Regularising the Factor Zoo with OWL

Chuanping Sun *

Preliminary draft. Please do not cite or circulate.

Abstract

Hundreds of anomaly variables have been proposed, claiming explanatory power to the cross-section of average returns in equity market. Cochrane (2011) dubs this phenomenon the "factor zoo" and further argues that the characteristic related factors to explain the average returns are in disarray. This paper introduces a newly developed machine learning tool, ordered and weighted L1 norm regularisation (OWL) to "regularise" this chaotic "factor zoo". The innovation of OWL is that high correlations among explanatory variables are permitted. Highly correlated variables will be identified simultaneously and grouped together. This is important because factor correlation prevails in high dimensionality and biases standard estimators (Fama-MacBeth regression, LASSO, etc.) yet it has not been discussed extensively in the literature.

Empirical evidence suggests that 'liquidity' related factors play an important role in explaining the cross-section of average returns. Further robustness check shows that OWL selected factors have superior performance in a broad range of criteria. Out-of-sample Sharpe ratio of hedge portfolio, formed using OWL selected factors as predictors in the past two decades, is around 3.5 (annualised) considering all stocks, and above 2.2 when excluding micro stocks.

JEL classification: C33 C49 C52 C55 C58 G12

Keywords: Factor Correlation, Asset Pricing, Machine Learning, LASSO, 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 to have explanatory power to the cross section of average returns. Harvey et al. (2015) documented 316 anomaly variables and raised the concern that many new factors are as a result of data snooping. Hou et al. (2017) replicate 447 anomaly variables and find 64% or 85% of which are insignificant depending on the choices of cut-off level. Mclean and Pontiff (2016) find many anomalies vanish once they are discovered and published. Cochrane (2011) dubs this phenomenon the "factor zoo" and he further argues that the characteristics related factors to explain the cross section of average returns are in disarray. He emphasises the importance of finding (useful) factors that can provide independent information about average returns, and distinguishing (redundant) factors that can be summarised by others or (useless) factors that have no explanatory powers to the cross section of average returns. Fama and French (2008) survey that portfolio sorting and Fama-MacBeth regression are traditionally employed to measure and test for factor's ability to explain the cross section of average return. 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, ordered and weighted L1 norm regularisation (OWL) to regularise the "factor zoo" which, to my best knowledge, is the first time applied in Finance. OWL is an extension of the LASSO family, studied by Zeng and Figueiredo (2015), Figueiredo and Nowak (2016) and applied in electronic engineering, for instance, to de-noise a satellite picture. The innovation of OWL, compared to other regularisers, is that it permits high correlation among explanatory variables. Correlation is important because it 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 characteristic and disregard the rest. DeMiguel et al. (2017) state that correlation between factors matters in a portfolio perspective and find six factors selected through their procedure are correlated. Asness et al. (2013) also find the negative correlation between 'momentum' and 'value' factors, and achieve superior portfolio performance by exploring this correlation.

Factor correlations are common, especially in high dimensional big data (see section 4 for a detailed illustration). Cochrane (2011) shows that in term of determining 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 in an SDF setting, or the regression of average returns on factor loadings (the sensitivities of factors with test assets) in a Fama-MacBeth framework. Kleibergen (2009) cautions that the estimation of risk premium that results from a Fama MacBeth two-pass regression is sensitive to collinearity of factor loadings. Factor correlation structure measured by factor loadings is usually much higher than that measured by its times series in the first-pass regression.

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 the question above. I will show analytically that when two useful factors are highly correlated, OWL estimator will assign similar coefficients to them. This statistical property allows one to identify highly correlated factors and ranking their contributions to explaining the cross section of average returns. Meanwhile, OWL can simultaneously shrink off useless and redundant factors.

Like other shrinkage based estimators, it is, however, challenging to make direct statistical inferences on OWL estimator. Following DeMiguel et al. (2017), Feng et al. (2017), I adopt a two stage (select and test) procedure to find factors that statistically contribute to explaining the cross section of average returns. 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. In order to deal with high correlation among factors after the first stage, I modify bootstrap design to bypass multicollinearity issues: instead of bootstrapping the distribution of true parameters, I bootstrap the null hypothesis that all factors have zero explanatory power. 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 rather than step-wisely, 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 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) argues 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, where some candidate factors are highly correlated and have the same true coefficients, I find OWL estimators can successfully group correlated factors together, while benchmark estimators like LASSO and adaptive LASSO assign random coefficients to highly correlated factors and give a very noisy estimation of true parameters. In terms of shrinking off useless and redundant factors while factors are correlated, OWL performs well while LASSO fails to set all useless factors' coefficients to zeros. In a different setting of the Monte Carlo experiment where the number of candidate factors are close to the number of assets available ($K \approx N$) , OWL can successfully group together highly correlated factors by assigning similar coefficients to them, while LASSO and adaptive LASSO assign random and noisy coefficients to factors: the noise (distance between the highest and lowest estimator) doubles the oracle value. This Monte Carlo experiment shows that OWL has superior performance to other estimators when high correlation is present among factors.

To illustrate, I give an example of applying OWL on some well known data sets. I choose 12 popular and prominent factors from well known models including Fama and

French (2015) 5 factors, Hou et al. (2014) q4 factors, 'momentum', 'quality-minus-junk', 'bet-against-beta', and 'hmldevil' from AQR's on-line data library. I consider these factors, because they are well established in the literature, and have available data libraries that have been intensively used and verified by many researchers. I also consider 7 sets of Fama French bi-variate sorted 25 portfolios plus the 49 industrial portfolios as test portfolios. I find, first of all, market factor is shown as the primary factor to explain the cross section of average returns in 8 out of 10 test portfolios I considered. This finding is consistent with Harvey and Liu (2017) in which they employ a statistical method on individual stocks and find market factor is a dominating factor. Second, when using bi-variate sorted 25 portfolios as test portfolios, very often the same characteristics used to forge anomaly factors are selected. This has been criticised in various papers, see Harvey et al. (2015)), Lewellen et al. (2010) and Ecker (2013). However, if we pool together all 25 portfolios (Feng et al. (2017)) as a single set of test portfolios, OWL estimators are less biased towards a small set of characteristics.

For completeness, I initially consider 100 firm characteristics documented in Green et al. (2017), using data set from CRSP and Compustat, 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), while deleting characteristics that have insufficient data to form sorted decile portfolios [1]. I obtain 80 anomaly factors which are spread returns between the top and bottom decile portfolios [2]. For test portfolios, I 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. Considering the possible combination of any two of 80 characteristics is large, I single out 'size' as a common characteristic to form bi-variate sorted portfolios

[1] 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 is insufficient stocks to form the decile portfolios, I delete the characteristic.

[2] Note that the sorting is always from high to low according to characteristics, and the anomaly variables are top decile return minus the bottom decile return. That will end up with some slight difference with some familiar notations. For instance, the famous size factor 'small-minus-big' in my factor library would be 'big-minus-samll', 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.

with the remaining ones (also see Feng et al. (2017)). I deem 'size' is an important factor in explaining the cross section of average stock returns since 'size' is included in many models like Fama and French (2015) 5-factor, Carhart (1997) 4-factor and Hou et al. (2014) q4-factor models. Asness et al. (2018) also shows 'size' matters when controlling other characteristics.

For a robustness check, I also consider different methods of sorting, including univariate sorting and all combinations of bi-variate 2 by 2 sorting, finding the OWL estimation is consistent in picking anomaly factors.

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

First, within 80 anomaly factors, I find strong correlation among some factors, measured by their time series. For instance, some beta related anomalies are highly correlated with other anomalies, including accruals, profitability, volatility and liquidities. 15% of the correlation coefficient matrix of all anomaly factors, measured by their time series, exhibit correlation coefficient higher than 0.5 (absolute value) and raises to 68% when factor correlation is measured by their factor loadings (important for second stage of Fama-MacBeth regression). This casts doubts about 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 to explain the cross section of 1927 bi-variate sorted portfolios using combinations of 'size' and other firm characteristics. This finding confirms the empirical evidence by Harvey and Liu (2017), and is consistent when I use either the valued weighted or equal weighted method, excluding micro stocks. However, when micro stocks are included, the importance of market factor drops several ranks. This is due to the fact that micro stocks, although only taking up less than 10% of market value, it constitutes 56% of all stocks in the database. It rings alarms about methodologies that may be biased by micro/small stocks, for instance, using individual stocks as test portfolios (even excluding micro stocks) or including micro stocks in portfolio sorting.

Third, liquidity related factors are the main contributors to the variation of cross sectional average returns. 'illiquidity' (Amihud (2002)) has the highest (absolute) coefficients among anomaly factors using value weighted method excluding micro stocks, followed by 'standard deviation of traded dollar volume' (Chordia et al. (2001)), 'percentage change of current ratio', consisting of the top 3 important anomaly factors. Then it is followed by 'return on invested capital', 'equity growth rate', 'percentage change of capital expenditure' and 'cash', which are related to 'profitability', 'investment' and 'growth'. This finding coincides with Hou et al. (2018) in which they update their q4-factor model into a q5-factor model by adding an asset growth factor.

Fourth, as the two stage procedure identifies a set of important anomaly factors, it is of interest to compare it with some well established factors in the literature. I first consider a wide range of in-sample comparison criteria including Sharpe ratio, Hansen Jagannathan distance, cross sectional R^2, and GRS statistics (see Gibbons et al. (1989)). First, I compare the first 3 most important OWL-selected-factors (including 'market'), call it "owl3" to compare it with the well known Fama and French (1992) 3-factor model, and similarly, "owl4" with Carhart (1997) 4-factor and Hou et al. (2014) q4-factor, "owl5" with Fama and French (2015) 5-factor, and "owl6" with Fama and French (2018) 6-factor (plus a momentum factor). I find that OWL selected factors dominate benchmarks in terms of performance score, achieving Sharpe ratios that are 3 to 5 times benchmark factors; HJ distance is also the smallest; cross sectional R^2 is about 20% to 40% higher; and GRS statistics are 40% to 50% lower.

Fifth, 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 of 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 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. On the contrary, 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) explode the investment perspectives and propose the q-theory model which includes an investment factor, a profitability factor, 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 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), before testing for significance. They bootstrap the null hypothesis that factor coefficient is zero by orthogonalising the factor with asset returns. 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 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 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 (L1 norm regularisation) that achieves dimension reduction within a convex optimisation problem. Since then, many modifications and improvement 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 L1 and L2 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 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 L1 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 explores the relation between risk price and risk premium and explains which one should be used to identify 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: first pass obtains the factor loadings by running time-series regressions of each asset; 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, in which case $b_i = 0$ (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:

$$e(b) = E[R_t m_t(b)] = E(R_t)E(m_t(b)) + cov(R_t, m_t(b)) \tag{3}$$

$$= E(R_t)E(m_t(b)) + r_0^{-1} cov(R_t, 1 - b(f - E(f))) \tag{4}$$

$$= r_0^{-1}(\mu_R - Cb) \tag{5}$$

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

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

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. Hence it produces stable estimators regardless of the choice of limited test assets. However, when facing a large universe of test assets (large N), we need a large number of time series observations (T) to estimate the second moment of returns. In practice, the length of T is limited, so the estimation of sample covariance will be near singular matrix. Ludvigson (2013) advocates the second choice of W_T: the identity matrix, if we have large N. Additionally, when test portfolios represent particular economic interests, for instance, firm characteristic sorted portfolios, the identity matrix will be a better choice, because identity matrix does not re-weight test portfolios and each characteristic sorted portfolios will be treated equally. We are interested to see which candidate factors can explain the entire test portfolios containing all anomaly information.

2.2 Challenges of high-dimensionality

Cochrane (2011) points out that traditional methods like portfolio sorting to identify useful factors, has fallen short in the high-dimensional world. Following Fama and French (1992), Fama and French (2008) to construct 5 by 5 portfolios and suppose 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 impossible to sort stocks into so many 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. Secondly, when many factors are included as explanatory variable in a regression, it is likely variables will be correlated. 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). Employing Fama-MacBeth regression in an environment where factors are correlated will be likely to pick up redundant factors which are unpriced but can be subsumed by other priced factors. Apart from that, Kleibergen (2009) cautions Fama-MacBeth regression faces multicollinearity issues in a high dimensional setting where factors are likely correlated.

2.3 Remedy through Sparsity

Empirical finance research has demonstrated strong evidence that many of the proposed factors are actually useless or redundant to explain the cross section of average returns, 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 the 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 L1 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 L1 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 use 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 a small number of principal components of characteristics based factors can approximate the SDF well.

The goal of this paper is to find a small subset of factors, potentially highly correlated, to explain the cross-section of average returns. To interpret that goal, I am interested in selecting factors that deliver good model-fit measures such as Hansen-Jagannathan distance, cross-sectional R^2, GRS statistics, etc, as well as out-of-sample Sharpe ratios. For that, I allow for correlation among factors and introduce the ordered and weighted L1 norm (OWL) regularisation to circumvent the curse of high dimensionality.

2.4 The Ordered and Weighted L1 (OWL) regularisation

In this subsection, I define OWL and explain the algorithm to solve the OWL optimisation problem. 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 from factor 1 to K: $\omega_i = \lambda_1 + (K - i)\lambda_2$, $i = 1, 2, ..., K$. Zeng and Figueiredo (2015), Figueiredo and Nowak (2016) show that by adopting a linearly decreasing weighting scheme, OWL maps to OSCAR (Bondell and Reich (2008)) setting, which has appealing properties to group highly correlated variables.

In order to solve (8), I use the proximal gradient descent algorithm. 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 order 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. [3]

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 $k = 1, 2, 3, ...$ is step of each iteration; $g(b) = \frac{1}{2}(\mu_R - Cb)' W_T(\mu_R - Cb)$ and sz_k is step size at each iteration k.

To achieve the optimal convergence rate, I consider the accelerated proximal gradient method, also regarded as the fast iterative soft-thresholding algorithm (FISTA, see appendix).

First initialise $b^{(0)} = b^{(-1)}$ and $t_0 = t_1 = 1$, repeat:

$$\begin{aligned} t_{k+1} &= (1+\sqrt{1+4t_k^2})/2 \\ u_{k+1} &= b_k + \frac{t_{k-1}}{t_{k+1}}(b_k - b_{k-1}) \\ b^{(k+1)} &= Prox_{\Omega_\omega}(u_k + sz_k \bigtriangledown g(b^{(k-1)})) \end{aligned}$$

until convergence.

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

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 regulariser. 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 statistic properties which distinguish it from other LASSO based shrinkage methods, and make OWL suitable in a highly correlated setting.

Theorem 2.1 (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.

This property implies that when factors are highly correlated, they are more likely to be grouped together by assigning them with the same coefficients. However, in standard estimators (Fama-MacBeth regression, LASSO, etc.) with orthogonal designs, that is assuming factors are uncorrelated, it may yield inconsistent estimators and bias economic interpretations. The grouping theorem safeguards correlated variables from being neglected and distorted. In the meanwhile, there are other elements that affect the grouping property. First, the weighting parameter λ_2 which controls the distance of neighbouring weights. A big gap between weights (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 to intersect with the contour coming from the unregularised quadratic minimisation solution. Second, the mean (μ_R) and standard deviation (σ_R) of test assets. A set of less informative assets (small μ_R and/or small σ_R) would result in factor clusterings: all factors assigned with the same (and small) coefficients. This is because factors, even if they are not highly correlated, are equally inadequate to explain a set of less informative test assets.

The grouping property is a major contribution of this paper and distinguishes it from other common approaches. Other orthogonal-design based estimators, for instance LASSO, will neglect factor correlation and distort factor interpretations. Fama-MacBeth regression will face multi-collinearity issues. For that, factor-trimming is usually required. Deleting factors with high correlations (for instance, see Green et al. (2017)) is a common practice among researchers. However, it is difficult to define a cut-off level of correlation to decide which factors to remove due to high correlation. Furthermore, it is difficult to decide which one to remove if they are highly correlated. Factors can be highly correlated due to common underlying risk.

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

Figueiredo and Nowak (2016) also show that the OWL estimator error is bounded by

$$E||\hat{b} - b^*||_2 = \mathcal{O}(||b^*||_2 \frac{\omega_1}{\bar{\omega}} \sqrt{\frac{SlogK}{N}})$$

where ω_1 is the first element of the weighting vector ω, and $\bar{\omega}$ is the mean of all elements of ω; b^* is the true value of risk price.

With the convergence rate of $\sqrt{\frac{S \log K}{N}}$, it is possible to estimate models with large numbers of factors, which alleviates the high-dimensionality problem greatly. However, OWL, like other shrinkage based estimators, is a biased estimator. The bias is proportional to the true parameter values. Informally, OWL shrinks more of prominent factors and less of weaker factors. This observation is of great importance: weak factors are shrunk less and thus likely to be retained after shrinkage. Following a similar approach by DeMiguel et al. (2017) and Feng et al. (2017), I propose a two-stage selection procedure to select and test factors.

2.7 Two-stage selection procedure

This section explains steps to test for factor significance after OWL selection. Since it is challenging to infer economic testing from OWL estimators, I propose a two-stage procedure to test for factor significance: 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 each of these 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 result: 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 (two sides) of t^*_{stat} distribution, I then declare the associated coefficient is significant.

3 Simulation and example for application

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. After simulation, I illustrate a simple application of OWL on some well known datasets.

3.1 Simulation design

Consider K candidate factors, $2K/3$ of them are useful factors, that is they are priced, i.e. $b \neq 0$, 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

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

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 and naive OLS. Then I compare these estimators with the oracle value of b, pre-specified.

3.2 Simulation result

In the first experiment I simulate an environment where the number of test assets is abundant ($N >> K$). Block 1 and 3 are useful factors (oracle value of the risk prices are non-zeros); block 2 are useless/redundant factors (oracle value of the risk prices are zeros); and $\rho_1 = 0.9; \rho_2 = 0.9; \rho_3 = 0$, so useful factors in block 1 and useless/redundant factors in block 2 are highly correlated and useful factors in block 3 are uncorrelated.

[Figure 1 about here.]

In figure (1), the upper left panel displays the overall plot of risk price estimators using all methods for all factors. The remaining three panels are plots for each of these three blocks.

The upper right panel displays the plot of all estimators of useful factors with high correlation. The bottom left panel displays the plot of all estimators of useless factors with high correlation. The bottom right panel displays the plot of all estimators of useful factors without correlation. In each plot, OWL estimator (red) is displayed along with LASSO, adaptive LASSO, naive OLS estimators and the oracle value of risk prices (black).

It suggests that OWL estimator can successfully group highly correlated variables by assigning the same coefficients to them, while other estimations failed to identify this feature. Furthermore, OWL estimator has the smallest estimation error and is the most efficient estimator compared to others. On the contrary, LASSO and adaptive LASSO neglect factor correlations and yield inefficient estimators. For useless/redundant factors, OWL and adaptive LASSO successfully set all risk price to zero but LASSO failed to shrink off all useless/redundant factors. For uncorrelated useful factors, OWL behaves similarly to LASSO and adaptive LASSO.

In the second experiment, I study the behaviour of OWL where only limited test assets are available. In particular, $N = 25$ and $K = 15$, and all the rest are the same as in the first experiment.

[Figure 2 about here.]

Figure (2) shows that in limited sample size, OWL replicates the performance in the first (larger sample size) experiment well, while LASSO and adaptive LASSO suffer greatly from a smaller sample size: the noise (deviation from the oracle value) of LASSO and adaptive LASSO estimators has increased more than 10 times compared to the first experiment.

These two Monte Carlo simulation experiments show that in a highly correlated environment, LASSO and adaptive LASSO fail to identify any correlations between highly correlated factors. The presence of correlation also makes these estimators inefficient (highly volatile). LASSO is also unfit to shrink off useless/redundant factors when factors are correlated. On the contrary, OWL estimator performs consistently well to identify highly correlated factors and shrink off useless/redundant factors, both in large and small sample sizes.

3.3 OWL application: an example

This section illustrates an example of OWL application on some well established datasets. I choose 12 popular factors, including 'market', 'momentum', 'smb', 'hml', 'rmw', 'cma' (Fama and French (2018)); 'me', 'ia', 'roe' (Hou et al. (2014)), 'qmj', 'bab' and 'hmldevil'. The first six factors are from Kenneth French's on-line data library, and the last three factors are from AQR's website [4]. I use 10 different sets of test portfolios. The first 7 sets of test portfolios are Fama French's bi-variate sorted 25 portfolios using different sorting characteristics; the 8th test portfolio is the 49 industrial portfolio; the 9th test portfolio is the combined 7 sets of 25 portfolios, amounting to 175 portfolios; the last set of test portfolios is the combined 175 portfolios and 49 industrial portfolio. All data are from Kenneth French's on-line data library.

3.3.1 Preliminary statistics

First of all, I present the correlation coefficient matrix of 12 factors, measured by their time series and their factor loadings, respectively.

[Figure 3 about here.]

In figure (3), left panel shows the heat map of correlation coefficients of factors measured by factor time series. High correlation is observed between the well known value factor 'hml' and investment factors ('cma' and 'ia'). 'qmj' is highly correlated with profitability and size factors. 'momentum' is negatively correlated with value factors. In the meanwhile, many other factors exhibit low correlation coefficients.

Right panel shows the heat map of correlation coefficients measured by factor loadings. Cochrane (2011) emphasises that *"we want to check whether the mean function of test assets lines up with the covariance function of factors"*, indicating that the factor correlation measured by its covariance with test assets (factor loadings) really matters, because testing an asset pricing model usually ends up with testing the regression of cross-sectional mean returns on factor loadings. Compared to the left panel, high correlation prevails among factors once correlation is measured by factor loadings: more than half correlation coefficients exceed 0.5 in absolute values.

This finding cautions against the methodologies for testing asset pricing models in a high dimensional setting where correlation is ignored.

3.3.2 Estimation result

[Figure 4 about here.]

[4] I thank Lu Zhang for providing the q4 factor data.

Figure (4) reports the heat map of estimated risk prices by OWL, in which all 12 factors enter as candidate factors in the penalised regression. The horizontal axis lists 10 sets of test portfolios, each column represents the OWL estimation result using a particular set of test portfolios.

Figure (4) suggests that market factor is a dominating factor, with largest magnitude in 8 out of 10 test portfolios and second largest in the rest. This is consistent with Harvey and Liu (2017), in which they employed an orthogonalised bootstrap method to select factors step-wisely. Second, when using Fama French 25 bi-variate sorted portfolios as test assets, OWL often selects the same characteristics that are used to sort test portfolios, a tautology criticised by Harvey et al. (2015). However, the combined 175 bi-variate sorted portfolios (as test assets, see also Feng et al. (2017)) suggests, when including large sets of portfolios, the bias towards some particular characteristics vanishes.

This example shows the limitations of using a small set of test portfolios to choose factors. Following suggestions of Lewellen et al. (2010) and Feng et al. (2017), I will consider a large set of combined portfolios as test assets in empirical analysis.

4 Empirical analysis

This section applies the two stage select-and-test 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 I consider, followed by a detailed account of the construction of anomaly factors and test portfolios. I consider value weighted and equal weighted methods, controlling firm sizes (removing small stocks), to gauge anomaly spread portfolios. 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 [5] to construct anomaly variables and test portfolios, because of their

[5] downloaded from the Wharton Research Data Services

availability and better data quality. The period spans from January 1980 to December 2017, total 456 months on all NYSE and NASDAQ listed stocks.

I consider 100 firm characteristics initially described in Green et al. (2017) [6], 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), Fama and French (2015). Mirco 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 that having clustered missing data and are insufficient to construct decile portfolios at every month will be dropped. Overall, I obtain 80 anomaly factors [7].

There is a debate in the literature about using either individual stocks or sorted portfolios as test assets. Harvey and Liu (2017) use individual stocks with bootstrap method to test for predictability of anomaly factors, and they find 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 in the estimator to mitigate EIV. However, empirical work shows that "Shanken's correction" is minimal in small samples. From another stream, 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. On the other hand, sorted portfolios are 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.

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

[7] see appendix for a detailed description.

However, the biggest drawback of using individual stocks stems 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 also challenging to implement in an out-of-sample framework.

On the other hand, using sorted portfolios 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 will guarantee continuity of portfolio returns, hence no missing data problem.

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.

On the other hand, portfolio sorting can circumvent this issue by using value weighted method, that is, 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 forming test portfolios. Lewellen et al. (2010) argue that the 25 size and value sorted portfolios are too low as 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 acknowledge 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 5 about here.]

[Figure 6 about here.]

Figure (5) displays the heat map of factor correlation coefficients matrix measured by their time series. It suggests that there are 16% factors whose correlation coefficients (absolute value) are 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 in candidate factors because of its high correlation profile with other factors.

Figure (6) displays the heat map of factor correlation coefficients matrix measured by factor loadings, which exhibits much higher correlation compared to figure (5): 64% correlation coefficients (absolute value) are greater than 0.5, implying serious multicollinearity issues if standard Fama-MacBeth regression is employed.

4.3 Which factors matter?

Facing 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.

[Table 1 about here.]

Table (1) 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 during 1980 and 2000. Interestingly, during 1980 and 2000 and 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 average returns of alternative test portfolios formed by different sorting methods.

First, I apply the uni-variate sorting method to sort all non-micro stocks into decile portfolios using each 80 characteristics, 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 anomaly 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 characteristics. 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 use 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.

[Table 2 about here.]

Table (2) reports the two-stage procedure result using four different sets of test assets (including the one used in empirical analysis). Factors selected by OWL are listed in descenting order by their coefficient magnitude. First, 'market' factor is a prominent factor in all estimations. Second, liquidity related factors (i.e. 'illiquidity', 'standard deviation of dollar volume') are consistently chosen as top anomaly factors to explain the cross section of stock returns. Amihud (2002)'s 'illiquidity' tops 'market' factor as the most prominent factor to explain the bi-variate sorted portfolios where all possible bi-variate combinations are considered.

Table (2) shows that singling out 'size' to form bi-variate sorted portfolios will not alter the result that liquidity related factors are primary factors to explain the cross section of average stock returns.

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 Comparison between OWL selected factors and alternatives

In this section, I will investigate the performance of OWL selected factors compared with some popular benchmarks such as Fama and French (2015) 5 factors and Hou et al. (2014) q4 factors.

I consider several criteria for in-sample comparison. I first compare Sharpe ratios of the long-short portfolios given a set of factors throughout the full sample size. Second, I compare the Hansen-Jagannathan distances of competing factors, which measure how far a model deviates from its true one. Third, I consider the cross sectional R^2, which is a conventional measure of model fit. Lastly, I consider the GRS statistic (see Gibbons et al. (1989)) which is based on model alphas and measures the degree of model misspecification. Test statistics are defined as follow:

$$SR = \sqrt{\mu_f' \Omega_f^{-1} \mu_f} \tag{16}$$

$$HJD = \sqrt{(m - m^*)' E(RR')^{-1} (m - m^*)} \tag{17}$$

$$CSR^2_{OLS} = 1 - \frac{RSS}{TSS} \tag{18}$$

$$GRS = \frac{T}{N} \frac{T - N - K}{T - K - 1} \frac{\alpha' \Sigma^{-1} \alpha}{1 + \mu_f' \Omega_f^{-1} \mu_f} \tag{19}$$

where μ_f is the mean of factors, and Ω_f is factor covariance matrix, $m - m^*$ is the deviation from an admissible SDF, $E(RR')$ is the second moment of test asset returns, α is the intercept of a regression of test assets on a set of factors, and Σ is the covariance of α.

[Table 3 about here.]

In table (3), I compare the first 3 factors selected by OWL with Fama and French (1992) 3 factors (FF3); likewise, I then compare the same number of factors selected by OWL with Carhart (1997) 4 factors , Hou et al. (2014) Q-theory factors, Fama and French (2015) 5 factors and Fama and French (2018) 6 factor. For robust results, not only do I consider the dataset constructed following Green et al. (2017), I also compare authors' original dataset downloaded from the authors' websites. [8]

Table (3) shows that OWL selected factors outperform benchmarks in all criteria. Sharpe ratio is between 3 and 5 times that of benchmark models, despite using either the authors' own dataset or the 80-anomaly-factor dataset. Note that performance scores

[8] Note that in order to compute HJ distance and GRS statistics, inverse of covariance matrix is needed, which requires $T > N$ to obtain non-singular inverse matrix. Under this case, I consider a 2 by 2 sorting method rather than a 5 by 5 sorting method in the empirical analysis, which brings down the dimension of N.

of benchmarks between these two datasets are similar, with Hou et al. (2014) Q-theory factors differ more than the Fama French factors. A possible explanation is that these 80 anomaly factors are sorted following Fama and French (1992) and Fama and French (2015), while Hou et al. (2014) employ a different sorting method. However, despite the score difference between two datasets while computing Hou et al. (2014)'s Q-theory factors, they are both underperformed by OWL selected factors. Incremental score by including more factors is minimal after 4 factors. Sharpe ratio, CSR R^2, and GRS statistics reach a plateau when OWL4 is considered. Including more factors increases GRS statistics (degree of model misspecification) and the incremental gain in Sharpe ratio and cross sectional R^2 are also minimal after OWL4. HJ distance is a measure of model misspecification. First, note that HJ distances of benchmarks using the 80-anomaly-factor dataset are smaller than that using the authors' own datasets, indicating the 80-anomaly-factor dataset is a better candidate to describe the test asset returns. Compared to benchmarks, OWL selected factors exhibit smallest HJ distances in all comparisons. Cross-sectional R^2 of OWL selected factors is typically 20% to 40% higher than benchmarks. The incremental score between 'owl3' and 'owl4' is highest, around 35%, while the incremental score between 'owl4' and 'owl5' (and between 'owl5' and 'owl6') is less than 1%. GRS statistic of OWL selected factors is around 40% to 50% smaller than that in benchmarks using either dataset.

Table (3) shows that OWL selected factors have superior in-sample performance to benchmarks. In the next subsection, I will investigate their out-of-sample performance.

4.7 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 [9] ensures an OOS metric by construction, the choice of factors are 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

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

another period.

I follow a similar procedure of 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.

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, [10] 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 4 about here.]

Table 4, 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).

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.

[10] excluding stocks whose size are less than 20 percentile of NYSE listed 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, which are 1.9714 and 1.8294 respectively.

Panel C shows 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.

Prevailing factors may change over time. A shift in economy may drive factors' contribution to explaining the cross section of stock returns varying. 'Profitability' factors drive asset returns in the first sub-period and 'liquidity' factors dominate the second sub-period, 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 faces multicollinearity issues when regressing on factor loadings. Among 80 anomaly factors I considered, I find more than 60% 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 L1 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, small HJ distance and GRS statistics. 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 2.2 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 the result of testing a hypothesis that some particular factors are priced if individual stocks are used. For that, Shanken (1992) proposed the Shanken's correction. However, there are two other major shortcomings while 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 all stocks (all NYSE and NASDAQ listed stocks) in terms of market capitalisation, they constitute 56% of all stocks. For that reason, 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 is 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) has 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 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. (2017) use firm characteristics as factors to form mean variance efficient portfolios, that is, instead of looking at the average returns, they investigate jointly the first two moments of asset returns. 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 L1 norm regularisation is a general tool useful for sparsity selection. It can be extended to explore portfolio selection strategies, where individual stock weights are regularised by OWL. Considering that correlations among stocks can be harnessed to achieve superior portfolio performance (see DeMiguel et al. (2014)), OWL will be particularly useful for forming portfolio selection strategies.

6 Appendix

6.1 A1: Anomaly names

[Table 5 about here.]

6.2 A2: FISTA algorithm

Algorithm 1: FISTA for OWL

```
Input: μ_R, C, ω
Output: b̂ in (8)
Initialisation: b_0 = b̂_OLS, t_0 = t_1 = 1, u_1 = b_0, k = 1, η ∈ (0,1), τ_0 ∈ (0, 1/L) [a]
while some stopping criterion not met do
    τ_k = τ_{k-1};
    b_k = Prox_{Ω_ω}(u_k + τ * C' * (μ_R − Cb))
    while (1/2)||μ_R − Cb_k||_2^2 > Q(b_k, u_k) [b] do
        τ_k = η * τ_k;
        b_k = Prox_{Ω_ω}(u_k + τ * C' * (μ_R − Cb))
    end
    t_{k+1} = (1 + sqrt(1 + 4t_k^2))/2
    u_{k+1} = b_k + (t_{k-1}/t_{k+1})(b_k − b_{k-1})
    k ← k + 1
end
Return: b_{k-1}
```

[a] L is a Lipschitz constant.

[b] $Q(b_k, u_k) = \frac{1}{2}||\mu_R - Cu_k||_2^2 - (b_k - u_k)'C'(\mu_R - Cu_k) + \frac{1}{2\tau_k}||b_k - u_k||_2^2$

6.3 A3: proof of Theorem (2.1)

The proof of theorem (2.1) relies on the Pigou-Dalton-transfer and directional derivative lemma.

Lemma 6.1 (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 6.2. *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*.*

Denote the objection 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 that 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:

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

Apply 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 weights has the same distance, that is $\Delta_\omega = \lambda_2$.

Using Cauchy-Schwarz inequality that is for any vector u and v, and $< u, v >$ 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 $||\mu_R - C\hat{b}||_2 \leq ||\mu_R||_2$. We have:

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

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

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

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:

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

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," .

Carhart, M. M. (1997): "On Persistence in Mutual Fund Performance," .

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

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

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

de Leeuw, J., K. Hornik, and P. Mair (2009): "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.

Ecker, F. (2013): "Asset Pricing Tests Using Random Portfolios," .

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," .

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.*

Gibbons, M. R., S. A. Ross, and J. Shanken (1989): "A Test of the Efficiency of a Given Portfolio," *Econometrica*, 57, 1121–1152.

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," .

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," .

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.

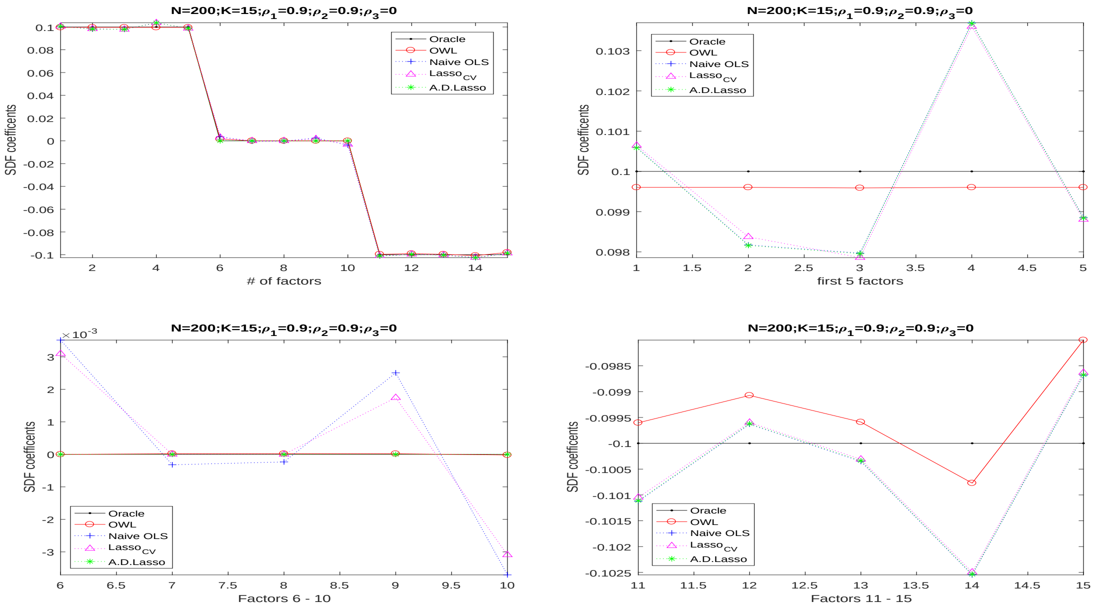

Figure 1. $N = 200, K = 15, \rho_1 = 0.9, \rho_2 = 0.9, \rho_3 = 0$

This figure reports the plot of OWL estimator along with other benchmark estimators. There are 200 test assets, 15 candidate factors, which are divided into 3 equal block, where their correlation within each blocks are $\rho_1 = 0.9, \rho_2 = 0.9, \rho_3 = 0$. The upper left panel displays the plots of risk price estimators using all methods for 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 exhibit high correlation. The bottom left panel displays the plot of all estimators of useless factors that are also highly correlated. The bottom right panel displays the plot of all estimators of useful factors but exhibits no correlation. In each plot, OWL estimator (red) is displayed along with LASSO, adaptive LASSO, native OLS estimators and the oracle value of risk prices (black).

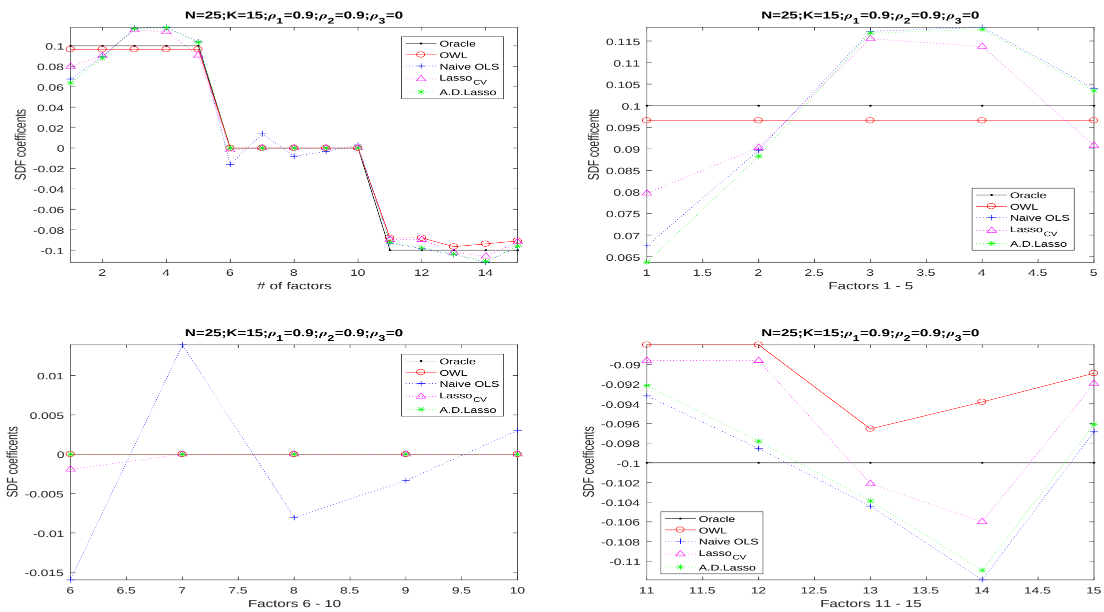

Figure 2. $N = 25, K = 15, \rho_1 = 0.9, \rho_2 = 0.9, \rho_3 = 0$

This figure reports the plot of OWL estimator along with other benchmark estimators. There are 25 test assets, 15 candidate factors, which are divided into 3 equal block, where their correlation within each blocks are $\rho_1 = 0.9, \rho_2 = 0.9, \rho_3 = 0$. The upper left panel displays the plots of risk price estimators using all methods for 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 exhibit high correlation. The bottom left panel displays the plot of all estimators of useless factors that are also highly correlated. The bottom right panel displays the plot of all estimators of useful factors but exhibits no correlation. In each plot, OWL estimator (red) is displayed along with LASSO, adaptive LASSO, native OLS estimators and the oracle value of risk prices (black).

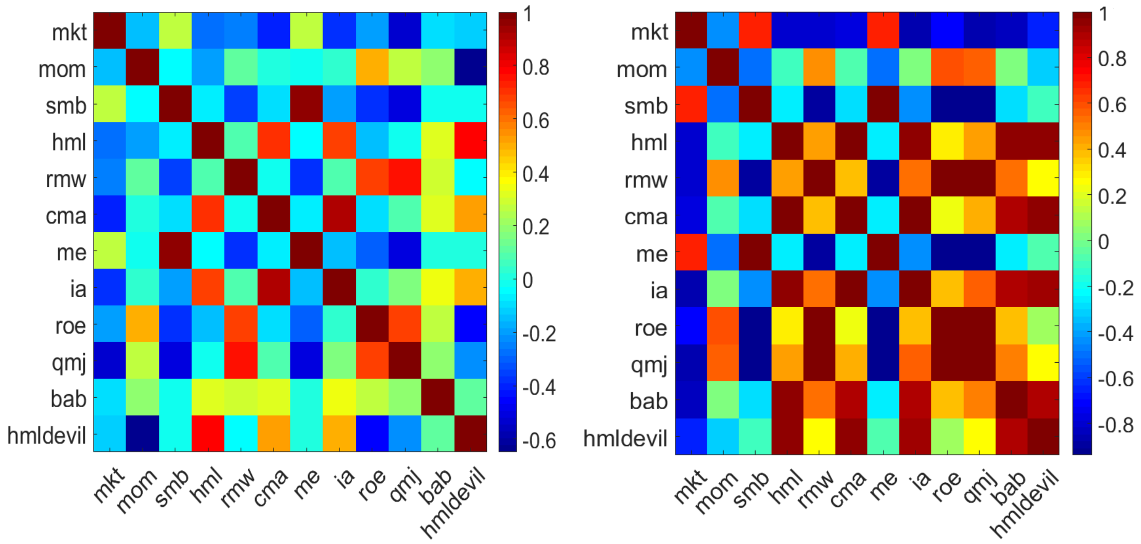

Figure 3. Correlation of factors

This figure reports the correlation coefficient matrix of factors. Left panel shows the heat map of correlation of factors measured by their time series; right panel shows the factor correlation measured by the covariance function with test assets (factor loadings). The test assets are the Fama-French 25 bi-variate sorted portfolios using size and value as sorting characteristics.

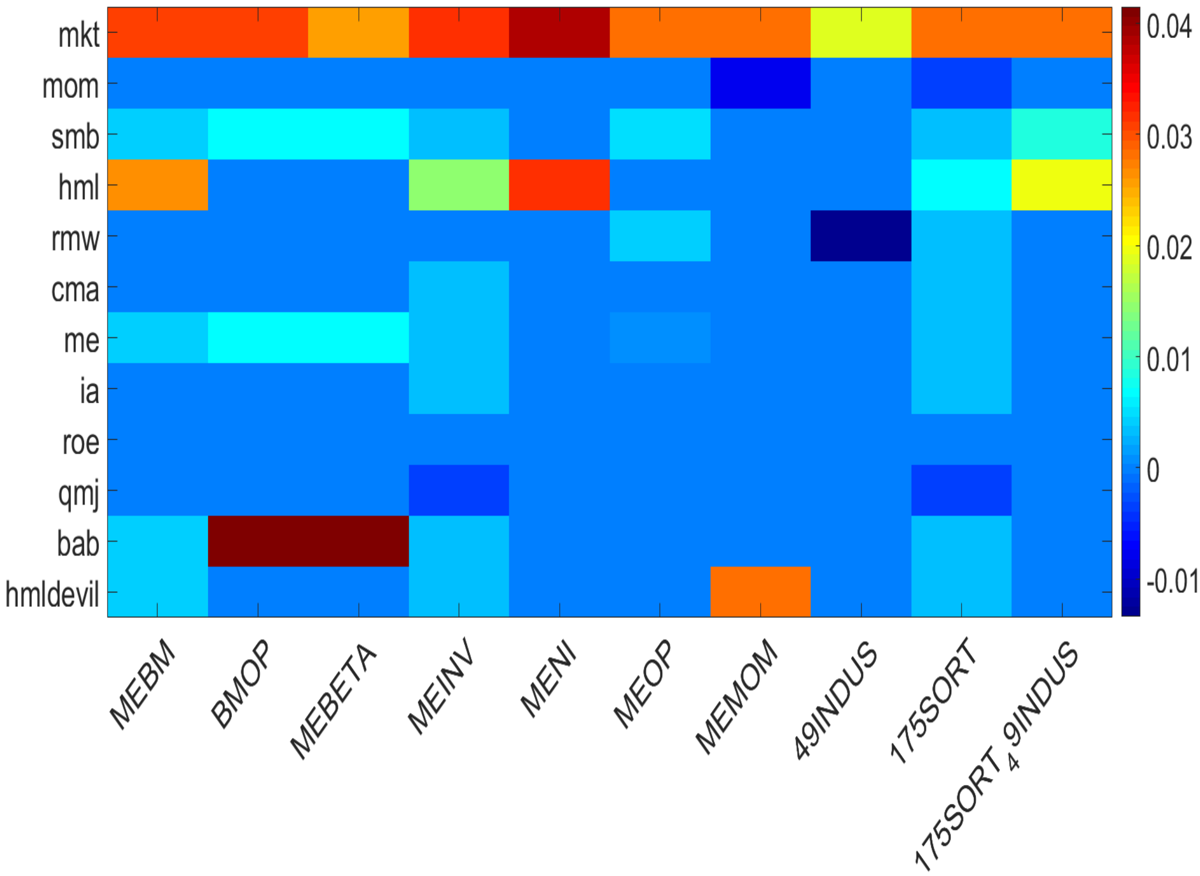

Figure 4. OWL estimation result

This figure reports OWL estimation on 10 sets of test assets, each column represents one set of test portfolios. Each row represents estimates of one particular factor. The first 1-7 sets of test assets are bi-variate sorted 25 portfolios using different sorting characteristics. the 8th test portfolio is the 49 industrial portfolio, the 9th test portfolio is the combined 7 sets of 25 portfolios, amounting 175 portfolios, the last set of test portfolio is the combined 175 portfolios and 49 industrial portfolio. All data are from Kenneth French's on-line data library. The estimates of risk prices are presented in head map where scale of coefficients maps to the scale of colours bar. The light blue exhibits at large area of the matrix are zeros.

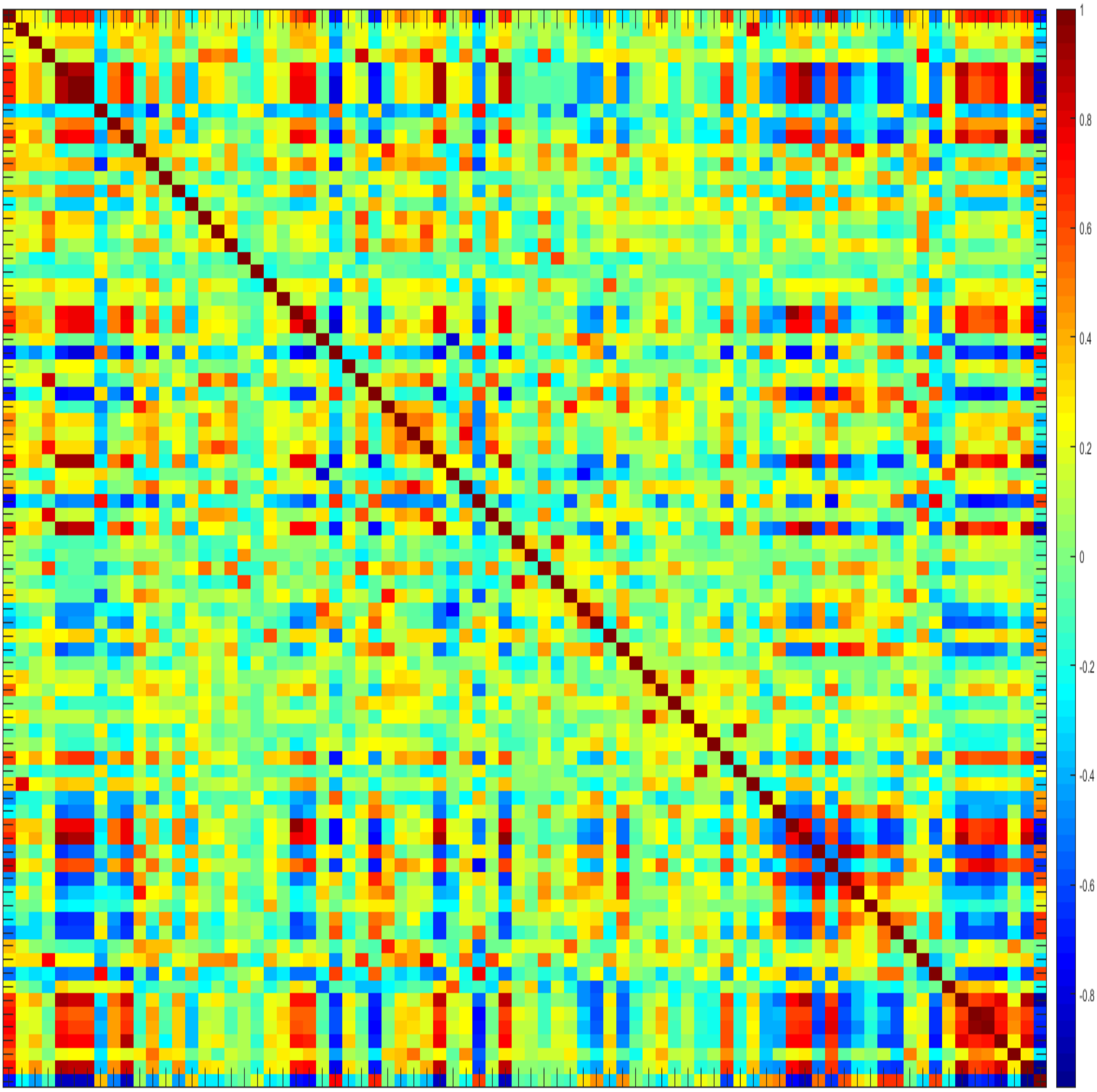

Figure 5. Factor correlation measured by times series

This heat map displays the correlation coefficients of all 80 anomaly factors, measured by factor times series from January 1980 to December 2017. Dark red or deep blue indicates of high correlation (positive or negative), while light colours indicate of low correlation.

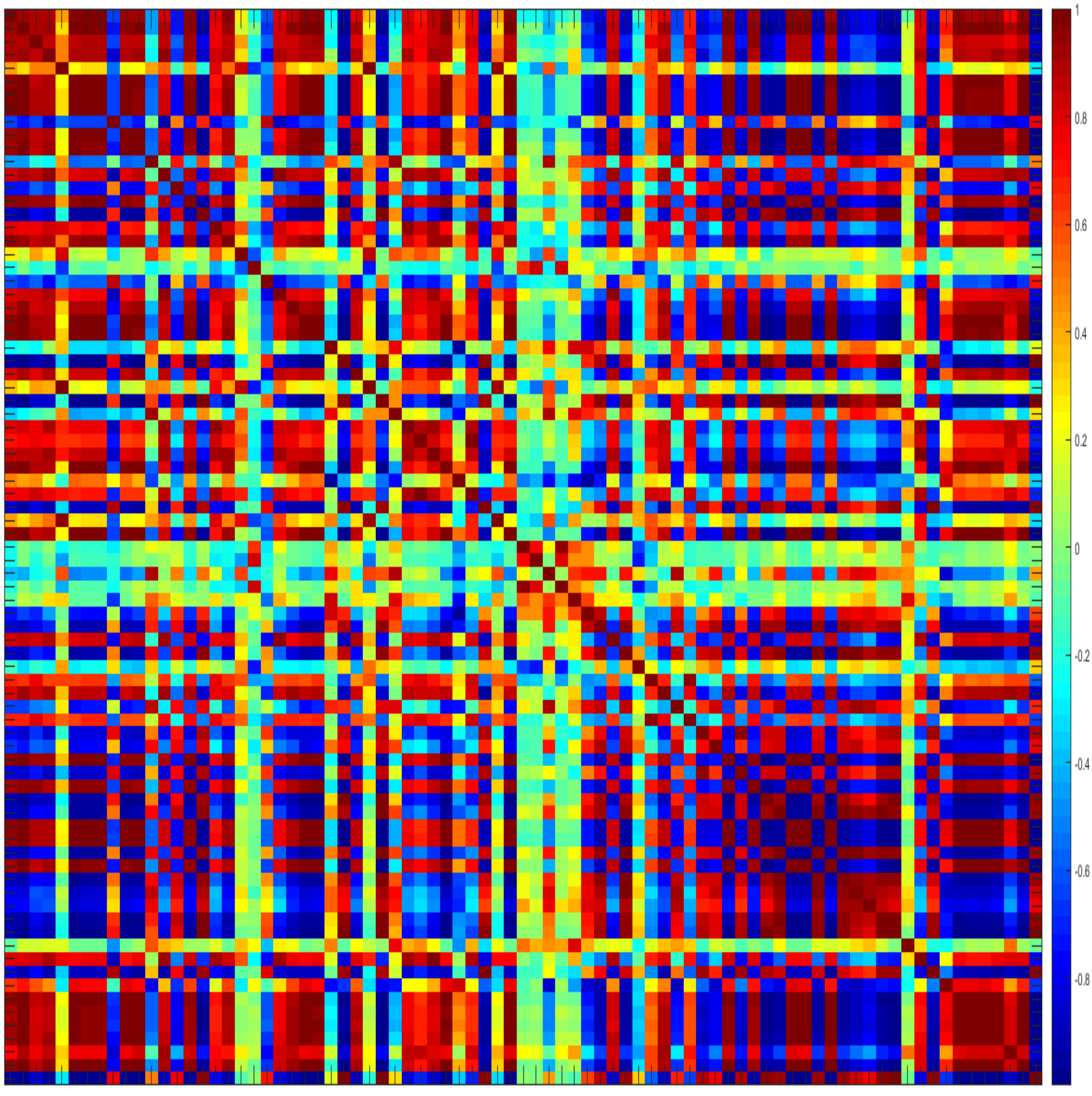

Figure 6. Factor correlation measured by factor loadings

This heat map displays the correlation coefficients of all 80 anomaly factors, measured by factor loadings, which is the covariance function between 80 anomaly factors and 1927 test portfolios. Dark red or deep blue indicates of high correlation (positive or negative), while light colours indicate of low correlation.

Table 1. 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	1	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 2. OWL estimation using different test portfolios

Panel A: Decile portfolios / uni-variate sorting

mkt	dolvol	ill	std_dolvol	mve_ia
0.072628	-0.02676	0.025792	0.0062255	-0.005078

Panel B: 2 by 2 bi-variate sorted portfolios using all combination of 80 characteristics

ill	mkt	dolvol	mve	bm
0.050849	0.045298	-0.01955	-0.004	0.0004368

Panel C: 3 by 3 bi-variate sorted portfolios using "size" and remaining characteristics

mkt	mve	ill	std_dolvol	pchcurrat	cashdebt
0.098621	0.066555	0.045616	0.035292	-0.029545	-0.013478

Panel D: 5 by 5 bi-variate sorted portfolios using "size" and remaining characteristics

mve	mkt	ill	std_dolvol	pchcurrat	roic
0.1189	0.1015	0.0987	0.0491	-0.0327	-0.0198

Table 3. Comparison between OWL selected factors and alternatives

'owln' indicates the first n factors selected by OWL using VW,excluding micro stocks. '*' indicates using the author's own data set. 'SR' and 'CSR^2' are using the 1927 bi-variate sorted portfolios with 5 by 5 sorting. 'HJ' and 'GRS_stat' are using 2 by 2 sorting to enable invertible covariance matrix.

	SR	HJ	CSR^2_{ols}	GRS_stat
owl3	1.0486	21.6959	0.3984	5.2038
FF3	0.1819	22.2079	0.1116	11.0327
FF3*	0.2137	23.1159	0.1569	9.0409
owl4	1.0616	21.6958	0.5384	5.1939
Carhart	0.1903	22.2036	0.1317	10.9881
Carhart*	0.2881	23.0676	0.1606	8.6685
Q-theory	0.1905	22.1539	0.3302	11.1928
Q-theory*	0.4211	22.9701	0.1762	7.9677
owl5	1.0616	21.5298	0.542	5.2924
FF5	0.2038	22.1535	0.3379	11.0664
FF5*	0.352	23.0554	0.242	8.5287
owl6	1.0665	21.4658	0.5431	5.6632
FF5+mom	0.2117	22.1469	0.3398	11.0304
FF5+mom*	0.3767	23.0261	0.2421	8.3609

Table 4. Out-of-sample Sharpe ratio of OWL selected factors

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.

Panel A

Sample	1980:01 - 2017:12		
OOS period	1990:01 - 2017:12		
Stocks	All stocks	>20 prctile	>40 prctile
Annualised OOS SR	3.1340	1.2086	0.8757

Panel B

Sample	1980:01 - 2000:12		
OOS period	1990:01 - 2000:12		
Stocks	All stocks	>20 prctile	>40 prctile
Annualised OOS SR	3.7603	1.9714	1.8294

Panel C

Sample	2001:01 - 2017:12		
OOS period	2011:01 - 2017:12		
Stocks	All stocks	> 20 prctile	> 40 prctile
Annualised OOS SR	3.5763	2.2309	2.3701

Table 5. 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