Learning the Stochastic Discount Factor

September 12, 2024

Abstract

We develop a statistical framework to learn the high-dimensional stochastic discount factor (SDF) from a large set of characteristic-based portfolios. Specifically, we provide statistical support to use the MAXSER method proposed in Ao et al. (2019) to screen for potentially useful factors, and develop a statistical inference theory for further factor selection and final SDF portfolio construction. Applying our approach to 194 characteristic-based portfolios, we find that our SDF estimator performs well in achieving a high Sharpe ratio and explaining the cross-section of expected returns of various portfolios.

JEL classification: C55, C58, G11, G12

Keywords: Stochastic discount factor; Factor models; High dimensions; Sparse regressions; Maximum Sharpe ratio regression

1 Introduction

Constructing the stochastic discount factor (SDF) or factor models is critical in asset pricing and factor investing. Decades of empirical studies have identified a fruitful set of factors and factor models. This large "zoo" of factors poses challenges as well as opportunities to better understand the compositions of the SDF, i.e., the sources of SDF. The central question of the SDF learning is, as Cochrane (2011) points out, "which characteristics provide truly independent useful information explaining the returns, and hence contribute to the SDF?" In this paper, we attempt to address this question by developing a statistical learning framework to construct the SDF and explore its properties.

Learning the SDF is challenging due to the high-dimensional nature of the problem. There are over hundreds of potential factors, while the number of monthly historical observations are only hundreds. 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 often constructed by portfolio sorting (e.g., based on some firm characteristics), and then factor models are tested against a set of test assets. This approach suffers from two limitations. First, it can only handle few (usually two or three) characteristics and cannot fully address the interaction among many different characteristics. Second, the choice of test assets is less explored, which influences the empirical evaluation. Especially, 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 (Kan and Zhang, 1999; Kleibergen, 2009; Bryzgalova, 2017; Gospodinov et al., 2017; Anatolyev and Mikusheva, 2022; Giglio et al., 2022). For example, many macroeconomic factors appear to be weak factors and test assets may be unable to identify risk premia of these weak factors.

To avoid the limitations of the above approach, in this paper, we follow the right-hand side approach suggested in Barillas and Shanken (2017) and Fama and French (2018). 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. 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 et al. (2020) apply this metric to empirically compare several

prevailing factor models. Following the right-hand side approach, we set our objective to be identifying the portfolio that is the one with the maximum Sharpe ratio on the mean-variance efficient frontier. Such an SDF portfolio is mean-variance efficient and explains 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.

Specifically, we build on a large set of potential factor portfolios in Hou et al. (2020) to construct the SDF portfolio. To tackle the high-dimensional SDF loading estimation problem, we apply the *maximum - Sharpe ratio estimated & sparse regression* (MAXSER) approach proposed by Ao et al. (2019) to screen for potentially useful factors for the SDF. As has been shown by Ao et al. (2019), MAXSER can asymptotically achieve the mean-variance efficiency asymptotically under the high-dimensional setting. We further show in this paper that the MAXSER estimator can consistently select the factors with nonzero weights in constructing the SDF, which guarantees MAXSER's screening effectiveness. With the number of factors significantly reduced by the MAXSER screening, we apply the plug-in method to estimate the optimal weights. We show both consistency and asymptotic normality of such estimated weights, which enable us to conduct further factor selection and construct the final SDF portfolio.

We apply our methodology to learning the SDF loadings on the monthly returns of 188 anomaly portfolios from Hou et al. (2020).[1] We also include the widely used six factors from the Fama-French five-factor model augmented with the momentum factor (Fama and French, 2018), so totally there are 194 factor portfolios. Our proposed SDF performs well in achieving a high Sharpe ratio and explaining the cross-section of expected returns of various portfolios. We find that important factors are mostly related to momentum and earnings, including the customer momentum, the cumulative abnormal stock return, the earnings predictability, and the quarterly earnings-to-price. Except the market factor, the Fama-French factors are not significant and not selected in our estimated SDF.

Next, 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 hereafer), Q4/Q5 models (Hou et al., 2015, 2021), the six-factor model in Barillas and Shanken (2018, hereafter, BS6), and behavioral models proposed in Stambaugh and Yuan (2017, hereafter, SY4) and Daniel et al. (2020, hereafter, DHS3). We also include the SDF estimator of Kozak et al. (2020, hereafter, KNS) as a benchmark case. The test results

[1]One might start with factors constructed from individual stocks by using firm characteristics. For example, Kozak and Nagel (2023) show that these could span SDF if the covariance of stock returns satisfies a specific structure. They argue that such requirement is more likely satisfied if a large number of firm characteristics are used.

show that our SDF significantly outperforms these benchmark factor models both statistically and economically. Our results are robust after taking into account the transaction costs.

Although we use anomaly portfolios to construct the SDF, it is important to note that this is not equivalent to using the most significant anomalies. It is possible that a particular factor has a zero alpha relative to some factor models but has a non-zero loading in our SDF because it helps to reduce the overall risk and achieve a higher overall Sharpe ratio due to the return dependence in characteristic-based portfolios. Therefore, it is important to consider the interactions among characteristics. That means, to construct the SDF, purely chasing factors that have significant alphas, as some studies do, is insufficient.

Our paper is closely related to the emerging literature on applying machine learning tools to identify factors and construct factor models.[2] Some papers apply machine learning techniques to estimate factor loadings. For example, Kelly et al. (2019a) develop an instrument principal component analysis (PCA) method to estimate loadings of latent factor models as linear functions of characteristics. Lettau and Pelger (2020b,a) propose a risk-premium PCA to estimate the SDF at the presence of weak factors. Kozak et al. (2020) impose economically motivated prior about SDF coefficients and apply shrinkage and selection methods to construct SDF. Gu et al. (2021) propose a nonlinear latent conditional factor model, based on asset characteristics learned with autoencoder neural networks. Giglio et al. (2022) suggest using a supervised PCA to select test assets and estimate risk premia for potentially weak factors. Bryzgalova et al. (2023b) emphasized the importance of considering the dependence of the characteristic and apply decision trees to build portfolios. Bryzgalova et al. (2023a) generalize principal component analysis to consider economic targets of cross-sectional and time-series moments which helps to detect weak factors. Meanwhile, some papers develop empirical tests. For example, Feng et al. (2020) propose a test for the significance of new factors under the existence of a factor model. Giglio et al. (2021) study multiple hypothesis testing problems in linear asset pricing models. 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. Second, we establish consistency and inference theory to support our SDF estimator.

Our study also relates to the empirical literature on identifying and interpreting characteristic-based portfolios. Numerous studies examine whether these portfolios are significant, robust, and replicable. For example, Hou et al. (2020) investigate over hun-

[2]Various machine learning approaches have been employed to predict asset returns. For example, Freyberger et al. (2020) use the adaptive group Lasso to select characteristics that provide incremental information for the cross-section of expected returns. Gu et al. (2020) use a variety of machine learning techniques to predict stock returns using characteristics.

dreds of anomalies and find that half of them fail to replicate with more recent data or a different portfolio construction scheme than the original paper. The factor zoo raises concerns about data mining, transaction costs, and replication (McLean and Pontiff, 2016; Harvey et al., 2016; Andrews and Kasy, 2019; Jacobs and Müller, 2020; Chen, 2021; Harvey and Liu, 2021a,b; Jensen et al., 2023). Different from these papers, our paper uses these portfolios to construct the pricing kernel, and the portfolios need not be anomaly portfolios. In fact, we find many important characteristic-based portfolios in our SDF do not have significant alphas relative to other factor models.

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 Black-Litterman's approach of expected return estimation (Black and Litterman, 1991), factor-model-based covariance matrix estimators (Fan et al., 2008, 2011, 2013), and shrinkage estimators (Ledoit and Wolf, 2017). Another approach is by adding constraints (e.g., no short sale constraint (Jagannathan and Ma (2003)) and gross exposure constraint (DeMiguel et al. (2009))). These approaches lead to improved numerical performance compared with plug-in portfolio but are suboptimal in achieving mean-variance targets. Ao et al. (2019) propose MAXSER 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 builds on Ao et al. (2019) and further establishes selection consistency of MAXSER and asymptotic normality to estimate SDF loadings.

The rest of this paper is organized as follows. We present our methodology and its statistical theory in Section 2. Section 3 presents main empirical results. Section 4 considers the impact of transaction cost, while Section 5 further explores the economic interpretations of our estimated SDF. Finally, we conclude in Section 6. The proofs are collected in Appendix A and additional empirical results are presented in Appendix B.

The following notation is used throughout the paper. For any matrix $\mathbf{A} = (a_{ij})$, its spectral norm is defined as $\|\mathbf{A}\| = \max_{\|\mathbf{x}\|\leq 1} \sqrt{\mathbf{x}^T\mathbf{A}\mathbf{x}}$, where $\|\mathbf{x}\| = \sqrt{\sum x_i}$ 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.

2 Methodology and theory

2.1 SDF and the mean-variance efficient portfolio

Under the beta-pricing models, factors ($\mathbf{R}_t$) can be constructed using characteristic-based stock portfolios, i.e., $\mathbf{R}_t = \mathbf{Z}_{t-1}^T \mathbf{R}_{s,t}$, where $\mathbf{R}_t$ is an $N \times 1$ vector of characteristic sorted factors, $\mathbf{Z}_{t-1}$ are $N_s \times N$ asset characteristics, $\mathbf{R}_{s,t}$ are $N_s \times 1$ excess stock returns. The stochastic discount factor (SDF) in this beta-pricing model can be written as

$$\begin{aligned} M_t &= \frac{1 - \mathbf{b}^T(\mathbf{R}_t - E(\mathbf{R}_t))}{R_{f,t}}, \\ E(M_t \mathbf{R}_t) &= 0, \end{aligned} \tag{2.1}$$

where $\mathbf{b}$ is the loading of the SDF on the characteristic sorted portfolio, and $R_{f,t}$ 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 $\mathbf{R}_t$.

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

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

where σ is the target risk constraint.[3] The optimal weight, $\boldsymbol{\omega}^* = (\omega_i^*)$, is

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

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

2.2 Mean-Variance optimization by MAXSER

First, we note that the returns of characteristic-based portfolios exhibit a clear factor structure; see Figure 1 in Section 3.2. For this reason, we suppose that the returns of the characteristic-based portfolios satisfy:

$$\mathbf{R}_t = \boldsymbol{\beta} \mathbf{f}_t + \mathbf{U}_t, \tag{2.4}$$

[3] 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 $\mathbf{f}_t$ is a $K \times 1$ vector of common factors, $\boldsymbol{\beta}$ is an $N \times K$ factor loadings, and $\mathbf{U}_t$ is the idiosyncratic return. We will use $(\mathbf{f}_t, \mathbf{R}_t)$ to construct the SDF.

We first recall the maximum-Sharpe-ratio estimated and sparse regression (MAXSER) approach in Ao et al. (2019). Given a risk constraint σ, using both $\mathbf{f}_t$ and $\mathbf{R}_t$ under model (2.4) to construct the optimal mean-variance portfolio, Proposition 3 of Ao et al. (2019) shows that the optimal weights, $\boldsymbol{\omega}^* = ((\boldsymbol{\omega}_f^*)^T, (\boldsymbol{\omega}_R^*)^T)^T$, is given by

$$\begin{aligned} \boldsymbol{\omega}_f^* &= \frac{\sigma}{\sqrt{\theta_{all}}} \boldsymbol{\Sigma}_f^{-1} \boldsymbol{\mu}_f - \frac{\sigma}{\sqrt{\theta_{all}}} \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\alpha}, \\ \boldsymbol{\omega}_R^* &= \frac{\sigma}{\sqrt{\theta_{all}}} \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\alpha}, \end{aligned} \tag{2.5}$$

where $\boldsymbol{\Sigma}_f$ and $\boldsymbol{\mu}_f$ are the covariance matrix and the mean of $\mathbf{f}_t$; $\boldsymbol{\Sigma}_u =: (\sigma_{u,ij})_{1 \le i,j \le N}$ and $\boldsymbol{\alpha}$ are the covariance matrix and the mean of $\mathbf{U}_t$; and $\theta_{all} = \boldsymbol{\mu}_f^T \boldsymbol{\Sigma}_f^{-1} \boldsymbol{\mu}_f + \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\alpha}$ is the squared Sharpe ratio of the optimal portfolio investing in factors and all assets.

The main difficulty in estimating $\boldsymbol{\omega}^*$ lies in estimating $\boldsymbol{\omega}_R^*$. Define

$$\boldsymbol{\omega}_u^* = \frac{1}{\sqrt{\theta_u}} \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\alpha},$$

where $\theta_u = \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_u^{-1} \boldsymbol{\alpha}$. $\boldsymbol{\omega}_u^*$ represents the optimal portfolio invested in the idiosyncratic components $\mathbf{U}$ with a risk constraint of one unit. We see that $\boldsymbol{\omega}_R^*$ is proportional to $\boldsymbol{\omega}_u^*$. MAXSER first obtains the optimal weights on idiosyncratic components, $\widehat{\boldsymbol{\omega}}_u^*$, using the following LASSO-type regression

$$\widehat{\boldsymbol{\omega}}_u^* = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} (\widehat{r}_{u,c} - \boldsymbol{\omega}^T \widehat{\mathbf{U}}_t)^2 + \lambda \|\boldsymbol{\omega}\|_1, \tag{2.6}$$

where $\|\mathbf{x}\|_1 := \sum |x_i|$ for any vector $\mathbf{x} = (x_i)$. In the above regression, $\widehat{r}_{u,c}$ is a consistent estimator of $\sqrt{\theta_u} + 1/\sqrt{\theta_u}$. Ao et al. (2019) further derive consistent estimators of the remaining components in (2.5) and obtain the final allocation. Ao et al. (2019) focus on the MAXSER portfolio's performance. In particular, they show that the estimated portfolio meets the risk constraints and asymptotically achieves the maximum Sharpe ratio.

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. Achieving these two goals requires deep investigations into the weights of MAXSER.

2.3 Sure screening property of MAXSER

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

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

We also recall the following assumptions from Ao et al. (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*

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

To study the estimated weights, we need the following assumptions. For any subset $S \subset \{1, ..., N\}$, $|S|$ denotes its cardinality of S and S^c denotes its complement set. Denote $S_{signal} = \{i : \omega^*_{u,i} \neq 0, 1 \le i \le N\}$.

Assumption 4 *There exists a set S_1 that satisfies $S_{signal} \subset S_1 \subset \{1, ..., N\}$, $|S_1|^{3/2}\sqrt{\log N/T} = o(1)$, and $\min_{i\in S_{signal}}|\omega^*_i| \gg \lambda\sqrt{|S_1|}/T$ with $\lambda \gg |S_1|\sqrt{(\log N)T}$ for λ in (2.6).*

Denote $\boldsymbol{\Sigma}_{u,11} = (\boldsymbol{\Sigma}_u)_{S_1,S_1}$, $\boldsymbol{\Sigma}_{u,21} = (\boldsymbol{\Sigma}_u)_{S_1^c,S_1}$, $\boldsymbol{\Sigma}_{u,22} = (\boldsymbol{\Sigma}_u)_{S_1^c,S_1^c}$, $\boldsymbol{\alpha}_1 = (\alpha_i)_{i\in S_1}$, $\boldsymbol{\alpha}_2 = (\alpha_i)_{i\in S_1^c}$, $\boldsymbol{\omega}^{*'}_R = (\boldsymbol{\omega}^*_R)_{S_1}$, and $S'_{signal} = \{i : i \le |S_1|, \text{ and } (\boldsymbol{\omega}^{*'}_R)_i \neq 0\}$. Further define the following set of vectors with elements being either 1 or -1:

$$V = \{(\delta_1, ..., \delta_{|S_1|})^T, \delta_i = \text{sign}((\boldsymbol{\omega}^*_u)_i) \text{ for } i \in S'_{signal}, \text{ and } \delta_i \in \{-1, 1\} \text{ for } i \in (S'_{signal})^c\}.$$

Assumption 5 $\max_{\tilde{\mathbf{1}}\in V}\|(\boldsymbol{\Sigma}_{u,21}+\boldsymbol{\alpha}_2\boldsymbol{\alpha}_1^T)(\boldsymbol{\Sigma}_{u,11}+\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T)^{-1}\tilde{\mathbf{1}}\|_{\max} < 1-\eta$ *for some constant $\eta > 0$, and $\|\boldsymbol{\Sigma}^{-1}_{u,11}\| < M$ for some constant $M < \infty$.*

Assumption 4, the sparsity assumption, is slightly different from Ao et al. (2019). We require the number of nonzero weights to be not too big. Moreover, the minimum nonzero weight is not too small. Assumption 5 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.

Zhao and Yu (2006) establish the sign consistency of the estimated coefficients from the LASSO regression. Wainwright (2009) shows 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. The difference lies in that in the LASSO regression, the inference theory is obtained by conditioning on the x-variable and utilizing the i.i.d. assumption imposed on the noise term. In our setting, if one conditions on the x-variable, 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-variable to develop asymptotic theory.

We study the composition of SDF. Note that there is no risk constraint as in Ao et al. (2019). In particular, we do not need to estimate the maximum squared Sharpe ratio as in Ao et al. (2019), which is the reason behind the assumption that $N < T$ in that paper. Instead, in (2.6), we can replace $\widehat{r}_{u,c}$ with an arbitrary constant c, say, $c = 1$. Specifically, one can use the following regression,

$$\widetilde{\boldsymbol{\omega}}_c^* = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} (1 - \boldsymbol{\omega}^T \widehat{\mathbf{U}}_t)^2 + \lambda \|\boldsymbol{\omega}\|_1. \tag{2.7}$$

The tuning parameter λ can be chosen by cross-validation using the criterion of maximizing Sharpe ratio.

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

Theorem 1 *Under Assumptions 1–5, the estimator* $\widetilde{\boldsymbol{\omega}}_c^*$ *in* (2.7) *satisfies*

$$P\Big(\operatorname{sign}((\widetilde{\boldsymbol{\omega}}_c^*)_i) = \operatorname{sign}((\boldsymbol{\omega}_R^*)_i) \quad \textit{for all } i \in S_{signal} \cup S_1^c\Big) \to 1. \tag{2.8}$$

Theorem 1 states that MAXSER approach achieves the sure screening property, namely, with probability tending to one, $S_{signal} \subset \hat{S}$, where $\hat{S} = \{i : (\widetilde{\boldsymbol{\omega}}_c^*)_i \neq 0, i \leq N\}$. In addition, $\hat{S} \subset S_1$, hence $|\hat{S}|$ is small. In summary, Theorem 1 guarantees that MAXSER can screen for true signals and effectively remove most of the irrelevant variables.

Proof: See Appendix A.1.

2.4 Statistical inference of SDF loadings

The sparsity of the estimated parameters from the LASSO regression depends on the tuning parameter λ. In 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 et al. (2006)). In addition, Leng

et al. (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., response being constant, their approach is not applicable to our setting. Another approach is to perform post inference based on variable selection from sparse regressions. Examples include Belloni et al. (2014) and Lee et al. (2016). In finance applications, Feng et al. (2020) develop factor testing methods based on OLS from the post-LASSO model selection. The framework of Feng et al. (2020) follows the ordinary regression setting as they use test asset returns as responses in the regression. Such inference results established for ordinary regression 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 using Eq. (2.2). 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); Fithian et al. (2014)). Specifically, we split the observations into two subsets of comparable sizes, apply MAXSER on one subsample of data, then conduct further estimation and inference using only the factors included in the MAXSER portfolio and the remaining subset of data. The independence between the two data sets guarantees that the distribution of data is not affected by the MAXSER-screening result, therefore ensuring the validity of asymptotic theory to be established in the second step.

We now explain the procedure in detail. For any deterministic 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 $\boldsymbol{\mu}_S$ and $\mathbf{\Sigma}_S$ are the mean and the covariance matrix matrix of $(\mathbf{f}_{it}^T, \mathbf{R}_{it}^T)_{i\in S}^T$. We aim to make inference about $\mathbf{b}_S$. To do so, we compute the plug-in estimator of $\mathbf{b}_S$:

$$\widehat{\mathbf{b}}_S = \widehat{\mathbf{\Sigma}}_S^{-1} \widehat{\boldsymbol{\mu}}_S, \tag{2.9}$$

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

Theorem 2 gives the central limit theorem of $\widehat{\mathbf{b}}_S$.

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

$\mathbf{A}$ *with* $\|\mathbf{A}\| = O(1)$*, we have*

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

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

Proof: See Appendix A.2.

Theorem 2 states that the SDF estimator $\widehat{\mathbf{b}}_S$ enjoys asymptotic normality, which allows us to perform inference on the SDF loadings. These theoretical results lay down the foundation of our methodology with statistical guarantee. The detailed procedure will be described in the next section with data.

3 Empirical tests

3.1 Data

We build our work on the fruitful research on return predictability and anomalies. Specifically, we use the characteristic-based portfolio data from Hou et al. (2020). The dataset includes monthly returns of 188 anomaly portfolios from January 1975 to December 2021. Hou et al. (2020) categorize these anomalies into 6 groups: frictions, intangibles, investment, momentum, profitability, and value-growth. We also include the Fama-French six factors (Fama and French (2018)), i.e., the market portfolio, SMB, HML, CMA, RMW and MOM portfolios. Monthly data of these six factors are obtained from French's data library.[4] Totally there are 194 test portfolios, each with 564 monthly returns.

3.2 SDF learning: Our approach

First, note that there exists a factor structure among testing portfolios, as is clearly shown in Figure 1. The market portfolio is correlated with all testing portfolios with a median magnitude of 0.15. For this reason, in implementing MAXSER, we choose the market portfolio as the base factor for the remaining factors. 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 due to that we intend to adopt the data splitting technique for further factor selection. The out-of-sample period is between 2000 and 2021.

[4] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

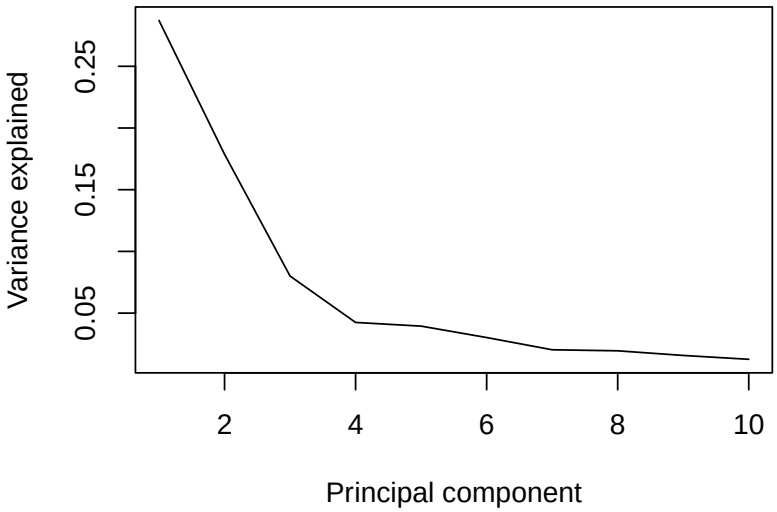


Figure 1: **Scree plot of PCA over 194 test portfolios.**

We further note that these factors are highly multi-collinear. To alleviate the multi-collinearity, in each time period when we estimate the SDF, we pre-screen the factors and exclude the ones that have high multi-collinearity with other factors. Specifically, we sequentially remove factors, one at a time, that has a variance inflation factor (VIF) higher than 10 and is the highest among all remaining factors. After the pre-screening, the VIFs of all factors in the training set are lower than 10.

We use MAXSER for a first step screening. Specifically, we split the sample into two subsamples according to even/odd years. Then we apply MAXSER to one subsample. We will then only focus on the subsect 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 fix this issue, we will conduct further factor selection using the inference theory that we build 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 (2.6), and the sub-models are nested and becoming smaller as λ increases (Efron et al. (2004)). Let $M_0 = \{i_{(1)}, i_{(2)}, \ldots, i_{(|M_0|)}\}$ denote the set of factors included in MAXSER, where the variables $(i_{(j)})_{j \leq |M_0|}$ are arranged in decreasing order of inclusion frequency in the collection of sub-models. For the last factor $i_{(|M_0|)}$, we test whether it is statistically significant using the

withheld subsample and Theorem 2. Specifically, we compute the p-value

$$2\left(1-\Phi\left(\frac{|\widehat{b}^{(M_0)}_{i_{(|M_0|)}}|}{\widehat{sd}(\widehat{b}^{(M_0)}_{i_{(|M_0|)}})}\right)\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $\widehat{b}^{(M_0)}_{i_{(|M_0|)}}$ is the plug-in estimator of the loading of $i_{(|M_0|)}$th variable from (2.9) for $i_{(|M_0|)} \in M_0$, and $\widehat{sd}(\widehat{b}^{(M_0)}_{i_{(|M_0|)}})$ is the standard deviation estimator from Theorem 2. If the p-value is greater than a pre-specified significance level τ_s, we repeat the procedure on the smaller sub-model $M_1 = \{i_{(1)}, i_{(2)}, \ldots, i_{(|M_0|-1)}\}$. We continue this process until the p-value of the last variable in the sub-model is smaller than τ_s. By doing so, we ensure that the last factor in the final sub-model is statistically significant at the chosen level τ_s. The argument for only testing the last variable rather than all the variables is because of the nesting nature of the LASSO solution path, which implies that 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 in the last sub-model, with the risk constraint of the market risk during the training period. We denote this portfolio as MAXSER-S(τ_s). For robustness check, we consider $\tau_s = 0.01, 0.05$ and 0.1. We summarize our proposed MAXSER-S algorithm for SDF learning as follows.

Algorithm

Input:	$\tilde{\mathbf{R}}_{t\leq n} = (\mathbf{R}_t, \mathbf{F}_t)_{t\leq n}$, σ, τ,τ_{VIF}
Output:	$\widehat{\boldsymbol{\omega}}$, $\widehat{\mathbf{b}}$
Step I.	Split sample into two subsets of comparable sizes $T_1 \cup T_2 = \{1, ..., n\}$.
Step II.	Get weight estimator $\hat{\boldsymbol{w}}^{MAXSER}$ via (2.7) based on $\tilde{\mathbf{R}}_{M',t\in T_1}$, and get $M_0 = \{i_{(j)}; \hat{w}^{MAXSER}_{i_{(j)}} \neq 0, 1 \leq i \in N + K\}$ with $i_{(1)}, ..., i_{(\lvert M_0\rvert)}$ arranged in decreasing order of inclusion of the solution path.
Step III.	Initialize $M = M_0$. While $\lvert M\rvert > 0$, do: Compute $p_{i_{(\lvert M\rvert)}}$ for variable $i_{(\lvert M\rvert)}$ based on $\tilde{\mathbf{R}}_{M,t\in T_2}$ and Theorem 2. If $p_{i_{(\lvert M\rvert)}} < \tau$, then go to Step IV. Else: $M \leftarrow \{i_{(1)}, ..., i_{(\lvert M\rvert-1)}\}$.
Step IV.	Compute plugin SDF weight estimator $\widehat{\mathbf{b}}$ via (2.9) and estimator of the optimal portfolio $\widehat{\boldsymbol{\omega}} = (\sigma/\sqrt{\widehat{\boldsymbol{\mu}}_M^T \widehat{\boldsymbol{\Sigma}}_M^{-1} \widehat{\boldsymbol{\mu}}_M})\ \widehat{\mathbf{b}}$ based on $\tilde{\mathbf{R}}_{M,t\leq n}$.

3.3 SDF learning: Benchmark models

For comparison, we consider the following approaches proposed in the literature to construct SDF. We use them as benchmark models.

- CAPM: the market portfolio.
- FF3/FF5/FF5+MOM (FF6): plug-in optimal portfolio using Mkt-Rf, SMB and HML factors from the Fama-French three-factor model (Fama and French (1992)), the Fama-French five-factor model with CMA and RMW portfolios (Fama and French (2015)), and further including MOM which we call the Fama-French six-factor model (Fama and French (2018)).
- Q4/Q5: plug-in optimal portfolio using Mkt-Rf, R_ME, R_IA and R_ROE factors from the Q4 factor model (Hou et al. (2015)) and additional R_EG factor for the Q5 factor model (Hou et al. (2021)).[5]
- BS6: plug-in optimal portfolio using Mkt-Rf, SMB, IA, ROE, MOM, HML(m) factors from Barillas and Shanken (2018).[6]

[5]Monthly data of the q-factors are from https://global-q.org/factors.html.
[6]HML(m) data are from https://pages.stern.nyu.edu/~afrazzin/data_library.htm.

- SY4: plug-in optimal portfolio using Mkt-Rf, SMB, PERF and MGMT factors from Stambaugh and Yuan (2017).[7]

- DHS3: plug-in optimal portfolio using Mkt-Rf, PEAD and FIN factors from Daniel et al. (2020).[8]

- KNS: we include the shrinkage SDF estimator in Kozak et al. (2020). The estimator is given by:

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

  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 et al. (2020), the tuning parameters are chosen by cross-validation with the criterion of maximizing the out-of-sample (oos) R^2:

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

  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.

3.4 Performance 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, Sharp ratio, maximum drawdown, together with different monthly distribution percentiles. In addition, we use the method in Ledoit and Wolf (2008)) for testing the difference in Sharpe ratios:

$$H_0 : SR_{MAXSER-S(0.05)} \leq SR_0 \qquad \text{vs.} \qquad H_1 : SR_{MAXSER-S(0.05)} > SR_0, \tag{3.1}$$

where $SR_{MAXSER-S(0.05)}$ is the Sharpe ratio of MAXSER-S(0.05) and SR_0 denotes the Sharpe ratio of another approach. The cumulative returns of the mean-variance optimal portfolios are plotted in Figure 2.

[7]Monthly data of SY4 factors are from https://finance.wharton.upenn.edu/~stambaug/.
[8]Monthly data of DHS factors are from https://sites.google.com/view/linsunhome.

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 and other benchmark 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 5%, 1% and 0.1%, 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 2021. We compare the original MAXSER, MAXSER with various post screening significance levels (0.01, 0.05, 0.1), the shrinkage SDF portfolio (KNS) by Kozak et al. (2020), and the SDF portfolio constructed from various factor models.

Model	Mean	SD	SR	SR-pvalue	MDD	NUM	P1	P25	P50	P75	P99
Panel A: MAXSER-based SDF											
MAXSER	2.26	5.08	0.45	0.125	15.53	47	-10.01	-0.62	2.15	4.98	19.22
MAXSER-S(0.01)	2.43	5.40	0.45	0.010*	30.46	18	-10.94	-0.63	2.58	5.80	16.99
MAXSER-S(0.05)	2.52	5.04	0.50	-	17.26	20	-10.04	-0.09	2.57	5.17	16.99
MAXSER-S(0.1)	2.44	5.01	0.49	0.029*	17.12	21	-10.19	-0.13	2.53	4.93	16.39
Panel B: SDF from other benchmark models											
KNS	1.63	4.80	0.34	0.001**	20.37	28	-9.86	-0.96	1.65	4.20	13.50
CAPM	0.61	4.51	0.13	0.000***	49.39	1	-10.49	-1.96	1.19	3.23	10.61
FF3	0.88	5.85	0.15	0.000***	66.38	3	-19.78	-1.65	1.11	4.17	13.86
FF5	1.96	6.60	0.30	0.002**	34.73	5	-11.52	-1.65	1.73	5.00	21.31
FF6	1.85	6.47	0.29	0.000***	37.16	6	-12.55	-1.31	1.60	4.39	18.76
Q4	1.55	5.63	0.28	0.000***	32.08	4	-9.92	-1.44	1.55	4.02	16.93
Q5	2.01	5.30	0.38	0.030*	24.41	5	-12.00	-0.77	1.68	4.77	16.87
BS6	1.67	6.38	0.26	0.000***	38.62	6	-13.35	-1.22	1.81	4.10	23.12
SY4	1.83	6.11	0.30	0.000***	27.36	4	-13.78	-1.11	1.88	4.37	20.05
DHS3	1.65	5.14	0.32	0.002**	37.00	3	-11.03	-1.42	1.40	4.45	13.93

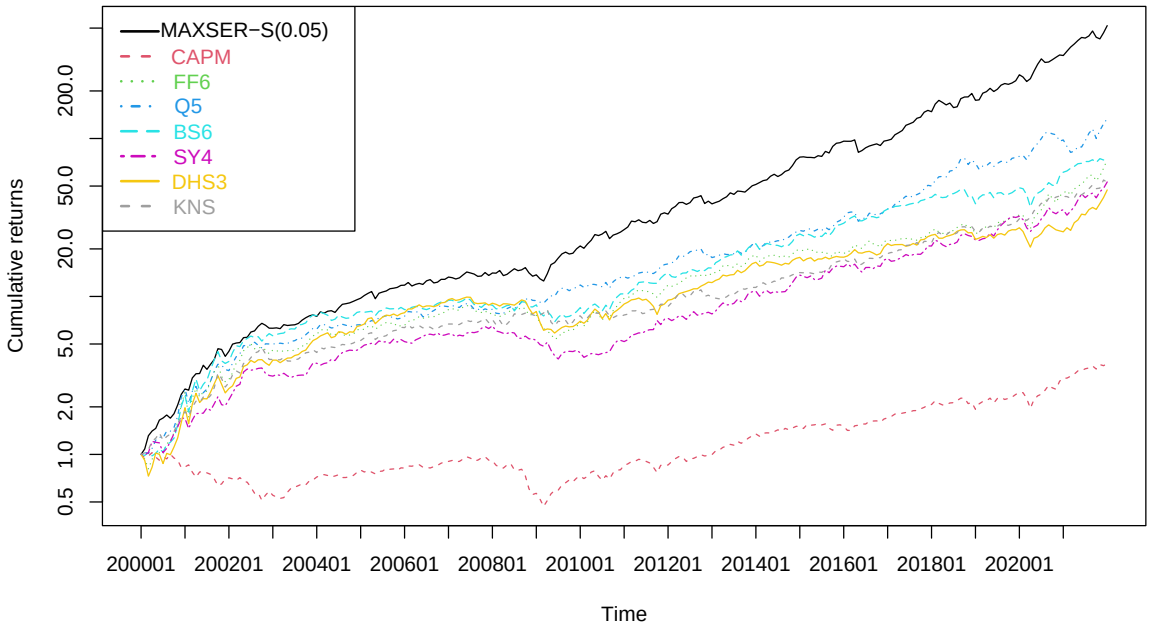


Figure 2: **Cumulative dollar returns of the SDF portfolios.** This figure plots the cumulative returns of various SDF portfolios from 2000 to 2021, starting with \$1. We include the SDF portfolios constructed from the MAXSER-S(0.05), CAPM, FF6, Q5, BS6, SY4, DHS3, and KNS. The portfolio risk is set to be the market risk.

We see from Panel A of Table 1 that the original MAXSER portfolio achieves a monthly Sharpe ratio of 0.45. Among the MAXSER portfolios with further selection, we see that as the significance level decreases from 0.10 to 0.05, the portfolios' monthly Sharpe ratio increases from 0.49 to 0.50, which is the highest among all approaches considered. When the significance level reaches 0.01, the performance deteriorates. The number of factors included for different significance levels are plotted in Figure 3. We see that, compared with the original MAXSER portfolio, MAXSER-S(0.05) reduces the number of factors usedwhile achieving a higher Sharpe ratio. The results demonstrate the advantage of utilizing the statistical inference theory for the SDF loadings established in Theorem 2.

The statistical tests (3.1) show that the differences in Sharpe ratios from various MAXSER-based portfolios are mostly statistically insignificant. For example, the difference between MAXSER and MAXSER-S(0.05) is not statistically significant at 0.1 level. This shows the robustness of our methodology. Balancing between the performance and the parsimony of the SDF, below we will focus on MAXSER-S(0.05).

Panel B of Table 1 shows that MAXSER-S(0.05) significantly outperforms all the benchmark methods. Among all the benchmark portfolios, the Q5 portfolio has the

highest monthly Sharpe Ratio of 0.38, which is still statistically significantly lower than that of MAXSER-S(0.05).

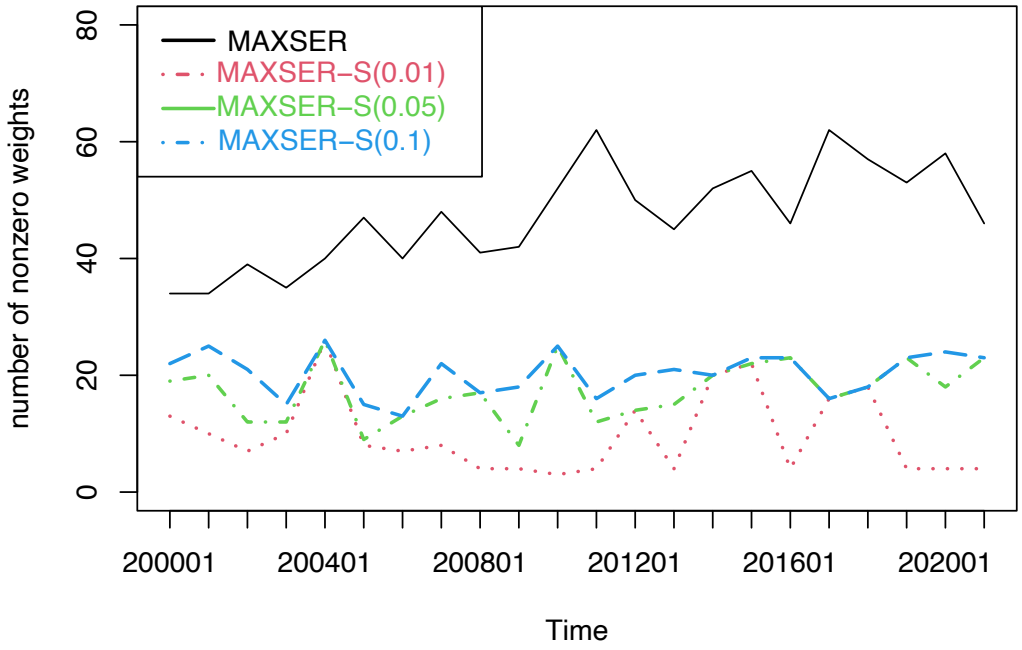


Figure 3: **Number of non-zero weights for original MAXSER portfolio and MAXSER-S(τ_s) portfolios for $\tau_s = 0.01, 0.05$ and 0.1.** This figure plots the number of non-zero weights for the portfolios estimated each year from 2000 to 2021.

Next, we perform asset pricing tests in Table 2 to examine the pricing power of various factor models. Specifically, we regress the 188 anomaly portfolios against various factor models to examine whether alpha is significant:

$$R_{it} = \alpha_i + \sum_{k=1}^{K} \beta_{ik} R_{f;kt} + \varepsilon_{it}, \tag{3.2}$$

where R_{it} is the return of the ith anomaly portfolio and $R_{f;kt}$ is the returns in the factor model. Newey-West robust standard deviation estimator (Newey and West (1987)) is used in calculating the t-statistics. Table 2 presents the number of significant alphas. We see that the proposed MAXSER-S(0.05) portfolio has the lowest number of significant alphas. Only 5 anomalies have significant alphas under MAXSER-S(0.05). On the other hand, all other benchmark portfolios have more rejections. The portfolio that has the second lowest number of rejections is Q5, which have 12 rejections, while the KNS portfolio has 13

Table 2: **The number of significant alphas from testing 188 anomalies against various SDFs.** We examine the explanatory power of various SDFs over 188 anomalies. This table summarizes the total number of significant alphas and the number of significant alphas in each characteristic group. The number in parenthesis is the total number of portfolios in each characteristic group. The number of significant alphas under the threshold $t > 1.96$ or $t > 3$ is reported. The evaluation period is between 2000 and 2021.

	All (188)		frictions (10)		intangibles (30)		investment (29)	
threshold of t	1.96	3	1.96	3	1.96	3	1.96	3
MAXSER-S(0.05)	5	0	0	0	0	0	0	0
KNS	13	0	0	0	5	0	1	0
CAPM	65	37	4	0	10	8	7	3
FF3	75	51	4	3	10	9	7	3
FF5	41	12	0	0	13	4	4	1
FF6	37	13	0	0	12	4	3	1
Q4	36	9	0	0	13	2	5	0
Q5	12	1	1	0	3	1	0	0
BS6	44	8	0	0	13	4	6	1
SY4	21	2	0	0	4	1	0	0
DHS3	30	4	0	0	14	1	2	0

	momentum (41)		profitability (46)		value-growth (32)	
threshold of t	1.96	3	1.96	3	1.96	3
MAXSER-S(0.05)	0	0	5	0	0	0
KNS	4	0	2	0	1	0
CAPM	9	2	34	24	1	0
FF3	14	6	37	30	3	0
FF5	6	1	17	6	1	0
FF6	5	1	17	7	0	0
Q4	1	0	16	7	1	0
Q5	2	0	4	0	2	0
BS6	6	0	17	3	2	0
SY4	6	0	5	0	6	1
DHS3	0	0	13	3	1	0

rejections.

Furthermore, we examine the explanatory power of the benchmark models over the SDF portfolio constructed from MAXSER-S(0.05). Specifically, we regress the out-of-sample return of the SDF portfolio from MAXSER-S(0.05) against the benchmark models:

$$R_{SDF-MAXSER;t} = \alpha + \sum_{k=1}^{K} \beta_{ik} R_{f;kt} + \varepsilon_{it}, \tag{3.3}$$

where $R_{SDF-MAXSER;t}$ is the return of MAXSER-S(0.05). Table 3 presents the alpha and its t-statistic from the regressions (3.3). Newey-West robust standard deviation estimator (Newey and West (1987)) is used in calculating the t-statistics.

Table 3: **Using benchmark models to explain MAXSER-S(0.05) SDF.** This table reports the regression results of the SDF portfolio from MAXSER-S(0.05) against benchmark models, using monthly returns from 2000 to 2021. Alpha in percentage and its corresponding t-statistic are reported.

	Alpha (%)	t-statistic
CAPM	2.42	5.83
FF3	2.33	7.55
FF5	1.85	5.46
FF6	1.81	5.46
Q4	1.81	5.95
Q5	1.29	4.63
BS6	1.85	5.80
DHS3	1.55	5.20
SY4	1.85	5.80
KNS	1.37	4.05

We see from Table 3 that all benchmark models fail to explain our estimated MAXSER-S(0.05) portfolio; the alpha of MAXSER-S(0.05) is both economically large and statistically significant under all benchmark models. Table 3 shows that the monthly alphas are all greater than 1% with a t-statistic higher than 4 across all models. These results suggest that the benchmark models can not capture the returns of our MAXSER-S(0.05) portfolio.

Finally, we switch the roles between benchmark models and our estimated SDF and examine whether MAXSER-S(0.05) is able to explain the commonly used pricing factors. Specifically, we regress various pricing factors against MAXSER-S(0.05):

$$xrR_{f;kt} = \alpha_k + \beta_k R_{MAXSER-SDF;t} + \varepsilon_{kt}, \tag{3.4}$$

where $R_{f;kt}$ is the factor return and $R_{MAXSER-SDF;t}$ is the return of MAXSER-S(0.05). Table 4 reports the alpha and its t-statistic from regressions (3.4). Newey-West estimator is used in computing the t-statistics.

Table 4: **Using MAXSER-S(0.05) SDF to explain commonly used pricing factors.** This table reports the regression results of commonly used pricing factors against MAXSER-S(0.05) SDF, using monthly returns from 2000 to 2021. Alpha in percentage and its corresponding t-statistic are reported.

Factor	Alpha (%)	t-statistic
Mkt-RF	0.26	0.68
SMB	0.01	0.05
HML	0.21	0.86
RMW	0.19	0.99
CMA	0.10	0.72
MOM	0.16	0.50
R_ME	0.06	0.28
R_IA	0.01	0.07
R_ROE	0.07	0.29
R_EG	0.16	0.50
HmLm	0.13	0.44
MGMT	0.07	0.35
PERF	0.25	0.49
PEAD	0.19	1.14
FIN	0.07	0.25
KNS	0.01	0.06

We see from Table 4 that all alphas are not significant at the t-statistic threshold level of 1.96. These results suggest that our SDF well captures those commonly used pricing factors.

4 Impacts of transaction costs

One naturally wonders if the above results still hold when transaction costs are taken into consideration. In this section, we evaluate the impact of transaction costs. To this end, we will add transaction cost to our MAXSER-SDF only, while leaving all benchmark models cost free. First, we compute the monthly turnover $\sum_{t=1}^{T}\sum_i |\omega_{it+} - \omega_{it+1}|/T$, where ω_{it+} is the weight of stock i at the end of month t, and ω_{it+1} is the weight of stock i at the beginning of month $t+1$. We find that the MAXSER-SDF has a monthly turnover of 45.3%. Then with the absolute monthly weight exposure defined as $\sum_{t=1}^{T}\sum_i |\omega_{it}|/T$, we find that MAXSER-S(0.05) has an absolute monthly weight exposure of 5.5. Therefore our SDF portfolio turnover per dollar exposure is 8%. Kan et al. (2022) suggests using 20 bps per dollar of transaction when evaluating the post-transaction cost performance of investment in anomaly portfolios. Following Kan et al. (2022), we check the performance

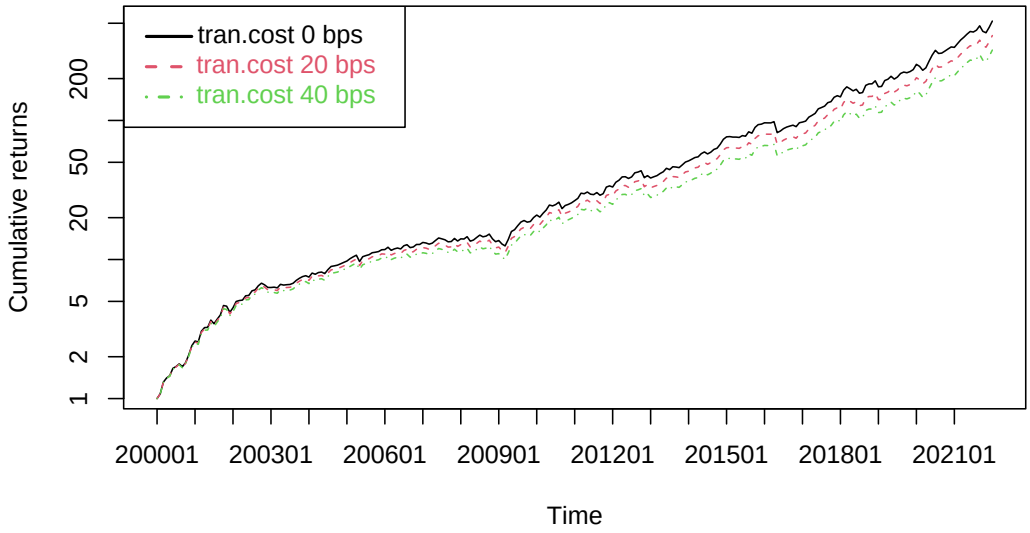


Figure 4: **Cumulative dollar returns of MAXSER-S(0.05) after adjusting for transaction costs.** The initial investment is \$1 in 2000. Transaction costs are assumed to be 20 bps or 40 bps between 2000 and 2021. The portfolio returns are computed every month.

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. The performance of our SDF estimator after adjusting for the transaction costs is summarized in Table 5. The cumulative returns of the SDF portfolio after transaction costs are plotted in Figure 4.

Table 5: **Performance of MAXSER-S(0.05) 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 2021. 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	2.43	5.03	0.48	18.05
40 bps	2.25	5.01	0.45	18.74

We see from Table 5 that, the portfolio maintains a high Sharpe ratio after adjusting

for the transaction costs. For example, the Sharpe ratio is 0.48 after removing 20 bps transaction cost per dollar, and is 0.45 when the cost is 40 bps/dollar. Similarly, Figure 4 shows that for \$1 initial investment in 2000, at the end of 2021, the cumulative return is \$518.27 when there is no transaction costs, while it is \$407.80 (\$320.64) after adjusting for a transaction cost of 20 bps (40 bps) per dollar.

Next, we repeat Table 3, adjusting for transaction costs. That is, we regress the MAXSER-S(0.05) SDF portfolio against various factor models (regressions (3.3)) in Table 6. For the variables from the factor models, we use the original data without removing the transaction cost. We see that again, our MAXSER-S(0.05) SDF portfolio has a statistically significant alpha against benchmark factor models.

Table 6: **Using benchmark models to explain MAXSER-S(0.05) SDF, adjusting for transaction costs.** This table presents the regression results of MAXSER-S(0.05) against various benchmark models, using the monthly returns from 2000 to 2021. The transaction costs of MAXSER-S(0.05) are assumed to be 20 bps or 40 bps per dollar of transaction. The pricing factors from various factor 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	2.32	5.63	2.16	5.39
FF3	2.24	7.25	2.08	6.82
FF5	1.76	5.18	1.59	4.64
FF6	1.71	5.21	1.55	4.72
Q4	1.72	5.61	1.56	5.14
Q5	1.20	4.27	1.04	3.68
BS6	1.76	5.52	1.59	4.99
SY4	1.29	3.63	1.12	3.14
DHS3	1.73	4.97	1.57	4.57

Last, we repeat Table 4 but take into account of transaction costs. We regress various pricing factors against our SDF after adjusting for transaction costs (regressions (3.4)). For the portfolios from the benchmark models, we use their original returns without removing the transaction cost. The results are summarized in Table 7. We see that pricing factors from benchmark models still do not have statistically significant alphas.

Table 7: **Using MAXSER-S(0.05) SDF to explain commonly used pricing factors, adjusting for transaction costs.** This table reports the regression results of various pricing factors against MAXSER-S(0.05) after adjusting for the transaction costs, using the monthly returns from 2000 to 2021. The transaction cost of MAXSER-S(0.05) 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	0.27	0.72	0.32	0.87
SMB	0.02	0.12	0.05	0.27
HML	0.19	0.79	0.15	0.64
RMW	0.20	1.07	0.20	1.14
CMA	0.11	0.81	0.13	0.96
MOM	0.15	0.47	0.15	0.48
R_ME	0.04	0.19	0.00	0.02
R_IA	0.00	0.01	0.02	0.13
R_ROE	0.08	0.34	0.08	0.35
R_EG	0.23	1.21	0.24	1.30
HmLm	0.14	0.49	0.18	0.63
MGMT	0.09	0.45	0.13	0.68
PERF	0.26	0.54	0.28	0.61
PEAD	0.19	1.46	0.20	1.55
FIN	0.09	0.31	0.11	0.42

5 Investigating the MAXSER-S(0.05) SDF

Previous results demonstrate the superior performance of MAXSER-S(0.05) SDF. One might wonder which factors are important. Note that we re-estimate the SDF every year between 2000 and 2021, resulting in 22 different trained portfolios. Therefore, the selected characteristics vary over time. We collect the portfolios that have ever been included in the estimated SDF portfolio for at least 30% of the times, i.e., at least 7 out of 22 years. Figure 5 gives the bar chart of the frequencies for 26 selected portfolios. Besides the market portfolio (Mkt-RF), the most important factors are usually related to momentum and earnings, including the quarterly earnings-to-price (epq_1), cumulative

abnormal stock return (abr_1), the average returns across 120 months (r10a), the earnings predictability (eprd) and the customer momentum (cm_1) , which are included for over 85% of the times.

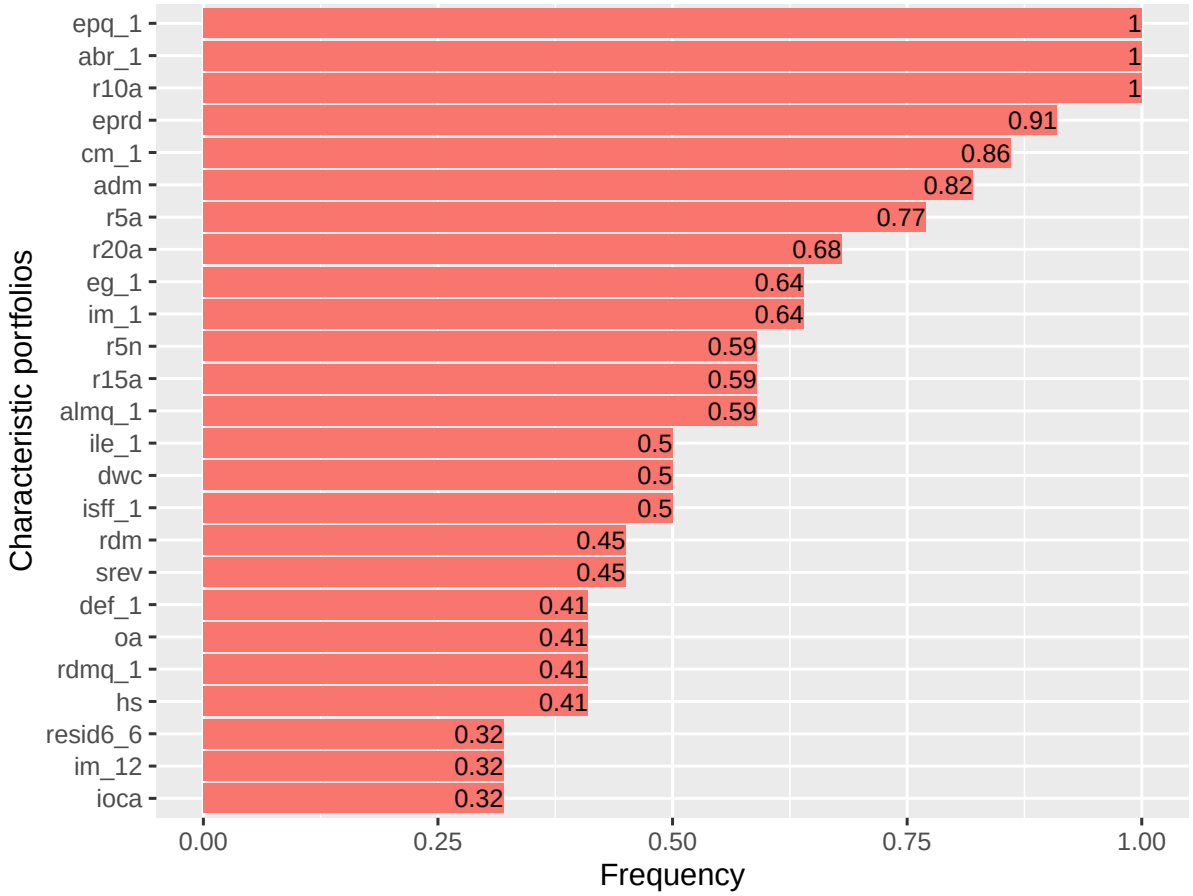


Figure 5: **Frequencies of portfolios included in the MAXSER-S(0.05) SDF.** This figure reports the inclusion frequency of portfolios in the MAXSER-S(0.05) SDF from all training periods with an inclusion frequency higher than 30%. The market portfolio is always included by default, hence it is omitted in the graph.

In Table 8, we further provide some summary statistics of the most frequently selected factors in the MAXSER-S(0.05) SDF. We report the t-statistics of alphas of these factors when regressing against various factor models.

Table 8: **Summary statistics of the most frequently selected portfolios in the MAXSER-S(0.05) SDF.** This table reports the summary statistics of the most frequently selected factors (other than the market portfolio) in constructing the MAXSER-S(0.05) SDF, including the frequency, group name, and the t-statistics of alphas when regressing this factor against some benchmark factor models.

			t-statistic of alphas from various factor models								
Variable	Frequency	Group	CAPM	FF3	FF5	FF6	Q4	Q5	SY4	DHS3	BS6
epq_1	1.00	value-growth	1.42	1.82	0.64	0.76	0.34	1.21	2.29	1.30	0.17
abr_1	1.00	momentum	2.21	2.28	2.21	2.20	1.75	1.94	0.08	0.13	1.98
r10a	1.00	intangibles	3.46	3.43	3.36	3.68	3.90	3.77	3.47	3.20	4.14
eprd	0.91	intangibles	3.39	4.84	4.45	3.65	3.15	2.83	1.95	2.76	3.68
cm_1	0.86	momentum	1.75	2.02	2.28	2.16	1.80	1.47	1.22	1.45	2.26
adm	0.86	intangibles	0.75	0.33	1.02	0.78	0.15	0.15	0.37	0.40	0.61
r5a	0.77	intangibles	1.14	1.26	1.19	1.05	1.35	1.50	1.29	0.81	1.30
r20a	0.68	intangibles	2.13	2.27	2.17	1.98	2.21	1.94	1.67	1.85	2.12
eg_1	0.64	profitability	4.25	4.85	2.96	2.74	3.08	0.08	1.75	2.79	3.29
im_1	0.64	momentum	2.76	2.59	1.68	1.34	1.38	1.08	0.83	0.80	1.79
r5n	0.59	intangibles	0.01	0.74	1.55	1.56	0.82	0.04	0.03	0.74	1.53
r15a	0.59	intangibles	1.05	0.98	1.08	1.14	0.76	0.65	0.64	0.51	0.78
almq_1	0.59	intangibles	1.05	0.77	0.35	0.52	0.65	1.56	0.78	1.18	0.21
ile_1	0.50	momentum	1.11	1.25	0.52	0.26	0.16	0.45	0.20	0.09	0.05
dwc	0.50	investment	0.27	0.56	0.22	0.23	0.55	0.01	0.44	0.17	0.59
isff_1	0.50	frictions	1.07	0.98	0.92	0.84	1.06	0.94	0.15	0.27	1.15
rdm	0.45	intangibles	1.63	1.36	2.04	2.18	2.32	1.05	0.10	2.61	2.60
srev	0.45	frictions	0.40	0.49	0.07	0.16	0.13	0.75	0.75	1.07	0.32
def_1	0.41	momentum	1.37	1.57	1.02	0.88	0.55	1.00	0.04	0.51	0.75
oa	0.41	investment	0.02	0.14	0.80	0.83	0.97	0.27	0.99	1.00	1.16
rdmq_1	0.41	intangibles	1.72	1.76	2.14	2.56	2.50	1.77	2.16	2.25	2.24
hs	0.41	intangibles	0.76	1.52	0.12	0.19	0.50	1.00	1.24	0.26	0.38
resid6_6	0.32	momentum	0.99	0.95	0.44	0.16	0.35	0.84	0.16	0.11	0.46
im_12	0.32	momentum	1.16	1.58	0.94	0.66	0.24	0.29	0.83	0.50	0.71
ioca	0.32	intangibles	3.35	3.64	1.88	1.72	2.11	1.21	0.97	2.86	1.97

We see from Table 8 that many frequently selected factors do not have statistically significant alphas. Take the CAPM as an example, only 6 out of 26 factors have a significant alpha with t-statistic > 1.96. These remaining factors are, however, instrumental in our estimated MAXSER-S(0.05) as they help to reduce the portfolio risk and hence improve the overall performance of the SDF. The results demonstrate the advantage of our approach in constructing the SDF. Instead of merely selecting factors which have significant alphas under various models (i.e., significant anomalies), our approach accounts for the cross-sectional dependence in returns and is able to select important factors which contribute to the Sharpe ratio of the SDF.

Finally, to provide more economic interpretation of the selected characteristics, we summarize them at the group-level, using the group definitions of Hou et al. (2020). Figure 6 reports the average number of selected portfolios in different groups.

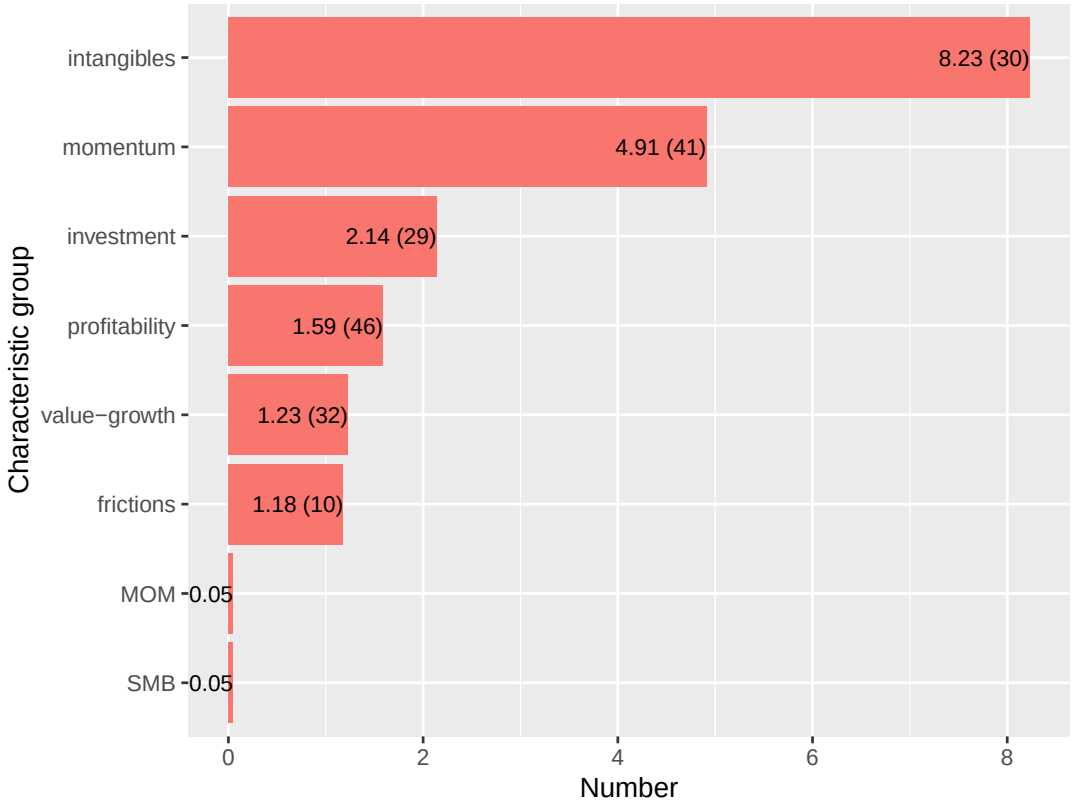


Figure 6: **Average number of portfolios from each characteristic group included in the MAXSER-S(0.05) SDF.** This barchart plots the average number of portfolios from each characteristic group included in the MAXSER-S(0.05) SDF. The number in parenthesis is the total number of factors in a characteristic group. The market portfolio is always included hence omitted in the graph.

We see from Figure 6 that the intangible group contributes the most in the MAXSER-S(0.05) SDF. On average, 8 out of 30 portfolios from the intangible group are selected. The second important group is momentum. On average, 5 out of 41 momentum portfolios are included in the SDF. The investment group, value-growth group and the profitability group each has on average 2 significant factors in the SDF, and the frictions group has one factor on average in the SDF. Examining the Fama-French factors, we see that (1) market portfolio is always selected; (2) SMB and MOM are included only about 5% of the times; (3) HML, RMW and CMA are not included in the MAXSER-S(0.05) SDF. That is, the Fama-French factors, except the market portfolio, are unimportant given other factors.

Figure 7 illustrates the evolution of the number of selected portfolios in each group over time. From Figure 7, we see that SMB and MOM no longer show up after Year 2001, and the factors from the friction group only show once (Year 2020) since Year 2015, while the other groups persistently contribute to the SDF composition. Overall, Figures 6 and 7 show that only a few groups are important. But within a group, several portfolios may be necessary and selected for the SDF, suggesting that these portfolios may not be constructed in an optimal way and complement each other.

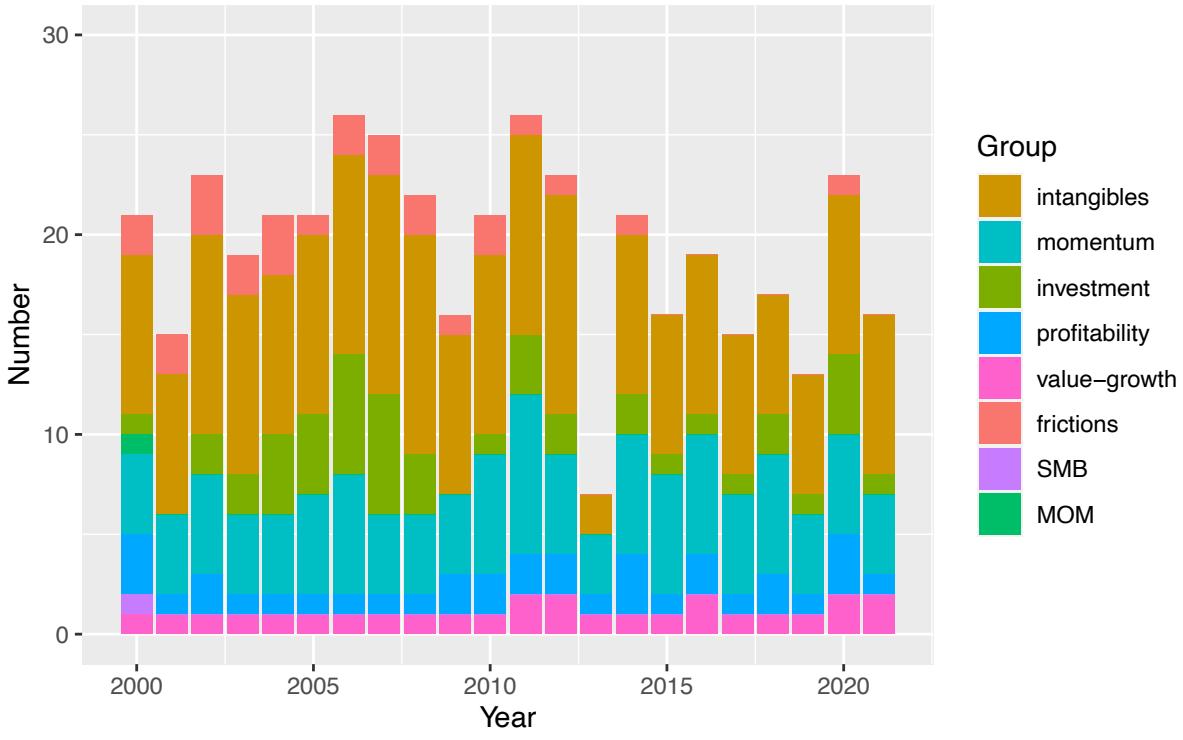


Figure 7: **Evolution path of the number of portfolios from each characteristic group included in the MAXSER-S(0.05) SDF.** This barchart plots the number of portfolios from each characteristic group which are included in the MAXSER-S(0.05) SDF in each year. The market portfolio is omitted.

Figures 5–7 show that the selected portfolios vary over time. One might concern why the estimated SDF is time varying. This might be due to several reasons. First, as we use rolling-window estimation, it naturally generates a time-varying SDF portfolio. Second, the U.S. economy might experience shocks to macro variables or even some structural changes over time (e.g., changes in industry structure, fiscal and monetary environment etc.), leading to a time-varying SDF. Third, the characteristic portfolios might contain some behavioral traits which could be time varying, e.g., market sentiment or extrapolative beliefs (see, e.g., Baker and Wurgler, 2006; Cassella and Gulen, 2018). Last, a time-varying SDF could be driven by the variations in economic primitives, as discussed in conditional asset pricing models (see, e.g., Nagel and Singleton, 2011; Kelly et al., 2019b).

The significance of the intangible group reflects its rising importance in the U.S. economy. More and more corporate investment has moved from physical capital to intangible capital over the last few decades (Papanikolaou, 2011; Buera and Kaboski, 2012; Eisfeldt and Papanikolaou, 2013, 2014; Crouzet and Eberly, 2023; Crouzet et al., 2022; Corrado et al., 2022; Belo et al., 2023).[9] In U.S., intangible capital grew quickly since middle 1980s. Intangible capital has a similar size as tangible capital, and firms invest more on R&D than capital expenditures since 2000s (Crouzet et al., 2022). Intangible capital impacts asset returns as it affects productivity, human capital, entrepreneurship, competitiveness, and even consumer behavior. However, as most intangible capital is off-balance sheet, balance sheet-based characteristics (i.e., factors used in the prevailing models) often fail to capture intangibles. Eisfeldt et al. (2020) and Gulen et al. (2024) show that accounting for intangibles can greatly improve the performances of the Fama-French three- and five-factor models and the q-model, especially during the last decades, which is consistent with our findings.

Figure 5 also shows that momentum is the second important source of the SDF. Earnings momentum such as cumulative abnormal stock return (abr_1) and price-based momentum such as the average returns across 120 months (r10a) and the customer momentum (cm_1) are included in the SDF for most of times, but not the momentum factor (MOM) as shown in Figure 6. This is in line with the recent findings that price momentum is largely driven by fundamentals (Hou et al., 2015, 2021).

[9]Intangible capitals are productive capitals without physical presence, including organization capital, firm-specific human capital, patents, brands, software and databases, and franchise.

6 Conclusions

We develop a statistical framework to construct the high-dimensional SDF. We propose to screen for potentially useful factors using MAXSER. We show the sure screening property of MAXSER, which guarantees that the screening step does not miss important factors. Moreover, to avoid the over-selection issue of Lasso, we develop a statistical inference theory, which allows us to conduct further factor selection. These two steps lead us to a much smaller set of characteristics, and one can simply use the plug-in method to get an unbiased estimator of the SDF. Our empirical tests show that the our SDF estimator delivers a high monthly Sharpe ratio of 0.50 and performs well in asset pricing tests.

References

Anatolyev, S. and Mikusheva, A. (2022). Factor models with many assets: Strong factors, weak factors, and the two-pass procedure. *Journal of Econometrics*, 229(1):103–126.

Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–94.

Ao, M., Li, Y., and Zheng, X. (2019). Approaching mean-variance efficiency for large portfolios. *Review of Financial Studies*, 32(7):2890–2919.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4):1645–1680.

Barillas, F., Kan, R., Robotti, C., and Shanken, J. (2020). Model comparison with Sharpe ratios. *Journal of Financial and Quantitative Analysis*, 55(6):1840–1874.

Barillas, F. and Shanken, J. (2017). Which alpha? *Review of Financial Studies*, 30(4):1316–1338.

Barillas, F. and Shanken, J. (2018). Comparing asset pricing models. *Journal of Finance*, 73(2):715–754.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Belo, F., Li, Y., Salomao, J., and Vitorino, M. A. (2023). The value of intangible capital around the world. Working Paper, INSEAD.

Black, F. and Litterman, R. B. (1991). Asset allocation: Combining investor views with market equilibrium. *Journal of Fixed Income*, 1(2):7–18.

Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *Journal of Finance*, 54(2):655–671.

Bryzgalova, S. (2017). Spurious factors in linear asset-pricing models. Working Paper, London Business School.

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

Bryzgalova, S., Pelger, M., and Zhu, J. (2023b). Forest through the trees: Building cross-sections of asset returns. Working Paper, London Business School.

Buera, F. J. and Kaboski, J. P. (2012). The rise of the service economy. *American Economic Review*, 102(6):2540–2569.

Cassella, S. and Gulen, H. (2018). Extrapolation bias and the predictability of stock returns by price-scaled variables. *Review of Financial Studies*, 31(11):4345–4397.

Chen, A. Y. (2021). The limits of p-hacking: Some thought experiments. *Journal of Finance*, 76(5):2447–2480.

Cochrane, J. H. (2011). Presidential address: Discount rates. *Journal of Finance*, 66(4):1047–1108.

Corrado, C., Haskel, J., Jona-Lasinio, C., and Iommi, M. (2022). Intangible capital and modern economies. *Journal of Economic Perspectives*, 36(3):3–28.

Crouzet, N. and Eberly, J. (2023). Rents and intangible capital: A q+ framework. *Journal of Finance*, 78(4):1873–1916.

Crouzet, N., Eberly, J. C., Eisfeldt, A. L., and Papanikolaou, D. (2022). The economics of intangible capital. *Journal of Economic Perspectives*, 36(3):29–52.

Daniel, K., Hirshleifer, D., and Sun, L. (2020). Short-and long-horizon behavioral factors. *Review of Financial Studies*, 33(4):1673–1736.

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statististics*, 32(1):407–499.

Eisfeldt, A. L., Kim, E., and Papanikolaou, D. (2020). Intangible value. Technical report, NBER Working Paper 28056.

Eisfeldt, A. L. and Papanikolaou, D. (2013). Organization capital and the cross-section of expected returns. *Journal of Finance*, 68(4):1365–1406.

Eisfeldt, A. L. and Papanikolaou, D. (2014). The value and ownership of intangible capital. *American Economic Review*, 104(5):189–194.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465.

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

Fama, E. F. and French, K. R. (2018). Choosing factors. *Journal of Financial Economics*, 128(2):234–252.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*, 39(6):3320.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(4):603–680.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3):1327–1370.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5):2326–2377.

Giglio, S., Liao, Y., and Xiu, D. (2021). Thousands of alpha tests. *Review of Financial Studies*, 34(7):3456–3496.

Giglio, S., Xiu, D., and Zhang, D. (2022). Test assets and weak factors. Working Paper, Yale University.

Gospodinov, N., Kan, R., and Robotti, C. (2017). Spurious inference in reduced-rank asset-pricing models. *Econometrica*, 85(5):1613–1628.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273.

Gu, S., Kelly, B., and Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450.

Gulen, H., Li, D., Peters, R. H., and Zekhnini, M. (2024). Intangible capital in factor models. *Management Science*, forthcoming.

Hansen, L. P. and Jagannathan, R. (1997). Assessing specification errors in stochastic discount factor models. *Journal of Finance*, 52(2):557–590.

Harvey, C. R. and Liu, Y. (2021a). Lucky factors. *Journal of Financial Economics*, 141(2):413–435.

Harvey, C. R. and Liu, Y. (2021b). Uncovering the iceberg from its tip: A model of publication bias and p-hacking. Working Paper, Duke University.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68.

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

Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *Review of Financial Studies*, 28(3):650–705.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133.

Jacobs, H. and Müller, S. (2020). Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics*, 135(1):213–230.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58(4):1651–1683.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *Journal of Finance*, 78(5):2465–2518.

Kan, R., Wang, X., and Zhou, G. (2022). Optimal portfolio choice with estimation risk: No risk-free asset case. *Management Science*, 68(3):2047–2068.

Kan, R. and Zhang, C. (1999). Two-pass tests of asset pricing models with useless factors. *Journal of Finance*, 54(1):203–235.

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

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

Kleibergen, F. (2009). Tests of risk premia in linear factor models. *Journal of Econometrics*, 149(2):149–173.

Kozak, S. and Nagel, S. (2023). When do cross-sectional asset pricing factors span the stochastic discount factor? NBER Working Paper 31275.

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

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859.

Ledoit, O. and Wolf, M. (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Review of Financial Studies*, 30(12):4349–4388.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the Lasso. *Annals of Statistics*, 44(3):907–927.

Leng, C., Lin, Y., and Wahba, G. (2006). A note on the Lasso and related procedures in model selection. *Statistica Sinica*, 16:1273–1284.

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

Lettau, M. and Pelger, M. (2020b). Factors that fit the time series and cross-section of stock returns. *Review of Financial Studies*, 33(5):2274–2325.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32.

Michaud, R. O. (1989). The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42.

Nagel, S. and Singleton, K. J. (2011). Estimation and evaluation of conditional asset pricing models. *Journal of Finance*, 66(3):873–909.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Papanikolaou, D. (2011). Investment shocks and asset prices. *Journal of Political Economy*, 119(4):639–685.

Stambaugh, R. F. and Yuan, Y. (2017). Mispricing factors. *Review of Financial Studies*, 30(4):1270–1315.

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(5):2183–2202.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 76(1):217–242.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Appendix A Proofs

A.1 Proof of Theorem 1

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 by $\mathbf{U}_{S_1}$ the first q_u rows, and $\mathbf{U}_{S_1^c}$ the remaining rows of $\mathbf{U} = (\mathbf{U}_1, ..., \mathbf{U}_T) =: (U_{it})$. Denote by $\widehat{\mathbf{U}}_{S_1}$ the first $|S_1|$ rows, and $\widehat{\mathbf{U}}_{S_1^c}$ the last remaining rows of $\widehat{\mathbf{U}} = (\widehat{\mathbf{U}}_1, ..., \widehat{\mathbf{U}}_T) =: (\widehat{U}_{it})$. Because $\boldsymbol{\omega}_R^*$ are proportional to $\boldsymbol{\omega}_u^*$, respectively, to show (2.8), it is equivalent to show

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

Recall that η is defined in Assumption 5, and $\boldsymbol{\omega}_u^* = \frac{1}{\sqrt{\theta_u}}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\alpha}$. Define $r_{u,c} = 1/\sqrt{\theta_u} + \sqrt{\theta_u}$, and

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

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

$$\begin{aligned}
&\sqrt{T}\left(C_{11}\Big(\widehat{\boldsymbol{\omega}}_c^* - \frac{1}{r_{u,c}}\boldsymbol{\omega}_u^*\Big)_{S_1}\right)_i - (W_1)_i = -\frac{\lambda}{2\sqrt{T}}\text{sign}((\boldsymbol{\omega}_u^*)_i) \text{ for all } i \in S_{signal}, \\
&\left|\Big(\widehat{\boldsymbol{\omega}}_c^* - \frac{1}{r_{u,c}}\boldsymbol{\omega}_u^*\Big)_i\right| < \frac{|(\boldsymbol{\omega}_u^*)_i|}{r_{u,c}} \text{ for all } i \in S_{signal}, \text{ and} \\
&\left\|\sqrt{T}C_{21}\Big(\widehat{\boldsymbol{\omega}}_c^* - \frac{1}{r_{u,c}}\boldsymbol{\omega}_u^*\Big)_{S_1} - W_2\right\|_{\max} < \frac{\lambda}{2\sqrt{T}},
\end{aligned}$$

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

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

where

$$\mathcal{A} = \left\{ \left|(C_{11}^{-1}W_1)_i\right| < \sqrt{T}\left(\frac{|(\boldsymbol{\omega}_u^*)_i|}{r_{u,c}} - \frac{\lambda}{2T}\left|\left(C_{11}^{-1}\text{sign}(\boldsymbol{\omega}_u^*)\right)_i\right| \right) \text{ for all } i = 1, ..., |S_{signal}| \right\},$$

$$\mathcal{B} = \left\{ \|(C_{21}C_{11}^{-1}W_1 - W_2\|_{\max} < \frac{\lambda}{4\sqrt{T}}\eta \right\}, \quad \text{and}$$

$$\mathcal{C} = \left\{ \max_{\tilde{\mathbf{1}} \in V} \|C_{21}C_{11}^{-1}\tilde{\mathbf{1}}\|_{\max} < 1 - \frac{\eta}{2} \right\},$$

where recall that $V = \{(\delta_1, ..., \delta_{|S_1|})^T, \delta_i = \text{sign}((\boldsymbol{\omega}_u^*)_i) \text{ for } i \in S_{signal}, \text{ and } \delta_i \in \{-1, 1\} \text{ for } i \in S_1 - S_{signal}\}$. Define for some $C > 0$, the following event

$$\begin{aligned}\mathcal{G} = &\left\{ \|(W_1^T, W_2^T)\|_{\max} < C\sqrt{\log N} \right\} \\ &\bigcap \left\{ \|C_{11}^{-1} - (\boldsymbol{\Sigma}_{u,11} + \boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T)^{-1}\| < C\sqrt{\frac{q_u \log N}{T}} \right\} \\ &\bigcap \left\{ \|C_{21} - \boldsymbol{\Sigma}_{u,21} - \boldsymbol{\alpha}_2\boldsymbol{\alpha}_1^T\|_{\max} < Cq_u\sqrt{\frac{\log N}{T}} \right\}.\end{aligned}$$

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 4, we have $q_u^2 \log N/T \to 0$. Assumption 5 implies that

$$\|(\boldsymbol{\Sigma}_{u,11} + \boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T)^{-1}\| = O(1).$$

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

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

Therefore,

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

Similarly

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

By Assumption 4 that $\min_{1 \le i \le |S_{signal}|} |\omega_i| \gg \lambda\sqrt{q_u}/T$ 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,

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

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 4 that $\lambda \gg q_u\sqrt{(\log N)T}$, we get

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

Therefore, (A.3) implies that

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

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

$$\begin{aligned}&\max_{\tilde{\mathbf{1}}\in V}\|C_{21}C_{11}^{-1}\tilde{\mathbf{1}}\|_{\max} \\ \leq & \max_{\tilde{\mathbf{1}}\in V}\|(\boldsymbol{\Sigma}_{u,21}+\boldsymbol{\alpha}_2\boldsymbol{\alpha}_1^T)(\boldsymbol{\Sigma}_{u,11}+\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T)^{-1}\tilde{\mathbf{1}}\|_{\max} \\ & + \max_{\tilde{\mathbf{1}}\in V}\|(C_{21}-\boldsymbol{\Sigma}_{u,21}-\boldsymbol{\alpha}_2\boldsymbol{\alpha}_1^T)C_{11}^{-1}\tilde{\mathbf{1}}\|_{\max} \\ & + \max_{\tilde{\mathbf{1}}\in V}\|(\boldsymbol{\Sigma}_{u,21}+\boldsymbol{\alpha}_2\boldsymbol{\alpha}_1^T)(C_{11}^{-1}-(\boldsymbol{\Sigma}_{u,11}+\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T)^{-1})\tilde{\mathbf{1}}\|_{\max} \\ =: & I+II+III.\end{aligned}$$

By Assumption 5, $I \leq 1-\eta$. In addition, under event G, by (A.4) and (A.8), we have

$$\begin{aligned}II \leq & \max_{\tilde{\mathbf{1}}\in V}\|C_{21}-\boldsymbol{\Sigma}_{u,21}-\boldsymbol{\alpha}_2\boldsymbol{\alpha}_1^T\|_{\max}\|C_{11}^{-1}\tilde{\mathbf{1}}\|_1 \\ \leq & \max_{\tilde{\mathbf{1}}\in V}\sqrt{q_u}\|C_{21}-\boldsymbol{\Sigma}_{u,21}-\boldsymbol{\alpha}_2\boldsymbol{\alpha}_1^T\|_{\max} \cdot \|C_{11}^{-1}\tilde{\mathbf{1}}\| \\ \leq & q_u\|C_{21}-\boldsymbol{\Sigma}_{u,21}-\boldsymbol{\alpha}_2\boldsymbol{\alpha}_1^T\|_{\max} \cdot \|C_{11}^{-1}\| \\ = & O\left(q_u\sqrt{\frac{\log N}{T}}\right).\end{aligned}$$

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

$$\begin{aligned} III \leq & q_u \|\boldsymbol{\Sigma}_{u,21} + \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_1^T\|_{\max} \cdot \|C_{11}^{-1} - (\boldsymbol{\Sigma}_{u,11} + \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^T)^{-1}\| \\ = & O\left(q_u^{3/2} \sqrt{\frac{\log N}{T}}\right). \end{aligned}$$

Under Assumptions 1 and 4, $q_u^{3/2}\sqrt{(\log N)/T} = o(1)$, 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). Recall that $r_{u,c} = 1/\sqrt{\theta_u} + \sqrt{\theta_u}$. Define $\overline{\widehat{\mathbf{U}}} = \widehat{\mathbf{U}}\mathbf{1}/T$, $\overline{\mathbf{U}} = \mathbf{U}\mathbf{1}/T$, and $W = (W_1^T, W_2^T)^T$. We have

$$\begin{aligned} W = & \sqrt{T}(\overline{\widehat{\mathbf{U}}} - \overline{\mathbf{U}}) \\ & - \frac{1}{r_{u,c}} \sqrt{\frac{1}{T}} (\widehat{\mathbf{U}} - \mathbf{U}) \mathbf{U}^T \boldsymbol{\omega}_u^* \\ & - \frac{1}{r_{u,c}} \sqrt{\frac{1}{T}} \mathbf{U} (\widehat{\mathbf{U}} - \mathbf{U})^T \boldsymbol{\omega}_u^* \\ & - \frac{1}{r_{u,c}} \sqrt{\frac{1}{T}} (\widehat{\mathbf{U}} - \mathbf{U})(\widehat{\mathbf{U}} - \mathbf{U})^T \boldsymbol{\omega}_u^* \\ & + \sqrt{\frac{1}{T}} \frac{1}{r_{u,c}} \mathbf{U} \left(r_{u,c} \mathbf{1} - \mathbf{U}^T \boldsymbol{\omega}_u^* \right) \\ =: & I + II + III + IV + V. \end{aligned}$$

About term I, under Assumptions 1–3, by Bernstein's inequality (Theorem 1 of Merlevède et al. (2011)),

$$P\left(\|\overline{\mathbf{U}} - \boldsymbol{\alpha}\|_{\max} > \sqrt{\frac{\log N}{T}} \right) = O\left(\frac{1}{T^2}\right).$$

Moreover, by the Cauchy-Schwarz inequality,

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

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

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

where $\check{\mathbf{F}} = (\check{\mathbf{f}}_{t,k})_{1\le t\le T, 1\le k\le K}$, $\check{\mathbf{f}}_t = (\mathbf{f}_{t,k} - \bar{f}_k)_{1\le k\le K}$, 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 $C > 0$,

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

and

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

Hence

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

By Bernstein's inequality again,

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

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,

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

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

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

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,

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

Note that $(\boldsymbol{\omega}_u^*)^T\mathbf{U}_t \overset{\text{i.i.d.}}{\sim} N(\sqrt{\theta_u},1)$, we get that for some $C>0$,

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

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

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

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

$$\|III\|_{\max}\leq\sqrt{\max_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}U_{it}^2}\sqrt{\sum_{t=1}^{T}\left((\boldsymbol{\omega}_u^*)^T(\widehat{\mathbf{U}}_t-\mathbf{U}_t)\right)^2}. \tag{A.15}$$

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

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

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

$$(\boldsymbol{\omega}_u^*)^T(\widehat{\mathbf{U}}_t-\mathbf{U}_t)=\left(\sum_{t=1}^{T}(\boldsymbol{\omega}_u^*)^T\mathbf{U}_t\check{\mathbf{f}}_t^T(\check{\mathbf{F}}^T\check{\mathbf{F}})^{-1}\right)\mathbf{f}_t.$$

Therefore,

$$\begin{aligned}\sum_{t=1}^{T}\left((\boldsymbol{\omega}_u^*)^T(\widehat{\mathbf{U}}_t-\mathbf{U}_t)\right)^2 \leq&\left\|\sum_{t=1}^{T}(\boldsymbol{\omega}_u^*)^T\mathbf{U}_t\check{\mathbf{f}}_t^T(\check{\mathbf{F}}^T\check{\mathbf{F}})^{-1}\right\|^2\left(\sum_{t=1}^{T}\|\mathbf{f}_t\|^2\right)\\ \leq&\left\|\sum_{t=1}^{T}(\boldsymbol{\omega}_u^*)^T\mathbf{U}_t\check{\mathbf{f}}_t^T\right\|^2\cdot\left\|(\check{\mathbf{F}}^T\check{\mathbf{F}})^{-1}\right\|^2\left(\sum_{t=1}^{T}\|\mathbf{f}_t\|^2\right).\end{aligned}$$

Because $(\boldsymbol{\omega}_u^*)^T\mathbf{U}_t\sim N(\sqrt{\theta_u},1)$, and independent with $\mathbf{f}_t$, we have

$$P\left(\left\|\sum_{t=1}^{T}(\boldsymbol{\omega}_u^*)^T\mathbf{U}_t\check{\mathbf{f}}_t^T\right\|>\sqrt{(\log N)T}\right)=O\left(\frac{1}{T^2}\right).$$

It follows from (A.12) that

$$P\left(\sum_{t=1}^{T}\left((\boldsymbol{\omega}_u^*)^T(\widehat{\mathbf{U}}_t-\mathbf{U}_t)\right)^2>C\log N\right)=O\left(\frac{1}{T^2}\right). \tag{A.17}$$

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

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

About term IV, we have

$$\begin{aligned}&\|(\widehat{\mathbf{U}}-\mathbf{U})(\widehat{\mathbf{U}}-\mathbf{U})^T\boldsymbol{\omega}_u^*\|_{\max}\\ \leq&\sqrt{\max_{1\leq i\leq N}\left(\sum_{t=1}^{T}(\widehat{U}_{it}-U_{it})^2\right)}\sqrt{\sum_{t=1}^{T}\left((\boldsymbol{\omega}_u^*)^T(\widehat{\mathbf{U}}_t-\mathbf{U}_t)\right)^2}.\end{aligned}$$

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

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

Finally, about term V, we have

$$\|V\|_{\max}=\frac{1}{\sqrt{T}r_{u,c}}\left(\max_{1\leq i\leq N}\left|\sum_{t=1}^{T}U_{it}\Big(r_{u,c}-\mathbf{U}_t^T\boldsymbol{\omega}_u^*\Big)\right|\right).$$

Note that

$$\begin{aligned}E\Big(\mathbf{U}_t(r_{u,c}-\mathbf{U}_t^T\boldsymbol{\omega}_u^*)\Big)&=\frac{1}{\sqrt{\theta_u}}\boldsymbol{\alpha}+\boldsymbol{\alpha}\sqrt{\theta_u}-\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^T+\boldsymbol{\Sigma_u}\right)\frac{1}{\sqrt{\theta_u}}\boldsymbol{\Sigma_u}^{-1}\boldsymbol{\alpha}\\&=\frac{1}{\sqrt{\theta_u}}\boldsymbol{\alpha}+\boldsymbol{\alpha}\sqrt{\theta_u}-\frac{1}{\sqrt{\theta_u}}\boldsymbol{\alpha}-\boldsymbol{\alpha}\sqrt{\theta_u}\\&=\mathbf{0}.\end{aligned}$$

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

$$\max_{1\leq i\leq N}E\left(\Big(U_{it}(r_{u,c}-\mathbf{U}_t^T\boldsymbol{\omega}_u^*)\Big)^2\right)\leq\max_{1\leq i\leq N}\sqrt{E\Big(U_{it}^4\Big)}\sqrt{E\Big((r_{u,c}-\mathbf{U}_t^T\boldsymbol{\omega}_u^*)^4\Big)}<C.$$

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

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

Combining the results above yields

$$P\Big(\|(W_1^T, W_2^T)\|_{\max} > C\sqrt{\log N}\Big) = O\Big(\frac{1}{T}\Big) \to 0.$$

Next, about term C_{11}, we have

$$\begin{aligned} C_{11} =& \frac{1}{T}(\widehat{\mathbf{U}}_{S_1} - \mathbf{U}_{S_1})(\widehat{\mathbf{U}}_{S_1} - \mathbf{U}_{S_1})^T \\ &+ \frac{1}{T}\mathbf{U}_{S_1}\mathbf{U}_{S_1}^T \\ &+ \frac{1}{T}(\widehat{\mathbf{U}}_{S_1} - \mathbf{U}_{S_1})\mathbf{U}_{S_1}^T \\ &+ \frac{1}{T}\mathbf{U}_{S_1}(\widehat{\mathbf{U}}_{S_1} - \mathbf{U}_{S_1})^T. \end{aligned} \tag{A.18}$$

By Assumptions 1–3 and (A.13), with probability approaching one,

$$\left\| \frac{1}{T}(\widehat{\mathbf{U}} - \mathbf{U})(\widehat{\mathbf{U}} - \mathbf{U})^T \right\|_{\max} < C\frac{\log N}{T},$$

and hence, by Assumption 4 that $\|\mathbf{\Sigma}_{u,11}^{-1}\| = O(1)$, we have

$$\left\| \frac{1}{T}\mathbf{\Sigma}_{u,11}^{-1/2}(\widehat{\mathbf{U}}_{S_1} - \mathbf{U}_{S_1})(\widehat{\mathbf{U}}_{S_1} - \mathbf{U}_{S_1})^T\mathbf{\Sigma}_{u,11}^{-1/2} \right\| = O\left(\frac{q_u \log N}{T}\right). \tag{A.19}$$

Assumption 2 implies that $\mathbf{\Sigma}_{u,11}^{-1/2}\mathbf{U}_{S_1,t} \sim N(\mathbf{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1, \mathbf{I})$. By Bernstein's inequality, we have, with probability tending one,

$$\|\mathbf{\Sigma}_{u,11}^{-1/2}\overline{\mathbf{U}}_{S_1} - \mathbf{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\|_{\max} = O\left(\sqrt{\frac{\log N}{T}}\right). \tag{A.20}$$

Assumption 3 implies that $\|\mathbf{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\| = \sqrt{\boldsymbol{\alpha}_1^T\mathbf{\Sigma}_{u,11}^{-1}\boldsymbol{\alpha}_1} = O(1)$. Therefore,

$$\begin{aligned} &\|\mathbf{\Sigma}_{u,11}^{-1/2}\overline{\mathbf{U}}_{S_1}\overline{\mathbf{U}}_{S_{11}}\mathbf{\Sigma}_{u,ss}^{-1/2} - \mathbf{\Sigma}_{u,ss}^{-1/2}\boldsymbol{\alpha}_s\boldsymbol{\alpha}_s^T\mathbf{\Sigma}_{u,11}^{-1/2}\| \\ \leq& 2\|\mathbf{\Sigma}_{u,11}^{-1/2}\overline{\mathbf{U}}_{S_1} - \mathbf{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_s\| \cdot \|\mathbf{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\| + \|\mathbf{\Sigma}_{u,11}^{-1/2}\overline{\mathbf{U}}_{S_1} - \mathbf{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\|^2 \\ =& O\left(\sqrt{\frac{q_u \log N}{T}}\right). \end{aligned} \tag{A.21}$$

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

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

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,

$$\left\|\frac{1}{T}\boldsymbol{\Sigma}_{u,11}^{-1/2}(\mathbf{U}_{S_1}-\overline{\mathbf{U}}_{S_1}\mathbf{1}^T)(\mathbf{U}_{S_1}-\overline{\mathbf{U}}_{S_1}\mathbf{1}^T)^T\boldsymbol{\Sigma}_{u,11}^{-1/2}-\mathbf{I}\right\|=O\left(\sqrt{\frac{q_u}{T}}\right). \tag{A.22}$$

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

$$\begin{aligned}&\left\|\frac{1}{T}\boldsymbol{\Sigma}_{u,11}^{-1/2}\mathbf{U}_{S_1}\mathbf{U}_{S_1}^T\boldsymbol{\Sigma}_{u,11}^{-1/2}-\mathbf{I}-\boldsymbol{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}_{u,11}^{-1/2}\right\|\\ \le&\left\|\frac{1}{T}\boldsymbol{\Sigma}_{u,11}^{-1/2}(\mathbf{U}_{S_1}-\mathbf{1}\overline{\mathbf{U}}_{S_1}^T)(\mathbf{U}_{S_1}-\mathbf{1}\overline{\mathbf{U}}_{S_1}^T)^T\boldsymbol{\Sigma}_{u,11}^{-1/2}-\mathbf{I}\right\|\\ &+\|\boldsymbol{\Sigma}_{u,11}^{-1/2}\overline{\mathbf{U}}_{S_1}\overline{\mathbf{U}}_{S_1}^T\boldsymbol{\Sigma}_{u,11}^{-1/2}-\boldsymbol{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}_{u,11}^{-1/2}\|.\end{aligned} \tag{A.23}$$

Combining (A.21), (A.22) and (A.23) yields

$$\left\|\frac{1}{T}\boldsymbol{\Sigma}_{u,11}^{-1/2}\mathbf{U}_{S_1}\mathbf{U}_{S_1}^T\boldsymbol{\Sigma}_{u,11}^{-1/2}-\mathbf{I}-\boldsymbol{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}_{u,11}^{-1/2}\right\|=O\left(\sqrt{\frac{q_u\log N}{T}}\right). \tag{A.24}$$

By $\|\boldsymbol{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\|=O(1)$, $q_u^2(\log N)=o(T)$ and (A.24), we have

$$\left\|\frac{1}{T}\boldsymbol{\Sigma}_{u,11}^{-1/2}\mathbf{U}_{S_1}\mathbf{U}_{S_1}^T\boldsymbol{\Sigma}_{u,11}^{-1/2}\right\|=O(1). \tag{A.25}$$

Therefore, by (A.22) and (A.25), we get that

$$\begin{aligned}&\left\|\frac{1}{T}\boldsymbol{\Sigma}_{u,11}^{-1/2}(\widehat{\mathbf{U}}_{S_1}-\mathbf{U}_{S_1})\mathbf{U}_{S_1}^T\boldsymbol{\Sigma}_{u,11}^{-1/2}\right\|\\ \le&\sqrt{\left\|\frac{1}{T}\boldsymbol{\Sigma}_{u,11}^{-1/2}(\widehat{\mathbf{U}}_{S_1}-\mathbf{U}_{S_1})(\widehat{\mathbf{U}}_{S_1}-\mathbf{U}_{S_1})^T\boldsymbol{\Sigma}_{u,11}^{-1/2}\right\|}\sqrt{\left\|\frac{1}{T}\boldsymbol{\Sigma}_{u,11}^{-1/2}\mathbf{U}_{S_1}^T\mathbf{U}_{S_1}^T\boldsymbol{\Sigma}_{u,11}^{-1/2}\right\|}\\ =&O\left(\sqrt{\frac{q_u\log N}{T}}\right),\end{aligned} \tag{A.26}$$

and similarly,

$$\|\frac{1}{T}\mathbf{\Sigma}_{u,11}^{-1/2}\mathbf{U}_{S_1}(\widehat{\mathbf{U}}_{S_1}-\mathbf{U}_{S_1})^T\mathbf{\Sigma}_{u,11}^{-1/2}\| = O\left(\sqrt{\frac{q_u\log N}{T}}\right). \tag{A.27}$$

Combining (A.18), (A.19), (A.24), (A.26) and (A.27) yields that

$$\|\mathbf{\Sigma}_{u,11}^{-1/2}C_{11}\mathbf{\Sigma}_{u,11}^{-1/2}-\mathbf{I}-\mathbf{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T\mathbf{\Sigma}_{u,11}^{-1/2}\| = O\left(\sqrt{\frac{q_u\log N}{T}}\right).$$

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

$$\|\mathbf{\Sigma}_{u,11}^{1/2}C_{11}^{-1}\mathbf{\Sigma}_{u,11}^{1/2}-(\mathbf{I}+\mathbf{\Sigma}_{u,11}^{-1/2}\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T\mathbf{\Sigma}_{u,11}^{-1/2})^{-1}\| = O\left(\sqrt{\frac{q_u\log N}{T}}\right).$$

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

$$\|C_{11}^{-1}-(\mathbf{\Sigma}_{u,11}+\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T)^{-1}\| = O\left(\sqrt{\frac{q_u\log N}{T}}\right).$$

Finally, about term C_{21}, we have

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

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

$$\left\|\frac{1}{T}\mathbf{U}\mathbf{U}^T-\mathbf{\Sigma}_u-\boldsymbol{\alpha}\boldsymbol{\alpha}^T\right\|_{\max} < C\sqrt{\frac{\log N}{T}},$$

and hence

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

Similarly,

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

We then get

$$\|C_{21} - \boldsymbol{\Sigma}_{u,21} - \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_1^T\|_{\max} \leq C\sqrt{\frac{\log N}{T}}.$$

The desired result (A.3) follows. □

A.2 Proof of Theorem 2

We have

$$\begin{aligned}
&\widehat{\boldsymbol{\Sigma}}_S^{-1}\widehat{\boldsymbol{\mu}}_S - \boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S \\
=&(\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1})(\widehat{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_S) \\
&+ \boldsymbol{\Sigma}_S^{-1}(\widehat{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_S) \\
&+ \boldsymbol{\Sigma}_S^{-1/2}(\boldsymbol{\Sigma}_S^{1/2}\widehat{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S^{1/2} - \mathbf{I})(\mathbf{I} - \boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2})\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S \\
&+ (\boldsymbol{\Sigma}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1})\boldsymbol{\mu}_S \\
=:&I + II + III + IV.
\end{aligned} \tag{A.28}$$

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

$$\|\widehat{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_S\| = O_p\left(\sqrt{\frac{|S|(\log|S|)}{T}}\right). \tag{A.29}$$

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

$$T\boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2} = \sum_{t=1}^{T-1}\mathbf{z}_t\mathbf{z}_t^T, \tag{A.30}$$

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,

$$\|\boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2} - \mathbf{I}\| = O_p\left(\sqrt{\frac{|S|}{T}}\right). \tag{A.31}$$

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

$$\|\boldsymbol{\Sigma}_S^{1/2}\widehat{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S^{1/2} - \mathbf{I}\| = O_p\left(\sqrt{\frac{|S|}{T}}\right) = o_p(1). \tag{A.32}$$

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

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

It follows that

$$\|I\| = O_p\left(|S|\frac{\sqrt{\log|S|}}{T}\right) = o_p\left(\frac{1}{\sqrt{T}}\right), \tag{A.34}$$

where the last equality holds by the assumptions that $|S|^2(\log|S|) = o(N)$ and $N \asymp T$, which imply that $S^2(\log|S|) = o(T)$.

About term II, by Assumption 2, we have

$$II \sim N\left(0, \frac{1}{T}\mathbf{\Sigma}_S^{-1}\right). \tag{A.35}$$

About term III, we have

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

Note that

$$\boldsymbol{\mu}_S^T\mathbf{\Sigma}_S^{-1}\boldsymbol{\mu}_S \leq \theta_{all} = \boldsymbol{\mu}_f^T\mathbf{\Sigma}_f^{-1}\boldsymbol{\mu}_f + \theta_u.$$

Hence by Assumption 3,

$$\|\mathbf{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\| = \sqrt{\boldsymbol{\mu}_S^T\mathbf{\Sigma}_S^{-1}\boldsymbol{\mu}_S} = O(1). \tag{A.36}$$

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

$$\|\mathbf{\Sigma}_S^{-1}\| = O(1). \tag{A.37}$$

By (A.31), (A.32) and (A.37), we then get

$$\|III\| = O_p\left(\frac{|S|}{T}\right) = o_p\left(\frac{1}{\sqrt{T}}\right). \tag{A.38}$$

About term IV, by (A.30), we have

$$IV = \mathbf{\Sigma}_S^{-1/2}\frac{1}{T}\sum_{t=1}^{T-1}(\mathbf{I} - \mathbf{z}_t\mathbf{z}_t^T)\mathbf{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S + \frac{1}{T}\mathbf{\Sigma}_S^{-1}\boldsymbol{\mu}_S. \tag{A.39}$$

By Assumption 3, we have $\|\boldsymbol{\mu}_S\| = O(\sqrt{|S|})$. Therefore, by (A.37), we have

$$\frac{1}{T}\|\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\| = O\left(\frac{1}{T}\|\boldsymbol{\Sigma}_S^{-1}\|\|\boldsymbol{\mu}_S\|\right) = O\left(\frac{\sqrt{|S|}}{T}\right) = o\left(\frac{1}{\sqrt{T}}\right). \tag{A.40}$$

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

$$E\Big(\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}(\mathbf{I} - \mathbf{z}_t\mathbf{z}_t^T)\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\Big) = \mathbf{0}. \tag{A.41}$$

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),

$$E(\mathbf{z}_t^T\mathbf{B}_1\mathbf{z}_t\mathbf{z}_t^T\mathbf{B}_2\mathbf{z}_t) = \operatorname{tr}(\mathbf{B}_1)\operatorname{tr}(\mathbf{B}_2) + \operatorname{tr}(\mathbf{B}_1\mathbf{B}_2) + \operatorname{tr}(\mathbf{B}_1\mathbf{B}_2^T). \tag{A.42}$$

By (A.42), we get that

$$\begin{aligned} &E\Big(e_i^T(\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t\mathbf{z}_t^T\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S)(\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t\mathbf{z}_t^T\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S)^T e_j\Big) \\ =&e_i^T(\boldsymbol{\mu}_S^T\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^T + 2\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^T)e_j, \end{aligned} \tag{A.43}$$

where e_i is a length-k vector with the ith element of one and zero elsewhere. By (A.41) and (A.43), we have

$$\begin{aligned} &\operatorname{Var}(\mathbf{A}\boldsymbol{\Sigma}_S^{-1/2}\mathbf{z}_t\mathbf{z}_t^T\boldsymbol{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S) \\ =&(\boldsymbol{\mu}_S^T\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S)\mathbf{A}\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^T \\ &+ \mathbf{A}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\boldsymbol{\Sigma}_S^{-1}\mathbf{A}^T. \end{aligned} \tag{A.44}$$

In addition, by the Cauchy-Schwarz inequality,

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

By Lemma 2.9 of Bai and Silverstein (1998),

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

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

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

where the last equality holds by (A.36). It follows that

$$E\Big(\|\mathbf{A}\mathbf{\Sigma}_S^{-1/2}\mathbf{z}_t\mathbf{z}_t^T\mathbf{\Sigma}_S^{-1/2}\boldsymbol{\mu}_S\|^4\Big) = O(1). \tag{A.45}$$

By the Lyapunov Central Limit Theorem,

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

where $\mathbf{\Sigma}_{A,IV} = \lim_{N\to\infty}(\boldsymbol{\mu}_S^T\mathbf{\Sigma}_S^{-1}\boldsymbol{\mu}_S)\mathbf{A}\mathbf{\Sigma}_S^{-1}\mathbf{A}^T + \mathbf{A}\mathbf{\Sigma}_S^{-1}\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\mathbf{\Sigma}_S^{-1}\mathbf{A}^T$. Using further the independence between $\widehat{\boldsymbol{\mu}}_S$ and $\widehat{\mathbf{\Sigma}}_S$, and (A.35), we obtain

$$\mathbf{A}(II + IV) \xrightarrow{\mathcal{L}} N(0, \mathbf{\Sigma}_A). \tag{A.46}$$

The desired result follows from (A.28), (A.34), (A.38) and (A.46). □

Appendix B Asset pricing tests of MAXSER-S(0.1)

In this section, we report the asset pricing tests of the MAXSER-S(0.05) SDF. First, similar to the test results reported in Table 2, we examine the pricing power of the MAXSER-S(0.05) SDF in explaining the returns of 188 anomaly portfolios. The results are reported in Table B.1. We see that there are totally 5 (0) anomalies with significant alphas for a threshold $t = 1.96$ ($t = 3$). Comparing with the performance of the MAXSER-S(0.05) SDF reported in Table 2, we see that MAXSER-S(0.1) has the same number of rejections.

Table B.1: **The number of significant alphas from testing 188 anomalies against MAXSER-S(0.1).** We examine the explanatory power of the MAXSER-S(0.1) SDF over 188 anomalies. This table summarizes the total number of significant alphas and the number of significant alphas in each characteristic group. The number in parenthesis is the total number of portfolios in each anomaly group. The number of significant alphas under the threshold $t > 1.96$ or $t > 3$ is reported. The evaluation period is between 2000 and 2021.

	All (188)		frictions (10)		intangibles (30)		investment (29)	
threshold of t	1.96	3	1.96	3	1.96	3	1.96	3
MAXSER-S(0.1)	5	0	0	0	0	0	1	0
	momentum (41)		profitability (46)		value-growth (32)			
threshold of t	1.96	3	1.96	3	1.96	3		
MAXSER-S(0.1)	4	0	0	0	0	0		

Next, similar to the test results reported in Table 3 based on the MAXSER-S(0.05), we evaluate the explanatory power of the benchmark models over the SDF constructed from MAXSER-S(0.1). Table B.2 reports the alpha and its t-statistic from the regressions.

Table B.2: **Using benchmark models to explain MAXSER-S(0.1) SDF.** This table reports the regression results of the SDF portfolio from MAXSER-S(0.1) against several benchmark models, using monthly returns from 2000 to 2021. Alpha in percentage and its corresponding t-statistic are reported.

	Alpha (%)	t-statistic
CAPM	2.34	5.80
FF3	2.26	7.44
FF5	1.77	5.20
FF6	1.73	5.36
Q4	1.74	5.78
Q5	1.23	4.39
BS6	1.77	5.57
DHS3	1.76	5.10
SY4	1.31	3.69
KNS	1.29	3.80

Table B.2 shows that the alpha of MAXSER-S(0.1) is both economically large and statistically significant under all benchmark models. However, comparing with the results

in Table 3, we see that the alphas of MAXSER-S(0.1) are all lower than those of MAXSER-S(0.05).

Finally, similar to the results reported in Table 4 for the MAXSER-S(0.1) SDF, we switch the roles between benchmark models and MAXSER-S(0.1), and check whether MAXSER-S(0.1) is able to explain the prevailing pricing factors. Table B.3 reports the alpha and its t-statistic from regressions. We see from Table B.3 that all the alphas are insignificant at the t-statistic of 1.96.

Table B.3: **Using MAXSER-S(0.1) SDF to explain commonly used pricing factors.** This table reports the regression results of various pricing factors against the MAXSER-S(0.1) SDF, using monthly returns from 2000 to 2021. Alpha in percentage and its corresponding t-statistic are reported.

Factor	Alpha (%)	t-statistic
Mkt-RF	0.29	0.77
SMB	0.03	0.14
HML	0.19	0.78
RMW	0.19	0.97
CMA	0.11	0.79
MOM	0.18	0.54
R_ME	0.03	0.16
R_IA	0.00	0.02
R_ROE	0.06	0.25
R_EG	0.22	1.16
HmLm	0.16	0.54
MGMT	0.09	0.45
PERF	0.24	0.50
PEAD	0.19	1.45
FIN	0.07	0.26
KNS	0.05	0.26

References

Bai, Z. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Annals of Probability*, 26(1):316–345.

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*.

Feldheim, O. N. and Sodin, S. (2010). A universality result for the smallest eigenvalues of certain sample covariance matrices. *Geometric And Functional Analysis*, 20(1):88–123.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151:435–474.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.