Regularising the Factor Zoo with OWL: A Correlation-Robust Machine Learning Approach

Chuanping Sun [*]

Abstract

Cochrane (2011) points out that the burgeoning characteristic-related "factor zoo" to explain the average returns in equity market are in disarray. This paper introduces a newly developed machine learning tool, ordered and weighted L_1 norm regularisation (OWL) to "regularise" this chaotic "factor zoo". OWL permits high correlations among explanatory variables, which is novel in the finance literature and of great importance. Factor correlation prevails in high dimensionality (factor zoo) and distorts standard estimators such as Fama-MacBeth (FM) regression, LASSO, etc. I show OWL estimator is consistent with finite factors and derive the convergence rate with infinite factors. I also derive conditions that OWL groups highly correlated variables, while shrinks off useless/redundant variables simultaneously. Monte Carlo experiments show OWL outperforms LASSO, adaptive LASSO and Elastic Net (EN) in various settings, particularly when factors are highly correlated. Empirical evidence suggests that *liquidity* related factors are primary to drive asset prices. Following Freyberger et al. (2017), out-of-sample Sharpe ratio of hedge portfolios, formed using OWL selected factors as predictors are considerably larger than that of LASSO, EN and FM.

Keywords: Factor Correlation, Cross-sectional Asset Pricing, Machine Learning, Bootstrap, LASSO, Elastic Net, Anomaly

[*] 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. Harvey et al. (2015) document 316 anomaly variables and raise concerns about data snooping. Hou et al. (2017) replicate 447 anomaly variables and find 64% to 85% are insignificant, depending on cut-off levels. Mclean and Pontiff (2016) find many anomalies vanish once they are discovered and published. Cochrane (2011) dubs this phenomenon the "factor zoo" and further argues that the characteristics related factors to explain the cross section of average returns are in disarray. He emphasises the importance of finding factors that can provide independent information about average returns, and of distinguishing factors that can be summarised by others or factors that have no explanatory powers to the cross section of average returns. Fama and French (2008) survey empirical methods used for dissecting anomalies and point out that portfolio sorting and Fama-MacBeth regression are traditionally employed to measure and test for factor's ability to drive asset prices. However, in high dimensionality, portfolio sorting will encounter the *curse of dimensionality*, while Fama-MacBeth regression will suffer from multicollinearity. 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 and weighted L_1 norm regularisation (OWL) to regularise this chaotic "factor zoo" which, to the best of my knowledge, is the first time applied in Finance. OWL permits high correlation among explanatory variables, which is of great importance. Correlation prevails in high dimensionality and can bias estimators severely if left neglected (see more details in section 2 and 3). Kozak et al. (2017), for instance, point out that the LASSO estimator will ignore correlations, and tends to pick one (highly correlated) characteristic and disregard the rest. DeMiguel et al. (2017) state that correlation between factors matters in a portfolio 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, and achieve superior portfolio performance by exploiting this correlation.

Factor correlations are common, especially in high dimensional big data. Yet, factor correlations measured by factor loadings [1] are usually much higher than those measured by their time series (see section 4 for a detailed illustration). 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, we want to check the regression of average returns on the covariance matrix of asset returns and anomaly factors (or regression of average returns on factor loadings in the Fama MacBeth framework). Kleibergen (2009) cautions that the estimation of risk premium that results from a Fama MacBeth regression is sensitive to collinearity of factor loadings.

The main empirical question of this paper is, under highly correlated anomaly factors, how to select useful factors and shrink off useless and redundant ones? OWL provides a unified solution to this question. I first show OWL estimator is consistent with finite factors. Then I derive the converge rate of OWL estimator when the number of factors goes to infinity, hence the conditions for a consistent estimator. I also show analytically that when two factors are highly correlated, OWL estimator will assign similar coefficients to them. This statistical property allows one to identify highly correlated factors and shrink off useless and redundant factors 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. (2017), I adopt a two stage (select and test) procedure to infer statistical significance of OWL estimators. In the first stage, I employ OWL to obtain a sparse set of useful factors. In the second stage, I propose a bootstrap based testing procedure to infer factor significance. At the presence of factor correlation, I bootstrap the orthogonalised asset returns to bypass multicollinearity issues (see section 2 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. However, I test factors jointly because I am interested in the joint factor inference.

DeMiguel et al. (2017) point out that firm characteristics based long-short returns and factors have different implications. Firm characteristics are computed using firm

[1] which is important in the secnod stage Fama-MacBeth regression.

specific data, for instance, accounting data or historical stock returns. Stocks are then sorted into decile portfolios according to their characteristics. Anomaly variables are obtained by computing the spread return between the top and bottom decile portfolios. Factors, on the other hand, command a common source of risk, for instance, the market return. Yet, they are closely related. Fama and French (1996) reckon that the return of a long-short hedging portfolio is a proxy for an underlying unknown risk. Kozak et al. (2018) argue there is no clear distinction between risk-factor pricing and behavioural asset pricing. The goal of this paper is to search for useful anomaly variables that explain the cross section of average returns. I make no distinction between risk-bearing factors and firm-specific characteristics based anomaly variables, and I refer to them all as factors.

In a Monte Carlo experiment, I consider 90 candidate factors with some are highly correlated. I compare OWL with LASSO, Elastic Net, adaptive LASSO, and OLS estimators. I do this experiment in two settings, one with small number of test assets ($N = 100$), another with large number of test assets ($N = 1000$). OWL is the only estimator that can successfully group together highly correlated factors by assigning them with similar coefficients; it also has the smallest error (zero, in the case of large N). On the other hand, benchmarks like LASSO, Elastic Net, and adaptive LASSO fail to identify any correlated factors and also yield very noisy estimators. When sample size is small, adaptive LASSO, influenced by the adaptive weight (i.e. the OLS estimator), performs poorly to shrink off useless/redundant factors. Nonetheless, adaptive LASSO does a good job for uncorrelated factors when sample size is large. This experiment shows that OWL has superior performance to other estimators when high correlation is present among factors.

Empirically, I initially consider 100 firm characteristics documented in Green et al. (2017), using CRSP and Compustat datasets, from January 1980 to December 2017. I first construct anomaly factors of each characteristic according to Fama and French (1992) and Fama and French (2015) [2]. I obtain 80 anomaly factors. [3]. For test portfolios, I

[2] I first discard any characteristics having more than 40% missing data. I 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, I delete the characteristic.

[3] 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

follow suggestions of Cochrane (2011), Lewellen et al. (2010) and Feng et al. (2017) by forming bi-variate sorted portfolios, and then combine them together as the grand set of test portfolios. [4]. For a robustness check, I also consider different methods of sorting, including uni-variate sorting and all combinations of bi-variate 2 by 2 sorting, finding OWL estimation is consistent in picking useful factors.

The empirical results complement and challenge some common stances in asset pricing literature.

First, I 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% correlation coefficients are higher than 0.5 (absolute value). It, however, rises to 68% when factor correlation is measured by their factor loadings (important for the second stage of Fama-MacBeth regression). This casts doubts on the validity of employing Fama-MacBeth regression to infer factor premiums. These alarmingly high correlations among factors echo Cochrane (2011)'s outcry: *in the high dimensional setting, we need to consider new methods.*

Second, OWL identifies 'market' as the primary factor important for the cross section of asset returns. This finding confirms the empirical evidence by Harvey and Liu (2017), and is consistent when 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.

Third, *liquidity* related factors are the main driver 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

difference with some familiar notations. For instance, the famous size factor 'small-minus-big' in my 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.

[4] I single out 'size' as a common characteristic to form bi-variate sorted portfolios with the remaining ones ,also see Feng et al. (2017).

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

correlation is identified by OWL by assigning them with similar coefficients. *Liquidity* related factors are more evident with smaller stocks, implying small firms face severe liquidity constraints, thus demand risk premiums to compensate for bearing the risk. Furthermore, *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. 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). 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 micro stocks (smaller than 40 percentile of NYSE listed), implying 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, OWL selected factors achieve impressive OOS Sharpe ratio for hedge portfolios using OWL selected factors as predictors. I follow a similar procedure to Freyberger et al. (2017) to conduct the OOS exercise and find that for the full sample selected factors, annualised OOS Sharpe ratio is around 3.13 when considering all stocks and around 1 once excluding micro-stocks, implying that micro stocks are main contributors to high OOS Sharpe ratio. However, the OOS Sharpe ratio is much higher once we split the full sample into two parts (before 2000 and after) and estimate each sub-sample with OWL separately. OWL selects different factors within these two sub-samples, indicating a shift in economic characteristics. The liquidity related factors are essential after the 2000 internet bubble burst, but insignificant before 2000. By contrast, momentum and profitability related factors drive asset prices primarily before 2000. Considering each sub-sample with unique OWL selected factors, the annualised OOS Sharpe ratio is above 3.5 for all stocks, and once removing micro-stocks, around 2 for the first sub-sample, and above 2.3 for the second sub-sample.

1.1 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 q-theory 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. 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 orthogonalising 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 the literature in several ways: I resort to an SDF setting instead of Fama-Macbeth regression to identify useful factors. It has important implications in terms of redundant factors: redundant factors are not priced but correlated with some priced factors and they usually have non-zero risk premiums. Under this circumstance, Fama-MacBeth regression would be ill-positioned to identify priced factors. Second, I restrict my test portfolios to sorted portfolios rather than individual stocks. The latter may suffer from missing data issues over a long period which could lead to imprecise estimation of covariances, particularly in an out-of-sample framework. Besides, micro/small stocks may dominate the result: although they only take up less than 10% of the market capitalisation, they consist of 56% of all stocks. Third and most importantly, to deal with high dimensionality with potential correlation among factors, which has not yet been discussed much in the literature, my shrinkage based estimator can identify highly correlated factors and group them together while removing useless/redundant factors simultaneously.

This paper also contributes to the vast growing literature using machine learning techniques to solve financial problems. Tibshirani (1996) proposed LASSO (L_1 norm regularisation) that achieves dimension reduction within a convex optimisation problem. Since then, many modifications and improvements have been made to achieve various targets. The LASSO family evolves rapidly. Yuan and Lin (2006) allow LASSO to shrink variables as groups by introducing the group LASSO. Freyberger et al. (2017) 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 consistent estimator (OLS estimator of risk premium) with factor loadings from the first pass of Fama-MacBeth regression. Feng et al. (2017) 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) 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.

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 closely related to Zeng and Figueiredo (2015), Figueiredo and Nowak (2016) in which they study the ordered and weighted L_1 regularisation (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 regularisation. Zeng and Figueiredo (2015) apply OWL on image processing and attain significant noise deduction.

2 Methodology

To study which factors jointly explain the cross section of average returns, I adopt the SDF method in Cochrane (2005). Section 2.1 explains the relation between risk price and risk premium and which one should be used to make factor inferences; section 2.2 points out limitations of traditional methods when facing high-dimensionality, and section 2.3 offers a remedy by imposing sparsity; sections 2.4 - 2.6 introduce OWL and discuss its statistical properties; section 2.7 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). A linear 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 $(K \times 1)$ is a 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 $(K \times 1)$ is the SDF coefficient, referred to as the risk price, it reflects whether a factor is priced or not.

I want to draw inferences on the risk prices of factors. Finding useful factors is the goal of this paper, that is factors with risk prices which are non-zero and 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 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 at each time. 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

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

Risk premium of a factor infers how much an investor demands to pay for bearing a certain risk. 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 interchangeable to use either risk price or risk premium to select factors. However, factors are likely correlated in a high dimensional setting, and our goal is to find useful factors to explain the cross-section of average returns, so we should infer on risk price rather than risk premium.

I observe a $T \times N$ matrix of test assets, denoted by R_t as excess returns. 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 no longer holds. The deviation from the equation is regarded as the pricing error. Let $m_t(b)$ be the unknown SDF which depends on the unknown risk price b. Pricing error $e(b)$ can be written and simplified as:

$$\begin{aligned} e(b) = E[R_t m_t(b)] &= E(R_t)E(m_t(b)) + cov(R_t, m_t(b)) && (3) \\ &= E(R_t)E(m_t(b)) + r_0^{-1} cov(R_t, 1 - b(f - E(f))) && (4) \\ &= r_0^{-1}(\mu_R - Cb) && (5) \end{aligned}$$

where $C = cov(R_t, f)$ is the covariance matrix ($N \times K$) of excess return and factors; μ_R ($N \times 1$) are 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. Let $Q(b)$ be the distance measure, we can recover b by minimising $Q(b)$:

$$\hat{b} = \underset{b}{argmin}\ Q(b) = \underset{b}{argmin}\ \frac{1}{2}(\mu_R - Cb)' W_T (\mu_R - Cb) \tag{6}$$

gives

$$\hat{b} = (C'W_TC)^{-1}C'W_T\mu_R \tag{7}$$

where W_T is a weighting matrix. r_0 is a constant, so it can be dropped out.

Ludvigson (2013) offers two choices of the weighting matrix W_T when comparing models. First, $E(RR')^{-1}$, the inverse of the second moment of test assets returns, which corresponds to the well known Hansen-Jaganathen (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 ($T > N$) to avoid near-singular matrix condition when estimating HJ distance, and 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 re-weight test portfolios and each characteristic sorted portfolio will be and should be treated equally.

2.2 Challenges 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 (1992) and Fama and French (2008) to construct 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, $n = 100$, it is infeasible to sort stocks into 5^{100} portfolios.

For the Fama-MacBeth regression, there are several complications in high dimensional setting too. First, the convergence rate of the risk premium estimator is $O(\sqrt{K/N})$, where N is the number of test assets and K is the number of factors. When K diverges ($K > N$), the Fama-MacBeth regression becomes infeasible. Second, variables are likely correlated under high-dimensinoality. 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.

2.3 Remedy through Sparsity

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). In this paper, I am going to impose a sparsity assumption on K candidate factors: there are only at most S useful factors, and $S << K$. With this assumption, the convergence rate of estimator becomes $\sqrt{\frac{S \log K}{N}}$, which greatly alleviates the high-dimensionality problem, and makes it feasible even in the case when $K > N$ (see section 2.6 for a detailed discussion).

Sparsity has been widely used in the machine learning literature. Tibshirani (1996) proposed the LASSO estimator which is a milestone to achieving sparsity. The LASSO penalty term takes the form of L_1 norm of parameters and it would set many coefficients to zero. Since Tibshirani (1996)'s ground-breaking work, many researchers have improved and extended LASSO to meet specific requirements. Zou (2006) added an adaptive weight (usually a first stage OLS estimator) for L_1 norm to derive the adaptive LASSO. Bryzgalova (2015) modified the adaptive LASSO to shrink off spurious factors by casting the adaptive LASSO in the Fama-MacBeth framework and using the factor loadings as adaptive weights to estimate risk premiums.

However, (adaptive) LASSO is derived from the assumption of orthogonal matrix design, which requires that factors are uncorrelated with each other. Thus, it is difficult to implement in high-dimensional setting, in which factors usually exhibit strong correlation (see section 4.2 for a detailed discussion).

Kozak et al. (2017) employed the ridge shrinkage and the elastic net in a Bayesian framework, which allows factors to be correlated. They found that a small number of principal components of characteristics based factors can approximate the SDF well.

This paper introduces a newly developed machine learning tool, the ordered and weighted L_1 norm (OWL) regularisation, to circumvent the curse of dimensionality while taking account of factor correlations.

2.4 The Ordered and Weighted L_1 (OWL) regularisation

OWL estimator is achieved by adding a penalty term in equation (6):

$$\hat{b} = \underset{b}{argmin}\ \frac{1}{2}(\mu_R - Cb)'W_T(\mu_R - Cb) + \Omega_\omega(b) \tag{8}$$

where $\Omega_\omega(b) = \omega'|b|_\downarrow$, and ω is a $K \times 1$ weighting vector, and $\omega \in \kappa$, where κ is a monotone non-negative cone, defined as $\kappa := \{x \in R^n : x_1 \geq x_2 \geq ... \geq x_n \geq 0\}$, $\omega_1 > \omega_K$. $|b|_\downarrow$ 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 (8) convex. The weighting vector ω is set to be linearly decreasing:

$$\omega_i = \lambda_1 + (K - i)\lambda_2, \quad i = 1, ..., K$$

Zeng and Figueiredo (2015), Figueiredo and Nowak (2016) show that by adopting a linear weighting scheme, OWL maps to the OSCAR (Bondell and Reich (2008)) setting, which has appealing properties of grouping highly correlated variables.

In order to solve (8), I use the proximal gradient descent algorithm. More details are in appendix.

2.5 Tuning parameters and cross-validation

OWL estimator is sensitive to the choice of the weighting vector ω. So finding appropriate values for tuning parameter λ_1 and λ_2, which pins down the weighting vector, is crucial. Following the machine learning literature, I 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, I first divide sample into five equal parts in their time series dimension. I use four parts to estimate the model with OWL. After obtaining the estimated model, I forecast the returns of the fifth part, and compute the out-of-sample root of mean squared forecast error (RMSE). I then repeat the same procedure five times by rotating the training samples and testing samples, and compute the average RMSE. Turning parameters are determined by the smallest RMSE on each point on the grid.

2.6 Statistical Properties

This section discusses OWL's statistical properties. I first show the case when the number of factors K is finite, the OWL estimator is a consistent estimator. Then I allow K to go infinity, with the sparsity assumption, I derive the convergence rate of OWL estimator, hence the conditions for a consistent estimator. Lastly, I show the grouping conditions under which two correlated factors will be assigned with the same coefficients.

When the weighting matrix W is an identity matrix [6], the model can be written as a regression model such that: $\mu_R = Cb^0 + \epsilon$, and the OWL estimator is given by $\hat{b} = \hat{b}_{OWL} = \underset{b}{argmin} \quad \frac{1}{N}||\mu_R - Cb||_2^2 + \frac{1}{N}\sum\limits_{i}[\lambda_1 + \lambda_2(K-i)]|b|_{[i]}$, where $|b|_{[1]} > |b|_{[2]} > ... > |b|_{[K]}$. I make following assumptions.

Assumption 1: Let $C_i = (C_{i,1}, ..., C_{i,j}, ..., C_{i,K})$ be a row vector of covariance matrix of returns and factors, where $i = 1, ..., N; j = 1, ..., K$ and C be normalised such that $\hat{\Sigma} = \dfrac{C'C}{N} \underset{d}{\rightarrow} \Sigma$, full rank, and $\hat{\Sigma}_{j,j} = 1$. ϵ follows a normal distribution such that $\epsilon \sim \mathbf{N}(0, \sigma^2)$, and $E(\epsilon' C^{(j)}) = 0$.

Theorem 2.1 (consistency of OWL)**.** *With assumption (1), for some* $t > 0$, $\quad \lambda_0 = 2\sigma\sqrt{\dfrac{t^2 + 2\log K}{N}}, \quad \dfrac{\lambda_1}{N} \geq \lambda_0, \quad \lambda_1 = o(N), \quad$ *and* $\quad \lambda_2 = o(N)$
with probability at least

$$\mathbf{P} = 1 - 2exp(-\frac{t^2}{2})$$

we have

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

if K *is finite,* $N \rightarrow \infty$ *then*

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

[6] When the weighting matrix 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 into the identify matrix format.

Proof: see appendix.

Theorem (2.1) shows that when K is finite, the OWL estimator converges to the true value when N goes to infinity. Assuming Gaussian distribution of the error term ϵ, and suitable conditions for λ_1 and λ_2 we can obtain the probability of this convergence is greater than $1 - 2exp(-\frac{t^2}{2})$ for some $t > 0$. The value of λ_0 is determined to utilise the Gaussian tail bounds.

Next, I show the convergence rate of OWL estimator when the number of factors K goes to infinity, and (hence) the conditions for a consistent estimator. To deal with infinite K, I impose the sparsity assumption and the compatibility assumption.

Assumption 2 (Sparsity): For $K, \quad (K \to \infty)$ number of factors, there are at most S factors that have non-zero coefficients, and $S << K$.

Assumption 3 (compatibility condition): For a set $s \subset \{1, ..., K\}, \quad b_S := b_i\mathbf{1}\{i \in s\}$, $b_{S^c} := b_i\mathbf{1}\{i \notin s\}$ and $b = b_S + b_{S^c}$. For some $\phi_0 > 0$, and for all b satisfying $||b_{S_0^c}||_1 \leq 3||b_{S_0}||_1$, it holds that: $||b_{S_0}||_1^2 \leq (b'\hat{\Sigma}b)S/\phi_0^2$.

Assumption 3 is similar to the restricted eigenvalues in Bickel et al. (2009), where ϕ_0^2 is a compatibility constant.

Theorem 2.2 (convergence rate of OWL)**.** *With assumptions (1), (2) and (3), for some* $t > 0$*, let* $\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log K}{N}}$, $\frac{\lambda_1}{N} \geq 2\lambda_0, \quad \lambda_1 = o(N)$ *, we have:*

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

with probability at least

$$\mathbf{P} = 1 - 2exp(-\frac{t^2}{2})$$

If $\lambda_2 = O(\frac{S\log K}{K})$

$$||\hat{b} - b^0||_2 = O(\sqrt{\frac{S\log K}{N}})$$

Proof: see appendix.

Theorem (2.2) shows when number of factors goes to infinity, if λ_1 and λ_2 satisfy some conditions, particularly λ_2 is of the same of smaller order of $\dfrac{S \log K}{K}$, then the convergence rate of OWL estimator is of $\sqrt{\dfrac{S \log K}{N}}$.

Theorem 2.3 (grouping)**.** *Let $\hat{b}(K \times 1)$ be a solution of (8), f_i and f_j (both $T \times 1$)be the ith and jth factors, so b_i and b_j are the coefficients in the SDF specification associated with the i^{th} and j^{th} factors. Let $\mu_R(N \times 1)$ be a vector of test asset means, and λ_2 be the tuning parameter in the weighting vector, 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.

Theorem (2.3) has several implications. First, when factors are highly correlated, they are more likely to be grouped together by assigning them with the same coefficients. It safeguards highly correlated variables from being neglected and distorted. Second, the weighting hyper parameter λ_2 which defines the distance of neighbouring weights in ω has direct impact on factor grouping. Large λ_2 encourages grouping. From a geometric perspective (more detailed geometric interpretation, see Zeng and Figueiredo (2015)), that is because a large λ_2 makes the atomic norm of OWL regulariser more pointy, thus more likely tangent to the contour coming from the unregularised quadratic minimisation solution. Third, the mean (μ_R) and standard deviation (σ_R) of test assets affect the grouping property. A set of less informative assets (small μ_R and/or small σ_R) will result in factor clusterings: all factors assigned with the same (and small) coefficients. Factors, even if they are not highly correlated, are equally inadequate to explain a set of less informative test assets.

On the other hand, orthogonal-design based estimators (that is assuming factors are uncorrelated, for instance, LASSO), will neglect factor correlation and distort factor interpretations (see section 3 for a detailed discussion). Fama-MacBeth regression usually requires deletion of factors with high correlations (for instance, see Green et al. (2017)). However, it is difficult to define a threshold to split factors.

OWL provides a unified solution to the issues faced by other estimators. No factor-trimming is required and factors with high correlation will be identified and grouped together, while useless/redundant factors will be shrunk off simultaneously.

2.7 Two stage selection procedure

In the first stage, OWL 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, I design a bootstrap test that is robust in collinearity. Instead of testing the slope coefficients by bootstrapping their standard errors, I bootstrap the null hypothesis, that is all factors have no explanatory power. 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 2 or 3 factors are tested as significant to explain the cross section of average returns. Instead, I test factor significance jointly, because I am interested in joint factor inferences.

In particular, suppose I obtain a sparse number of factors from OWL (after the first stage), I first compute the covariance of survival factors and test assets: let's denote this covariance matrix as C. Let μ_R denote the average returns of test assets. I first regress μ_R on C to obtain t_{stat} of estimated slopes and the residual series e. I then draw sub-samples with replacement from C and e, call them C^* and e^*. 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. I then compare t_{stat} estimated from real data with t^*_{stat} distribution. If t_{stat} exceeds 95 percentile of t^*_{stat} distribution, I then declare the associated coefficient is significant.

3 Simulation

This section studies the finite sample performance of OWL estimator together with other benchmarks in a Monte Carlo simulation experiment, where factors can be highly correlated.

3.1 Simulation design

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 $(K \times K)$ denote the correlation coefficient matrix of the covariance matrix of asset return and factors $C = cov(R, f)$. Let $\rho_1, \rho_2, \rho_3 \in (-1, 1)$ and ρ is divided into 3 blocks such that:

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

and

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

In bk_1 (block 1) the diagonal of matrix are ones, elsewhere are ρ_1; similarly for bk_2 and bk_3 where off-diagonal elements are ρ_2 and ρ_3, respectively. Then these three blocks constitute the diagonal direction of matrix ρ, and elsewhere is filled with zeros.

This setting implies 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.

I specify the values of ρ_1, ρ_2 and ρ_3 (some are zeros and some are non-zeros), and randomly generate an $N \times K$ matrix using the i.i.d. Gaussian distribution. Then multiply it with the Choleski decomposition of ρ to obtain the covariance matrix C, denoted as

$simC$.

I further specify an oracle value for b (risk price). Then I simulate the cross-section of average returns as $\mu_R = simC * b + e$, where e is a $N \times 1$ i.i.d. error vector with the scale about 10% of $simC$.

Finally, I estimate risk price with simulated data $simC$ and μ_R using OWL, LASSO, adaptive LASSO, Elastic Net, and naive OLS [7]. Then I compare these estimators with the oracle value of b, pre-specified.

3.2 Simulation result

In this Monte Carlo experiment, I consider 90 candidate factors ($K = 90$). 30 of them (block 1) are useful factors which are also highly correlated, with correlation coefficient $\rho_1 = 0.9$; 30 of them (block 2) are useless/redundant factors, which are also highly correlated ($\rho_2 = 0.9$); and 30 of them (block 3) are useful factors but not correlated ($\rho_3 = 0$). In the first experiment the number of assets $N = 100$ and in the second experiment the number of assets is $N = 1000$.

[Figure 1 about here.]

Figure (1) reports the plot of OWL estimator along with other benchmarks and the oracle value (black). There are 100 test assets. 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. At the presence of high correlation, LASSO performs poorly with highest estimation errors. EN is a hybrid estimator between LASSO and Ridge regression, the component of Ridge regression makes EN slightly better than LASSO. Adaptive LASSO is strongly governed by the adaptive weights which is the OLS estimator, thus adaptive LASSO and OLS estimator behave similarly. OWL produces the smallest estimation error and is the only estimator that groups together highly correlated variables by assigning them with similar coefficients, while other estimators, particularly LASSO and EN, are adversely affected by high correlation, yielding noisy and inaccurate estimators.

[7] See appendix for a concise introduction of LASSO, adaptive LASSO, and Elastic Net (EN).

The bottom left panel displays all estimators of useless/redundant factors. In terms of shrinking off useless/redundant factors, LASSO, EN, and OWL all perform well: LASSO and EN have a few outliers, but overall they set most of useless factors to zeros. OWL performs the best in which it has the smallest estimation errors. 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. OWL is the most efficient estimator. LASSO together with EN have the biggest estimation errors. Adaptive LASSO and OLS provide unbiased but more volatile estimators. Note that in the case of uncorrelated variables, both LASSO, EN and OWL are biased towards zeros, which is a trade-off between shrinkage and consistence in many shrinkage based estimators.

[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. However, for the same reason, it performs poorly to shrink off useless/redundant factors when sample size is small. LASSO and EN offer noisy and most biased estimators among all benchmarks. With highly correlated factors, OWL produces the most accurate estimation (zero errors, in this case), and is the only estimator which correctly identifies highly correlated factors.

These two experiments confirm the poor performance of LASSO. LASSO was developed for uncorrelated factor structure, if incorrectly employed for highly correlated variables, it can cause disastrous results. Adaptive LASSO is strongly influenced by the adaptive weight (OLS estimator), which makes it problematic to shrink off useless/redundant factors when sample size is small. OWL is the only estimator that can identify highly correlated factors.

4 Empirical analysis

This section applies the two stage procedure on 80 anomaly factors to infer which are priced and can explain the cross section of average returns in stock market. I first introduce the datasets, followed by a detailed account of the construction of anomaly factors and test portfolios. I consider both value weighted and equal weighted methods controlling micro stocks. Following a similar line of Feng et al. (2017), I construct pooled bi-variate sorted portfolios as test assets.

4.1 Data

I use the U.S. stock data from the Center for Research in Security Prices (CRSP) and Compustat database [8] 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.

I consider 100 firm characteristics described in Green et al. (2017) [9], while deleting characteristics that have more than 40% missing data. Then, for each remaining characteristic, I sort portfolios into deciles at each month, according to Fama and French (1992) and Fama and French (2015). 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 should be treated with caution. 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. Overall, I obtain 80 anomaly factors, see table (1) for a detailed description.

[Table 1 about here.]

There is a debate in the literature about using either individual stocks or sorted

[8] downloaded from the Wharton Research Data Services

[9] I am grateful to Jeremiah Green for providing SAS code to compute firm characteristics. I modified the SAS code to cope with only CRSP and Compustat database.

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 2-3 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. (2017) 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 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 a 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), Fama and French (2008), Fama and French (2015), used bi-variate sorting to create the 5 by 5 test portfolios and they have now become popular choices of test assets. However, Harvey et al. (2015) caution that when only a small set of sorted portfolios are considered, 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. (2017) 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 5 by 5 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), Fama and French (2015), Hou et al. (2014), Carhart (1997) all include the 'size' and the 'market' factors in their models. Asness et al. (2018) find size matters while controlling other variables.

To strike a balance between using sorted portfolios and individual stocks as test assets, I follow Feng et al. (2017) by singling 'size' out as a common characteristic, together with the remaining characteristics to form bi-variate sorted 25 portfolios. I drop any test portfolios which have insufficient stocks (due to missing data) to sort. Finally, I group them together, which amounts to 1927 test portfolios.

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.

4.2 Factor correlation

[Figure 3 about here.]

Figure (3) 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 4 about here.]

Figure (4) displays the heat map of factor correlation coefficients matrix measured by factor loadings. It exhibits much higher correlation compared to figure (3): 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.

4.3 Which factors matter?

Considering high correlation among factors, I apply the two stage procedure to select useful factors from the 80 candidate factors. I first employ OWL to shrink off useless/redundant factors, obtaining a sparse number of survival factors. In the second stage I use bootstrap method described in section 2 to test survival factors.

[Figure 5 about here.]

Figure (5) shows the convergence of OWL estimation using Fista-OWL with backtracking algorithm (see appendix). Vertical axis shows the distance between the k^{th} estimation and the optimiser. Horizontal axis shows the number of iterations (steps) until a stopping criterion is met. I set a tight stopping criterion of OWL searching for the optimiser, 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 optimiser because of a tight stopping criterion.

[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 with the full sample, ranging from January 1980 to December 2017; columns 6-7 report results from 1980 to 2000, and columns 8-9 from 2001-2017. Both the value weighted (vw) and equal weighted (ew) methods are considered. In order to gauge the impact of small stocks, I consider three thresholds for micro stocks. Before sorting test portfolios, I screen out stocks with market capitalisation smaller than 20, 30 or 40 percentile of all NYSE listed stocks. This table lists all anomaly factors selected by the two-stage procedure in each estimation. It also reports the ordinal number after each factor selected by OWL (in the bracket), indicating the importance of each factor (smaller number implies bigger impact).

'Size' (mve) has been selected as the most important factor in most of these estimations, which is not surprising. 'Size' characteristic has multiple entries in forming test portfolios, thus 'size' impact prevails in test portfolios. For this reason I exclude 'size' factor as a competing factor, yet I include it in the table to show that OWL can correctly identify relevant 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 (i.e. removing stocks that are smaller than 20 or 30 percentile) of NYSE listed stocks exhibit higher importance of 'illiquidity' (smaller ordinal number after OWL selection) 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. In the next subsection, I will show that portfolios can achieve high Sharpe ratios by taking advantage of their correlation.

'Asset growth rate' (agr) follows 'illiquidity' and 'standard deviation of dollar volume' as the third most important 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)). I also find 'asset growth rate' is more prominent with smaller stocks, with equal weighted method showing 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.

Column 6 and 7 report estimations using the 1980-2000 sub-sample and column 8 and 9 report estimations using the 2001-2017 sub-sample. I find 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) market shows no strong evidence of liquidity related factors to drive asset prices. On the contrary, 'momentum' and 'profitability' are the most important factors between 1980 and 2000.

Interestingly, during 1980 to 2000, with 20-percentile-micro-stocks excluded, I 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 was evident again, which implies the vanishing of size effect is likely to be caused by some small "junk" stocks. When removing these junk stocks, size effect resurfaces again, which echoes the discovery by Asness et al. (2018): *size matters, if you control your junk.*

4.4 Robustness check

In this section, I want to check whether liquidity related factors are robust in explaining the cross section of asset returns and also how small stocks affect factors' implications.

For the first task, I will consider four types of sorting methods for constructing test portfolios and check whether liquidity related factors are consistent to drive asset prices. First, I 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, I consider bi-variate sorting, but with all possible combinations of 80 characteristics, that is 3240 possible combinations. To reduce the dimension of test portfolios, I consider the 2 by 2 (instead of 5 by 5) sorting, that is I sort all stocks into high and low groups where the threshold is the median of each characteristic. I obtain 3240×4, total 12960 test portfolios. Third, I consider a similar method in empirical analysis, that is singling out 'size' as a common characteristic, and using it with the remaining characteristics to form bi-variate sorted portfolios; however, instead of forming the 5 by 5 portfolios, I form 3 by 3 portfolios. Fourth, I consider the sorting method used in empirical analysis, that is 5 by 5 bi-variate sorting between 'size' and the remaining characteristics.

For the second task, I use the same sorting method as in empirical analysis, but I 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). I want to check how factors' implications vary within each scenario.

[Figure 6 about here.]

Figure (6) 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.

[Figure 7 about here.]

Figure (7) reports the heat map of OWL estimation before bootstrap test. We find some clear patterns. First, micro stocks alter market factor's interpretation to drive asset prices. When micro stocks are all included to form test portfolios, I find 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, I 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.

4.5 Liquidity as a risk factor

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. That implies 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). 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.

4.6 Out-Of-Sample Sharpe ratio

In this subsection, I 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 overfit. Freyberger et al. (2017) point out that OOS exercise ensures that in-sample overfit does not explain superior performance. Although the 5-fold cross validation method used for evaluating OWL hyper parameters [10] 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.

I follow a similar procedure to Freyberger et al. (2017) to form hedge portfolios using a rolling window scheme, to predict returns of each test assets, that is the bi-variate sorted portfolios, OOS. Rolling window size is 120 months (10 years). Specifically, at the end of the estimation window, I regress each test asset on factors selected by the two-stage procedure, but one period back. For instance, at time t, I regress each test asset return from $t - 120 - 1$ to t on selected factors from $t - 120 - 2$ to $t - 1$, and obtain $\hat{\beta}$. I then forecast each test asset's next period return (at $t + 1$) by multiplying $\hat{\beta}$ and selected factors at t. I then sort stocks by their predicted returns into decile portfolios. I then long the top decile and short the bottom decile. At the next period $(t + 1)$, when returns are realised, I can compute the spread portfolio return. Then roll the window one period forward and repeat the steps until the end of period. In the end I compute the Sharpe ratio based on the OOS returns.

[10] use 4 folds to estimate the model and 1 fold to evaluate the model performance OOS.

For the fact that OWL selects some different factors for some sub-periods, I also evaluate the OOS performance for two sub-samples. OWL selected factors may differ in each sub-period. In particular, in the 1980 to 2000 sub-sample, [11] the top 3 OWL selected factors are 'momentum', 'return on asset' and 'sales cash ratio' which are distinguished from other periods. The second sub-sample estimation suggests 'illiquidity' related factors are most important to explain the cross section of average returns.

[Table 3 about here.]

Table 3, panel A reports that annualised OOS Sharpe ratio of all stocks is 3.1340, where OWL selected factors are 'illiquidity' related factors. But when excluding small stocks, OOS Sharpe ratio declines drastically: excluding stocks smaller than 20 percentile of NYSE listed stocks, OOS Sharpe ratio drops by around half; and drops a further third when excluding stocks smaller than 40 percentile of NYSE listed stocks. This finding is consistent with Freyberger et al. (2017) and Lewellen (2015) that micro stocks contribute largely to the high out-of-sample Sharpe ratio.

Panel B shows that in the first sub-sample, where prevailing factors are 'momentum' and 'profitability' related factors, annualised OOS Sharpe ratio is 3.7603 for all stocks. OOS Sharpe ratios for stocks larger than 20 or 40 percentile of NYSE listed stocks did not drop as much as in the full sample: 1.9714 and 1.8294 respectively.

Panel C shows that in the second sub-sample, where 'illiquidity' related factors mainly drive the cross-sectional asset returns, annualised OOS Sharpe ratio is 3.5763 for all stocks, and declines even less for larger stocks: 2.2309 and 2.3701 for stocks larger than 20 and 40 percentile of NYSE listed stocks, respectively.

On the other hand, out-of-sample Sharpe ratios of LASSO, EN and FM are smaller than OWL. FM is typically the worst performer. EN and LASSO select the same factors in many cases, which is not surprising. Because the shrinkage component of EN is a combination of LASSO and Ridge shrinkage.

Sub-sample analysis suggests prevailing factors may change over time. A shift in economy may drive factors' contribution to explaining the cross section of stock returns to vary. 'Profitability' factors drive asset returns in the first sub-period and 'liquidity'

[11] excluding stocks whose size are less than 20 percentile of NYSE listed stocks

factors dominate the second sub-period that is after the 2000 internet bubble burst. In a full sample estimation, prevailing factors in the first sub-sample are suppressed by 'liquidity' related factors which are essential to explain the second half of sample. That explains why the OOS Sharpe ratio increases dramatically after splitting into two sub-samples.

5 Conclusion

In the zoo of factors, traditional methods to find useful factors that can explain the cross section of stock average returns face tremendous challenges. Correlation in the factor zoo makes the challenge even harsher. Yet, factor correlation should not be neglected, as it causes severe consequences in standard analytical tools. For instance, (Adaptive) LASSO ignores factor correlation and picks up a small set of highly correlated variables randomly while discarding the rest. LASSO also fails to shrink off useless/redundant factors when factors are highly correlated. In a high-dimensional setting, Fama-MacBeth regression faces multicollinearity issues. Among 80 anomaly factors I considered, I find 64% are highly correlated (absolute value of correlation coefficient is greater than 0.5) when investigating factor loadings.

I introduce a newly developed machine learning tool, the ordered and weighted L_1 norm (OWL) regularisation, which is designed to cope with high correlations among explanatory variables. OWL groups together highly correlated variables by assigning them with similar coefficients.

Empirical analysis shows that 'illiquidity' related factors play an important role in explaining the cross section of average stock returns. A small set of (3 or 4) OWL selected factors, usually highly correlated, explains a bulk of the cross section of average returns, demonstrating strong Sharpe ratios (in-sample and out-of-sample), high cross sectional R^2, and small HJ distance and GRS statistic. Out-of-sample Sharpe ratio of hedge portfolios formed by using OWL selected factors as predictors is around 3.5 (annualised) for all stocks, and above 2.3 for non-micro stocks in the past two decades.

However, it is worth stressing the importance of using sorted portfolios rather than individual stocks as test assets. Many papers have argued that error-in-variables (EIV) will bias testing if individual stocks are used. For that, Shanken (1992) proposed the Shanken's correction. However, there are two other major shortcomings of using individual stocks as test assets.

First, micro stocks (market capitalisation smaller than 20 percentile of NYSE listed) will dominate the estimation result. Although micro stocks comprise less than 10 percent of aggregated stock capitalisation, they constitute 56% of all stocks. Hence, if individual stocks are used, estimation will primarily explain only a small fraction of the market value.

Second, individual stocks face tremendous challenges of missing data. The typical treatment is to delete stocks with many missing data. For example, deleting stocks with any missing data will lead to only a handful of stocks surviving over a long period. Alternatively, a threshold of missing data can be set to determine which stocks to keep, for instance, deleting stocks with more than 20% missing data. Then when evaluating historical covariance matrix, delete rows with any missing data. This treatment, however, would have extra challenge within an out-of-sample estimation framework. With a smaller (than full sample) rolling window, after deleting rows with missing data, the estimation of covariance matrix is inaccurate, and very often leads to non-invertible covariance matrix. Sorted portfolios, on the contrary, bypass all the shortcomings of individual stocks. Sorting portfolios at each point of time avoids missing data issues. Before sorting, micro stocks can be removed (or set up thresholds to control the effect of small stocks) to mitigate the issue with small stocks. Additionally, value weighted method can further alleviate small stock impact. Fama and French (2008) have already shown how sorted portfolios can alleviate the error-in-variables.

Finally, note that the purpose of this paper is not to find a parsimonious asset pricing model (since OWL selected factors are usually highly correlated), but to identify a set of sparse factors to explain the cross section of average returns. With that in mind, my procedure is particularly useful for factor investing: OWL can identify correlated factors that jointly drive stock returns, which can then be utilised to form portfolio strategies. Asness et al. (2013) find 'momentum' and 'value' are negatively correlated, and this correlation can be further exploited to achieve high-performance portfolio strategies.

DeMiguel et al. (2014) employed a VAR(1) model to explore the correlation among stocks and find consistent superior out-of-sample performance. Ordered and weighted L_1 norm regularisation is a general tool useful for sparsity selection which permits correlations. Since stocks are usually correlated, future research can be extended to explore portfolio selection strategies, where individual stock weights are regularised by OWL.

Appendix A Solve the OWL optimisation problem

A.1 OWL proximal function

First define the proximal function as

$$Prox_{\Omega_\omega}(b) = argmin_x \frac{1}{2}||x - b||_2^2 + \Omega_\omega(x) \tag{9}$$

With the definition of $\Omega_\omega(b)$, we have:

$$\Omega_\omega(b) = \Omega_\omega(|b|) \tag{10}$$

It is easy to show that

$$||b - sign(b) \odot |x|||_2^2 \leq ||b - x||_2^2 \tag{11}$$

where $sign(.)$ is a function to retrieve signs from a vector, with elements in $\{1, -1, 0\}$. $\odot$ is a point-wise production operator.
(10) and (11) infer:

$$Prox_{\Omega_\omega}(b) = sign(b) \odot Prox_{\Omega_\omega}(|b|) \tag{12}$$

Let P be a permutation matrix that orders a vector decreasingly, we have $||P(x - b)||_2^2 = ||x - b||_2^2$, and with the definition of $\Omega_\omega(b)$, we have: $\Omega_\omega(b) = \Omega_\omega(Pb)$. These two equations imply:

$$Prox_{\Omega_\omega}(b) = sign(b) \odot P'(|b|)Prox_{\Omega_\omega}(|b|_\downarrow) \tag{13}$$

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 $|b|$.

For any $|b|_\downarrow \in \kappa$, where κ is a monotone non-negative cone, defined above:

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

where $x^* \in \kappa$. It infers: $Prox_{\Omega_\omega}(|b|_\downarrow) \in \kappa$, and $\Omega_\omega(x) = \omega' x$, Further, we have:

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

which is the projection of $(|b|_{\downarrow} - \omega)$ onto κ, Then equation (13) can be written as:

$$Prox_{\Omega_\omega}(b) = sign(b) \odot (P'(|b|) Proj_{\kappa}(|b|_{\downarrow} - \omega)) \tag{14}$$

where $Proj_{\kappa}(.)$ is the projection operator onto κ. [12]

After solving the proximal function, we can employ the iterative soft-thresholding algorithm.

First initialise $b^{(0)}$, then repeat:

$$b^{(k+1)} = prox_{\Omega_\omega}(b^{(k)} - sz_k \bigtriangledown g(b^{(k)})) \tag{15}$$

until convergence. 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 each iteration of k.

To achieve the optimal convergence rate, I consider the accelerated proximal gradient method, also regarded as the fast iterative soft-thresholding algorithm (FISTA), which has a much faster rate to converge.

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

A.2 FISTA algorithm

Algorithm 1: FISTA-OWL

```
Input: $\mu_R, C, \omega$
Output: $\hat{b}$ in (8)
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] $Q(b_k, u_k) = \frac{1}{2}||\mu_R - Cu_k||_2^2 - (b_k - u_k)'C'(\mu_R - Cu_k) + \frac{1}{2\tau_k}||b_k - u_k||_2^2$

Appendix B Proof of Theorem (2.1)

The OWL estimator is minimising the function such that:

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

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

$$\frac{1}{N}||\mu_R - C\hat{b}||_2^2 + \frac{1}{N}\sum_i [\lambda_1 + \lambda_2(K-i)]|\hat{b}|_{[i]} \leq \frac{1}{N}||\mu_R - Cb^0||_2^2 + \frac{1}{N}\sum_i [\lambda_1 + \lambda_2(K-i)]|b^0|_{[i]} \tag{16}$$

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 $\mu_R = Cb^0 + \epsilon$, (16) can be simplified as:

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

in which,

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

Let λ_0 be a constant, such that

$$\frac{1}{N} \underset{1 \leq j \leq K}{max} 2|\epsilon' C^{(j)}| \leq \lambda_0 \tag{18}$$

(17) can be written as:

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}||_1 \leq \lambda_0||\hat{b} - b^0||_1 + \frac{1}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1 \tag{19}$$

Use triangle inequality, we have:

$$||\hat{b} - b^0||_1 \leq ||\hat{b}||_1 + ||b^0||_1$$

(19) can be written as:

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 + (\frac{\lambda_1}{N} - \lambda_0)||\hat{b}||_1 \leq [\lambda_0 + \frac{\lambda_1 + \lambda_2(K-1)}{N}]||b^0||_1 \tag{20}$$

Since $\frac{\lambda_1}{N} - \lambda_0 \geq 0$ and $\lambda_1 = o(N), \lambda_2 = o(N)$ we have:

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 \leq [\lambda_0 + \frac{\lambda_1 + \lambda_2(K-1)}{N}]||b^0||_1 \tag{21}$$

$$(\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0) \leq [\lambda_0 + \frac{\lambda_1 + \lambda_2(K-1)}{N}]||b^0||_1 \tag{22}$$

Now we compute the probability of this inequality.
Let $\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log K}{N}}$, for any $t > 0$ and $V_j := \epsilon' C^{(j)}/\sqrt{N\sigma^2} \backsim \mathbf{N}(0,1)$, since C has been normalised such that $\hat{\sigma_j^2} = 1$.
Using the Gaussian tail bound and union bound, we have:

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

Consequently, the probability of (18) is

$$\mathbf{P} \geq 1 - 2exp(-\frac{t^2}{2})$$

Appendix C Proof of Theorem (2.2)

Using the "argmin" inequality on OWL, we have:

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 + \frac{1}{N}\lambda_1||\hat{b}||_1 \leq \lambda_0||\hat{b} - b^0||_1 + \frac{1}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1 \tag{23}$$

Let $\frac{\lambda_1}{N} \geq 2\lambda_0$, (23) can be written as:

$$\frac{2}{N}||C(\hat{b}-b^0)||_2^2 + \frac{2}{N}\lambda_1||\hat{b}||_1 \leq \frac{\lambda_1}{N}||\hat{b}-b^0||_1 + \frac{2}{N}[\lambda_1 + \lambda_2(K-1)]||b^0||_1 \tag{24}$$

in which,

$$||\hat{b}||_1 = ||\hat{b}_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1 \geq ||b^0_{s_0}||_1 - ||\hat{b}_{s_0} - b^0_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1$$

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

then (24) can be written as:

$$\frac{2}{N}||C(\hat{b}-b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}_{s_0^c}||_1 \leq 3\frac{\lambda_1}{N}||\hat{b}_{s_0} - b^0_{s_0}||_1 + \frac{2\lambda_2(K-1)}{N}||b^0||_1 \tag{25}$$

Using (25), we can obtain below inequality:

$$\begin{aligned}\frac{2}{N}||C(\hat{b}-b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}-b^0||_1 &= \frac{2}{N}||C(\hat{b}-b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}_{s_0} - b^0_{s_0}||_1 + \frac{\lambda_1}{N}||\hat{b}_{s_0^c}||_1 \quad (26)\\ &\leq 4\frac{\lambda_1}{N}||\hat{b}_{s_0} - b^0_{s_0}||_1 + \frac{2\lambda_2(K-1)}{N}||b^0||_1 \quad (27)\end{aligned}$$

Using compatibility condition on $||\hat{b}_{s_0} - b^0_{s_0}||_1$ we have:

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

where the second inequality is using $4ab \leq 4a^2 + b^2$. So (27) can be written as:

$$\frac{1}{N}||C(\hat{b}-b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}-b^0||_1 \leq 4(\frac{\lambda_1}{N})^2 S/\phi_0^2 + \frac{2\lambda_2(K-1)}{N}||b^0||_1 \tag{28}$$

Give $\frac{\lambda_1}{N} = 2\lambda_0 = 4\hat{\sigma}\sqrt{\frac{t^2 + 2\log K}{N}}$ and $\lambda_2 = O(\frac{S\log K}{K})$, both two terms on the right hand side of (28) are $O(\frac{S\log K}{N})$, so

$$||\hat{b}-b^0||_2 = O(\sqrt{\frac{S\log K}{N}})$$

Appendix D Proof of Theorem (2.3)

The proof of theorem (2.3) relies on the Pigou-Dalton-transfer and directional derivative lemma.

Lemma 1 (Pigou-Dalton-Transfer(P.D.T))**.** *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 8, and Δ_ω is the smallest gap in weighting vector ω.

Lemma 2. *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 $x^* \in argmin(f)$, if and only if $f'(x^*, u) \geq 0$ for any u.

Proof: Denote the object function as $Q = \frac{1}{2}(\mu_R - Cb)' W_T (\mu_R - Cb) + \Omega_\omega(b)$. Let $\hat{b}$ be a solution of (8).

Suppose

$$\sigma_{f_i - f_j} < \frac{\lambda_2}{||\mu_R||_2 ||\sigma_R||_2}$$

and

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

assume $\hat{b}_i > \hat{b}_j$ without loss of the generality (we want to find a condition under which this assumption is violated, and thus we have a contradiction between the implied condition and the assumption).

The directional derivative of Q at $\hat{b}$ with $u_i = -1, u_j = 1, u_k = 0, \forall k \neq i, j$, is:

$$\begin{aligned} Q'(\hat{b}, u) &= \lim_{\alpha \to 0^+} \frac{||\mu_R - C\hat{b} + \alpha(C_i - C_j)||_2^2 - ||\mu_R - C\hat{b}||_2^2}{2\alpha} + \lim_{\alpha \to 0^+} \frac{\Omega_\omega(\hat{b} + \alpha u) - \Omega_\omega(\hat{b})}{\alpha} \\ &= (\mu_R - C\hat{b})(C_i - C_j) + \lim_{\alpha \to 0^+} \frac{\Omega_\omega(\hat{b} + \alpha u) - \Omega_\omega(\hat{b})}{\alpha} \end{aligned}$$

Applying the Pigou-Dalton-transfer on the OWL norm, we have:

$$\begin{aligned} Q'(\hat{b},u) &\leq (\mu_R - C\hat{b})(C_i - C_j) - \lim_{\alpha \to 0^+} \frac{\Delta_\omega \alpha}{\alpha} \\ &= (\mu_R - C\hat{b})(C_i - C_j) - \Delta_\omega \\ &= (\mu_R - C\hat{b})(C_i - C_j) - \lambda_2 \end{aligned}$$

In the linear weighting scheme of OWL, each neighbouring weight has the same distance, that is $\Delta_\omega = \lambda_2$.

Cauchy-Schwarz inequality states that for any vector u and v, $< u, v >$ is the inner product of vector u and v, we have:

$$| < u, v > | \leq ||u||_2 ||v||_2$$

And since $\mu_R - C\hat{b}$ is a pricing error, we can establish $||\mu_R - C\hat{b}||_2 < ||\mu_R||_2$. We have:

$$\begin{aligned} Q'(\hat{b},u) &\leq ||\mu_R - C\hat{b}||_2 ||C_i - C_j||_2 - \lambda_2 \\ &< ||\mu_R||_2 ||cov(R, f_i - f_j)||_2 - \lambda_2 \end{aligned}$$

Using the Cauchy-Schwarz inequality again on the covariance term:

$$\begin{aligned} Q'(\hat{b},u) &< ||\mu_R||_2 ||\sigma_R||_2 \sigma_{f_i - f_j} - \lambda_2 \\ &< 0 \end{aligned}$$

which violates the directional derivative lemma. Hence there is a contradiction. So if $\hat{b}$ is an optimiser of $Q(\hat{b},u)$ we must have:

$$\hat{b_i} = \hat{b_j}$$

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

Let y denote a vector of responses which is a $N \times 1$ column 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 below problem:

$$\hat{\beta}_{a.d.Lasso} = \underset{\beta}{argmin} \quad ||y - X\beta||^2 + \lambda||\beta||_1 \tag{29}$$

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

Elastic net solves the below problem:

$$\hat{\beta}_{a.d.Lasso} = \underset{\beta}{argmin} \quad ||y - X\beta||^2 + \lambda\alpha||\beta||_1 + \lambda(1-\alpha)||\beta||_2^2 \tag{30}$$

Elastic net is combines the L_1 (or Lasso) penalty and the L_2 (or Ridge) penalty together, which gives more robust results when variables are correlated; however, it is not designed to select highly correlated factors.

Adaptive Lasso minimise the following problem:

$$\hat{\beta}_{a.d.Lasso} = \underset{\beta}{argmin} \quad ||y - X\beta||^2 + \lambda \sum_{i=1}^{K} \frac{1}{|\hat{\beta}_{ols}|^{\gamma}} |\beta_i| \tag{31}$$

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 adds an adaptive weight, for instance, the first stage OLS estimator, to each of the element of the Lasso penalty. The variables with small (absolute value) OLS estimated coefficients receive stronger penalty.

OSCAR (Octagonal shrinkage and clustering algorithm for regression) solves the below problem:

$$\hat{\beta}_{OSCAR} = \underset{\beta}{argmin} \quad ||y - X\beta||^2 + \lambda_1||\beta||_1 + \lambda_2\sum_{i<j} max\{|\beta_i|, |\beta_j|\} \tag{32}$$

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) shows that by adopting a linear decreasing weighting vector, OWL 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.

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.

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.

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): "Time series for macroeconomics and finance," *Manuscript, University of Chicago*, 1–136.

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

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.

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.

——— (1996): "Multifactor explanations of asset pricing anomalies," *Journal of Finance*, 51, 55–84.

——— (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 (2017): “Taming the Factor Zoo,” *SSRN Electronic Journal*, 1–56.

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 (2017): “Dissecting Characteristics Nonparametrically,” *Stockholm School of Economics TAU Finance Conference.*

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 U.S. Monthly Stock Returns,” *The Review of Financial Studies*, 1–80.

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.

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

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

——— (2018): "Interpreting Factor Models," *Journal of Finance*, LXXIII.

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.

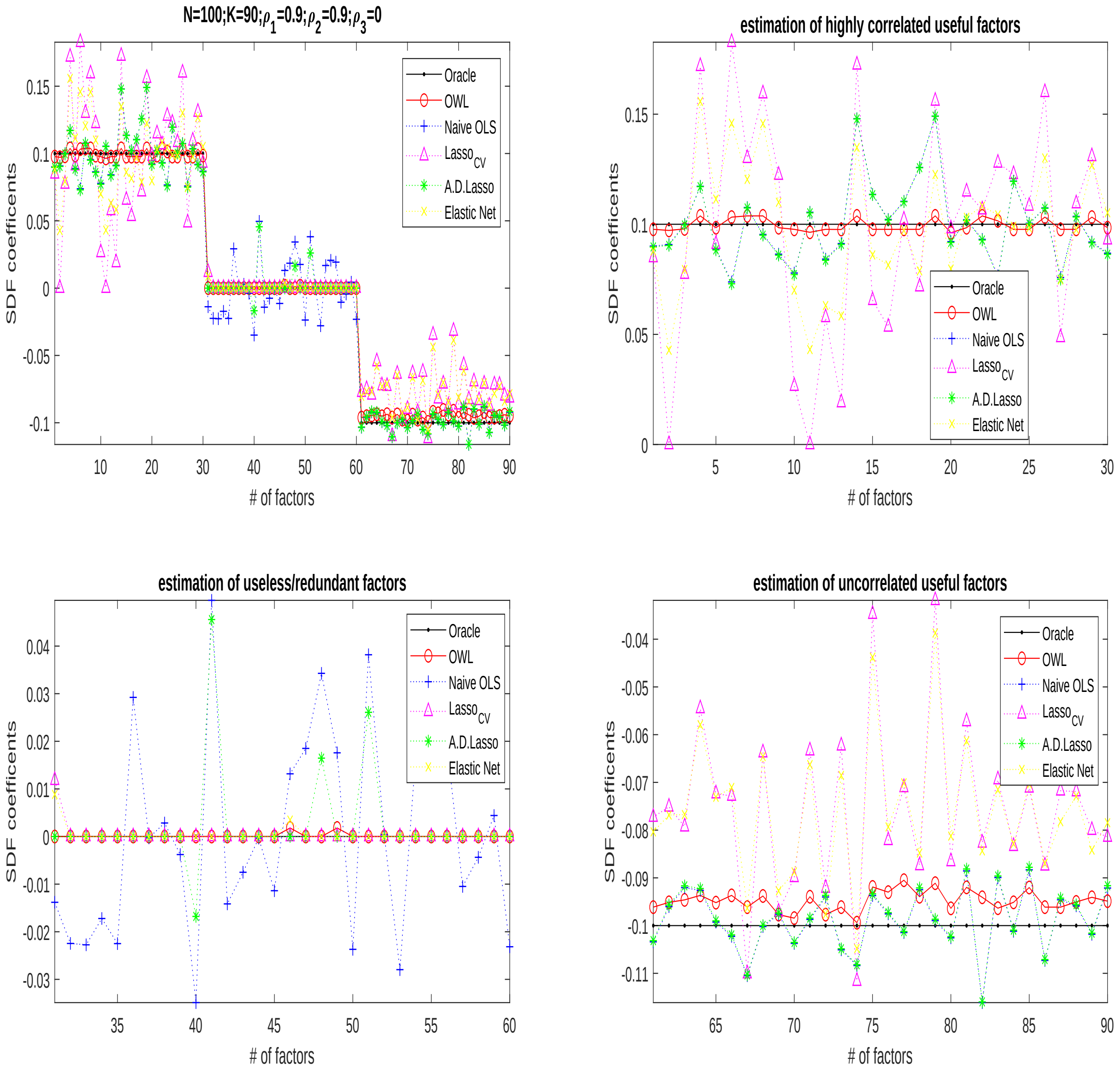


Figure 1. Simulation: $N = 100, K = 90$

This figure reports the plot of OWL estimator 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 (red) is displayed along with LASSO, adaptive LASSO, Elastic Net, and native OLS estimators.

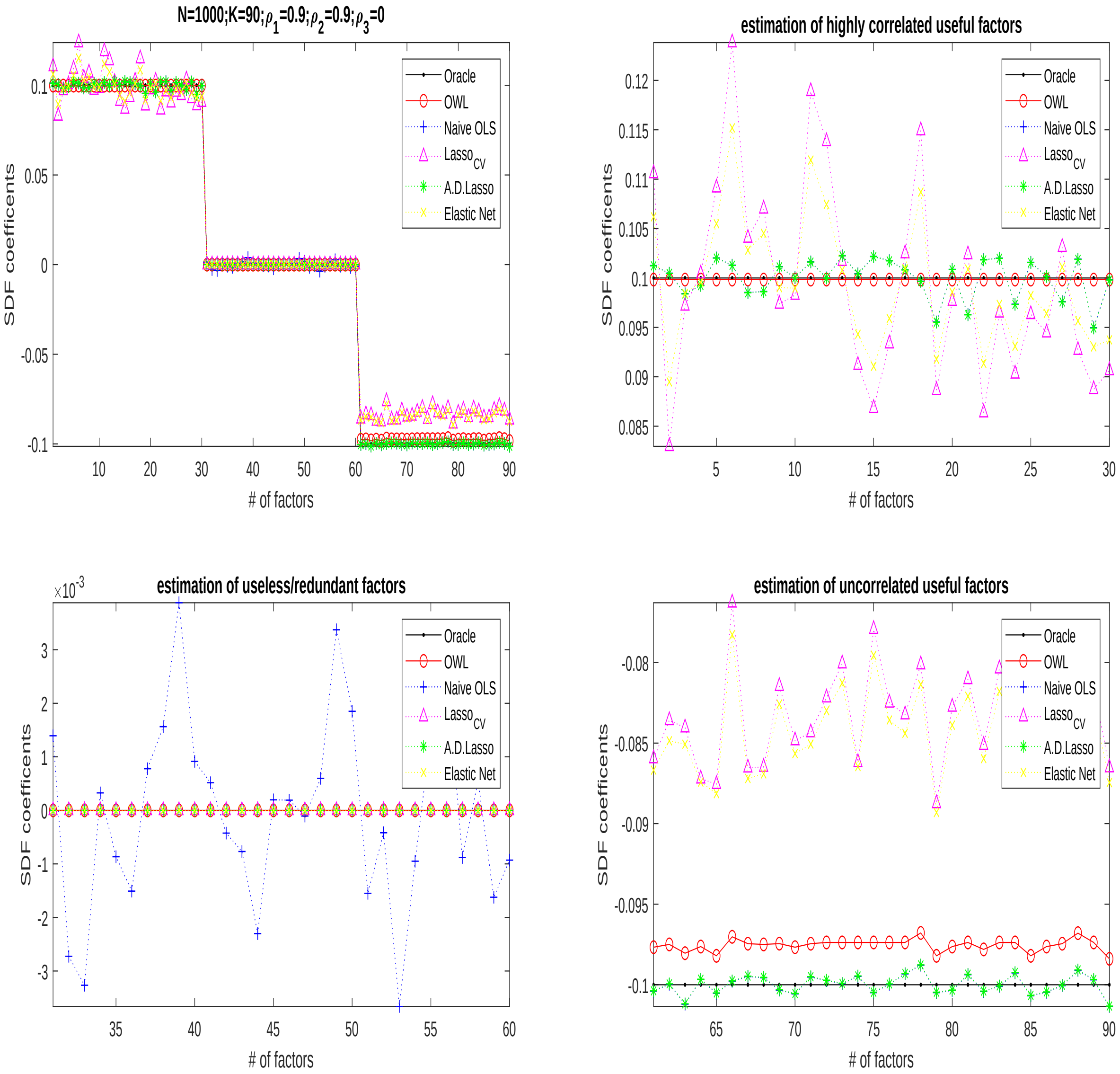


Figure 2. Simulation: $N = 1000, K = 90$

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.

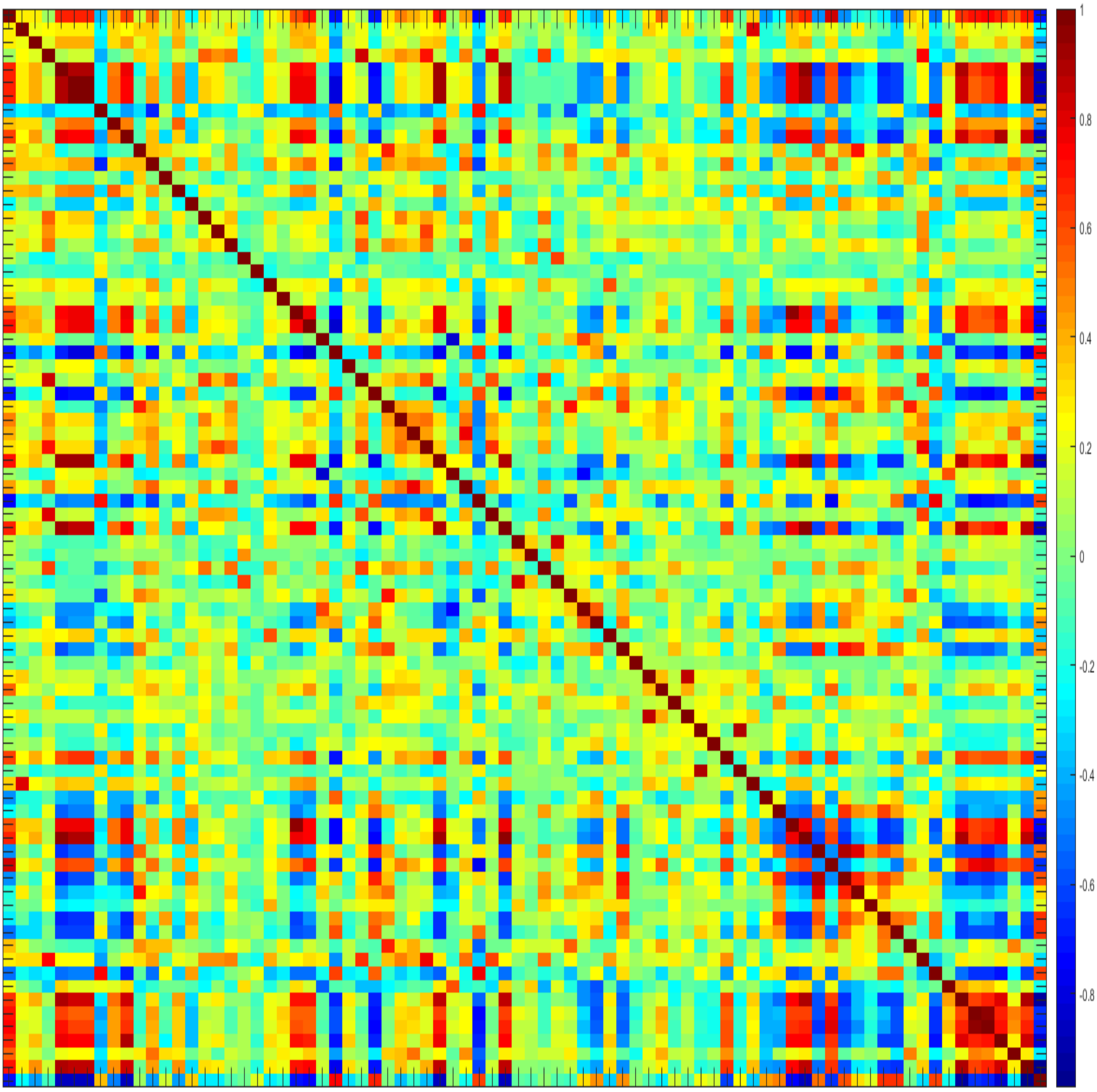


Figure 3. Factor correlation measured by times series

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

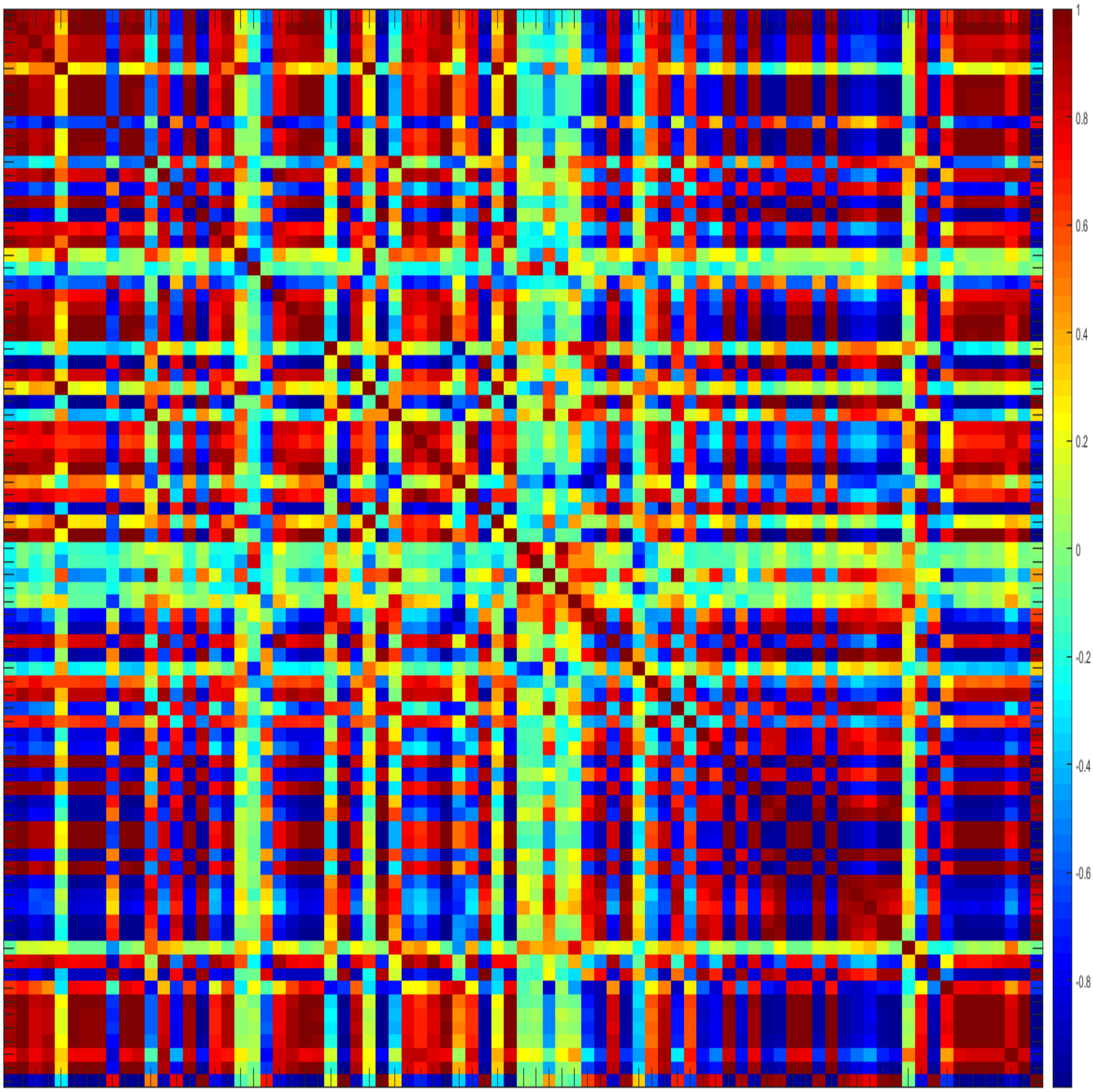


Figure 4. Factor correlation measured by factor loadings

This heat map displays the correlation coefficients measured by factor loadings, all the rest are the same in figure (3).

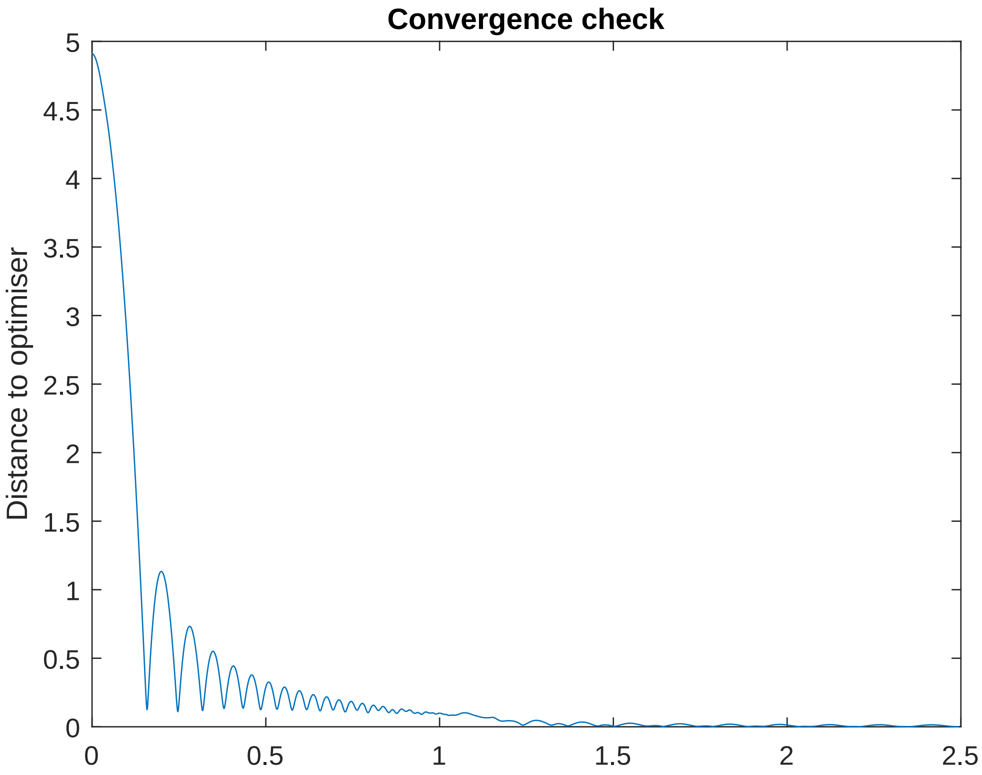


Figure 5. Convergence check

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.

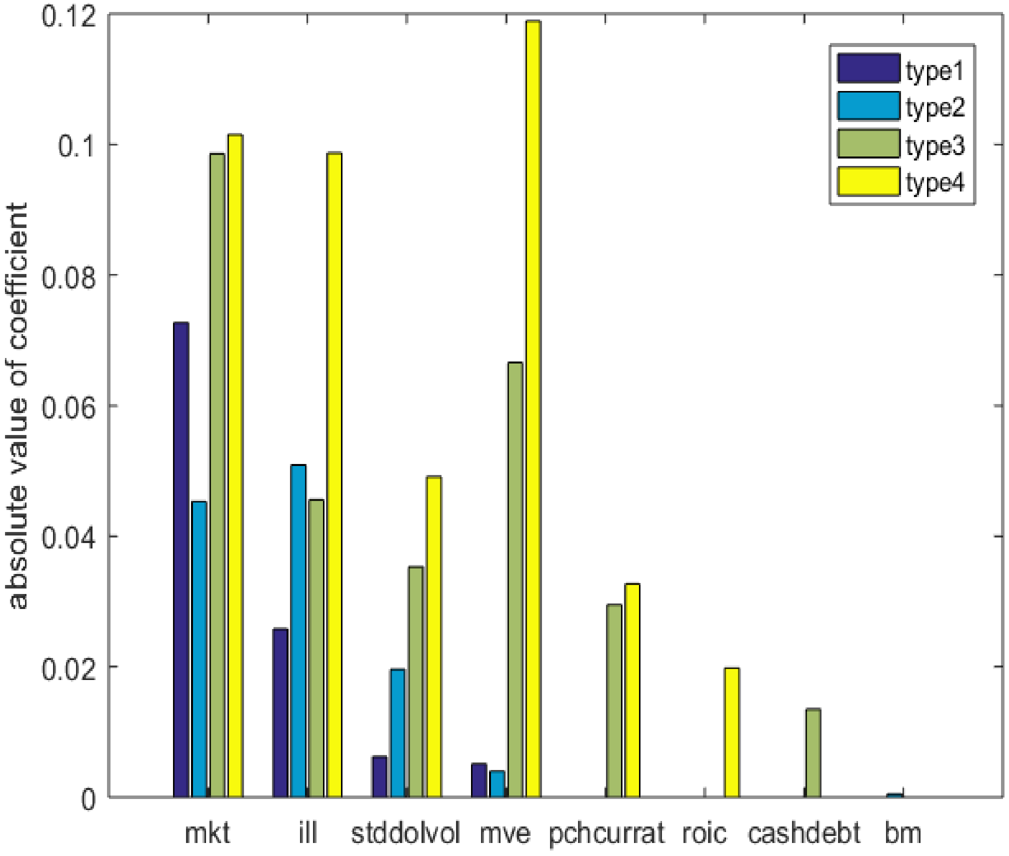
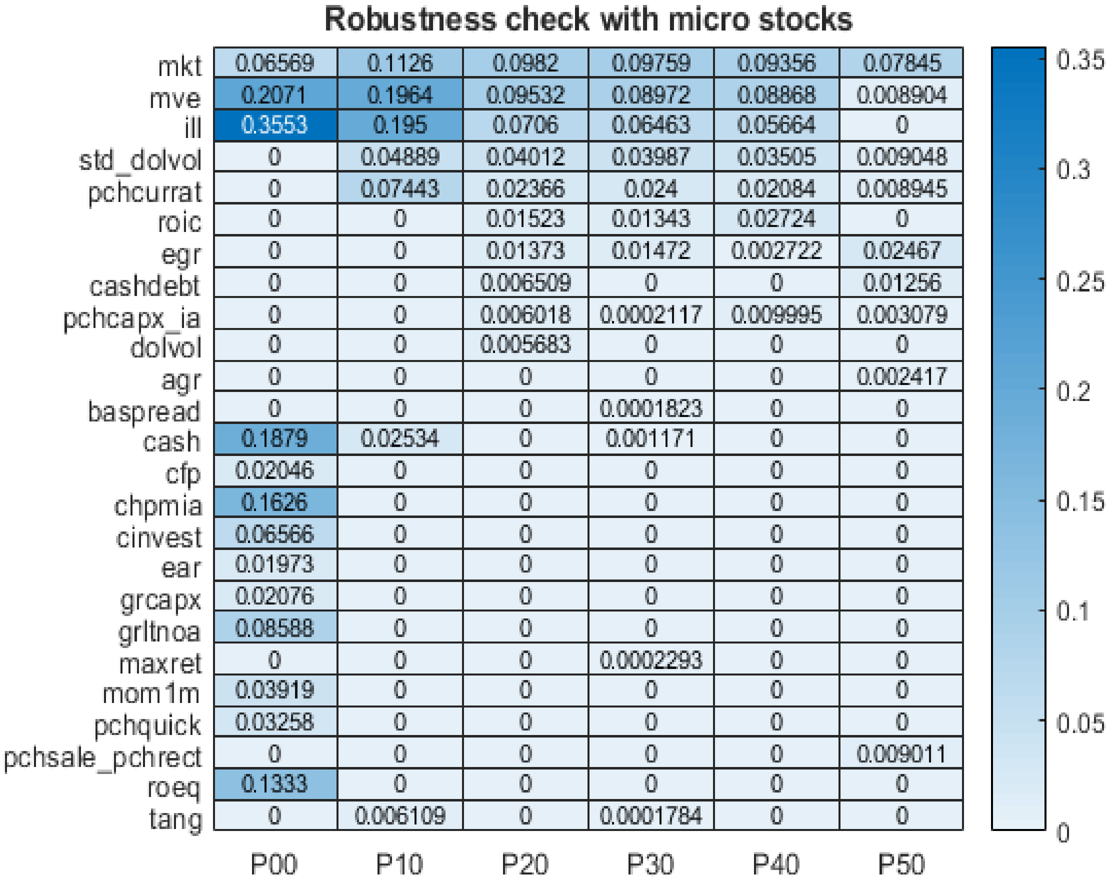


Figure 6. Robustness check using different sorting methods

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.



	P00	P10	P20	P30	P40	P50
mkt	0.06569	0.1126	0.0982	0.09759	0.09356	0.07845
mve	0.2071	0.1964	0.09532	0.08972	0.08868	0.008904
ill	0.3553	0.195	0.0706	0.06463	0.05664	0
std_dolvol	0	0.04889	0.04012	0.03987	0.03505	0.009048
pchcurrat	0	0.07443	0.02366	0.024	0.02084	0.008945
roic	0	0	0.01523	0.01343	0.02724	0
egr	0	0	0.01373	0.01472	0.002722	0.02467
cashdebt	0	0	0.006509	0	0	0.01256
pchcapx_ia	0	0	0.006018	0.0002117	0.009995	0.003079
dolvol	0	0	0.005683	0	0	0
agr	0	0	0	0	0	0.002417
baspread	0	0	0	0.0001823	0	0
cash	0.1879	0.02534	0	0.001171	0	0
cfp	0.02046	0	0	0	0	0
chpmia	0.1626	0	0	0	0	0
cinvest	0.06566	0	0	0	0	0
ear	0.01973	0	0	0	0	0
grcapx	0.02076	0	0	0	0	0
grltnoa	0.08588	0	0	0	0	0
maxret	0	0	0	0.0002293	0	0
mom1m	0.03919	0	0	0	0	0
pchquick	0.03258	0	0	0	0	0
pchsale_pchrect	0	0	0	0	0	0.009011
roeq	0.1333	0	0	0	0	0
tang	0	0.006109	0	0.0001784	0	0

Figure 7. 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).

Table 1. Anomaly factors and their acronyms

Acronym	Firm Characteristics	Acronym	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. I 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: I remove stocks that is smaller than 20 (30 and 40) percentile of NYSE listed stocks. Within each estimation I 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. Out-of-sample Sharpe ratio of OWL and alternative factor selection strategies

This table reports the out-of-sample (OOS) Sharpe ratio of portfolios using a rolling window scheme. Rolling window size is of 120 months, at the end of estimation window, I regress each test asset (bi-variate sorted portfolios) on factors selected by the two-stage procedure, but one period back. Suppose at time t, I regress each test asset return from $t-120-1$ to t on selected factors from $t-120-2$ to $t-1$, and obtain estimated β. I then forecast each test asset's next period return (at $t+1$) by multiply estimated β and selected factors at t. I then sort test assets by their predicted returns into decile portfolios, I long the top decile and short the bottom decile, at next period $(t+1)$ when returns are realised, I can compute the OOS portfolio returns and its Sharpe ratio. Panel A reports the full sample estimation. Panel B and Panel C reports two sub-sample estimations. Factor selection strategies include OWL, LASSO, Elastic Net (EN), and two-pass Fama-MacBeth regression (FM).

Panel A

Sample		1980:01 - 2017:12	
OOS period		1990:01 - 2017:12	
Stocks	All stocks	>20 prctile	>40 prctile
OWL	3.1340	1.2086	0.8757
LASSO	2.4654	1.0943	0.8253
EN	2.5055	1.0943	0.8253
FM	2.4742	1.0448	0.7826

Panel B

Sample		1980:01 - 2000:12	
OOS period		1990:01 - 2000:12	
Stocks	All stocks	>20 prctile	>40 prctile
OWL	3.7603	1.9714	1.8294
LASSO	3.4189	1.9006	1.1144
EN	3.4189	1.9006	1.1144
FM	2.9419	1.7426	0.9200

Panel C

Sample		2001:01 - 2017:12	
OOS period		2011:01 - 2017:12	
Stocks	All stocks	> 20 prctile	> 40 prctile
OWL	3.5763	2.2309	2.3701
LASSO	3.2726	1.9877	2.0819
EN	3.2315	1.9877	1.7862
FM	3.0998	2.2030	2.0604