^{1/2}-\mathbf{I})(\mathbf{I}-\boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2})\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\| \\ \le& \|\boldsymbol{\Sigma}_S^{-1/2}\|\|\boldsymbol{\Sigma}_S^{1/2}\widehat{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S^{1/2}-\mathbf{I}\|\|\mathbf{I}-\boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2}\|\|\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\|.\end{aligned}$$

Note that

$$\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S \le \theta_{all} = \boldsymbol{\mu}_f^\top\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\mu}_f + \theta_u.$$

Hence by Assumption 3,

$$\|\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\| = \sqrt{\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S} = O(1). \tag{A.38}$$

The assumption that $\|\boldsymbol{\Sigma}_{full}^{-1}\| = O(1)$ implies

$$\|\boldsymbol{\Sigma}_S^{-1}\| = O(1). \tag{A.39}$$

By (A.33), (A.34) and (A.39), we then get

$$\|III\| = O_p\left(\frac{|S|}{T}\right) = o_p\left(\frac{1}{\sqrt{T}}\right). \tag{A.40}$$

About term IV, by (A.32), we have

$$IV = \boldsymbol{\Sigma}_S^{-1/2}\frac{1}{T}\sum_{t=1}^{T-1}(\mathbf{I} - \mathbf{z}_t\mathbf{z}_t^\top)\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S + \frac{1}{T}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S. \tag{A.41}$$

By Assumption 3, we have $\|\boldsymbol{\mu}_S\| = O(\sqrt{|S|})$. Therefore, by (A.39), we have

$$\frac{1}{T}\|\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\| = O\left(\frac{1}{T}\|\boldsymbol{\Sigma}_S^{-1}\|\|\boldsymbol{\mu}_S\|\right) = O\left(\frac{\sqrt{|S|}}{T}\right) = o\left(\frac{1}{\sqrt{T}}\right). \tag{A.42}$$

For any non-random matrix $\mathbf{A}$,

$$E\Big(\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}(\mathbf{I} - \mathbf{z}_t\mathbf{z}_t^\top)\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\Big) = \mathbf{0}. \tag{A.43}$$

In addition, for any non-random $|S| \times |S|$ matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ (that are no need to be symmetric or positive definite),

$$E(\mathbf{z}_t^\top\mathbf{B}_1\mathbf{z}_t\mathbf{z}_t^\top\mathbf{B}_2\mathbf{z}_t) = \mathrm{tr}(\mathbf{B}_1)\,\mathrm{tr}(\mathbf{B}_2) + \mathrm{tr}(\mathbf{B}_1\mathbf{B}_2) + \mathrm{tr}(\mathbf{B}_1\mathbf{B}_2^\top). \tag{A.44}$$

By (A.44), we get that

$$\begin{aligned} &E\Big(e_i^\top(\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S)(\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S)^T e_j\Big) \\ =&e_i^\top\big((\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S)\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^\top + 2\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^\top\big)e_j, \end{aligned} \tag{A.45}$$

where e_i is a length-k vector with the ith element of one and zero elsewhere. By (A.43) and (A.45), we have

$$\begin{aligned} &\mathrm{Cov}(\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S) \\ =&(\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S)\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^\top \\ &+\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^\top. \end{aligned} \tag{A.46}$$

In addition, by the Cauchy-Schwarz inequality,

$$
\begin{aligned}
&E\Big(\|\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\|^4\Big)\\
=&E\Big((\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\mathbf{A}^\top\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t)^2(\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t)^2\Big)\\
\leq&\sqrt{E\Big((\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\mathbf{A}^\top\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t)^4\Big)}\sqrt{E\Big((\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t)^4\Big)}.
\end{aligned}
$$

By Lemma 2.9 of Bai and Silverstein (1998),

$$
\begin{aligned}
&E\Big((\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\mathbf{A}^\top\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t)^4\Big)\\
\leq&16E\Big(\big(\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\mathbf{A}^\top\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t-\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^\top)\big)^4\Big)\\
&+16\Big(\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^\top)\Big)^4\\
\leq&Ck^4\|\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^\top\|^4=O(1),
\end{aligned}
$$

where the last equality holds by the assumptions that $\|\boldsymbol{\Sigma}_S^{-1}\|=O(1)$ and $\|\mathbf{A}\|=O(1)$. Similarly,

$$
E\Big((\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t)^4\Big)=O\Big((\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S)^4\Big)=O(1),
$$

where the last equality holds by (A.38). It follows that

$$
E\Big(\|\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t\mathbf{z}_t^\top\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\|^4\Big)=O(1). \tag{A.47}
$$

By the Lyapunov Central Limit Theorem,

$$
\mathbf{A}(IV)\xrightarrow{\mathcal{L}}N(0,\boldsymbol{\Sigma}_{A,IV}),
$$

where $\boldsymbol{\Sigma}_{A,IV}=\lim_{N\to\infty}(\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S)\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^\top+\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^\top\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^\top$. Using further the independence between $\widehat{\boldsymbol{\mu}}_S$ and $\widehat{\boldsymbol{\Sigma}}_S$, and (A.37), we obtain

$$
\mathbf{A}(II+IV)\xrightarrow{\mathcal{L}}N(0,\boldsymbol{\Sigma}_A). \tag{A.48}
$$

The desired result follows from (A.30), (A.36), (A.40) and (A.48). □

Appendix B Additional empirical results: MAXSER-S by long-short characteristic-sorted portfolios

In this section, we further examine the MAXSER-S SDF, which is constructed by long-short characteristic-sorted portfolios.

B.1 Spanning tests

We repeat the test in Table 2 to examine whether MAXSER-S can be priced by benchmark models. We see from Table B1 that all benchmark models fail to price our estimated MAXSER-S SDF portfolio; the alpha of MAXSER-S is both economically large and statistically significant under all benchmark models. Table B1 shows that the monthly alphas are greater than 0.86% with a t-statistic higher than 2.9 across all models. These results suggest that the benchmark models can not capture the returns of our MAXSER-S portfolio.

Table B1: **Testing the pricing of MAXSER-S with benchmark models.** This table reports alphas from regressions of the SDF portfolio from MAXSER-S against various benchmark models, using monthly returns from 2000 to 2023. The MAXSER-S is constructed by long-short characteristic-sorted portfolios. Alpha in percentage and its corresponding t-statistic are reported.

	Alpha (%)	t-statistic
CAPM	1.91	4.60
FF3	1.87	5.11
FF5	1.56	4.76
FF6	1.52	2.95
Q	1.58	3.55
Q5	1.31	3.30
BS6	1.69	3.26
SY4	1.60	3.61
DHS3	1.60	3.89
KNS	0.86	2.96

B.2 Encompassing tests

Next, we repeat the test in Table 3. Table B2 reports the alpha and its t-statistic from regression (4.3). We see from Table B2 that all alphas are insignificant at the t-statistic threshold level of 1.96. These results suggest that our MAXSER-S SDF reasonably captures those commonly used pricing factors. But, compared with the results in Table 3, Table B2 suggests that the MAXSER-S constructed from long-short characteristic-sorted portfolios performs worse than that constructed from cross-sectional regressions.

Table B2: **Testing the pricing of commonly used factors with MAXSER-S.** This table reports alphas from regressions of commonly used pricing factors against MAXSER-S, using monthly returns from 2000 to 2023. The MAXSER-S SDF is constructed by long-short characteristic-sorted portfolios. Alpha in percentage and its corresponding t-statistic are reported.

Factor	Alpha (%)	t-statistic
Mkt.RF	0.47	1.39
SMB	0.19	1.02
HML	0.06	0.19
RMW	0.30	1.84
CMA	0.16	1.05
Mom	-0.19	-0.45
R_ME	0.23	1.23
R_IA	0.10	0.62
R_ROE	0.19	0.79
R_EG	0.34	1.91
HmLm	0.23	0.53
MGMT	0.19	1.01
PERF	0.33	0.63
PEAD	0.27	1.78
FIN	0.25	0.92
KNS	0.27	1.55

B.3 Asset pricing tests

Finally, we test the pricing power of MAXSER-S over the 153 anomaly portfolios.

Table B3: **The number of significant alphas from testing 153 anomalies against various SDFs.** We examine the pricing power of various SDFs over 153 anomalies. The MAXSER-S is constructed by long-short characteristic-sorted portfolios. This table summarizes the total number of significant alphas and the number of significant alphas in each characteristic theme. The number in parenthesis is the total number of portfolios in each characteristic theme. The number of significant alphas under the threshold $t > 1.96$ or $t > 3$ is reported. The evaluation period is between 2000 and 2023.

	All (153)		Accruals (6)		Debt Issuance (7)		Investment (22)	
Threshold of t	1.96	3	1.96	3	1.96	3	1.96	3
MAXSER-S	9	2	1	0	0	0	1	0
KNS	6	2	0	0	0	0	0	0
CAPM	59	21	0	0	1	1	4	0
FF3	75	45	2	0	2	1	6	1
FF5	38	13	2	0	1	1	2	0
FF6	39	11	2	0	1	1	3	0
Q	37	12	2	1	1	1	2	0
Q5	21	1	2	0	0	0	0	0
BS6	50	21	2	1	1	1	4	0
SY4	24	3	0	0	0	0	1	0
DHS3	28	4	1	0	1	0	1	0

	Low Leverage (11)		Low Risk (18)		Momentum (8)		Profit Growth (12)	
Threshold of t	1.96	3	1.96	3	1.96	3	1.96	3
MAXSER-S	0	0	0	0	0	0	3	2
KNS	0	0	0	0	0	0	3	2
CAPM	1	0	16	5	2	0	4	1
FF3	3	2	17	13	2	0	4	1
FF5	5	2	3	0	0	0	3	1
FF6	5	1	3	0	0	0	3	1
Q	1	0	3	0	0	0	2	2
Q5	1	0	3	0	0	0	3	1
BS6	7	2	3	0	0	0	2	2
SY4	0	0	1	0	1	0	1	1
DHS3	3	0	1	0	0	0	2	0

Table B3: **The number of significant alphas from testing 153 anomalies against various SDFs, continued.**

	Profitability (11)		Quality (17)		Seasonality (12)		Short-Term Reversal (5)	
Threshold of t	1.96	3	1.96	3	1.96	3	1.96	3
MAXSER-S	0	0	3	0	0	0	0	0
KNS	0	0	1	0	0	0	0	0
CAPM	8	6	12	7	3	0	2	0
FF3	9	7	15	13	2	0	2	2
FF5	4	1	12	5	1	1	1	1
FF6	4	0	11	5	1	0	1	1
Q	5	2	14	5	1	0	2	0
Q5	1	0	5	0	2	0	1	0
BS6	6	2	14	11	3	1	1	0
SY4	3	0	5	0	2	0	0	0
DHS3	3	0	9	2	2	0	1	1
	Size (5)		Value (18)					
Threshold of t	1.96	3	1.96	3				
MAXSER-S	1	0	0	0				
KNS	1	0	1	0				
CAPM	1	0	5	1				
FF3	1	0	10	5				
FF5	1	1	2	1				
FF6	1	1	3	1				
Q	1	1	3	0				
Q5	2	0	1	0				
BS6	1	1	6	0				
SY4	3	0	7	2				
DHS3	1	1	3	0				

In Table B3, we see that again, the MAXSER-S portfolio has the smallest number of significant alphas. There are 9 (2) anomalies have significant alphas for a threshold of $t = 1.96$ ($t = 3$). KNS has 6 (2) significant alphas for a threshold of $t = 1.96$ ($t = 3$). Overall, we see that MAXSER-S performs better in the seasonality and low risk themes than Q5 and KNS. The performance of MAXSER-S in Table B3 is similar to that in Table 4.

B.4 Impacts of transaction costs

We also evaluate the impact of transaction costs on the performance of the MAXSER-S. Kan, Wang, and Zhou (2022) suggests using 20 bps per dollar of transaction when evaluating the post-transaction cost performance of investment in anomaly portfolios. Following Kan, Wang, and Zhou (2022), we check the performance of the SDF portfolio after deducting the transaction costs, assuming a cost of 20 bps per dollar of transaction. In addition, we also check the results assuming a higher cost of 40 bps per dollar of transaction. Because we use the long-short characteristic-sorted portfolios, we are not able to compute the stock-level turnover for the MAXSER-S. Therefore, to be conservative, here we assume a higher transaction cost than that in Table 7. The performance of our SDF estimator after adjusting for the transaction costs is summarized in Table B4.

Table B4: **Performance of MAXSER-S after adjusting for transaction costs.** The transaction cost is assumed to be 20 bps (40 bps) per dollar of transaction. The portfolios are re-estimated every year from 2000 to 2023. The summary statistics include the monthly mean return (Mean), standard deviation (SD), Sharpe ratio (SR), and the maximum drawdown (MDD) of the SDF portfolio. All are reported in percentage except the Sharpe ratio.

transaction cost	Mean	SD	SR	MDD
20 bps	1.71	5.15	0.33	32.95
40 bps	1.49	5.20	0.29	35.16

We see from Table B4 that, the portfolio maintains a high Sharpe ratio after adjusting for the transaction costs. For example, the Sharpe ratio is 0.33 after removing 20 bps transaction cost per dollar, and is 0.29 when the cost is 40 bps per dollar.

Next, we repeat Table B1, adjusting for transaction costs. That is, we regress the MAXSER-S portfolio against various factor models (regression (4.2)) in Table B5. For the variables from other benchmark models, we use the original data without removing the transaction cost. We see that again, our MAXSER-S portfolio has a statistically significant alpha against benchmark models.

Table B5: **Testing the pricing of MAXSER-S with benchmark models, adjusting for transaction costs.** This table presents alphas from regressions of MAXSER-S against various benchmark models, using the monthly returns from 2000 to 2023. The transaction costs of MAXSER-S are assumed to be 20 bps or 40 bps per dollar of transaction. The pricing factors from various benchmark models are the original data without removing the transaction costs. The portfolios are re-estimated every year. We report the alpha in percentage and t-statistic.

Transaction cost	20 bps		40 bps	
Model	Alpha (%)	t-statistic	Alpha (%)	t-statistic
CAPM	1.69	5.52	1.47	3.72
FF3	1.65	5.40	1.43	4.04
FF5	1.36	4.33	1.18	3.68
FF6	1.32	4.28	1.14	2.26
Q	1.36	4.42	1.15	2.61
Q5	1.10	3.50	0.89	2.20
BS6	1.48	4.69	1.28	2.56
SY4	1.40	3.72	1.19	2.84
DHS3	1.38	4.45	1.17	3.05
KNS	0.72	3.11	0.58	2.19

Last, we repeat Table B2 but take into account of transaction costs. We regress various pricing factors against our SDF after adjusting for transaction costs (regression (4.3)). For the portfolios from other benchmark models, we use their original returns without removing the transaction cost. The results are summarized in Table B6. We see results similar to those reported in Table B2. That is, alphas of pricing factors from other benchmark models are still insignificant except for RMW, R_EG, and $PEAD$ at the presence of the MAXSER-S.

Table B6: **Testing the pricing of commonly used factors with MAXSER-S, adjusting for transaction costs.** This table reports alphas from regressions of various pricing factors against MAXSER-S after adjusting for the transaction costs, using the monthly returns from 2000 to 2023. The transaction cost of MAXSER-S is assumed to be 20 bps or 40 bps per dollar of transaction. The pricing factors from various models are the original returns without removing the transaction costs. Alpha in percentage and its t-statistic are reported.

Transaction cost	20 bps		40 bps	
Factor	Alpha (%)	t-statistic	Alpha (%)	t-statistic
Mkt.RF	0.49	1.48	0.51	1.56
SMB	0.21	1.12	0.22	1.20
HML	0.08	0.31	0.11	0.42
RMW	0.33	1.97	0.35	2.07
CMA	0.19	1.23	0.21	1.40
Mom	-0.16	-0.41	-0.13	-0.35
R_ME	0.24	1.36	0.26	1.48
R_IA	0.13	0.78	0.15	0.94
R_ROE	0.20	0.88	0.22	0.99
R_EG	0.35	2.02	0.37	2.12
HmLm	0.24	0.59	0.26	0.65
MGMT	0.24	1.17	0.28	1.34
PERF	0.38	0.73	0.43	0.86
PEAD	0.28	1.86	0.28	1.97
FIN	0.30	1.09	0.34	1.26
KNS	0.27	1.66	0.29	1.75