Dissecting the Factor Zoo: A Correlation-Robust Approach *

Chuanping Sun †

Abstract

This paper sheds light on a new perspective of the "factor zoo enigma", in which *factor correlation* prevails and worsens the *"curse of dimensionality"*. We introduce the Ordered-Weighted-LASSO (OWL) estimator, which circumvents complications from correlations: it can identify and group correlated factors while shrinking off useless/redundant ones. We show that OWL estimator is consistent and we derive the grouping condition for correlated factors. Empirical evidence suggests that *liquidity* related factors are primary to drive asset prices but cautions about a time-varying nature in factor selections. Out-of-sample analysis confirms superior performance of OWL-hedged portfolios compared with LASSO, Elastic-Net and Fama-MacBeth regression.

JEL classification: C38 C55 G12

Keywords: Factor Correlation, Cross-sectional Asset Pricing, Machine Learning, LASSO, Anomaly, Firm Characteristics, Stochastic Discount Factor

*I am indebted to Mario Figueiredo for his generous help and support. I am also grateful to Thomas Sargent, George Kapetanios, Anders Kock, Liudas Giraitis, Marcelo Fernandes, Emmanuel Guerre, Dick Van Dijk (discussant) and Kazuhiro Hiraki for discussion and insightful suggestions. I would like to thank participants at the Frontiers of Factor Investing conference 2018 at Lancaster University; the 29th EC^2 on big data econometrics with applications conference at the Bank of Italy; the Royal Economic Society annual meeting 2019 at Warwick University; the Econometric Institution international PhD conference 2019 at Erasmus University Rotterdam; the Econometric Society Asian meeting and European meeting 2019 at Xiamen University and Manchester University; the SoFiE annual meeting 2019 at Fudan University; the EFMA annual meeting 2019 at the University of Azores and the SITE 2019 workshop on Asset Pricing Theory and Computation at Stanford University for valuable comments and inspiring discussions. All remaining errors are mine.

†School of Economics and Finance, Queen Mary University of London, chuanping.sun@qmul.ac.uk

1 Introduction

In the past few decades, hundreds of anomaly variables have been proposed, claiming explanatory power to the cross section of average returns. However, Harvey et al. (2015), Mclean and Pontiff (2016) and Hou et al. (2017) find strong evidence that many of them are spurious and not replicable. Cochrane (2011) dubs this phenomenon the "factor zoo" and further argues that using characteristics related factors to explain the cross section of average returns is in disarray. He emphasizes the importance of finding factors that can *provide independent information about average returns*, and of distinguishing factors that can be summarized by others. Fama and French (2008) survey empirical methods for dissecting anomalies and point out that portfolio sorting and Fama-MacBeth regression (Fama and Macbeth (1973)) are traditionally employed to measure and test for factors' ability to drive asset prices. However, in the zoo of factors, portfolio sorting will encounter the *curse of dimensionality*, while Fama-MacBeth regression will suffer from multicollinearity.[1] Kleibergen (2009) cautions that the estimation of risk premium that results from a Fama-MacBeth regression is sensitive to collinearity of factor loadings. In the most recent development, a new strand of literature using machine learning techniques to solve high dimensional financial problems becomes prosperous. In particular, using the LASSO estimator (Tibshirani (1996)) to choose factors becomes the new mainstream in finance literature. However, it is often neglected that LASSO estimator is built on the assumption that covariates are *uncorrelated*. The mere fact that correlation prevails in the factor zoo brings in severe complications: Kozak et al. (2017) and Figueiredo and Nowak (2016) show that, with correlated factors, LASSO tends to pick a few (highly correlated) factors and wrongly shrink the rest to zeros. *Correlation* in high dimensionality deepens the "factor zoo enigma", so Cochrane (2011) points out: *"How to address these questions in the zoo of new variables, I suspect we will have to use different methods."*

This paper introduces a newly developed machine learning tool, *the Ordered-Weighted-LASSO* (OWL) to regularize this chaotic "factor zoo" which, to the best of our knowledge, is the first time applied in Finance.[2] OWL *permits* correlation among explanatory vari-

[1] In particular, for the second stage Fama-MacBeth regression, factor correlations measured by factor loadings are usually much higher than those measured by their time series (see Section 4 for a detailed illustration).

[2] OWL is built upon LASSO in the sense of ℓ_1-norm constraint, but they differ substantially in terms

ables, which distinguishes it from standard machine learning tools like LASSO. Factor correlations are common in high dimensional big data and they are of great importance in financial implications. DeMiguel et al. (2017) show that correlation between factors matters in a portfolio-optimisation perspective and find that six factors selected through their procedure are correlated. Asness et al. (2013) also find a negative correlation between 'momentum' and 'value' factors, which leads to superior portfolio performance. Cochrane (2005) points out that factor correlations dampen the implications of using risk premiums to infer priced factors. Cochrane (2011) shows that to determine which factors are useful to explain the cross section of average returns, we need to check *whether expected returns line up with the covariances of returns with factors.* In other words, it is the covariance measured by factor loadings, which is typically highly correlated, that really matters to infer priced factors. Hence, in the quest of finding useful factors to explain the cross section of average returns, factor correlations play an important role and should not be neglected.

The main empirical question of this paper is, under potentially highly correlated factors, how to select useful factors and shrink off useless and redundant ones? OWL provides a unified solution to this question. We first show OWL estimator is consistent with finite factors. We then derive the converge rate of OWL estimator when the number of factors grows to infinity. We also derive conditions under which correlated factors will be grouped together. This allows for factor-correlation identification and sparsity shrinkage, simultaneously. Like other shrinkage based estimators, it is, however, challenging to make direct statistical inferences on OWL estimator. Following DeMiguel et al. (2017) and Feng et al. (2019), we adopt a two stage (select and test) procedure to infer statistical significance of OWL estimators. In the first stage, we employ OWL to obtain a sparse set of useful factors. In the second stage, we propose a bootstrap based testing procedure to infer factor significance. In the presence of factor correlation, we bootstrap the orthogonalized asset returns to bypass multicollinearity issues (see Section 2.5 for a detailed discussion). This method is in line with Harvey and Liu (2017) in which they design a step-wise bootstrap testing method to select useful factors one by one. However, we test factors jointly because we are interested in their joint factor inference.

of handling correlated variables and optimization algorithms. More details see Section 2.

In a Monte Carlo experiment, we consider 90 candidate factors ($K = 90$) with correlations taken into account. We compare OWL with LASSO, Elastic Net, adaptive LASSO, and OLS estimators. We do this experiment in three settings: one with the number of test assets marginally larger than the number of factors ($N = 100$); one with a large number of test assets ($N = 1000$, $N \gg K$) which represents a low-dimensional setting; and finally, one with a small number of assets ($N = 70$, $N < K$) which represents a high-dimensional setting. OWL is the best performer especially when factors are correlated. Adaptive LASSO performs well in the low-dimensional setting, but it performs the worst in the high-dimensional setting: its performance depends heavily on a consistent estimator as an adaptive weight. LASSO, on the other hand, typically performs worst, especially when factors are correlated. LASSO estimator is severely affected by factor correlations, producing very unstable estimation and wrongly shrinking some useful factors to zeros, which is also pointed out by Kozak et al. (2017) and Figueiredo and Nowak (2016). Elastic Net does improve the performance of LASSO when factors are correlated, it stabilizes factor selections and reduces estimation errors. However, it is still substantially outperformed by OWL. When factors are not correlated, Elastic Net performs very similarly to LASSO. This experiment shows that in the high-dimensional factor zoo where factors are correlated, OWL is the best candidate.

Empirically, we initially consider 100 firm characteristics documented in Green et al. (2017), using CRSP and Compustat datasets, from January 1980 to December 2017. We first construct anomaly factors of each characteristic according to Fama and French (1992, 2015).[3] We obtain 80 anomaly factors. For test portfolios, we follow suggestions of Cochrane (2011), Lewellen et al. (2010) and Feng et al. (2019) by forming bi-variate sorted portfolios, and then combine them together as the grand set of test portfolios.[4] For robustness check, we also consider different methods of sorting and controlling for various scales of micro stocks, finding OWL is consistent in picking useful factors when a reasonable amount of micro stocks are removed.

[3]We first discard any characteristics having more than 40% missing data. We then use non-micro stocks to form decile portfolios at each point of time. If at any point of time, there are insufficient stocks to form the decile portfolios, we delete the characteristic.

[4]We single out 'size' as a common characteristic to form bi-variate sorted portfolios with the remaining ones, also see Feng et al. (2019).

The empirical results complement and challenge some common stances in asset pricing literature. First, we find moderate correlation among 80 anomaly factors, measured by their time series. Some beta related anomalies are highly correlated with other anomalies, including accruals, profitability, volatility and liquidities.[5] 15% of the correlation coefficients are higher than 0.5 (absolute value). However, that rises to 68% when factor correlations are measured by their factor loadings. So Kleibergen (2009) raises the concerns of multicollinearity issue for Fama-MacBeth estimator. Furthermore, from a different perspective, using Fama-MacBeth regression to test for factor risk premiums when factors are correlated is ill-positioned: it is inadequate to remove *redundant* factors, which contain no pricing information but earn positive risk premium (see Section 2.1 for a detailed illustration). Fama-MacBeth test would wrongly retain redundant factors. Cochrane (2011) emphasized the importance of finding factors that can provide independent information about average returns and of distinguishing from factors that can be summarized by others (i.e., redundant factors). These alarmingly high correlations among factors echo his outcry: in the zoo of new variables, we need to consider new methods.

Second, treatment for micro stocks is crucial for empirical interpretation. OWL identifies 'market' as the primary factor for the cross section of asset returns. This finding confirms the empirical evidence by Harvey and Liu (2017), and is consistent using either the value weighted or equal weighted method, *excluding* micro stocks. However, when micro stocks are included, the importance of market factor plummets. Micro stocks, although only taking up less than 10% of market capitalisation, constitute 56% of all stocks in the database. That rings alarms about methodologies using individual stocks as test assets: they may bias results because of the abundance of small stocks and their inferiority in aggregated market capitalisation. Hence, we adopt and advocate the use of sorted and pooled portfolios as the grand set of test portfolio as in Feng et al. (2019), controlling micro stocks. Sorted portfolios can efficiently avoid: 1) the "error in variable" bias; 2) missing data from individual stocks which is a great challenge for covariance matrix estimation; 3) micro stock issues caused by inferior stocks that little represent the market but dominate the estimation result.

[5]For this reason, Green et al. (2017) discard beta related anomalies in their factor library.

Third, *liquidity* related factors are the main drivers of the variation of cross sectional average returns. 'Illiquidity' (Amihud (2002)) is the most important anomaly factor, followed by 'standard deviation of traded dollar volume' (Chordia et al. (2001)). Their high correlation is identified by OWL. *Liquidity* related factors are more evident with smaller stocks, implying small firms face severe liquidity constraints. However, we should caution about the time-varying nature of factors that drive asset prices. Sub-sample estimation reveals that *liquidity* related factors are particularly evident after the 2000 internet-bubble bursts, while before that (1980 - 2000), 'profitability' and 'momentum' are the most important factors to drive asset prices, indicating a shift in economic characteristics. Since the object of this paper is to answer the question: *over a given period of time, what are the useful factors that drive asset prices?* It is immune to the time-varying nature of factors' ability to drive asset prices; however, it would be unfit if we use historical estimated factors (particularly over a long period) to predict future returns. In addition, some *'asset growth rate', 'profitability'* and *'investment'* related factors are also significant to explain the cross section of average returns. This finding is consistent with Hou et al. (2018).[6] Interestingly, the 'size effect' disappears during the 1980-2000 period, which is well documented: the size effect weakened after its discovery in the early 1980s (see Amihud (2002), van Dijk (2011) and Asness et al. (2018)). However, it 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, size effect resurfaces again, which echoes the discovery by Asness et al. (2018): *size matters, if you control your junk.*

Fourth, from an out-of-sample (OOS) perspective, we follow a similar procedure to Freyberger et al. (2019) to conduct the OOS exercise and we find strong evidence of time-varying factors between sub-samples: Sharpe ratios have substantially improved in the sub-sample estimations than that in the full-sample and this trend happens in all methods we considered. Comparing OWL with LASSO, Elastic Net and Fama-MacBeth regression methods, we find that, typically, Fama-MacBeth regression yields the worst Sharpe ratio due to factor correlations. When micro stocks are included, all methods produce similar Sharpe ratios. Meanwhile, the returns of hedge portfolios of all methods

[6]They add *'asset growth rate'* in their *q4* factor model and propose the *q5* factor model.

considered are left-skewed and yield exceedingly high kurtosis when micro stocks are included but, once they are excluded, the hedge portfolio returns are more "normal-like". OWL produces 20% to 30% higher out-of-sample Sharpe ratio than other methods once micro stocks are removed.

Related literature

This paper naturally builds on a series of papers devoted to identifying pricing factors. Fama and French (1992) propose the three-factor model consisting of a market return factor, a size, and a value factor that achieves enormous success. Carhart (1997) adds the momentum factor in Fama-French's three factor model that makes it the new standard among practitioners. Hou et al. (2014) explore the investment perspectives and propose the *q4* model which includes an investment factor, a profitability factor, and a size factor along with the market factor. Fama and French (2015) develop their own version of investment and profitability factors and expand the three-factor model to a five-factor model. Fama and French (2018) argue that an extra 'momentum' factor increases Sharpe ratio according to Barillas and Shanken (2018), and they suggest a six-factor model. Now after over half a century since the CAPM of Sharpe (1964) and Lintner (1965), hundreds of anomaly factors have been proposed, claiming explanatory power to the cross section of average returns. Harvey et al. (2015) document 316 factors and find most of them are results of data-snooping. Hou et al. (2017) try to replicate 447 anomaly factors, and find 64% to 85% of them are not replicable.

This paper also relates to a series of econometric papers devoted to asset pricing model testing. Fama and Macbeth (1973) put forward the two-pass regression method that has now become a standard practice in finance. Green et al. (2017) use Fama-MacBeth regression to find significant 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. 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 model misspecification robust test to tackle spurious factors, using a step-wise test to remove useless factors one by one. Kelly et al. (2019) propose

the instrumented PCA (IPCA) analysis by introducing observable characteristics that instrument for unobservable dynamic loadings. 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 (2017) suggest a step-wise bootstrap method to test for factors. In particular, at each step they pick a factor that has the best statistics (for instance, the t-stat), then bootstrap the null hypothesis that factor has no explanatory power by orthogonalizing asset returns with the factor. Pukthuanthong et al. (2018) propose a protocol to select factors: all factors should be correlated with principal components of test assets covariance matrix. However, this paper differs from other approaches by allowing correlations among factors, which is little discussed in the literature. The OWL estimator achieves sparsity selection and correlation identification simultaneously.

This paper also contributes to the vastly growing literature using machine learning techniques to solve financial problems. Tibshirani (1996) proposed LASSO that achieves dimension reduction within a convex optimisation problem. Since then, many adaptations and improvements have been made to achieve various targets. The literature about the LASSO family evolves rapidly. Yuan and Lin (2006) allow LASSO to shrink variables as groups by introducing the group LASSO. Freyberger et al. (2019) employ the adaptive group LASSO to find pervasive factors to explain the cross section of average returns. Zou (2006) introduces the adaptive LASSO by adding a consistent estimator as the weight of LASSO which makes the adaptive LASSO estimator consistent and enjoys the oracle property. Bryzgalova (2015) modifies the adaptive LASSO by replacing the adaptive weight (OLS estimator of risk premium) with factor loadings from the first pass of Fama-MacBeth regression. Feng et al. (2019) adopt the double selection LASSO of Belloni et al. (2014). In the first step they use LASSO to choose controlling factors with test assets; in the second step they use LASSO again to choose controlling factors with candidate factors yet to be tested; in the third step, they run OLS regression of test assets on the union of candidate and controlling factors selected from the first two steps. They make statistical inferences on the candidate factors in the third step. Fan and Li (2001) propose the smoothly clipped absolute deviation (SCAD) estimator so that it bridges the hard-thresholding and soft-thresholding. Ando and Bai (2015) employ SCAD to find Chinese stock predictors. Zou and Hastie (2005) combine the L_1 and L_2 norm and propose the

elastic net (EN), which achieves clustering selection of correlated variables. Kozak et al. (2017) employ EN in a Bayesian framework and find that sparse principle components can largely explain the cross section of the average returns. Gu et al. (2019) compare popular machine learning techniques used in empirical asset pricing problems and demonstrate large economic gains to investors using regressing trees and neuron networks.

Bondell and Reich (2008) propose the octagonal shrinkage and clustering algorithm for regression (OSCAR) by exploring the L_∞ norm of parameters pair-wisely to achieve clustered selection when variables are highly correlated. This paper is also closely related to Zeng and Figueiredo (2015), Figueiredo and Nowak (2016) in which they study the ordered and weighted LASSO (OWL) and reveal the close connection between OWL and OSCAR: by adopting a linear decreasing weighting scheme for the penalty term, OWL encompasses the OSCAR regularization. Zeng and Figueiredo (2015) apply OWL on image processing and attain significant noise deduction.

Bogdan et al. (2015) study the sorted L_1 penalized estimator (SLOPE) which is closely related to OWL. In fact, their design, before we define the weighting vector ω, is exactly the same. However, in their paper, the weighting vector for SLOPE is a function of normal distribution. The largest distinction between their paper and OWL is that they assume orthogonal design, that all variables are not correlated with each other, and they investigate the false discovery rate (FDR) to select factors. OWL differs from SLOPE in two distinctive perspectives: 1) a linear weighting vector for ω maps OWL to OSCAR, which allows clustering identification; 2) OWL permits correlation among variables, and can achieve correlation identification and factor shrinkage simultaneously.

2 Methodology

To study which factors jointly explain the cross section of average returns, we adopt the SDF method in Cochrane (2005). Section 2.1 explores the relation between risk price and risk premium and explains which one should be used to infer priced factors; Section 2.2 points out limitations of traditional methods when facing high-dimensionality and offers a remedy by imposing sparsity; Sections 2.3 and 2.4 introduce OWL estimator and discuss its statistical properties; Section 2.5 proposes a two stage testing procedure to

validate selected factors.

2.1 Risk price or risk premium?

Let m_t denote the stochastic discount factor (SDF)

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

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 mimic portfolio returns of non-traded factors. $f - E(f)$ is the demeaned factor return. b is a $K \times 1$ vector of the SDF coefficient, referred to as the *risk price*, a non-zero (zero) entry of b means the corresponding factor is (not) priced and b' is the transpose of vector b.

We want to draw inferences on the risk prices of factors. Finding useful factors is the goal of this paper, that is factors with non-zero risk prices and that directly drive the variation of SDF and contain pricing information. More specifically, they reflect the marginal utility of factors to explain the cross-section of average returns. Factors can also be useless or redundant. Useless factors are those whose risk prices are zero and which are uncorrelated with test assets. Redundant factors also have zero risk prices but they are correlated with some useful factors. In other words, they can be subsumed by other useful factors.

Risk premium refers to the free parameter in the second pass Fama-MacBeth regression: the first pass obtains the factor loadings by running time-series regressions of each asset; the second pass runs cross-sectional regressions of asset returns on factor loadings. Risk price and risk premium are directly related through the covariance matrix of factors, yet they differ substantially in their interpretation. Cochrane (2005) shows that b (risk price) and λ (risk premium) are related by equation:

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

Risk premium of a factor infers how much an investor demands to pay for bearing 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 with each other, that is, $E(ff')$ is a diagonal matrix, $b_i = 0$ (the i^{th} factor is not priced) implies $\lambda_i = 0$, and vice verse. However, this is not true when factors are correlated. Risk premium of a factor can be non-zero while the factor is not priced. A factor can earn positive risk premium by being correlated with a useful factor, even though its risk price is indeed zero. To give an example, suppose we have two factors f_1 and f_2, the covariance matrix is $E(ff') = \begin{pmatrix} 10 & 1 \\ 1 & 10 \end{pmatrix}$, and the first factor is priced and the second is not, that is $b_1 = 1 \neq 0$ and $b_2 = 0$, according to (2), $\lambda_1 = 10$, $\lambda_2 = 1$. Although factor f_2 is not priced it earns non-zero risk premium 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, and our goal is to find factors that are useful for the cross-sectional asset pricing, so we should use *risk price* to infer priced factors.

Let R_t denote the excess returns of a vector of N test assets, and $t = 1, ..., T$. The fundamental asset pricing equation states: $E(m_t R_t) = 0$ for any admissible SDF, m_t. However, when m_t is unknown and estimated from a model, the fundamental equation may not hold. The deviation from zero of the above equation is regarded as the pricing error. Let $m_t(b)$ denote the unknown SDF which depends on the unknown risk price b. Pricing error $e(b)$ can be written and simplified as:

$$\begin{aligned} e(b) &= E[R_t m_t(b)] = E(R_t)E(m_t(b)) + cov(R_t, m_t(b)) \\ &= E(R_t)E(m_t(b)) + r_0^{-1} cov(R_t, 1 - b'(f - E(f))) \\ &= r_0^{-1}(\mu_R - Cb), \end{aligned} \tag{3}$$

where $C = cov(R_t, f)$ is the $N \times K$ covariance matrix of excess return and factors and μ_R is the $N \times 1$ vector of the expectations of excess returns of test assets.

A quadratic form of the pricing error measures how far the candidate model deviates from the true model, which is similar to the Hansen-Jagannathan (HJ) distance. Let

$$Q(b) = \hat{e}(b)' \, W_T \, \hat{e}(b)$$

be the distance measure, where W_T is a weighting matrix and $\hat{e}(b)$ is an estimation of pricing error using sample analogues[7] of μ_R and C. Then we can recover b by minimizing $Q(b)$:

$$\hat{b} = \arg\min_b Q(b) = \arg\min_b (\mu_R - Cb)' W_T (\mu_R - Cb), \tag{4}$$

which gives

$$\hat{b} = (C' W_T C)^{-1} C' W_T \mu_R. \tag{5}$$

Since r_0 in (3) is a constant, it is irrelevant to the minimisation problem so it can be dropped out. Note that for the estimation of $\hat{b}$, one should use the sample analogues of μ_R and C, but for the ease of expression, we do not change their notations.

For the weighting matrix W_T, Ludvigson (2013) offers two choices for comparing models. First, $E(RR')^{-1}$, the inverse of the second moment of test assets returns, which corresponds to the well known HJ distance. The use of HJ distance is more appealing when facing limited asset choices (small N). The weighting matrix $E(RR')^{-1}$ accounts for and offsets the variations of test assets, producing stable estimators regardless of limited test assets. It is, however, challenging to obtain HJ distance when N is large: large T is required ($T \gg N$) to avoid near-singular matrix condition when estimating HJ distance, but the length of T is usually limited. Ludvigson (2013) advocates the second choice of W_T: the identity matrix if N is large. Additionally, when test portfolios represent particular economic interests, for instance, firm characteristic sorted portfolios, the identity matrix will be a better choice. Identity matrix does not tilt the weight to favour any test portfolios, each characteristic sorted portfolio will be and should be treated equally. In the empirical analysis, since we have plenty of sorted portfolios as test assets, we do not want to tilt the weight to favour any particular characteristics, so we will use the identity matrix as the weighting matrix.

2.2 Challenges and/or blessings of high-dimensionality

Cochrane (2011) points out that traditional methods like portfolio sorting to identify useful factors have fallen short in the high-dimensional world. Following Fama and French

[7] In the scope of this paper we are not interested in the population value of μ_R and C, so for the ease of expression, we use μ_R and C to denote their sample values.

(1992, 2008) to construct, for instance, 5 by 5 portfolios, and supposing n characteristics based anomaly factors need to be tested, we have to sort all stocks into 5^n portfolios. When n is small, for instance $n = 2$, it is handy to sort portfolios, and check the marginal distribution of returns on each characteristic. However, when n is large, for instance, $n = 10$, it is infeasible to sort stocks into $5^{10} \approx 9.8$ *million* portfolios. In reality, there are hundreds (or even thousands, if we consider the interaction between them) of factors been proposed.

For the Fama-MacBeth regression, there are several complications in high dimensional setting too. First, to obtain a consistent estimator one needs the number of test assets much larger than the number of factors ($N \gg K$), which is hard to be satisfied in a high-dimensional world. On the contrary, K is likely to diverge ($K > N$) under high-dimensionality, then the Fama-MacBeth regression becomes infeasible. Second, variables are likely correlated under high-dimensionality. As discussed in Section 2.1, when factors are correlated, unpriced factors can earn positive risk premium if they are correlated with priced factors (redundant factors), so Fama-MacBeth regression is likely to pick up redundant factors. Third, Kleibergen (2009) also cautions that Fama-MacBeth regression faces multicollinearity issues under high-dimensionality.

Nonetheless, empirical finance research has demonstrated strong evidence that many of the proposed factors are actually useless or redundant, see Harvey et al. (2015), Mclean and Pontiff (2016) and Hou et al. (2017). Thus, the sparsity assumption which originates from the machine learning literature becomes popular in finance to solve high-dimensional problems. Sparsity assumes that for K candidate factors, there are at most S of them which are non-zero (S is unknown and $S \ll K$). Tibshirani (1996) proposed the LASSO estimator which is a milestone to achieve sparsity. LASSO makes it even possible to estimate models even when $K > N$, which is infeasible by traditional methods.

However, LASSO is derived from the *orthogonal matrix design* assumption, which assumes all factors are uncorrelated with each other. In reality, however, correlation prevails between factors in high-dimensionality (see Section 4.3 for a detailed discussion). Kozak et al. (2017) and Figueiredo and Nowak (2016) have demonstrated that when factors are correlated, LASSO estimator is problematic: it yields unstable results and wrongly shrinks some useful factors to zero.

To circumvent the curse of dimensionality while taking account of factor correlations, we introduce a newly developed machine learning tool, the ordered and weighted LASSO (OWL) regularization, which explicitly allows for factor correlations.

2.3 The Ordered-Weighted-LASSO (OWL) estimator

OWL estimator is achieved by adding a penalty term in equation (4):

$$\hat{b} = \arg\min_{b} \frac{1}{2}(\mu_R - Cb)'W_T(\mu_R - Cb) + \Omega_\omega(b), \qquad \Omega_\omega(b) = \omega'|b|_\downarrow, \tag{6}$$

where $\omega \in \kappa$ is a pre-specified $K \times 1$ weighting vector, κ is a monotone non-negative cone, defined as $\kappa := \{x \in R^n : x_1 \geq x_2 \geq ... \geq x_n \geq 0\}$. $|b|_\downarrow$ is the absolute value of risk price, decreasingly ordered by its magnitude.

The weighting vector ω is restricted in a monotone non-negative cone, which makes the optimisation problem in (6) convex, and it is set to be linearly decreasing:

$$\omega_i = \lambda_1 + (K - i)\lambda_2, \qquad i = 1, ..., K, \tag{7}$$

where λ_1 and λ_2 are two tuning parameters. Zeng and Figueiredo (2015), Figueiredo and Nowak (2016) show that by adopting a linear weighting scheme, the OWL estimator maps to the OSCAR (Bondell and Reich (2008)) estimator, which has appealing properties of grouping highly correlated variables.[8] In order to solve (6), we use the proximal gradient descent algorithm. More details about this algorithm are included in the Online appendix.

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 five-fold cross-validation method to find tuning parameters. Given the grid values of λ_1 and λ_2, at each point on the grid, we first divide sample into five equal parts in their time series dimension. We use four parts (training sample) to estimate the model with OWL. After obtaining the estimated model, we forecast the returns of the fifth part (testing sample), and compute the out-of-sample root of mean squared forecast error (RMSE). We then repeat the same

[8] See the online Appendix for an introduction of OSCAR.

procedure five times by rotating the training samples and testing samples, and compute the average RMSE for this point on the grid. Turning parameters are determined by the smallest average RMSE on the grid.[9]

2.4 Statistical properties

This section discusses the statistical properties of the OWL estimator. We first show that, with some regularity conditions, when the number of factors K is finite, the OWL estimator is consistent. Then we allow K to go to infinity, with the sparsity assumption and restricted eigenvalue condition, we derive the convergence rate of OWL estimator, and hence the condition for consistent OWL estimation. Lastly, we derive the grouping condition under which two correlated factors will be grouped together. We also show the consistency of model selection using a thresholding method and other auxiliary results, but because of limited displaying space, we include more theoretical results in the appendix.

Suppose that

$$\mu_R = Cb^0 + \epsilon. \tag{8}$$

When the weighting matrix W in (6) is an identity matrix,[10] (6) can be written as[11]

$$\hat{b} = \arg\min_b \quad \frac{1}{N}||\mu_R - Cb||_2^2 + \frac{1}{N}\sum_{i=1}^{K} \lambda_1 + \lambda_2(K-i)]|b|_{[i]}, \tag{9}$$

where $|b|_{[1]} \geq |b|_{[2]} \geq ... \geq |b|_{[K]}$.

In order to derive the next theorem, we make some assumptions.

Assumption 1: The $N \times K$ covariance matrix of returns and factors C is normalized, such that $\hat{\Sigma} = \frac{C'C}{N} \underset{d}{\to} \Sigma$, where Σ is a full rank matrix, $\hat{\Sigma}_{j,j} = 1, \forall j \in \{1, ..., K\}$.

Assumption 1 requires a full rank Gram matrix $\hat{\Sigma}$, which restricts applications to a low dimensional case where the number of factor K is smaller than the number of assets

[9] In practice, once the tuning parameters are in a suitable region, the model selection is stable. In the empirical analysis, this region for tuning parameters is between 10^{-7} and 10^{-6}.

[10] When the weighting matrix W is not identity matrix, as long as it is a semi-positive definite matrix, we can use Cholesky decomposition of W, then we can map it into the identify matrix format.

[11] Note that the scaler "2" on the second term of (9) is dropped because it is negligible when turning parameter $\frac{\lambda_1}{N} \asymp \sqrt{\frac{\log K}{N}}$, which will be introduced in the next theorem.

N. Theorem 2.1 below is built on Assumption 1, which delivers the consistency property of OWL estimator in a typical low dimensional case ($K < N$).

Assumption 2: Suppose that ϵ in (8) follows a normal distribution such that $\epsilon \sim i.i.d.\, \mathbf{N}(0, \mathbf{I}\sigma^2)$, and $E(\epsilon' C^{(j)}) = 0$, where $C^{(j)}$ is the j^{th} column of C matrix.

The Gaussianity assumption imposed on ϵ is for the simplification of computing the probability p, which we will introduce in the theorem below. It can be relaxed to allow fat tails and correlations at the expense of more sophisticated statistical tools.

Theorem 2.1 (Consistency of OWL)**.** *Let Assumptions 1 and 2 be satisfied. Suppose that* $t > 0$, $\lambda_0 = 2\sigma\sqrt{\dfrac{t^2 + 2\log K}{N}}$, λ_1 *and* λ_2 *are such that* $\dfrac{\lambda_1}{N} \geq \lambda_0$, $\lambda_1 = o(N)$ *and* $\lambda_2 = o(N)$.

Then with probability at least

$$p = 1 - 2\exp(-\frac{t^2}{2}), \tag{10}$$

the estimator $\hat{b}$ *satisfies*

$$(\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0) \leq \left(\lambda_0 + \frac{\lambda_1 + \lambda_2(K-1)}{N}\right)||b^0||_1. \tag{11}$$

In addition, if K *is fixed, in* λ_0, $t \to \infty$, *and* $N \to \infty$, *then*

$$||\hat{b} - b^0||_2 \to \mathbf{0}.$$

Proof: see appendix A.1.

Theorem 2.1 shows the consistency of OWL estimator when K is finite and offers an upper bound in (11) of the estimation error of the OWL estimator $(\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0)$. It is derived in a low dimensional setting where the number of factors is small while the number of observable assets is large ($K \ll N$).

Next, we consider the high dimensional setting, where we allow the number of factors K to grow to infinity and we obtain the convergence rate of OWL estimator and the conditions for consistent estimation. With $K \gg N$, the Gram matrix $\hat{\Sigma}$ will be singular. In order to derive the converge rate, we impose Assumption 3 and 4.

Assumption 3 (Sparsity): Denoted by S the number of non-zero parameters in $b^0 = \{b_1^0, b_2^0, \cdots, b_K^0\}$. We assume that $S \ll K$, and $S\sqrt{\frac{\log K}{N}} = o(1)$.

Let $s_0 \subset \{1, \cdots, K\}$, denote $|s_0|$ 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 4 (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

$$\phi_0^2 := \min_{\substack{s_0 \subset \{1,\ldots,K\} \\ |s_0| < K}} \min_{\substack{b \in R^K \setminus \{0\} \\ ||b_{s_0^c}||_1 < 3||b_{s_0}||_1}} \frac{b' \hat{\Sigma} b}{||b_{s_0}||_2^2} > 0. \tag{12}$$

The restricted eigenvalue condition implies the compatibility condition in Buhlmann and Van de Geer (2011) which is the key requirement to establish Theorem 2.2 below. See the on-line Appendix for the motivation and the derivation of the compatibility condition.

Theorem 2.2 (Convergence rate of OWL)**.** *Let Assumptions 2, 3 and 4 be satisfied. Suppose that* $t > 0$, $\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log K}{N}}$, λ_1 *are such that* $\frac{\lambda_1}{N} \geq 2\lambda_0$, $\lambda_1 = o(N)$. *Then with probability at least*

$$p = 1 - 2\exp(-\frac{t^2}{2}), \tag{13}$$

$\hat{b}$ *satisfies*

$$(\hat{b} - b^0)' \hat{\Sigma} (\hat{b} - b^0) + \frac{\lambda_1}{N} ||\hat{b} - b^0||_1 \leq 4(\frac{\lambda_1}{N})^2 S/\phi_0^2 + 2\frac{\lambda_2}{N}(K-1)||b^0||_1. \tag{14}$$

In addition if $\lambda_2 = O(\frac{S \log K}{K})$ *, then*

$$||\hat{b} - b^0||_1 = O(\sqrt{\frac{\log K}{N}} S), \qquad ||\hat{b} - b^0||_2 = O(\sqrt{\frac{S \log K}{N}}). \tag{15}$$

Proof: see appendix A.2.

Theorem 2.2 establishes the convergence rate of OWL estimator in a high dimensional setting, where the number of factors can be much larger than observable assets (i.e. the factor zoo).

To this point, we have obtained convergence rate of OWL estimator when both K and N go to infinity. Following a similar argument from Kock and Callot (2015) by utilizing the L_∞ bound and the convergence rate, we can introduce a thresholding estimator $\tilde{b}$ based on $\hat{b}$, in which $\tilde{b}$ can select non-zero coefficients as the true ones. See appendix B for more details on definition and properties of $\tilde{b}$.

Next, we investigate the grouping condition under which correlated factors will be grouped together, i.e. elements of $\hat{b}$ in (6) will be assigned with similar coefficients.

Theorem 2.3 (Grouping). *Let f_i and f_j be the i^{th} and j^{th} factors, and $\hat{b}_i$ and $\hat{b}_j$ are OWL estimates of risk prices of factor f_i, f_j. Let $\sigma(f_i - f_j)$ denote the sample standard deviation of the vector $f_i - f_j$, and μ_R, σ_R be the sample mean and standard deviation of test assets. If*

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

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

Proof: see appendix A.3.

Theorem 2.3 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, $\hat{b}_i \approx \hat{b}_j$), which is consistent with economic intuitions. Second, the hyper parameter λ_2 in (7) has direct impact on factor grouping: large λ_2 encourages grouping.[12] Third, the sample mean (μ_R) and sample 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: weak factors are equally inadequate to explain a set of less informative test assets.

Theorem 2.3 shows that OWL can identify and group correlated factors. On the other hand, estimators that assume orthogonal structure of factors, for instance LASSO, is severely affected by factor correlations (see Section 3 for a detailed illustration). Green et al. (2017) delete highly correlated factors from the factor zoo in a Fama-MacBeth regression framework when they investigate which factors are useful to price the US

[12] From a geometric perspective (see Zeng and Figueiredo (2015)), large λ_2 makes the atomic norm of OWL more pointy, which encourages grouping. A geometric interpretation and more details are included in the online appendix.

stock market. However, it is difficult to define a threshold to screen out factors by their correlations and deletion of those factors lacks some economic justification.

We want to emphasize again that the main contribution of this paper is that we show OWL estimator *permits* high correlations among factors: it is robust to factor correlations and it can further identify those correlations and group together highly correlated factors. No factor-trimming/screening is required. OWL achieves correlation identification and sparsity shrinkage simultaneously.

2.5 Two stage selection procedure

This section proposes a two stage selection procedure to test the significance of factor coefficients following a similar approach of Harvey and Liu (2017). In the first stage, OWL estimator allows to selects a sparse number of factors; in the second stage, a bootstrap testing procedure will be implemented to infer factor significance.

Considering high correlation between OWL selected factors, we propose a bootstrap test that is robust in collinearity: instead of bootstrapping asset returns and factors to get the standard error of the slope coefficients, we first orthogonalize asset returns with factors, then we bootstrap the orthogonalized returns and factors. This method is in line with Harvey and Liu (2017) in which they use an orthogonal bootstrap method to select factors step by step. However, their step-wise selection method usually yields very conservative results: only two or three factors are tested as significant to explain the cross section of average returns. Instead, we test factor significance jointly, because we are interested in their joint inferences.

In particular, suppose we obtain a sparse number of factors with non-zero coefficients from OWL (after the first stage). We first compute the covariance matrix of survival factors and test assets denoted by C. Let μ_R denote the average returns of test assets. We first regress μ_R on C to obtain the t-statistic (t_{stat}) of estimated slopes and the residual series e. We then draw sub-samples with replacement from C and e, and call them C^* and e^*. Then we regress e^* on C^*, compute and save t^*_{stat}. Since e is orthogonal to C, t^*_{stat} represents the t_{stat} distribution under the null hypothesis, that is factors can not explain the correspondent variable e^*. We then compare t_{stat} estimated from real data with t^*_{stat} distribution. If t_{stat} exceeds 95 percentile of t^*_{stat} distribution, we then

declare the associated coefficient is significant.

3 Simulation

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

3.1 Simulation design

In our experiment, 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, $K/3$ are highly correlated, and $K/3$ are uncorrelated.

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

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

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 themselves they are correlated with a correlation coefficient ρ_1, ρ_2 or ρ_3, but factors in different blocks are uncorrelated with each other.

We set the values of ρ_1, ρ_2 and ρ_3, and then randomly generate an $N \times K$ matrix C, that each column of C follows a standard normal distribution. Then multiply it with the Choleski decomposition of ρ to obtain the covariance matrix C, denoted as $simC$.

We further set an oracle value for b (risk price). Then we simulate the cross section of average returns as $\mu_R = simC * b + e$, where e is an error vector, with the scale about 10% of $simC$, $e \sim N(0, 0.01)$.

Finally, we estimate risk price with simulated data $simC$ and μ_R using OWL, LASSO, adaptive LASSO, Elastic Net, and naive OLS estimators.[13] Then we compare these estimators with the pre-specified oracle value of b.

3.2 Simulation results

In the first experiment, we consider 90 candidate factors ($K = 90$). 30 of them (block 1) are useful factors which are also highly correlated ($b \neq 0,\ \rho_1 = 0.9$); 30 of them (block 2) are useless/redundant factors, which are also highly correlated ($b = 0,\ \rho_2 = 0.9$); and 30 of them (block 3) are useful factors but not correlated ($b \neq 0,\ \rho_3 = 0$). There are 100 test assets ($N = 100$).

[Figure 1 about here.]

Figure 1 reports the plot of OWL estimator over 90 factors along with other benchmarks and the oracle value (black). The upper left panel displays the plots of all factors. The remaining three panels display the detailed plot for each of these three blocks. The upper right panel displays the plot of useful factors which are highly correlated. In the presence of high correlation, LASSO estimator performs poorly with highest estimation errors. Adaptive LASSO is strongly governed by the adaptive weights which is set as the OLS estimator: it exhibits very similar behaviour to 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

[13] See the online Appendix for an introduction of LASSO, adaptive LASSO, and Elastic Net (EN) estimators. OLS estimator is only included if $N > K$.

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 useless/redundant factors which are also 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., OLS estimator) and fails to set many useless/redundant factors to zeros.

The bottom right panel displays all estimators of useful factors which are not correlated. Again, LASSO and Elastic Net are the worst performers which yield 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.

For the robustness check of this experiment, we repeat the simulation multiple times, and report the deviation of each estimator from the oracle values. Because of limited display space, we put the robustness check in the online appendix.

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 2 about here.]

Figure 2 reports the plot of OWL estimator along with 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 to estimate 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 factors, OWL produces the most accurate estimation (zero errors, in this case). With uncorrelated factors, OLS and adaptive LASSO is undoubtedly the best estimators, followed tightly 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 weight 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 observations.

[Figure 3 about here.]

Figure 3 reports estimation results of each methods along with the oracle value. Once $K > N$ the naive OLS becomes infeasible, thus we remove it from the candidate list. Meanwhile, we use the LASSO estimator as the adaptive weight for adaptive LASSO estimator. 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 estimators and wrongly shrinking many useful factors to zero. Interestingly, we find significant improvement of Elastic Net compared to LASSO. However, despite of substantial improvement comparing to LASSO, Elastic Net is still substantially outperformed by OWL. Adaptive LASSO, using the LASSO estimator as the adaptive weight, performs even worse than LASSO: it wrongly shrinks many useful factors to zeros.

In case of useless factors, all machine learning methods do a good job to shrink useless factors to zeros, except a couple of outliers for LASSO, Elastic net and OWL, however, the deviation from zero is moderate and the majority of useless factors are shrunk to zeros.

In case of uncorrelated useful factors, OWL becomes the best performer producing the smallest estimation error, whereas the adaptive LASSO using LASSO as first stage weight, performs the worst. It shrinks many useful factors to zeros and produces very volatile estimation. Adaptive LASSO is strongly influenced by the adaptive weight: in a low-dimensional setting where $N \gg K$, adaptive LASSO combines the merits of OLS and LASSO estimators, producing unbiased yet sparse estimators. However, in a high-dimensional setting, where OLS is infeasible, adaptive LASSO amplifies the shrinkage effect from LASSO, producing over-sparse estimators. In case of uncorrelated factors, Elastic Net has very moderate improvement on LASSO. They are producing very similar

estimators.

These experiments confirm the poor performance of LASSO when factors are correlated. Elastic Net does improve the performance of LASSO when factors are correlated. Adaptive LASSO is a good candidate in a low-dimensional setting where $N \gg K$; however, it performs poorly (the worst) in a high-dimensional setting where $K > N$ when OLS becomes infeasible. OWL estimator is consistently performing well particularly in the correlated settings, indicating that OWL is the best candidate in the high-dimensional "factor zoo".

4 Empirical analysis

This section applies the two-stage procedure on 80 anomaly factors to infer which factors are priced and can explain the cross section of average returns in stock market. We first introduce the datasets, followed by a detailed account of the construction of anomaly factors and test portfolios. We consider both value weighted and equal weighted methods, controlling micro stocks. Following a similar method of Feng et al. (2019), we construct pooled bi-variate sorted portfolios as test assets.

4.1 Data

We use the U.S. stock data from the Center for Research in Security Prices (CRSP) and Compustat database[14] to construct anomaly variables and test portfolios because of their availability and better data quality. 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 online data library. All anomaly variables are demeaned and scaled to have the same standard deviation with the market factor.

We consider 100 firm characteristics described in Green et al. (2017),[15] while deleting characteristics that have more than 40% missing data. Then, for each remaining char-

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

[15]We are grateful to Jeremiah Green for providing SAS code to compute firm characteristics. We modified the SAS code to cope with only CRSP and Compustat database.

acteristic, we sort stocks into decile portfolios at each month using uni-variate sorting. Micro stocks, defined as market capitalisation smaller than the 20 percentile of NYSE listed stocks, are removed. Although micro stocks only account for less than 10% of aggregated market capitalisation, they constitute about 56% of all stocks in the database, implying that small stocks would distort the interpretation of the aggregated market capitalisation if not removed, also see Hou et al. (2014) and Fama and French (2015). Then, anomaly factors are computed as the spread returns between the top and the bottom decile portfolios. Characteristics having 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 anomaly variables are top decile return minus the bottom decile return. That will end up with some slight difference with 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. The interpretation of the sign of coefficients should be looked at together with the sorting order when forming anomaly variables. Overall, we obtain 80 anomaly factors. See Table 1 for a detailed description.

[Table 1 about here.]

4.2 Bi-variate sorted portfolios as test assets

Regarding test assets, there is a debate in the literature about using either individual stocks or sorted portfolios as test assets. Harvey and Liu (2017) use individual stocks with bootstrap method to test for predictability of anomaly factors, and they find that only two or three anomaly factors can significantly predict asset returns. Lewellen (2015) employed Fama-MacBeth to test for anomaly factors with individual stocks. However, others argue that individual stocks 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, empirical work shows

that "Shanken's correction" is minimal in small samples. On the other hand, Fama and French (2008), Hou et al. (2014), Feng et al. (2019) advocate sorted portfolios as test assets. Individual 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 some similar characteristics, which would mitigate the EIV problem. Hence, using sorted portfolios as test assets is an alternative (arguably better) way to avoid EIV.

Yet, the biggest drawbacks of using individual stocks stem from missing data and micro stocks. It is inevitable, over a long period, to have new firms entering and old firms exiting, that will result in continuous missing data. Discontinuity of data can bias the estimation of covariance matrix of factors and test assets, which is essential for factor inference. A possible remedy could be deleting all stocks with any missing data. However, that will leave only 375 stocks during the period between January 1980 and December 2017, which is insufficient to represent the stock market. A less extreme treatment could be setting up a threshold for missing data: first, delete stocks with many missing data while keeping stocks with a few (depending on the threshold) missing data then, when estimating covariance matrix, delete rows with any missing data. However this treatment will lead to imprecise estimation of covariance matrix. It is particularly challenging to implement in an out-of-sample framework.

Using sorted portfolios, however, can circumvent this shortcoming. Portfolios are formed at each point of time according to certain characteristics, then portfolio returns are weighted averages of (varying) stocks in each portfolio, that guarantees continuity of portfolio returns.

Micro stocks bring up another 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 capitalisation. If using individual stocks to gauge factor impact, it is inevitable to distort the market implications: micro stocks, as long as individual stocks are concerned for test assets, will dominate the estimation result. Big stocks which have much larger impact on market price fluctuation will be out-weighted by the large number of small stocks.

Portfolio sorting, however, can circumvent this issue by using the value weighted method. First, micro stocks can be removed before sorting. Then returns of each sorted

portfolio can be computed by the weighted average of stocks returns where the weights reflect their market capitalisations.

Fama and French (1992, 2008, 2015) 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, for instance, the bi-variate sorted 25 portfolios, 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. Feng et al. (2019) construct a large set of combined portfolios as test assets. In particular, they single out 'size' characteristic and combine it with the remaining characteristics to form five by five bi-variate sorted portfolios and pool them together. 'Size' has been widely acknowledged as an important characteristic in asset pricing literature. Fama and French (1992, 2015), Hou et al. (2014), Carhart (1997) all include the 'size' and the 'market' factors in their models.

To strike a balance between using sorted portfolios and individual stocks as test assets, we follow Feng et al. (2019) by singling 'size' out as a common characteristic, together with the remaining characteristics to form bi-variate sorted 25 portfolios. We drop any test portfolios which have insufficient stocks (due to missing data) to sort. Finally, we group them together, which amounts to 1927 test portfolios.

4.3 Factor correlation

[Figure 4 about here.]

Figure 4a displays the heat map of factor correlation coefficients matrix measured by their time series. It suggests that 16% of factors exhbit correlation coefficients (absolute value) greater than 0.5. In particular, beta related characteristics are highly correlated with factors associated with liquidity, profitability, investment, and other financial ratios. Green et al. (2017) excluded beta related factors as candidate factors because of their high correlation profile with other factors.

Figure 4b displays the heat map of factor correlation coefficients matrix measured by

factor loadings. It exhibits much higher correlation compared to Figure 4a: 64% correlation coefficients (absolute value) are greater than 0.5, implying serious multicollinearity issues if standard Fama-MacBeth regression is employed. Cochrane (2011) points out that *we need to find whether expected returns line up with covariances of returns with factors*, implying that correlation measured by factor loadings really matters to infer priced factors and Fama-MacBeth regression would be ill-positioned to infer priced factors.

4.4 Which factors matter?

Considering high correlation among factors, we apply the two-stage procedure to select useful factors from the 81 candidate factors.[16] We first employ OWL to shrink off useless/redundant factors, obtaining a sparse number of survival factors. In the second stage we use bootstrap method described in Section 2.5 to find significant factors.

[Table 2 about here.]

Table 2 reports the result of the two-stage procedure to find factors that explain the cross section of average returns. The first 5 columns are estimated using the full sample, ranging from January 1980 to December 2017; columns 6-7 report results from 1980 to 2000, and columns 8-9 from 2001 to 2017. Both the value weighted (vw) and equal weighted (ew) methods are considered. In order to gauge the impact of small stocks, we consider three thresholds for micro stocks. Before sorting test portfolios, we screen out stocks with market capitalisation smaller than 20, 30 or 40 percentile of NYSE listed stocks. This table lists all anomaly factors selected by the two-stage procedure 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).

'Size' (mve) 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 show that OWL can correctly identify relevant factors.

[16] We include market factor along with 80 anomaly factors, total 81 candidate factors.

Amihud (2002)'s 'illiquidity' (ill) is the most important factor that drives variations of test asset returns. Its explanatory power 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. That implies small firms face severer liquidity constraints, and demand risk premiums to compensate for bearing the risk.

'Standard deviation of dollar volume' (std_dolvol) follows 'illiquidity', becoming the second most important anomaly factor. 'Standard deviation of dollar volume' is strongly correlated with 'illiquidity'. Both are proxies for liquidity risk. Recognising their high correlation, OWL groups them together by assigning them with similar coefficients.

Liquidity as a risk source for stocks that commands risk premiums has been documented extensively in the literature. Pástor and Stambaugh (2003) show that market-wide liquidity is a state variable important for asset pricing. Average returns on stocks with high sensitivities to liquidity exceed that for stocks with low sensitivities by 7.5%, while controlling for 'market', 'size', 'value' and 'momentum' factors. Acharya and Pedersen (2005) unified several empirical findings on liquidity in an equilibrium model, where illiquidity is modelled by per-share cost of selling security. They decompose liquidity risk premium into three components: 1) the covariance of individual stock's illiquidity to the aggregated market illiquidity, which implies that an investor requires risk premium for a stock that is illiquid while the market is illiquid; 2) the covariance between individual stock's return and market-wide illiquidity, which is consistent with Pástor and Stambaugh (2003)'s findings; 3) the covariance between individual stock's illiquidity and market returns, which implies investors are willing to pay a premium for stock that is liquid while the market return is low.

'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. (2018)'s new $q5$ model, in which they add 'asset growth rate' as a fifth factor after their famous $q4$ model (see Hou et al. (2014)). We also find 'asset growth rate' is more prominent with smaller stocks: equal weighted method shows stronger impact of 'asset growth rate' on stock returns.

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 related to liquidity risk. Beyond that, 'Return on invested capital' (roic), and 'return on assets' (roaq) are profitability related factors and are also significant to explain the cross section of average stock returns.

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 to explain 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. Meanwhile, 'momentum' and 'profitability' related factors primarily drive asset prices between 1980 and 2000.

Interestingly, during 1980 to 2000, with 20-percentile-micro-stocks excluded, we find 'size' (mve) is not selected by OWL, which makes it the only exception from all estimations. This phenomenon is well documented in the literature (see Amihud (2002), van Dijk (2011) and Asness et al. (2018)): the size effect weakened after its discovery in the early 1980s. However, when removing 40-percentile-micro-stocks, size effect is evident again, which implies the vanishing of 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.*

4.5 Robustness check

In this section, we check whether liquidity related factors are robust in explaining the cross section of asset returns as well as how small stocks affect factors' interpretations. Because of the limitation of space, we place Section 4.5 in the online Appendix.

4.6 Out-Of-Sample analysis

In this subsection, we will evaluate the performance of OWL selected factors in an out-of-sample (OOS) context. OOS method is less prone to data mining and gains robustness against in-sample over-fit. Freyberger et al. (2019) point out that OOS exercise ensures

that in-sample over-fit does not explain superior performance in model selection. Although the five-fold cross validation method used for evaluating OWL hyper parameters ensures an OOS metric by construction, the choice of factors is based on the overall sample. It is possible that factors selected to explain the cross-sectional returns for one period do not hold well for another period.

Considering the time-varying nature of factors, we also evaluate the OOS performance for two sub-samples, divided before and after 2000. We report the first five factors with highest estimated coefficients (absolute value). Concerning over-fitting in out-of-sample exercise typically yields poor performance, we consider a five-factor model for out-of-sample prediction.[17]

[Table 3 about here.]

Table 3 shows the five most important factors selected using various methods in different samples, controlling for micro stocks. We consider both the full sample estimation and the sub-sample estimation. We can find obvious difference in selected factors between full-sample and sub-samples, as well as between sub-samples. In addition, controlling micro stocks has big impact on factor selection too. While including all micro stocks (P00), OWL and together with other methods select a mixture of 'liquidity', 'profitability' and 'momentum' related factors. However, once remove micro stocks (P20 and P40), we can find some patterns in selected factors: OWL finds 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. Moreover, 'mkt' as a primary factor selected by OWL when excluding micro stocks, is missing by other methods, which is counter-intuitive. Once removing micro stocks, market factor should be the dominating factor driving asset prices since idiosyncratic risks have been largely reduced. However, LASSO, Elastic Net and

[17]We also considered a four-factor and a three-factor model for robustness check. We find a four-factor model performs slightly better than the five-factor model in predictions. However, due to limited reporting space, we do not include them and they are available on request.

Fama-MacBeth all fail to identify 'mkt' as an important factor, which is due to the high correlation between 'mkt' and other factors.

After selecting the driving factors in each samples. We want to compare the out-of-sample performance between various methods. In particular, we follow a similar procedure to Freyberger et al. (2019) to form hedge portfolios using a rolling window scheme to predict returns of each test assets. Rolling window size is 120 months (10 years). Specifically, at the end of the estimation window, we regress each test asset on factors selected by each method, but one period lagged. 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 hedge portfolio returns in the out-of-sample period as well as the Sharpe ratio. In addition, we also compute the out-of-sample slope and R^2 following Freyberger et al. (2019). Specifically, we regress realized returns of test assets on the predicted returns at each out-of-sample period, and we report the average slope and R^2. Slope value close to one or having a high R^2 indicates good predictive power.

[Table 4 about here.]

Table 4 reports performance scores of OWL, LASSO, Elastic Net and Fama-MacBeth estimators in full/sub-samples while controlling micro stocks. The first five columns of data are reporting the out-of-sample Sharpe ratio and the first four moments of the hedge portfolio returns. The last two columns are reporting the average slope and R^2 by regressing the realized returns on the predicted returns month by month.

In the case of including micro stocks, all methods report similarly high Sharpe ratios, which is also reported by Freyberger et al. (2019) and Lewellen (2015); however, the differences between each methods are minor and negligible. Once removing stocks that are smaller than 20/40 percentile of NYSE listed stocks, we can find OWL reports superior Sharpe ratios than other methods. In addition, sub-sample estimations while controlling

for micro stocks report higher Sharpe ratios than in the full sample, indicting a changing of characteristics of driving factors in the full sample period. Also, we find the superior performance of OWL while removing micro stocks are more evident in the sub-samples.

Since Elastic Net encompassed LASSO, they yield very similar scores while removing micro stocks and in sub-sample estimations. Generally speaking Elastic Net produces more robust result than the LASSO, indicating unstable LASSO selection in factors: as shown in the Monte Carlo simulation, LASSO performs poorly when factors are correlated. Fama-MacBeth regression typically performs the worst: Fama-MacBeth regression is severely affected by factor correlation and the robustness of t-statistics is corrupted by weak factors (Kleibergen (2009)).

For Low risk-aversion investors, OWL hedged portfolios also produce highest mean returns once micro stocks are removed. Generally speaking, in the full sample estimation, all method are skewed left once including micro stocks and skewed right once excluding micro stocks and yield exceedingly high kurtosis. However, hedge portfolio returns are more "normal-like" once considering the sub-sample estimations.

The out-of-sample slope and R^2 do not differ much between methods in the full-sample estimation. However, once removing micro stocks, OWL method shows superior fit for the slope, particularly in the first sub-sample period with stocks smaller than 40 percentile of NYSE listed been removed.

5 Conclusion

In the zoo of factors, using traditional methods to find factors that can provide independent information about average returns faces tremendous challenges. In addition, correlations make the matter worse: among 80 anomaly factors we considered, 64% of them exhibit correlation coefficient (absolute value) greater than 0.5 in the second stage Fama-Macbeth regression. Factor correlations cause severe complications in some popular estimators. LASSO yields very unstable estimators of factor coefficients and wrongly shrinks some useful factors to zeros.[18] In the presence of correlated factors, Cochrane (2005) shows that Fama-MacBeth regression is inadequate to remove redun-

[18] See Kozak et al. (2017) and Figueiredo and Nowak (2016).

dant factors. Furthermore, Kleibergen (2009) cautions about the multicollinearity issues in Fama-MacBeth estimator.

The OWL estimator, on the other hand, circumvents the complications from correlated factors. In particular, OWL permits correlated variables and achieves correlation identification and sparsity shrinkage simultaneously. Monte Carlo experiments confirm the superior performance of OWL against other methods, especially when factors are correlated. Empirical analysis shows that 'liquidity' related factors play an important role to drive asset prices. Furthermore, sub-sample estimations imply a shift in economic characteristics and reveal a time-varying nature in factor selections: between 1980 and 2000, 'profitability' and 'momentum' are primary factors that drive asset prices, while 'liquidity' related factors are dominant between 2000 and 2017. An out-of-sample analysis in line of Freyberger et al. (2019) confirms the time-varying nature in factor selections: the out-of-sample Sharpe ratio increases substantially in sub-samples than that in the full-sample across all methods, while micro-stocks are removed. Meanwhile, OWL method demonstrates superior performance over LASSO, Elastic Net and Fama-MacBeth regression in the out-of-sample analysis.

Finally, note that the purpose of this paper is not to find a parsimonious asset pricing model, but to identify a set of sparse factors, potentially highly correlated, to explain the cross section of average returns given a certain period. With that in mind, our procedure is particularly useful for factor investing: OWL can identify correlated factors that jointly drive stock returns, which can be further utilized to form portfolio strategies, see Asness et al. (2013) for instance. Note that a shift in economic characteristics can affect factors' capability and interpretation to drive asset prices, so the interpretability of factors to explain the cross section of average returns is pertained in a suitable time frame. Possible future works can explore a time-varying version of OWL estimator, which allows factor coefficient to vary across time. Furthermore, OWL is a general tool useful for sparse selection and correlation identification in high-dimensional variables. Some stock returns exhibit high correlation and can be utilized for superior portfolio performance (see DeMiguel et al. (2014) for instance), so future research can explore portfolio selection strategies, where individual stocks are regularized by OWL.

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.

Ando, T. and J. Bai (2015): "Asset Pricing with a General Multifactor Structure," *Journal of Financial Econometrics*, 13, 556–604.

Asness, C. S., A. Frazzini, R. Israel, T. J. Moskowitz, and L. H. Pedersen (2018): "Size Matters, If You Control Your Junk," *Journal of Financial Economics*, 0, 1–31.

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

Barillas, F. and J. Shanken (2018): "Comparing Asset Pricing Models," *The Journal of Finance*, LXXIII, 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.

Bogdan, M., E. van den Berg, C. Sabatti, W. Su, and E. J. Candès (2015): "Slopeadaptive variable selection via convex optimization," *Annals of Applied Statistics*, 9, 1103–1140.

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

Bryzgalova, S. (2015): "Spurious Factors in Linear Asset Pricing Models," *LSE Working Paper*, 1–78.

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

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

Chordia, T., R. Roll, and A. Subrahmanyam (2001): "Market Liquidity and Trading Activity," *The Journal of Finance*, 56, 501–530.

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

——— (2011): "Discount Rates; Presidential Address: Discount Rates," *the Journal of Finance @Bullet*, LXVI, 1047–1108.

DeMiguel, V., A. Martin-Utrera, F. J. Nogales, and R. Uppal (2017): "A Portfolio Perspective on the Multitude of Firm Characteristics," *SSRN Electronic Journal*.

DeMiguel, V., F. J. Nogales, and R. Uppal (2014): "Stock return serial dependence and out-of-sample portfolio performance," *Review of Financial Studies*, 27, 1031–1073.

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.

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

——— (2018): "Choosing Factors," *Journal of Financial Economics*, 128, 234–252.

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

Fan, J. and R. Li (2001): "Variable Selection via Nonconcave Penalized," *Journal of the American Statistical Association*, 96, 1348–1360.

Feng, G., S. Giglio, and D. Xiu (2019): "Taming the Factor Zoo," *Chicago Booth working paper*.

Figueiredo, M. A. T. and R. D. Nowak (2016): "Ordered weighted L1 regularized regression with strongly correlated covariates: Theoretical aspects," *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 41, 930–938.

Freyberger, J., A. Neuhierl, and M. Weber (2019): "Dissecting Characteristics Nonparametrically," *Review of Financial Studies*, forthcoming.

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, 2139–2170.

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

Gu, S., B. Kelly, and D. Xiu (2019): "Empirical Asset Pricing Via Machine Learning," *Review of Financial Studies*, Forthcoming.

Harvey, C. R. and Y. Liu (2017): "Lucky Factors," *National Bureau of Economic Research - Working Paper*.

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

Hou, K., H. Mo, C. Xue, L. Zhang, C. Haitao Mo, and E. J. Ourso (2018): "Motivating Factors," *SSRN eLibrary*.

Hou, K., C. Xue, and L. Zhang (2014): "Digesting Anomalies: An Investment Approach," *Review of Financial Studies*, 28, 650–705.

——— (2017): "Replicating anomalies," *SSRN eLibrary*.

Kan, R. and C. Zhang (1999): "Two-Pass Tests of Asset Pricing Models with Useless Factors," *The Journal of Finance*, 54, 203–235.

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

Kleibergen, F. (2009): "Tests of Risk Premia in Linear Factor Models," *Journal of Econometrics*, 149, 149–173.

Kock, A. B. and L. Callot (2015): "Oracle inequalities for high dimensional vector autoregressions," *Journal of Econometrics*, 186, 325–344.

Kozak, S., S. Nagel, and S. Santosh (2017): "Shrinking the Cross Section," *NBER Working Paper*, 0–42.

Lewellen, J. (2015): "The Cross-section of Expected Stock Returns," *Critical Finance Review*, 1–14.

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.

Ludvigson, S. C. (2013): "Advances in Consumption-Based Asset Pricing : Empirical Tests," *Handbook of the economics of Finance*, 2, 799–906.

Mclean, R. D. and J. Pontiff (2016): "Does Academic Research Destroy Stock Return Predictability?" *Journal of Finance*, 71, 5–32.

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

Pukthuanthong, K., R. Roll, and A. Subrahmanyam (2018): "A Protocol for Factor Identification," *Review of Financial Studies*, forthcoming.

Shanken, J. (1992): "On the Estimation of Beta Pricing Models," *The Review of Financial Studies*, 5, 1–33.

Sharpe, W. F. (1964): "Capital Asset Prices: A Theroy 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*, 58, 267–288.

van Dijk, M. A. (2011): "Is Size Dead? A Review of the Size Effect in Equity Returns," *Journal of Banking and Finance*, 35, 3263–3274.

Yuan, M. and Y. Lin (2006): "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68, 49–67.

Zeng, X. and M. A. T. Figueiredo (2015): "The Ordered Weighted L1 Norm: Atomic Formulation, Projections, and Algorithms," .

Zou, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and T. Hastie (2005): "Regularization and Variable Selection via the Elastic-Net," *Journal of the Royal Statistical Society*, 67, 301–320.

Appendix

A Proof of theorems

A.1 Proof of Theorem 2.1

By definition the OWL estimator is minimizing the function

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

where $|b|_{[.]}$ 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

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

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

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

Together with $\mu_R = Cb^0 + \epsilon$, this implies that (A.1) can be simplified as:

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

where

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

Consider the event

$$\frac{1}{N}\max_{1\leq j\leq K} 2|\epsilon' C^{(j)}| \leq \lambda_0, \tag{A.3}$$

where $\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log K}{N}}$ by assumption. Then (A.2) can be bounded as:

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}||_1 \leq \lambda_0||\hat{b} - b^0||_1 + \frac{1}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1. \tag{A.4}$$

By triangle inequality, $||\hat{b} - b^0||_1 \leq ||\hat{b}||_1 + ||b^0||_1$.

Therefore (A.4) can be written as:

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 + (\frac{\lambda_1}{N} - \lambda_0)||\hat{b}||_1 \leq [\lambda_0 + \frac{\lambda_1 + \lambda_2(K-1)}{N}]||b^0||_1. \tag{A.5}$$

By assumption of the theorem, $\frac{\lambda_1}{N} - \lambda_0 \geq 0$ and $\lambda_1 = o(N), \lambda_2 = o(N)$, we obtain:

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 \leq [\lambda_0 + \frac{\lambda_1 + \lambda_2(K-1)}{N}]||b^0||_1. \tag{A.6}$$

Since $\hat{\Sigma} = \frac{C'C}{N}$, we have

$$(\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0) \leq [\lambda_0 + \frac{\lambda_1 + \lambda_2(K-1)}{N}]||b^0||_1. \tag{A.7}$$

This completes the proof of (11).

We obtained (A.7) assuming (A.3), now we compute the probability of this inequality to be true.

By assumption $\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log K}{N}}$, $t > 0$ and by assumption 1 and 2, $V_j := \epsilon' C^{(j)}/\sqrt{N\sigma^2} \backsim \mathbf{N}(0,1)$.

Using the Gaussian tail bound for independent normal variable V_j, we have:

$$\begin{aligned}
\mathbf{P}(\frac{1}{N}\max_{1\leq j\leq K} 2|\epsilon' C^{(j)}|) \geq \lambda_0) &= \mathbf{P}(\max_{1\leq j\leq K}|V_j| > \sqrt{t^2 + 2\log K}) \\
&\leq \sum_{i=1}^{K} \mathbf{P}(|V_j| > \sqrt{t^2 + 2\log K})) \\
&\leq 2K\exp(-\frac{t^2 + 2\log K}{2}) \\
&= 2\exp(-\frac{t^2}{2}).
\end{aligned}$$

Consequently, (A.3) is valid with probability

$$p \geq 1 - 2exp(-\frac{t^2}{2}).$$

This complete the proof of (10).

If K is fixed, $N \to \infty$, $\hat{\Sigma}$ is a positive definite matrix and the right-hand-side of (11) $\to 0$. Further, if $t \to \infty$, (A.3) holds with probability $p \to 1$. Then it follows trivially that

$$\|\hat{b} - b^0\|_2 \to \mathbf{0}.$$

This completes the proof of the last claim of Theorem 2.1.

A.2 Proof of Theorem 2.2

Using the "argmin" property, we have:

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 + \frac{1}{N}\lambda_1||\hat{b}||_1 \leq \lambda_0||\hat{b} - b^0||_1 + \frac{1}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1. \tag{A.8}$$

By assumption, $\frac{\lambda_1}{N} \geq 2\lambda_0$, then (A.8) can be written as:

$$\frac{2}{N}||C(\hat{b} - b^0)||_2^2 + \frac{2}{N}\lambda_1||\hat{b}||_1 \leq \frac{\lambda_1}{N}||\hat{b} - b^0||_1 + \frac{2}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1. \tag{A.9}$$

Note that

$$||\hat{b}||_1 = ||\hat{b}_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1 \geq ||b^0_{s_0}||_1 - ||\hat{b}_{s_0} - b^0_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1, \tag{A.10}$$

$$||\hat{b} - b^0||_1 = ||\hat{b}_{s_0} - b^0_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1. \tag{A.11}$$

Therefore (A.9) can be written as:

$$\frac{2}{N}||C(\hat{b} - b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}_{s_0^c}||_1 \leq 3\frac{\lambda_1}{N}||\hat{b}_{s_0} - b^0_{s_0}||_1 + \frac{2\lambda_2(K-1)}{N}||b^0||_1. \tag{A.12}$$

Using (A.11), we have

$$\frac{2}{N}||C(\hat{b}-b^0)||_2^2+\frac{\lambda_1}{N}||\hat{b}-b^0||_1=\frac{2}{N}||C(\hat{b}-b^0)||_2^2+\frac{\lambda_1}{N}||\hat{b}_{s_0}-b^0_{s_0}||_1+\frac{\lambda_1}{N}||\hat{b}_{s_0^c}||_1.$$

Utilizing (A.12), we obtain inequality

$$\frac{2}{N}||C(\hat{b}-b^0)||_2^2+\frac{\lambda_1}{N}||\hat{b}-b^0||_1\leq 4\frac{\lambda_1}{N}||\hat{b}_{s_0}-b^0_{s_0}||_1+\frac{2\lambda_2(K-1)}{N}||b^0||_1. \tag{A.13}$$

By assumption 4, the restricted eigenvalue condition states that

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

which implies

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

where S is defined in assumption 3.

Rearrange above inequality, we have

$$||b_{s_0}||_1^2\leq b'\hat{\Sigma}bS/\phi_0^2, \tag{A.14}$$

which is also regarded as the *compatibility condition* in Buhlmann and Van de Geer (2011).

Applying (A.14) on $||\hat{b}_{s_0}-b^0_{s_0}||_1$ and using $\hat{\Sigma}=\frac{C'C}{N}$, we have

$$||\hat{b}_{s_0}-b^0_{s_0}||_1^2\leq(\hat{b}-b^0)'\hat{\Sigma}(\hat{b}-b^0)S/\phi_0^2=||C(\hat{b}-b^0)||_2^2S/(N\phi_0^2).$$

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

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

So (A.13) can be written as:

$$\frac{1}{N}||C(\hat{b}-b^0)||_2^2+\frac{\lambda_1}{N}||\hat{b}-b^0||_1 \leq 4(\frac{\lambda_1}{N})^2 S/\phi_0^2+\frac{2\lambda_2(K-1)}{N}||b^0||_1. \tag{A.15}$$

Note that $\frac{1}{N}||C(\hat{b}-b^0)||_2^2=(\hat{b}-b^0)'\hat{\Sigma}(\hat{b}-b^0)$, so (A.15) completes the proof of (14).

By assumption $\frac{\lambda_1}{N}=2\lambda_0=4\hat{\sigma}\sqrt{\frac{t^2+2\log K}{N}}$ and $\lambda_2=O(\frac{S\log K}{K})$, both two terms on the right hand side of (A.15) are $O(\frac{S\log K}{N})$. So

$$\frac{1}{N}||C(\hat{b}-b^0)||_2^2=O\left(\frac{S\log K}{N}\right), \tag{A.16}$$

$$||\hat{b}-b^0||_1=O\left(S\sqrt{\frac{\log K}{N}}\right). \tag{A.17}$$

So (A.17) completes the proof of the first claim of (15). Observe that

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

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 population value of $\hat{\Sigma}$, so $\Lambda_{min}>0$. Moreover,

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

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 $\max_{1\leq i,j\leq K}|\hat{\Sigma}_{i,j}-\Sigma_{i,j}|=O_p(\sqrt{\frac{\log K}{N}})$. So (A.16) can be written

$$\begin{aligned} O\left(\frac{S\log K}{N}\right) &= \frac{1}{N}||C(\hat{b}-b^0)||_2^2 \\ &\geq \Lambda_{min}^2||\hat{b}-b^0||_2^2-O_p\left(S^2\left(\frac{\log K}{N}\right)^{3/2}\right). \end{aligned} \tag{A.18}$$

Rearrange, 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 $S\sqrt{\frac{\log K}{N}} = o(1)$ and $\frac{1}{\Lambda_{min}^2} = O(1)$, therefore we obtain

$$||\hat{b} - b^0||_2 = O_p(\sqrt{\frac{S \log K}{N}}). \tag{A.19}$$

Here we complete the proof of the second claim of (15). Lastly, the proof of (13) follows the same method in Theorem 2.1.

A.3 Proof of Theorem 2.3

The proof of theorem 2.3 relies on the Pigou-Dalton-transfer lemma and the directional derivative lemma, and it follows a similar argument with Figueiredo and Nowak (2016), except that we are dealing with both the time-series and cross-sectional dimensions.

Lemma 1 (Pigou-Dalton-Transfer). *A vector $x \in R_+^p$, and its two components x_i, x_j 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 $z_k = x_k$, $\forall k \neq i, j$, then*

$$\Omega_\omega(x) - \Omega_\omega(z) \geq \Delta_\omega \epsilon,$$

where $\Omega_\omega(.)$ is the OWL norm defined in (6), and $\Delta_\omega := \min_{i=1,\cdots,p-1} \omega_{i+1} - \omega_i$.

Lemma 2 (Directional derivative). *The directional derivative of a real valued convex function f at $x \in dom(f)$, $f(x) \neq \infty$, is*

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

Then $x^* \in \arg\min(f)$, if and only if $f'(x^*, u) \geq 0$ for any u.

Proof of Theorem 2.3: Denote the object function in (6) as

$$Q := \frac{1}{2}(\mu_R - Cb)' W_T (\mu_R - Cb) + \Omega_\omega(b) = \frac{1}{2}||\mu_R - Cb||_2^2 + \Omega_\omega(b),$$

where W_T is an identity matrix [19]. Suppose

$$\sigma(f_i - f_j) < \frac{\lambda_2}{||\mu_R||_2||\sigma_R||_2}, \tag{A.20}$$

and assume

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

Without loss of the generality, assume $\hat{b_i} > \hat{b_j}$ (we want to find a contradiction between the assumption $\hat{b_i} \neq \hat{b_j}$ and (A.20)). The directional derivative of Q at $\hat{b}$ with $u_i = -1, u_j = 1, i < j, u_k = 0, \forall k \neq i, j$, is:

$$Q'(\hat{b}, u) = \underbrace{\lim_{\alpha \to 0^+} \frac{||\mu_R - C\hat{b} + \alpha(C_i - C_j)||_2^2 - ||\mu_R - C\hat{b}||_2^2}{2\alpha}}_{\text{quadratic loss part}} + \underbrace{\lim_{\alpha \to 0^+} \frac{\Omega_\omega(\hat{b} + \alpha u) - \Omega_\omega(\hat{b})}{\alpha}}_{\text{regularization part}},$$

where C_i and C_j are the i^{th} and j^{th} columns of the factor-return covariance matrix.
Next, we investigate the quadratic loss part of the above equation.

$$\begin{aligned}\text{quadratic loss part} &= \lim_{\alpha \to 0^+} \frac{||\mu_R - C\hat{b}||^2 + 2\alpha(\mu_R - C\hat{b})(C_i - C_j) + \alpha^2||C_i - C_j||_2^2 - ||\mu_R - C\hat{b}||_2^2}{2\alpha} \\ &= (\mu_R - C\hat{b})(C_i - C_j).\end{aligned}$$

Apply the Pigou-Dalton-transfer lemma on the regularization part, where $\epsilon = \alpha$, we have

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

So the regularization part follows:

$$\text{regularization part} \leq -\lim_{\alpha \to 0^+} \frac{\Delta_\omega \alpha}{\alpha} = -\Delta_\omega.$$

[19] As long as W_T is a semi-positive definite matrix, we can perform Cholesky decomposition and map it back to the identity matrix format.

By the definition of ω in (7), $\Delta_\omega = \lambda_2$. Then it follows that

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

Using Cauchy-Schwarz inequality we have

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

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

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

Now we further utilize the covariance inequality

$$cov(R, f_i - f_j) \leq \sqrt{var(R) var(f_i - f_j)} = \sigma_R \sigma(f_i - f_j),$$

where σ_R is a $N \times 1$ vector (the sample standard deviation for N test assets, each has $T \times 1$ times series observations), and $\sigma(f_i - f_j)$ is a scaler, which is the standard deviation of a $T \times 1$ difference sequence between factor i and j.
Now we have

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

Together with (A.20) we obtain

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

which violates the directional derivative lemma if $\hat{b}$ is a minimizer of Q. Hence there is a contradiction between $\hat{b}_i \neq \hat{b}_j$ and (A.20). So we must have:

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

B Consistency of model selection

In Section 2.4 we derived the convergence rate of OWL estimator when both K and N grow large. Starting from theorem 2.2, we can establish that $||\hat{b} - b^0||_\infty \leq ||\hat{b} - b^0||_2 \leq C\sqrt{\frac{S \log K}{N}}$ with arbitrarily large probability by selecting a constant C sufficiently large. To fix ideas, the thresholding estimator $\tilde{b}$ is defined as

$$\tilde{b}_j = \begin{cases} \hat{b}_j, & if \quad |\hat{b}_j| \geq H \\ 0, & if \quad |\hat{b}_j| \leq H \end{cases}, \tag{B.21}$$

where H is the hard thresholding parameter. Let $J(b_{s_0}) = \{j = 1, 2, ..., K : b_j^0 \neq 0\}$ and $J(\tilde{b}_s) = \{j = 1, 2, ..., K : \tilde{b}_j \neq 0\}$, we can establish the following corollary on consistency of model selection by thresholding, i.e. $\tilde{b}$ can select the non-zero coefficients of factors as the true ones.

Corollary B.1 (Consistency of model selection by thresholding)**.** *Let assumptions 2, 3 and 4 be satisfied and assume that* $\min_{j \in J(b_{S_0})} |b_{s_0,j}| > 3C\sqrt{\frac{S \log K}{N}}$*, where* C *is a constant. Then, for all* $\epsilon > 0$ *there exist a* C *such that* $H = 2C\sqrt{\frac{S \log K}{N}}$*, one has* $P\left(J(\tilde{b}_s) = J(b_{s_0})\right) > 1 - \epsilon$ *as* $n \to \infty$*.*

Corollary B.1 shows that consistency in model selection by thresholding is possible. However, finding a suitable C is *empirically challenging*, especially in small samples. Although thresholding together with the OWL estimator offers a theoretic foundation for consistency in model selection, we resort to a more traditional model testing procedure, similar to Feng et al. (2019) and DeMiguel et al. (2017). More details about model selection procedures are discussed in Section 2.5.

Proof of Corollary B.1

From theorem 2.2 we can establish that for every $\epsilon > 0$ there exists a $C > 0$ such that $||\hat{b} - b^0||_\infty \leq C\sqrt{\frac{S \log K}{N}}$ with probability at least $1 - \epsilon$ for sufficiently large C. Consider separately the zero and non-zero coefficients of b^0. For the first type (true coefficients are

zeros), suppose $j \in J(b_{s_0^c}) = \{j = 1, 2, ..., K : b_{s_0^c, j} = 0\}$ and we have

$$\max_{j \in J(b_{s_0^c})} |\hat{b}_j| \leq C\sqrt{\frac{S \log K}{N}} < 2C\sqrt{\frac{S \log K}{N}} = H.$$

By definition, $\tilde{b}_{s_0^c} = 0$.

Next, for the second type (true coefficients are non-zeros), suppose $j \in J(b_{s_0})$ and we have

$$|\hat{b}_j| \geq |b_{s_0,j}| - |\hat{b}_j - b_{s_0,j}| \geq \min_{j \in J(b_{s_0})} |b_{s_0,j}| - |\hat{b}_j - b_{s_0,j}| \geq 3C\sqrt{\frac{S \log K}{N}} - C\sqrt{\frac{S \log K}{N}} = H,$$

again, by definition, $\tilde{b}_{s_0} = \hat{b}_{s_0} \neq 0$.

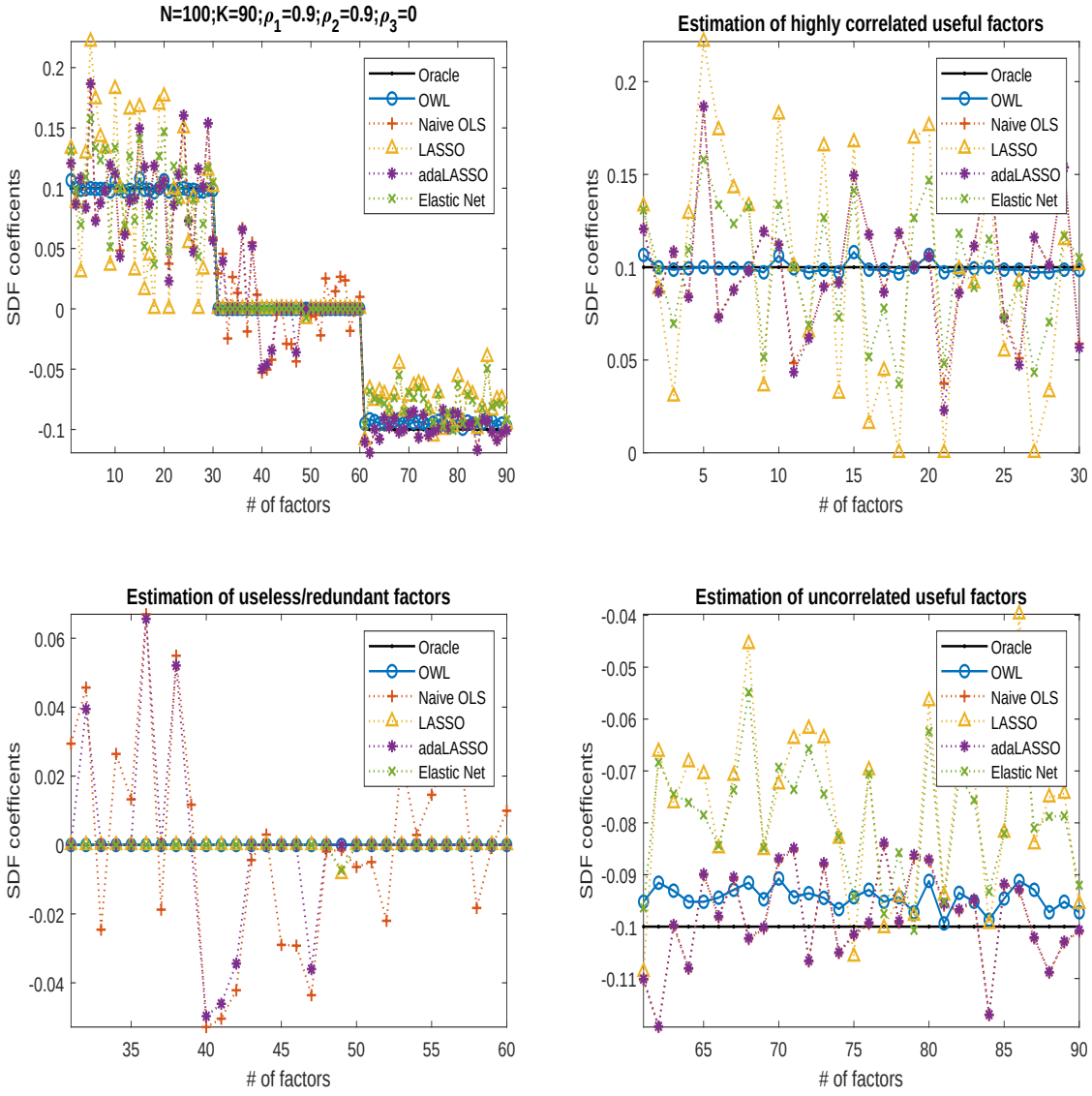


Figure 1. Simulation: $N = 100, K = 90$

This figure reports the plot of 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 block, where their correlation coefficients within each block are $\rho_1 = 0.9, \rho_2 = 0.9, \rho_3 = 0$. The upper left panel displays the plots of all factors. The remaining three panels are detailed plot for each of these three blocks. The upper right panel displays the plot of all estimators of useful factors that are highly correlated. The bottom left panel displays the plot of all estimators of useless/redundant factors. The bottom right panel displays the plot of all estimators of useful factors but not correlated. In each plot, OWL estimator is displayed along with LASSO, adaptive LASSO, Elastic Net, and native OLS estimators.

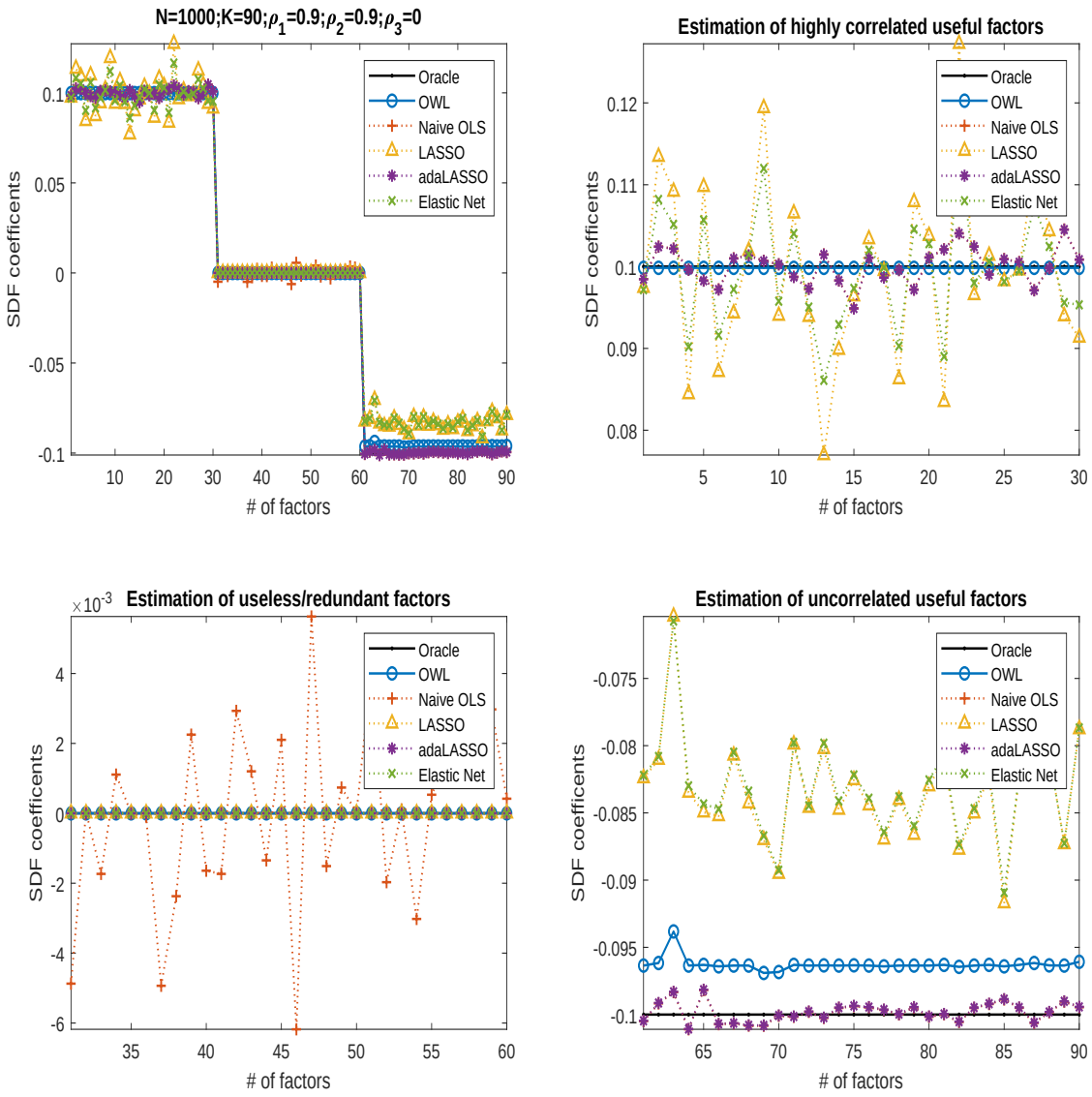


Figure 2. Simulation: $N = 1000, K = 90$

This figure reports the plot of OWL estimator along with other benchmarks. The number of assets is 1000, all the rest are the same with the first experiment.

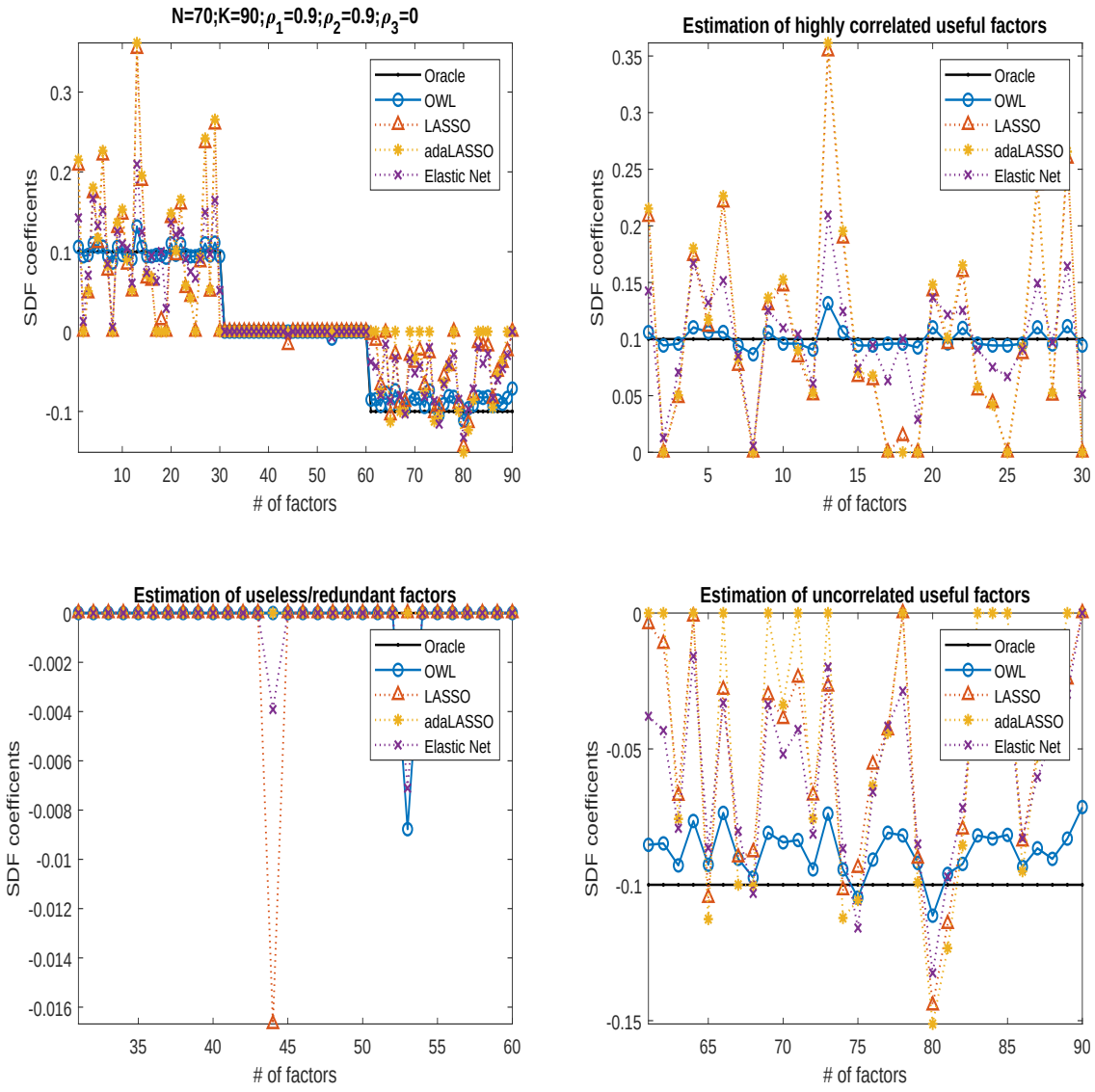
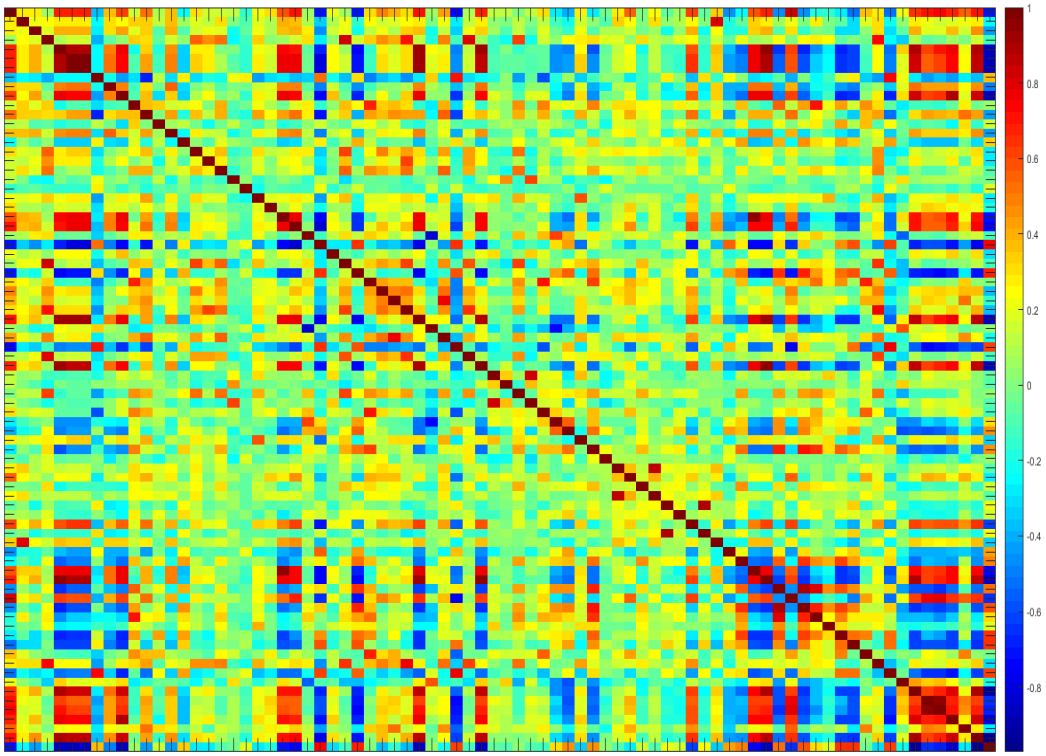
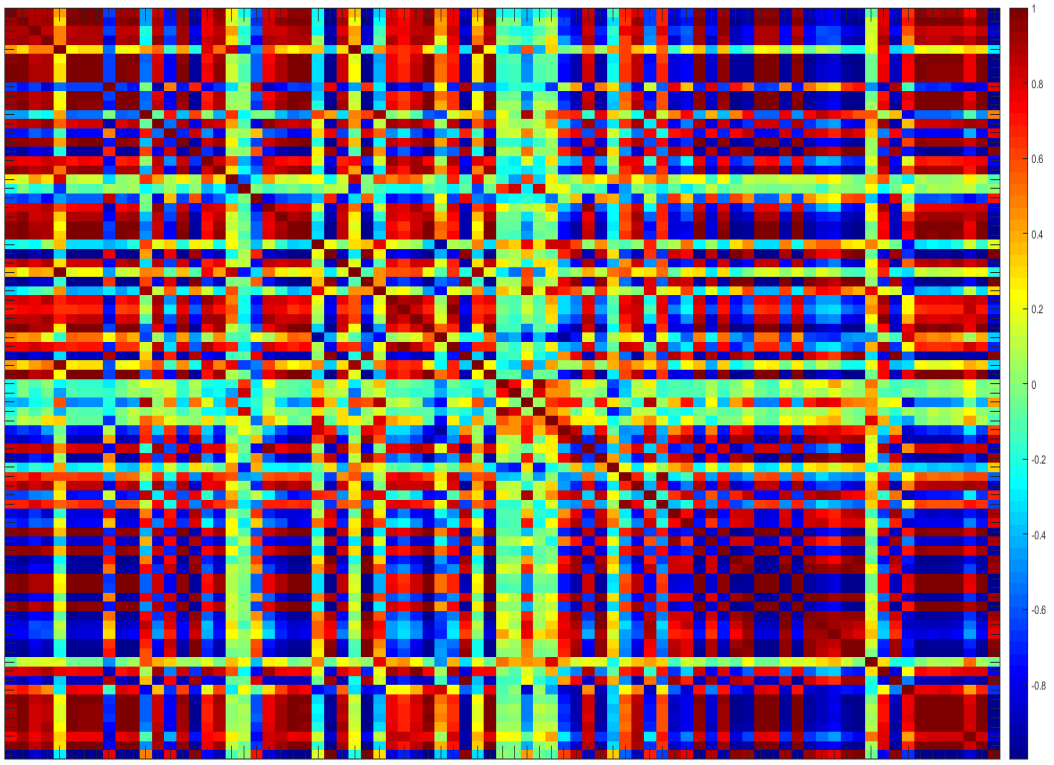


Figure 3. Simulation: $N = 70, K = 90$

This figure reports the plot of OWL estimator along with other benchmarks. The number of assets is 70 and the number of factors is 90. Adaptive LASSO is using the LASSO estimator as its adaptive weight.



(a) Factor correlation measured by times series



(b) Factor correlation measured by factor loadings

Figure 4. Factor correlation coefficient matrix

This heat map displays the correlation coefficients of all 80 anomaly factors. Dark red and deep blue indicate high correlation (positive or negative) while light colours indicate low correlation.

Table 1. Anomaly factors and their acronyms

This table lists all 80 factors considered in our factor library. The abbreviation is consistent with Green et al. (2017). 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

Table 2. Robust estimation of Two-step selection procedure

This table reports the two-stage select-and-test procedure to find anomaly factors that explains the cross section of average stock returns. We consider the full sample size from 1980 to 2017 and two sub sample sizes breaks on year 2000. equal weighted (ew) and valued weighted (vw) methods are both considered. Three measures of micro stock impact are employed: we remove stocks that is smaller than 20 (30 and 40) percentile of NYSE listed stocks. Within each estimation we list all selected factors, where in the bracket is the ordinal number it selected by OWL (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)		

Table 3. Full/sub-sample factor selection using various methods

This table reports the first five factors selected with greatest magnitude 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, which are divided before and after 2000 (sub1 and sub2). We also control micro stocks. We consider all stocks (P00), or remove micro stocks' market capitalisation which are smaller than 20/40 percentile of NYSE listed stocks (P20/P40).

		First five selected factors (decreasingly) ordered by its magnitude				
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'
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'
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'

Table 4. Out-of-sample hedge 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. Rolling window size is of 120 months. Factor selection strategies include OWL, LASSO, Elastic Net (EN), and two-pass Fama-MacBeth regression (FM). The upper panel is using the full sample; the middle and lower panels are using sub-samples.

		SR	Mean	Std	Skewness	Kurtosis	Slope	R2
full_P00	OWL	2.44	6.54	9.29	-1.01	15.50	0.69	0.26
	LASSO	2.42	6.89	9.86	-0.95	16.69	0.70	0.27
	EN	2.46	6.69	9.42	-1.24	17.01	0.69	0.27
	FM	2.37	6.61	9.68	0.07	8.49	0.67	0.27
full_P20	OWL	1.21	2.17	6.24	-0.07	9.48	0.56	0.11
	LASSO	1.01	2.13	7.30	2.21	31.09	0.57	0.12
	EN	1.04	2.26	7.52	1.71	27.70	0.57	0.12
	FM	0.96	1.96	7.08	2.88	37.37	0.53	0.10
full_P40	OWL	0.90	1.59	6.13	1.39	25.06	0.46	0.08
	LASSO	0.77	1.48	6.62	4.09	57.11	0.42	0.08
	EN	0.82	1.52	6.39	3.17	46.12	0.44	0.09
	FM	0.72	1.41	6.79	3.68	49.89	0.43	0.08
sub1_P00	OWL	3.51	7.46	7.35	0.42	4.48	0.84	0.29
	LASSO	3.42	7.33	7.43	0.50	4.46	0.77	0.29
	EN	3.35	7.12	7.36	0.60	4.37	0.76	0.28
	FM	3.03	7.80	8.92	0.59	5.50	0.80	0.33
sub1_P20	OWL	2.10	2.54	4.18	0.10	3.41	0.66	0.12
	LASSO	1.87	2.09	3.87	0.10	3.48	0.66	0.10
	EN	1.87	2.09	3.87	0.10	3.48	0.66	0.10
	FM	1.66	1.92	4.01	0.65	5.45	0.54	0.09
sub1_P40	OWL	1.35	1.34	3.44	-0.03	4.37	0.46	0.06
	LASSO	1.03	1.13	3.82	0.02	3.67	0.29	0.08
	EN	1.03	1.13	3.82	0.02	3.67	0.29	0.08
	FM	0.75	0.75	3.50	-0.21	5.62	0.25	0.06
sub2_P00	OWL	3.20	6.04	6.54	0.04	3.55	0.66	0.23
	LASSO	3.23	5.98	6.42	0.26	3.35	0.67	0.23
	EN	3.26	6.09	6.48	0.46	3.96	0.68	0.23
	FM	3.20	5.87	6.36	-0.31	3.88	0.66	0.24
sub2_P20	OWL	2.10	2.43	4.67	1.02	8.72	0.54	0.11
	LASSO	1.91	2.10	3.80	0.16	3.51	0.54	0.09
	EN	1.91	2.10	3.80	0.16	3.51	0.54	0.09
	FM	1.78	1.80	3.49	-0.48	3.82	0.48	0.09
sub2_P40	OWL	2.11	2.04	3.34	0.62	5.83	0.56	0.08
	LASSO	1.80	1.69	3.27	0.58	6.16	0.53	0.06
	EN	1.69	1.59	3.25	0.37	4.44	0.50	0.06
	FM	1.80	1.75	3.35	0.13	2.91	0.53	0.07

Online Appendix

Chuanping Sun

A Solve the OWL optimization problem

This section 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 computes the optimizer at each step. In the second subsection we introduce a fast-iterative-soft-thresholding-algorithm (FISTA) to find the global optimizer, where we also implement a backtracking line search condition which speeds up computation greatly.

A.1 OWL proximal function

First define the proximal function as

$$Prox_{\Omega_\omega}(b) = \arg\min_x \frac{1}{2}||x-b||_2^2 + \Omega_\omega(x). \tag{A.1}$$

$\Omega_\omega(x) = \omega'|x|_\downarrow$, where $\omega \in \kappa$, and κ is a monotone non-negative cone, defined as $\kappa := \{v \in R^n : v_1 \geq v_2 \geq \cdots \geq v_n \geq 0\}$. $|x|_\downarrow$ is the absolute value of vector x, decreasingly ordered. So by the definition of $\Omega_\omega(b)$, we have

$$\Omega_\omega(b) = \Omega_\omega(|b|). \tag{A.2}$$

It is easy to show that

$$||b - sign(b) \odot |x|||_2^2 \leq ||b-x||_2^2, \tag{A.3}$$

where $sign(.)$ is a function to retrieve signs from a vector, with elements in $\{1, -1, 0\}$ and $\odot$ is a point-wise production operator.

Therefore, (A.2) and (A.3) imply

$$Prox_{\Omega_\omega}(b) = sign(b) \odot Prox_{\Omega_\omega}(|b|). \tag{A.4}$$

Let P be a permutation matrix that orders a vector decreasingly, by the property of permutation matrix, we obtain

$$||P(x-b)||_2^2 = ||x-b||_2^2, \tag{A.5}$$

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

$$\Omega_\omega(b) = \Omega_\omega(Pb). \tag{A.6}$$

So (A.5) and (A.6) imply that (A.4) can be written as

$$Prox_{\Omega_\omega}(b) = sign(b) \odot P'(|b|) Prox_{\Omega_\omega}(|b|_\downarrow), \tag{A.7}$$

where $|b|_\downarrow$ is a vector of decreasingly ordered absolute value of coefficients, and $P'(|b|)$ is the transpose of the permutation matrix, which recovers the order of $|b|$.

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

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

It implies that $Prox_{\Omega_\omega}(|b|_\downarrow) \in \kappa$, and $\Omega_\omega(x) = \omega' x$.

Then, we have

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

which is the projection of $(|b|_\downarrow - \omega)$ onto κ. Then equation (A.7) can be written as

$$Prox_{\Omega_\omega}(b) = sign(b) \odot (P'(|b|) Proj_\kappa(|b|_\downarrow - \omega)), \tag{A.8}$$

where $Proj_{\kappa}(.)$ is the projection operator onto κ. [1]

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

$$b^{(k+1)} = prox_{\Omega_{\omega}}(b^{(k)} - sz_k \bigtriangledown g(b^{(k)})) \tag{A.9}$$

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

A.2 FISTA algorithm

Fast computation is achieved by using the backtracking line condition and the acceleration in u (step 12). The backtracking line condition (step 7) 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 while approaching the global optimizer.

[1] The projection onto κ can be obtained by using the Pool-Adjacent-Violators algorithm. See de Leeuw et al. (2009).

Algorithm 1: FISTA-OWL

Input: μ_R, C, ω

Output: OWL estimator $\hat{b}$

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]

while *some stopping criterion not met* **do**

  $\tau_k = \tau_{k-1}$;

  $b_k = Prox_{\Omega_\omega}(u_k + \tau * C' * (\mu_R - Cb))$

  while $\frac{1}{2}||\mu_R - Cb_k||_2^2 > Q(b_k, u_k)$ [b] **do**

    $\tau_k = \eta * \tau_k$;

    $b_k = Prox_{\Omega_\omega}(u_k + \tau * C' * (\mu_R - Cb))$

  end

  $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$

  $u_{k+1} = b_k + \frac{t_{k-1}}{t_{k+1}}(b_k - b_{k-1})$

  $k \leftarrow k + 1$

end

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.

B Motivating the "restricted eigenvalue condition"

The following lemma motivates the restricted eigenvalue condition. For a matrix that satisfies the restricted eigenvalue condition, it suffices that this matrix is close to a matrix whose restricted eigenvalues are strictly positive.

Let $\Sigma = E(\hat{\Sigma}) = E(\frac{C'C}{N})$ be the population value of the scaled Gram matrix and its restricted eigenvalues are strictly positive, that is $\phi_\Sigma^2 > 0$.

Lemma 1. *Suppose S is the sparsity parameter, $\delta = \max_{1 \le i,j \le N} |\Sigma_{i,j} - \hat{\Sigma}_{i,j}|$, then for any vector b satisfies $||b_{s_0^c}||_1 \le 3||b_{s_0}||_1$, one also has*

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

Proof.

$$\begin{aligned} b'\Sigma b - b'\hat{\Sigma} b &\leq |b'\Sigma b - b'\hat{\Sigma} b| = |b'(\Sigma - \hat{\Sigma})b| \\ &\leq ||b||_1 ||(\Sigma - \hat{\Sigma})b||_\infty \leq \delta ||b||_1^2 \\ &\leq \delta(||b_{s_0^c}||_1 + ||b_{s_0}||_1)^2 \leq 16\delta ||b_{s_0}||_1^2 \\ &\leq 16S\delta ||b_{s_0}||_2^2. \end{aligned}$$

Rearrange above inequality, we have

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

Equivalently,

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

□

Lemma 1 shows that for the restricted eigenvalue condition to be satisfied, it suffices to show that δ is small, or the Gram matrix is close to a positive definite matrix with high probability.

The following lemma shows the "Restricted eigenvalue condition" implies the compatibility condition in Buhlmann and Van de Geer (2011).

Lemma 2 (Compatibility condition)**.** *If the scaled Gram matrix $\hat{\Sigma}$ satisfies the restricted eigenvalue condition, then*

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

Proof. From the definition of restricted eigenvalues, we have

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

which implies

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

Rearrange, we have

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

□

C Robustness Check

In this section, we check 1) whether liquidity related factors are robust in explaining the cross section of asset returns; 2) how small stocks affect factors' implications; 3) whether the FISTA-OWL algorithm has a sound convergence property.

C.1 Are 'liquidity' factors robust?

For the first task, we consider three types of alternative sorting methods for constructing test portfolios and compare them with the sorting method in the empirical analysis to check whether liquidity related factors are consistently chosen by OWL. First, we apply the uni-variate sorting method to sort all non-micro stocks into decile portfolios using each characteristic, and combine them together to obtain 800 test portfolios. Compared to the test portfolio in empirical analysis, all characteristics are treated equally. In other words, 'size', like any other anomaly factor, is a candidate factor. Second, we consider bi-variate sorting, but with all possible combinations of 80 characteristics, that is 3160 possible combinations. To reduce the dimension of test portfolios, we consider the 2 by 2 (instead of 5 by 5) sorting: we sort all stocks into high and low groups where the threshold is the median of each characteristic. We 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 A.1 about here.]

Figure 1 reports the two-stage procedure result using four different sets of test assets (including the one used in empirical analysis). 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' top the chart of anomaly factors. Second, the impact of 'size' factor (mve) on test assets dropped 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 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.

C.2 How small stocks affect the interpretation of 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 factors' implications vary within each scenario.

[Figure A.2 about here.]

Figure 2 reports the heat map of OWL estimation before bootstrap test. First, micro stocks alter market factor's interpretation drastically. When micro stocks are all included to form test portfolios, market factor only plays a mediocre role for asset prices; however, liquidity related factors dominating 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 to drive asset prices, particularly with the inclusion of small stocks. It shows that small firms face severe liquidity constraint, 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 which case, profitability and growth related factors, after liquidity related factors, become the second tier of factors to drive asset prices.

C.3 Convergence using FISTA-OWL algorithm

Figure 3 shows the convergence of Fista-OWL with backtracking algorithm (see appendix A) in the empirical analysis using 81 factors (80 anomaly factors plus the market factor). Vertical axis shows the distance between the k^{th} estimation and the optimizer. Horizontal axis shows the number of iterations (steps) until a stopping criterion is met. Following the machine learning literature (see Zeng and Figueiredo (2015)), we set a tight stopping criterion which is $\frac{||b(k)-b(k-1)||_2}{||b(k)||_2} < 10^{-6}$, $b(k)$ is the OWL estimation of the risk price at the k^{th} iteration. This figure shows that FISTA-OWL algorithm has a sound convergence property: it converges quickly at the first 1000 steps, then it gradually converges to the optimizer because of a tight stopping criterion.

[Figure A.3 about here.]

D Monte Carlo Simulation

For the robustness check of Monte Carlo experiments, we repeat simulation experiments in various settings multiple times, and we report the deviation of each estimator from oracle values.

Figure 4 shows the estimation error of each methods with multiple repetitions. We repeat the first experiment five times, in which we consider 90 candidate factors ($K = 90$) and 100 test assets ($N = 100$). [2] First of all, the patterns are consistent among the repeated exercises. LASSO is the worst performer especially when factors are correlated. Elastic Net does improve the performance of LASSO when factors are correlated, while yields similar results to LASSO when factors are not correlated. OLS is an unbiased estimator yet it does not produce sparsity. Adaptive LASSO was influenced by OLS failing to shrink all useless factors to zero but, performs the best in the uncorrelated setting. OWL, on the contrary, is the best performer when factors are correlated. And in the uncorrelated setting, though outperformed by adaptive LASSO, it is substantially better than LASSO and Elastic Net.

[2] We repeat the experiment 5 times and it is for the convenience and clarity of displaying the figure. Repetitions of large numbers are also available upon request.

[Figure A.4 about here.]

In the second experiment which represents a typical low-dimensional setting ($N = 1000, K = 90,\ N \gg K$), figure 5 plots the estimation error of each estimators with five repetitions. It shows a similar pattern to figure 4. OWL performs best when factors are correlated. In this low-dimensional world, adaptive LASSO is the best estimator when factors are not correlated and it also does a good job to shrink off useless factors. LASSO and Elastic Net are worst performers.

[Figure A.5 about here.]

In the third experiment which represents typically a high-dimensional setting ($N = 70, K = 90,\ N < K$), figure 6 plots estimation errors of each estimators with five repetitions when $K > N$. The best performer is OWL followed by Elastic Net and LASSO. The worst performer is adaptive LASSO which uses the LASSO estimator as the adaptive weight.

[Figure A.6 about here.]

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

Let y denote a vector of responses which is a $N \times 1$ vector. x is a data matrix of size $N \times K$. β_i is the i^{th} element of parameter vector β of size $K \times 1$.

LASSO solves the problem

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \quad ||y - X\beta||^2 + \lambda||\beta||_1, \tag{E.1}$$

where $||\beta||_1$ is the summation of the absolute values of the parameter vector β, or the L_1 norm of β. LASSO estimator achieves sparsity selection by shrink the coefficients of many unimportant covariates to zeros.

Elastic net solves this problem

$$\hat{\beta}_{EN} = \arg\min_{\beta} \quad ||y - X\beta||^2 + \lambda\alpha||\beta||_1 + \lambda(1-\alpha)||\beta||_2^2. \tag{E.2}$$

Elastic net combines the L_1 (or LASSO) penalty and the L_2 (or Ridge) penalty together, which gives more robust results when variables are correlated.

Adaptive LASSO minimize the following problem:

$$\hat{\beta}_{adaLASSO} = \arg\min_{\beta} \quad ||y - X\beta||^2 + \lambda \sum_{i=1}^{K} \frac{1}{|\hat{\beta}_{ols}|^\gamma} |\beta_i|, \tag{E.3}$$

where $\frac{1}{|\hat{\beta}_{ols}|^\gamma}$, $\gamma > 0$, is the adaptive weight. $\hat{\beta}_{ols}$ is a consistent estimator of β. $|\beta_i|$ is the absolute value of the i^{th} element of the parameter vector. Essentially, adaptive LASSO assigns an adaptive weight, for instance, the first stage OLS estimator, to each coefficient's penalty. In other words, each $|\beta_i|$ receives a different penalty, while in LASSO estimator all $|\beta_i|$ receive the same penalty λ . The variables with small (absolute value) OLS estimated coefficients receive stronger penalty.

OSCAR (Octagonal shrinkage and clustering algorithm for regression) solves this problem

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

where $\sum_{i<j} \max\{|\beta_i|, |\beta_j|\}$ is a L_∞ norm. Bondell and Reich (2008) show that OSCAR's octagonal atomic norm encourages factor clustering when they are correlated. Figueiredo and Nowak (2016) illustrate that by adopting a linear decreasing weighting vector, OWL estimator maps to OSCAR exactly. Starting from the OSCAR penalty,

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

With $\omega = \lambda_1 + \lambda_2(K - i)$, OWL encompasses OSCAR. Further, if we set $\lambda_2 = 0$, OWL encompasses LASSO.

[Figure A.7 about here.]

A geometric interpretation of the OWL/OSCAR norm is illustrated in figure (7).

References

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*.

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.

Figueiredo, M. A. T. and R. D. Nowak (2016): "Ordered weighted L1 regularized regression with strongly correlated covariates: Theoretical aspects," *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 41, 930–938.

Zeng, X. and M. A. T. Figueiredo (2015): "The Ordered Weighted L1 Norm: Atomic Formulation, Projections, and Algorithms," .

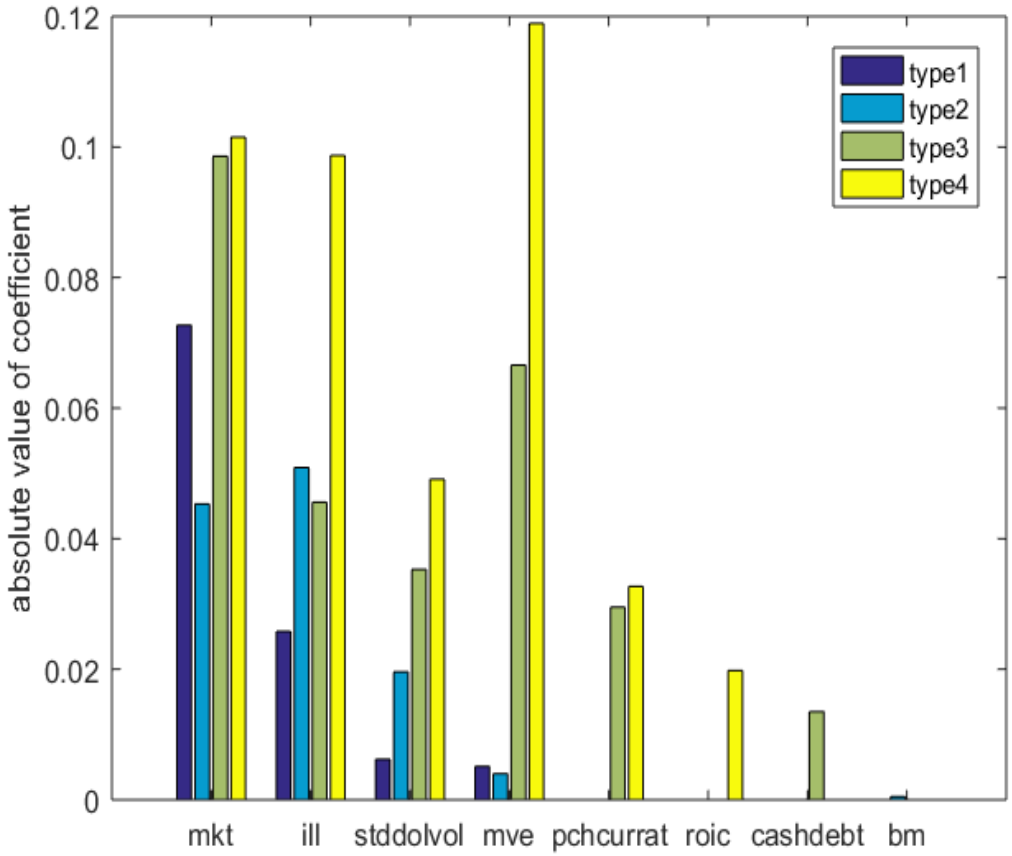


Figure 1. Robustness check using different sorting methods

This figure reports the absolute value of coefficients estimated by OWL using different sorting methods. 'type1' is the uni-variate sorting method; 'type2' is 2 by 2 bi-variate sorting, considering all possible combinations of 80 characteristics; 'type3' is 3 by 3 bi-variate sorting by singling out 'size' to form bi-variate sorting with the remaining characteristics; 'type4' is the 5 by 5 bi-variate sorting in empirical analysis.

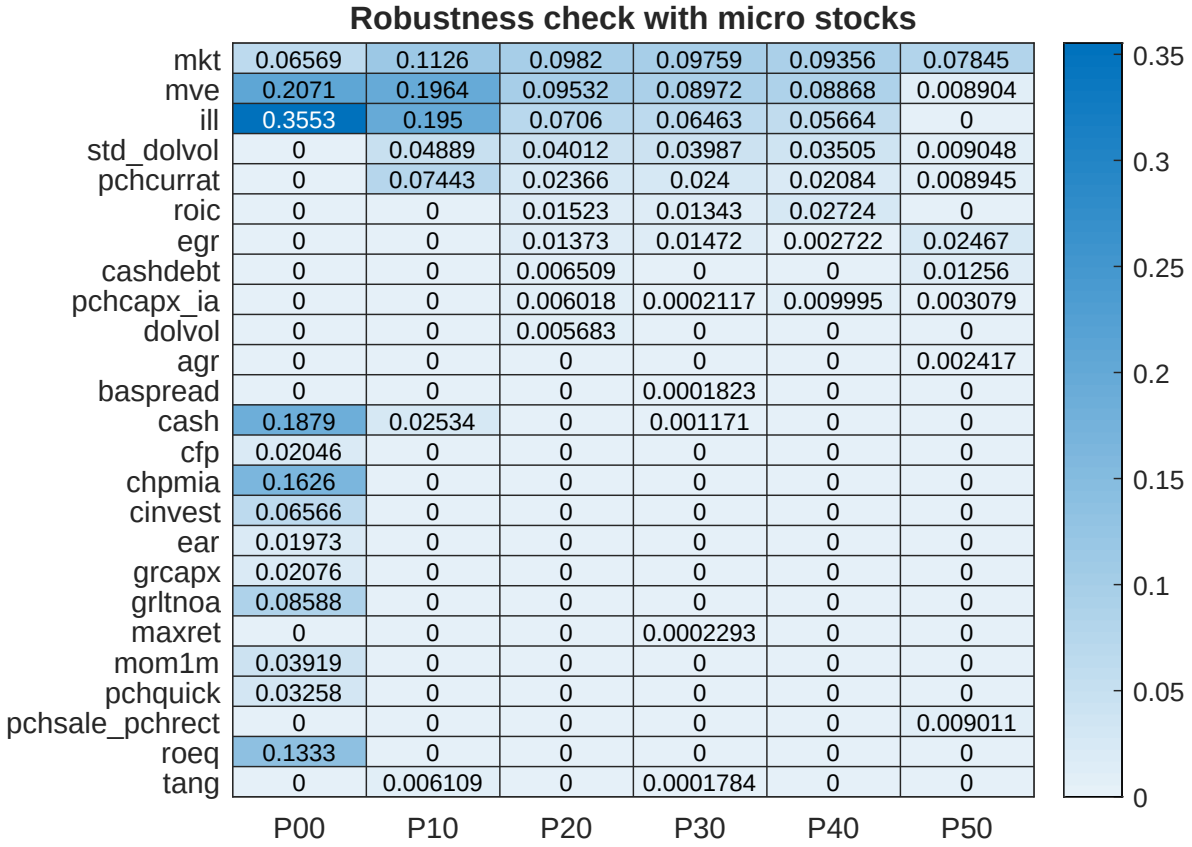


Figure 2. Robustness check with micro stocks

This figure reports six OWL estimations (before bootstrap test) with different treatments 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).

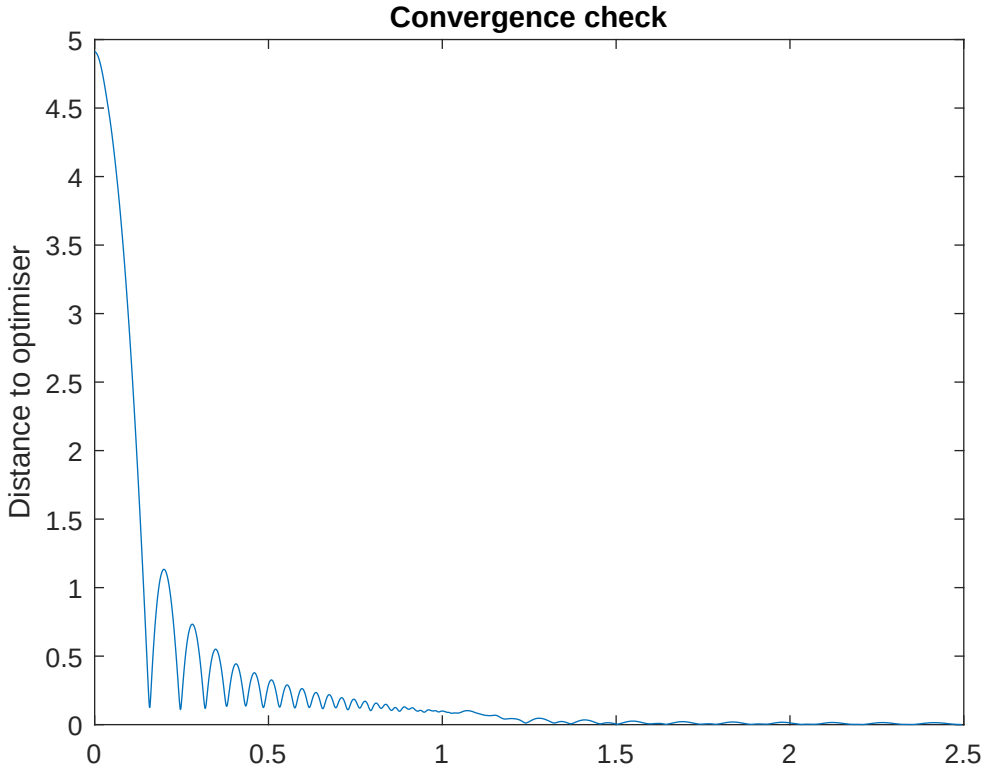


Figure 3. Convergence check

This figure shows the convergence of OWL estimation using Fista-OWL algorithm, where the stopping criterion is $\frac{||b(k)-b(k-1)||_2}{||b(k)||_2} < 10^{-6}$, in which k is the number of iterations. and $b(k)$ is the OWL estimation of risk price at the k^{th} iteration.

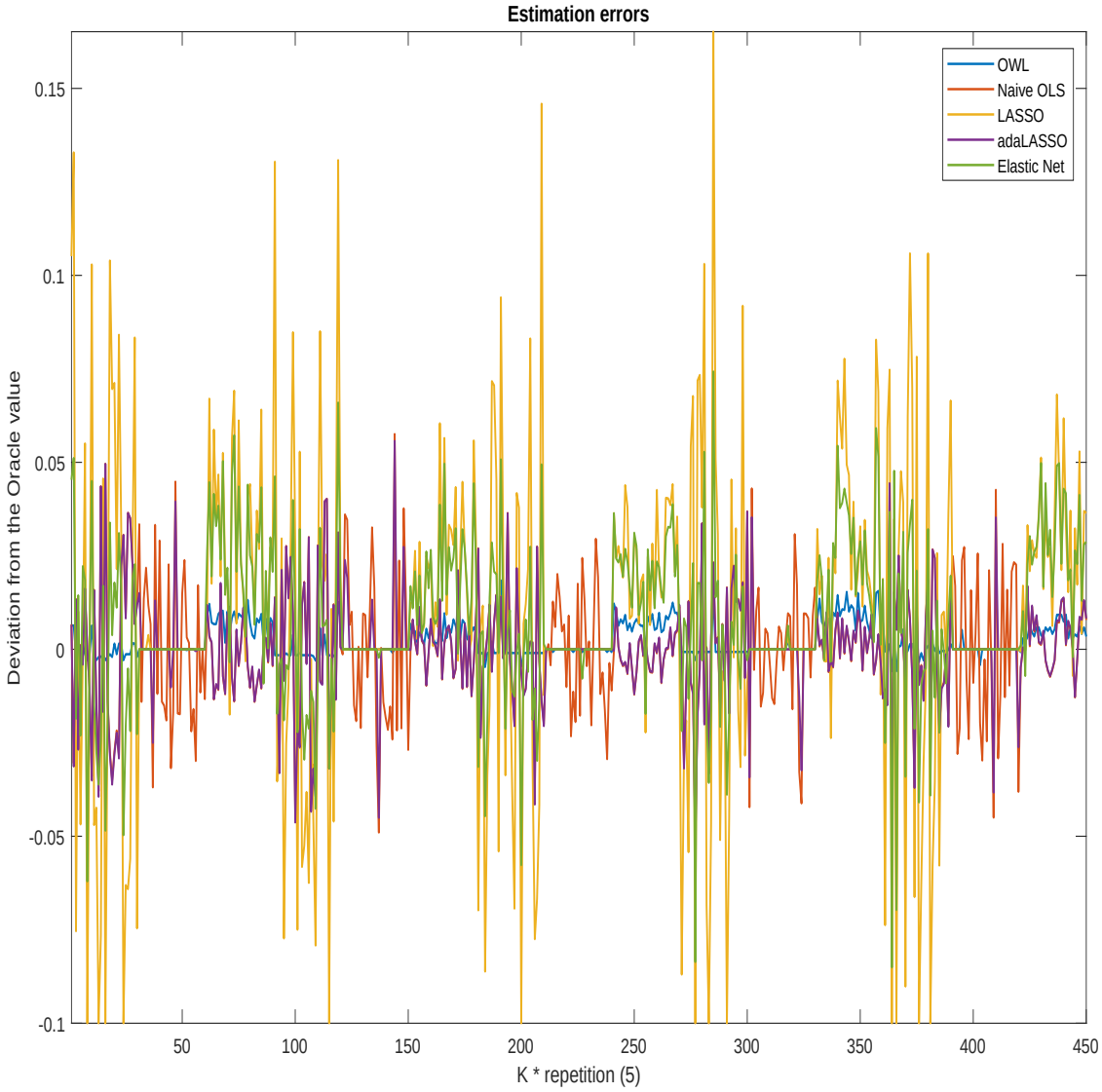


Figure 4. Simulation: distance to the oracle values

This figure shows the estimator error of each methods, measured by the distance between the oracle value and estimators, where $N = 100$. We repeat the simulation 5 times and stack estimation errors for each method.

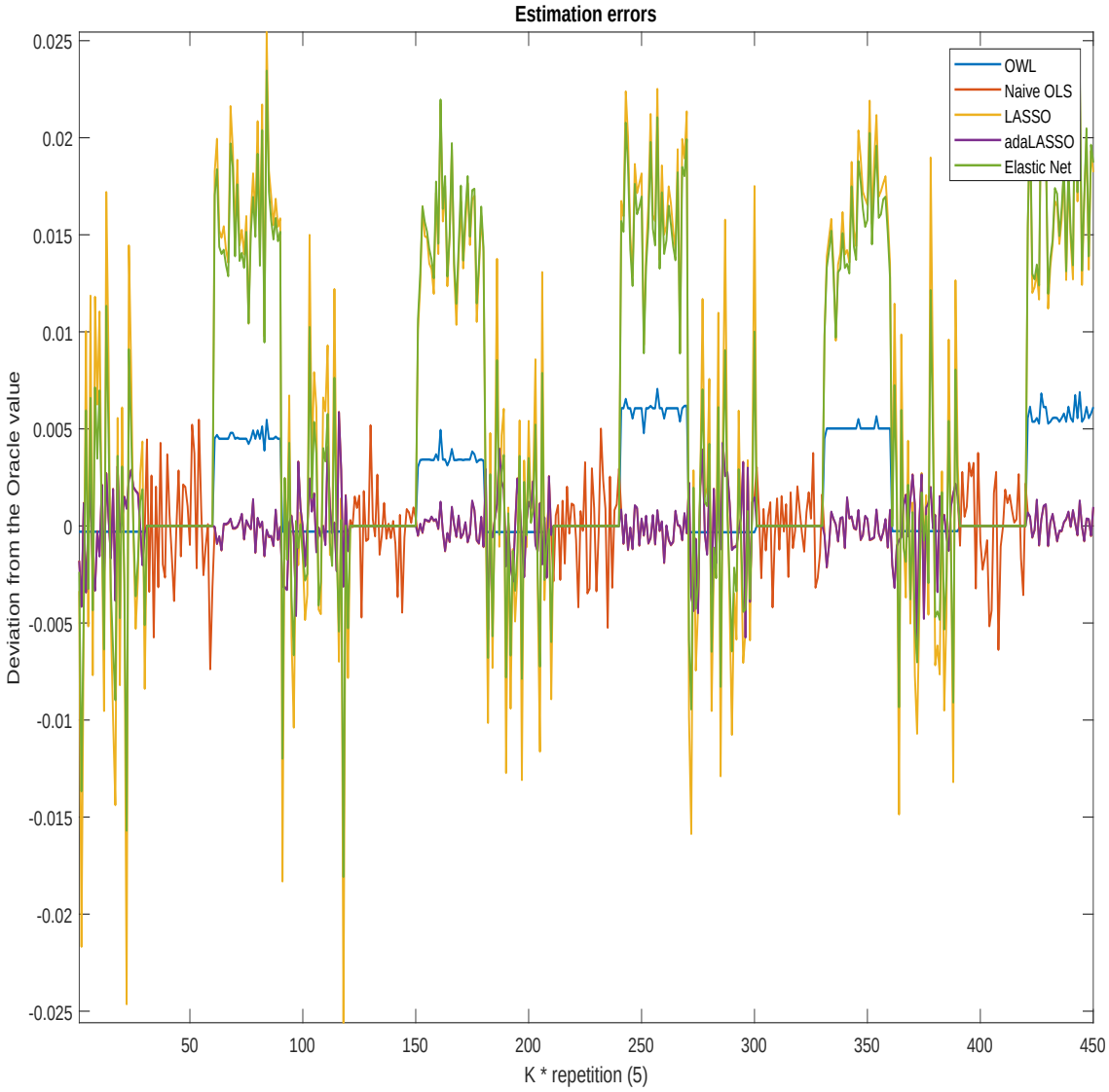


Figure 5. Simulation: distance to the oracle values

This figure shows the estimator error of each methods, measured by the distance between the oracle value and estimators, where $N = 1000$ and $K = 90$. It represents a low-dimensional setting in which $N \gg K$. Everything else is the same as in figure 4.

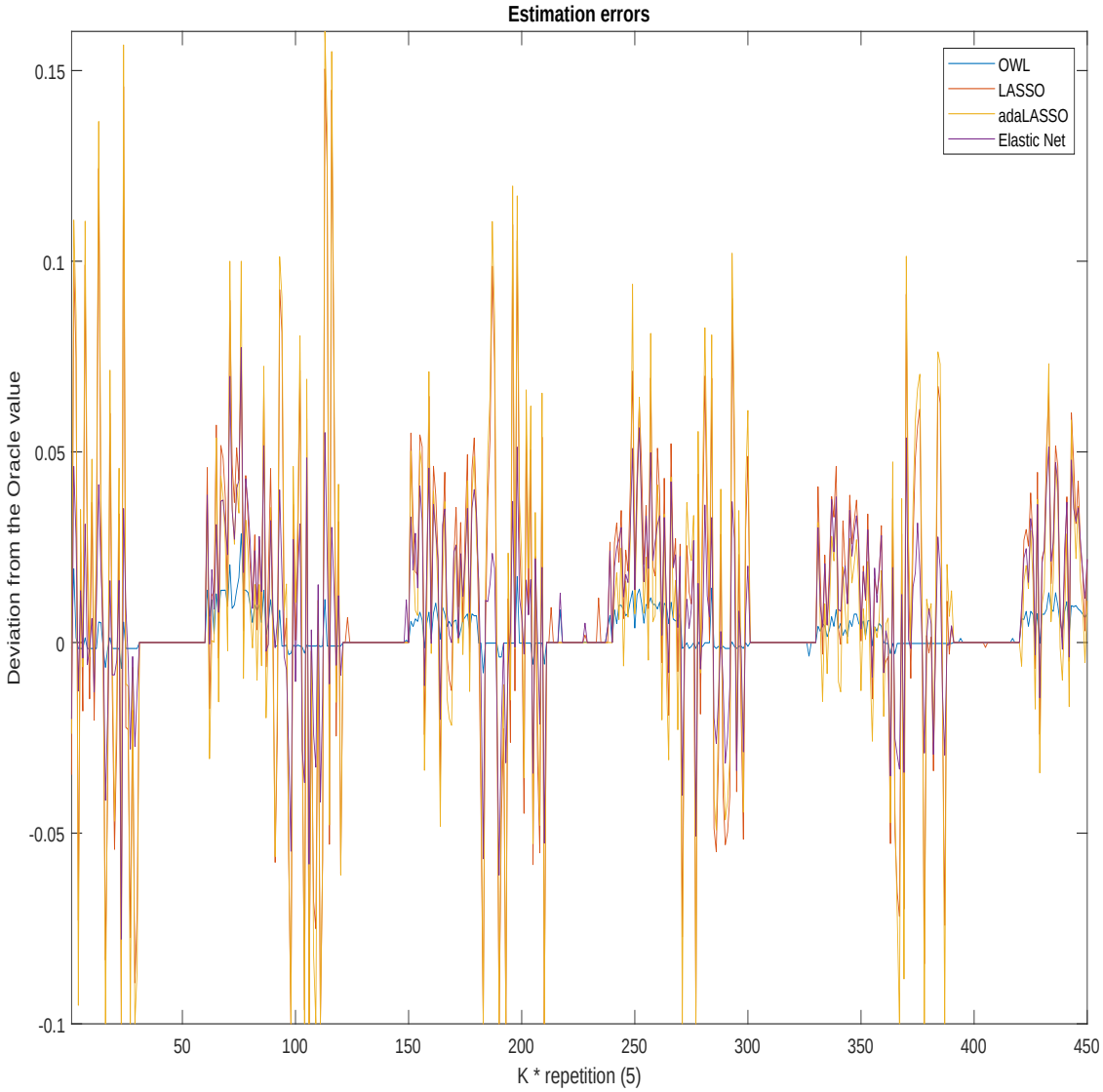


Figure 6. Simulation: Distance to the oracle values

This figure shows the estimator error of each methods, measured by the distance between the oracle value and estimators, where $N = 70$ and $K = 90$. It represents a high-dimensional setting in which $K > N$. Everything else is the same as in figure (4).

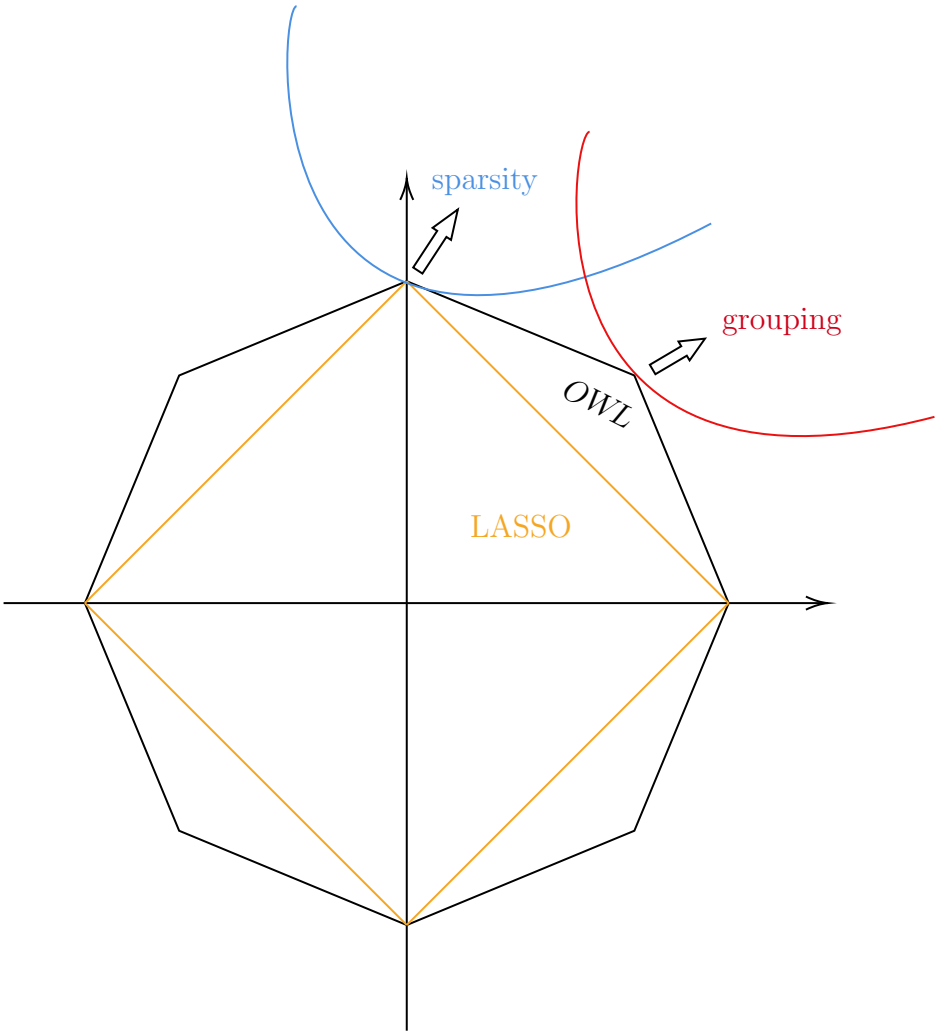


Figure 7. OWL/OSCAR norm