Factor Correlation and the Cross Section of Asset Returns: a Correlation-robust Approach [*]

Chuanping Sun [†]

This version: January 2023

Abstract

Factor correlation is an important consideration when selecting factors for explaining the cross-sectional asset returns, and ignoring it often compromises the robustness and validity in standard methods for such an exercise. This paper investigates high-dimensional factor models for cross-sectional asset returns with a focus on robust estimation when factors are (highly) correlated. We utilize a newly developed Machine Learning method to select factors and to disentangle correlated factors without imposing structural assumptions. We develop asymptotic properties for this estimator with relaxed assumptions and show that it is consistent under mild conditions. Empirically, we illustrate that our method is robust with correlated factors and consistently identifies that the 'market' factor is the most important factor for cross-sectional asset returns, while other benchmarks are adversely affected by factor correlations, rendering the 'market' factor redundant.

JEL classification: C38 C55 G12
Keywords: Factor Correlation, Factor Investing, Cross-sectional Asset Pricing, Machine Learning, LASSO, Firm Characteristics, Stochastic Discount Factor

[*] I am indebted to Mario Figueiredo for his help. I am also grateful to Thomas Sargent, Liudas Giraitis, Marcelo Fernandes, Emmanuel Guerre, Anders Kock, Dick Van Dijk and Kazuhiro Hiraki for discussions and suggestions. I would like to thank participants at the RES 2019 meeting; the Econometric Institution international PhD conference 2019; the Econometric Society Asian meeting and European meeting 2019; the SoFiE 2019 meeting; the EFMA 2019 meeting; the SITE 2019 workshop, the EEA 2020 meeting and the Econometric Society China meeting 2021 for valuable comments and discussions.

[†] Faculty of Finance, Bayes Business School (formerly Cass), City, University of London, 106 Bunhill Row, London EC1Y 8TZ, United Kingdom, chuanping.sun@city.ac.uk

1 Introduction

There is a burgeoning literature that attempts to examine and dissect the relation between high dimensional firm characteristics (or firm characteristic related factors, a.k.a factor zoo) and the cross-sectional asset returns, see Cochrane (2011), Harvey et al. (2015), Green et al. (2017), Hou et al. (2020), Feng et al. (2020), Freyberger et al. (2020) and among others for example. However, a vigorous discussion regarding the factor correlations and their impact and implications on cross-sectional asset returns has fallen short in the related literature. In a high-dimensional setting, many firm characteristics (or the related factor zoo) are often close cousins and highly correlated between each other. Ignoring factor correlations would compromise the robustness of standard models and therefore cast doubts on the validity of such models. For example, in our empirical analysis, we show that standard econometric models, such as the Fama-Macbeth two-step procedure and the LASSO shrinkage method, both failed to identify the 'market' factor as an important factor in driving cross-sectional asset returns due to the fact that the 'market' factor is highly correlated with many factors in the "factor zoo".

In this paper, we utilize and further develop a LASSO-type estimator, namely the *Ordered-Weighted-LASSO* (OWL, Figueiredo and Nowak (2016)) estimator, to find prevailing factors for the cross sectional asset returns with a focus on disentangling correlated factors. In other words, how do we robustly choose important factors when they are highly correlated? Our contribution to the related literature is twofold. First, we further develop the asymptotic properties of the OWL estimator under relaxed assumptions which are more suitable for economic and finance related research questions compared to Figueiredo and Nowak (2016) and show the consistency of such estimator under some mild conditions. In addition, we combine the OWL shrinkage method with the stochastic discount factor (SDF) method (Cochrane, 2005) to choose factors. Based on that, we derive the grouping property of factor selection for the cross section of asset returns (i.e., we quantify conditions of identifying highly correlated factors for cross-sectional asset returns). Monte Carlo simulation exercise shows favorable results of the OWL shrinkage method compared to other machine learning benchmarks such as the LASSO, the adaptive LASSO and the Elastic Net models, especially when factors are highly correlated.

The second fold of our contribution offers new insights on some puzzling questions in the cross-sectional asset pricing literature. High correlation among factors often attenuates statistical significance of many factors, and in particular, renders the 'market' factor insignificant to drive asset prices. Traditional tactic to avoid such problem often involves excluding highly correlated factors in the model. For example, Green et al. (2017) exclude 'beta' related factors in their Fama-MacBeth framework before finding factors that drive cross-sectional returns in the US stock market. However, such procedure requires a threshold level to decide which factors to be dropped and, such decision often lacks rigorous justification. Furthermore, if two factors are highly correlated, it is not trivial to decide which one to drop and which one to keep. By contrast, our method is robust to correlated factors - highly correlated factors receive similar coefficients in our model. Thus, we avoid any ad hoc screenings of factors before applying our model. Furthermore, we provide ample evidence to show that the 'market' factor, despite often being deemed insignificant in standard testing procedures due to high-correlation with other factors, is indeed an important factor in driving cross-sectional asset returns in our estimation framework. This coincides with a new finding made in Harvey and Liu (2021) utilizing a bootstrap based testing procedure. We discuss this in detail in the following text.

For empirical analysis, we consider 100 factors documented in Green et al. (2017) for factor investing, then we form hedging portfolios following Freyberger et al. (2020) using a sparse set of factors selected by various methods. To do this, we first construct anomaly factors (i.e., factors other than the 'market' factor) for each firm characteristic via portfolio sorting. Then, we follow Feng et al. (2020) to form thousands of bi-variate sorted portfolios as our test assets.[1] It is worth stressing that we are using sorted portfolios (by firm characteristics) as test assets instead of individual stocks. This is because large proportion of individual stocks are small stocks which takes an insignificant weight in the aggregated market value, whereas some small number of large stocks take a large proportion of the aggregated market value. Therefore, factor selection, if using individual stocks as test assets, will biased towards explaining mainly small stocks, rather than the aggregated market. On the other hand, using (value weighted) sorted portfolios as test assets can

[1] For robustness check (see Appendix F), we also consider other methods of constructing test portfolios while controlling for micro stocks.

effectively test for prevailing factors for cross-sectional asset returns at the aggregated level.

Our empirical findings complement and challenge some common stances in the asset pricing literature. First, when we implement a Fama-MacBeth regression procedure we find serious correlations among factor loadings: 68% of the correlation coefficients are higher than 0.5 (absolute value), which casts doubts on the validity of using standard estimation methods such as Fama-MacBeth regression and the LASSO shrinkage method: we find that the Fama-MacBeth, the LASSO and the Elastic Net shrinkage methods all failed to identify the 'market' factor as an important factor to drive the cross section of asset returns. That is because the 'market' factor is highly correlated with many characteristics-based factors and correlation between factors erodes the valid inference for those methods. On the contrary, the OWL shrinkage method can consistently identify the 'market' factor as the most important driver for cross-sectional returns. This finding coincides with the empirical evidence by Harvey and Liu (2021) showing that the 'market' factor is the primary factor to drive asset prices following a bootstrap motivated test.

Second, we find that *'liquidity'*, *'asset growth rate'*, *'profitability'* and *'investment'* related factors are the main drivers of the variation of cross-sectional average returns. This finding is consistent with Hou et al. (2020, 2021). Interestingly, we also find that the 'size effect' disappears during the 1980-2000 period, which is well documented in the literature, see Amihud (2002) and Asness et al. (2018) for an example. Nonetheless, the size effect becomes evident again after removing more small stocks (smaller than 40 percentile of the NYSE listed), implying that the vanishing size effect is likely to be caused by some small "junk" stocks. Once "junk" stocks are removed, the size effect resurfaces, which echoes the discovery by Asness et al. (2018): *"size matters, if you control your junk"*.[2]

Third, we follow a similar procedure to Freyberger et al. (2020) to conduct an out-of-sample exercise to find which method selects factors that can best predict the cross-sectional asset returns and use those factors to construct hedged portfolios. We compare hedged portfolios and find that the hedged portfolio using OWL selected factors produces 20% to 30% higher out-of-sample Sharpe ratios than those of other methods, suggesting that the

[2] Asness et al. (2018) add some controlling factors measuring the "junk-level" of stocks in their regression model and find that the size-effect is evident again after controlling the "junk".

OWL shrinkage method, compared to other benchmarks, can pick the best factors that contribute to the cross section of asset returns in an environment where factor correlation is prevalent.

Finally, it is worth stressing that the OWL estimator, like many other machine learning methods, is biased. Therefore, it is unfit to conduct statistical tests without further developing an unbiased version and deriving its asymptotic properties. We elaborate on this subject in Section 2.4. Having mentioned that, we want to point out that our main object in this paper is not to conduct statistical tests to determine a parsimonious asset pricing model - although such research agenda is crucially important, it can be a topic of future research. Nonetheless, our paper focuses on the robust estimation and selection of correlated factors in high-dimensional factor models. Then, we postulate a sparse model, say a five-factor model,[3] and we compare such selection of factors with other benchmarks. Our empirical result coincides with some recent empirical findings and also offers explanations to some puzzling questions faced in the cross-sectional asset pricing literature.

Related literature

This paper naturally builds on a series of papers devoted to identifying pricing factors for cross-sectional asset returns, for example see Fama and French (1992), Carhart (1997) , Hou et al. (2014), Fama and French (2016), Fama and French (2018) among others. Now after over half a century since the CAPM of Sharpe (1964) and Lintner (1965), hundreds of anomaly factors have been proposed to explain the cross-sectional asset returns. However, Harvey et al. (2015) document 316 factors and find most of them are the result of data-snooping. Hou et al. (2020) try to replicate 447 anomaly factors, and find 64% to 85% of them cannot be replicated depending on the choice of a significance level. Kan and Zhang (1999) caution that the presence of useless factors bias test results, leading to a lower than normal threshold to accept priced factors. Gospodinov et al. (2014) develop a model mis-specification robust test to tackle suprious factors, using a step-wise test to

[3] Note that we do not assume that the true model is a five-factor model. Instead, we use various methods to choose important factors and we restrict all benchmarks to have the same number of factors to conduct a prediction exercise, such that we ensure each benchmark does not suffer from the over-fitting problems in this out-of-sample exercise. We also implement such an exercise with a four factor model and the result is similar, thus it is not reported.

remove useless factors one by one. Fama and French (2018) use Sharpe ratio and employ the Right-Hand-Side method of Barillas and Shanken (2018) to "*choose factors*". Harvey and Liu (2021) suggest a step-wise bootstrap method to test for factors and find that the 'market' factor is the most important factor for cross-sectional asset returns.

This paper also relates to the strand of research focuses on methodologies of selecting and testing factors. Fama and MacBeth (1973) put forward the two-pass regression method which is commonly used to test for factors with significant risk premiums. Green et al. (2017) use Fama-MacBeth regression procedure to find significant factors among 100 candidate factors for the US stock market. Lewellen (2015) studies the cross sectional properties of return forecasts derived from the Fama-MacBeth regression and finds that forecasts vary substantially across stocks and have strong predictive power for actual returns.

This paper also contributes to the rapidly growing literature using machine learning techniques for financial research questions. Tibshirani (1996) proposes the LASSO estimator which receives huge success and becomes a new norm nowadays in dealing with high-dimensional data-sets. Since then, many adaptations and improvements have been made to achieve various targets. The literature about the LASSO family evolves rapidly. Belloni et al. (2014) devised the double LASSO selection procedure for causal inference while having large number of control variables. Feng et al. (2020) employ the double LASSO selection procedure to recursively evaluate (in a chronological order) if factors are significant to explain the cross-sectional stock returns. Yuan and Lin (2006) propose the group LASSO estimator which allows for correlated factors - we can put correlated factors or factors share similar characteristics in groups. During estimation, the group LASSO estimator can shrink off all variables together in a group if such group is deemed unimportant by the group LASSO estimator. Freyberger et al. (2020) employ the adaptive group LASSO to find pervasive factors to predict cross-sectional stock returns.[4] Babii et al. (2021) utilize the sparse group LASSO estimator for nowcasting GDP. Our paper is closely related to the group-LASSO shrinkage method in the sense that both estimation methods allow for factor correlation. However, the crucial difference is that group LASSO method requires

[4] It is worth mentioning that Freyberger et al. (2020) use the group LASSO to put all terms of the non-parametric transformation of each firm characteristic in a group, while potential correlations between firm characteristics are not discussed.

prior knowledge of how to cluster factors in various groups before implementing the model. On the contrary, the OWL shrinkage method does not require such prior knowledge and the OWL estimator can identify highly correlated factors during estimation. This identification of correlated factors happens simultaneously with the shrinkage of useless factors. Zou and Hastie (2005) propose the Elastic Net estimator, which stabilizes factor selection among correlated variables. Kozak et al. (2020) employ the Elastic Net estimator in a Bayesian framework and find that sparse principle components can largely explain the cross-section of the average returns. Gu et al. (2020) compare popular machine learning techniques used in empirical asset pricing literature and demonstrate large economic gains using random forest and neuron networks. van Binsbergen et al. (2022) show that, using a sophisticated random forest algorithm, 'machine' wins the contest against 'man' in the contest of predicting earnings of stocks. On the other hand, Cao et al. (2021) argue that 'man' wins the competition against 'machine' when firms are complex and with intangible assets, while 'machine' wins the contest when information is transparent and voluminous - combining them, however, yields the best result in forecasting stock prices.

Finally, the shrinkage method used in this paper is built directly upon the work of Figueiredo and Nowak (2016) and Zeng and Figueiredo (2014). Our innovation in this paper compared to their work is two folds: first, we further develop statistical properties of the OWL estimator under less restrictive assumptions (we relax the Gaussianity assumption and instead, impose tail bounds on the distribution of random variables). Second, we combine the OWL shrinkage method with the stochastic discount factor (SDF) method in finance to search for prevailing factors that drives cross-sectional asset returns.

The remainder of this paper is organized as follows: Section 2 presents the methodology to find prevailing factors that drive cross-sectional returns. Then, we move on to introduce the OWL shrinkage method and discuss its statistical properties. Section 3 use Monte Carlo simulation experiments to evaluate the performance of the OWL shrinkage method under various settings and compare it with some benchmarks that are commonly employed in economic and finance research for high dimensional data-sets. Section 4 presents empirical findings and discuss our contributions to the asset pricing literature. Section 5 presents conclusions.

2 Methodology

We adopt the SDF method in Cochrane (2005) to infer factors that drive cross-sectional asset returns. Section 2.1 compare the SDF method and the Fama-MacBeth two-pass regression method and point out that the former should be adopted when factors are correlated; Section 2.2 introduces the model and discusses challenges and opportunities in high-dimensional financial applications. Sections 2.3 and 2.3.2 introduce the OWL shrinkage method and discuss its statistical properties.

2.1 Risk price or risk premium?

Let m denote the stochastic discount factor (SDF)

$$m = r_0^{-1}(1 - b'(f - \mathrm{E}(f))), \tag{1}$$

where r_0 is the zero beta rate which is a constant, f is a $K \times 1$ vector of K factor returns, which can be either traded factors or mimicking portfolio returns of non-traded factors. b is a $K \times 1$ vector of the unknown SDF coefficient, referred to as the *risk price*; a non-zero (zero) entry of b means the corresponding factor is (not) priced. We want to draw inferences on the risk prices of factors. Finding useful factors (i.e., factors with non-zero risk prices) is our target. Useful factors drive the variation of SDF, thus contain pricing information: they reflect the marginal utility of factors to explain the cross-section of average returns.

On the other hand, factors can be useless or redundant. Useless factors are those whose risk prices are zero and uncorrelated with other useful factors. Redundant factors also have zero risk prices but they are correlated with some useful factors. The difference between useless factors and redundant factors plays an important role when choosing between the SDF method and the two-pass Fama-MacBeth regression method to search for prevailing factors.

A closely related concept to risk price is the risk premium. It refers to the slope coefficient in the second pass Fama-MacBeth regression.[5] Cochrane (2005) shows that risk price and risk premium are directly related through the covariance matrix of factors

$$\lambda = \mathrm{E}(ff')b, \tag{2}$$

where b is a vector of risk prices and λ is a vector of risk premiums. However, they differ substantially in their interpretation. Risk premium of a factor infers how much an investor demands to pay for bearing the risk of the factor. Risk price implies whether a factor is useful to explain the cross-section of average asset returns. When factors are uncorrelated, that is, $\mathrm{E}(ff')$ is a diagonal matrix. Then, $b_i = 0$ (the i^{th} factor is not priced) implies $\lambda_i = 0$ (the i^{th} factor earns zero risk premium), and vice verse. In this case, using risk premium to infer prevailing factors for cross-sectional returns yields the same result as do risk prices. However, this is not true when factors are correlated: an unpriced factor can earn positive risk premium by being correlated with a useful factor. To give an example, suppose we have two factors f_1 and f_2, the covariance matrix is $\mathrm{E}(ff') = \begin{pmatrix} 10 & 1 \\ 1 & 10 \end{pmatrix}$, the first factor is priced and the second is not, that is $b_1 = 1 \neq 0$ and $b_2 = 0$. Then, according to (2), we have $\lambda_1 = 10$ and $\lambda_2 = 1$. So we find that the unpriced factor f_2 (i.e. $b_2 = 0$) earns non-zero risk premium (i.e. $\lambda_2 \neq 0$) by simply being correlated with a useful factor f_1. As discussed before, if factors are uncorrelated it is valid to use either risk price (SDF method) or risk premium (Fama-MacBeth regression) to select factors. However, factors are typically correlated in a high dimensional setting, so we should use *risk price* to infer priced factors under such circumstance.

2.2 Model

Denote by R the excess returns of a vector of N test assets. Define $Y = (f', R')'$, so $\mathrm{Var}(Y) = \begin{pmatrix} \mathrm{Var}(f) & \mathrm{Cov}(R,f)' \\ \mathrm{Cov}(R,f) & \mathrm{Var}(R) \end{pmatrix}$, where $\mathrm{Var}(f)$ and $\mathrm{Var}(R)$ are the $K \times K$ and

[5] The two-pass Fama-MacBeth regression procedure involves the following two steps: the first pass regression obtains the factor loadings by running time-series regressions of each asset on factors; the second pass runs cross-sectional regressions of asset returns on factor loadings.

$N \times N$ variance-covariance matrices of factors f and test asset returns R, respectively. $\mathrm{Cov}(R, f)$ is the $N \times K$ covariance matrix of returns and factors. The fundamental asset pricing equation states that $\mathrm{E}(Rm) = \mathbf{0}$ for any admissible SDF. However, the fundamental equation may not hold when m is unknown and is estimated from a model. The deviation from zero of the above equation is regarded as the pricing error. Let $m(b)$ denote the unknown SDF which depends on the unknown risk price b. Pricing error $e(b)$ can be written and simplified as

$$\begin{aligned} e(b) &= \mathrm{E}[Rm(b)] = \mathrm{E}(R)\mathrm{E}(m(b)) + \mathrm{Cov}(R, m(b)) \\ &= r_0^{-1}\mathrm{E}(R)\mathrm{E}(1 - b'(f - E(f))) + r_0^{-1}\mathrm{Cov}(R, 1 - b'(f - E(f))) \\ &= r_0^{-1}[\mathrm{E}(R) - \mathrm{Cov}(R, f)b] \\ &= r_0^{-1}(\mu_R - Cb), \end{aligned} \tag{3}$$

where $\mu_R := \mathrm{E}(R)$ is the $N \times 1$ vector of the expectation of excess returns of test assets and $C := \mathrm{Cov}(R, f)$. A quadratic form of the pricing error can be defined as

$$Q(b) = e(b)' \, W \, e(b), \tag{4}$$

where W is a $N \times N$ weighting matrix. Then we can estimate b by minimizing $Q(b)$:[6]

$$\hat{b} = \arg\min_b \, Q(b) = \arg\min_b \, (\mu_R - Cb)'W(\mu_R - Cb), \tag{5}$$

which gives

$$\hat{b} = (C'WC)^{-1}C'W\mu_R, \tag{6}$$

For the weighting matrix W, Ludvigson (2013) offers two choices of W for comparing models. First, $W = \mathrm{E}(RR')^{-1}$, which connects $Q(b)$ to the well known Hansen-Jagannathan (HJ) distance. Ludvigson (2013) points out that the use of HJ distance accounts for and offsets the variations of test assets, leading to stable estimators. Therefore, it is preferred when small number of test assets are available. On the other hand, when test assets are prolific, Ludvigson (2013) advocates the second choice of W: the identity matrix. She

[6] Since r_0 in (3) is a constant, it can be dropped out in the minimization problem.

argues that using the identity matrix does not tilt the weight to favour any subset of test assets, especially when test assets represent particular economic interests. In our application, the test assets consist of firm characteristic sorted portfolios, hence we do not want to tilt the weights to favour any firm characteristics, so the identity matrix will be used as the weighting matrix throughout this paper.

Cochrane (2011) points out that traditional methods to identify useful factors have fallen short in the high-dimensional world. On the other hand, recent finance research has demonstrated ample evidence that many firm-characteristics based factors are not priced. Thus, the sparsity assumption which originates from the machine learning literature becomes a useful tool to handle these problems. The LASSO estimator (Tibshirani, 1996) is a powerful tool to achieve sparse models and gains immense popularity in recent years in the finance related literature. However, the LASSO estimator is also well known for its poor performance when covariates are correlated. To circumvent the curse of dimensionality while taking account of factor correlations, we introduce a newly developed machine learning tool, the Ordered-Weighted-LASSO (OWL) estimator (Figueiredo and Nowak, 2016), and tailor it to select factors from the (highly correlated) factor zoo under the SDF framework.

2.3 The Ordered-Weighted-LASSO (OWL) estimator

The OWL estimator is achieved by adding a penalty term in equation (5)[7]

$$\hat{b} = \arg\min_{b} \frac{1}{2}(\mu_R - Cb)'(\mu_R - Cb) + \Omega_\omega(b), \qquad \Omega_\omega(b) = \omega'|b|_\downarrow, \tag{7}$$

where $|b|_\downarrow := (|b|_{[1]}, |b|_{[2]}, \cdots, |b|_{[K]})'$ and $|b|_{[1]} \geq |b|_{[2]} \geq \cdots \geq |b|_{[K]}$, is a vector of the absolute values of risk prices, decreasingly ordered by their magnitude. ω is a pre-specified $K \times 1$ weighting vector, defined as

$$\omega_i = \lambda_1 + (K - i)\lambda_2, \qquad i = 1, ..., K, \tag{8}$$

[7] We use the identity matrix for the weighting matrix W.

where λ_1 and λ_2 are two tuning parameters. In order to solve (7), we use the proximal gradient descent algorithm. More details about this algorithm are included Appendix C. The OWL estimator is sensitive to the choice of the weighting vector ω. So finding appropriate values for tuning parameters λ_1 and λ_2, which pin down the weighting vector, is crucial. Following the machine learning literature, we use a ten-fold cross-validation method to find tuning parameters.[8]

In Appendix A, we present a geometric interpretation of the OWL penalty and a comprehensive comparison between the OWL and the LASSO shrinkage methods, following an argument typically employed in the machine learning literature. In the next section, we start to discuss statistical properties of the OWL estimator.

2.3.1 The grouping property

Next, we present the grouping property, which quantifies the condition for identifying correlated factors - this is a key property of the OWL estimator which enables correlation-robust estimation.

Theorem 2.1 (Grouping). *Let f_i and f_j denote the i^{th} and j^{th} factor (both of size $T \times 1$). $\hat{b}_i$ and $\hat{b}_j$ are OWL estimates of risk prices of factor i and j. Let $\sigma(f_i - f_j)$ denote the standard deviation of the vector $f_i - f_j$, and μ_R, σ_R be the $N \times 1$ vectors collecting the mean and standard deviation of N test assets. If*

$$\sigma(f_i - f_j) < \frac{\lambda_2}{\|\mu_R\|_2 \ \|\sigma_R\|_2},$$

then $\quad \hat{b}_i = \hat{b}_j$.

Proof: see Appendix B.1.

[8] Specifically, given a grid of values for λ_1 and λ_2, at each point on the grid, we first divide the sample into ten equal parts in its time series dimension. Then, we use nine parts (training sample) to estimate the model with the OWL estimator. After obtaining the estimated model, we forecast the returns of the tenth part (testing sample), and compute the root of mean squared forecast error (RMSE). We repeat the same procedure ten times by rotating the training samples and testing samples, and therefore compute the average RMSE for this point on the grid. Tuning parameters are determined by the smallest average RMSE on the grid.[9]

Corollary 2.1. *Let $f_i, f_j, \lambda_2, \mu_R, \sigma_R$ be the same as in Theorem 2.1. If*

$$\sigma(f_i + f_j) < \frac{\lambda_2}{\|\mu_R\|_2 \ \|\sigma_R\|_2},$$

then $\quad \hat{b}_i = -\hat{b}_j$.

Proof: see Appendix B.2.

Theorem 2.1 has several implications. First, when factors are highly correlated (i.e. $\sigma(f_i - f_j)$ is small) they are more likely to be grouped together (i.e. receive similar coefficients, $\hat{b}_i \approx \hat{b}_j$): two factors exhibiting high correlation could be the result of the same unobservable underlying factor that dictates these observable factors simultaneously. Thus, they should share similar magnitude in explaining asset returns which are driven by the same unobservable underlying factor. Second, the hyper parameter λ_2 in (8) has direct impact on the grouping property: large λ_2 encourages grouping. This property comes with the key design of the OWL shrinkage method as discussed above, more detailed discussion can be found in Appendix A. Third, the mean (μ_R) and standard deviation (σ_R) of test assets affect the grouping property. A set of less informative assets (small μ_R and/or small σ_R) will result in factor grouping: factors are equally weak to explain a set of test assets whose returns vary little across time. Corollary 2.1 extends the grouping property of the OWL estimator to *negatively* correlated factors: highly but *negatively* correlated factors will receive a similar magnitude in the coefficient but with opposite signs.

It is worth mentioning that the grouping property distinguishes the OWL estimator from other related machine learning methods, such as the LASSO and Elastic Net estimators, and it is the main reason why we argue that the OWL estimator should be employed when factors are highly correlated. Theorem 2.1 shows that the OWL estimator *permits* correlations among factors and assigns them with similar coefficients. On the other hand, the LASSO estimator may arbitrarily set some highly correlated factors to zeros while keeping others as non-zeros, resulting in unstable estimation results.[10]

[10] In Appendix A, we present a detailed analysis on the problems that the LASSO estimator may encounter when factors are highly correlated.

2.3.2 Asymptotic properties

This section discusses the asymptotic properties of the OWL estimator under less restrictive assumptions made in Figueiredo and Nowak (2016). We allow the number of factors K to diverge and potentially $K \gg N$. Under some regularity conditions, we derive the oracle inequality (error bounds) and the convergence rate of the OWL estimator, and hence the conditions for consistent OWL estimation. Consider a linear high-dimensional asset pricing model and suppose that

$$\mu_R = Cb^0 + \epsilon, \tag{9}$$

where b^0 is the true risk price coefficients and ϵ is the pricing error from (3) after scaling a constant r_0^{-1}. Equation (7) is a penalized estimator of model (9) and it can be written as[11]

$$\hat{b} = \arg\min_{b} \quad \frac{1}{N}||\mu_R - Cb||_2^2 + \frac{1}{N}\sum_{i=1}^{K} \lambda_1 + \lambda_2(K-i)]|b|_{[i]}, \tag{10}$$

where$|b|_{[i]}$ is the i^{th} element of $|b|_{\downarrow} := (|b|_{[1]}, |b|_{[2]}, \cdots, |b|_{[K]})'$ and $|b|_{[1]} \geq |b|_{[2]} \geq ... \geq |b|_{[K]}$.

First, we use the following notations and assumptions to derive our theoretical results. Denote by $\zeta_j := \epsilon' C^{(j)} := \sum_{i=1}^{N} \epsilon_i C_i^{(j)} := \sum_{i=1}^{N} \zeta_{i,j}$, where $C^{(j)}$ is the j^{th} column of C and ϵ is defined in (9). We denote $\hat{\Sigma} = \frac{1}{N} C'C$ as the scaled Gram Matrix of C. For any scalar $y \in R$, we denote $|y|$ the absolute value of y. For any vector $x \in R^N$, we denote $\|x\|_2 = (\sum_{i=1}^{N} x_i^2)^{1/2}$, $\|x\|_1 = \sum_{i=1}^{N} |x_i|$ and $\|x\|_\infty = \max_{1 \leq i \leq N} |x_i|$. In order to derive the next theorem, we make the following assumptions.

Assumption 1 (Random Variables). $\{\zeta_{i,j}\}_{i=1}^{N}$ *are identically and independently distributed and* $\mathrm{E}(\zeta_{i,j}) = 0$ *for* $i = 1, \cdots, N$ *and* $j = 1, \cdots, K$. *The distributions of variable* $C_i^{(j)}$ *and* ϵ_i *for all* $i = 1, \cdots, N$ *are uniformly subgaussian such that* $\sup_{i,j} \mathbb{P}(|C_i^{(j)}| > a) \leq c_1 \exp[-c_2 a^2]$ *and* $\sup_i \mathbb{P}(|\epsilon_i| > a) \leq c_1 \exp[-c_2 a^2]$ *for all* $i = 1, \cdots, N$, $a > 0$ *and some* $c_1, c_2 > 0$ *which do not depend on* a, i, j.

Assumption 1 outlines the conditions for random variables. Note that Assumption

[11] Note that the scalar "2" on the second term of (10) is dropped because it is negligible for tuning parameter $\frac{\lambda_1}{N} \asymp \sqrt{\frac{\log K}{N}}$, which will be introduced in the next theorem.

1 is more relaxed compared to Figueiredo and Nowak (2016) (i.e., *i.i.d. Gaussian*) and these assumptions on random variables are commonly assumed in the high-dimensional econometric literature according to Kock (2016).

Assumption 2 (Sparsity)**.** *Denote by S the number of non-zero parameters in $b^0 = \{b^0_1, b^0_2, \cdots, b^0_K\}$. We assume that $S\sqrt{\frac{\log K}{N}} = o(1)$ when $N, K \to \infty$.*

Let s_0 denote a subset, $s_0 \subset \{1, \cdots, K\}$, and $|s_0|$ the cardinality of s_0. For $b = \{b_1, \cdots, b_K\} \in \mathbf{R}^K$, denote $b_{s_0} := b_i\mathbf{1}\{i \in s_0, i = 1, \cdots, K\}$, $b_{s_0^c} := b_i\mathbf{1}\{i \notin s_0, i = 1, \cdots, K\}$. Then $b = b_{s_0} + b_{s_0^c}$.

Assumption 3 (Restricted eigenvalue condition, Bickel et al. (2009))**.** *For all b such that $||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1$, $\hat{\Sigma}$ satisfies the restricted eigenvalue condition*

$$\phi_0^2 := \min_{\substack{s_0 \subset \{1,\ldots,K\} \\ |s_0| < K}} \quad \min_{\substack{b \in R^K \setminus \{0\} \\ ||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1}} \frac{b'\hat{\Sigma}b}{||b_{s_0}||_2^2} > 0. \tag{11}$$

Assumption 2 and 3 are necessary conditions for deriving asymptotic properties in high-dimensional statistics. Assumption 2 is often referred to as the approximate sparsity assumption which is a rather mild assumption - it only requires that the log-rate of the high-dimensional parameter K, scaled by the sparsity parameter S grows slower than the rate of the number of observations N. Yet the exact sparsity parameter S is not assumed. The restricted eigenvalue condition in Assumption 3 circumvents the issues stemmed from a degenerate scaled gram matrix under high-dimensional factor models. See Appendix D for a detailed discussion on restricted eigenvalue condition and its implications.

Theorem 2.2 (Oracle inequality)**.** *Let Assumptions 1, 2 and 3 be satisfied. Suppose that $\lambda_0 = \kappa\sqrt{\frac{\log K}{N}} = o(1)$, where κ is a positive constant. Let $\frac{\lambda_1}{N} = 2\lambda_0$ and $\frac{\lambda_2}{N} = O(\frac{S\log K}{NK})$. Then, by selecting a sufficiently large κ, as $N, K \to \infty$, with probability tending to one, $\hat{b}$ satisfies*

$$(\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0) + \frac{\lambda_1}{N}||\hat{b} - b^0||_1 \leq 4(\frac{\lambda_1}{N})^2\frac{S}{\phi_0^2} + 2\frac{\lambda_2}{N}(K-1)||b^0||_1. \tag{12}$$

Proof: see Appendix B.3.

Note that the oracle inequality of (12) can be further developed to offer upper bounds separately for the prediction error $(\hat{b}-b^0)'\hat{\Sigma}(\hat{b}-b^0) := \|C(\hat{b}-b^0)\|_2^2/N$ and the estimation error $\|\hat{b}-b^0\|_1$. Therefore, we further utilize these bounds for the prediction error and estimation error to obtain the convergence rate of the OWL estimator.

Corollary 2.2 (Convergence rate of OWL). *Let all conditions in Theorem 2.2 be satisfied, then*

$$||\hat{b}-b^0||_1 = O\left(S\sqrt{\frac{\log K}{N}}\right), \qquad ||\hat{b}-b^0||_2 = O\left(\sqrt{\frac{S\log K}{N}}\right). \tag{13}$$

Proof: see Appendix B.4.

Corollary 2.2 shows that the convergence rate of the OWL estimator is the same as the LASSO estimator if we restrict the rate of λ_2 is slower than λ_1 as specified in Theorem 2.2. Furthermore, with Assumption 2, Corollary 2.2 also implies that the OWL estimator is consistent.

2.4 Discussion on the cross-sectional asset returns and the factor zoo

Cochrane (2011) brought up the "factor zoo" enigma. Since then, it has attracted considerable attention and spurred development of methodological contributions to dissect the factor zoo for cross-sectional asset returns. Green et al. (2017) employed the Fama-MacBeth two-step regression method to select factors from the "factor zoo" for US stock returns. They removed some factors (i.e., the 'beta' related factors) before conducting their analysis, as highly correlated factors would dampen the robustness of the inference in the Fama-Macbeth framework. However, such ad hoc treatment of screening out correlated factors is not a trivial task - a vigorous discussion on the criteria used for such a procedure is much needed. On the other hand, the fast-evolving technology developed in Statistic and Machine Learning literature has shed light on new methods which can be utilized for dissecting the factor zoo. Feng et al. (2020) employed the double-LASSO selection procedure devised by Belloni et al. (2014) to recursively test factors for driving cross-sectional asset returns. The double-LASSO selection method is devised to test a small number of factors which are of interest to economists while having a large number of controlling factors. Belloni et al. (2014) show that the double LASSO selection method can overcome the omitted variable bias. Having said that, our paper is closely related to Feng et al. (2020) in terms of the research question (i.e., the factor zoo and cross-sectional asset returns). However, the focus of the research question and the methods utilised are drastically different. First, the focus of our paper is correlation-robust selection: the OWL shrinkage method will assign similar coefficients to factors if they are highly correlated, while the LASSO shrinkage method will likely shrink some highly correlated factors to zeros while keeping others as non-zeros, resulting in unstable factor selections when factors are highly correlated. Second, conducting statistical testing is the focus in Feng et al. (2020): the double LASSO selection method is employed to test the significance of factors that were proposed in a specific calendar year while having all previously proposed factors as controlling factors. They conduct such tests recursively, based on each calendar year. On the other hand, conducting statistical testing is beyond the scope of this paper. Note that the OWL estimator is biased in small samples. To conduct statistical tests for the OWL shrinkage method, a debiased version of the OWL estimator and its asymptotic properties need to be developed, which can be addressed in a future research agenda. Third, since the double LASSO selection procedure only conducts test on a small number of factors (i.e., it does not make inference on the large number of controlling factors), we need to hand-pick a small subset of factors to be tested (in the case of Feng et al. (2020) they use each calendar year as the criterion to form such a subset). On the other hand, the OWL shrinkage method is not restricted by such constraints.

3 Simulation

This section studies the performance of the OWL estimator together with other benchmark estimators in various Monte Carlo simulation experiments.

3.1 Simulation design

In our experiment, we consider K candidate factors, $2K/3$ of them are useful factors, that is they are priced ($b \neq 0$), and $K/3$ of them are useless or redundant factors ($b = 0$). Within these useful factors, we set half of them (1/3 of total factors) are correlated, while the remaining half are uncorrelated. In this setting, we include correlated factors, uncorrelated factors and useless factors in our models.

Let ρ denote the $K \times K$ correlation coefficient matrix of C ($N \times K$) defined in (3). We suppose that $\rho_1, \rho_2, \rho_3 \in (-1, 1)$ and ρ is divided into 3 blocks such that:

$$bk_1 = \underbrace{\begin{pmatrix} 1 & \dots & \rho_1 \\ \vdots & \ddots & \vdots \\ \rho_1 & \dots & 1 \end{pmatrix}}_{K/3}; \quad bk_2 = \underbrace{\begin{pmatrix} 1 & \dots & \rho_2 \\ \vdots & \ddots & \vdots \\ \rho_2 & \dots & 1 \end{pmatrix}}_{K/3}; \quad bk_3 = \underbrace{\begin{pmatrix} 1 & \dots & \rho_3 \\ \vdots & \ddots & \vdots \\ \rho_3 & \dots & 1 \end{pmatrix}}_{K/3}$$

and

$$\rho = \begin{pmatrix} bk_1 & & \text{\Large 0} \\ & bk_2 & \\ \text{\Large 0} & & bk_3 \end{pmatrix}.$$

In the block bk_1 (block 1) the diagonal elements are ones and off-diagonal elements are ρ_1; similarly for the block bk_2 and bk_3 where off-diagonal elements are ρ_2 and ρ_3, respectively. These three blocks constitute the diagonal direction of matrix ρ, and elsewhere ρ is filled with zeros. This setting allows three blocks of factors. Within each block, factors are correlated with a correlation coefficient ρ_1, ρ_2 or ρ_3, but factors in different blocks are uncorrelated.

We first set the values of ρ_1, ρ_2 and ρ_3, and then randomly generate an $N \times K$ matrix C, denoted as $simC$, which has the correlation coefficient matrix of ρ. We use block 3 to represent uncorrelated useful factors, thus ρ_3 is set to be zero in our experiments. We consider different correlation coefficient values for ρ_1 and ρ_2. Then, we specify the oracle value for b (risk price) before simulating the cross section of average returns as $\mu_R = simC * b + e$, where e is the pricing error. We use block 2 to represent useless factors, therefore the oracle value of b in block 2 are set to be zeros. Finally, we estimate risk price with simulated data $simC$ and μ_R using OWL, LASSO, adaptive LASSO, Elastic Net, and naive OLS estimators.[12] Then we compare these estimates with the oracle value of b.

3.2 Simulation results

We consider 90 candidate factors ($K = 90$). We set the first block of 30 factors as useful factors ($b = 0.1$) and they are correlated with correlation coefficient ρ_1; the second block of 30 factors are useless/redundant factors ($b = 0$) and they are correlated with correlation coefficient ρ_2; the third block of 30 factors (block 3) are uncorrelated useful factors (we set $b = -0.1$ and $\rho_3 = 0$). For simplicity, we set $\rho_1 = \rho_2$ and they are chosen from the set $\{0.3,\ 0.5,\ 0.9\ \}$, which gives various profiles of correlation structure between factors. We also consider how the variation of N (the number of assets) compared to K (the number of factors, which is set to be 90 in our experiments) affects model comparisons. To do that, we choose N from the set $\{70,\ 100,\ 1000\ \}$. When $N \gg K$ it represents a near-asymptotic setting. On the other hand, if $N \sim K$ or $N < K$, it approximate a setting where the number of factors can be larger than the number of observations - a common characteristic shared in high-dimensional data-sets. We run 500 trails in our simulation experiment and we use the mean squared estimation error (MSE) of candidate models as our comparison criterion. To fix ideas, for the i^{th} model, MSE for all factors is defined as follows

$$MSE^i = \frac{1}{500K} \sum_{rep=1}^{500} \|\hat{b}^i_{rep} - b^0\|_2,$$

where rep is the index for the trail in our simulation experiment. To better under the performance of candidate models under various settings, we look into the MSE for each blocks, such that

$$MSE^i_{bk1} = \frac{1}{500 * 30} \sum_{rep=1}^{500} \|\hat{b}^i_{bk1,rep} - b^0_{bk1}\|_2,$$

Similarly, MSE for block 2 and block 3 factors can be defined accordingly.

[12] See Appendix G for an introduction to LASSO, adaptive LASSO, and Elastic Net (EN) estimators. The OLS estimator is comparable to the Fama-MacBeth regression method. It is only included as a benchmark if $N > K$.

Table 1. Simulation results

		Useful factors in bk1				Useless factors in bk2				All factors			
		OWL	LASSO	adaLASSO	EN	OWL	LASSO	adaLASSO	EN	OWL	LASSO	adaLASSO	EN
N=70	rho = 0.3	0.0008	0.0047	0.0072	0.0037	0.0006	0.0005	0.0003	0.0005	0.0008	0.0035	0.0049	0.0030
	rho = 0.5	0.0006	0.0054	0.0080	0.0041	0.0005	0.0004	0.0002	0.0003	0.0006	0.0036	0.0051	0.0029
	rho = 0.9	0.0002	0.0100	0.0120	0.0057	0.0002	0.0000	0.0000	0.0000	0.0002	0.0043	0.0058	0.0027
N = 100	rho = 0.3	0.0001	0.0007	0.0008	0.0005	0.0001	0.0001	0.0000	0.0001	0.0001	0.0006	0.0005	0.0005
	rho = 0.5	0.0001	0.0010	0.0011	0.0008	0.0001	0.0001	0.0000	0.0001	0.0001	0.0007	0.0006	0.0006
	rho = 0.9	0.0001	0.0040	0.0038	0.0024	0.0001	0.0000	0.0015	0.0000	0.0001	0.0017	0.0020	0.0011
N = 1000	rho = 0.3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	rho = 0.5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	rho = 0.9	0.0000	0.0001	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0001

Note: this table reports the Mean Squared Error (MSE) from candidate models. The number of factors K is 90. We consider three different settings of N: $N = 70$ high-dimensional setting where $N < K$; $N = 100$, where N is marginally larger than K and $N = 1000$, low-dimensional setting where $N \gg K$. We also consider different correlation setting between factors which is indicated by the parameter ρ. This table reports the MSE for the useful factors in block 1, useless factors in block 2 and all factors and MSE for all factors. Note that MSE for factors in block 3 is not reported but they can be inferred from this table.

Table 1 reports the simulation results by comparing four candidate models including the OWL, the LASSO, the adaptive LASSO and the Elastic Net estimators.[13] For the first and second block of factors we allow correlation coefficients between factors to vary $\rho \in \{0.3, 0.5, 0.9\}$, whereas the third block of factors are set to be uncorrelated. We also consider three different settings for the value of N. When $N = 70$, it represents an environment when the number of observations is smaller than the number of factors, which typically resembles the characteristics of a high-dimensional setting. When $N = 100$, it represents an environment where the number of observations is about the same size with the number of factors. When $N = 1000$ it represents an ideal setting where the number of observation is much larger than the number of factors, which approximates a near-asymptotic setting. The left panel reports the MSE of the first block of factors which are useful factors with correlation specified in the table. The middle panel reports the second block of factors which are useless factors. The right panel reports the MSE for all factors. Note that the third block of factors are not reported, but it can be inferred from the table.

[13] Note that the LASSO and EN models are evaluated using a 3-fold cross-validation method for their tuning parameters. However, we set $\lambda_1 = \lambda_2 = 10^{-6}$ as fixed to ease the computational burden. It implies that the OWL estimator potentially can perform better if we use cross-validation method to find optimal tuning parameters for each simulated trials. Furthermore, for the adaptive LASSO we use the OLS estimate as the adaptive weights when $N \geq K$, while we use the LASSO estimate as the adaptive weights when $N < K$, since OLS estimation would be infeasible in this case.

First, we investigate the performance of correlated useful factors (i.e. factors in block 1) as well as "all factors". Table 1 shows that when $N = 70$ (which approximate a high-dimensional setting where $K > N$) and when $N = 100$ (i.e. $N \sim K$), we find that the OWL estimator achieves the smallest MSE for the correlated useful factors compared to other benchmarks. Furthermore, we find that other benchmarks such as LASSO, adaptive LASSO and EN, their performance deteriorate when factor correlation increases and by contrast, the OWL estimator is less affected. The performance of MSE for all factors also suggests that the OWL estimator outperforms other benchmarks yielding the smallest MSEs for all settings for ρ. When $N = 1000$ (i.e. $N \gg K$ which approximates the near-asymptotic setting), the OWL estimator achieves MSE close to zero in all settings, which confirms the theoretical result in Corollary 2.2 that the OWL estimator is a consistent estimator. Although other benchmarks are also consistent estimators, we find that the OWL estimator achieves smaller MSE compared to LASSO and EN when $\rho = 0.9$, suggesting that the OWL estimator consistently outperforms the LASSO and EN when factors are highly correlated.

It is worth noting that for factors in block 2, the OWL estimator is less effective to shrink off useless factors compared to the LASSO and EN estimators, with MSE marginally larger than LASSO and EN in the $N = 70$ and $N = 100$ settings. On the other hand, when $N = 1000$ (at the near-asymptotic setting) the OWL estimator, along with all other shrinkage estimators, successfully shrink off all useless factors. [14]

These findings suggest that, when factors are correlated, the OWL estimator is the preferred estimator especially in the high-dimensional setting. The performance of the LASSO estimator deteriorates when factor correlation increases. The Elastic Net model does improve the performance of the LASSO model when factor correlation increases, but it is substantially outperformed by the OWL estimator.

Table 1 summarizes the performance of four candidate models under various settings by comparing their average MSEs under 500 trials. To have a better view on how those candidate models perform for each block of factors, we randomly chose one trail and plot the estimates from candidate models along with the oracle value. We present our results

[14] This under-performance of the OWL estimator to shrink off useless factors in small samples can be overcome by applying thresholding (i.e. set small estimates to zero according to a validated threshold level) or by applying statistical test utilising the de-biased estimator and its asymptotic properties. However, those subjects are beyond the scope of this paper, which can be left in the future research agenda.

and analysis in Appendix E.

4 Empirical analysis

In this section we employ the OWL shrinkage method to find useful factors among 80 (potentially highly correlated) factors. We first introduce the data-sets, followed by a detailed account of constructing anomaly factors and test portfolios using portfolio sorting. Then, we reveal the estimation results and discuss their implications.

4.1 Data

We use the U.S. stock data from the Center for Research in Security Prices (CRSP) and Compustat database[15] to construct anomaly factors and test portfolios. The period spans from January 1980 to December 2017, totalling 456 months on all NYSE, AMEX and NASDAQ listed common stocks. Risk-free rate and market excess returns are downloaded from Kenneth French's on-line data library. All anomaly factors are demeaned and scaled to have the same standard deviation with the market factor.

4.2 Constructing the factor zoo

We consider 100 firm characteristics described in Green et al. (2017),[16] while deleting characteristics that have more than 40% missing data. Then, for each remaining characteristic, we sort stocks into decile portfolios at each month using uni-variate sorting. Micro stocks, defined as having market capitalization smaller than the 20 percentile of NYSE listed stocks, are removed.[17] Then, a characteristic-based factor is computed as the spread returns between the top and the bottom decile portfolios with respect to each firm characteristic after screening. [18] Overall, we obtain 80 anomaly factors which are listed in Table 2. See Green

[15] CRSP and Compustat data are downloaded from the Wharton Research Data Services.

[16] We are grateful to Jeremiah Green for providing SAS code to compute firm characteristics.

[17] Although micro stocks only account for less than 10% of aggregated market capitalization, they constitute about 56% of all stocks in the database, implying that small stocks would distort the interpretation of the aggregated market capitalization if not removed, also see Hou et al. (2014) and Fama and French (2016) for a similar treatment.

[18] Characteristics that have insufficient data to construct decile portfolios at every month will be dropped. Note that the sorting is always from high to low according to characteristics, and the factors are computed

et al. (2017) for a detailed description of each characteristic.

Next, we conduct a preliminary analysis by checking correlations between factors. Figure 1a displays the heat map of factor correlation coefficients matrix measured by their time series. It shows that 16% of factors exhibit correlation coefficients (absolute value) greater than 0.5. In particular, 'beta' related factors are highly correlated with 'liquidity', 'profitability', 'investment' and other factors. For that reason, Green et al. (2017) exclude 'beta' related factors from candidate factors before they employ the Fama-MacBeth method to find significant factors that drives the cross-sectional returns for the US stock market. Figure 1b displays the heat map of factor correlation coefficients matrix measured by factor loadings (i.e. the correlation coefficients of explanatory variables for the second stage Fama-MacBeth regression). It exhibits much higher correlation compared to Figure 1a: 64% correlation coefficients (absolute value) are greater than 0.5, implying a serious multicollinearity issue if the standard Fama-MacBeth regression is employed.

This preliminary examination of factor correlations shows that many factors are highly correlated, suggesting severe complications would occur if traditional methods such as the Fama-MacBeth regression procedure or the LASSO regression model are employed to infer useful factors. Therefore, a correlation-robust estimation method is much needed.

4.3 Constructing test assets

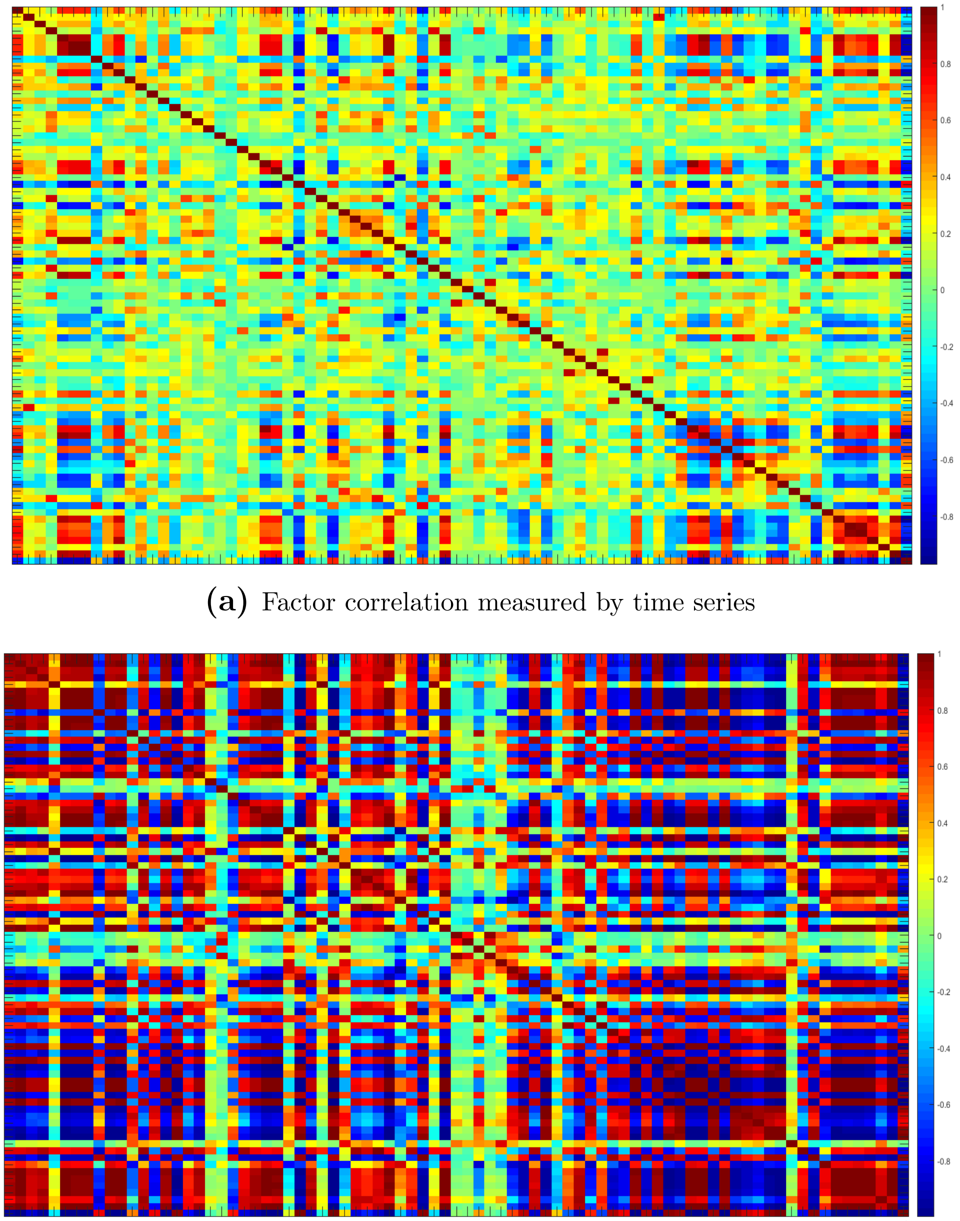
Regarding test assets, there is a debate in the literature about using either individual stocks or sorted portfolios as test assets. The main concern in the literature of using individual stocks as test assets is that it will introduce errors in variables (EIV). When regression is made on estimated variables, i.e. factor loadings, the pre-estimated factor loadings would incur estimation errors. Shanken (1992) modified the estimator by introducing the "Shanken's correction" term to mitigate EIV. However, others argue that "Shanken's correction" is minimal in small samples. On the other hand, Fama and French (2008), Hou et al. (2014), Feng et al. (2020) advocate sorted portfolios as test assets. Individual

as the top decile return minus the bottom decile return. That will end up with some differences compared to some familiar notations. For instance, the famous size factor 'small-minus-big' in our factor library would be 'big-minus-small', however, they are essentially the same after giving a negative sign. In estimation, we only care about the coefficient magnitude (i.e., the absolute value). The interpretation of the sign of coefficients should be looked at together with the sorting order when forming anomaly variables.

Table 2. Anomaly factors

This table lists all 80 factors considered in our factor library. The abbreviation is consistent with Green et al. (2017). For a more detailed description of each factor, including the original paper where it is proposed, please refer to Green et al. (2017).

Abbreviation	**Firm Characteristics**	**Abbreviation**	**Firm Characteristics**
’absacc’	absolute accruals	’mom1m’	1 month momentum
’acc’	working capital accruals	’mom36m’	36 month momentum
’aeavol’	abnormal earnings announcement volume	’mom6m’	6 month momentum
’agr’	asset growth	’ms’	financial statement score
’baspread’	bid-ask spread	’mve’	size
’beta’	beta	’mve_ia’	industry adjusted size
’betasq’	beta squared	’nincr’	number of earnings increases
’bm’	book-to-market	’operprof’	operating profitability
’bm_ia’	industry adjusted book-to-market	’pchcapx_ia’	i.a. %change in capital expenditures
’cash’	cash holding	’pchcurrat’	% change in current ratio
’cashdebt’	cash flow to debt	’pchdepr’	% change in depreciation
’cashpr’	cash productivity	’pchgm_pchsale’	% change in gross margin - %change in sales
’cfp’	cash flow to price ratio	’pchquick’	%change in quick ratio
’cfp_ia’	industry adjusted cfp	’pchsale_pchinvt’	% change in sale - % change in inventory
’chatoia’	industry adjusted change in asset turnover	’pchsale_pchrect’	% change in sale - % change in A/R
’chcsho’	change in share outstanding	’pchsale_pchxsga’	% change in sale - % change in SG&A
’chempia’	industry adjusted change in employees	’pchsaleinv’	% change in sales-to-inventory
’chinv’	change in inventory	’pctacc’	percent accruals
’chmom’	change in 6-month momentum	’pricedelay’	price delay
’chpmia’	industry adjusted change in profit margin	’ps’	financial statement score
’chtx’	change in tax expense	’quick’	quick ratio
’cinvest’	corporate investment	’retvol’	return volatility
’currat’	current ratio	’roaq’	return on assets
’depr’	depreciation	’roavol’	earning volatility
’dolvol’	dollar trading volume	’roeq’	return on equity
’dy’	dividend to price	’roic’	return on invested capital
’ear’	earnings announcement return	’rsup’	revenue surprise
’egr’	growth in common shareholder equity	’salecash’	sales to cash
’ep’	earnings to price	’saleinv’	sales to inventory
’gma’	gross profitability	’salerec’	sales to receivables
’grcapx’	growth in capital expenditure	’sgr’	sales growth
’grltnoa’	growth in long term net operating assets	’sp’	sales to price
’hire’	employee growth rate	’std_dolvol’	volatility of liquidity (dollar trading volume)
’idiovol’	idiosyncratic return volatility	’std_turn’	volatility of liquidity (share turnover)
’ill’	illiquidity	’stdacc’	accrual volatility
’invest’	capital expenditure and inventory	’stdcf’	cash flow volatility
’lev’	leverage	’tang’	debt capacity/firm tangibility
’lgr’	growth in long term debt	’tb’	Tax income to book income
’maxret’	max daily return	’turn’	share turnover
’mom12m’	12 month momentum	’zerotrade’	zero trading days



(a) Factor correlation measured by time series

(b) Factor correlation measured by factor loadings

Figure 1. Factor correlation coefficients

This heat map displays the matrix of correlation coefficients of all 80 anomaly factors. Dark red and deep blue colors signal high correlation (positive or negative) while light colours indicate low correlation. There are N test assets and K factors, each asset/factor has T time series observations. “Factor correlation measured by time series” means the correlation coefficients matrix is computed through the $T \times K$ factor time series data. “Factor correlation measured by factor loadings” means the correlation coefficients matrix is computed through the $N \times K$ factor loadings after the first stage of Fama-MacBeth regression.

stocks are usually noisy and exhibit outliers, which are the main source of EIV. Sorted portfolios are (weighted) mean returns of a group of stocks sharing similar characteristics, which would be less affected by the EIV problem.

Nonetheless, the biggest drawbacks of using individual stocks stem from missing data and micro stocks. It is inevitable, over a long period of time, to have new firms entering and old firms exiting, which frequently causes missing data in our data-sets. Discontinuity of data leads to imprecise estimation of the covariance matrix of returns and factors, which is essential for factor inference. On the other hand, sorted portfolios are constructed at each point of time while sorting (possibly varying) stocks that share similar characteristics into portfolios, guaranteeing that they are immune to the missing data problem.

Small stocks bring up another major concern of using individual stocks as test assets. Small stocks take up the majority of all stocks, while only a few big stocks constitute a large share of total market capitalization. If individual stocks are used to gauge factor impact, it is inevitable that they will distort the market implications: small stocks will dominate the estimation result if individual stocks are used for test assets - big stocks which have much larger impact on the market capitalization will be out-weighed by the large number of small stocks. Sorted portfolios, on the other hand, can circumvent this issue by using the value weighted sorting method, in which portfolio returns are computed by the weighted average of stocks returns where the weights reflect their market capitalization.

Fama and French (1992, 2016) use bi-variate sorting to create the five by five test portfolios which have now become popular choices for test assets. However, Harvey et al. (2015) caution that when only a small set of sorted portfolios are considered for test assets, factor selection is biased towards the same characteristics that are used to form test portfolios. Lewellen et al. (2010) argue that the 25 size and value sorted portfolios are too low a threshold to test factors. They recommend adding other portfolios in test assets. Following their advice, Feng et al. (2020) construct a large set of combined bi-variate sorted portfolios as test assets. In particular, they use the 'size' characteristic and the other remaining characteristics to form five by five bi-variate sorted portfolios and pool them together as the grand set of test assets. We follow Feng et al. (2020) to construct bi-variate sorted test portfolios and we obtain 1927 test portfolios as our grand set of test assets.[19]

4.4 Estimation results: which factors matter?

We use the SDF method described in Section 2.2 to estimate the risk prices for all factors using the OWL shrinkage method. We use sample analogs of C and μ_R in our estimation. Specifically, $\hat{C} = \widehat{\text{Cov}(R, f)} = \frac{1}{T}\sum_{t=1}^{T}(R_t - \hat{\mu}_R)(f_t - \hat{\mu}_f)'$, $\hat{\mu}_f = \frac{1}{T}\sum_{t=1}^{T} f_t$ and $\hat{\mu}_R = \frac{1}{T}\sum_{t=1}^{T} R_t$. For robust estimation, we look into the following cases: first, we consider the full sample estimation as well as its sub-samples, to check the time-varying treand of factor selection. Second, we compare different weighting methods (equal weighted or value weighted) for sorting portfolios and investigate their impact on the estimation results - value weighted portfolios would put more weights on large stocks whereas equal weighted portfolios are dominated by small stocks. Third, we use different percentile (20, 30 and 40 percentiles, respectively) to remove micro stocks before sorting portfolios. A larger percentile used to remove micro stocks means larger proportion of small stocks are removed before sorting portfolios, resulting in factor selection that is more influenced by larger stocks.

Table 3 reports the estimation results. The first 5 columns are estimated using the full sample, columns 6-7 report results of the first half sample, from 1980 to 2000, and columns 8-9 reports the second half sample, from 2001 to 2017. Both the value weighted (vw) and equal weighted (ew) methods are considered. Also, three levels to partition micro stocks are considered. This table lists all anomaly factors selected in each estimation. It also reports how many times each factor has been selected by all estimations and the ordinal number (in the bracket) for each factor in a separate estimation, which indicates the importance of the factor (smaller number implies greater importance).[20]

Table 3 shows that ‘size’ (mve) factor has been selected as the most important factor in most of these estimations which, however, is not surprising. ‘Size’ characteristic has multiple entries in forming test portfolios, thus ‘size’ impact prevails in test portfolios. For this reason we exclude ‘size’ factor as a competing factor, yet we include it in the table to

[19] We drop any test portfolios which have insufficient stocks to sort at any time, due to missing data.

[20] It is worth stressing that the factor selection in each estimation does not imply a true parsimonious asset pricing model, as such an implication would require statistical tests on factors. Instead, we are focusing on the important factors selected by the OWL estimator - since all factors are scaled to have the same mean and variance, their estimated coefficient can be interpreted as the importance of factors.

Table 3. Estimation results

This table reports the selected factors using the OWL shrinkage method. We consider the full sample from 1980 to 2017 and two sub samples divided by year 2000. Equal weighted (ew) and valued weighted (vw) sorting methods are both considered. Three treatments of micro stocks are considered: we remove stocks that are smaller than 20 (30 or 40) percentile of NYSE listed stocks. For each combination of the sample size, weighting method and micro-stock treatment, we list all selected factors with the ordinal numbers in the bracket (smaller means more important).

Sample size		full	full	full	full	full	1980:2000	1980:2000	2001:2017	2001:2017
Weighting		vw	vw	vw	ew	ew	vw	vw	vw	vw
Micro stock		20 prctile	30 prctile	40 prctile	20 prctile	40 prctile	20 prctile	40 prctile	20 prctile	40 prctile
	# selected									
agr	5		agr (8)	agr (8)	agr (5)	agr (4)	agr (5)			
baspread	2	baspread (7)								baspread (4)
beta	2					beta (1)				beta (1)
betasq	3				betasq (4)	betasq (2)				betasq (2)
cash	3	cash (6)	cash (7)				cash (6)			
cashdebt	4		cashdebt (6)	cashdebt (2)	cashdebt (7)			cashdebt (2)		
dolvol	3			dolvol (10)	dolvol (6)	dolvol (6)				
egr	3		egr (4)	egr (3)				egr (9)		
ill	7	ill (2)	ill (2)	ill (6)	ill (2)	ill (5)			ill (2)	ill (6)
invest	2						invest (7)	invest (10)		
mom12m	1							mom12m (3)		
mom6m	2						mom6m (1)	mom6m (4)		
mve	8	mve (1)	mve (1)	mve (1)	mve (1)	mve (3)		mve (1)	mve (1)	mve (5)
pchcapx_ia	1			pchcapx_ia (5)						
pchcurrat	4	pchcurrat (4)	pchcurrat (3)	pchcurrat (9)			pchcurrat (4)			
pchquick	2			pchquick (11)					pchquick (4)	
retvol	1									retvol (3)
roaq	2						roaq (2)			roaq (7)
roic	3	roic (5)		roic (7)					roic (5)	
salecash	1						salecash (3)			
saleinv	1							saleinv (5)		
sp	1							sp (6)		
std_dolvol	6	std_dolvol (3)	std_dolvol (5)	std_dolvol (4)	std_dolvol (3)	std_dolvol (7)			std_dolvol (3)	
stdcf	1							stdcf (7)		
turn	1							turn (8)		

show that the OWL estimator can correctly identify relevant factors.

The 'illiquidity' (ill) factor (Amihud, 2002) is the most important anomaly factor for cross-sectional returns. Its importance is particularly evident with smaller stocks. Portfolios sorted with size greater than 20 or 30 percentile of NYSE listed stocks exhibit higher importance of 'illiquidity' than those with 40 percentile, implying that small firms face severer liquidity constraints. 'Standard deviation of dollar volume' (std_dolvol) (Chordia et al., 2001) which is another proxy for liquidity risk, follows 'illiquidity', becoming the second most important anomaly factor. Meanwhile, its high correlation with 'illiquidity' is also identified by the OWL estimator. Liquidity as a risk source that influences cross-sectional asset returns has been documented extensively in the literature, see Pastor and Stambaugh (2003) and Acharya and Pedersen (2005) for example.

'Asset growth rate' (agr) follows 'illiquidity' and 'standard deviation of dollar volume' as the third most frequently selected anomaly factor. This finding coincides with Hou et al. (2021) in which they propose an augmented $q5$ model, adding 'asset growth rate' as a fifth factor into their well celebrated $q4$ factor model (Hou et al., 2014). Other anomaly factors that have been selected multiple times include 'beta', 'beta squared' (betasq), 'cash to debt ratio', and 'percentage change in current ratio' (pchcurrat), which are also correlated with liquidity risks. Beyond that, 'momentum', 'return on invested capital' (roic), 'return on assets' (roaq) and other profitability related factors are also selected by the OWL estimator as useful factors, particularly for the first half of the sample period.

Columns 6 and 7 report estimations using the 1980 - 2000 sub-sample and columns 8 and 9 report estimations using the 2001 - 2017 sub-sample. We find that liquidity constraint only appears in the second sub-sample (2001 - 2017), where liquidity related factors ('baspread', 'standard deviation of dollar volume', 'change in quick ratio', etc...) play an important role in explaining the cross section of average returns. However, in the first sub-sample (1980 - 2000), columns 6 and 7 show no strong evidence that liquidity related factors drive asset prices. On the other hand, 'momentum' and 'profitability' factors are the most important ones to drive asset prices between 1980 and 2000. This implies a time-varying trend of factors that dictates cross-sectional returns.

Interestingly, from 1980 to 2000, with 20-percentile-micro-stocks excluded, we find 'size' (mve) is not selected as an useful factor to drive cross-sectional returns, which makes it the only exception from all estimations. This phenomenon is well documented in the literature, see Amihud (2002) and Asness et al. (2018) for example. The size effect weakened after its discovery in the early 1980s. However, our estimation results suggest that the size effect is evident during this period after removing 40-percentile-micro-stocks, which implies that the vanishing size effect is likely to be caused by some small "junk" stocks. Once removing these junk stocks, size effect resurfaces again, which echoes the discovery by Asness et al. (2018): *size matters, if you control your junk.* Asness et al. (2018) shows that when adding some controlling variables (measuring the junk-level of stocks) in their regression model, they find that the size-effect is significant even in the early 1980s.

4.5 Robustness check

In this section, we check whether, and to what extend, different sorting methods and different treatments of micro stocks would affect our estimation results and factor selections. Because of the limitation of space, we place Section 4.5 in Appendix F.

4.6 Out-Of-Sample analysis

In this section, we use the selected factors to construct hedging portfolios in an out-of-sample framework following Freyberger et al. (2020). This exercise aims to compare the OWL shrinkage method with other benchmarks by evaluating their hedging portfolio returns (based on different factors selected by each model). Apart from the OWL shrinkage method, we consider three other benchmarks, including the LASSO shrinkage method, the Elastic Net model, and the two-pass Fama-MacBeth (FM) procedure. To enable fair comparison, we consider a five-factor model and select only the five most important factors determined by each method to form hedging portfolios.

To offer some insights on the time-varying trend in factor selection, we also consider two sub-samples, divided before and after 2000. We report the five most important factors selected by each method.[21]

[21] For robustness check, we also experimented on a three-factor and a four-factor model for out-of-sample prediction. We find that conclusions are similar.

It is worth noting that we put the market factor (mkt) together with 80 anomaly factors as the grand set of factors we can choose from. Because the market factor is often highly correlated with many anomaly factors, its importance, though backed by financial theory, is often compromised by the high-correlation with other factors. Therefore, it is frequently deemed as an unimportant factor when using traditional methods. We further investigate its implications in the following text.

Table 4 reports the five most important factors selected using various methods in different samples while controlling micro stocks. We find that selected factors vary substantially between different sample periods, signalling time-varying trend in prominent factors that drive cross-sectional asset returns. In addition, controlling micro stocks has a big impact on factor selection too. While including all micro stocks (P00), all methods select a mixture of 'liquidity', 'profitability' and 'momentum' related factors. However, once we remove micro stocks (P20 and P40), we can find some patterns in selected factors: OWL suggests that the most important factors to drive asset prices in the first sub-sample are 'momentum' and 'profitability' related factors while 'liquidity' related factors are relatively unimportant. However, the implication is reversed in the second sub-sample, where 'liquidity' related factors mainly drive asset prices. On the other hand, LASSO and other methods do not show a clear pattern of change in characteristics.

Interestingly, we find that the OWL shrinkage method is the only method that consistently identifies the market factor as an important factor to drive cross-sectional asset returns, especially when micro stocks are removed before portfolio sorting. This finding is consistent with numerous finance literature related to the CAPM model of Lintner (1965) and Sharpe (1964). On the contrary, other benchmarks such as LASSO, Elastic Net and Fama-MacBeth estimators all fail to identify the market factor as an important factor to drive cross-sectional returns even when 40 percentile of micro stocks are removed. As discussed before, this is caused by high correlation between the market factor and other factors. This finding reiterates the merit of using the OWL shrinkage method when factors are highly correlated.

Next, we want to compare the out-of-sample performance between various methods. In particular, we follow a similar procedure to Freyberger et al. (2020) to form factor-hedged

Table 4. Full/sub-sample factor selection using various methods

This table reports the first five factors selected with greatest magnitude of $\hat{b}$ using methods including OWL, LASSO, Elastic Net (EN), and two-pass Fama-MacBeth regression (FM). We do factor selection either on the full sample (full) or two sub-samples, divided by year 2000 (sub1 and sub2). We also control micro stocks: we consider all stocks (P00), or remove micro stocks' market capitalisation which is smaller than 20/40 percentile of NYSE listed stocks (P20/P40).

		First five selected factors (decreasingly) ordered by their magnitude of $\hat{b}$				
Panel A: Full sample estimation						
full_P00	OWL	’ill’	’mve’	’cash’	’chpmia’	’roeq’
	LASSO	’idiovol’	’mve’	’mom6m’	’zerotrade’	’operprof’
	EN	’idiovol’	’mve’	’mom6m’	’ill’	’pctacc’
	FM	’idiovol’	’maxret’	’ill’	’betasq’	’beta’
full_P20	OWL	’mve’	’ill’	’mkt’	’std_dolvol’	’pchcurrat’
	LASSO	’idiovol’	’mve’	’ill’	’mom36m’	’ms’
	EN	’mve’	’idiovol’	’ill’	’mom36m’	’bm’
	FM	’idiovol’	’baspread’	’ill’	’beta’	’betasq’
full_P40	OWL	’mkt’	’mve’	’cashdebt’	’egr’	’std_dolvol’
	LASSO	’mve’	’idiovol’	’ill’	’operprof’	’roavol’
	EN	’mve’	’idiovol’	’ill’	’operprof’	’mkt’
	FM	’idiovol’	’baspread’	’ill’	’betasq’	’beta’
Panel B: sub-sample estimation (1980:2000)						
sub1_P00	OWL	’pchcurrat’	’sp’	’bm’	’mkt’	’absacc’
	LASSO	’dy’	’turn’	’acc’	’mve’	’sp’
	EN	’dy’	’turn’	’acc’	’mve’	’ill’
	FM	’maxret’	’retvol’	’idiovol’	’betasq’	’mom1m’
sub1_P20	OWL	’mkt’	’mom6m’	’roaq’	’salecash’	’pchcurrat’
	LASSO	’baspread’	’dy’	’gma’	’mve’	’ill’
	EN	’baspread’	’dy’	’gma’	’mve’	’ill’
	FM	’idiovol’	’betasq’	’beta’	’ep’	’baspread’
sub1_P40	OWL	’mkt’	’mve’	’cashdebt’	’mom12m’	’mom6m’
	LASSO	’mve’	’mve_ia’	’std_turn’	’invest’	’turn’
	EN	’mve’	’mve_ia’	’std_turn’	’invest’	’turn’
	FM	’idiovol’	’beta’	’betasq’	’baspread’	’retvol’
Panel C: sub-sample estimation (2001:2017)						
sub2_P00	OWL	’ill’	’mve’	’cash’	’mkt’	’roeq’
	LASSO	’mve’	’ill’	’stdacc’	’gma’	’pctacc’
	EN	’mve’	’ill’	’pctacc’	’stdacc’	’agr’
	FM	’ill’	’idiovol’	’dolvol’	’baspread’	’std_dolvol’
sub2_P20	OWL	’mve’	’ill’	’mkt’	’std_dolvol’	’pchquick’
	LASSO	’mve’	’pchquick’	’idiovol’	’ill’	’pchcurrat’
	EN	’mve’	’pchquick’	’ill’	’idiovol’	’pchcurrat’
	FM	’ill’	’baspread’	’idiovol’	’std_dolvol’	’dolvol’
sub2_P40	OWL	’mkt’	’beta’	’betasq’	’retvol’	’baspread’
	LASSO	’mve’	’ill’	’roavol’	’tang’	’pchquick’
	EN	’mve’	’ill’	’sgr’	’pchquick’	’salerec’
	FM	’idiovol’	’baspread’	’ill’	’betasq’	’beta’

portfolios using a rolling window scheme to predict returns. First of all, we choose five most prominent factors as shown in Table 4 while considering different samples and different treatment of micro stocks (i.e. remove micro stocks at the 20- and 40-percentile levels). Then we use a rolling window scheme (rolling window size is 120 months) to evaluate the performance of the factor-hedged portfolios with each method. Specifically, at the end of each estimation window, we regress each test asset on factors selected by each method, but with one period lagged behind. For instance, at time t, we regress each test asset return from $t-120-1$ to t on selected factors from $t-120-2$ to $t-1$, and obtain $\hat{\beta}$. We then forecast each test asset's next period return (at $t+1$) by multiplying $\hat{\beta}$ and selected factors at t. We then sort stocks by their predicted returns into decile portfolios and long the top decile and short the bottom decile. At the next period $(t+1)$, when returns are realized, we can compute the spread portfolio return. Subsequently, we roll the window one period forward and repeat the steps until the end of period. In the end we compute four moments of the factor-hedged portfolio returns in the out-of-sample period as well as the Sharpe ratio.

Table 5 reports performance scores including the Sharpe ratios and the four moments of out-of-sample returns of hedged portfolios using the OWL, LASSO, Elastic Net and Fama-MacBeth methods. Panel A suggests that in the full sample estimation, the OWL estimator produces about 20% higher Sharpe ratio than other benchmarks. In addition, we find that the skewness and the kurtosis of the OWL hedged portfolio are much smaller those of other benchmarks. Fama-MacBeth estimator typically performs the worst. We reckon that it is severely affected by factor correlations and estimation result is eroded by weak factors in the second pass Fama-MacBeth regression, see Kleibergen (2009) for a detailed discussion on this matter.

In sub-sample estimations, we find that the Sharpe ratios of the factor-hedged portfolios are typically much higher than that of the full-sample estimation in all methods we considered. Furthermore, we find that the skewness and the kurtosis of hedged portfolio returns are significantly reduced compared to the full-sample estimation, making the distribution of the out-of-sample returns more "normal" alike. This trend signals a time-varying nature in prominent factors that drive asset prices. In addition, we find that the Sharpe

Table 5. Out-of-sample portfolio performance with a five-factor model

This table reports the out-of-sample portfolio performance using a rolling window scheme while controlling micro stocks (P20/P40: only include stocks are larger than 20/40 percentile of the NYSE listed stocks). Factor selection strategies include OWL, LASSO, Elastic Net (EN), and Fama-MacBeth regression (FM). The upper panel is obtained using the full sample; the middle and lower panels are obtained using sub-samples.

		SR	Mean	Std	Skewness	Kurtosis
		Panel A: full sample estimation				
full_P20	OWL	1.21	2.17	6.24	-0.07	9.48
	LASSO	1.01	2.13	7.30	2.21	31.09
	EN	1.04	2.26	7.52	1.71	27.70
	FM	0.96	1.96	7.08	2.88	37.37
full_P40	OWL	0.90	1.59	6.13	1.39	25.06
	LASSO	0.77	1.48	6.62	4.09	57.11
	EN	0.82	1.52	6.39	3.17	46.12
	FM	0.72	1.41	6.79	3.68	49.89
		Panel B: sub-sample estimation (1980:2000)				
sub1_P20	OWL	2.10	2.54	4.18	0.10	3.41
	LASSO	1.87	2.09	3.87	0.10	3.48
	EN	1.87	2.09	3.87	0.10	3.48
	FM	1.66	1.92	4.01	0.65	5.45
sub1_P40	OWL	1.35	1.34	3.44	-0.03	4.37
	LASSO	1.03	1.13	3.82	0.02	3.67
	EN	1.03	1.13	3.82	0.02	3.67
	FM	0.75	0.75	3.50	-0.21	5.62
		Panel C: sub-sample estimation (2001:2017)				
sub2_P20	OWL	2.10	2.43	4.67	1.02	8.72
	LASSO	1.91	2.10	3.80	0.16	3.51
	EN	1.91	2.10	3.80	0.16	3.51
	FM	1.78	1.80	3.49	-0.48	3.82
sub2_P40	OWL	2.11	2.04	3.34	0.62	5.83
	LASSO	1.80	1.69	3.27	0.58	6.16
	EN	1.69	1.59	3.25	0.37	4.44
	FM	1.80	1.75	3.35	0.13	2.91

ratio of the OWL-hedged portfolios are consistently higher than other benchmarks in all sample periods. In the first sub-sample, the OWL shrinkage method picked 'momentum' and 'profitability' related factors as the most important factors, whereas the other benchmarks picked up mostly 'liquidity' and 'beta' related factors.[22] Figure 5 Panel B shows that those 'momentum' and 'profitability' related factors, selected by the OWL shrinkage method, predict stock returns better than other benchmarks: resulting in 20% to 30% higher Sharpe ratios. A similar result can be found in Panel C for the second half sample estimation: the OWL-hedged portfolios yield the highest Sharpe ratios compared to other methods.

5 Conclusion

In the zoo of factors, using traditional methods, such as Fama-MacBeth regression and the LASSO shrinkage method to find factors that drive cross-sectional asset returns, faces tremendous challenges due to factor correlations. Nonetheless, we find that among 80 anomaly factors we considered, 64% of them exhibit correlation coefficients greater than 0.5 (absolute value), which casts doubt on the validity of using these traditional methods. By contrast, the OWL shrinkage method permits factor correlations and achieves correlation identification and sparsity shrinkage simultaneously. We derive the statistical properties for the OWL estimator and show that the OWL estimator is a consistent estimator under some regularity conditions. Monte Carlo experiments confirm the superior performance of the OWL estimator against other benchmarks when factors are correlated. Empirical analysis reveals that the OWL shrinkage method consistently chooses the 'market' factor as the most important factor to drive cross-sectional asset returns, while other benchmarks all failed to identify the 'market' factor as an important factor due to its high-correlation with other factors. In addition, out-of-sample analysis shows that the OWL shrinkage method can select factors that yield the highest Sharpe ratios in the factor-hedged portfolios compared to other benchmarks.

Finally, note that the purpose of this paper is not to find a "true" parsimonious asset pricing model, but to *robustly* identify a set of sparse factors to drive cross-sectional asset

[22] Note that 'beta' related factors are highly correlated with many other factors.

returns under potentially highly correlated factors. Bearing that in mind, our procedure is particularly useful for factor investing: the OWL shrinkage method can identify correlated factors that jointly drive stock returns, and can be further utilized to form portfolio strategies, see Asness et al. (2013) for an example. However, finding a "true" parsimonious asset pricing model remains as a key research question in the finance research. Such task is achievable once a de-biased version of the OWL estimator is developed - and that can be considered as a future research subject.

Appendix

A The OWL penalty, geometric interpretation and comparison with the LASSO

This section explains why the LASSO estimator is problematic whereas the OWL estimator is robust with correlated factors. Recall that the OWL penalty is defined as

$$\Omega_\omega(b) = \omega' |b|_\downarrow = \sum_{i=1}^{K} \omega_i |b|_{[i]}, \tag{A.1}$$

where $|b|_{[i]}$ and ω_i are specified in (7) and (8), respectively. Consider a simple two dimensional case, where $K = 2$. Then, the atomic norm of the LASSO and the OWL penalty can be written as

$$\Omega_{\omega,LASSO}(b) = \lambda|b_1| + \lambda|b_2| \leq 1, \tag{A.2}$$

$$\Omega_{\omega,OWL}(b) = \omega_1|b|_{[1]} + \omega_2|b|_{[2]} \leq 1, \tag{A.3}$$

respectively. Recall that $|b|_{[1]} = \max(|b_1|, |b_2|)$ and $|b|_{[2]} = \min(|b_1|, |b_2|)$. Therefore, equation (A.3) can be written as

$$\Omega_{\omega,OWL}(b) = \begin{cases} \omega_1|b_1| + \omega_2|b_2| \leq 1, & \text{if} \quad |b_1| \geq |b_2|, \\ \omega_1|b_2| + \omega_2|b_1| \leq 1, & \text{if} \quad |b_1| < |b_2|, \end{cases} \tag{A.4}$$

which implies that the geometric interpretation of the atomic norm of the LASSO and OWL penalties can be shown as in Figure 2.

Next, we will compare the LASSO and OWL penalties following a geometric argument, typically illustrated in the machine learning literature. From Figure 2, we can see that the LASSO norm has vertexes on both axes, which makes the LASSO estimator enjoy the sparse selection property (i.e., it shrinks one variable to zero while keeping the other non-zero). During estimation, the tangent point between the penalty norm and the contour coming

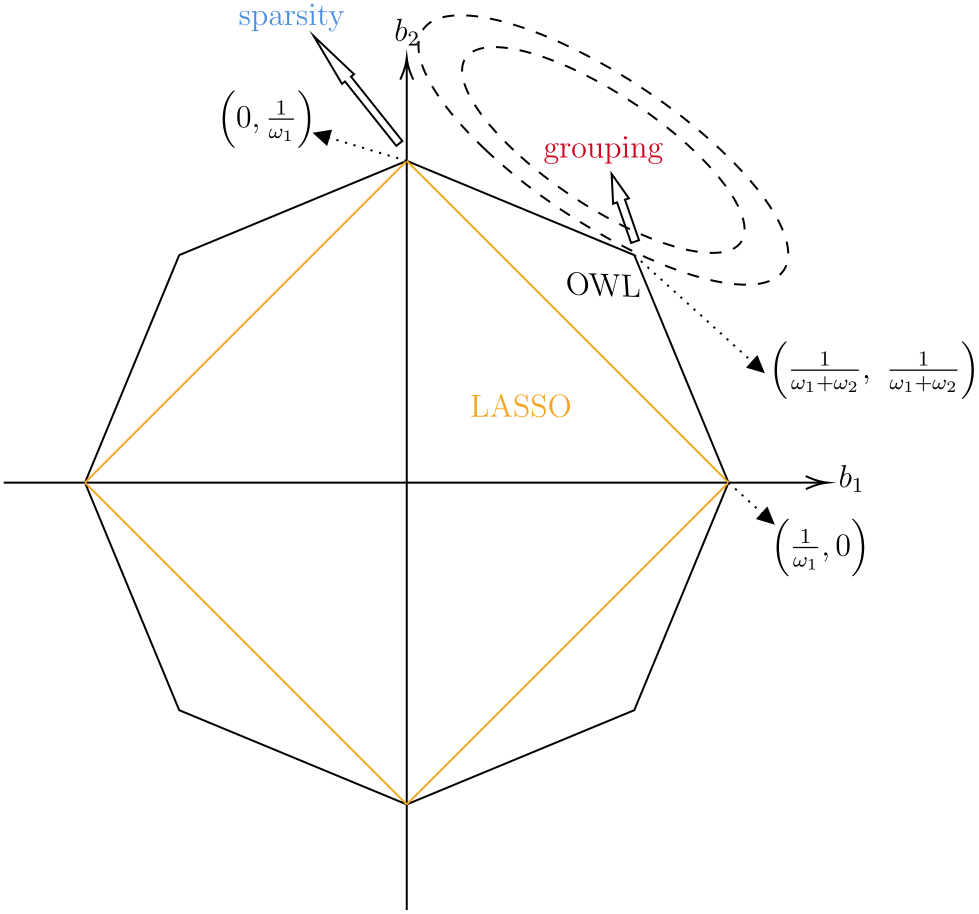

sparsity
b_2
$\left(0, \frac{1}{\omega_1}\right)$
grouping
OWL
$\left(\frac{1}{\omega_1+\omega_2}, \ \frac{1}{\omega_1+\omega_2}\right)$
LASSO
b_1
$\left(\frac{1}{\omega_1}, 0\right)$

Figure 2. Geometric interpretation of OWL and LASSO penalties

from the un-regularized solution determines the estimation results. However, when two variables are highly correlated, the frontier of the contour coming from the un-regularized solution is flat. Given the shapes of the LASSO norm and the contour under correlated factors, it is very unstable in determining which variable to shrink. A slight estimation error from the un-regularized solution can easily produce opposite inferences on factors selections. On the other hand, the OWL norm not only has vertexes on both axes, it also has vertexes on the ±45 degree lines. Those vertexes on the axes produce sparse selection like the LASSO estimator, while those on the ±45 degree lines yield grouping property which ensures robust factor selection while factors are correlated. When factors are highly correlated, they will be assigned with similar coefficients.

Also, note that

$$\begin{aligned} |b|_{[1]} &= \max(|b_1|, |b_2|) = \frac{1}{2}(|b_1| + |b_2| + \Big||b_1| - |b_2|\Big|), \\ |b|_{[2]} &= \min(|b_1|, |b_2|) = \frac{1}{2}(|b_1| + |b_2| - \Big||b_1| - |b_2|\Big|). \end{aligned}$$

Then the OWL penalty can be written as

$$\Omega_{\omega,LASSO}(b) = \omega_1|b|_{[1]} + \omega_2|b|_{[2]} = \frac{\omega_1 + \omega_2}{2}(|b_1| + |b_2|) + \frac{\overbrace{\omega_1 - \omega_2}^{\lambda_2}}{2}\Big||b_1| - |b_2|\Big|,$$

which suggests that the OWL penalty term can be decomposed into two components: first, $|b_1| + |b_2|$, which is the same as the LASSO shrinkage method, which produces sparse factor selection; second, $\Big||b_1| - |b_2|\Big|$, which shrinks $|b_1| \neq |b_2|$, where the shrinkage intensity is controlled by $(\omega_1 - \omega_2)/2$. Note that, by the definition of ω, we have $\omega_1 - \omega_2 = \lambda_2$. Therefore, the turning parameter λ_2 has a direct impact on the grouping property of the OWL estimator and thus can be controlled to achieve desirable grouping intensity.

B Proof of theorems

B.1 Proof of Theorem 2.1

The proof of Theorem 2.1 relies on the Pigou-Dalton transfer principle and the directional derivative lemma at the minimum of a convex function. It follows using a similar argument as in Figueiredo and Nowak (2016), except that we are dealing with both the time-series and cross-sectional dimensions.

Lemma 1 (Pigou-Dalton transfer principle)**.** *Let be given vector $x \in R^p_+$, and its two components x_i, x_j are such that $x_i > x_j$. Let $\epsilon \in (0, (x_i - x_j)/2)$, $z_i = x_i - \epsilon$, $z_j = x_j + \epsilon$, and $z_k = x_k$, for $k \neq i, j$. Set $\Omega_\omega(x) = \omega' x$, where $\omega \in R^p_+$, and $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_p$. It holds*

$$\Omega_\omega(x) - \Omega_\omega(z) \geq \Delta_\omega \epsilon, \qquad \Delta_\omega := \min_{i=1,\cdots,p-1} \omega_{i+1} - \omega_i.$$

Lemma 2 (Directional derivative)**.** *The directional derivative of function $f : R^K \to R$ at $x \in dom(f)$, in the direction $u \in R^K$ is given by*

$$f'(x, u) = \lim_{\alpha \to 0^+} [f(x + \alpha u) - f(x)]/\alpha, \quad \alpha > 0.$$

If f is a convex function, then $x^* \in \arg\min(f)$ if and only if $f'(x^*, u) \geq 0$ for any direction $u \in R^K$.

Proof of Theorem 2.1 . Denote the objective function in (7) as $Q(b) := \frac{1}{2}||\mu_R - Cb||_2^2 + \Omega_\omega(b)$. By definition, $\hat{b}$ is the minimizer of $Q(b)$, $Q(\hat{b}) \leq Q(b)$ for all b. Thus by Lemma 2, for any u,

$$Q'(\hat{b}, u) \geq 0. \tag{B.5}$$

Suppose

$$\sigma(f_i - f_j) < \frac{\lambda_2}{||\mu_R||_2 ||\sigma_R||_2}, \tag{B.6}$$

and assume

$$\hat{b}_i \neq \hat{b}_j.$$

We will show a contradiction between the assumption $\hat{b}_i \neq \hat{b}_j$ and (B.6). Without loss of the generality, assume $\hat{b}_i > \hat{b}_j$, $i < j$. First we define a special direction vector $u = (u_1, \cdots, u_K)$.

Set $u_i = -1$, $u_j = 1$, $u_k = 0$, for $k \neq i, j$. The directional derivative of Q at $\hat{b}$ with such u is

$$Q'(\hat{b}, u) = \lim_{\alpha \to 0^+} \left(QL_\alpha(\hat{b}, u) + RP_\alpha(\hat{b}, u) \right), \tag{B.7}$$

where

$$\begin{aligned} QL_\alpha(\hat{b}, u) &= \frac{||\mu_R - C(\hat{b} + \alpha u)||_2^2 - ||\mu_R - C\hat{b}||_2^2}{2\alpha}, \\ RP_\alpha(\hat{b}, u) &= \frac{\Omega_\omega(\hat{b} + \alpha u) - \Omega_\omega(\hat{b})}{\alpha}. \end{aligned}$$

By definition of u, we have $-\alpha C u = \alpha(C_i - C_j)$, where C_i and C_j are the i^{th} and j^{th} columns of the factor-return covariance matrix C. Hence $QL_\alpha(\hat{b}, u)$ can be written as

$$QL_\alpha(\hat{b}, u) = \frac{||\mu_R - C\hat{b} + \alpha(C_i - C_j)||_2^2 - ||\mu_R - C\hat{b}||_2^2}{2\alpha}.$$

Observe that

$$\begin{aligned} QL_\alpha(\hat{b}, u) &= \frac{||\mu_R - C\hat{b}||^2 + 2\alpha(\mu_R - C\hat{b})(C_i - C_j) + \alpha^2 ||C_i - C_j||_2^2 - ||\mu_R - C\hat{b}||_2^2}{2\alpha} \\ &\to (\mu_R - C\hat{b})(C_i - C_j) \quad \text{as } \alpha \to 0. \end{aligned}$$

Applying the Pigou-Dalton transfer principle on $RP_\alpha(\hat{b}, u)$ with $\epsilon = \alpha$, we obtain

$$-RP_\alpha(\hat{b}, u)\alpha = \Omega_\omega(\hat{b}) - \Omega_\omega(\hat{b} + \alpha u) \geq \Delta_\omega \alpha.$$

So for any α and u,

$$RP_\alpha(\hat{b}, u) \leq -\frac{\Delta_\omega \alpha}{\alpha} = -\Delta_\omega.$$

By the definition of ω in (8), $\Delta_\omega = \lambda_2$. Therefore, applying the above bound in (B.7), we obtain

$$\begin{aligned} Q'(\hat{b}, u) &\leq (\mu_R - C\hat{b})(C_i - C_j) - \Delta_\omega \\ &= (\mu_R - C\hat{b})(C_i - C_j) - \lambda_2. \end{aligned} \tag{B.8}$$

Using Cauchy-Schwarz inequality, we have $(\mu_R - C\hat{b})(C_i - C_j) \leq ||\mu_R - C\hat{b}||_2 \; ||C_i - C_j||_2$. So (B.8) becomes

$$Q'(\hat{b}, u) \leq ||\mu_R - C\hat{b}||_2 \; ||C_i - C_j||_2 - \lambda_2.$$

Since $\mu_R - C\hat{b}$ is a pricing error, then $||\mu_R - C\hat{b}||_2 < ||\mu_R||_2$, while by definition $\text{cov}(R, f_i - f_j) = C_i - C_j$. Then we have

$$Q'(\hat{b}, u) < ||\mu_R||_2 \; ||\text{cov}(R, f_i - f_j)||_2 - \lambda_2. \tag{B.9}$$

Now we further utilize the covariance inequality. For any $n = 1, \cdots, N$, R_n is the n^{th} column of the return matrix R, we have

$$\text{cov}(R_n, f_i - f_j) \leq \sqrt{\text{var}(R_n)\text{var}(f_i - f_j)} = \sigma_{R_n}\sigma(f_i - f_j), \tag{B.10}$$

where σ_{R_n} is the standard deviation of the n^{th} test asset. Apply (B.10) in (B.9), we have

$$\begin{aligned} Q'(\hat{b}, u) &< ||\mu_R||_2 \; ||\sigma_R \sigma(f_i - f_j)||_2 - \lambda_2 \\ &= ||\mu_R||_2 \; ||\sigma_R||_2 \; \sigma(f_i - f_j) - \lambda_2, \end{aligned} \tag{B.11}$$

where σ_R is a $N \times 1$ vector collecting the standard deviations of N test assets. So (B.11) together with (B.6) implies

$$Q'(\hat{b}, u) < 0,$$

which violates (B.5). Hence there is a contradiction between $\hat{b}_i \neq \hat{b}_j$ and (B.6). So we must have

$$\hat{b}_i = \hat{b}_j,$$

which completes the proof. □

B.2 Proof of Corollary 2.1

Proof. The proof of corollary 2.1 follows the same method as in Appendix B.1, except we choose a special vector for u where we set $u_i = 1$, $u_j = 1$, $u_k = 0$, for $k \neq i, j$. The rest of the proof follows trivially. □

B.3 Proof of Theorem 2.2

Proof. By definition the OWL estimator is minimizing the function

$$\hat{b} = \hat{b}_{OWL} = \arg\min_{b} \quad \frac{1}{N}||\mu_R - Cb||_2^2 + \frac{1}{N}\sum_{i=1}^{K}[\lambda_1 + \lambda_2(K-i)]|b|_{[i]},$$

where $|b|_{[\cdot]}$ denotes the element of the decreasingly ordered vector of $|\mathbf{b}|$, such that $|b|_{[1]} \geq |b|_{[2]} \geq ... \geq |b|_{[K]}$. Let b^0 be the vector of true values of risk prices, and $\mu_R = Cb^0 + \epsilon$. According to the "argmin" property, definition of $\hat{b}$ implies

$$\frac{1}{N}||\mu_R - C\hat{b}||_2^2 + \frac{1}{N}\sum_i[\lambda_1 + \lambda_2(K-i)]|\hat{b}|_{[i]} \leq \frac{1}{N}||\mu_R - Cb^0||_2^2 + \frac{1}{N}\sum_i[\lambda_1 + \lambda_2(K-i)]|b^0|_{[i]}. \tag{B.12}$$

Since $\omega_i = \lambda_1 + \lambda_2(K-i)$ is in a monotone non-negative cone and $\omega_1 \geq \omega_2 \geq ... \geq \omega_K$, we have

$$\sum_i[\lambda_1 + \lambda_2(K-i)]|\hat{b}|_{[i]} \geq \omega_K||\hat{b}||_1 = \lambda_1||\hat{b}||_1,$$
$$\sum_i[\lambda_1 + \lambda_2(K-i)]|b^0|_{[i]} \leq \omega_1||b^0||_1 = [\lambda_1 + \lambda_2(K-1)]||b^0||_1.$$

Together with $\mu_R = Cb^0 + \epsilon$, this implies that (B.12) can be simplified as:

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}||_1 \leq \frac{2}{N}\epsilon' C(\hat{b} - b^0) + \frac{1}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1. \tag{B.13}$$

Note that

$$2|\epsilon' C(\hat{b} - b^0)| \leq \left(\max_{1\leq j\leq K} 2|\epsilon' C^{(j)}|\right) ||\hat{b} - b^0||_1. \tag{B.14}$$

Hence, (B.13) can be written as

$$\frac{1}{N}||C(\hat{b}-b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}||_1 \leq \left(\frac{1}{N}\max_{1\leq j\leq K} 2|\epsilon' C^{(j)}|\right) ||\hat{b}-b^0||_1 + \frac{1}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1. \tag{B.15}$$

Consider the event

$$E := \left\{\frac{1}{N}\max_{1\leq j\leq K} 2|\epsilon' C^{(j)}| \leq \lambda_0\right\}, \tag{B.16}$$

where $\lambda_0 = \kappa\sqrt{\frac{\log K}{N}}$ and κ is a positive constant. Then, in view of (B.16), (B.15) can be bounded as

$$\frac{1}{N}||C(\hat{b}-b^0)||_2^2 + \frac{1}{N}\lambda_1||\hat{b}||_1 \leq \lambda_0||\hat{b}-b^0||_1 + \frac{1}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1. \tag{B.17}$$

By assumption, $\frac{\lambda_1}{N} = 2\lambda_0$. Therefore, (B.17) can be written as

$$\frac{2}{N}||C(\hat{b}-b^0)||_2^2 + \frac{2}{N}\lambda_1||\hat{b}||_1 \leq \frac{\lambda_1}{N}||\hat{b}-b^0||_1 + \frac{2}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1. \tag{B.18}$$

Note that

$$||\hat{b}||_1 = ||\hat{b}_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1 \geq ||b^0_{s_0}||_1 - ||\hat{b}_{s_0} - b^0_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1, \tag{B.19}$$

$$||\hat{b}-b^0||_1 = ||\hat{b}_{s_0} - b^0_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1. \tag{B.20}$$

Therefore, using (B.19) and (B.20), (B.18) can be written as

$$\begin{aligned}\frac{2}{N}||C(\hat{b}-b^0)||_2^2 &+ \frac{2\lambda_1}{N}(||b^0_{s_0}||_1 - ||\hat{b}_{s_0} - b^0_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1) \\ &\leq \frac{\lambda_1}{N}(||\hat{b}_{s_0} - b^0_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1) + \frac{2\lambda_1}{N}\|b^0\|_1 + \frac{2\lambda_2(K-1)}{N}||b^0||_1.\end{aligned} \tag{B.21}$$

Note that $\|b^0_{s_0}\|_1 = \|b^0\|_1$, so (B.21) can be written as

$$\frac{2}{N}||C(\hat{b}-b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}_{s_0^c}||_1 \leq 3\frac{\lambda_1}{N}||\hat{b}_{s_0} - b^0_{s_0}||_1 + \frac{2\lambda_2(K-1)}{N}||b^0||_1. \tag{B.22}$$

By (B.20), $||\hat{b}_{s_0^c}||_1 = ||\hat{b}-b^0||_1 - ||\hat{b}_{s_0} - b^0_{s_0}||_1$. Utilizing this in (B.22), we obtain

$$\frac{2}{N}||C(\hat{b}-b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}-b^0||_1 \leq 4\frac{\lambda_1}{N}||\hat{b}_{s_0} - b^0_{s_0}||_1 + \frac{2\lambda_2(K-1)}{N}||b^0||_1. \tag{B.23}$$

By Assumption 3, the restricted eigenvalue condition states that

$$\phi_0^2 := \min_{\substack{s_0 \subset \{1,\ldots,K\} \\ |s_0|<K}} \quad \min_{\substack{b \in R^K \setminus \{0\} \\ ||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1}} \frac{b'\hat{\Sigma}b}{||b_{s_0}||_2^2} > 0,$$

which implies that for any b,

$$\phi_0^2 \leq \frac{b'\hat{\Sigma}b}{||b_{s_0}||_2^2} \leq \frac{b'\hat{\Sigma}bS}{||b_{s_0}||_1^2},$$

where S is defined in Assumption 2 and the second inequality follows by utilizing the norm inequality $\|b_{s_0}\|_1 \leq \sqrt{S}\|b_{s_0}\|_2$. Rearranging the above inequality, we have

$$||b_{s_0}||_1^2 \leq b'\hat{\Sigma}bS/\phi_0^2, \tag{B.24}$$

which is called the *compatibility condition* in Buhlmann and Van De Geer (2011) pp. 106. Applying (B.24) on $||\hat{b}_{s_0} - b_{s_0}^0||_1$ and using $\hat{\Sigma} = \dfrac{C'C}{N}$, we have

$$\begin{aligned} ||\hat{b}_{s_0} - b_{s_0}^0||_1^2 &\leq (\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0)S/\phi_0^2 = ||C(\hat{b} - b^0)||_2^2 S/(N\phi_0^2), \\ ||\hat{b}_{s_0} - b_{s_0}^0||_1 &\leq ||C(\hat{b} - b^0)||_2\sqrt{S}/(\sqrt{N}\phi_0). \end{aligned}$$

Therefore, using inequality $4ab \leq a^2 + 4b^2$, we obtain

$$\begin{aligned} 4\frac{\lambda_1}{N}||\hat{b}_{s_0} - b_{s_0}^0||_1 &\leq 4\left(\frac{||C(\hat{b} - b^0)||_2}{\sqrt{N}}\right)\left(\frac{\lambda_1}{N}\frac{\sqrt{S}}{\phi_0}\right) \\ &\leq \frac{1}{N}||C(\hat{b} - b^0)||_2^2 + 4(\frac{\lambda_1}{N})^2\frac{S}{\phi_0^2}. \end{aligned}$$

So (B.23) can be written as

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b} - b^0||_1 \leq 4(\frac{\lambda_1}{N})^2\frac{S}{\phi_0^2} + \frac{2\lambda_2(K-1)}{N}||b^0||_1. \tag{B.25}$$

Note that $\dfrac{1}{N}||C(\hat{b} - b^0)||_2^2 = (\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0)$, so (B.25) completes the proof of (12).

Now we have obtained (12) assuming (B.16). In the next step we want to evaluate the probability of the inequality (B.16) to be true, i.e. $\mathbb{P}(E)$. By a union bound and using the notation $\zeta_j = \epsilon' C^{(j)} = \sum_{i=1}^{N} \epsilon_i C_i^{(j)} = \sum_{i=1}^{N} \zeta_{i,j}$, we obtain

$$\mathbb{P}(E^C) = \mathbb{P}(\frac{1}{N}\max_{1\leq j\leq K} 2|\epsilon' C^{(j)}|) \geq \lambda_0) \leq \sum_{j=1}^{K} \mathbb{P}(\frac{1}{N}|\zeta_j| \geq \frac{\lambda_0}{2}). \tag{B.26}$$

Note that both $\{\epsilon_i\}_{i=1}^N$ and $\{C_i^{(j)}\}_{i=1}^N$ for all $i = 1, \cdots, N$ and $j = 1, \cdots, K$ are uniformly subgaussian variables. Therefore, variables $\{\zeta_{i,j}\}_{i=1}^N$ are uniformly subexponentially distributed. Hence, applying Corollary 5.17 in Vershynin (2012) and utilizing $\lambda_0 = \kappa\sqrt{\frac{\log K}{N}}$, we obtain

$$\begin{aligned}\mathbb{P}(E^C) &\leq K \max_{1\leq j\leq K} \mathbb{P}(\frac{1}{N}|\zeta_j| \geq \frac{\lambda_0}{2}) = K \max_{1\leq j\leq K} \mathbb{P}(\frac{1}{N}\left|\sum_{i=1}^N \zeta_{i,j}\right| \geq \frac{\lambda_0}{2}) \\ &\leq 2K \exp[-c\kappa^2 \log K] = 2K^{1-c\kappa^2}.\end{aligned}$$

where c and κ are positive constants. Therefore, selecting κ such that $c\kappa^2 > 1$, we have the following property for (B.16):

$$\mathbb{P}(E) = 1 - \mathbb{P}(E^C) \geq 1 - 2K^{1-c\kappa^2} \to 1, \tag{B.27}$$

as $N, K \to \infty$. This completes the proof of Theorem 2.2. □

B.4 Proof of Corollary 2.2

Proof. By assumption of theorem $\frac{\lambda_1}{N} = 2\lambda_0 = 2\kappa\sqrt{\frac{\log K}{N}}$, where κ is a positive constant, and $\frac{\lambda_2}{N} = O(\frac{S\log K}{NK})$. Therefore, both two terms on the right hand side of (B.25) are $O(\frac{S\log K}{N})$. Hence, (B.25) implies

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 = O\left(\frac{S\log K}{N}\right), \tag{B.28}$$

$$||\hat{b} - b^0||_1 = O\left(S\sqrt{\frac{\log K}{N}}\right). \tag{B.29}$$

So (B.29) proves the first claim of (13). Observe that

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 = (\hat{b} - b^0)'(\hat{\Sigma} - \Sigma)(\hat{b} - b^0) + (\hat{b} - b^0)'\Sigma(\hat{b} - b^0), \tag{B.30}$$

Notice that

$$(\hat{b} - b^0)'\Sigma(\hat{b} - b^0) \geq \Lambda_{min}^2||\hat{b} - b^0||_2^2,$$

where Λ_{min} denotes the smallest eigenvalue of Σ, and Σ is the true value of $\hat{\Sigma}$, so $\Lambda_{min} > 0$. Moreover in (B.30), it holds

$$(\hat{b} - b^0)'(\hat{\Sigma} - \Sigma)(\hat{b} - b^0) \geq -||\hat{\Sigma} - \Sigma||_\infty ||\hat{b} - b^0||_1^2,$$

where $||\hat{\Sigma} - \Sigma||_\infty := \max_{1\leq i,j \leq K} |\hat{\Sigma}_{i,j} - \Sigma_{i,j}|$. Using Lemma 14.12 in Buhlmann and Van De Geer (2011), we have $\max_{1\leq i,j \leq K} |\hat{\Sigma}_{i,j} - \Sigma_{i,j}| = O_p(\sqrt{\frac{\log K}{N}})$. Hence (B.28) can be rewritten as

$$\begin{aligned} O\left(\frac{S \log K}{N}\right) &= \frac{1}{N}||C(\hat{b} - b^0)||_2^2 \\ &\geq \Lambda_{min}^2 ||\hat{b} - b^0||_2^2 - ||\hat{\Sigma} - \Sigma||_\infty ||\hat{b} - b^0||_1^2 \\ &\geq \Lambda_{min}^2 ||\hat{b} - b^0||_2^2 - O_p\left(S^2 \left(\frac{\log K}{N}\right)^{3/2}\right). \end{aligned} \tag{B.31}$$

Rearranging it, we have

$$||\hat{b} - b^0||_2^2 \leq \frac{1}{\Lambda_{min}^2} O(\frac{S \log K}{N}) + \frac{1}{\Lambda_{min}^2} O_p\left(S^2 \left(\frac{\log K}{N}\right)^{3/2}\right).$$

By Assumption 2, $S\sqrt{\frac{\log K}{N}} = o(1)$. Together with $\frac{1}{\Lambda_{min}^2} = O(1)$, we obtain

$$||\hat{b} - b^0||_2^2 = O_p(\frac{S \log K}{N}), \tag{B.32}$$

which proves the second claim of (13). □

C Solving the OWL optimization problem

This section follows similar arguments in Zeng and Figueiredo (2014) and explains how to use the proximal gradient descent algorithm to solve the optimization problem of the OWL estimator. The first subsection introduces the OWL proximal function which is used to compute the optimizer at each step. The second subsection outlines the fast-iterative-soft-thresholding-algorithm (FISTA) used to find the global optimizer, together with a backtracking line search condition which speeds up computation greatly.

C.1 OWL proximal function

Denote by $b = (b_1, \cdots, b_n)'$, $x = (x_1, \cdots, x_n)'$ column vectors. First we define the proximal function as

$$Prox_{\Omega_\omega}(b) = \arg\min_x [\frac{1}{2}||x - b||_2^2 + \Omega_\omega(x)], \quad \Omega_\omega(x) = \omega'|x|_\downarrow \tag{C.33}$$

where $\omega \in \kappa$, takes values from a monotone non-negative cone, defined as $\kappa := \{v \in R^n : v_1 \geq v_2 \geq \cdots \geq v_n \geq 0\}$, $|x|_\downarrow = (|x|_{[1]}, |x|_{[2]}, \cdots, |x|_{[n]})'$ and $|x|_{[1]} \geq |x|_{[2]} \geq \cdots \geq |x|_{[n]}$, is the vector of absolute values of elements of vector x, decreasingly ordered. By the definition of $\Omega_\omega(b)$, we have

$$\Omega_\omega(b) = \Omega_\omega(|b|), \tag{C.34}$$

where $|b| = (|b_1|, \cdots, |b_n|)'$. It is easy to show that

$$||b - \text{sign}(b) \odot |x|||_2^2 \leq ||b - x||_2^2, \tag{C.35}$$

where $\text{sign}(b) = (\text{sign}(b_1), \cdots, \text{sign}(b_n))'$ is a function that retrieves signs from a vector, with elements in $\{1, -1, 0\}$ and $\odot$ is a point-wise production operator. Therefore, (C.34) and (C.35) imply

$$Prox_{\Omega_\omega}(b) = \text{sign}(b) \odot Prox_{\Omega_\omega}(|b|). \tag{C.36}$$

Let P be a permutation matrix that orders elements of a vector in decreasing order. Then permutation matrix has property

$$||P(x - b)||_2^2 = ||x - b||_2^2, \tag{C.37}$$

and by the definition of $\Omega_\omega(b)$,

$$\Omega_\omega(b) = \Omega_\omega(Pb). \tag{C.38}$$

So (C.37) and (C.38) imply that (C.36) can be written as

$$Prox_{\Omega_\omega}(b) = sign(b) \odot P' \, Prox_{\Omega_\omega}(|b|_\downarrow), \tag{C.39}$$

where $|b|_\downarrow$ is defined as $|x|_\downarrow$, and P' is the transpose of the permutation matrix, which recovers the order of $|b|_\downarrow$, i.e. $P|b| = |b|_\downarrow$, $P'|b|_\downarrow = |b|$ and $P'P = I$, where I is an identity matrix.

Since $|b|_\downarrow \in \kappa$, for any $x^* \in \kappa$ and any $x \in R^n$, we have $|b|'_\downarrow x \leq |b|'_\downarrow x^*$. Therefore,

$$\begin{aligned}\frac{1}{2}||x - |b|_\downarrow||_2^2 + \Omega_\omega(x) &= \frac{1}{2}||x||_2^2 + \frac{1}{2}|||b|_\downarrow||_2^2 - |b|'_\downarrow x + \Omega_\omega(x)\\ &\geq \frac{1}{2}||x^*||_2^2 + \frac{1}{2}|||b|_\downarrow||_2^2 - |b|'_\downarrow x^* + \Omega_\omega(x^*)\\ &= \frac{1}{2}||x^* - |b|_\downarrow||_2^2 + \Omega_\omega(x^*).\end{aligned}$$

Note that $Prox_{\Omega_\omega}(|b|_\downarrow) = \arg\min_x[\frac{1}{2}||x - |b|_\downarrow||_2^2 + \Omega_\omega(x)]$, and $\frac{1}{2}||x^* - |b|_\downarrow||_2^2 + \Omega_\omega(x^*) \leq \frac{1}{2}||x - |b|_\downarrow||_2^2 + \Omega_\omega(x)$. It implies that $Prox_{\Omega_\omega}(|b|_\downarrow) \in \kappa$, and $Prox_{\Omega_\omega}(|b|_\downarrow) = \arg\min_{x\in\kappa}[\frac{1}{2}||x - |b|_\downarrow||_2^2 + \omega' x]$. Completing the square, we have

$$Prox_{\Omega_\omega}(|b|_\downarrow) = \arg\min_{x\in\kappa}(\frac{1}{2}\|x - |b|_\downarrow\|_2^2 + \omega' x) = \arg\min_{x\in\kappa}\frac{1}{2}\|x - (|b|_\downarrow - \omega)\|_2^2,$$

which is the projection of $(|b|_\downarrow - \omega)$ onto κ [23]. Then equation (C.39) can be written as

$$Prox_{\Omega_\omega}(b) = \text{sign}(b) \odot P' \, Proj_\kappa(|b|_\downarrow - \omega)), \tag{C.40}$$

where $Proj_\kappa(.)$ is the projection operator onto κ.

After solving the proximal function, we can employ the iterative soft-thresholding algorithm to find the global optimizer. First, we initialize $b^{(0)}$, [24] then repeat

$$b^{(k+1)} = prox_{\Omega_\omega}(b^{(k)} - sz_k \bigtriangledown g(b^{(k)})) \tag{C.41}$$

[23] The projection onto κ is an isotonic optimization problem and can be obtained by using the Pool-Adjacent-Violators algorithm in de Leeuw et al. (2009).

[24] For instance, we use the OLS estimate as initialization in our application but it can be any random vector, which will results in the same global minimizer for b since it is a convex minimization problem. However, a good choice of initialization can reduce computation time greatly.

until a stopping criterion is met, where $k = 1, 2, 3, ...$ are steps of each iteration, $g(b) = \frac{1}{2}(\mu_R - Cb)'W(\mu_R - Cb)$ and sz_k is the step size at the k^{th} iteration.

C.2 FISTA algorithm

Algorithm 1 is based on Zeng and Figueiredo (2014) and fast computation is achieved by using the backtracking line condition (step 7) and the acceleration in u (step 12). The backtracking line condition allows large step sizes if optimizer stays in the right direction, otherwise shrinks step sizes. Steps 11 to 12 accelerate computation by moving the optimizer further towards the global optimizer at early iterations, while this acceleration diminishes when approaching the global optimizer.

Algorithm 1: FISTA-OWL

```
1  **Input:** $\mu_R, C, \omega$
2  **Output:** **OWL estimator** $\hat{b}$
3  **Initialisation:**$b_0 = \hat{b}_{OLS}, t_0 = t_1 = 1, u_1 = b_0, k = 1, \eta \in (0, 1), \tau_0 \in (0, 1/L)$ [a]
4  **while** *some stopping criterion not met* **do**
5      $\tau_k = \tau_{k-1}$;
6      $b_k = Prox_{\Omega_\omega}(u_k + \tau * C' * (\mu_R - Cb))$
7      **while** $\frac{1}{2}||\mu_R - Cb_k||_2^2 > Q(b_k, u_k)$ [b] **do**
8          $\tau_k = \eta * \tau_k$;
9          $b_k = Prox_{\Omega_\omega}(u_k + \tau * C' * (\mu_R - Cb))$
10     **end**
11     $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$
12     $u_{k+1} = b_k + \frac{t_{k-1}}{t_{k+1}}(b_k - b_{k-1})$
13     $k \leftarrow k + 1$
14 **end**
15 **Return:** $b_{k-1}$
```

[a] L is a Lipschitz constant.

[b] $Q(b_k, u_k) := \frac{1}{2}||\mu_R - Cu_k||_2^2 - (b_k - u_k)'C'(\mu_R - Cu_k) + \frac{1}{2\tau_k}||b_k - u_k||_2^2$ is the backtracking line condition.

D Motivating the "restricted eigenvalue condition"

The following lemma motivates the restricted eigenvalue condition. A matrix $\hat{\Sigma}$ that satisfies the restricted eigenvalue condition

$$\phi^2_{\hat{\Sigma}} := \min_{\substack{s_0 \subset \{1,\ldots,K\} \\ |s_0|<K}} \min_{\substack{b \in R^K \setminus \{0\} \\ ||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1}} \frac{b'\hat{\Sigma}b}{||b_{s_0}||_2^2} > 0, \tag{D.1}$$

if it is close to a matrix whose restricted eigenvalues are strictly positive. Let $\Sigma = E(\hat{\Sigma}) = E(\dfrac{C'C}{N})$ be the population value of the scaled Gram matrix. Since Σ is a non-singular matrix, its restricted eigenvalues are strictly positive: $\phi^2_{\Sigma} > 0$.

Lemma 3. *Suppose S is the sparsity parameter, $\delta = \max\limits_{1 \leq i,j \leq N} |\Sigma_{i,j} - \hat{\Sigma}_{i,j}|$, then for any vector b that satisfies $||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1$, it holds*

$$\phi^2_{\hat{\Sigma}} > \phi^2_{\Sigma} - 16S\delta.$$

Proof.

$$\begin{aligned} b'\Sigma b - b'\hat{\Sigma}b &\leq |b'\Sigma b - b'\hat{\Sigma}b| = |b'(\Sigma - \hat{\Sigma})b| \\ &\leq ||b||_1||(\Sigma - \hat{\Sigma})b||_\infty \leq \delta||b||_1^2 \end{aligned}$$

Recall that $b = b_{s_0} + b_{s_0^c}$, so $\|b\|_1 \leq \|b_{s_0}\|_1 + \|b_{s_0^c}\|_1$. Together with the assumption $||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1$, we have $||b||_1^2 \leq (||b_{s_0^c}||_1 + ||b_{s_0}||_1)^2 \leq 16||b_{s_0}||_1^2$. Hence we have

$$b'\Sigma b - b'\hat{\Sigma}b \leq 16\delta||b_{s_0}||_1^2.$$

Rearranging the above inequality, we have

$$\frac{b'\hat{\Sigma}b}{||b_{s_0}||_2^2} \geq \frac{b'\Sigma b}{||b_{s_0}||_2^2} - 16S\delta.$$

By the definition of restricted eigenvalues in (D.1), we have

$$\phi_{\hat{\Sigma}}^2 \geq \phi_{\Sigma}^2 - 16S\delta.$$

□

Lemma 3 shows that for the restricted eigenvalue condition to be satisfied, i.e. $\phi_{\hat{\Sigma}}^2 > 0$, it suffices to show that δ is small, or that the Gram matrix $\hat{\Sigma}$ is close to a positive definite matrix Σ. The following lemma shows that the "Restricted eigenvalue condition" implies the compatibility condition in Buhlmann and Van De Geer (2011) (pp. 106), which will be used for deriving the error bound in Theorem 2.2.

Lemma 4 (Compatibility condition)**.** *If the scaled Gram matrix $\hat{\Sigma}$ satisfies* (D.1)*, then*

$$||b_{s_0}||_1^2 \leq (b'\hat{\Sigma}b)S/\phi_0^2.$$

Proof. From the definition of restricted eigenvalues, we have

$$\phi_0^2 = \min_{\substack{s_0 \in \{1,\ldots,K\} \\ |s_0| < K}} \quad \min_{\substack{b \in R^K \setminus \{0\} \\ ||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1}} \frac{b'\hat{\Sigma}b}{||b_{s_0}||_2^2} > 0.$$

By the norm inequality, $\sqrt{S}\|b_{s_0}\|_2 \geq \|b_{s_0}\|_1$. Hence for any b, it holds

$$\phi_0^2 \leq \frac{b'\hat{\Sigma}b}{||b_{s_0}||_2^2} \leq \frac{b'\hat{\Sigma}bS}{||b_{s_0}||_1^2}.$$

Rearranging, we obtain

$$||b_{s_0}||_1^2 \leq (b'\hat{\Sigma}b)S/\phi_0^2.$$

□

E Simulation

Table 1 summarizes the performance of four candidate models under various settings by comparing their average MSEs under 500 trials. To have a better view on how those candidate models perform for each block of factors, we randomly chose one trail and plot the estimates from candidate models along with the oracle value. In particular, we focus on the highly correlated setting (i.e. $\rho = 0.9$).

Figure 3 looks into the setting when $N \sim K$ and it reports the plot of the OWL estimate over 90 factors along with other benchmarks and the oracle value (black). The upper left panel displays the plots of estimated coefficients for all factors. The remaining three panels display the detailed plot of estimates for each of these three blocks of factors. The upper right panel displays the plot of all estimates of useful factors that are correlated. In the presence of high correlation, the LASSO estimator performs poorly with highest estimation errors. Adaptive LASSO is strongly governed by the adaptive weights which is set to be the OLS estimate. So adaptive LASSO exhibits very similar behaviour to the OLS estimator. Elastic Net, as a hybrid estimator between LASSO and Ridge regression, is designed to stabilize LASSO selections in the presence of correlation. Although Elastic Net does improve the performance of LASSO in the context of correlated factors, it is still substantially outperformed by OWL. OWL produces the smallest estimation error and is the only estimator that groups together highly correlated variables by assigning them with similar coefficients. The bottom left panel displays the plot of all estimates of useless/redundant factors which are highly correlated. In terms of shrinking off useless/redundant factors, LASSO, EN, and OWL all perform well: they set most of useless factors to zeros. By contrast, adaptive LASSO is affected by the adaptive weights (i.e., the OLS estimate) and fails to set many useless/redundant factors to zeros. The bottom right panel displays the plot of all estimates of useful factors which are not correlated. Again, LASSO and Elastic Net are the worst performers yielding the largest estimation error. Also note that in the uncorrelated setting Elastic Net performs similarly to LASSO. In the ideal world where factors are uncorrelated, OLS and adaptive LASSO are the best performers, which is tightly followed by OWL. Note that OWL, LASSO and Elastic Net are biased towards zero, which is typically observed for shrinkage-estimators in small samples.

In the second experiment, there are 1000 test assets ($N = 1000$, $N \gg K$) and everything else is the same as in the first experiment. This setting typically represents a low-dimensional world.

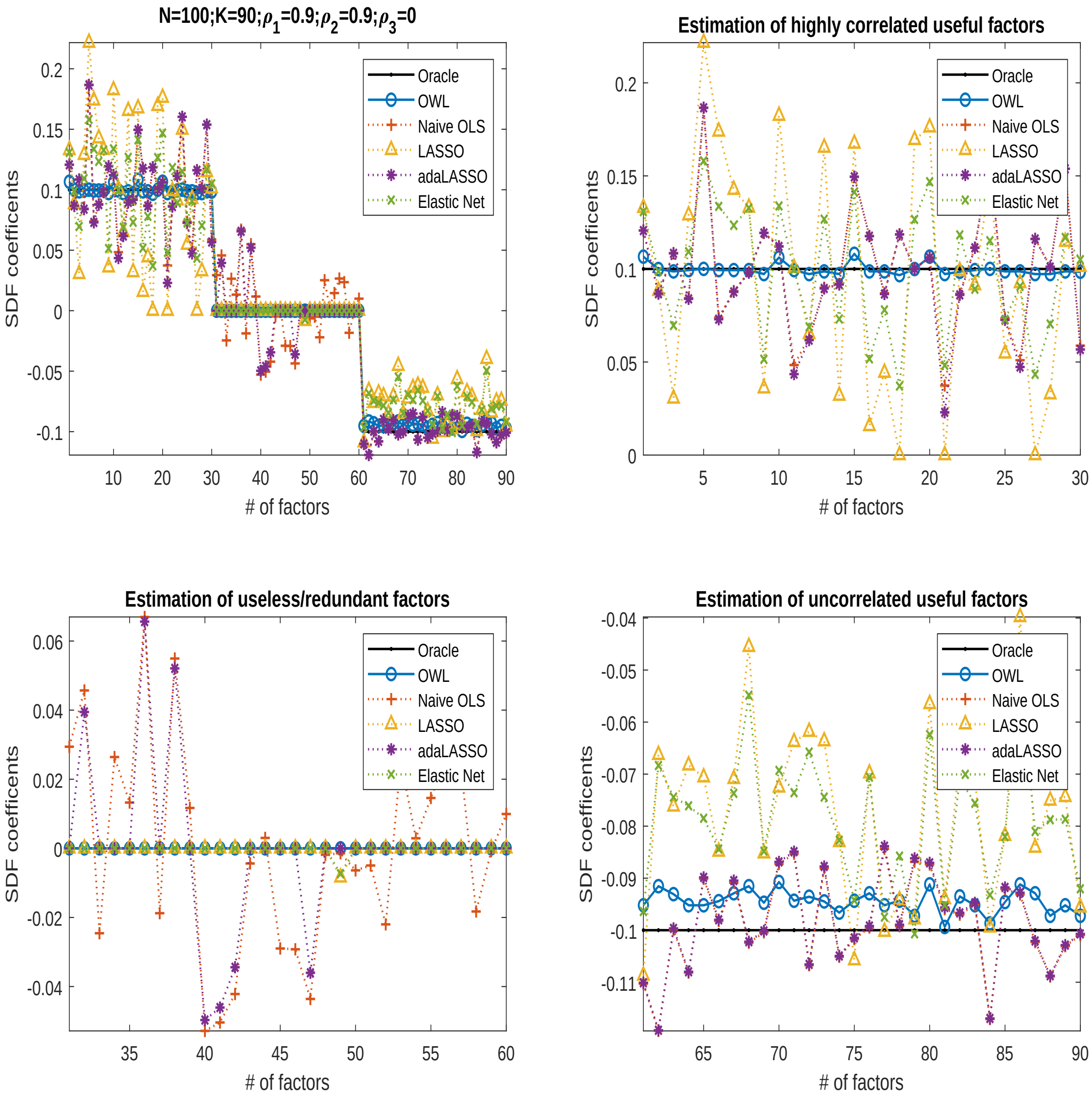


Figure 3. Estimation of SDF coefficients: $N = 100, K = 90$

This figure reports the values of the OWL estimator over 90 factors along with other benchmarks and the oracle value (black). There are 100 test assets, 90 candidate factors, which are divided into 3 equal blocks, where correlation coefficients of factors within each block are $\rho_1 = 0.9, \rho_2 = 0.9, \rho_3 = 0$. The upper left panel displays the plot of estimated SDF coefficients for all factors. The remaining three panels are detailed plot of estimates for each of these three blocks of factors. The upper right panel displays the plot of all estimates of useful factors that are highly correlated. The bottom left panel displays the plot of all estimates of useless/redundant factors. The bottom right panel displays the plot of all estimates of useful factors that are not correlated. In each plot, OWL estimator is displayed along with LASSO, adaptive LASSO, Elastic Net, and naive OLS estimator.

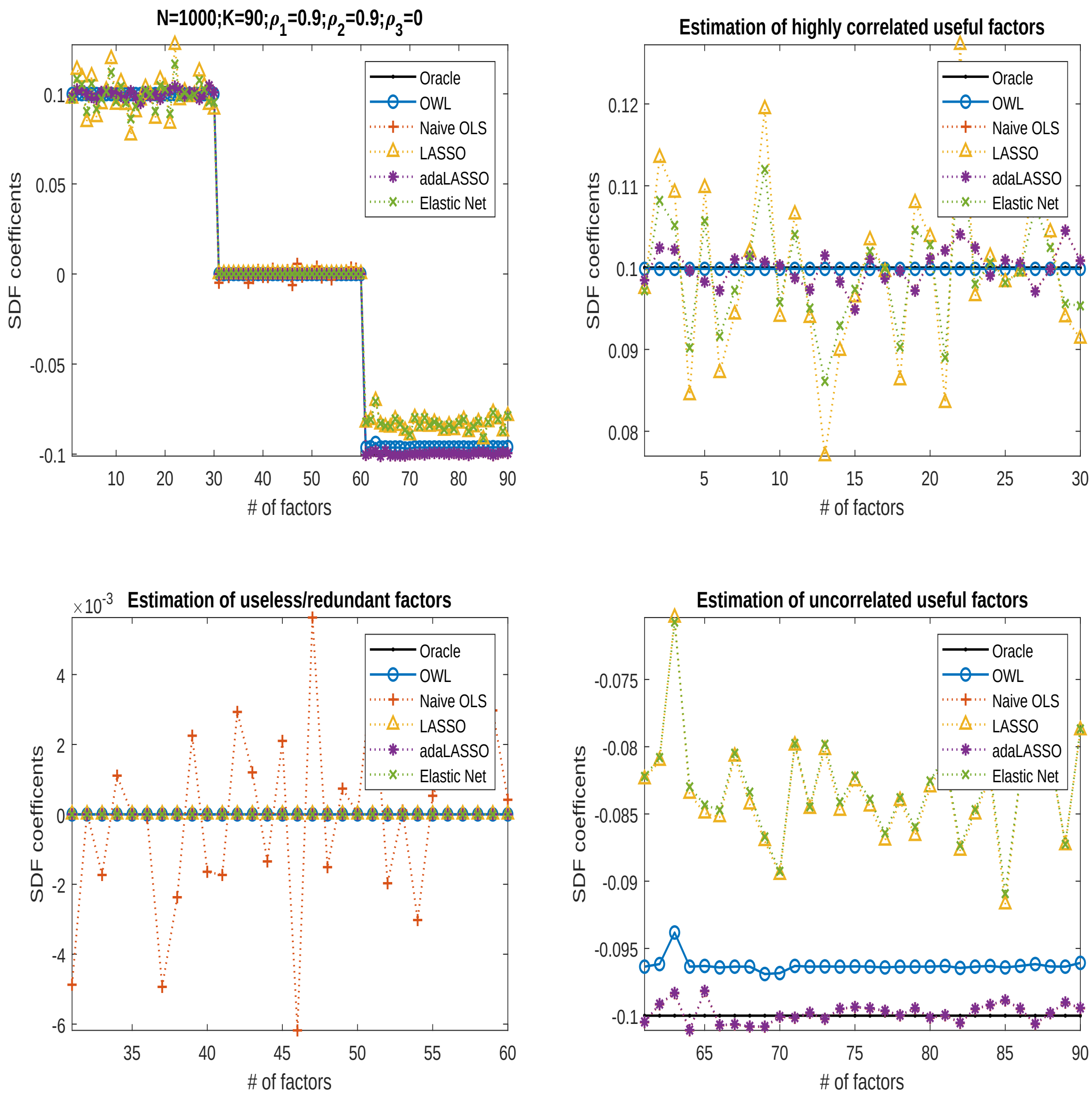


Figure 4. Estimation of SDF coefficients: $N = 1000, K = 90$

This figure reports the plot of the values of the OWL estimator along with other benchmark estimators. The number of assets is 1000. The rest are the same with the first experiment in Figure 3.

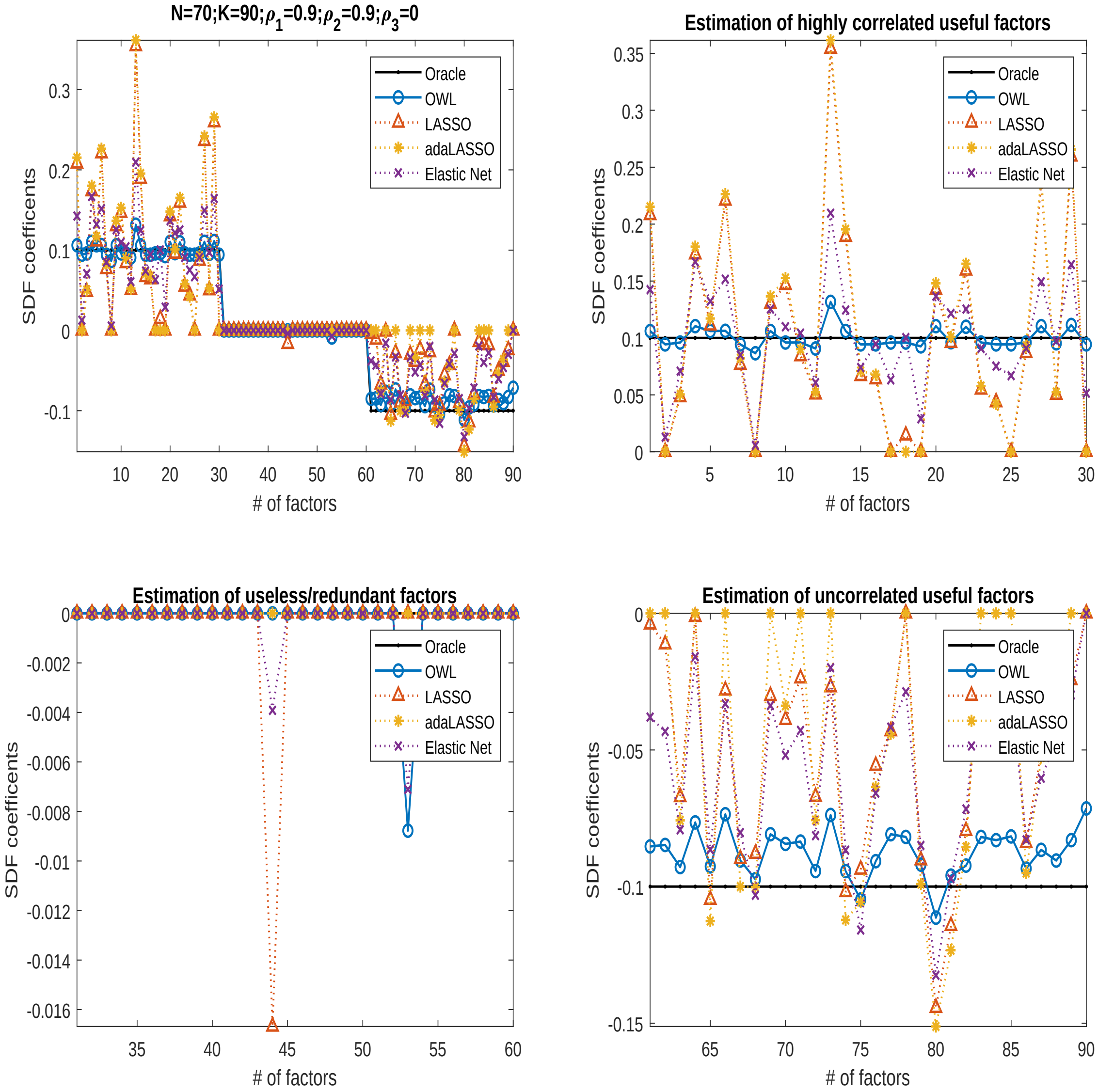


Figure 5. Estimation of SDF coefficients: $N = 70, K = 90$

This figure reports the plot of the values of the OWL estimator along with other benchmark estimators. Adaptive LASSO is using the LASSO estimate as its adaptive weight. The number of assets is 70. The rest are the same as in the first experiment in Figure 3.

Figure 4 reports the plot of estimated SDF coefficients using OWL and other benchmarks with 1000 test assets. When test assets are abundant, all shrinkage based estimators do a good job to shrink off useless/redundant factors. Adaptive LASSO performs the best at estimating uncorrelated factors: governed by the OLS weights, it is the only unbiased estimator among shrinkage based estimators. LASSO and Elastic Net produce the most biased estimators among all benchmarks. With highly correlated useful factors, OWL produces the most accurate estimation. With uncorrelated factors, OLS and adaptive LASSO are undoubtedly the best estimators, followed closely by OWL. For that reason, adaptive LASSO would be a good estimator in a low dimensional world where $N \gg K$. However, in a world of many factors, where $K > N$, OLS will be infeasible, hence the adaptive LASSO using OLS weighting is also improbable.

In the third experiment, there are 70 test assets ($N = 70,\ N < K$), everything else is the same as in the first two experiments. This setting represents a high-dimensional world, where the number of factors is greater than the number of test assets.

Figure 5 reports estimation results of each method along with the oracle value. Once $K > N$ the naive OLS estimator becomes infeasible, thus we remove it from the benchmarks. Meanwhile, we use the LASSO estimate as the adaptive weight for adaptive LASSO estimator. As for useless factors, all machine learning methods do a good job to shrink most useless factors to zeros. For the highly correlated useful factors, OWL is still the best estimator, producing the smallest estimation error while LASSO and adaptive LASSO are the worst performers producing very volatile estimates and wrongly shrinking many useful factors to zero. Interestingly, we find that Elastic Net performs significantly better compared to LASSO. However, despite this, Elastic Net is still substantially outperformed by OWL. For the useful factors (both correlated and uncorrelated), adaptive LASSO, using the LASSO estimate as the adaptive weight, performs the worst. The adaptive weight exacerbates the estimation severely.

These three experiments confirm that the LASSO estimator performs poorly when factors are correlated. Elastic Net does improve the performance of LASSO under such circumstance, however, it is still substantially outperformed by the OWL estimator, which makes the OWL estimator the best candidate when factors are correlated. Adaptive LASSO is a good choice in a low-dimensional setting where $N \gg K$, however, it performs the worst in a high-dimensional setting where $K > N$ (i.e., the OLS estimate becomes infeasible).

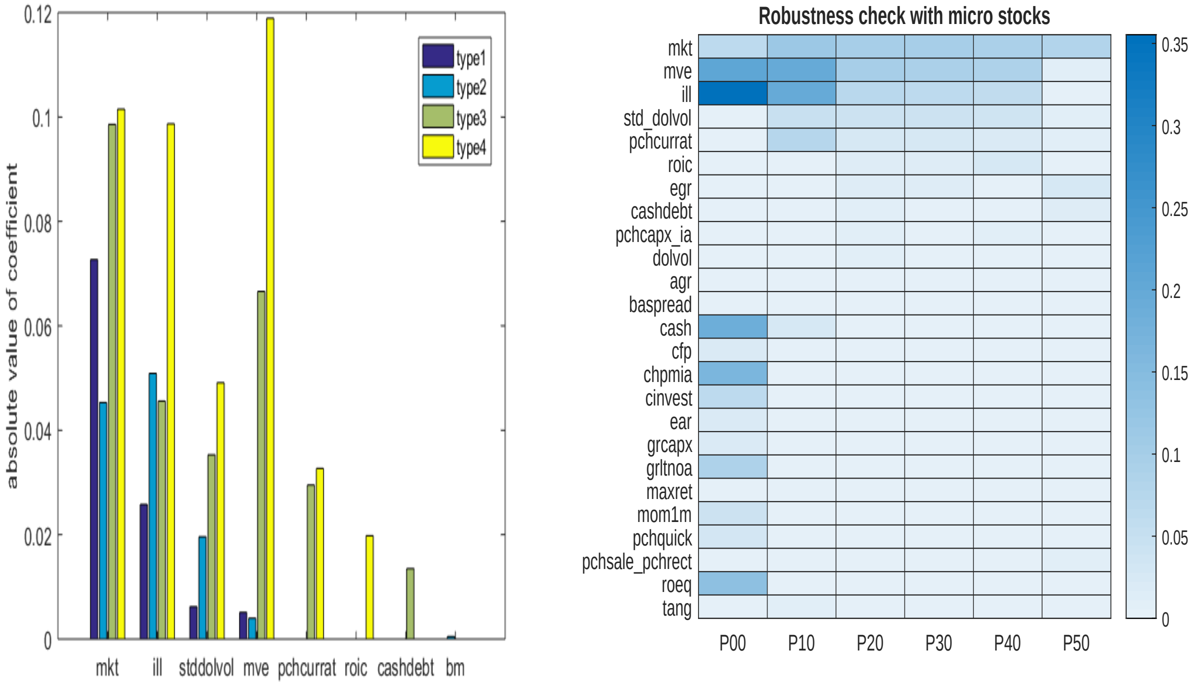
F Robustness check

In this section, we check whether various sorting methods would alter our estimation results and investigate how small stocks affect priced factors;

F.1 Using various sorting methods for test assets

For the first task, we consider three additional types of sorting method for constructing test portfolios and compare them with the sorting method used in the main body of this paper to check whether liquidity related factors are consistently chosen. First, we apply the uni-variate sorting method to sort all non-micro stocks into decile portfolios before combining them together to obtain 800 test portfolios. Compared to the test portfolio in empirical analysis, all characteristics are treated equally. Second, we consider the bi-variate sorting method, but using all possible combinations of two out of 80 characteristics, that is $80 \times 79/2 = 3160$ possibilities. To reduce the dimension of test portfolios, for each possible combination, we consider the 2 by 2 (instead of 5 by 5) sorting method: we sort stocks into 'high' and 'low' groups by each of these two characteristics where the thresholds are the medians of these characteristics. We then obtain 3160×4, total 12640 test portfolios. Third, we consider a similar method in the empirical analysis, that is singling out 'size' as a common characteristic, and using it with the remaining characteristics to form bi-variate sorted portfolios; but instead of forming the 5 by 5 portfolios, we form 3 by 3 portfolios.

Figure 6a reports the estimation results using four different sets of test assets (including the one used in the main body of this paper). First, 'market' along with 'illiquidity' and 'standard deviation of dollar volume' are consistently chosen as the most important factors to drive asset prices, with 'illiquidity' topping the chart of anomaly factors. Second, the impact of 'size' factor (mve) on test assets decreased colossally once it is not singled out to form bi-variate sorted portfolios. We can conclude that in 'type3' and 'type4' where 'size' effect tops the chart, it is artificially caused by portfolio sorting methods. However in



(a) Robustness check with test assets

(b) Robustness check with micro stocks

Figure 6. Robustness check with alternative test assets

Figure 6a reports the absolute value of SDF coefficients estimated by OWL using four types of test assets. Figure 6b reports the OWL estimates with six different treatments of micro stocks.

empirical analysis (‘type4’), ‘size’ is not a competing factor. Third, although singling out ‘size’ to form bi-variate sorted portfolios may alter the ‘size’ effect, it does not alter other factors’ implications: liquidity related factors are primary factors driving asset prices.

F.2 How small stocks affect the inference of priced factors

For the second task, we use the same sorting method as in the empirical analysis, but we consider six types of treatment of micro stocks: 1) keep all micro stocks (P00); 2) remove stocks that are smaller than 10 percentile of NYSE listed stocks (P10); 3-6) similarly, remove stocks that are smaller than (20-50) percentile of NYSE listed stocks (P20-P50). We investigate how factor-selection varies between different scenarios.

Figure 6b reports the heat map of estimated risk prices using the OWL estimator while controlling stock sizes. First, micro stocks alter the market factor’s interpretation drastically. When micro stocks are all included to form test portfolios, market factor only plays a moderate role for asset prices; however, liquidity related factors dominate the chart. Market factor nonetheless consistently becomes the primary factor to drive asset prices once micro stocks are removed (at P20 and above levels). Second, liquidity related factors consistently top the chart in driving asset prices, particularly with the inclusion of small stocks. It shows that small firms face severe liquidity constrains, and investors demand risk premiums to bear that risk. Third, to be consistent with the finance literature, we consider the typical 20 percentile cut-off level to remove micro stocks. In this case, profitability and growth related factors, after liquidity related factors, become the second tier of factors that drive asset prices.

G Introduction of LASSO, adaptive LASSO, Elastic Net and OSCAR

Denote by y a $N \times 1$ vector of responses, by X a $N \times K$ data matrix and by $\beta = (\beta_1, \cdots, \beta_K)'$ a $K \times 1$ parameter vector. The LASSO (Tibshirani, 1996) estimator solves the problem

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \quad \left[\frac{1}{2}||y - X\beta||^2 + \lambda||\beta||_1\right], \tag{G.1}$$

where $||\beta||_1 = \sum\limits_{i=1}^{K} |\beta_i|$. The LASSO estimator can shrink the coefficients β_i of unimportant covariates to zeros. The Elastic net (EN) (Zou and Hastie, 2005) method solves the problem

$$\hat{\beta}_{EN} = \arg\min_{\beta} \quad \left[\frac{1}{2}||y - X\beta||^2 + \lambda\alpha||\beta||_1 + \lambda(1-\alpha)||\beta||_2^2 \right], \tag{G.2}$$

where $||\beta||_2^2 = \sum\limits_{i=1}^{K} \beta_i^2$. Elastic net combines the ℓ_1 norm (LASSO) and the ℓ_2 norm (Ridge) penalty together, which stabilizes the LASSO selections of β's when variables are correlated. Here, $\alpha \in (0,1)$ is a tuning parameter used to tilt the weight between the ℓ_1- and ℓ_2- shrinkage components. The adaptive LASSO (Zou, 2006) method minimizes the following function

$$\hat{\beta}_{adaLASSO} = \arg\min_{\beta} \quad \left[\frac{1}{2}||y - X\beta||^2 + \lambda \sum_{i=1}^{K} \frac{1}{|\hat{\beta}_{i,ada}|^{\gamma}} |\beta_i| \right], \tag{G.3}$$

where $\gamma > 0$ and $|\hat{\beta}_{i,ada}|$ is an adaptive weight for the i^{th} element in β, which is obtained through a first-stage estimation and typically based on the OLS estimate when it is feasible. Variables with small magnitudes in fist-stage estimated coefficients (i.e., small $|\hat{\beta}_{i,ada}|$) receive stronger penalty and γ controls the intensity of penalty for small parameters. λ controls the overall penalty level. The OSCAR (Octagonal shrinkage and clustering algorithm for regression) (Bondell and Reich, 2008) method solves this problem

$$\hat{\beta}_{OSCAR} = \arg\min_{\beta} \quad \left[\frac{1}{2}||y - X\beta||^2 + \lambda_1||\beta||_1 + \lambda_2 \sum_{i<j} \max\{|\beta_i|, |\beta_j|\} \right], \tag{G.4}$$

where $\sum\limits_{i<j} \max\{|\beta_i|, |\beta_j|\}$ compares all elements in β pair-wisely and penalizes more on the larger one. Bondell and Reich (2008) show that OSCAR method encourages factor clustering when they are correlated. Zeng and Figueiredo (2014) illustrate that by adopting a linear weighting scheme for ω, the OWL estimator encompasses the OSCAR estimator.

To gain some insights in this claim, we start from the OSCAR penalty term. Note that

$$\begin{aligned}\Omega_{OSCAR}(\beta) &= \lambda_1||\beta||_1 + \lambda_2 \sum_{i<j} \max\{|\beta_i|, |\beta_j|\} \\ &= \sum_i \underbrace{\lambda_1 + \lambda_2(K-i)}_{\text{linear decreasing weights}} \quad |\beta|_{[i]} = \sum_i \omega_i |\beta|_{[i]} \\ &= \omega' |\beta|_{\downarrow} \; = \; \Omega_{OWL}(\beta),\end{aligned}$$

With a linear weighting scheme for ω, the OWL penalty term encompasses the OSCAR penalty term. Furthermore, if we set $\lambda_2 = 0$, OWL encompasses LASSO.

References

Acharya, V. V. and L. H. Pedersen (2005): "Asset pricing with liquidity risk," *Journal of Financial Economics*, 77, 375–410.

Amihud, Y. (2002): "Illiquidity and stock returns: Cross-section and time-series effects," *Journal of Financial Markets*, 5, 31–56.

Asness, C., A. Frazzini, R. Israel, T. J. Moskowitz, and L. H. Pedersen (2018): "Size matters, if you control your junk," *Journal of Financial Economics*, 129, 479–509.

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013): "Value and Momentum Everywhere," *Journal of Finance*, 68, 929–985.

Babii, A., E. Ghysels, and J. Striaukas (2021): "Machine Learning Time Series Regressions With an Application to Nowcasting," *Journal of Business and Economic Statistics*, 1–23.

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

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

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of lasso and dantzig selector," *Annals of Statistics*, 37, 1705–1732.

Bondell, H. D. and B. J. Reich (2008): "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, 64, 115–123.

Buhlmann, P. and S. Van De Geer (2011): *Statistics for High-Dimensional Data - Methods, Theory and Applications*, Springer.

Cao, S. S., W. Jiang, J. L. Wang, and B. Yang (2021): "From Man vs. Machine to Man + Machine: The Art and Ai of Stock Analyses," *Columbia Business School Research Paper*.

Carhart, M. M. (1997): "On Persistence in Mutual Fund Performance," *The Journal of Finance*, 52, 57–82.

Chordia, T., R. Roll, and A. Subrahmanyam (2001): "American Finance Association Market Liquidity and Trading Activity Author (s): Tarun Chordia , Richard Roll and Avanidhar Subrahmanyam Source : The Journal of Finance , Vol . 56 , No . 2 (Apr ., 2001), pp . 501-530 Published by : Wiley for the America," *Journal of Finance*, 56, 501–530.

Cochrane, J. H. (2005): *Asset Pricing*, Princeton University Press.

——— (2011): "Presidential Address: Discount Rates," *Journal of Finance*, 66.

de Leeuw, J., K. Hornik, and P. Mair (2009): "Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods," *Journal of Statistical Software*, 32.

Fama, E. F. and K. R. French (1992): "The Cross-Section of Expected Stock Returns," *The Journal of Finance*, 47, 427–465.

——— (2008): "Dissecting Anomalies," *The Journal of Finance*, 63, 1653–1678.

——— (2016): "Dissecting Anomalies with a Five-Factor Model," *Review of Financial Studies*, 29.

——— (2018): "Choosing factors," *Journal of Financial Economics*, 128.

Fama, E. F. and J. MacBeth (1973): "Risk , Return , and Equilibrium : Empirical Tests," *Journal of Political Economy*, 81, 607–636.

Feng, G., S. Giglio, and D. Xiu (2020): "Taming the Factor Zoo: A Test of New Factors," *Journal of Finance*, 75, 1327–1370.

Figueiredo, M. and R. Nowak (2016): "Ordered weighted L1 Regularized Regression with Strongly Correlated Covariates: Theoretical Aspects," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 930–938.

Freyberger, J., A. Neuhierl, and M. Weber (2020): "Dissecting Characteristics Nonparametrically," *The Review of Financial Studies*, 33, 2326–2377.

Gospodinov, N., R. Kan, and C. Robotti (2014): "Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors," *Review of Financial Studies*, 27.

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

Gu, S., B. Kelly, and D. Xiu (2020): "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, 33, 2223–2273.

Harvey, C. R. and Y. Liu (2021): "Lucky factors," *Journal of Financial Economics*, 141.

Harvey, C. R., Y. Liu, and H. Zhu (2015): "... and the Cross-Section of Expected Returns," *The Review of Financial Studies*, 29, 5–68.

Hou, K., H. Mo, C. Xue, and L. Zhang (2021): " An Augmented q -Factor Model with Expected Growth* ," *Review of Finance*, 25.

Hou, K., C. Xue, and L. Zhang (2014): "Digesting anomalies: An investment approach," *Review of Financial Studies*, 28.

——— (2020): "Replicating Anomalies," *The Review of Financial Studies*, 33, 2019–2133.

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

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

Kock, A. B. (2016): "Oracle inequalities , variable selection and uniform inference in high-dimensional correlated random effects panel data models," *Journal of Econometrics*, 195, 71–85.

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

Lewellen, J. (2015): "The Cross-section of Expected Stock Returns," *Critical Finance Review*, 4, 1–44.

Lewellen, J., S. Nagel, and J. Shanken (2010): "A skeptical appraisal of asset pricing tests," *Journal of Financial Economics*, 96, 175–194.

Lintner, J. (1965): "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *The Review of Economics and Statistics*, 47, 13–37.

Ludvigson, S. C. (2013): "Advances in Consumption-Based Asset Pricing: Empirical Tests," in *Handbook of the Economics of Finance*, vol. 2.

Pastor, L. and R. F. Stambaugh (2003): "Liquidity Risk and Expected Stock Returns," *Journal of Political Economy*, 111, 642–685.

Shanken, J. (1992): "On the Estimation of Beta-Pricing Models," *Review of Financial Studies*, 5.

Sharpe, W. F. (1964): "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *The Journal of Finance*, 19, 425–442.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, 58, 267–288.

van Binsbergen, J. H., X. Han, and A. Lopez-Lira (2022): "Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases," *Review of Financial Studies.*

Vershynin, R. (2012): "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications.*

Yuan, M. and Y. Lin (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society. Series B*, 68, 49–67.

Zeng, X. and M. A. T. Figueiredo (2014): "The Ordered Weighted L1 Norm: Atomic Formulation, Projections, and Algorithms," *arXiv: Data Structures and Algorithms.*

Zou, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B*, 67, 301–320.