Machine Learning in Finance: Estimation, Inference and Financial Applications for Correlated Data

Chuanping Sun

Submitted in partial fulfillment of the requirements of
the Degree of Doctor of Philosophy (PhD) in Economics

School of Economics and Finance
Queen Mary University of London
United Kingdom

September 2020

Statement of originality

I, Chuanping Sun, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third partys copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Details of collaboration

Author contributions and additional collaborators are listed below for each chapter.

Introduction: solo work by Chuanping Sun.

Chapter 1: solo work by Chuanping Sun.

Chapter 2: solo work by Chuanping Sun.

Chapter 3: joint work of Chuanping Sun and Kazuhiro Hiraki.

Signature: **Chuanping Sun**

Acknowledgment

First and foremost, I would like to express my sincere gratitude to Professor Mario Figueiredo, who offered me unreserved help and support when I felt most uncertain and dispirited. Without his generosity, I would not have had the strength to overcome the challenges of PhD study and accomplish my research projects.

I am also grateful to my supervisors Professor Liudas Giraitis and Professor George Kapetanios for their support and guidance, which helped me immensely to improve my research projects and polish this thesis. I would like to thank Professor Emmanuel Guerre, who offered me much feedback and many good suggestions during internal seminars at Queen Mary University of London. I would like to thank Professor Marcelo Fernandes and Professor Thomas Sargent who kindly invited me to visit them at FGV, Sao Paulo and New York University, respectively. These visiting positions greatly enriched my experience of conducting international academic dialogue and learning additional research skills from various departments.

I have benefited hugely from attending workshops and conferences such as the Frontiers of Factor Investing conference 2018 at Lancaster University, the RES 2019 meeting at Warwick University, the Econometric Institution international PhD conference 2019 at Erasmus University Rotterdam, the Econometric Society Asian Meeting 2019 at Xiamen University, the EFMA 2019 meeting at Azores University, the Econometric Society European Meeting 2019 at Manchester University, the SoFiE 2019 meeting at Fudan University, the SITE 2019 workshop at Stanford University, the 2nd FinTech conference at NHH, and the EEA 2020 virtual meeting. I would like to thank the organizers and participants for offering me valuable feedback on my work, as well as nurturing my social skills and boosting my confidence in my research projects.

I would also like to extend my gratitude to my colleagues and friends I have made in various conferences and workshops. Especially, I would like to thank Kazuhiro for his precious friendship and fruitful academic collaborations.

Last but not least, I am very grateful to my mother, Fangying, my father, Yuzhong, my dearest friends and family, Richard and Guy, for their unreserved support and encouragement. They give me hope and make me believe in myself. Without them, I could not have accomplished this piece of work. I would like to dedicate this thesis to Richard and Guy and to my parents.

Abstract

Modern technologies have generated big data at an unprecedented scale and speed, which has spurred remarkable progress in high-dimensional statistical research and offers alternative solutions to some prominent financial research questions facing the "curse of dimensionality". This thesis endeavors to utilize some newly developed statistical methods to address the "curse of dimensionality" in financial research, while providing new perspectives on the economic and financial implications. For instance, Chapter One of this thesis addresses the "factor zoo enigma" while taking account of high correlations observed between factors. I introduce a newly developed machine learning method to dissect this chaotic factor zoo: the OWL estimator, which is not only efficient in dimension reduction but also robust with correlated variables. Chapter Two extends the econometric theory of the OWL estimator I derived in Chapter One, and mainly concerns the underlying statistical properties of the OWL estimator under less restrictive conditions. Furthermore, I utilize the nodewise LASSO technique to identify and quantify the bias in the OWL estimator and I propose the de-biased OWL estimator before deriving its asymptotic normality property. Chapter Three employs the OWL shrinkage method in the portfolio optimization problems, to exploit contemporaneous relations between stocks. I also develop a flexible algorithm which can incorporate bespoke constraints on portfolio weights should investors have any prior information on individual stocks.

This thesis covers a broad range of research areas spanning between empirical asset pricing and econometric inferences. It contributes to the literature concerning high-dimensional statistics, with an emphasis on the LASSO-type estimators, while taking account of correlated variables. It also contributes to the empirical asset pricing literature: this thesis sheds light on new perspectives of the "factor zoo enigma", where the importance of factor correlations is highlighted. It also enriches the literature

pertaining to portfolio optimization problems. The OWL shrinkage method offers an extension to the existing LASSO shrinkage method while further exploiting stocks' contemporaneous relations.

Contents

List of Figures

List of Tables

Introduction

Modern technologies have generated big data at an unprecedented scale and speed, which has spurred remarkable progress in high-dimensional statistical research and offers alternative solutions to some prominent financial research questions facing the "curse of dimensionality". Like many other disciplines, empirical asset pricing literature has seen an explosion of growth in factors claiming to have the explanatory power to the cross section of the expected asset returns. Hundreds of anomaly factors have been documented and cross-tested. However, the majority of them face intense scrutiny for p-hacking and data-mining. Hence, the quest to find an appropriate method to identify useful factors that drive the cross section of asset prices in high dimensional datasets is much needed. Traditional methods such as portfolio sorting and Fama-MacBeth regression used to find useful factors in the factor zoo suffer from the "curse of dimensionality". On the other hand, a new estimation method from the machine learning literature, namely the LASSO estimator (Tibshirani, 1996) has become a new trend in financial applications because of its efficiency in dimension reduction. Despite its huge success and popularity in dealing with high-dimensional big data, the LASSO estimator is often criticized for suffering severe complications from correlated data. Therefore, the objective of this thesis is to study high dimensional financial problems while taking account of correlations in big data.

In Chapter One, I attempt to decipher the "factor zoo enigma" where factors are highly correlated. I begin with a linear setup for a stochastic discount factor (SDF) model before deriving the SDF method to estimate risk prices for factors. I show that the SDF method and the Fama-Macbeth regression are directly related but have different implications on redundant factors, which are defined as factors that contain no pricing information but earn positive risk premiums due to the correlations between factors. Hence, I follow Cochrane (2005) to use the SDF method to infer priced factors. For estimation method, I introduce a newly developed machine learning tool, the Ordered-Weighted-LASSO (OWL) estimator, together with the SDF method, to dissect this chaotic factor zoo. The OWL estimator achieves *sparsity shrinkage* and *correlation identification* simultaneously. Specifically, the OWL estimator encompasses the LASSO shrinkage method as a special case and thus enjoys the *sparsity shrinkage* property of the LASSO estimator. On the other hand, the OWL estimator is robust to correlated factors: highly correlated variables will be identified during estimation and will be assigned with similar coefficients by the OWL estimator (grouping). I also study statistical properties of the OWL estimator. First, I derive the estimation error bound and the consistency property of the OWL estimator under a finite number of factors. Second, I move on to derive the convergence rate of the OWL estimator with an infinite number of factors under the Gaussian assumption and other regularity conditions. Third, I show that by introducing a thresholded estimator based on the OWL estimate, it achieves consistency in model selection, i.e. the thresholded estimator can pick the factors with non-zero risk prices as the true ones. Fourth, I derive the grouping condition under which correlated factors will be assigned with the same coefficients. Using simulated data, I show that the OWL estimator outperforms other benchmarks such as LASSO, adaptive LASSO, Elastic Net, and OLS estimators, particularly when factors are highly correlated. For empirical analysis, I use granular data such as the Compustat and CRSP datasets to construct 80 anomaly factors via portfolio sorting. Similarly, I use the bi-variate sorting method to construct thousands of test portfolios, while controlling micro stocks. The empirical results reveal some interesting findings. First, the factor loading of 80 candidate factors exhibits high correlations, which makes the traditional Fama-MacBeth regression method ill-conditioned to estimate risk premiums. Strong correlations in factor loading will not only result in the weak factor identification problem (Kleibergen, 2009) in Fama-MacBeth regression, but also lead to distortions in the interpretation of redundant factors. Second, micro and small stocks have a strong impact on factors' interpretations. Micro stocks, defined as stocks whose market capitalization is smaller than the 20 percentile of all stocks listed in the New York Stock Exchange, comprise less than 10% of total market capitalization but constitute 56% of all common stocks traded on NYSE, NASDAQ and AMEX. This casts doubts on applications using individual stocks as test assets to find factors that drive asset prices in the cross section of the stock market. Third, the OWL estimator suggests that 'liquidity' related factors primarily drive asset prices, and their high correlations are also identified by the OWL estimator. However, sub-sample estimations reveal a possible time-varying trend in factor selections during different periods. We find that 'profitability' and 'momentum' are prominent factors driving asset prices between 1980 and 2000, while 'liquidity' related factors become crucial determinants of the cross section of expected asset returns between 2001 and 2017. An out-of-sample exercise shows substantial gains in Sharpe ratios by comparing sub-samples and the full sample. Therefore, a theoretical extension of the OWL estimator enabling time-varying parameters would be of great interest as a future research subject, though it is not in the scope of this thesis. On the other hand, we also compare the Sharpe ratios of portfolios constructed using factors selected by the OWL, LASSO, Elastic Net and Fama-MacBeth estimators, and find that the factor-hedged portfolios using the OWL estimation method yield 20% to 30% higher Sharpe ratios than other benchmarks.

In Chapter Two, I focus on deriving robust inferences of the OWL estimator under less restrictive assumptions. In Chapter One, I derived the convergence rate of the OWL estimator under the i.i.d. Gaussian assumption. Now, in Chapter Two, I impose a less restrictive mixing condition and allow fatter tails for variables. I first derive the estimation error bound (oracle inequality) for the OWL estimator. Then, by utilizing a Bernstein-type inequality and some exponential inequalities studied by Dendramis et al. (2019) under the mixing condition, I show that the oracle inequality holds with probability tending to one if the number of factors grows to infinity. I also derive the closed-form formula for this probability with a finite number of factors, which reveals that both the correlation structure and tail distribution of random variables influence this probability. For a long while, the LASSO-type estimators have been criticized for being biased and incapable of statistical inference. Recent developments in the nodewise LASSO technique (see Van De Geer et al. (2014) and Kock (2016) for example) make this task possible. Chapter Two also introduces the de-biased OWL estimator by implementing a nodewise LASSO technique to remove biases from the OWL estimator. I first give a detailed account of identifying and quantifying the bias of the OWL estimator. Then I derive the asymptotic normality property under mixing conditions for the de-biased OWL estimator, which enables inference and testing. Using simulated data, I find that the de-biased OWL estimator can greatly decrease the estimation error in the OWL estimate, while including the true values in the 95% confidence interval with satisfying rates. Empirically, I use 80 factors constructed in Chapter One to predict returns for 15 large stocks (with no missing data) from the Dow Jones Industrial Average index. Note that this empirical analysis is different from Chapter One, where I investigated factors that contribute to the cross-sectional asset prices, whereas here, I implement a lagged time-series regression to find strong predictors for each stock's return. I find that 'sales' related factors are strong predictors for many stocks. However, the results vary substantially between different stocks.

In Chapter Three (joint work with Kazuhiro Hiraki), we focus on the portfolio optimization problem. We start from a general mean-variance efficient portfolio optimization framework and point out the challenges and remedies that have been proposed in the literature to improve its empirical performance. This paper extends the norm shrinkage method of DeMiguel et al. (2009a) and a VAR(1) model implemented in DeMiguel et al. (2014) to catch serial correlations between stocks. We propose a novel norm shrinkage (the OWL shrinkage) method which explicitly exploits contemporaneous correlations between stocks. First, we derive the grouping conditions (i.e. stocks will be assigned with similar portfolio weights) in relation to stocks' contemporaneous correlations. Second, we devise an ADMM (Alternating Direction Method of Multiplier) algorithm and tailor it to solve the optimization problem with the OWL shrinkage method. This algorithm is flexible to incorporate bespoke bounds constraints on portfolio weights should investors have prior beliefs about individual stocks. Empirically, we apply the OWL shrinkage method with various constraints on five different asset classes, including the Fama-French 25 portfolios, S&P 500 and S&P 100 stocks with daily and monthly returns, and randomly selected 100 and 500 stocks from the CRSP dataset with daily and monthly returns. We find strong evidence that the OWL shrinkage method yields very similar portfolio weights to (but not the same as) those of the 1/N portfolio strategy (DeMiguel et al., 2009b) due to the grouping property, but outperforms the 1/N portfolio strategy in terms of both the Sharpe ratio and turnovers. We also find that the OWL-based portfolio strategies work better in asset classes where assets exhibit higher correlations (e.g. the S&P 100 stocks with monthly returns rather than the randomly selected stocks from CRSP dataset with daily returns).

Next, I will briefly discuss how this thesis relates and contributes to a few strands of literature. First, this thesis contributes to the rich asset pricing literature devoted to identifying factors that drive cross-sectional asset prices. Hundreds of factors and anomalies have been claimed to capture and explain the cross section of expected stock returns, see Fama and French (1992, 2015), Carhart (1997), Hou et al. (2014) for some examples of the most celebrated factors in the asset pricing literature. Harvey et al. (2015) document 316 factors that have been published since the CAPM of Sharpe (1964) and Lintner (1965) in the 1960s, and find that the majority of them face intense scrutiny for p-hacking and data-mining. Thus, they suggest raising the bar for testing pricing factors using the t-statistic at the cutoff value of 3 instead of 1.96. Hou et al. (2018b) replicate 447 anomaly factors and find 64% of them are not replicable using the traditional t-statistic, and that rises to 85% if using the cutoff t-value of 3. Traditionally, Fama-MacBeth (FM) regression (Fama and MacBeth, 1973) is employed to find factors with significant risk premiums. However, Kleibergen (2009) shows that the FM regression suffers from multicolinearity problems and the standard statistical inference is distorted under correlated factors. From a different perspective, Cochrane (2005) shows that correlated factors will lead to FM regression incapable of removing redundant factors, which contain no pricing information but are correlated to priced factors, and thus earn positive risk premiums. Gospodinov et al. (2014) develop a model misspecification robust test to tackle useless factors, using a step-wise test to remove useless factors one by one. Kelly et al. (2019) propose the instrumented principal component analysis by introducing observable characteristics as instruments for unobservable dynamic loadings. Fama and French (2018) use Sharpe ratio and employ the Right-Hand-Side method of Barillas and Shanken (2018) to “choose factors”. Harvey and Liu (2017) suggest a step-wise bootstrap method to test for factors’ ability to explain stock 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, these methods mainly concern a low-dimensional setting, where the number of candidate factors is (substantially) less than the number of observations, and they pay little attention to the correlations between factors. In this thesis, I introduce a newly developed machine learning method, the Ordered-Weighted-LASSO estimator, which is tailored to deal with high dimensional problems (the number of factors can be larger than observations, if needed) with correlated factors. To the best of my knowledge, this thesis is the first attempt to dissect the factor zoo while taking account of the correlations between factors.

Chapter Two of this thesis is closely related to the high-dimensional econometric and machine learning literature. The LASSO estimator (Tibshirani, 1996) has long been celebrated for achieving dimension reduction within a convex optimization problem. Many adaptations and improvements have been made to achieve various targets. The literature about the LASSO family evolves rapidly. To name a few, Yuan and Lin (2006) allow the LASSO estimator to shrink variables as groups by introducing the group LASSO estimator. Zou (2006) introduces the adaptive LASSO by adding a consistent estimate as the adaptive weight of the LASSO penalty, making the adaptive LASSO estimator enjoy the oracle property. Zou and Hastie (2005) combine the LASSO and Ridge shrinkage and propose the Elastic Net estimator, which stabilizes factor selection among correlated variables. Yet it was only recently that the bias in the LASSO-type estimators was addressed. Belloni et al. (2014) and Van De Geer et al. (2014) propose the double-LASSO estimator and the de-sparsified LASSO estimator which can identify and correct the bias in the LASSO estimate. In Chapter Two, I follow the ideas of Van De Geer et al. (2014) to implement the nodewise LASSO technique to identify and correct the bias in the OWL estimator, before deriving the asymptotic normality property for the bias-corrected OWL estimator. Meanwhile, researchers have been making strenuous efforts to better understand the asymptotic properties of the LASSO estimator and have made remarkable progress while assuming the i.i.d. process for random variables. For instance, Kock (2016) studies the LASSO estimator for panel data and derives the oracle inequality for the LASSO estimator under the i.i.d. sub-Gaussian assumption. In Chapter Two, I relax the i.i.d. assumption and replace it with a less restrictive α-mixing condition. In addition, I also relax the sub-Gaussian assumption on tail distributions of random variables. Instead, I leave a parameter q, which controls the fatness of tail distributions, in the formula concerning the asymptotic properties of the OWL estimator. I utilize some exponential inequality results from Dendramis et al. (2019) to derive the oracle inequality (i.e. the upper bound of the estimation error of the OWL estimator) and show that the oracle inequality holds with probability tending to one if the number of factors grows to infinity. Meanwhile I provide a closed-form solution for this probability when the number of factors are finite. Therefore, Chapter Two contributes to the high-dimensional statistical literature, where the theoretical results of LASSO-type estimators rely on less restrictive assumptions.

Chapter Three of this thesis contributes to the voluminous literature pertaining to portfolio optimization problems. The mean-variance efficient portfolio theory put forward by Markowitz (1952), despite its theoretical elegance, performs poorly in empirical applications, due to the difficulties of precisely estimating two important ingredients in the portfolio optimization problem: the expected returns and covariances. Empirical applications usually use the sample analogs of these two ingredients in practice. Michaud (1989) looks into the "Markowitz optimization enigma" and finds that the mean variance optimization is in fact "error maximization": the estimation errors in the sample analogs are so large that they erode all the gains from optimization. Subsequently, many researchers have attempted to mitigate the estimation errors in those sample estimates of the expected returns and covariance matrix. Ledoit and Wolf (2003) propose a shrinkage based estimation method for the covariance matrix that shrinks the sample covariance matrix to a target matrix (for instance the identity matrix), and they find substantial gains in out-of-sample Sharpe ratio of the minimum variance portfolio. Jagannathan and Ma (2003) suggest a simple no-short-sale constraint on all stocks and find significant improvement in out-of-sample Sharpe ratio for the minimum variance portfolio. They argue that the no-short-sale constraint can effectively prevent large upward biases in the sample covariance matrix. DeMiguel et al. (2009a) consider the LASSO shrinkage method on portfolio weights for the minimum variance portfolio and achieve competitive Sharpe ratio compared to other benchmarks. DeMiguel et al. (2014) implement a VAR(1) model capturing serial correlations between stocks and find substantial gains in Sharpe ratio. DeMiguel et al. (2009b) compare the naive 1/N portfolio strategy with 14 other optimization-based portfolio strategies and find superior performance in the naive diversification portfolio strategy. The novel portfolio optimization method in Chapter Three of this thesis extends the work of DeMiguel et al. (2009a) and DeMiguel et al. (2014), while the empirical results in Chapter Three relate it to DeMiguel et al. (2009b). Specifically, the OWL shrinkage method encompasses and accounts for the sparse selection property of the LASSO shrinkage method in DeMiguel et al. (2009a). It also explicitly exploits contemporaneous relations between stocks, which is a nice extension of the VAR(1) method in DeMiguel et al. (2014), which however is a reduced model and leaves the contemporaneous correlations between stocks unexplained. On the other hand, our empirical results reveal that the OWL-based portfolio strategies yield very similar portfolio weights to the 1/N portfolio in DeMiguel et al. (2009b) due to the grouping property, but outperform the 1/N portfolio strategy in both Sharpe ratio and turnovers (transaction cost). So our OWL shrinkage method for portfolio optimization problems complements this strand of literature and offers an alternative interpretation for the naive 1/N portfolio strategies.

Chapter 1

Dissecting the Factor Zoo: A Correlation-Robust Approach

1.1 Introduction

Hundreds of anomaly variables have been proposed in the past few decades, claiming explanatory power to the cross section of average returns. Yet many of them are found spurious and not replicable, see Harvey et al. (2015), Mclean and Pontiff (2016) and Hou et al. (2018b) for a detailed discussion. Cochrane (2011) dubs this phenomenon the "factor zoo" and further argues that using characteristics related factors to explain the cross section of average returns is in disarray. He emphasizes the importance of finding factors that can *provide independent information about average returns*, and of distinguishing factors that can be summarized by others. Fama and French (2008) survey empirical methods for dissecting anomalies and point out that portfolio sorting and Fama-MacBeth regression (Fama and MacBeth, 1973) are traditionally employed to find useful factors that drive asset prices. However, in the zoo of factors, portfolio sorting will encounter the *curse of dimensionality*, while Fama-MacBeth regression will suffer from multicollinearity.[1] Kleibergen (2009) cautions that the estimation of risk premium that results from a Fama-MacBeth regression is sensitive to collinearity of factor loadings. In the most recent development, a new strand of literature using machine learning techniques to solve high dimensional fi-

[1] In particular, for the second stage Fama-MacBeth regression, factor correlations measured by factor loadings are usually much higher than those measured by their time series (see Section 1.4 for a detailed illustration).

nancial problems attracts great attention. In particular, using the LASSO estimator (Tibshirani, 1996) to choose factors becomes the new mainstream in finance literature. However, it is well known that the LASSO estimator performs poorly when covariates are *correlated*. Yet the mere fact that correlation prevails in the factor zoo brings in severe complications: Kozak et al. (2020) and Figueiredo and Nowak (2016) show that, with correlated factors, LASSO tends to yield unstable estimate and wrongly shrink off some useful factors. *Correlation* in high dimensionality deepens the "factor zoo enigma", so Cochrane (2011) points out: *"How to address these questions in the zoo of new variables, I suspect we will have to use different methods."*

This paper introduces a newly developed machine learning tool, *the Ordered-Weighted-LASSO* (OWL), to dissect this chaotic factor zoo. OWL *permits* correlation among explanatory variables, which distinguishes it from standard machine learning tools like LASSO. Factor correlations are common in high dimensional big data and they are of great importance in financial implications. For instance, Asness et al. (2013) find a negative correlation between 'momentum' and 'value' factors, which leads to superior portfolio performance. Cochrane (2005) also points out that factor correlations jeopardize the implications of using risk premiums to infer priced factors. Cochrane (2011) shows that to determine which factors are useful in explaining the cross section of average returns, we need to check *whether expected returns line up with the covariances of returns with factors*. In other words, it is the covariance measured by factor loadings, which is typically highly correlated, that really matters to infer priced factors. Hence, in the quest to find useful factors to explain the cross section of average returns, factor correlations play an important role and should not be neglected.

The main empirical question of this paper is, in the high dimensional and potentially highly correlated factor zoo, how to select useful factors and disentangle correlations between factors? OWL provides a unified solution to this question. We first show that the OWL estimator is consistent with finite factors. Then, we allow the number of factors to diverge and derive the convergence rate of the OWL estimator before we devise a thresholded estimator that is consistent in model selection. We also derive conditions under which correlated factors will be grouped together. This allows for factor-correlation identification and sparsity shrinkage, simultaneously.

In a Monte Carlo experiment, we consider 90 candidate factors ($K = 90$) with correlations taken into account. We compare OWL with LASSO, Elastic Net, adaptive LASSO, and OLS estimators. We do this experiment in three settings: one with the number of test assets marginally larger than the number of factors ($N = 100$); one with a large number of test assets ($N = 1000$, $N \gg K$) which represents a low-dimensional setting; and finally, one with a small number of assets ($N = 70$, $N < K$) which represents a high-dimensional setting. In general, OWL is the best performer, especially when factors are correlated. Adaptive LASSO performs well in the low-dimensional setting, but performs the worst in the high-dimensional setting: its performance depends heavily on a consistent estimator as an adaptive weight. LASSO, on the other hand, typically performs worst, especially when factors are correlated. LASSO estimator is severely affected by factor correlations, producing very unstable estimation and wrongly shrinking some useful factors to zeros, which is also pointed out by Kozak et al. (2020) and Figueiredo and Nowak (2016). Although Elastic Net does improve on the performance of LASSO when factors are correlated, stabilizing factor selections and reducing estimation errors, it is still substantially outperformed by OWL. This experiment shows that in the high-dimensional factor zoo where factors are correlated OWL is the best candidate.

Empirically, we initially consider 100 firm characteristics documented in Green et al. (2017), using CRSP and Compustat datasets, from January 1980 to December 2017. We first construct anomaly factors of each characteristic according to Fama and French (1992, 2015).[2] We obtain 80 anomaly factors. For test portfolios, we follow suggestions of Cochrane (2011), Lewellen et al. (2010) and Feng et al. (2020) by forming bi-variate sorted portfolios, and then combine them together as the grand set of test portfolios.[3]

The empirical results complement and challenge some common stances in asset pricing literature. First, we find moderate correlation among 80 anomaly factors, measured by their time series. Some beta related anomalies are highly correlated with

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

[3] For robustness check (included in the online appendix), we also consider other methods of constructing test portfolios while controlling for micro stocks, and we find that OWL is consistent in picking useful factors when a reasonable number of micro stocks are removed.

other anomalies, including accruals, profitability, volatility and liquidities.[4] 15% of the correlation coefficients are higher than 0.5 (absolute value). However, that rises to 68% when factor correlations are measured by their factor loadings. So Kleibergen (2009) raises concerns about the multicollinearity issue for the Fama-MacBeth estimator. Furthermore, from a different perspective, using Fama-MacBeth regression to test for factor risk premiums when factors are correlated is ill-positioned: it is inadequate to remove *redundant* factors, which contain no pricing information but earn positive risk premiums (see Section 1.2.1 for a detailed illustration). Cochrane (2011) emphasizes the importance of finding factors that can provide independent information about average returns and of distinguishing from factors that can be summarized by others (i.e., redundant factors). These alarmingly high correlations among factors echo his concerns: in the zoo of variables, we should consider new methods.

Second, treatment of micro stocks is crucial for empirical interpretation. OWL identifies 'market' as the primary factor for the cross section of asset returns. This finding confirms the empirical evidence by Harvey and Liu (2017). However, when micro stocks are included, the importance of the market factor plummets. Micro stocks, although only taking up less than 10% of market capitalization, constitute 56% of all stocks in the database. That rings alarms about methodologies using individual stocks as test assets: they may bias results because of the abundance of small stocks and their inferiority in aggregated market capitalization. Hence, we adopt and advocate the use of sorted and pooled portfolios as the grand set of test portfolios as in Feng et al. (2020) while controlling micro stocks. Sorted portfolios can efficiently avoid: 1) the "error in variable" bias; 2) missing data problems from individual stocks; 3) problems caused by inferior stocks that little represent the market but dominate the estimation result.

Third, *liquidity* related factors are the main drivers of the variation of cross sectional average returns. 'Illiquidity' (Amihud, 2002) is the most important anomaly factor, followed by 'standard deviation of traded dollar volume' (Chordia et al., 2001). Their high correlation is identified by OWL. In addition, some *'asset growth rate'*, *'profitability'* and *'investment'* related factors are also significant to explain the cross section of average returns. This finding is consistent with Hou et al. (2018a): they

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

add *'asset growth rate'* in their *q4* factor model and propose the *q5* factor model. Interestingly, the 'size effect' disappears during the 1980-2000 period, which is well documented (Amihud, 2002; van Dijk, 2011; Asness et al., 2018). However, it becomes evident again after removing more small stocks (smaller than 40 percentile of the NYSE listed), implying that the vanishing size effect is likely to be caused by some small "junk" stocks. Once "junk" stocks are removed, the size effect resurfaces, which echoes the discovery by Asness et al. (2018): *size matters, if you control your junk.*

Fourth, from an out-of-sample (OOS) perspective, we follow a similar procedure to Freyberger et al. (2020) to conduct the OOS exercise to compare hedged portfolios using factors selected by either the OWL, LASSO, Elastic Net or Fama-MacBeth estimator. We find that the hedged portfolio using OWL selected factors produces 20% to 30% higher out-of-sample Sharpe ratios than other methods. Meanwhile, subsample estimations reveal that *liquidity* related factors are particularly evident after 2000, while before that (1980 - 2000) 'profitability' and 'momentum' are the most important factors to drive asset prices, indicating a shift in economic characteristics. Furthermore, the Sharpe ratios rise substantially while the skewness and kurtosis of portfolio returns are reduced greatly in sub-samples compared to the full sample estimation. This trend urges us to caution about a possible time-varying nature in prominent factors that drive asset prices.

Related literature

This paper naturally builds on a series of papers devoted to identifying pricing factors. Fama and French (1992) propose the three-factor model, consisting of a market return factor, a size and a value factor, that achieves enormous success. Carhart (1997) adds the momentum factor in Fama-French's three factor model that makes it the new standard among practitioners. Hou et al. (2014) explore the investment perspectives and propose the *q4* model which includes an investment factor, a profitability factor, and a size factor along with the market factor. Fama and French (2015) develop their own version of investment and profitability factors and expand the three-factor model to a five-factor model. Fama and French (2018) argue that an extra 'momentum' factor increases Sharpe ratio according to a new test method proposed by 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 the result of data-snooping. Hou et al. (2018b) try to replicate 447 anomaly factors, and find 64% to 85% of them are not replicable.

This paper also relates to a series of econometric papers devoted to asset pricing model testing. Fama and MacBeth (1973) put forward the two-pass regression method that has now become a standard practice in finance. Green et al. (2017) use Fama-MacBeth regression to find significant factors for the US stock market. Lewellen (2015) studies the cross sectional properties of return forecasts derived from the Fama-MacBeth regression and finds that forecasts vary substantially across stocks and have strong predictive power for actual returns. Kan and Zhang (1999) caution that the presence of useless factors bias test results, leading to a lower than normal threshold to accept priced factors. Gospodinov et al. (2014) develop a model misspecification robust test to tackle spurious factors, using a step-wise test to remove useless factors one by one. Kelly et al. (2019) propose the instrumented PCA (IPCA) analysis by introducing observable characteristics that instrument for unobservable dynamic loadings. Fama and French (2018) use Sharpe ratio and employ the Right-Hand-Side method of Barillas and Shanken (2018) to *"choose factors"*. Harvey and Liu (2017) suggest a step-wise bootstrap method to test for factors.[5] Pukthuanthong et al. (2018) propose a protocol to select factors: all factors should be correlated with principal components of test assets covariance matrix. However, our paper differs from other approaches by allowing correlations among factors, which is little discussed in the literature. The OWL estimator achieves sparsity selection and correlation identification simultaneously.

This paper also contributes to the rapidly growing literature using machine learning techniques to solve financial problems. Tibshirani (1996) proposes the LASSO estimator which achieves dimension reduction within a convex optimization problem. Since then, many adaptations and improvements have been made to achieve vari-

[5] In Harvey and Liu (2017), at each step they pick a factor that has the best statistics (for instance, the t-stat), then bootstrap the null hypothesis that factor has no explanatory power by orthogonalizing asset returns with the factor.

ous targets. The literature about the LASSO family evolves rapidly. Yuan and Lin (2006) allow LASSO to shrink variables as groups by introducing the group LASSO. Freyberger et al. (2020) 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 using factor loadings from the first pass Fama-MacBeth regression as the adaptive weight to estimate risk premiums in the second pass regression. Feng et al. (2020) adopt the double selection LASSO of Belloni et al. (2014) to "tame" the factor zoo. Fan and Li (2001) propose the smoothly clipped absolute deviation (SCAD) estimator so that it bridges hard-thresholding and soft-thresholding. Ando and Bai (2015) employ SCAD to find Chinese stock predictors. Zou and Hastie (2005) combine the ℓ_1 and ℓ_2 norm and propose the elastic net (EN), which stabilizes factor selection among correlated variables. Kozak et al. (2020) employ EN in a Bayesian framework and find that sparse principle components can largely explain the cross section of the average returns. Gu et al. (2020) compare popular machine learning techniques used in empirical asset pricing literature and demonstrate large economic gains using regressing trees and neuron networks.

Bondell and Reich (2008) propose the octagonal shrinkage and clustering algorithm for regression (OSCAR) that achieves clustered selection when variables are highly correlated. Zeng and Figueiredo (2015) reveal the close connection between OWL and OSCAR: by adopting a linear weighting scheme for ω, OWL encompasses the OSCAR regularization. Bogdan et al. (2015) study the sorted ℓ_1 penalized estimator (SLOPE) which is closely related to OWL. In fact, their design, before we define the weighting vector ω, is exactly the same. However, the weighting vector for SLOPE is non-linear and they assume that all variables are not correlated before implementing the false discovery rate (FDR) to select factors. OWL differs from SLOPE in the sense that it permits correlations among variables and a linear ω maps OWL to OSCAR, which enables clustering identification.

1.2 Methodology

To study which factors jointly explain the cross section of average returns, we adopt the SDF method in Cochrane (2005). Section 1.2.1 explores the relationship between risk price and risk premium and explains which one should be used to infer priced factors; Section 1.2.2 points out limitations of traditional methods when facing high-dimensionality and offers a remedy by imposing sparsity; Sections 1.2.3 and 1.2.4 introduce the OWL estimator and discuss its statistical properties.

1.2.1 Risk price or risk premium?

Let m denote the stochastic discount factor (SDF)

$$m = r_0^{-1}(1 - b'(f - \mathrm{E}(f))), \tag{1.1}$$

where r_0 is the zero beta rate which is a constant, f is a $K \times 1$ vector of K factor returns, which can be either traded factors or mimicking portfolio returns of non-traded factors. b is a $K \times 1$ vector of the SDF coefficient, referred to as the *risk price*; a non-zero (zero) entry of b means the corresponding factor is (not) priced and b' is the transpose of vector b.

We want to draw inferences on the risk prices of factors. Finding useful factors is the goal of this paper, that is factors with non-zero risk prices and that directly drive the variation of SDF and contain pricing information. More specifically, they reflect the marginal utility of factors to explain the cross-section of average returns. Factors can also be useless or redundant. Useless factors are those whose risk prices are zero and which are uncorrelated with other useful factors. Redundant factors also have zero risk prices but they are correlated with some useful factors. In other words, they can be summarized by other useful factors. Risk premium refers to the free parameter in the second pass Fama-MacBeth regression: the first pass obtains the factor loadings by running time-series regressions of each asset; the second pass runs cross-sectional regressions of asset returns on factor loadings. Cochrane (2005) shows that risk price and risk premium are directly related through the covariance matrix of factors

$$\lambda = \mathrm{E}(ff')b, \tag{1.2}$$

where b is a vector of risk prices and λ is a vector of risk premiums. However, they differ substantially in their interpretation. Risk premium of a factor infers how much an investor demands to pay for bearing the risk of the factor. Risk price implies whether a factor is useful to explain the cross-section of average asset returns. When factors are uncorrelated, that is, $\mathrm{E}(ff')$ is a diagonal matrix. Then, $b_i = 0$ (the i^{th} factor is not priced) implies $\lambda_i = 0$ (the i^{th} factor earns zero risk premium), and vice verse. However, this is not true when factors are correlated: an unpriced factor can earn positive risk premium by being correlated with a useful factor. To give an example, suppose we have two factors f_1 and f_2, the covariance matrix is $\mathrm{E}(ff') = \begin{pmatrix} 10 & 1 \\ 1 & 10 \end{pmatrix}$, the first factor is priced and the second is not, that is $b_1 = 1 \neq 0$ and $b_2 = 0$. Then, according to (1.2), we have $\lambda_1 = 10$ and $\lambda_2 = 1$. So we find that the unpriced factor f_2 (i.e. $b_2 = 0$) earns non-zero risk premium (i.e. $\lambda_2 \neq 0$) by simply being correlated with a useful factor f_1. As discussed before, if factors are uncorrelated it is valid to use either risk price (SDF method) or risk premium (Fama-MacBeth regression) to select factors. However, factors are typically correlated in a high dimensional setting, so we should use *risk price* to infer priced factors.

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

$$\begin{aligned} e(b) &= \mathrm{E}[Rm(b)] = \mathrm{E}(R)\mathrm{E}(m(b)) + \mathrm{Cov}(R, m(b)) \\ &= r_0^{-1}\mathrm{E}(R)\mathrm{E}(1 - b'(f - E(f))) + r_0^{-1}\mathrm{Cov}(R, 1 - b'(f - E(f))) \\ &= r_0^{-1}[\mathrm{E}(R) - \mathrm{Cov}(R, f)b] \\ &= r_0^{-1}(\mu_R - Cb), \end{aligned} \tag{1.3}$$

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

$$Q(b) = e(b)' W e(b), \tag{1.4}$$

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

$$\hat{b} = \arg\min_b Q(b) = \arg\min_b (\mu_R - Cb)'W(\mu_R - Cb), \tag{1.5}$$

which gives

$$\hat{\hat{b}} = (\hat{C}'\hat{W}\hat{C})^{-1}\hat{C}'\hat{W}\hat{\mu}_R, \tag{1.6}$$

where $\hat{C} = \widehat{\mathrm{Cov}(R, f)} = \frac{1}{T}\sum_{t=1}^{T}(R_t - \hat{\mu}_R)(f_t - \hat{\mu}_f)'$, $\hat{\mu}_f = \frac{1}{T}\sum_{t=1}^{T} f_t$ and $\hat{\mu}_R = \frac{1}{T}\sum_{t=1}^{T} R_t$. $\hat{\hat{b}}$ is an empirical estimate of $\hat{b}$ where we use sample estimates of C and μ_R.[7] For the weighting matrix W, Ludvigson (2013) offers two choices of W for comparing models. First, $W = \mathrm{E}(RR')^{-1}$, which connects $Q(b)$ to the well known Hansen-Jagannathan (HJ) distance. Ludvigson (2013) points out that the use of HJ distance is more appropriate with limited asset choices (small N, large T), in which case the weighting matrix $\mathrm{E}(RR')^{-1}$ accounts for and offsets the variations of test assets, leading to stable estimators. On the other hand, when test assets are prolific, Ludvigson (2013) advocates the second choice of W: the identity matrix. She argues that using the identity matrix does not tilt the weight to favor any subset of test assets, especially when test assets represent particular economic interests. In our application, the test assets consist of firm characteristic sorted portfolios, hence we do not want to tilt the weights to favor any firm characteristics, so the identity matrix will be used as the weighting matrix throughout this paper.

1.2.2 Challenges and/or blessings of high-dimensionality

Cochrane (2011) points out that traditional methods like portfolio sorting to identify useful factors have fallen short in the high-dimensional world. For instance, following Fama and French (1992, 2008) to construct 5 by 5 portfolios, and supposing there are n characteristics, we have to sort all stocks into 5^n portfolios. When n is small, say

[6] Since r_0 in (1.3) is a constant, it can be dropped out in the minimization problem.

[7] Note that in Section 1.2.4, the statistical properties of $\hat{b}$ are built upon assumptions on C and μ_R. In order to have consistent estimator for $\hat{\hat{b}}$, we need to impose an additional condition that $T \gg N$ and $T \gg K$ to ensure consistent estimates of $\hat{C}$ and $\hat{\mu}_R$.

$n = 2$, it is handy to sort portfolios and check the marginal distribution of returns on each characteristic. However, when n is large, for instance $n = 10$, it is infeasible to sort stocks into $5^{10} \approx 9.8$ *million* portfolios. Yet, there are hundreds of anomaly based factors having been proposed in empirical asset pricing literature, see Harvey et al. (2015) and Hou et al. (2018b) for examples. For the Fama-MacBeth regression method, there are several complications too. First, K is likely to diverge $(K > N)$ in the high-dimensional world, in which case the Fama-MacBeth regression becomes infeasible. Second, variables are likely correlated under high-dimensionality. As discussed in Section 1.2.1, when factors are correlated, unpriced factors can earn positive risk premiums if they are correlated with priced factors. In this case, Fama-MacBeth regression is likely to pick up redundant factors. Third, Kleibergen (2009) cautions that the second pass Fama-MacBeth regression faces the weak factor identification problem when factors are correlated.

Nonetheless, empirical finance research has demonstrated strong evidence that many of those proposed factors are actually useless or redundant. Thus, the sparsity assumption which originates from the machine learning literature becomes a useful tool to handle these problems. Approximate sparsity assumes that for K candidate factors, there are at most S of them which are useful ($S \ll K$) while the exact number and location of these useful factors need not to be known ex ante. Tibshirani (1996) proposed the LASSO estimator which is a milestone in achieving sparsity within a convex optimization problem and subsequently has been widely used to solve high dimensional financial problems, see Chinco et al. (2019) for example. However, the LASSO estimator is also well known for its poor performance when covariates are correlated. Kozak et al. (2020), Figueiredo and Nowak (2016) and Zou and Hastie (2005) have demonstrated that when factors are correlated, the LASSO estimator is unstable and wrongly shrinks some useful factors to zeros.

To circumvent the curse of dimensionality while taking account of factor correlations, we introduce a newly developed machine learning tool, the Ordered-Weighted-LASSO (OWL) estimator (Figueiredo and Nowak, 2016), which explicitly allows for factor correlations.

1.2.3 The Ordered-Weighted-LASSO (OWL) estimator

The OWL estimator is achieved by adding a penalty term in equation (1.5)[8]

$$\hat{b} = \arg\min_{b} \frac{1}{2}(\mu_R - Cb)'(\mu_R - Cb) + \Omega_\omega(b), \qquad \Omega_\omega(b) = \omega'|b|_\downarrow, \tag{1.7}$$

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

$$\omega_i = \lambda_1 + (K - i)\lambda_2, \qquad i = 1, ..., K, \tag{1.8}$$

where λ_1 and λ_2 are two tuning parameters. In order to solve (1.7), we use the proximal gradient descent algorithm. More details about this algorithm are included in the Online appendix. The OWL estimator is sensitive to the choice of the weighting vector ω. So finding appropriate values for tuning parameters λ_1 and λ_2, which pin down the weighting vector, is crucial. Following the machine learning literature, we use a five-fold cross-validation method to find tuning parameters.[9]

1.2.4 Statistical properties

This section discusses the statistical properties of the OWL estimator. We first show that, with some regularity conditions, when the number of factors K is finite, the OWL estimator is consistent. Then we allow K to go to infinity and, with the sparsity assumption and restricted eigenvalue condition, we derive the convergence rate of the OWL estimator, and hence the conditions for consistent OWL estimation. Next, we devise a thresholded estimator based on the OWL estimate that can achieve consistency in model selection. Finally, we derive the grouping condition under which two correlated factors will be grouped together.

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

[9] Given the grid values of λ_1 and λ_2, at each point on the grid, we first divide the sample into five equal parts in their time series dimension. We use four parts (training sample) to estimate the model with OWL. After obtaining the estimated model, we forecast the returns of the fifth part (testing sample), and compute the root of mean squared forecast error (RMSE). We then repeat the same procedure five times by rotating the training samples and testing samples, and compute the average RMSE for this point on the grid. Tuning parameters are determined by the smallest average RMSE on the grid.

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

Suppose that

$$\mu_R = Cb^0 + \epsilon, \tag{1.9}$$

where ϵ is the pricing error from (1.3) after scaling a constant r_0^{-1}. Then (1.7) can be written as[10]

$$\hat{b} = \arg\min_b \quad \frac{1}{N}||\mu_R - Cb||_2^2 + \frac{1}{N}\sum_{i=1}^{K} \lambda_1 + \lambda_2(K-i)]|b|_{[i]}, \tag{1.10}$$

where$|b|_{[i]}$ is the i^{th} element of $|b|_{\downarrow} := (|b|_{[1]}, |b|_{[2]}, \cdots, |b|_{[K]})'$ and $|b|_{[1]} \geq |b|_{[2]} \geq ... \geq |b|_{[K]}$. In order to derive the next theorem, we make the following assumptions.

Assumption 1.2.1 (Gram matrix). *The $N \times K$ covariance matrix of returns and factors C is normalized, such that $\hat{\Sigma} = \dfrac{C'C}{N} \rightarrow_p \Sigma$, where Σ is a full rank matrix, $\hat{\Sigma}_{j,j} = 1$, for all $j \in \{1, ..., K\}$.*

Assumption 1.2.1 requires a full rank Gram matrix $\hat{\Sigma}$, which restricts applications to a low dimensional case where the number of factors K is smaller than the number of assets N. Theorem 1.2.1 below is built on Assumption 1.2.1, which delivers the consistency property of the OWL estimator in a typical low dimensional case ($K < N$).

Assumption 1.2.2 (Normality). *Suppose that ϵ in (1.9) follows a normal distribution such that $\epsilon \sim i.i.d.\, \mathbf{N}(0, \mathbf{I}\sigma^2)$, and $E(\epsilon' C^{(j)}) = 0$, where $C^{(j)}$ is the j^{th} column of C.*

The *i.i.d.* normal assumption imposed on ϵ is for the sake of obtaining the probability measures in (1.11) and (1.14). We recognize that this assumption is rather restrictive and we leave it as a further research agenda which we could explore under $\alpha-$mixing condition (weak correlation) and fat tails.

Theorem 1.2.1 (Consistency of OWL). *Let Assumptions 1.2.1 and 1.2.2 be satisfied. Suppose that $t > 0$, $\quad \lambda_0 = 2\sigma\sqrt{\dfrac{t^2 + 2\log K}{N}} = o(1)$, $\quad \lambda_1$ and λ_2 are such that $\dfrac{\lambda_1}{N} \geq \lambda_0$, $\quad \lambda_1 = o(N) \quad$ and $\quad \lambda_2 = o(N)$. Then with probability at least*

$$p = 1 - 2\exp(-\frac{t^2}{2}), \tag{1.11}$$

[10] Note that the scaler "2" on the second term of (1.10) is dropped because it is negligible when tuning parameter $\frac{\lambda_1}{N} \asymp \sqrt{\frac{\log K}{N}}$, which will be introduced in the next theorem.

the estimator $\hat{b}$ satisfies

$$(\hat{b}-b^0)'\hat{\Sigma}(\hat{b}-b^0) \leq \left(\lambda_0 + \frac{\lambda_1+\lambda_2(K-1)}{N}\right)||b^0||_1. \tag{1.12}$$

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

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

Proof: see Appendix 1.A.1.1.

Theorem 1.2.1 shows the consistency of the OWL estimator when K is finite and offers an upper bound in (1.12) of the estimation error of the OWL estimator $(\hat{b}-b^0)'\hat{\Sigma}(\hat{b}-b^0)$. It is derived in a low dimensional setting where the number of factors is small while the number of observable assets is large ($K \ll N$).

Next, we consider the high dimensional setting, where we allow the number of factors K to grow to infinity. Then we derive the convergence rate of the OWL estimator and the conditions for consistent estimation. With $K \gg N$, the Gram matrix $\hat{\Sigma}$ will be singular. In order to derive the convergence rate, we impose Assumptions 1.2.3 and 1.2.4.

Assumption 1.2.3 (Sparsity)**.** *Denote by S the number of non-zero parameters in $b^0 = \{b^0_1, b^0_2, \cdots, b^0_K\}$. We assume that $S\sqrt{\frac{\log K}{N}} = o(1)$ when $N, K \to \infty$.*

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

Assumption 1.2.4 (Restricted eigenvalue condition, Bickel et al. (2009))**.** *For all b such that $||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1$, $\hat{\Sigma}$ satisfies the restricted eigenvalue condition*

$$\phi_0^2 := \min_{\substack{s_0 \subset \{1,\ldots,K\} \\ |s_0|<K}} \min_{\substack{b \in R^K \setminus \{0\} \\ ||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1}} \frac{b'\hat{\Sigma}b}{||b_{s_0}||_2^2} > 0. \tag{1.13}$$

The restricted eigenvalue condition implies the compatibility condition in Buhlmann and Van de Geer (2011) (pp. 106), which is the key requirement to establish Theorem 1.2.2 below. See the online Appendix for the motivation and the derivation of the compatibility condition.

Theorem 1.2.2 (Convergence rate of OWL). *Let Assumptions 1.2.2, 1.2.3 and 1.2.4 be satisfied. Suppose that $t > 0$, $\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log K}{N}} = o(1)$ and let $\frac{\lambda_1}{N} = 2\lambda_0$. Then with probability at least*

$$p = 1 - 2\exp(-\frac{t^2}{2}), \tag{1.14}$$

$\hat{b}$ satisfies

$$(\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0) + \frac{\lambda_1}{N}||\hat{b} - b^0||_1 \leq 4(\frac{\lambda_1}{N})^2\frac{S}{\phi_0^2} + 2\frac{\lambda_2}{N}(K-1)||b^0||_1. \tag{1.15}$$

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

$$||\hat{b} - b^0||_1 = O(S\sqrt{\frac{\log K}{N}}), \qquad ||\hat{b} - b^0||_2 = O(\sqrt{\frac{S\log K}{N}}). \tag{1.16}$$

Proof: see Appendix 1.A.1.2.

Theorem 1.2.2 establishes the convergence rate of the OWL estimator in a high dimensional setting, where both K and N go to infinity. Following a similar argument from Kock and Callot (2015), by utilizing the ℓ_∞ bound and the convergence rate we can introduce a thresholded estimator $\tilde{b}$ that is consistent in model selection. From (1.16), we obtain

$$||\hat{b} - b^0||_\infty \leq ||\hat{b} - b^0||_2 \leq C\sqrt{\frac{S\log K}{N}}$$

with probability close to one by selecting a constant C sufficiently large. Given the OWL estimator $\hat{b}$, we define the thresholded estimator $\tilde{b}$ as

$$\tilde{b}_j = \begin{cases} \hat{b}_j & \text{if} \quad |\hat{b}_j| \geq H, \\ 0 & \text{if} \quad |\hat{b}_j| < H, \end{cases} \tag{1.17}$$

where H is the hard thresholding parameter. We set

$$H = c\sqrt{\frac{S\log K}{N}}, \tag{1.18}$$

where $c > 0$ is any positive fixed constant. Recall that by Assumption 1.2.3, $\sqrt{(\log K)/N} = o(1)$, so $H = o(1)$. In the following theorem we show that estimator $\tilde{b}$ is a consistent estimator of b^0 and can select the true non-zero coefficients of factors with probability tending to one. We assume that $K \to \infty$ and b^0 has property

$$\|b^0\|_2^2 = \sum_{j=1}^{K} (b_j^0)^2 \to \sum_{j=1}^{\infty} (b_j^0)^2 < \infty, \quad \text{as } K \to \infty. \tag{1.19}$$

Theorem 1.2.3 (Consistency of model selection by thresholding)**.** *Let Assumptions 1.2.2, 1.2.3 and 1.2.4 be satisfied and* (1.19) *holds. Then the following is true.*

(a) The thresholded estimator $\tilde{b}$ computed with H, (1.18)*, has property*

$$\|\tilde{b} - b^0\|_2 = o_p(1), \quad \textit{as } N \to \infty. \tag{1.20}$$

(b) $\tilde{b}$ has property

$$\mathbb{P}(\tilde{b}_j = 0, j \in \{k : b_k^0 = 0\}) \to 1, \quad \textit{as } N \to \infty. \tag{1.21}$$

(c) For any $\xi_n \to \infty$ such that $\xi_n H = o(1)$,

$$\mathbb{P}(\tilde{b}_j \neq 0, j \in \{k : |b_k^0| > \xi_n H\}) \to 1, \quad \textit{as } N \to \infty. \tag{1.22}$$

Proof: see Appendix 1.A.1.3.

Theorem 1.2.3 shows that thresholded estimator $\tilde{b}$ offers a theoretical foundation for consistency in model selection, in which $\tilde{b}$ will select the true useful factors as non-zeros while shrinking off all useless factors, with probability tending to one. However, finding a suitable threshold H is an *empirically challenging* problem, especially in small samples and this task is beyond the scope of this paper. After all, the goal of this paper is not to find a new parsimonious asset pricing model, but to identify a set of useful (and potentially highly correlated) factors that drive asset prices.

Next, we investigate the grouping condition under which correlated factors will be grouped together, i.e. assigning similar values to the coefficients of correlated factors.

Theorem 1.2.4 (Grouping)**.** *Let f_i and f_j be the i^{th} and j^{th} factor returns (both of size $T \times 1$). $\hat{b}_i$ and $\hat{b}_j$ are OWL estimates of risk prices of factor i and j. Let $\sigma(f_i - f_j)$ denote the standard deviation of the vector $f_i - f_j$, and μ_R, σ_R be the $N \times 1$ vectors* *collecting the mean and standard deviation of N test assets. If*

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

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

Proof: see Appendix 1.A.1.4.

Corollary 1.2.1. *Let $f_i, f_j, \lambda_2, \mu_R, \sigma_R$ be the same as in Theorem 1.2.4. 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 1.A.1.5.

Theorem 1.2.4 has several implications. First, when factors are highly correlated (i.e. $\sigma(f_i - f_j)$ is small) they are more likely to be grouped together (i.e. receive similar coefficients, $\hat{b}_i \approx \hat{b}_j$): two factors exhibiting high correlation could be the result of the same unobservable underlying factor that dictates these observable factors simultaneously. Thus, they should share similar magnitude in explaining asset returns which are driven by the same unobservable underlying factor. Second, the hyper parameter λ_2 in (1.8) has direct impact on factor grouping: large λ_2 encourages grouping.[11] Third, the mean (μ_R) and standard deviation (σ_R) of test assets affect the grouping property. A set of less informative assets (small μ_R and/or small σ_R) will result in factor grouping: weak factors are equally inadequate to explain a set of less informative test assets. Corollary 1.2.1 shows that the OWL estimator can also group negatively correlated factors and assign opposite signs to those factors.

Theorem 1.2.1 and Theorem 1.2.2 establish the consistency property of the OWL estimator under some regularity conditions and Theorem 1.2.3 establishes the theoretical foundation that a thresholded estimator based on the OWL estimator can achieve consistency in model selection. Theorem 1.2.4 shows that the OWL estimator *permits* correlations among factors and can *group* correlated factors, which distinguishes it from the LASSO estimator that suffers severely from correlated variables.

[11] A geometric interpretation of the OWL norm is included in the online appendix, and more details about how λ_2 affects the grouping property can be found in Zeng and Figueiredo (2015).

1.3 Simulation

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

1.3.1 Simulation design

In our experiment, consider K candidate factors, $2K/3$ of them are useful factors, that is they are priced ($b \neq 0$), and $K/3$ of them are useless or redundant factors ($b = 0$). Within these useful factors, $K/3$ are highly correlated, and $K/3$ are uncorrelated.

Let ρ denote the $K \times K$ correlation coefficient matrix of C ($N \times K$) defined in (1.3). We suppose that $\rho_1, \rho_2, \rho_3 \in (-1, 1)$ and ρ is divided into 3 blocks such that:

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

and

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

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

We first set the values of ρ_1, ρ_2 and ρ_3, and then randomly generate an $N \times K$ matrix C, denoted as $simC$, which has the correlation coefficient matrix of ρ.[12] We further set an oracle value for b (risk price). Then we simulate the cross section of average returns as $\mu_R = simC * b + e$, where e is a pricing error vector, with the

[12] In particular, we first randomly generate an $N \times K$ matrix where each column follows a standard normal distribution. Then multiply it with the Cholesky decomposition of ρ.

scale about 10% of $simC$, i.e. $e \sim N(0, 0.01)$. Finally, we estimate risk price with simulated data $simC$ and μ_R using OWL, LASSO, adaptive LASSO, Elastic Net, and naive OLS estimators.[13] Then we compare these estimators with the pre-specified oracle value of b.

1.3.2 Simulation results

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

Figure 1.1 reports the plot of the OWL estimator over 90 factors along with other benchmarks and the oracle value (black). The upper left panel displays the plots of estimated SDF coefficients for all factors. The remaining three panels display the detailed plot of estimates for each of these three blocks of factors. The upper right panel displays the plot of all estimates of useful factors that are highly correlated. In the presence of high correlation, the LASSO estimator performs poorly with highest estimation errors. Adaptive LASSO is strongly governed by the adaptive weights and is set to be the OLS estimate. So adaptive LASSO exhibits very similar behaviour to the OLS estimator. Elastic Net, as a hybrid estimator between LASSO and Ridge regression, is designed to stabilize LASSO selections in the presence of correlation. Although Elastic Net does improve the performance of LASSO in the context of correlated factors, it is still substantially outperformed by OWL. OWL produces the smallest estimation error and is the only estimator that groups together highly correlated variables by assigning them with similar coefficients. The bottom left panel displays the plot of all estimates of useless/redundant factors which are highly correlated. In terms of shrinking off useless/redundant factors, LASSO, EN, and OWL all perform well: they set most useless factors to zeros. By contrast, adaptive LASSO is affected by the adaptive weights (i.e., the OLS estimate) and fails to set many useless/redundant factors to zeros. The bottom right panel displays the plot of all

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

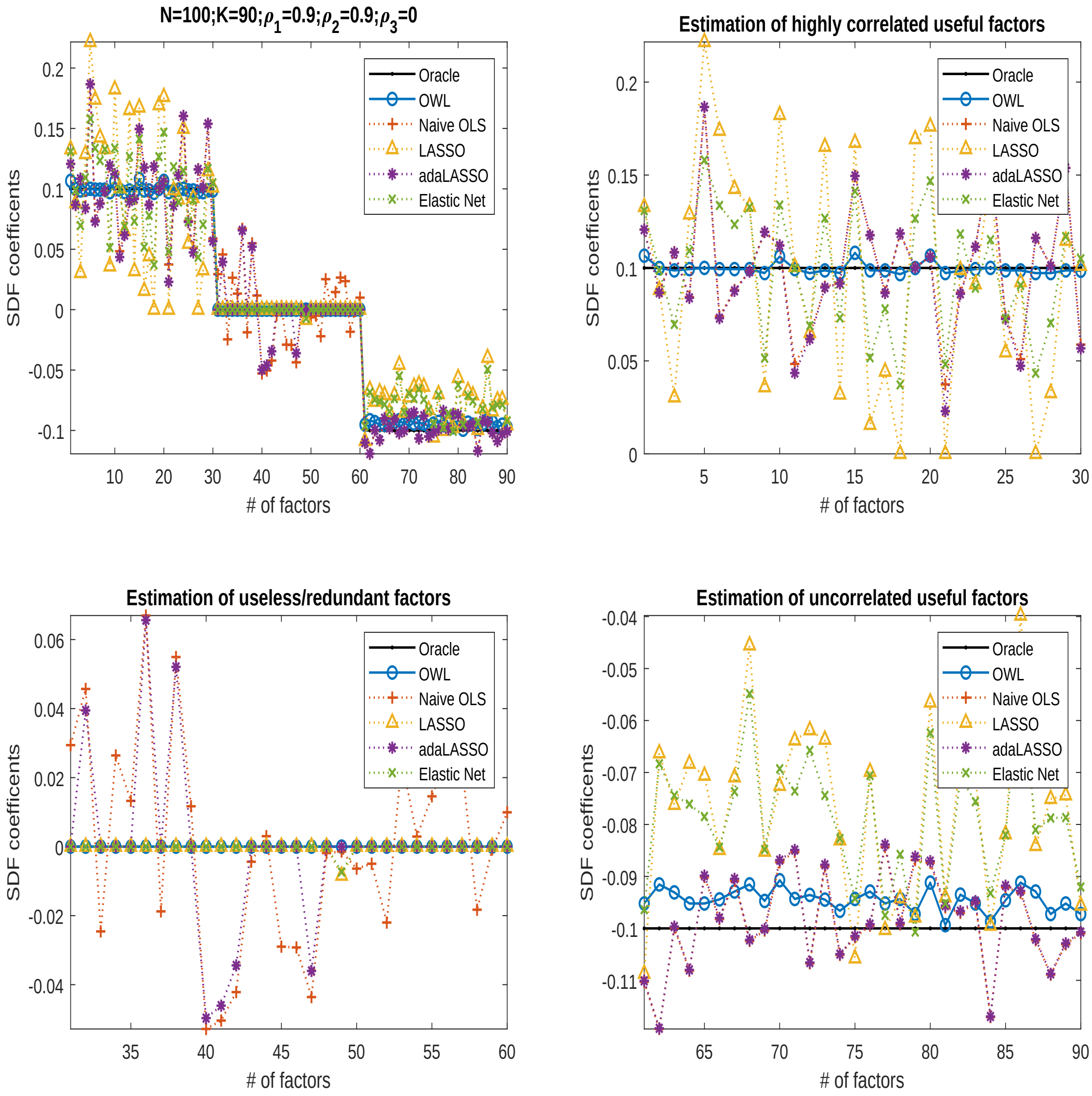


Figure 1.1. Estimation of SDF coefficients: $N = 100, K = 90$

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

estimates of useful factors which are not correlated. Again, LASSO and Elastic Net are the worst performers, yielding the largest estimation errors. Also note that in the uncorrelated setting Elastic Net performs similarly to LASSO. In the ideal world where factors are uncorrelated, OLS and adaptive LASSO are the best performers, which is tightly followed by OWL. Note that OWL, LASSO and Elastic Net are biased towards zero, which is typically observed for shrinkage-estimators in small samples.

For the robustness check of this experiment, we repeat the simulation multiple times and report the deviation of each estimator from the oracle values. Because of limited display space, we put the robustness check in Appendix 1.A.4.

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

Figure 1.2 reports the plot of estimated SDF coefficients using OWL and other benchmarks with 1000 test assets. When test assets are abundant, all shrinkage based estimators do a good job to shrink off useless/redundant factors. Adaptive LASSO performs the best at estimating uncorrelated factors: governed by the OLS weights, it is the only unbiased estimator among shrinkage based estimators. LASSO and Elastic Net produce the most biased estimators among all benchmarks. With highly correlated useful factors, OWL produces the most accurate estimation. With uncorrelated factors, OLS and adaptive LASSO are undoubtedly the best estimators, followed closely by OWL. For that reason, adaptive LASSO would be a good estimator in a low dimensional world where $N \gg K$. However, in a world of many factors, where $K > N$, OLS will be infeasible, hence the adaptive LASSO using OLS weighting is also improbable.

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

Figure 1.3 reports estimation results of each method along with the oracle value. Once $K > N$ the naive OLS estimator becomes infeasible, thus we remove it from the benchmarks. Meanwhile, we use the LASSO estimate as the adaptive weight for adaptive LASSO estimator. As for useless factors, all machine learning methods do a good job to shrink most useless factors to zeros. For the highly correlated useful factors

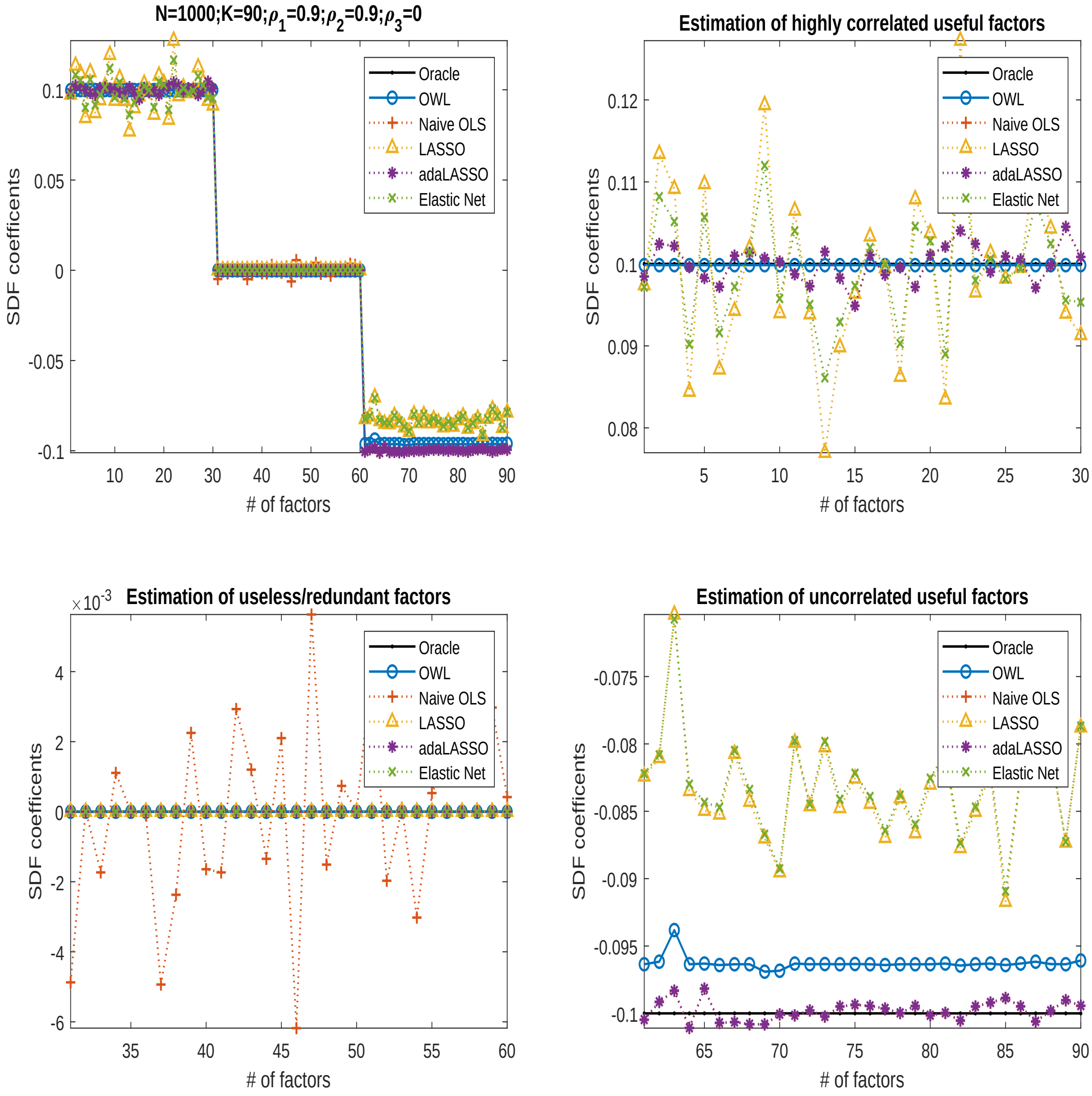


Figure 1.2. Estimation of SDF coefficients: $N = 1000, K = 90$

This figure reports the plot of the values of the OWL estimator along with other benchmark estimators. The number of assets is 1000. The rest are the same as in the first experiment in Figure 1.1.

OWL is still the best estimator, producing the smallest estimation error, while LASSO and adaptive LASSO are the worst performers producing very volatile estimates and wrongly shrinking many useful factors to zero. Interestingly, we find that Elastic Net performs significantly better compared to LASSO. However, despite this, Elastic Net is still substantially outperformed by OWL. For the useful factors (both correlated and uncorrelated), adaptive LASSO, using the LASSO estimate as the adaptive weight, performs the worst. The adaptive weight exacerbates the estimation severely.

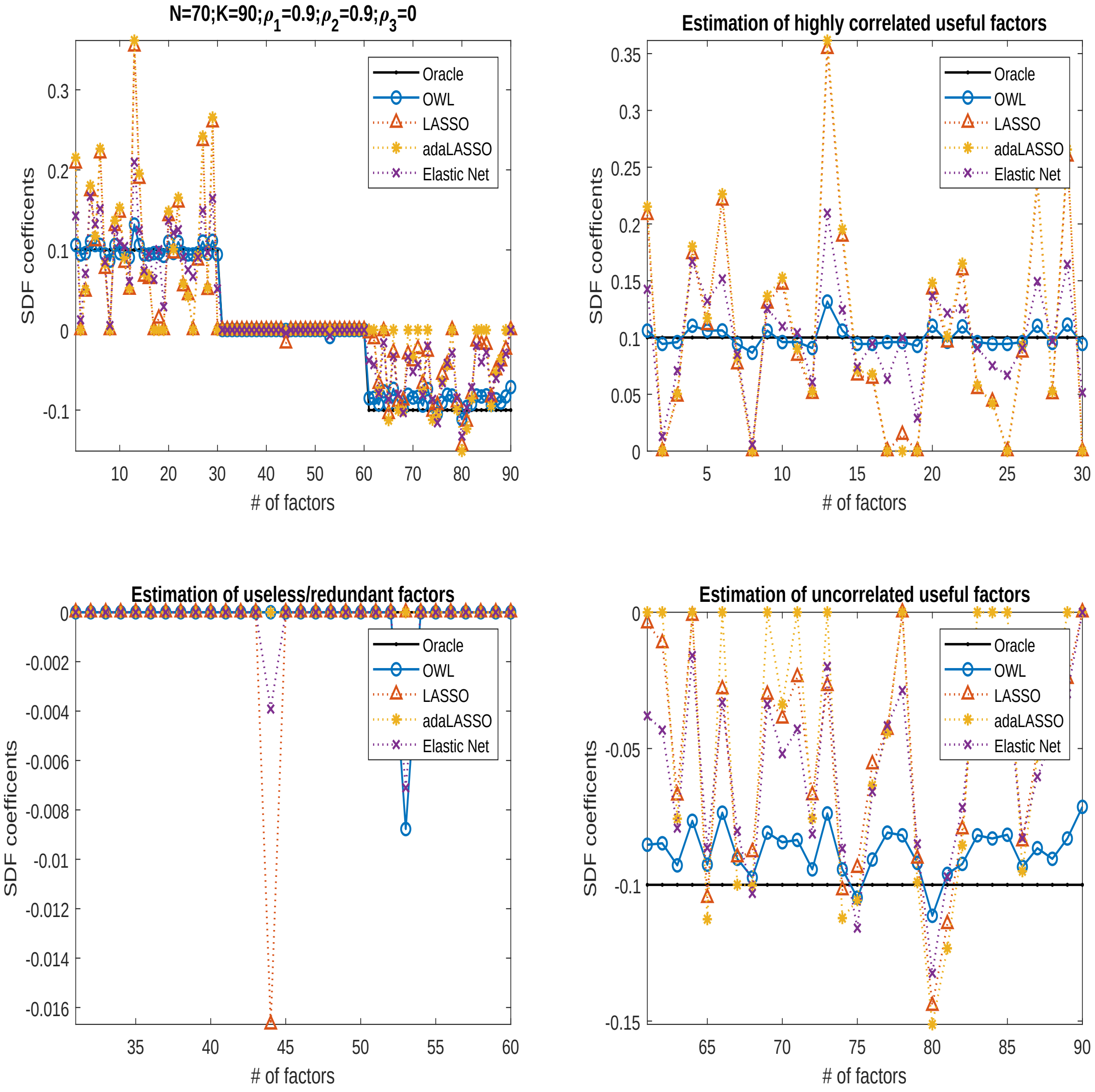


Figure 1.3. Estimation of SDF coefficients: $N = 70, K = 90$

This figure reports the plot of the values of the OWL estimator along with other benchmark estimators. Adaptive LASSO is using the LASSO estimate as its adaptive weight. The number of assets is 70. The rest are the same with the first experiment in Figure 1.1.

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

1.4 Empirical analysis

This section applies the OWL estimator while using the SDF method to find useful factors among 80 anomaly factors that drive the cross section of average returns in stock market. We first introduce the datasets, followed by a detailed account of the construction of anomaly factors and test portfolios. We consider both value weighted and equal weighted methods, controlling micro stocks. We construct pooled bi-variate sorted portfolios as test assets following a similar method to Feng et al. (2020).

1.4.1 Data

We use the U.S. stock data from the Center for Research in Security Prices (CRSP) and Compustat database[14] to construct anomaly variables and test portfolios. The period spans from January 1980 to December 2017, totalling 456 months on all NYSE, AMEX and NASDAQ listed common stocks. Risk-free rate and market excess returns are downloaded from Kenneth French's on-line data library. All anomaly variables are demeaned and scaled to have the same standard deviation with the market factor.

1.4.2 Constructing the anomaly factors

We consider 100 firm characteristics described in Green et al. (2017),[15] while deleting characteristics that have more than 40% missing data. Then, for each remaining characteristic, we sort stocks into decile portfolios at each month using uni-variate sorting. Micro stocks, defined as having market capitalization smaller than the 20 percentile of NYSE listed stocks, are removed. Although micro stocks only account for less than 10% of aggregated market capitalisation, they constitute about 56% of all stocks in the database, implying that small stocks would distort the interpretation of the aggregated market capitalization if not removed, also see Hou et al. (2014) and Fama and French (2015). Then, anomaly factors are computed as the spread returns between the top and the bottom decile portfolios. Characteristics having insufficient data to construct decile portfolios at every month will be dropped. Note that the sorting is always from high to low according to characteristics, and the anomaly

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

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

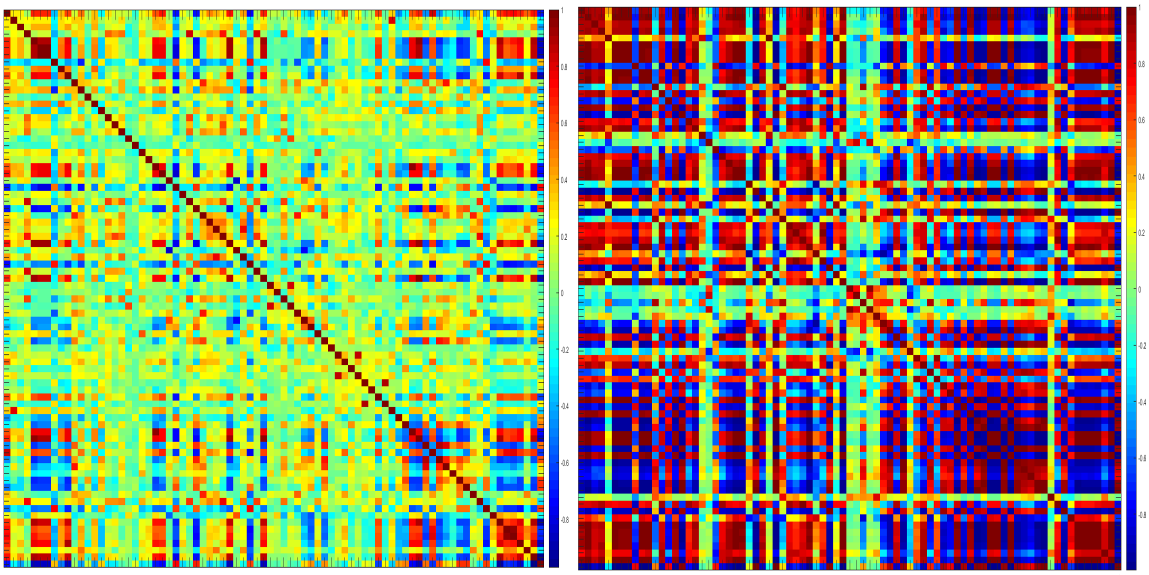
variables are top decile return minus the bottom decile return. That will end up with some slight difference with some familiar notations. For instance, the famous size factor ‘small-minus-big’ in our factor library would be ‘big-minus-small’; however, they are essentially the same after giving a negative sign. In estimation, we only care about the coefficient magnitude. The interpretation of the sign of coefficients should be looked at together with the sorting order when forming anomaly variables. Overall, we obtain 80 anomaly factors which are listed in Table 1.1. See Green et al. (2017) for a detailed description of each characteristic.

Table 1.1. Anomaly factors

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

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

Figure 1.4a displays the heat map of factor correlation coefficients matrix mea-



(a) Factor correlation measured by time series **(b)** Factor correlation measured by factor loadings

Figure 1.4. Factor correlation coefficients

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

sured by their time series.[16] It suggests that 16% of factors exhibit correlation coefficients (absolute value) greater than 0.5. In particular, ‘beta’ related factors are highly correlated with ‘liquidity’, ‘profitability’, ‘investment’, and other factors. For that reason, Green et al. (2017) exclude ‘beta’ related factors in the factor zoo. Figure 1.4b displays the heat map of factor correlation coefficients matrix measured by factor loadings. It exhibits much higher correlation compared to Figure 1.4a: 64% correlation coefficients (absolute value) are greater than 0.5, implying serious multicollinearity issue if the standard Fama-MacBeth regression is employed. Cochrane (2011) points out that *we need to find whether expected returns line up with covariances of returns with factors*, implying that correlation measured by factor loadings really matters for inferring priced factors.

[16] “Factor correlation measured by time series” means the correlation coefficients matrix is computed through the $T \times K$ factor time series data. “Factor correlation measured by factor loadings” means the correlation coefficients matrix is computed through the $N \times K$ factor loadings after the first stage of Fama-MacBeth regression.

1.4.3 Bi-variate sorted portfolios as test assets

Regarding test assets, there is a debate in the literature about using either individual stocks or sorted portfolios as test assets. Harvey and Liu (2017) use individual stocks with bootstrap method to test for predictability of anomaly factors, and they find that only two or three anomaly factors can significantly predict asset returns. Lewellen (2015) employed Fama-MacBeth to test for anomaly factors with individual stocks. However, others argue that individual stocks will introduce errors in variables (EIV). When regression is made on estimated variables, i.e. factor loadings, the pre-estimated factor loadings would incur estimation errors. Shanken (1992) modified the estimator by introducing the "Shanken's correction" term to mitigate EIV. However, others argue that "Shanken's correction" is minimal in small samples. On the other hand, Fama and French (2008), Hou et al. (2014), Feng et al. (2020) advocate sorted portfolios as test assets. Individual stocks are usually noisy and exhibit outliers, which are the main source of EIV. Sorted portfolios are (weighted) mean returns of a group of stocks sharing some similar characteristics, which would mitigate the EIV problem. Hence, using sorted portfolios as test assets is an alternative (arguably better) way to avoid EIV.

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

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 capitalization. If individual stocks are used to gauge factor impact, it is inevitable that they will 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 the market will be out-weighted by the large number of small stocks. Portfolio sorting, however, can circumvent this issue by using the value weighted method, in which portfolio returns are computed by the weighted average of stocks returns where the weights reflect their market capitalization.

Fama and French (1992, 2015) use bi-variate sorting to create the five by five test portfolios which have now become popular choices for test assets. However, Harvey et al. (2015) caution that when only a small set of sorted portfolios are considered for test assets, factor selection is biased towards the same characteristics that are used to form test portfolios. Lewellen et al. (2010) argue that the 25 size and value sorted portfolios are too low a threshold to test factors. They recommend adding other portfolios in test assets. To strike a balance between using sorted portfolios and individual stocks as test assets, Feng et al. (2020) construct a large set of combined portfolios as test assets. In particular, they single out 'size' characteristic and combine it with the remaining characteristics to form five by five bi-variate sorted portfolios and pool them together as the grand set of test assets.[17] We follow Feng et al. (2020) to construct test portfolios and we obtain 1927 test portfolios as the grand set of test assets.[18]

1.4.4 Which factors matter?

Considering high correlation among factors, we apply the OWL estimator while using the SDF method to select useful factors from the 81 candidate factors.[19]

Table 1.2 reports the result of the OWL estimation. The first 5 columns are estimated using the full sample, ranging from January 1980 to December 2017; columns 6-7 report results from 1980 to 2000, and columns 8-9 from 2001 to 2017. Both the value weighted (vw) and equal weighted (ew) methods are considered. In order to gauge the impact of small stocks, we consider three thresholds for micro stocks. This table lists all anomaly factors selected in each estimation. It also reports how many times each factor has been selected by all estimations and the ordinal number (in the bracket) for each factor in a separate estimation, which indicates the importance of the factor (smaller number implies greater importance).

[17] 'Size' has been widely acknowledged as an important characteristic in asset pricing literature. Fama and French (1992, 2015), Hou et al. (2014), Carhart (1997) all include the 'size' and the 'market' factors in their models.

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

[19] We include the market factor together with 80 anomaly factors, total 81 candidate factors.

Table 1.2. Estimation results of the OWL estimator

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

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

Table 1.2 shows that 'size' (mve) has been selected as the most important factor in most of these estimations which, however, is not surprising. 'Size' characteristic has multiple entries in forming test portfolios, thus 'size' impact prevails in test portfolios. For this reason we exclude 'size' factor as a competing factor, yet we include it in the table to show that OWL can correctly identify relevant factors.

The 'illiquidity' (ill) factor (Amihud, 2002) is the most important factor that drives variations in test asset returns. Its explanatory power is particularly evident with smaller stocks. Portfolios sorted with size greater than 20 or 30 percentile of NYSE listed stocks exhibit higher importance of 'illiquidity' than those with 40 percentile. That implies small firms face severer liquidity constraints, and demand risk premiums to compensate for bearing the risk. 'Standard deviation of dollar volume' (std_dolvol)

(Chordia et al., 2001) which is another proxy for liquidity risk, follows ‘illiquidity’, becoming the second most important anomaly factor. Meanwhile, its high correlation with ‘illiquidity’ is also identified by the OWL estimator. Liquidity as a risk source that 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. ‘Asset growth rate’ (agr) follows ‘illiquidity’ and ‘standard deviation of dollar volume’ as the third most frequently selected anomaly factor. This finding coincides with Hou et al. (2018a) in which they propose a new $q5$ model, adding ‘asset growth rate’ as a fifth factor into their famous $q4$ model (Hou et al., 2014). Other anomaly factors that have been selected multiple times include ‘beta’, ‘beta squared’ (betasq), ‘cash to debt ratio’, and ‘percentage change in current ratio’ (pchcurrat), which are also related to liquidity risk. Beyond that, ‘momentum’, ‘return on invested capital’ (roic), ‘return on assets’ (roaq) and other profitability related factors are also selected by the OWL estimator multiple times.

Columns 6 and 7 report estimations using the 1980 - 2000 sub-sample and columns 8 and 9 report estimations using the 2001 - 2017 sub-sample. We find that liquidity constraint only appears in the second sub-sample (2001 - 2017), where liquidity related factors (‘baspread’, ‘standard deviation of dollar volume’, ‘change in quick ratio’, etc...) play an important role in explaining the cross section of average returns. However, in the first sub-sample (1980 - 2000), columns 6 and 7 show no strong evidence that liquidity related factors drive asset prices. Meanwhile, ‘momentum’ and ‘profitability’ related factors primarily drive asset prices between 1980 and 2000.

Interestingly, from 1980 to 2000, with 20-percentile-micro-stocks excluded, we find ‘size’ (mve) is not selected by the OWL estimator, which makes it the only exception from all estimations. This phenomenon is well documented in the literature (see Amihud (2002), van Dijk (2011) and Asness et al. (2018)): the size effect weakened after its discovery in the early 1980s. However, when removing 40-percentile-micro-stocks, size effect is evident again, which implies the vanishing of size effect is likely to be caused by some small “junk” stocks. Once removing these junk stocks, size effect resurfaces again, which echoes the discovery by Asness et al. (2018): *size matters, if you control your junk.*

1.4.5 Robustness check

In this section, we check whether liquidity related factors are robust in explaining the cross section of asset returns as well as how small stocks affect factors' interpretations. Because of the limitation of space, we place this section in Appendix 1.A.4.

1.4.6 Out-Of-Sample analysis

Freyberger et al. (2020) point out that out-of-sample (OOS) exercise ensures that in-sample over-fit does not explain superior performance in model selection. In this subsection, we will evaluate the OOS performance of portfolios hedged with OWL selected factors, and compare it with other benchmarks. To offer some insights to the possible time-varying trend in prominent factors, we also consider two sub-samples, divided before and after 2000. We report the first five factors with highest estimated coefficients (absolute value).[20]

Table 1.3 shows the five most prominent factors selected using various methods in different samples, controlling micro stocks. We consider both the full sample estimation and the sub-sample estimations. We can find obvious differences in selected factors between full-sample and sub-samples, as well as between sub-samples. In addition, controlling micro stocks has a big impact on factor selection too. While including all micro stocks (P00), OWL and other methods select a mixture of 'liquidity', 'profitability' and 'momentum' related factors. However, once we remove micro stocks (P20 and P40), we can find some patterns in selected factors: OWL suggests that the most important factors to drive asset prices in the first sub-sample are 'momentum' and 'profitability' related factors while 'liquidity' related factors are relatively unimportant. However, the implication is reversed in the second sub-sample, where 'liquidity' related factors mainly drive asset prices. On the other hand, LASSO and

[20] Concerning over-fitting typically yields poor performance in out-of-sample exercise, we consider a five-factor model for out-of-sample prediction. We also consider a four-factor and a three-factor model for robustness check. We find that a four-factor model performs slightly better than the five-factor model in predictions. However, due to limited reporting space, we do not include them and they are available on request.

Table 1.3. Full/sub-sample factor selection using various methods

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

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

other methods do not show a clear pattern of change in characteristics. Moreover, 'mkt' as a primary factor selected by OWL when excluding micro stocks, is missing by other methods, which is counter-intuitive. The market factor should be the dominating factor driving asset prices when micro stocks are removed since idiosyncratic risks have been largely reduced. However, LASSO, Elastic Net and Fama-MacBeth estimators all fail to identify 'mkt' as an important factor, due to the high correlation between 'mkt' and other factors.

Next, we want to compare the out-of-sample performance between various methods. In particular, we follow a similar procedure to Freyberger et al. (2020) to form factor-hedged portfolios using a rolling window scheme to predict returns. First of all, we choose five most prominent factors as in Table 1.3 for the full sample and two sub-samples, while controlling micro stocks at the 20- and 40-percentile levels. Then we use a rolling window scheme (rolling window size is 120 months) to evaluate the performance of the factor-hedged portfolios with each method. Specifically, at the end of each estimation window, we regress each test asset on factors selected by each method, but one period lagged. For instance, at time t, we regress each test asset return from $t - 120 - 1$ to t on selected factors from $t - 120 - 2$ to $t - 1$, and obtain $\hat{\beta}$. We then forecast each test asset's next period return (at $t + 1$) by multiplying $\hat{\beta}$ and selected factors at t. We then sort stocks by their predicted returns into decile portfolios and long the top decile and short the bottom decile. At the next period $(t+1)$, when returns are realized, we can compute the spread portfolio return. Subsequently, we roll the window one period forward and repeat the steps until the end of period. In the end we compute four moments of the factor-hedged portfolio returns in the out-of-sample period as well as the Sharpe ratio.

Table 1.4 reports performance scores including the Sharpe ratio and the four moments of out-of-sample returns using the OWL, LASSO, Elastic Net and Fama-MacBeth estimators while controlling micro stocks. Panel A suggests that in the full sample estimation, the OWL estimator produces about 20% higher Sharpe ratio than other benchmarks. In addition, we find that the skewness and the kurtosis of the OWL hedged portfolio are much smaller than those of other benchmarks. Fama-MacBeth estimator typically performs the worst. We reckon that it is severely affected by factor correlations and estimation result is eroded by weak factors in the second pass Fama-

Table 1.4. Out-of-sample portfolio performance with a five-factor model

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

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

MacBeth regression (Kleibergen, 2009). In sub-sample estimations, we find that the Sharpe ratios are typically much higher than that of the full-sample estimation in all methods we considered. Meanwhile, we find that skewness and kurtosis are significantly reduced compared to the full-sample estimation, making the distribution of the out-of-sample returns more "normal" alike. This trend signals a possible time-varying nature in prominent factors which drives asset prices.

1.5 Conclusion

In the zoo of factors, using traditional methods to find factors that provide independent information about average returns faces tremendous challenges. In addition, correlations make the matter worse: among 80 anomaly factors we considered, 64% of factor loadings exhibit correlation coefficients greater than 0.5 (absolute value). However, factor correlations cause severe complications in the LASSO estimator (Kozak et al., 2020; Figueiredo and Nowak, 2016) and in the Fama-MacBeth regression (Kleibergen, 2009; Cochrane, 2005). The OWL estimator, on the other hand, permits correlated variables and achieves correlation identification and sparsity shrinkage simultaneously. We show that the OWL estimator is a consistent estimator under some regularity conditions, and we derive the grouping conditions for correlated factors. Monte Carlo experiments confirm the superior performance of the OWL estimator against other benchmarks, especially when factors are correlated. Empirical analysis shows that 'liquidity' related factors play an important role to drive asset prices, meanwhile sub-sample estimations suggest a shift in economic characteristics and reveal a time-varying nature in factor selections.

Finally, note that the purpose of this paper is not to find a parsimonious asset pricing model, but to identify a set of sparse factors, potentially highly correlated, to explain the cross section of average returns given a certain period. With that in mind, our procedure is particularly useful for factor investing: OWL can identify correlated factors that jointly drive stock returns, and can be further utilized to form portfolio strategies, see Asness et al. (2013) for instance. Meanwhile, we notice that there is a time-varying trend in prominent factors that drive asset prices, which argues for a time-varying model to be placed on the future research agenda. Future work can also extend the statistical theories established in this paper, for instance to derive the asymptotic properties of the OWL estimator under more general conditions. Furthermore, developing a de-biased version of the OWL estimator is possible following the recent development of the de-sparsified LASSO estimator as in Van De Geer et al. (2014) and Kock (2016), which will enable robust inferences.

1.A Appendix

1.A.1 Technical proofs

1.A.1.1 Proof of Theorem 1.2.1

Proof. By definition the OWL estimator is minimizing the function

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

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

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

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

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

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

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

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

where

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

Consider the event

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

where $\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log K}{N}}$ by assumption. Then in view of (1.A.3), (1.A.2) can be bounded as

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

By triangle inequality, $||\hat{b}-b^0||_1 \leq ||\hat{b}||_1 + ||b^0||_1$. Therefore (1.A.4) can be written as

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

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

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

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

$$(\hat{b}-b^0)'\hat{\Sigma}(\hat{b}-b^0) = \frac{1}{N}\|C(\hat{b}-b^0)\|_2^2 \leq [\lambda_0 + \frac{\lambda_1 + \lambda_2(K-1)}{N}]||b^0||_1. \tag{1.A.7}$$

This completes the proof of (1.12).

We obtained (1.A.7) assuming (1.A.3). Now we compute the probability of inequality (1.A.3) to be true.

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

Using the Gaussian tail bound, $\mathbf{P}(|V_j| > x) \leq 2\exp(-x^2/2)$, we have

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

Consequently, (1.A.3) is valid with probability

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

This completes the proof of (1.11).

Let K be fixed. By Assumption 1, $\hat{\Sigma}$ is a positive definite matrix, therefore $(\hat{b} - b^0)'\hat{\Sigma}(\hat{b}-b^0) \geq \Lambda_{min}\|\hat{b}-b^0\|_2^2$, where Λ_{min} is the smallest eigenvalue of $\hat{\Sigma}$, and $\Lambda_{min} > 0$. Note that as $N \to \infty$, the right-hand-side of (1.12) tends to 0. By assumptions of the theorem, $\lambda_0 = o(1), \frac{\lambda_1}{N} = o(1)$ and $\frac{\lambda_2 K}{N} = o(1)$. Further if $t \to \infty$, (1.A.3) holds with probability $p \to 1$. Then it follows trivially that

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

This completes the proof of the last claim of Theorem 1.2.1. □

1.A.1.2 Proof of Theorem 1.2.2

Proof. Using the "argmin" property, we have

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

By assumption, $\frac{\lambda_1}{N} = 2\lambda_0$. Then (1.A.8) can be written as

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

Note that

$$||\hat{b}||_1 = ||\hat{b}_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1 \ge ||b^0_{s_0}||_1 - ||\hat{b}_{s_0} - b^0_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1, \quad (1.A.10)$$

$$||\hat{b} - b^0||_1 = ||\hat{b}_{s_0} - b^0_{s_0}||_1 + ||\hat{b}_{s_0^c}||_1. \quad (1.A.11)$$

Therefore, using (1.A.10) and (1.A.11), (1.A.9) can be written as

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

Note that $\|b^0_{s_0}\|_1 = \|b^0\|_1$, so (1.A.12) can be written as

$$\frac{2}{N}||C(\hat{b} - b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b}_{s_0^c}||_1 \le 3\frac{\lambda_1}{N}||\hat{b}_{s_0} - b^0_{s_0}||_1 + \frac{2\lambda_2(K-1)}{N}||b^0||_1. \quad (1.A.13)$$

By (1.A.11), $||\hat{b}_{s_0^c}||_1 = ||\hat{b} - b^0||_1 - ||\hat{b}_{s_0} - b^0_{s_0}||_1$. Utilizing this in (1.A.13), we obtain

$$\frac{2}{N}||C(\hat{b} - b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b} - b^0||_1 \le 4\frac{\lambda_1}{N}||\hat{b}_{s_0} - b^0_{s_0}||_1 + \frac{2\lambda_2(K-1)}{N}||b^0||_1. \quad (1.A.14)$$

By Assumption 4, the restricted eigenvalue condition states that

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

which implies that for any b,

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

where S is defined in Assumption 3. Rearranging the above inequality, we have

$$||b_{s_0}||_1^2 \leq b'\hat{\Sigma}bS/\phi_0^2, \tag{1.A.15}$$

which is called the *compatibility condition* in Buhlmann and Van de Geer (2011) pp. 106.

Applying (1.A.15) on $||\hat{b}_{s_0} - b_{s_0}^0||_1$ and using $\hat{\Sigma} = \frac{C'C}{N}$, we have

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

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

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

So (1.A.14) can be written as

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 + \frac{\lambda_1}{N}||\hat{b} - b^0||_1 \leq 4(\frac{\lambda_1}{N})^2\frac{S}{\phi_0^2} + \frac{2\lambda_2(K-1)}{N}||b^0||_1. \tag{1.A.16}$$

Note that $\frac{1}{N}||C(\hat{b} - b^0)||_2^2 = (\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0)$, so (1.A.16) completes the proof of (1.15).

By assumption of theorem $\frac{\lambda_1}{N} = 2\lambda_0 = 4\sigma\sqrt{\frac{t^2 + 2\log K}{N}}$ and $\lambda_2 = O(\frac{S\log K}{K})$. Therefore, for fixed t or $t = O(\sqrt{\log K})$, both two terms on the right hand side of (1.A.16) are $O(\frac{S\log K}{N})$. Hence, (1.A.16) implies

$$\frac{1}{N}||C(\hat{b} - b^0)||_2^2 = O\left(\frac{S\log K}{N}\right), \tag{1.A.17}$$

$$||\hat{b} - b^0||_1 = O\left(S\sqrt{\frac{\log K}{N}}\right). \tag{1.A.18}$$

So (1.A.18) proves the first claim of (1.16). Observe that

$$\frac{1}{N}||C(\hat{b}-b^0)||_2^2 = (\hat{b}-b^0)'(\hat{\Sigma}-\Sigma)(\hat{b}-b^0) + (\hat{b}-b^0)'\Sigma(\hat{b}-b^0), \tag{1.A.19}$$

Notice that

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

where Λ_{min} denotes the smallest eigenvalue of Σ, and Σ is the true value of $\hat{\Sigma}$, so $\Lambda_{min} > 0$. Moreover in (1.A.19), it holds

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

where $||\hat{\Sigma}-\Sigma||_\infty := \max_{1\leq i,j\leq K}|\hat{\Sigma}_{i,j}-\Sigma_{i,j}|$. Using Lemma 14.12 in Buhlmann and Van de Geer (2011), we have $\max_{1\leq i,j\leq K}|\hat{\Sigma}_{i,j}-\Sigma_{i,j}| = O_p(\sqrt{\frac{\log K}{N}})$. Hence (1.A.17) can be rewritten as

$$\begin{aligned} O\left(\frac{S\log K}{N}\right) &= \frac{1}{N}||C(\hat{b}-b^0)||_2^2 \\ &\geq \Lambda_{min}^2||\hat{b}-b^0||_2^2 - ||\hat{\Sigma}-\Sigma||_\infty||\hat{b}-b^0||_1^2 \\ &\geq \Lambda_{min}^2||\hat{b}-b^0||_2^2 - O_p\left(S^2\left(\frac{\log K}{N}\right)^{3/2}\right). \end{aligned} \tag{1.A.20}$$

Rearranging it, we have

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

By Assumption 3, $S\sqrt{\frac{\log K}{N}} = o(1)$. Together with $\frac{1}{\Lambda_{min}^2} = O(1)$, we obtain

$$||\hat{b}-b^0||_2^2 = O_p(\frac{S\log K}{N}), \tag{1.A.21}$$

which proves the second claim of (1.16). Lastly, the claim of (1.14) follows using the same argument as in the proof of Theorem 1.2.1. □

1.A.1.3 Proof of Theorem 1.2.3

Proof. We can bound

$$||\tilde{b}-b^0||_2 \leq ||\hat{b}-b^0||_2 + ||\tilde{b}-\hat{b}||_2. \tag{1.A.22}$$

By (1.16) of Theorem 1.2.2,

$$\|\hat{b} - b^0\|_2 = O_p(\sqrt{\frac{S \log K}{N}}) = O_p(H) = o_p(1). \tag{1.A.23}$$

Thus to prove (1.20), it suffices to show $\|\tilde{b} - \hat{b}\|_2 = o_p(1)$. By definition of $\tilde{b}$ in (1.17),

$$\|\tilde{b} - \hat{b}\|_2^2 = \sum_{j=1, b_j^0 \neq 0}^{K} \hat{b}_j^2 \mathbf{1}(|\hat{b}_j| < H).$$

Since $\hat{b}_j^2 \leq 2(\hat{b}_j - b_j^0)^2 + 2{b_j^0}^2$, we have

$$\|\tilde{b} - \hat{b}\|_2^2 \leq 2\|\hat{b} - b^0\|_2^2 + 2\mathbf{I}_K,$$

where $\mathbf{I}_K := \sum_{j=1, b_j^0 \neq 0}^{K} (b_j^0)^2 \mathbf{1}(|\hat{b}_j| < H)$. It remains to show that

$$\mathbf{I}_K = o_p(1). \tag{1.A.24}$$

Let $M \to \infty$ be a large number and $MH \to 0$. Denote

$$\begin{aligned} A_{1,j} &= \{|\hat{b}_j - b_j^0| \geq MH\}, \\ A_{2,j} &= \{|\hat{b}_j - b_j^0| < MH,\ |\hat{b}_j| \leq H,\ |b_j^0| \geq (M+2)H\}, \\ A_{3,j} &= \{|b_j^0| < (M+2)H,\ b_j^0 \neq 0\}. \end{aligned}$$

Then

$$\{|\hat{b}_j| < H,\ b_j^0 \neq 0\} \subset A_{1,j} \cup A_{2,j} \cup A_{3,j}.$$

So

$$\begin{aligned} I_K &\leq \sum_{l=1}^{3} \sum_{j=1}^{K} {b_j^0}^2 \mathbf{1}(A_{l,j}) \\ &:= I_{K,1} + I_{K,2} + I_{K,3}. \end{aligned}$$

To prove (1.A.24), it suffices to show

$$I_{K,l} = o_p(1), \quad \text{for } l = 1, 2, 3. \tag{1.A.25}$$

By (1.19), $\max_j |b_j^0| \le c_0 < \infty$, so

$$\begin{aligned} I_{K,1} &\le c_0^2 \sum_{j=1}^{K} \mathbf{1}(|\hat{b}_j - b_j^0| \ge MH) \le c_0^2 \sum_{j=1}^{K} \frac{(\hat{b}_j - b_j^0)^2}{(MH)^2} \\ &= \frac{1}{M^2} c_0^2 \frac{\|\hat{b} - b^0\|_2^2}{H^2} = \frac{1}{M^2} O_p(1) = o_p(1), \end{aligned}$$

as $M \to \infty$ by (1.A.23). Notice that $A_{2,j} = \emptyset$ is an empty set. Indeed, in $A_{2,j}$

$$\begin{aligned} |\hat{b}_j| &= |\hat{b}_j - b_j^0 + b_j^0| \ge |b_j^0| - |\hat{b}_j - b_j^0| \\ &\ge (M+2)H - MH \ge 2H, \end{aligned}$$

which contradicts $|\hat{b}_j| < H$. Therefore, $I_{K,2} = 0$.

Finally, by (1.19) and the definition of $A_{3,j}$,

$$\begin{aligned} I_{K,3} &= \sum_{j=1}^{K} {b_j^0}^2 \mathbf{1}(|b_j^0| < (M+2)H) \\ &\le \sum_{j=1}^{\infty} {b_j^0}^2 \mathbf{1}(|b_j^0| < (M+2)H) \\ &= o_p(1) \quad \text{as } N \to \infty, \end{aligned}$$

for any M, because $MH \to 0$. This proves (1.20) and completes the proof of part (a) of Theorem 1.2.3.

To prove part (b) it suffices to show that

$$\max_{j:|b_j^0|=0} |\tilde{b}_j| = o_p(1).$$

We have

$$\max_{j:|b_j^0|=0} |\tilde{b}_j| = \max_{j:|b_j^0|=0} |\tilde{b}_j - b_j^0| \le \|\tilde{b} - b^0\|_2 = o_p(1),$$

by part (a), which completes the proof of part (b).

Now we turn to part (c). Take any $j \in \{1, \cdots, K\}$, and let $|b_j^0| \ge \xi_n H$. Then

$$\begin{aligned} |\hat{b}_j| &= |\hat{b}_j - b_j^0 + b_j^0| \ge |b_j^0| - |\hat{b}_j - b_j^0| \ge |b_j^0| - \|\hat{b} - b^0\|_\infty \\ &\ge \xi_n H - \|\hat{b} - b^0\|_2 = \xi_n H - O_p(H) = (\xi_n - O_p(1))H. \end{aligned}$$

Therefore, with $\xi_n \to \infty$,

$$\mathbb{P}(\min_{|b_j^0| \geq \xi_n H} |\tilde{b}_j| = 0) = \mathbb{P}(\min_{|b_j^0| \geq \xi_n H} |\hat{b}_j| < H) \leq \mathbb{P}((\xi_n - O_p(1))H < H) \to 0,$$

which implies (1.22). Here we complete the proof of part (c). □

1.A.1.4 Proof of Theorem 1.2.4

The proof of Theorem 1.2.4 relies on the Pigou-Dalton transfer principle and the directional derivative lemma at the minimum of a convex function. It follows using a similar argument as in Figueiredo and Nowak (2016), except that we are dealing with both the time-series and cross-sectional dimensions.

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

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

Lemma 1.A.2 (Directional derivative)**.** *The directional derivative of function $f : R^K \to R$ at $x \in dom(f)$, in the direction $u \in R^K$ is given by*

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

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

Proof of Theorem 1.2.4 . Denote the objective function in (1.7) as $Q(b) := \frac{1}{2}||\mu_R - Cb||_2^2 + \Omega_\omega(b)$. By definition, $\hat{b}$ is the minimizer of $Q(b)$, $Q(\hat{b}) \leq Q(b)$ for all b. Thus by Lemma 1.A.2, for any u,

$$Q'(\hat{b}, u) \geq 0. \tag{1.A.26}$$

Suppose

$$\sigma(f_i - f_j) < \frac{\lambda_2}{||\mu_R||_2 ||\sigma_R||_2}, \tag{1.A.27}$$

and assume

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

We will show a contradiction between the assumption $\hat{b}_i \neq \hat{b}_j$ and (1.A.27). Without loss of the generality, assume $\hat{b}_i > \hat{b}_j$, $i < j$. First we define a special direction vector $u = (u_1, \cdots, u_K)$. Set $u_i = -1$, $u_j = 1$, $u_k = 0$, for $k \neq i, j$. The directional derivative of Q at $\hat{b}$ with such u is

$$Q'(\hat{b}, u) = \lim_{\alpha \to 0^+} \left(QL_\alpha(\hat{b}, u) + RP_\alpha(\hat{b}, u) \right), \tag{1.A.28}$$

where

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

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

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

Observe that

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

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

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

So for any α and u,

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

By the definition of ω in (1.8), $\Delta_\omega = \lambda_2$. Therefore, applying the above bound in (1.A.28), we obtain

$$\begin{aligned} Q'(\hat{b}, u) &\leq (\mu_R - C\hat{b})(C_i - C_j) - \Delta_\omega \\ &= (\mu_R - C\hat{b})(C_i - C_j) - \lambda_2. \end{aligned} \tag{1.A.29}$$

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

$C_j||_2$. So (1.A.29) becomes

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

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

$$Q'(\hat{b},u) < ||\mu_R||_2 \; ||\text{cov}(R, f_i - f_j)||_2 - \lambda_2. \tag{1.A.30}$$

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

$$\text{cov}(R_n, f_i - f_j) \leq \sqrt{\text{var}(R_n)\text{var}(f_i - f_j)} = \sigma_{R_n}\sigma(f_i - f_j), \tag{1.A.31}$$

where σ_{R_n} is the standard deviation of the n^{th} test asset. Apply (1.A.31) in (1.A.30), we have

$$\begin{aligned} Q'(\hat{b},u) &< ||\mu_R||_2 \; ||\sigma_R\sigma(f_i - f_j)||_2 - \lambda_2 \\ &= ||\mu_R||_2 \; ||\sigma_R||_2 \; \sigma(f_i - f_j) - \lambda_2, \end{aligned} \tag{1.A.32}$$

where σ_R is a $N \times 1$ vector collecting the standard deviations of N test assets. So (1.A.32) together with (1.A.27) implies

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

which violates (1.A.26). Hence there is a contradiction between $\hat{b}_i \neq \hat{b}_j$ and (1.A.27). So we must have

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

which completes the proof. □

1.A.1.5 Proof of corollary 1.2.1

Proof. The proof of corollary 1.2.1 follows the same method as in Appendix 1.A.1.4, except we choose a special vector for u where we set $u_i = 1$, $u_j = 1$, $u_k = 0$, for $k \neq i, j$. The rest of the proof follows trivially. □

1.A.2 Solving the OWL optimization problem

This section follows similar arguments in Zeng and Figueiredo (2015) and explains how to use the proximal gradient descent algorithm to solve the optimization problem of the OWL estimator. The first subsection introduces the OWL proximal function which is used to compute the optimizer at each step. The second subsection outlines the fast-iterative-soft-thresholding-algorithm (FISTA) used to find the global optimizer, together with a backtracking line search condition which speeds up computation substantially.

1.A.2.1 OWL proximal function

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

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

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

$$\Omega_\omega(b) = \Omega_\omega(|b|), \tag{1.A.34}$$

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

$$||b - \text{sign}(b) \odot |x|||_2^2 \leq ||b - x||_2^2, \tag{1.A.35}$$

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

$$Prox_{\Omega_\omega}(b) = \text{sign}(b) \odot Prox_{\Omega_\omega}(|b|). \tag{1.A.36}$$

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

$$||P(x-b)||_2^2 = ||x-b||_2^2, \tag{1.A.37}$$

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

$$\Omega_\omega(b) = \Omega_\omega(Pb). \tag{1.A.38}$$

So (1.A.37) and (1.A.38) imply that (1.A.36) can be written as

$$Prox_{\Omega_\omega}(b) = \text{sign}(b) \odot P' \, Prox_{\Omega_\omega}(|b|_\downarrow), \tag{1.A.39}$$

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

For any $x^* \in \kappa \subset R^n$, $x \in R^n$ and $x^* = x_\downarrow$ (i.e., x^* and x are two vectors having the same elements but with possibly different ordering of elements), since $|b|_\downarrow \in \kappa$, we have $|b|'_\downarrow x \leq |b|'_\downarrow x^*$. Therefore,

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

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

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

which is the projection of $(|b|_\downarrow - \omega)$ onto κ [21]. Then equation (1.A.39) can be written as

$$Prox_{\Omega_\omega}(b) = \text{sign}(b) \odot P' \, Proj_\kappa(|b|_\downarrow - \omega)), \tag{1.A.40}$$

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

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

After obtaining (1.A.40), we can employ the iterative soft-thresholding algorithm to find the global optimizer of (1.7). First, we initialize $b^{(0)}$,[22] then repeat

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

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

1.A.2.2 FISTA algorithm

The FISTA-OWL algorithm[23] below is based on Zeng and Figueiredo (2015). Fast computation is achieved by using the backtracking line condition (step 7) and the acceleration in u (step 12). The backtracking line condition allows large step sizes if optimizer stays in the right direction, otherwise shrinks step sizes. Steps 11 to 12 accelerate computation by moving the optimizer further towards the global optimizer at early iterations, while this acceleration diminishes when approaching the global optimizer.

Algorithm 1: FISTA-OWL

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

[22] For instance, we use the OLS estimate of (1.5) as initialization but it can be any random vector, which will result in the same global minimizer for b since (1.7) is a convex minimization problem. However, a good choice of initialization $b^{(0)}$ can reduce computation time greatly.

[23] In the initialization step of the FISTA-OWL algorithm below, L is a Lipschitz constant.

1.A.3 Motivating the "restricted eigenvalue condition"

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

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

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

Lemma 1.A.3. *Let S be the sparsity parameter and $\delta = \max_{1 \leq i,j \leq N} |\Sigma_{i,j} - \hat{\Sigma}_{i,j}|$. Let $\phi_{\hat{\Sigma}}^2$ and ϕ_{Σ}^2 be defined as in (1.A.1). Then for any vector $b \in R^K \setminus \{0\}$ that satisfies $||b_{s_0^c}||_1 \leq 3||b_{s_0}||_1$, it holds*

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

Proof. It is easy to show that

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

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

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

Rearranging the above inequality and using the norm property $\sqrt{S}||b_{s_0}||_2 \geq ||b_{s_0}||_1$, we have

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

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

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

□

Lemma 1.A.3 shows that for $\hat{\Sigma}$ to satisfy the restricted eigenvalue condition (1.A.1), i.e. $\phi^2_{\hat{\Sigma}} > 0$, it suffices to show that δ is small enough so that $\phi^2_{\Sigma} - 16S\delta > 0$, or that the Gram matrix $\hat{\Sigma}$ is close to a positive definite matrix Σ. The following lemma shows that the "Restricted eigenvalue condition" implies the compatibility condition (1.A.2) used in Buhlmann and Van de Geer (2011) (pp. 106).

Lemma 1.A.4 (Compatibility condition)**.** *Let* $\phi^2_0 := \phi^2_{\hat{\Sigma}}$*, if the scaled Gram matrix* $\hat{\Sigma}$ *satisfies* (1.A.1)*, then*

$$||b_{s_0}||_1^2 \leq (b'\hat{\Sigma}b)S/\phi_0^2. \tag{1.A.2}$$

Proof. From the definition of restricted eigenvalues, we have

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

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

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

Rearranging, we obtain the compatibility condition in Buhlmann and Van de Geer (2011)

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

□

1.A.4 Robustness check

In this section, we 1) check whether liquidity related factors are robust in explaining the cross section of asset returns; 2) investigate how small stocks affect estimation results of priced factors; 3) look into the convergence property of the FISTA-OWL algorithm; 4) compare estimation errors of all candidate methods using simulated data.

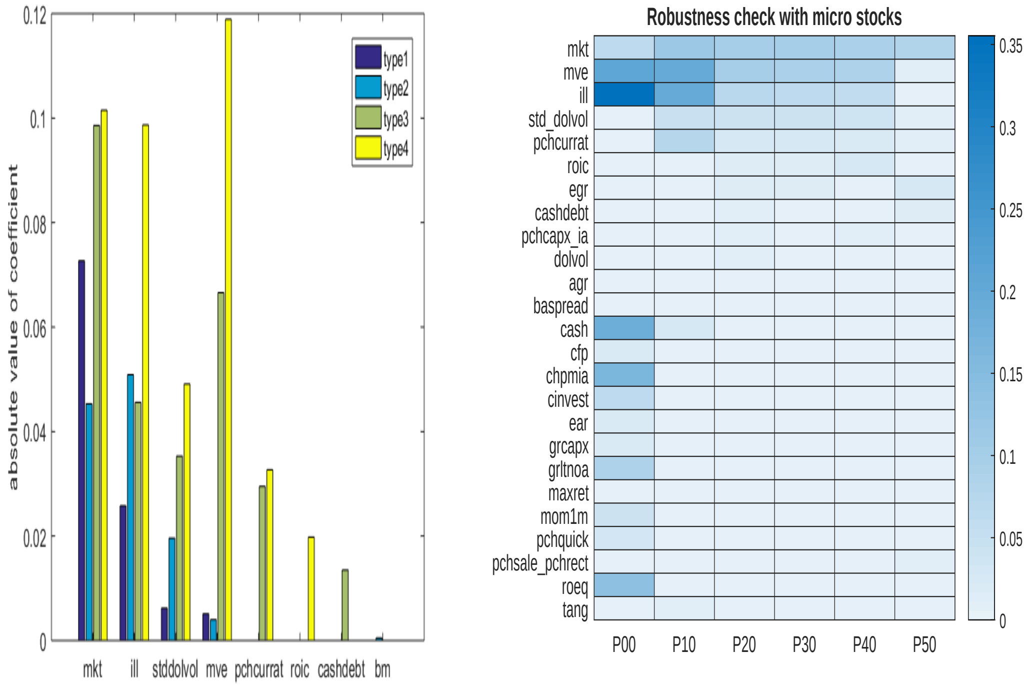
1.A.4.1 Robustness check with alternative test assets: are 'liquidity' factors robust?

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

Figure 1.5a reports the two-stage estimation procedure result using four different sets of test assets (including the one used in empirical analysis). First, 'market' along with 'illiquidity' and 'standard deviation of dollar volume' are consistently chosen as the most important factors to drive asset prices, with 'illiquidity' topping the chart of anomaly factors. Second, the impact of 'size' factor (mve) on test assets decreases colossally once it is not singled out to form bi-variate sorted portfolios. We can conclude that in 'type3' and 'type4' where 'size' effect tops the chart, it is artificially caused by portfolio sorting methods. However in empirical analysis ('type4'), 'size' is not a competing factor. Third, although singling out 'size' to form bi-variate sorted portfolios may alter the 'size' effect, it does not alter other factors' implications: liquidity related factors are primary factors driving asset prices.

1.A.4.2 Robustness check with micro stocks

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



(a) Robustness check with test assets

(b) Robustness check with micro stocks

Figure 1.5. Robustness check

Figure 1.5a reports the absolute value of SDF coefficients estimated by OWL using four types of test assets. Figure 1.5b reports the OWL estimates with six different treatments of micro stocks.

Figure 1.5b reports the heat map of the estimated risk prices using the OWL estimator while controlling stock sizes. First, micro stocks alter the market factor's interpretation drastically. When micro stocks are all included to form test portfolios, market factor only plays a moderate role for asset prices; however, liquidity related factors dominate the chart. Market factor nonetheless consistently becomes the primary factor to drive asset prices once micro stocks are removed (at P20 and above levels). Second, liquidity related factors consistently top the chart in driving asset prices, particularly with the inclusion of small stocks. It shows that small firms face severe liquidity constraints, and investors demand risk premiums to bear that risk. Third, to be consistent with the finance literature, we consider the typical 20 percentile cut-off level to remove micro stocks. In this case, profitability and growth related factors, after liquidity related factors, become the second tier of factors that drive asset prices.

1.A.4.3 Convergence using FISTA-OWL algorithm

Figure 1.6 shows the convergence of FISTA-OWL with backtracking algorithm (see Appendix 1.A.2) in the empirical analysis using 81 factors (80 anomaly factors plus the market factor). Vertical axis shows the distance between the k^{th} estimation and the optimizer. Horizontal axis shows the number of iterations (steps) until a stopping criterion is met. Following the machine learning literature (see Zeng and Figueiredo (2015)), we set a tight stopping criterion which is $\frac{||b(k)-b(k-1)||_2}{||b(k)||_2} < 10^{-6}$, $b(k)$ is the OWL estimation of the risk price at the k^{th} iteration. This figure shows that FISTA-OWL algorithm has a sound convergence property: it converges quickly at the first 1000 steps, then it gradually converges to the optimizer because of a tight stopping criterion.

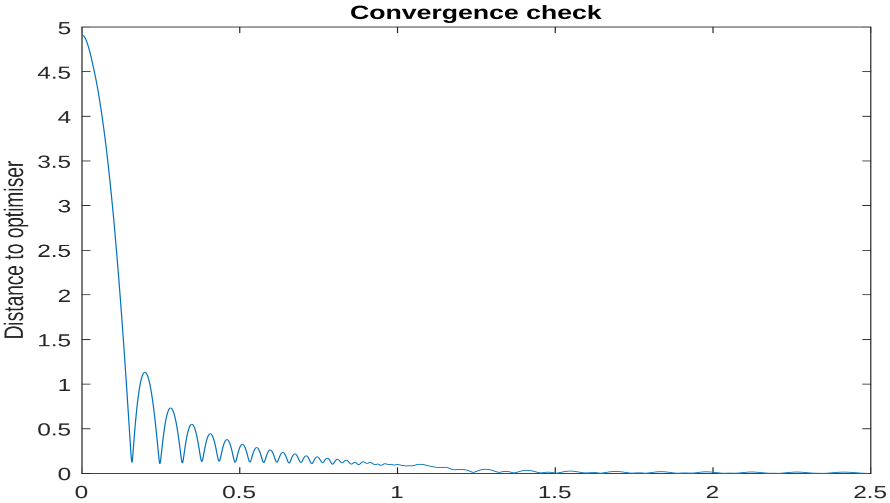


Figure 1.6. Convergence check for the FISTA-OWL algorithm
The stopping criterion is $\frac{||b(k)-b(k-1)||_2}{||b(k)||_2} < 10^{-6}$, where k is the number of iterations and $b(k)$ is the OWL estimate of risk price at the k^{th} iteration.

1.A.4.4 Monte Carlo Simulation

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

Figure 1.7a shows the estimation error of each method with multiple repetitions. We repeat the first experiment five times, in which we consider 90 candidate factors

$(K = 90)$ and 100 test assets $(N = 100)$.[24] First of all, the patterns are consistent among the repeated exercises. LASSO is the worst performer especially when factors are correlated. Elastic Net does improve the performance of LASSO when factors are correlated, while yielding similar results to LASSO when factors are not correlated. OLS is an unbiased estimator yet it does not produce sparsity. Adaptive LASSO was influenced by OLS failing to shrink most useless factors to zero but performs the best in the uncorrelated setting. OWL, by contrast, is the best performer when factors are correlated, and in the uncorrelated setting, though outperformed by adaptive LASSO, it is substantially better than LASSO and Elastic Net.

In the second experiment, which represents a typical low-dimensional setting $(N = 1000, K = 90,\ N \gg K)$, Figure 1.7b plots the estimation errors of each estimator with five repetitions. It shows a similar pattern to Figure 1.7a. The OWL estimator performs the best when factors are correlated. However, in this low-dimensional world, the adaptive LASSO estimator using OLS estimate as adaptive weight is the best candidate when factors are not correlated. LASSO and Elastic Net are the worst performers.

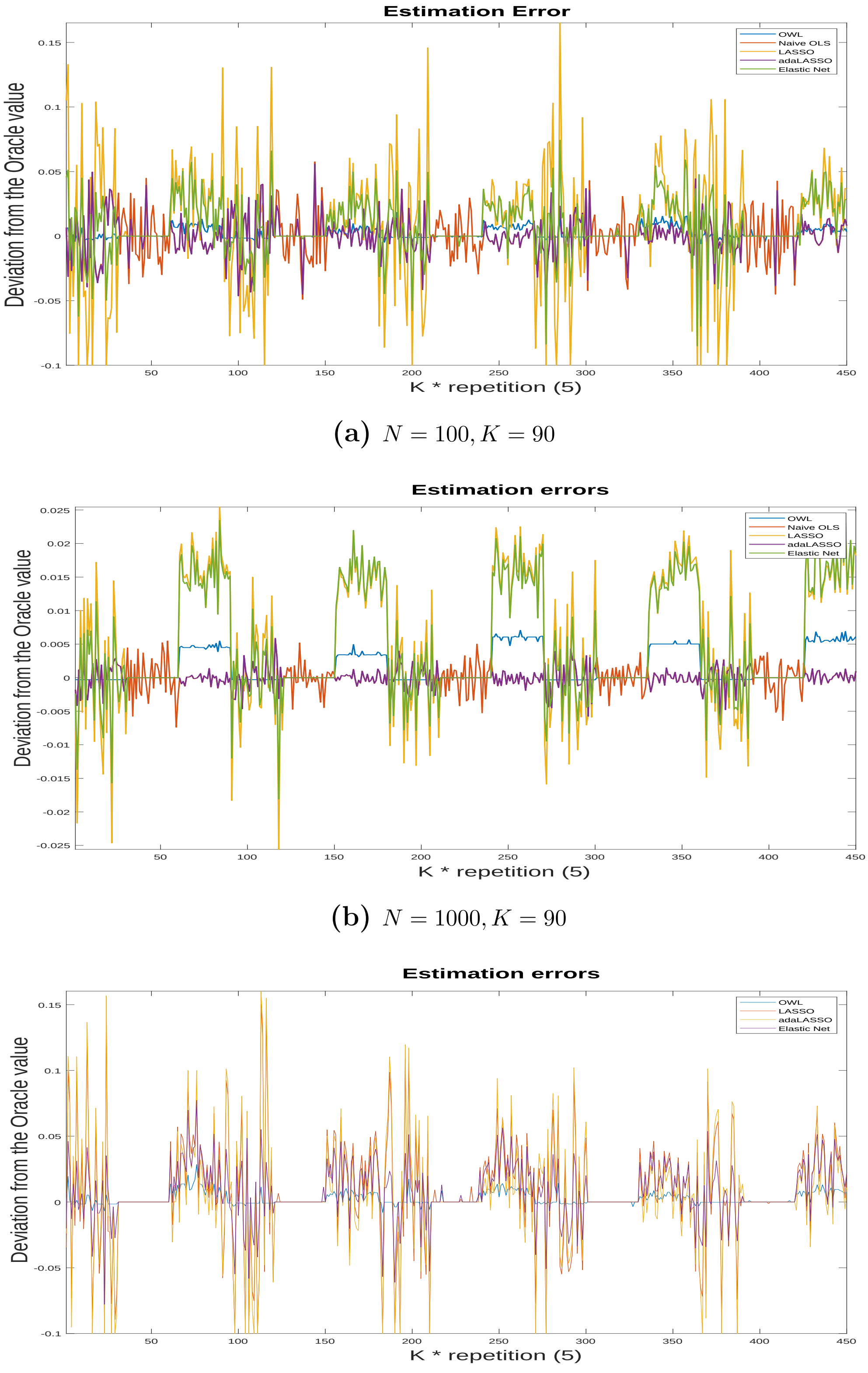
In the third experiment, which represents a high-dimensional setting $(N = 70, K = 90,\ N < K)$, Figure 1.7c plots estimation errors of each estimator with five repetitions. The best performer is OWL followed by Elastic Net and LASSO. The worst performer is the adaptive LASSO estimator using the LASSO estimate as the adaptive weight.

1.A.5 Introduction of LASSO, adaptive LASSO, Elastic Net and OSCAR

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

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \quad \left[\frac{1}{2} ||y - X\beta||^2 + \lambda ||\beta||_1 \right], \tag{1.A.3}$$

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



(a) $N = 100, K = 90$

(b) $N = 1000, K = 90$

(c) $N = 70, K = 90$

Figure 1.7. Simulation: estimation errors

This figure shows the estimator error of each method, measured by the distance between the oracle value and estimators. We repeat the simulation 5 times.

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

$$\hat{\beta}_{EN} = \arg\min_{\beta} \quad \left[\frac{1}{2}||y - X\beta||^2 + \lambda\alpha||\beta||_1 + \lambda(1-\alpha)||\beta||_2^2\right], \qquad (1.A.4)$$

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

$$\hat{\beta}_{adaLASSO} = \arg\min_{\beta} \quad \left[\frac{1}{2}||y - X\beta||^2 + \lambda \sum_{i=1}^{K} \frac{1}{|\hat{\beta}_{i,ada}|^{\gamma}} |\beta_i|\right], \qquad (1.A.5)$$

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

$$\hat{\beta}_{OSCAR} = \arg\min_{\beta} \quad \left[\frac{1}{2}||y - X\beta||^2 + \lambda_1||\beta||_1 + \lambda_2 \sum_{i<j} \max\{|\beta_i|, |\beta_j|\}\right], \qquad (1.A.6)$$

where $\sum_{i<j} \max\{|\beta_i|, |\beta_j|\}$ compares all elements in β pair-wisely and penalizes more on the larger one. Bondell and Reich (2008) show that OSCAR method encourages factor clustering when they are correlated.

Chapter 2

Robust Inference of the Ordered-Weighted-LASSO Estimator

2.1 Introduction

Economic and financial research topics related to the LASSO (Tibshirani, 1996) estimator have burgeoned and evolved rapidly in the past decade as high-dimensional big datasets become more available. For some examples, see Feng et al. (2020), Freyberger et al. (2020), Kozak et al. (2020) among others. However, as pointed out by Babii et al. (2019): *"...the bulk of machine learning methods assume i.i.d. regressors and residuals."*. They further argue that time series data are usually correlated and, as a remedy, they utilize a structured group-LASSO estimator using mixed frequency time series data.[1] Nonetheless, empirical evidence has suggested that correlations are also commonly observed in the cross-sectional dimension,[2] yet we often encounter insufficient information to impose structural restrictions on cross-sectional covariates. Consequently, it is not straightforward to implement the group-LASSO method while the cross-sectional dimension is large and potentially highly correlated. Conversely,

[1] In particular, each group consists of lagged values of either the dependent variable or a single explanatory variable, which means in effect, correlations on the time-series dimension are all retained in separate groups.

[2] Asness et al. (2013) find negative correlation between value and momentum factors which can be utilized to achieve superior portfolio performance. Kleibergen (2009) cautions about the collinearity between factor loadings when implementing a Fama-MacBeth regression.

we resort to a newly developed machine learning tool, the Ordered-Weighted-LASSO (OWL) estimator, which is structure-free (needless to define group structures ex ante) and entirely data-driven to exploit cross-sectional correlations. Figueiredo and Nowak (2016) demonstrated that the OWL estimator explicitly permits correlations among covariates and achieves correlation identification and sparsity shrinkage simultaneously. Sun (2019) further established the consistency property of the OWL estimator under i.i.d Gaussian assumptions and applied the OWL estimator to dissect the factor zoo.

This paper focuses on developing robust inference of the OWL estimator under more general conditions. First, we relax the usual i.i.d. assumption for regressors and instead impose less restrictive weak dependence conditions among high dimensional covariates before we derive the non-asymptotic bounds for the prediction error and the parameter estimation error. In particular, we assume $\alpha-$mixing conditions and potentially fatter (than sub-Gaussian) tails on variables and their distributions. We leave a free parameter q that controls the fatness of the tail distribution and we derive the probability measure of the validity of the oracle inequality in relation to q. Furthermore, we do not rely on an upper bound assumption for any random variable, which is usually required before implementing a Bernstein type inequality. Instead, we follow Dendramis et al. (2019) to truncate random variables at a level which will be specified later to bring together a refined bound for Bernstein type inequality under strong mixing conditions. In this respect, our theoretical framework requires much less restrictive assumptions and explicitly allows researchers to investigate cross-sectional correlations.

Second, following recent development of the de-sparsified LASSO estimator, for instance see Van De Geer et al. (2014), Belloni and Chernozhukov (2012), Kock (2016), Caner and Kock (2018), Kock and Tang (2019) among others, we extend Figueiredo and Nowak (2016) and Sun (2019) to develop the de-biased OWL estimator using the nodewise LASSO technique. The OWL estimator has appealing properties of grouping together highly correlated variables without pre-specifying any factor structures. Although Sun (2019) shows that the OWL estimator is consistent under some regularity conditions, it is biased in small samples. The de-biased OWL estimator bridges that gap. We show that after bias-correction, it is asymptotically normal and we derive the confidence intervals for each parameter.

Empirically, we apply the de-biased OWL estimator on 15 large stocks in the Dow Jones industrial average index with 80 factors constructed using accounting data. We implement a portfolio sorting method to obtain our factor zoo library.[3] It is worth stressing that we are not implementing a two-pass Fama-MacBeth type of regression or a stochastic discount factor (SDF) method[4] to identify true factors that drive asset prices, which are most commonly studied in the cross-sectional asset pricing literature. Instead, this exercise focuses on forecasting. We implement a simple one-pass time series regression to predict stock returns directly from lagged values of factors, which are high dimensional and potentially correlated.[5] We are interested in whether the de-biased OWL estimator can outperform other benchmarks in an out-of-sample framework in terms of predicting asset returns given a set of test assets. Empirical evidence suggests that the de-biased OWL estimator yields higher out-of-sample Sharpe ratios compared to standard LASSO and OLS methods. In addition, the de-biased OWL estimator illustrates a clear pattern of time-varying nature of factor selections during different periods, while LASSO and OLS do not show strong evidence of such pattern.

This paper builds naturally on the active and expanding literature pertaining to the LASSO estimator, in both the machine learning and empirical asset pricing literature. Tibshirani (1996) proposes the LASSO estimator that achieves efficient dimension reduction within a convex optimization problem, which enjoys huge success. Since then voluminous research has evolved to broaden the scope of the LASSO estimator. Yuan and Lin (2006) allow covariates sharing similar characteristics to be grouped together as a unit and propose the group LASSO estimator that performs sparse selection among groups. Freyberger et al. (2020) apply the adaptive group LASSO method to find pervasive firm characteristics to predict stock returns while Babii et al. (2019) implement the group LASSO estimator with mixed-frequency time

[3] In particular, we sort stocks (after removing micro-stocks) from the NYSE, NASDAQ and AMEX into decile portfolios according to a large number of firm characteristics at each point of time. For each characteristic we compute the spread returns between the top and bottom decile portfolios at each point of time.

[4] See Sun (2019) for an example of implementing the SDF method to find pervasive factors on the cross-section of stock returns.

[5] Nonetheless, the de-biased OWL estimator can also be implemented for the Fama-Macbeth regression (or SDF method) to identify pricing factors for a universe of stocks.

series data for nowcasting GDP growth. Belloni and Chernozhukov (2012) and Belloni et al. (2014) propose the three-pass double LASSO estimation method to de-bias LASSO coefficients of a set of factors that are of primary interest to researchers. Feng et al. (2020) adopt the double LASSO selection procedure to "tame" the factor zoo. Zou and Hastie (2005) combine the ℓ_1 and ℓ_2 norm regularization and propose the elastic net (EN), which stabilizes LASSO coefficients especially when covariates exhibit correlations. Kozak et al. (2020) employ EN in a Bayesian framework and find that sparse components can largely explain the cross section of average returns. Bondell and Reich (2008) propose the octagonal shrinkage and clustering algorithm for regression (OSCAR) method by exploring the ℓ_∞ norm of parameters pair-wisely to achieve clustered selections when covariates are highly correlated. Zeng and Figueiredo (2015) and Figueiredo and Nowak (2016) promote the Ordered-Weighted-LASSO (OWL) estimator, which is closely related to the SLOPE (Sorted ℓ_1 Penalized Estimator) by Bogdan et al. (2015): both assign a fixed and decreasing weighting vector to penalized coefficients (by contrast, LASSO estimator assigns the same penalty to all coefficients), with the larger coefficients (absolute value) receiving larger penalty. Bogdan et al. (2015) continue to specify a normal CDF based (non-linear) design for the decreasingly ordered weighting vector, before using the false discovery rate (FDR) to infer significance in the multi-testing framework assuming i.i.d. covariates. On the other hand, the OWL estimator, although having the same design in the regularization as the SLOPE, differs substantially in the weighting vector specification. Figueiredo and Nowak (2016) specify a *linear* weighting vector, and they further find that, by adopting a linear weighting vector, the OWL estimator encompasses the OSCAR regularization, which has appealing properties to group together highly correlated variables without imposing any structural restrictions ex ante. Van De Geer et al. (2014) developed the de-sparsified LASSO estimator using the nodewise LASSO technique, which enables them to find a way to approximate the usually un-invertible scaled Gram matrix to identify and quantify the bias of the LASSO estimator. The de-sparsified LASSO estimator enjoys asymptotic normality. Kock (2016), Caner and Kock (2018) and Kock and Tang (2019) expand the de-sparsified LASSO estimator on panel data and develop statistical properties under sub-Gaussian assumption. Babii et al. (2019) extend the nodewise LASSO technique to group-LASSO estimator using mixed frequency time-series data. This paper marries the OWL estimator and the nodewise LASSO technique to propose the de-biased version of the OWL estimator. Meanwhile, this paper relaxes the usual i.i.d. and sub-Gaussian assumptions to derive (non)asymptotic properties of the estimator. In particular, we allow for weak dependence (α-mixing) between covariates and fatter (than sub-Gaussian) tails.

In the remainder of this paper, Section 2.2 outlines the OWL estimation framework and we study its (non)asymptotic properties and further discuss a de-biased version of the OWL estimator and its asymptotic normality property. Section 2.3 studies Monte Carlo experiments with various settings in dimensions and correlations. Section 3.4 applies the de-biased OWL estimator on 15 large stocks to find the best predictors from a factor zoo library constructed from accounting data.

2.2 Model

In this section, we define the Ordered-Weighted-LASSO (OWL) estimator and derive its theoretical properties under mixing and some other regularity assumptions. Then we develop the de-biased OWL estimator, and show that it has asymptotically normal distribution.

Notation

Throughout this paper, X is a $n \times p$ matrix, and y is a $n \times 1$ vector. We denote by $\hat{\Sigma} = \dfrac{1}{n}X'X$ the scaled Gram Matrix of X, while $\Sigma = E(\hat{\Sigma})$ is the expected (true) value of the scaled Gram matrix. For any $x, y \in R^n$, we denote $\|x\|_2 = (\sum\limits_{i=1}^{n} x_i^2)^{1/2}$, $\|x\|_1 = \sum\limits_{i=1}^{n} |x_i|$, $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$, $\|x\|_0$ the cardinality of x, and $x \odot y$ the Hadamard (point-wise) production of two vectors. For matrix $\mathbb{M} \in R^{n \times n}, \Lambda_{min}$ and Λ_{max} denote the smallest and largest eigenvalues of $\mathbb{M}$. For two sequences x_n and y_n, we write $x_n \asymp y_n$ if there exist $0 < a \leq b < \infty$, such that $ay_n \leq x_n \leq by_n$ and we write $x_n \lesssim y_n$ if $x_n \leq by_n$ for some $0 < b < \infty$. For any set s, s^c denotes the complimentary set. For two scalers p and q, $p \vee q := \max(p, q)$ and $p \wedge q := \min(p, q)$. For any $\beta = \{\beta_1, \cdots, \beta_p\} \in R^p$, we denote $|\beta|_\downarrow := (|\beta|_{[1]}, |\beta|_{[2]}, \cdots, |\beta|_{[p]})'$, where $|\beta|_{[1]} \geq |\beta|_{[2]} \geq \cdots \geq |\beta|_{[p]}$ and $|\beta|_{[j]}$ is the j^{th} element of $|\beta|_\downarrow$.

2.2.1 OWL estimator and the oracle inequality

Consider a linear model

$$y = X\beta^0 + \epsilon, \tag{2.1}$$

where $X := (X_1, \cdots, X_p)$ and $\beta^0 = (\beta_1^0, \cdots, \beta_p^0)'$. Note that in the high-dimensional case, we allow $p \gg n$ and $X_j's$ can be correlated for $j = 1, \cdots, p$. The OWL estimator $\hat{\beta}$ minimizes the objective function

$$\hat{\beta} = \arg\min_{\beta} \quad \left[\frac{1}{n}||y - X\beta||_2^2 + \frac{1}{n}\omega'|\beta|_{\downarrow}\right], \qquad \omega'|\beta|_{\downarrow} = \sum_{j=1}^{p} \omega_j |\beta|_{[j]}, \tag{2.2}$$

where $\omega = (\omega_1, \cdots, \omega_p)'$, $\omega_j = \lambda_1 + \lambda_2(p - j),\ j = 1, \cdots, p$ and $\lambda_1, \lambda_2 \geq 0$ are tuning parameters.

Zeng and Figueiredo (2015) have shown that the OWL estimator has sparsity selection and correlation identification properties. In particular, the tuning parameter λ_1 controls the overall level of penalty while λ_2 influences the grouping property: large (small) λ_2 encourages (discourages) correlated variables to be grouped together by assigning them with similar coefficients, see Figueiredo and Nowak (2016) for a detailed discussion. We want to stress here that we do not impose any factor structure restrictions in our model, for instance defining groups ex ante to encapsulate correlated variables. Correlation identification is entirely data-driven. On the other hand, the OWL penalty term encompasses the LASSO setup. Setting $\lambda_2 = 0$, the OWL estimator will collapse to the standard LASSO estimator. A gradient proximal algorithm can be implemented to solve the optimization problem in (2.2), see Sun (2019) for technical details.

Before we derive the statistical properties for the OWL estimator $\hat{\beta}$, we make the following assumptions which are the foundation for building the theoretical framework and add novelty to our contributions. Assumption 2.2.1 states restriction on random variables, including cross-sectional dependence and on tails of their distributions. Assumption 2.2.2 is a standard requirement for developing asymptotic theory for LASSO type estimators in high dimensions. Assumption 2.2.3 specifies some rates on s, n and p required to obtain consistent estimators.

Assumption 2.2.1 (Random variables, Dendramis et al. (2019)).

(a) $\{X_{i,j}\}$ *and* $\{X_{i,j}\epsilon_i\}$, $i = 1, \cdots, n,\ j = 1, \cdots, p$ *are* $\alpha-$*mixing sequences, which*

are not necessarily stationary. The mixing coefficients have property $\alpha_k \leq c_* \phi^k$, $c > 0$, $0 < \phi < 1$, $k \geq 1$,

(b) $\sup_{i,j} \mathbb{P}(|X_{i,j}| > a) \leq c_1 \exp[-c_2 a^q]$ *and* $\sup_{i,j} \mathbb{P}(|X_{i,j}\epsilon_i| > a) \leq c_1 \exp[-c_2 a^q]$ *for all* $a > 0$ *, for some* $q > 0$ *and* $c_1, c_2 > 0$ *which do not depend on* a, i, j,

(c) $\mathrm{E}(\epsilon_i | X_{i,j}) = 0$, *and* $\max_{i,j} \mathrm{E}(X_{i,j}^4) < \infty$.

Assumption 2.2.1(a) relaxes the i.i.d condition which is usually assumed on X_j in the bulk of LASSO related literature, for instance see Kock (2016), Van De Geer et al. (2014) and Belloni and Chernozhukov (2012). Instead, we allow variables X_j to be weakly dependent, i.e. $\alpha-$mixing. Furthermore, mixing condition permits heteroscedasticity which is typically exhibited in empirical data. Assumption 2.2.1(b) further specifies tail bounds of distributions of X_j and ϵ. Although we use an exponential type of bound, it allows tails to be fatter than in the sub-Gaussian case. The tail parameter q controls the fatness of the tails, and it encompasses the sub-Gaussian tail ($q = 2$) as a special case. Assumption 2.2.1(c) is a standard assumption stating that the error term is orthogonal to covariates, in other words $\{X_{i,j}\epsilon_i\}$ is a zero mean sequence. Note that we do not assume random variables to be bounded which is typically assumed when implementing a Bernstein type inequality. To this end, our assumptions are more general and less restrictive than many of those in the literature which typically consider sub-Gaussian i.i.d. random variables.

Assumption 2.2.2 (Restricted eigenvalue condition on $\hat{\Sigma}$, Bickel et al. (2009))**.** *Let* $s_0 \subset \{1, \cdots, p\}$ *be a subset and* $s := |s_0|$ *the cardinality of* s_0. *For* $\beta = \{\beta_1, \cdots, \beta_p\}$, *denote* $\beta_{s_0} := \beta_i \mathbf{1}\{i \in s_0, i = 1, \cdots, p\}$, $\beta_{s_0^c} := \beta_i \mathbf{1}\{i \notin s_0, i = 1, \cdots, p\}$, *so that* $\beta = \beta_{s_0} + \beta_{s_0^c}$. *We suppose that for all* β *such that* $\|\beta_{s_0^c}\|_1 \leq 3\|\beta_{s_0}\|_1$, $\hat{\Sigma}$ *satisfies the restricted eigenvalue condition*

$$\phi_0^2 = \min_{\substack{s_0 \subset \{1,\cdots,p\} \\ s<p}} \quad \min_{\substack{\beta \in R^p \setminus \{0\} \\ \|\beta_{s_0^c}\|_1 \leq 3\|\beta_{s_0}\|_1}} \frac{\beta' \hat{\Sigma} \beta}{\|\beta_{s_0}\|_2^2} > 0. \tag{2.3}$$

Assumption 2.2.2 is a cornerstone to many theoretical results related to LASSO estimation. First of all, it allows us to specify the approximate sparsity condition as follows: only for a subset s_0, the true parameter vector has non-zero values ($\beta_i^0 \neq 0 : \forall i \in s_0$), while the complement contains only zeros ($\beta_i^0 = 0 : \forall i \notin s_0$). The cardinality s of such subset s_0 does not need to be known ex ante nor its elements, though we restrict it so that $s \ll p$. The restricted eigenvalue condition implies the compatibility condition of Buhlmann and Van de Geer (2011) (see below Lemma 2.2.1), which is an essential element in the proof of Theorem 2.2.1.

Lemma 2.2.1 (Compatibility condition for $\hat{\Sigma}$, Buhlmann and Van de Geer (2011)**).** *If the scaled Gram matrix $\hat{\Sigma}$ satisfies the restricted eigenvalue condition in* (2.3)*, then for any β*

$$\|\beta_{s_0}\|_1^2 \leq (\beta' \hat{\Sigma} \beta) s / \phi_0^2.$$

Proof: see Appendix 2.A.1.4

Assumption 2.2.3 (Rates on n, p and s)**.** *Denote by $s := |s_0|$ the sparsity parameter indicating the number of non-zero elements in $\hat{\beta}$ as in* (2.2) *and s_j the sparsity parameter in* (2.15) *by regressing the j^{th} column of X on the remaining columns of X. For any $j \in \{1, \cdots, p\}$, we assume*

(a) $(s \vee s_j)\sqrt{\dfrac{\log p}{n}} = o(1)$,

(b) $s_j\sqrt{\dfrac{\log^2 p}{n}} = o(1)$.

Assumption 2.2.3 specifies some rates on n, p, s and s_j which lead to consistent estimators. The rate that is required in 2.2.3(a) is rather standard and similar to that used in Kock (2016) and Van De Geer et al. (2014). The other requirement in 2.2.3(b) is typically weaker than in Kock (2016).

2.2.1.1 Statistical properties

Next, we investigate some statistical properties for the OWL estimator. Theorem 2.2.1 establishes oracle inequality for the prediction error and parameter estimation error. The probability we obtained is based on the assumption of weak dependence. Its proof uses Bernstein type inequalities for $\alpha-$mixing variables obtained in Dendramis et al. (2019).

Theorem 2.2.1 (Oracle inequality)**.** *Suppose Assumption 2.2.1 and 2.2.2 hold. Set $\lambda_0 = \kappa\sqrt{\dfrac{\log p}{n}}$, where κ is a positive constant. Let $\dfrac{\lambda_1}{n} = 2\lambda_0$ and assume $\dfrac{\lambda_2}{n} = O_p(\dfrac{s \log p}{np})$. Suppose that for some $\delta > 0$, $p \lesssim n^\delta$.*

1. *Let $n, p \to \infty$. Then for sufficiently large κ,*

$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2 \lesssim 4\lambda_0\sqrt{s}/\phi_0 + \lambda_0\sqrt{2s\|\beta^0\|_1} \tag{2.4}$$

$$||\hat{\beta} - \beta^0||_1 \lesssim 8\lambda_0 s/\phi_0^2 + \lambda_0 s\|\beta^0\|_1, \tag{2.5}$$

   with probability at least $1 - c_0' p^{-\epsilon} \to 1$, for some $\epsilon > 0$, where c_0' is a positive constant which is independent on n and p.

2. *Let p be bounded. Then (2.4) and (2.5) hold with probability at least*

$$1 - pc_0\left[\exp(-\frac{c_1'}{4}\kappa^2\log p) + \exp\left(-c_2'\left(\frac{\kappa\sqrt{n\log p}}{2\log^2 n}\right)^{\zeta}\right)\right], \tag{2.6}$$

   where $\zeta = q/(q+1)$ and c_0, c_1', c_2' are some positive constants which are independent on n and p.

Proof: see Appendix 2.A.1.1.

Remark 1 Theorem 2.2.1 offers bounds for the prediction error $||X(\hat{\beta} - \beta^0)||_2/n$ and parameter estimation error $||\hat{\beta} - \beta^0||_1$ for the OWL estimator under strong mixing conditions. Once we further incorporate Assumption 2.2.3, we will derive consistency and the convergence rate for the OWL estimator. See Corollary 2.2.1.

Remark 2 We analyze the probability of (2.4) and (2.5) to hold under two scenarios. First, when $n, p \to \infty$, we find that those inequalities hold with probability tending to one once a sufficiently large κ is chosen. Second, when p is fixed, we find that the probability of (2.4) and (2.5) to hold converges to $1 - pc_0 \exp(-c_1''\kappa^2 \log p)$ as $n \to \infty$, where c_0 and $c_1'' = c_1'/4$ are some positive constants which depend only on the mixing coefficient α_k in Assumption 2.2.1. Then we need to select κ sufficiently large to ensure that $pc_0 \exp(-c_1''\kappa^2 \log p)$ is close to zero.

Remark 3 Our results on the probability measures are obtained under the general assumption of exponential decaying tails on random variable $z_{i,j} := X_{i,j}\epsilon_i$. If $\zeta = 2/3$ (i.e., $q = 2$), equation (2.6) encompasses the sub-Gaussian case, which is a frequent assumption in related literature, see Kock (2016) and Kock and Tang (2019) for example. In addition, it also accommodates fatter tails, i.e. $0 < q < 2$. However, when both p and n are bounded, the probability of (2.4) and (2.5) to hold depends also on the tail parameter q. The thinner is the tail of the distribution (i.e., q is larger) of the random variable $z_{i,j}$, the closer the probability in (2.6) is to one.

To this end, we want to emphasize that our results in Theorem 2.2.1 are based on less restrictive assumptions, where we allow for weak dependence between random variables $X_j's$ and we further relax the sub-Gaussian tail restriction where we leave a parameter q that controls the fatness of the tail distributions.

Corollary 2.2.1 (Convergence rate)**.** *Suppose Assumption 2.2.3 is satisfied and assume $n, p \to \infty$. Then for sufficiently large κ, with probability tending to one,*

$$\|\hat{\beta} - \beta^0\|_2 = O_p\left(\sqrt{\frac{s \log p}{n}}\right) = o_p(1), \qquad \|\hat{\beta} - \beta^0\|_1 = O_p\left(s\sqrt{\frac{\log p}{n}}\right) = o_p(1). \quad (2.7)$$

Proof: see Appendix 2.A.1.2.

Corollary 2.2.1 establishes the convergence rate in ℓ_1 and ℓ_2 norm of the OWL estimator $\hat{\beta}$. After specifying some growth rate for n and p in Assumption 2.2.3, we show that the OWL estimator is consistent.

2.2.1.2 Choice of penalty parameters

It is well recognized that the choice of penalty level has huge impact on the performance of LASSO type estimators. In the machine learning literature, cross-validation is the most commonly implemented method for choosing penalty parameters. However, cross-validation can be computationally expensive to implement, for instance, in a recursively estimated application.[6] Hence, it would be useful if we can infer an appropriate penalty level based on the statistical properties of the estimator. Belloni and Chernozhukov (2012) argue that we should choose a penalty level that is sufficiently large to cancel noises coming from estimation errors (i.e. $\mathbb{P}(\lambda_0 > 2\|X'\epsilon\|_\infty/n)$ is large), yet not overly large to write off signals from variables. To achieve that, we propose the rule of thumb about penalty choice below based on a similar argument to Belloni et al. (2012) but incorporating our unique setting for random variables (weak dependence and exponential tails).

[6] Taking the commonly used 10-fold cross-validation as an example, at each step of the recursive exercise (for instance, a rolling window estimation procedure) we need to split the sample into 10 folds, while holding one tenth of the sample as testing sample and the remaining as estimation sample to evaluate and test the model, then swap positions of testing/estimation samples to re-evaluate the model (10 times). Suppose we have two tuning parameters and we want to search for a best fit in a 5×5 grid, and suppose the rolling window requires T recursive estimations. Then the 10-fold cross-validation method would require the model to be run $5^2 * 10 * T$ times.

Proposition 2.2.1. *Let Assumption 2.2.1 be satisfied, $\Phi^{-1}(\cdot)$ denote the inverse of the standard normal distribution function. We propose the following values for tuning parameters λ_1 and λ_2 in (2.2).*

$$\frac{\lambda_1}{n} = \frac{4}{\sqrt{n}}\sigma^*(1+\frac{1}{\log n})^{1/2}\Phi^{-1}(1-\frac{\alpha}{2p}), \qquad \frac{\lambda_2}{n} = \frac{\lambda_1}{n}\frac{\sqrt{\log p}}{\sqrt{n}\ p}, \tag{2.8}$$

where we evaluate σ^* recursively similar to Algorithm A.1 in Belloni et al. (2012) and α is a significance level.

Proof: see Appendix 2.A.1.5.

Note that α is selected to ensure the probability that the penalty is large enough to cancel out noises is close to one, that is $\mathbb{P}(\lambda_0 > 2\|X'\epsilon\|_\infty/n) \geq 1-\alpha$. So a smaller value of α will result in a larger penalty level. Proposition 2.2.1 offers a guideline for penalty choices when cross-validation is too expensive to implement. Equation (2.8) suggests that the penalty level depends on four elements. First, the noise level σ^* affects penalty level. Large variance of the error term requires a higher penalty level to cancel out noises. We evaluate σ^* recursively: we first evaluate the model and obtain the residuals while setting $\sigma^* = 1$, then update σ^* with the empirical residual variance and re-evaluate the model. Second, large n reduces the penalty level. Note that the total penalty is determined by λ_1/n and λ_2/n in (2.2), so large n commands smaller values for λ_1/n and λ_2/n. From a different perspective, we can view that large n leads to smaller variance σ^2, which requires less penalty on parameters. Third, the dimension of covariates p dictates the optimal penalty level. Large p requires a higher level of penalty to shrink off more irrelevant variables. Fourth, the significant parameter α affects the penalty level as discussed above.

2.2.2 De-biased OWL estimator

Although Theorem 2.2.1 shows that the OWL estimator is consistent under some regularity conditions, it is biased in small samples. In this section, we discuss a bias-corrected version of the OWL estimator using the nodewise LASSO method introduced in Van De Geer et al. (2014). Then we develop the asymptotic normal approximation result for the de-biased OWL estimator.

2.2.2.1 Identifying the bias of the OWL estimator

For the convenience of expression, the OWL estimator defined in (2.2) can be written as

$$\hat{\beta} = \arg\min_{\beta} \left[\|y - X\beta\|_2^2/n + 2\omega'|\beta|_{\downarrow}/n \right], \tag{2.9}$$

where we extract 2 out of the weighting vector ω.[7] The first order condition of minimization of (2.9) gives

$$-X'(y - X\hat{\beta})/n + \omega \odot \hat{\tau}/n = 0, \qquad \hat{\tau} = \begin{cases} 1 & \text{if} \quad \hat{\beta} > 0 \\ [-1, 1] & \text{if} \quad \hat{\beta} = 0 \\ -1 & \text{if} \quad \hat{\beta} < 0. \end{cases} \tag{2.10}$$

where $\odot$ denotes the point-wise product of two vectors, and $\hat{\tau}$ is the definition of sub-gradient of $|\hat{\beta}|_{\downarrow}$. We further utilize the equality $y = X\beta^0 + \epsilon$ and $\hat{\Sigma} = X'X/n$. Then (2.10) can be written as

$$\hat{\Sigma}(\hat{\beta} - \beta^0) + \omega \odot \hat{\tau}/n = X'\epsilon/n. \tag{2.11}$$

Since $\hat{\Sigma}$ is not invertible when $p > n$, we are using a relaxed form $\hat{\Theta}$ suggested by Van De Geer et al. (2014) to approximate the unobservable Σ^{-1}, where Σ is the population value of $\hat{\Sigma}$. Suppose such $\hat{\Theta}$ exists. Then we can write

$$\hat{\beta} - \beta^0 + \hat{\Theta}\omega \odot \hat{\tau}/n = \hat{\Theta}X'\epsilon/n - \Delta/\sqrt{n}, \tag{2.12}$$

$$\Delta = \sqrt{n}(\hat{\Theta}\hat{\Sigma} - I)(\hat{\beta} - \beta^0), \tag{2.13}$$

where we will show later that $\hat{\Theta}X'\epsilon/n$ is asymptotically normal and the approximation error, Δ, is negligible. Then we obtain the de-biased OWL estimator

$$\hat{b} = \hat{\beta} + \hat{\Theta}\omega \odot \hat{\tau}/n = \hat{\beta} + \hat{\Theta}X'(Y - X\hat{\beta})/n, \tag{2.14}$$

where the second equation holds in view of (2.10). So the bias is identified as $\widehat{bias} = \hat{\Theta}\omega \odot \hat{\tau}/n = \hat{\Theta}X'(Y - X\hat{\beta})/n$. In the next subsection, we construct the required approximation $\hat{\Theta}$.

[7] Note that ω is exactly pinned down by λ_1 and λ_2 which can be determined according to (2.8). So for the convenience of expression, we keep the same notation here for ω.

2.2.2.2 Construction of $\hat{\Theta}$

We follow Van De Geer et al. (2014) and Kock (2016) and use the nodewise LASSO technique to obtain $\hat{\Theta}$. First, the nodewise LASSO estimator is defined as

$$\hat{\gamma}_j = \arg\min_{\gamma \in R^{p-1}} \left(\|X_j - X_{-j}\gamma\|_2^2/n + 2\lambda_j \|\gamma_j\|_1 \right), \tag{2.15}$$

where $\hat{\gamma}_j := \{\hat{\gamma}_{j,k} \,:\, j,k = 1,\cdots,p,\, k \neq j\} \in R^{p-1}$ is a row vector of the nodewise LASSO estimator by regressing X_j (the j^{th} column of matrix X) on X_{-j} (which denotes the remaining columns of X) with LASSO penalty λ_j. Define a $p \times p$ matrix $\hat{C}$ and a $p \times p$ diagonal matrix $\hat{T}^2$ as

$$\hat{C} := \begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}, \qquad \hat{T}^2 := \mathrm{diag}(\hat{\delta}_1^2, \hat{\delta}_2^2, \cdots, \hat{\delta}_p^2,), \tag{2.16}$$

where for $j = 1, \cdots, p$,

$$\hat{\delta}_j^2 = \|X_j - X_{-j}\hat{\gamma}_j\|_2^2/n + \lambda \|\hat{\gamma}_j\|_1. \tag{2.17}$$

Then $\hat{\Theta}$ is constructed by setting

$$\hat{\Theta} := \hat{T}^{-2}\hat{C}. \tag{2.18}$$

For a close consideration of whether $\hat{\Theta}$ is a good approximation of Σ^{-1}, see Appendix 2.A.2.

2.2.2.3 Inference on the de-biased OWL estimator

Denote $\Sigma_{X\epsilon} := \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)']$, $\hat{\Sigma}_{X\epsilon} := \frac{1}{n}\sum_{i=1}^{n}[(X_i'\hat{\epsilon}_i)(X_i'\hat{\epsilon}_i)']$ and $\Theta := \Sigma^{-1}$. For any $l \in \{1, ..., p\}$, let $\hat{\Theta}_l$ (Θ_l) be the l^{th} row of the $\hat{\Theta}$ (Θ) matrix, written as a column vector.

Theorem 2.2.2. *Let $\hat{b}$ and $\hat{\Theta}$ be defined as in (2.14) and (2.18), respectively. Then* *the following hold:*

$$\sqrt{n}(\hat{b} - \beta^0) = \hat{\Theta} X'\epsilon/\sqrt{n} + o_p(1), \tag{2.19}$$

$$\hat{\Theta}_l' X'\epsilon/\sqrt{n} \to \mathbb{N}(0, \Theta_l' \Sigma_{X\epsilon} \Theta_l), \tag{2.20}$$

Furthermore, a uniformly valid point-wise confidence interval based on the t-statistics for β_l^0 where $l = 1, \cdots, p$ is given by

$$[\hat{b}_l - C(\alpha, \hat{\Theta}_l, \hat{\Sigma}_{X\epsilon}), \quad \hat{b}_l + C(\alpha, \hat{\Theta}_l, \hat{\Sigma}_{X\epsilon})], \tag{2.21}$$

where $C(\alpha, \hat{\Theta}_l, \hat{\Sigma}_{X\epsilon}) = \Phi^{-1}(1 - \alpha/2)\sqrt{\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon} \hat{\Theta}_l / n}$ and α is the confidence level.
Proof: see Appendix 2.A.1.3.

Theorem 2.2.2 arrives at the asymptotic normality property for the de-biased OWL estimator $\hat{b}$ and allows uniformly valid test for β_l^0 (i.e. the confidence interval applies to all $l = 1, \cdots, p$). The confidence interval is derived through the t-statistics based on the asymptotically normal property of the de-biased OWL estimator $\hat{b}$. Alternatively, a related Wald test can be subsequently developed. However, in this paper, we focus on the t-statistics and using (2.21) for testing the significance of the de-biased OWL estimator in our empirical exercises.

Next, we investigate the performance of the de-biased OWL estimator using simulated data.

2.3 Simulation

This section studies the performance of the de-biased OWL estimator alongside other benchmark estimations using simulated data. First, let us consider a toy example of 300 test assets ($N = 300$) and 20 covariates ($K = 20$). The oracle (true) values of the first six coefficient parameters of covariates are non-zeros and the rest are all zeros. Specifically, we set $\beta_0 = \{10, \frac{10}{2}, \frac{10}{3}, ..., \frac{10}{6}, 0, 0, ..., 0\} \in R^{20}$. Variables are not correlated.

Figure 2.1 displays the plots of estimated coefficients using various methods, alongside the true values ($b0$, blue line). The shaded area is the confidence interval for the de-biased OWL estimator. First of all, we find the OWL estimator (red/circle) exhibits good sparse-selection property: it shrinks the coefficients of all useless factors

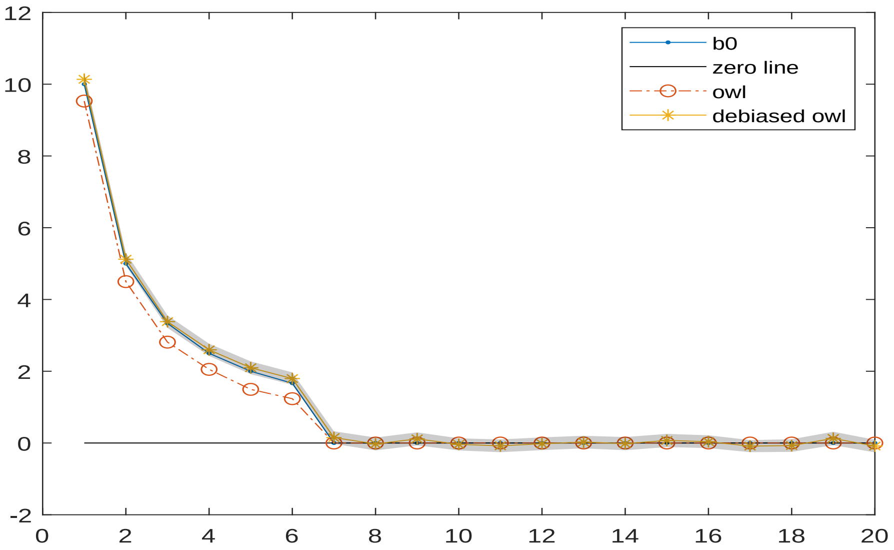


Figure 2.1. A toy example

This graph plots the estimated coefficients using OWL and its de-biased version, alongside the true values ($b0$, blue line). There are total 20 covariates, the first six (true value) are non-zeros, while the reminder are zeros. The shaded area is the confidence interval for the de-biased OWL estimator. Variables are uncorrelated.

to zeros. We also find that the OWL estimates for the non-zero coefficients are all biased towards zero, which is a common pitfall of many LASSO related estimators in small samples. On the other hand, we find that the de-biased OWL estimator (yellow/asterisk) *corrects* the bias: the bias-corrected estimates are much closer to the oracle values (blue line), with the oracle values lying inside the confidence interval (shaded area). On the flip side, the de-biased OWL estimates lose the sparse-selection property: all those useless factors now have non-zero coefficients using the de-biased OWL estimator. However, this incorrect de-biasing is bounded by the confidence intervals. We find that the true values (zeros) of the coefficients of these useless factors lie inside the confidence interval. Hence, we can easily remove those useless factors by running a t-test. This simple toy example illustrates the nice properties of the de-biased OWL estimator. Next, we run a sequence of Monte Carlo experiments to investigate how dimensions of data-set, correlations and other aspects would affect the performance of the de-biased OWL estimation.

We set the dimension of covariates X such that $K = dim(X) \in \{100, 1000\}$ and the number of observations $N \in \{60, 800, 1000\}$. We allow covariates in X to be correlated, and their covariance structure is defined as

$$\text{Corr}_{i,j}(X) = \Sigma_{i,j}(\rho) = \rho^{|i-j|}, \qquad i, j \in \{1, 2, ..., K\}, \qquad \rho \in \{0, 0.3, 0.5, 0.7\},$$

where Corr is the correlation coefficient function. The true oracle value for β is set to be

$$\beta_0 = \{10, \frac{10}{2}, \frac{10}{3}, ..., \frac{10}{6}, 0, 0, ..., 0\} \in R^K.$$

The first six elements are non-zeros, and the rest are zeros. The covariates matrix X and the response y are generated through the following distribution

$$X = Z * chol(\Sigma), \quad Z \sim \mathbf{N}(0, 1) \in R^{N\times K},$$
$$y = X\beta_0 + \epsilon, \quad \epsilon \sim \mathbf{N}(0, 0.01) \in R^{N\times 1},$$

where $chol(.)$ is the lower triangle matrix of the Cholesky decomposition. We use the de-biased OWL estimator to obtain estimated coefficients. The penalty hyper-parameters of λ_1 and λ_2 are chosen according to the optimal level discussed in Section 2.2.1.2.

$$\begin{aligned}\lambda_1/N =& \tilde{\sigma}(1 + \frac{1}{\log N})^{1/2}\Phi^{-1}(1 - \frac{\alpha}{2K})/\sqrt{N},\\ \lambda_2/N =& (\lambda_1/N)\sqrt{\log K}/(\sqrt{N}K),\end{aligned}$$

where $\Phi^{-1}(\cdot)$ is the inverse of a normal cumulative distribution function and $\alpha = 5\%$. We set $\tilde{\sigma} = 4\sigma^* = 0.01$ to gain computation speed.[8] We compare the de-biased OWL estimator with other benchmarks, including the OLS (when it is feasible) and the LASSO estimators. The number of the Monte Carlo repetition is 500 ($rep = 500$) for all set-ups. We report four estimated coefficients of $\hat{\beta}$, of which two have the true value of non-zeros: $\{\beta_3, \beta_6\}$, the other two have true values of zeros: $\{\beta_{12}, \beta_{20}\}$. We report the performance table of $\hat{\beta}$ using the following criteria:

1. Coverage rate for de-biased OWL. We compute the confidence interval of de-biased OWL according to (2.21). The coverage rate is the rate of the true value of the parameter included in the confidence interval throughout all Monte Carlo

[8] We opt for this easy choice of σ^* to gain computation speed, especially in high-dimensional cases. The de-biased OWL estimates may be sub-optimal, and a carefully cross-validated choice of σ^* can potentially improve the de-biased OWL estimates.

repetitions. We compute the coverage rate for each of these four parameters.

2. The width of confidence intervals (CI) for the de-biased OWL estimates. We compute the average width of confidence intervals of de-biased OWL throughout all Monte Carlo repetitions.

3. MAE (Mean Absolute Errors). We compare the mean absolute estimation errors between the de-biased OWL, LASSO and OLS estimates. The MAE for each coefficient $j \in \{3, 6, 12, 20\}$ is defined as $\text{MAE}^{j}_{benchmark} = \sum\limits_{i=1}^{rep} |\beta^{i}_{j,0} - \hat{\beta}^{benchmark,i}_{j}|/rep$, and the average MAE across all coefficients of $j \in \{3, 6, 12, 20\}$ for each benchmark is defined as $\text{MAE}_{benchmark} = \sum\limits_{i=1}^{rep} \sum\limits_{j} |\beta^{i}_{j,0} - \hat{\beta}^{benchmark,i}_{j}|/(4rep)$.

Table 2.1. Simulation results

Panel A: Coverage rate, CI width and MAE comparison between benchmarks

	Coverage rate of dowl				Width of CI of dowl				Average MAE			
	β_3	β_6	β_{12}	β_{20}	β_3	β_6	β_{12}	β_{20}	dowl	ols	lasso	lasso_cv
K = 50, N = 60												
$\rho = 0$	0.9360	0.9350	0.9600	0.9360	0.1016	0.0665	0.0820	0.0942	0.0112	0.0263	0.0819	0.0819
$\rho = 0.3$	0.9560	0.9300	0.9280	0.9320	0.1316	0.0948	0.1138	0.1424	0.0143	0.0347	0.0657	0.0657
$\rho = 0.5$	0.9560	0.9320	0.9420	0.9560	0.1209	0.1271	0.2372	0.1142	0.0154	0.0396	0.0894	0.0894
$\rho = 0.7$	0.9780	0.9780	0.9620	0.9500	0.1857	0.1782	0.1897	0.1504	0.0185	0.0495	0.0663	0.0663
K = 50, N = 1000												
$\rho = 0$	0.9420	0.9480	0.9380	0.9600	0.0129	0.0123	0.0127	0.0121	0.0015	0.0026	0.0689	0.0689
$\rho = 0.3$	0.9480	0.9600	0.9480	0.9540	0.0139	0.0139	0.0137	0.0137	0.0016	0.0028	0.0813	0.0813
$\rho = 0.5$	0.9640	0.9380	0.9280	0.9520	0.0158	0.0170	0.0162	0.0161	0.0019	0.0033	0.0758	0.0758
$\rho = 0.7$	0.9380	0.9600	0.9420	0.9400	0.0214	0.0207	0.0210	0.0211	0.0025	0.0044	0.0755	0.0755
K = 1000, N = 800												
$\rho = 0$	0.9080	0.9340	0.9400	0.9300	0.0939	0.1000	0.0907	0.0726	0.0131	N/A	0.0738	0.0738
$\rho = 0.3$	0.9460	0.9360	0.9280	0.9460	0.0823	0.0804	0.0996	0.0925	0.0105	N/A	0.0777	0.0777
$\rho = 0.5$	0.9620	0.9580	0.9460	0.9420	0.0878	0.0889	0.0832	0.0762	0.0096	N/A	0.0806	0.0806
$\rho = 0.7$	0.9720	0.9400	0.9400	0.9680	0.0756	0.0837	0.0882	0.0840	0.0089	N/A	0.0776	0.0776

Panel B: MAE comparison of each coefficient

	MAE_dowl				MAE_ols				MAE_lasso			
	β_3	β_6	β_{12}	β_{20}	β_3	β_6	β_{12}	β_{20}	β_3	β_6	β_{12}	β_{20}
K = 50, N = 60												
$\rho = 0$	0.0219	0.0173	0.0019	0.0036	0.0257	0.0243	0.0328	0.0223	0.1645	0.1629	0.0000	0.0000
$\rho = 0.3$	0.0266	0.0198	0.0048	0.0060	0.0405	0.0277	0.0276	0.0428	0.0562	0.2064	0.0000	0.0000
$\rho = 0.5$	0.0238	0.0266	0.0082	0.0029	0.0292	0.0313	0.0678	0.0299	0.1857	0.1721	0.0000	0.0000
$\rho = 0.7$	0.0337	0.0315	0.0043	0.0045	0.0488	0.0572	0.0492	0.0429	0.0576	0.2075	0.0000	0.0000
K = 50, N = 1000												
$\rho = 0$	0.0026	0.0026	0.0005	0.0003	0.0026	0.0026	0.0026	0.0024	0.1403	0.1352	0.0000	0.0000
$\rho = 0.3$	0.0028	0.0027	0.0004	0.0004	0.0028	0.0027	0.0028	0.0028	0.1445	0.1807	0.0000	0.0000
$\rho = 0.5$	0.0031	0.0036	0.0007	0.0004	0.0031	0.0036	0.0034	0.0032	0.1017	0.2014	0.0000	0.0000
$\rho = 0.7$	0.0045	0.0041	0.0007	0.0008	0.0045	0.0041	0.0044	0.0046	0.0603	0.2417	0.0000	0.0000
K = 1000, N = 800												
$\rho = 0$	0.0228	0.0231	0.0033	0.0030	N/A	N/A	N/A	N/A	0.1465	0.1488	0.0000	0.0000
$\rho = 0.3$	0.0164	0.0182	0.0044	0.0030	N/A	N/A	N/A	N/A	0.1392	0.1717	0.0000	0.0000
$\rho = 0.5$	0.0157	0.0174	0.0026	0.0027	N/A	N/A	N/A	N/A	0.0995	0.2231	0.0000	0.0000
$\rho = 0.7$	0.0130	0.0180	0.0032	0.0016	N/A	N/A	N/A	N/A	0.0783	0.2319	0.0000	0.0000

Panel A of Table 2.1 shows the results of the coverage rate and the confidence interval (CI) width of the de-biased OWL estimator, as well as the average MAE (mean absolute error) of each method. For the LASSO estimator we consider two methods for tuning the penalty parameter: one is by a ten-fold cross-validation (lasso_cv), which is widely used in machine learning literature; another one is by specifying the maximum number of non-zero coefficients we want to obtain.[9] We consider three settings in our experiment about the dimension of the dataset. First, we consider the case where $K = 50, N = 60$ $(N \approx K)$. Second, we look into the near asymptotic case where $K = 50, N = 1000$ $(N \gg K)$. Third, we investigate the high-dimensional case where $K = 1000, N = 800$ $(K > N)$. First of all, we find that the coverage rates of the de-biased OWL estimates for all cases are above 90%. In particular, the coverage rate for the near asymptotic case is near the correct size (95%) when correlation is not too high $(\rho < 0.5)$. Comparing coverage rates with different correlation profiles within each setting suggests that the coverage rate is typically higher when correlation is high $(\rho = 0.7)$. However, we find that this is a result of enlarged confidence interval width rather than improved estimation accuracy. The width of confidence interval at the near asymptotic case suggests that when the correlation coefficient increases (ρ increases from 0 to 0.7), the width of confidence interval enlarges, particularly when ρ changes from 0.5 to 0.7. Meanwhile, an increase in ρ is also associated with a decrease in estimation accuracy: the average MAE for the de-biased OWL estimate increases steadily when ρ increases. Also, comparing the average MAE of four coefficients $(\beta_3, \beta_6, \beta_{12}, \beta_{20})$ between the de-biased OWL, OLS and LASSO estimators, we find that the de-biased OWL estimate yields the lowest estimation errors in all cases.

Panel B of Table 2.1 gives a detailed illustration of MAE comparison between benchmarks for each coefficient. We find that the OLS estimator is good at estimating β_3 and β_6 because the OLS estimator is unbiased. However, the OLS estimation error is large when estimating β_{12} and β_{20} when their true values are zeros. The performance of the LASSO estimator is the opposite: it correctly shrinks β_{12} and β_{20} to zeros (in which case there is no estimation error for β_{12} and β_{20}) but the LASSO estimates for β_3 and β_6 are biased, and the estimation errors are large compared to the OLS estimates.

[9] We specify the maximum number of non-zero coefficients as ten to ensure sparse selection. After evaluation, we find both methods for choosing the LASSO penalty parameter tend to yield the same result.

The de-biased OWL estimate combines the merits of the OLS and LASSO estimators: it achieves unbiased estimation for the non-zero coefficients but also shrinks zero coefficients. In the cases where $K = 50$ (the OLS estimator is feasible), the de-biased OWL estimates for β_3 and β_6 are very close to the OLS estimates, especially in the near asymptotic case. Meanwhile, the de-biased OWL estimates for β_{12} and β_{20} are close to LASSO estimates, performing sparsity shrinkage for useless covariates (whose true coefficients are zeros). In the high-dimensional case where $K = 1000$, we find that the MAE of the de-biased OWL estimates are substantially smaller than those of the LASSO estimates while the OLS estimates become infeasible.

This Monte Carlo experiment shows that, in both the low- and high-dimensional cases, the de-biased OWL estimator delivers unbiased estimation for useful covariates (whose true coefficients are non-zeros) as good as the OLS estimator while shrinking off useless covariates almost as good as the LASSO estimator.

2.4 Empirical application on factor investing

In this section, we apply the de-biased OWL method to predict stock returns using firm-characteristic based factors. We first introduce the dataset and the empirical method before conducting the empirical analysis.

2.4.1 Data and empirical method

We use the U.S. stock data from the Center for Research in Security Prices (CRSP) and Compustat database, both downloaded from the Wharton Research Data Service. The data spans January 1980 to December 2017, totalling 456 months on all NYSE, AMEX and NASDAQ listed common stocks. Risk-free rate and market returns are downloaded from Kenneth French's on-line data library.[10] For predicting stock returns, we use a factor library which contains 80 anomaly factors constructed using characteristic-sorted portfolios. More details of constructing those anomaly factors can be found in Sun (2019). We consider 30 stocks in the Dow Jones Industrial Average index as test assets while deleting stocks having any missing data between January 1980 and December 2017, which leaves 15 stocks as test assets. We then use

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

these characteristic-based factors to predict stock returns for each of those 15 stocks.

Suppose we use the lagged factor returns to predict individual stock return. The predicted return of any stock i at time t is

$$\hat{R}^i_{t+1} = f_t\tilde{\beta}_t, \qquad \tilde{\beta}_t \in \{\tilde{\beta}^{dOWL}_t, \tilde{\beta}^{OWL}_t, \tilde{\beta}^{LASSO}_t, \tilde{\beta}^{OLS}_t\}, \tag{2.22}$$

$$\tilde{\beta}^{dOWL}_t = \tilde{\beta}^{OWL}_t + \hat{\Theta}(R^i_t - f_{t-1}\tilde{\beta}^{OWL}_t)/n, \tag{2.23}$$

$$\tilde{\beta}^{OWL}_t = \arg\min_{\beta} \|R^i_t - f_{t-1}\beta\|^2_2 + \Omega(\beta), \tag{2.24}$$

$$\tilde{\beta}^{LASSO}_t = \arg\min_{\beta} \|R^i_t - f_{t-1}\beta\|^2_2 + \lambda\|\beta\|_1, \tag{2.25}$$

$$\tilde{\beta}^{OLS}_t = \arg\min_{\beta} \|R^i_t - f_{t-1}\beta\|^2_2, \tag{2.26}$$

where $\tilde{\beta}_t$ includes the de-biased OWL ('dOWL') estimator as well as benchmarks such as the OWL, OLS and LASSO estimators. $\hat{\Theta}$ is constructed in (2.18). $\tilde{\beta}^{OWL}_t$ is the OWL estimator in (2.2) and $\tilde{\beta}^{dOWL}_t$ is the de-biased OWL estimator in (2.14). λ is a hyper parameter for the LASSO estimator and we use two methods to determine its value: either by a 10-fold cross-validation (CV) method or restricting its maximum non-zero coefficients to ten (DFmax = 10) to ensure sparsity.

2.4.2 A stock-by-stock analysis

In this subsection, we look at each stock and find which factors are the best predictors using the full sample estimation. However, it is worth stressing that our target here is to predict stock returns using a potentially large number of predictors. Since our approach is stock-specific, the selected factors for each stock should not be interpreted as cross-sectionally valid true factors. Cross-sectional stock returns are typically investigated through the Fama-MacBeth regression method or the SDF method, see Sun (2019) for more details on dissecting the factor zoo for cross-sectional asset returns.

Figure 2.2 shows the contour plot of the estimated de-biased OWL coefficients (absolute value).[11] The vertical axis lists all the factors considered in the factor library and the horizontal axis shows 15 stocks as test assets. The left panel displays the estimated coefficient before testing, while the right panel displays the contour

[11] Note that we excluded 'betasq' in the factor library because the correlation coefficient between 'beta' and 'betasq' is more than 0.9. Including both of them in the factor zoo leads to serious estimation problems for OLS and LASSO estimators. For that reason, we exclude 'betasq' from the factor library.

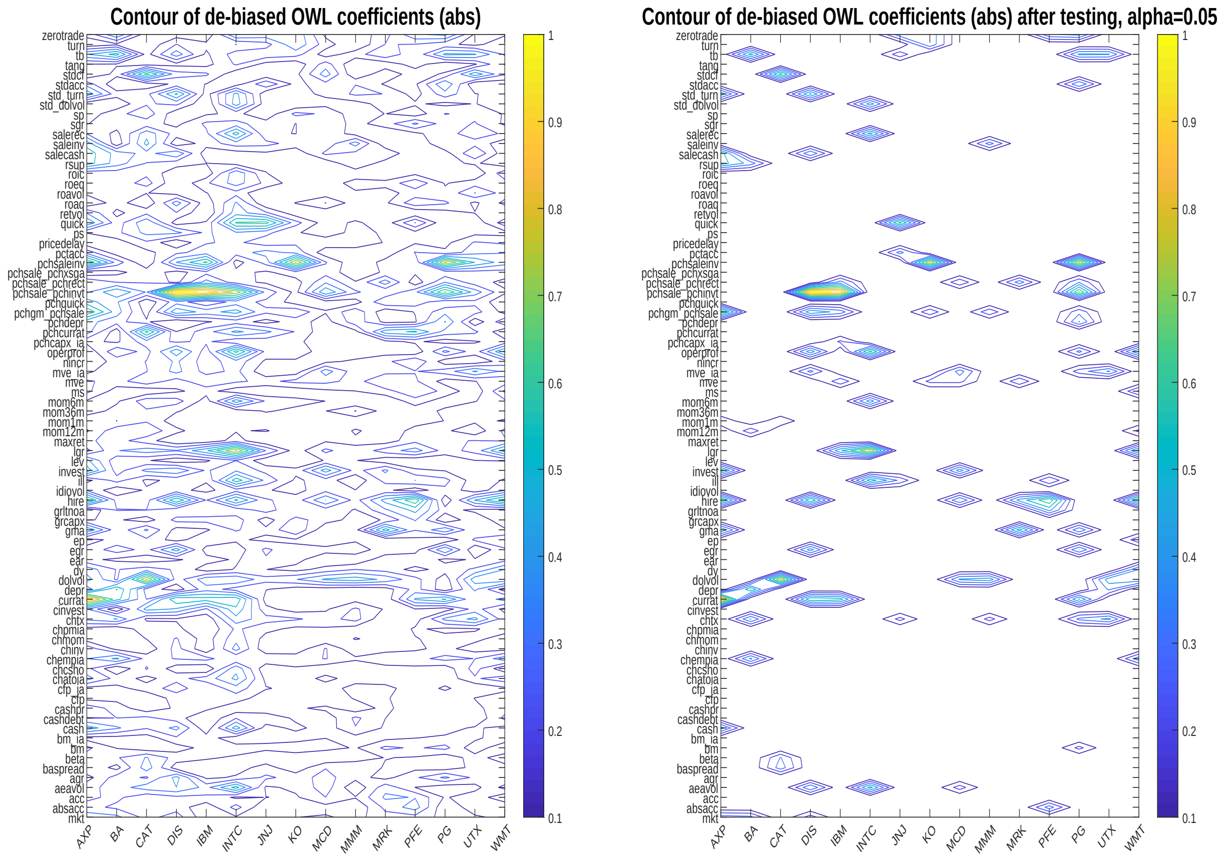


Figure 2.2. Contour plot of the de-biased OWL estimator of factor loadings
The left panel is before testing and the right panel is after testing. Yellow areas represent higher values, blue areas represent lower values and blank areas are zeros.

plot of the de-biased OWL estimate after removing insignificant ones by applying the confidence interval in (2.21) at a significance level $\alpha = 5\%$. We first find that 'sales' related factors are typically selected as strong predictors for many stocks, while 'profitability' and 'investment' related factors form the second tier of strong predictors for stocks returns. The right panel confirms that most of those strong predictors are significant while many other minor predictors are removed after applying the confidence interval. Meanwhile, it also suggests that some stocks, for instance 'KO' and 'MMM', are sensitive to only very few (less than five) factors in our factor library, while others like 'J&J' and 'DIS' have many (more than ten) significant predictors.

Next we choose a random stock, for instance 'DIS', to compare the estimation results using different methods. Figure 2.3 shows the plot of estimated coefficients using the de-biased OWL (blue), OWL (red), LASSO (yellow, with DFmax = 10), LASSO_CV (purple, with 10-fold cross validation) and OLS (green) estimators. The grey area displays the confidence intervals for the de-biased OWL estimator at a significance level $\alpha = 5\%$.

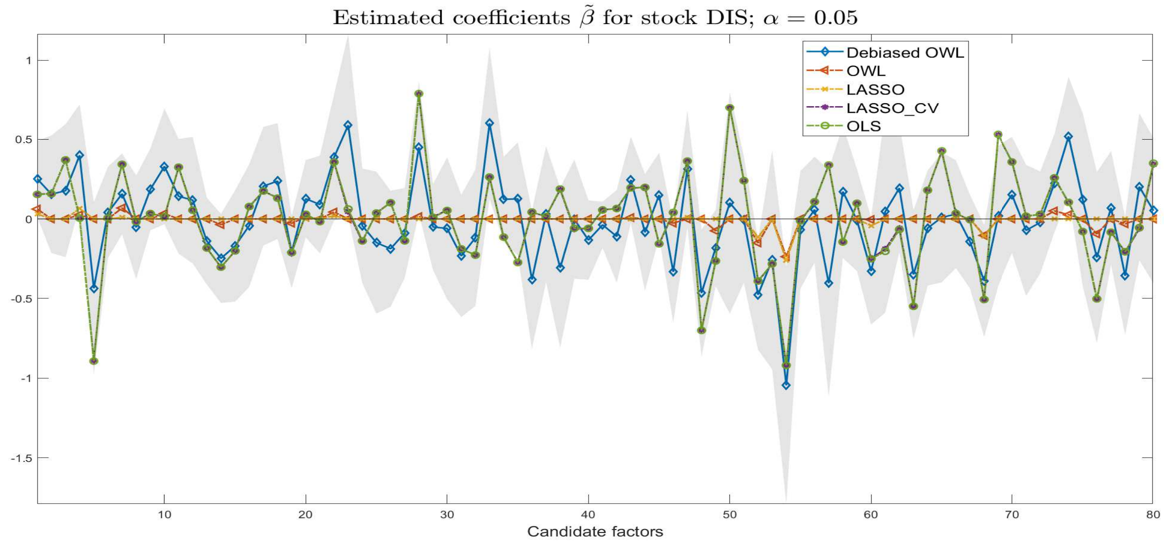


Figure 2.3. Estimated factor loadings of 'DIS'

This figure plot the estimated factor loadings using the de-biased OWL, OWL, LASSO and OLS estimators. The shaded area is the confidence interval for the de-biased OWL estimator.

Figure 2.3 shows that the OWL estimator yields very similar results to the LASSO estimator (with maximum number of non-zero coefficients restricted to ten to ensure sparsity) for the sparsity property, i.e., they both shrink many factors' coefficients to zeros, yet they differ in some of the survival factors (i.e., factors having non-zero estimated coefficients). Meanwhile, the estimated coefficients of survival factors of both the OWL and the LASSO (yellow) estimators are very close zero, which is caused by an inward bias pulling the coefficients towards zeros. The cross validated LASSO estimator yields very similar results to the OLS estimator. Cross validation method suggests all factors are useful to predict stock returns and thus shrink no factors and yield almost the same result as the OLS estimator. The de-biased OWL estimator corrects that bias for the OWL estimator. We find that after bias-correction, the de-biased OWL estimate displays a similar trend to the OLS estimator, although the magnitude of estimated coefficients varies on some factors compared to the OLS estimator. Meanwhile, the de-biased OWL estimator loses the sparsity property (i.e. no factors receive zero coefficients for the de-biased OWL estimator), but we find that many of those factors receiving zero coefficients in the OWL estimation are insignificant in the de-biased OWL estimation after applying the confidence intervals. In addition, to preserve the sparsity property of the OWL estimator while correcting the bias for survival factors, we can selectively de-bias these estimated non-zero coefficients of the OWL estimator.

2.5 Conclusion

In high dimensional datasets where covariates exhibit high correlations, Zou and Hastie (2005) and Figueiredo and Nowak (2016) have shown that the LASSO estimator performs poorly. Figueiredo and Nowak (2016) introduced the Ordered-Weighted-LASSO (OWL) estimator which is specifically tailored to deal with correlations between covariates. Sun (2019) introduced the OWL estimator to dissect the factor zoo for the cross sectional asset returns and further developed asymptotic properties for the OWL estimator. Although Sun (2019) shows that the OWL estimator is consistent, it is biased in small samples. This paper extends Figueiredo and Nowak (2016) and Sun (2019) to study the (non)asymptotic properties of the OWL estimator with *less restrictive* assumptions and further proposes a bias-corrected version of the OWL estimator. Monte Carlo experiments show that, in both the low- and high-dimensional settings, the de-biased OWL estimator delivers unbiased estimation for useful covariates as good as the OLS estimator, while shrinking off useless covariates almost as good as the LASSO estimator. In the empirical analysis, we implement the de-biased OWL estimation to predict returns for 15 stocks from the Dow Jones Industrial Average index using 80 factors. We find some ‘sales’, ‘profitability’ and ‘investment’ related factors are strong predictors for many stock returns.

2.A Appendix

2.A.1 Technical proofs

2.A.1.1 Proof of Theorem 2.2.1

Proof. The proof of Theorem 2.2.1 consists of two parts. In the first part we derive the oracle inequality (2.4) and (2.5) under the event E, which will be specified below in (2.A.4). In the second part we will derive the probability of this event $\mathbb{P}(E)$ to be true.

Part I. According to the "argmin" property,

$$\frac{1}{n}||y - X\hat{\beta}||_2^2 + \frac{1}{n}\sum_{j=1}^{p}\omega_j|\hat{\beta}|_{[j]} \leq \frac{1}{n}||y - X\beta^0||_2^2 + \frac{1}{n}\sum_{j=1}^{p}\omega_j|\beta^0|_{[j]}. \tag{2.A.1}$$

Since $(\omega_1, \cdots, \omega_p)'$ where $\omega_j = \lambda_1 + \lambda_2(p-j)$, $j = \{1, \cdots, p\}$ is in a monotone non-negative cone, so $\omega_1 \geq \omega_2 \geq ... \geq \omega_p$. Then we have

$$\sum_{j=1}^{p}\omega_j|\hat{\beta}|_{[j]} \geq \omega_p||\hat{\beta}||_1 = \lambda_1||\hat{\beta}||_1,$$
$$\sum_{j=1}^{p}\omega_j|\beta^0|_{[j]} \leq \omega_1||\beta^0||_1 = [\lambda_1 + \lambda_2(p-1)]||\beta^0||_1.$$

Together with $y = X\beta^0 + \epsilon$, this implies that (2.A.1) can be simplified as follow:

$$\frac{1}{n}||\epsilon - X(\hat{\beta} - \beta^0)||_2^2 + \frac{1}{n}\omega_p||\hat{\beta}||_1 \leq \frac{1}{n}||\epsilon||_2^2 + \frac{1}{n}\omega_1||\beta^0||_1 \tag{2.A.2}$$
$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta}||_1 \leq \frac{2}{n}\epsilon' X(\hat{\beta} - \beta^0) + \frac{1}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1. \tag{2.A.3}$$

Note that $\epsilon' X(\hat{\beta} - \beta^0) \leq \|\epsilon' X\|_\infty ||\hat{\beta} - \beta^0||_1$. Let $\lambda_0 > 0$ and consider an event

$$E := \left\{\frac{1}{n}\|\epsilon' X\|_\infty \leq \frac{\lambda_0}{2}\right\}, \tag{2.A.4}$$

where $\lambda_0 = \kappa\sqrt{\frac{\log p}{n}}$, where $\kappa > 0$ is a constant. Then in view of (2.A.4), (2.A.3) can be bounded as

$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta}||_1 \leq \lambda_0||\hat{\beta} - \beta^0||_1 + \frac{1}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1. \tag{2.A.5}$$

By assumption of the theorem, $\frac{\lambda_1}{n} = 2\lambda_0$. So we obtain

$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta}||_1 \le \frac{\lambda_1}{2n}||\hat{\beta} - \beta^0||_1 + \frac{1}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1. \tag{2.A.6}$$

By the definition of s_0, $\hat{\beta} = \hat{\beta}_{s_0} + \hat{\beta}_{s_0^c}$. Utilizing the triangle inequality $\|a\|_1 + \|b\|_1 \ge \|a+b\|_1$ for any vector a and b, we obtain

$$||\hat{\beta}||_1 = ||\hat{\beta}_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1 \ge ||\beta^0_{s_0}||_1 - ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1, \tag{2.A.7}$$

$$||\hat{\beta} - \beta^0||_1 = ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1. \tag{2.A.8}$$

Therefore, using (2.A.7) and (2.A.8), (2.A.6) can be written as

$$\begin{aligned}\frac{2}{n}||X(\hat{\beta} - \beta^0)||_2^2 &+ \frac{2\lambda_1}{n}(||\beta^0_{s_0}||_1 - ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1)\\ &\le \frac{\lambda_1}{n}(||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1) + \frac{2}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1.\end{aligned} \tag{2.A.9}$$

Note that $\|\beta^0_{s_0}\|_1 = \|\beta^0\|_1$, so (2.A.9) can be written as

$$\frac{2}{n}||X(\hat{\beta} - \beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta}_{s_0^c}||_1 \le \frac{3\lambda_1}{n}||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + \frac{2\lambda_2(p-1)}{n}||\beta^0||_1. \tag{2.A.10}$$

By (2.A.8), $||\hat{\beta}_{s_0^c}||_1 = ||\hat{\beta} - \beta^0||_1 - ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1$. Utilizing this in (2.A.10), we obtain

$$\frac{2}{n}||X(\hat{\beta} - \beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta} - \beta^0||_1 \le \frac{4\lambda_1}{n}||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + \frac{2\lambda_2(p-1)}{n}||\beta^0||_1. \tag{2.A.11}$$

Utilizing the compatibility condition $\|\beta_{s_0}\|_1^2 \le (\beta'\hat{\Sigma}\beta)s/\phi_0^2$ given in Lemma 2.2.1 on $||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1$ and using definition $\hat{\Sigma} = \frac{X'X}{n}$, we obtain

$$\begin{aligned}||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1^2 \le (\hat{\beta} - \beta^0)'\hat{\Sigma}(\hat{\beta} - \beta^0)s/\Phi_0^2 &= \|X(\hat{\beta} - \beta^0)\|_2^2 s/(n\Phi_0^2),\\ ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 &\le \|X(\hat{\beta} - \beta^0)\|_2\sqrt{s}/(\sqrt{n}\Phi_0).\end{aligned} \tag{2.A.12}$$

Therefore, applying inequality $4ab \le a^2 + 4b^2$, we obtain

$$\begin{aligned}\frac{4\lambda_1}{n}||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 &\le 4\left(\frac{\|X(\hat{\beta} - \beta^0)\|_2}{\sqrt{n}}\right)\left(\frac{\lambda_1}{n}\frac{\sqrt{s}}{\Phi_0}\right)\\ &\le \frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + 4(\frac{\lambda_1}{n})^2\frac{s}{\Phi_0^2}.\end{aligned}$$

So (2.A.11) can be written as

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2^2+\frac{\lambda_1}{n}||\hat{\beta}-\beta^0||_1 \leq 4(\frac{\lambda_1}{n})^2\frac{s}{\Phi_0^2}+\frac{2\lambda_2(p-1)}{n}||\beta^0||_1. \tag{2.A.13}$$

By assumption of the theorem, $\frac{\lambda_1}{n}=2\lambda_0 \asymp \sqrt{\frac{\log p}{n}}$, and $\frac{\lambda_2}{n} \lesssim \frac{s\log p}{np} \asymp \frac{s\lambda_0^2}{p}$. Therefore, (2.A.13) can be written as

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2^2+2\lambda_0||\hat{\beta}-\beta^0||_1 \lesssim 16\lambda_0^2 s/\Phi_0^2+2\lambda_0^2 s\|\beta^0\|_1. \tag{2.A.14}$$

Using $\sqrt{a^2+b^2}\leq a+b$, for all $a,b>0$, (2.A.14) implies

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2 \lesssim 4\lambda_0\sqrt{s}/\Phi_0+\lambda_0\sqrt{2s\|\beta^0\|_1}, \tag{2.A.15}$$

$$||\hat{\beta}-\beta^0||_1 \lesssim 8\lambda_0 s/\Phi_0^2+\lambda_0 s\|\beta^0\|_1. \tag{2.A.16}$$

This shows that (2.4) and (2.5) in Theorem 2.2.1 are valid, assuming that (2.A.4) holds.

Part II. Next we calculate $\mathbb{P}(E)$. We have

$$\begin{aligned}\mathbb{P}(E^c) &= \mathbb{P}(\frac{1}{n}\|X'\epsilon\|_\infty > \frac{\lambda_0}{2}) = \mathbb{P}(\frac{1}{n}\max_{j=1,\cdots,p}|\sum_{i=1}^n X_{i,j}\epsilon_i| > \frac{\lambda_0}{2}) \\ &\leq \sum_{j=1}^p \mathbb{P}(\frac{1}{\sqrt{n}}\sum_{i=1}^n |X_{i,j}\epsilon_i| > \frac{\lambda_0\sqrt{n}}{2}) = p\max_{j=1,\cdots,p}\mathbb{P}(\frac{1}{\sqrt{n}}\sum_{i=1}^n |X_{i,j}\epsilon_i| > \frac{\lambda_0\sqrt{n}}{2}).\end{aligned} \tag{2.A.17}$$

By Assumption 2.2.1, sequence $z_{i,j}:=X_{i,j}\epsilon_i,\ i=1,\cdots,n,\ j=1,\cdots,p$ is $\alpha-$mixing with exponential decaying mixing coefficients, and

$$\mathbb{P}(|z_{i,j}|\geq a)\leq c_1\exp(-c_2a^q), \quad a,q>0,$$

for all i and j. It also has zero-mean, i.e. $\mathrm{E}(z_{i,j})=0$. Thus, by Lemma 1 in Dendramis et al. (2019), for all $j=1,\cdots,p$,

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n z_{i,j}\right|\geq\xi\right)\leq c_0\left[\exp(-c_1'\xi^2)+\exp\left(-c_2'\left(\frac{\xi\sqrt{n}}{\log^2 n}\right)^\zeta\right)\right],$$

where $\zeta=q/(q+1)$ and constants c_0,c_1',c_2' do not depend on ξ, i and j.

Note that $\lambda_0 = \kappa\sqrt{\log p/n}$. Setting $\xi = \lambda_0\sqrt{n}/2 = \kappa\sqrt{\log p}/2$, we obtain

$$\begin{aligned} p\mathbb{P}(\frac{1}{\sqrt{n}}|z_{i,j}| > \frac{\lambda_0\sqrt{n}}{2}) &\leq pc_0\exp(-c_1'(\frac{\kappa}{2})^2\log p) + pc_0\exp(-c_2'(\frac{\kappa\sqrt{n\log p}}{2\log^2 n})^\zeta) \\ &:= r_p + r_{p,n}'. \end{aligned} \tag{2.A.18}$$

Now we consider two cases of different rates of p and n.

Case 1: $n, p \to \infty$.

Selecting $\kappa > 0$, such that $c_1'(\kappa/2)^2 > 1 + \epsilon$ for some small number $\epsilon > 0$, we obtain

$$r_p \leq pc_0\exp[-(1+\epsilon)\log p] = c_0 p^{-\epsilon} \to 0, \quad \text{as } p \to \infty. \tag{2.A.19}$$

By Assumption $p = O(n^\delta)$ for some $\delta > 0$, we have $n^{1/4} \geq p^{1/(4\delta)}$. Also, $n^{1/4} > 2\log^2 n$ as $n \to \infty$. Then

$$c_2'(\frac{\kappa\sqrt{n\log p}}{2\log^2 n})^\zeta \geq c_2'(\kappa p^{1/(4\delta)}\sqrt{\log p})^\zeta > (1+\epsilon)\log p, \quad \text{as } p \to \infty. \tag{2.A.20}$$

Therefore, equation (2.A.19) and (2.A.20) imply that

$$r_{p,n}' \leq r_p \to 0, \quad \text{as } n, p \to \infty.$$

Then by (2.A.17) and (2.A.18), we obtain

$$\begin{aligned} \mathbb{P}(E^c) &= r_p + r_{p,n}' \leq 2r_p \leq 2c_0 p^{-\epsilon}, \\ \mathbb{P}(E) &= 1 - \mathbb{P}(E^c) \geq 1 - c_0' p^{-\epsilon} \to 1, \quad \text{as } n, p \to \infty, \end{aligned} \tag{2.A.21}$$

where $c_0' = 2c_0$. This proves the first probability claim in part one of Theorem 2.2.1.

Case 2: p is bounded.

In this case, $\log p$ is also bounded, then r_p and $r_{p,n}'$ in (2.A.18) can be bounded as

$$r_p = pc_0\exp(-\frac{c_1'}{4}\kappa^2\log p), \qquad r_{p,n}' = pc_0\exp\left(-c_2'\left(\frac{\kappa\sqrt{n\log p}}{2\log^2 n}\right)^\zeta\right).$$

Therefore,

$$\mathbb{P}(E) = 1 - \mathbb{P}(E^c) = 1 - pc_0\left[\exp(-\frac{c_1'}{4}\kappa^2\log p) + \exp\left(-c_2'\left(\frac{\kappa\sqrt{n\log p}}{2\log^2 n}\right)^\zeta\right)\right], \tag{2.A.22}$$

which complete the proof of Theorem 2.2.1.

□

2.A.1.2 Proof of corollary 2.2.1

Proof. Note that $\lambda_0 = \kappa\sqrt{\log p/n}$, where $\kappa > 0$ is a tuning parameter. By (2.5) in Theorem 2.2.1 and Assumption 2.2.3(a), it follows naturally that

$$\|\hat{\beta} - \beta^0\|_1 = O_p(s\sqrt{\frac{\log p}{n}}) = o_p(1), \tag{2.A.23}$$

which proves the second claim of (2.7). Utilizing $\hat{\Sigma} = X'X/n$, we obtain

$$\begin{aligned} \|X(\hat{\beta} - \beta^0)\|_2^2/n &= (\hat{\beta} - \beta^0)'\hat{\Sigma}(\hat{\beta} - \beta^0) \\ &= (\hat{\beta} - \beta^0)'(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0) + (\hat{\beta} - \beta^0)'\Sigma(\hat{\beta} - \beta^0). \end{aligned} \tag{2.A.24}$$

Note that $\Sigma = \mathrm{E}(\hat{\Sigma})$ is non-singular, so

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

where Λ_{min} is the smallest eigenvalue of Σ, and $\Lambda_{min} > 0$. Moreover, the first part of the r.h.s of (2.A.24) has the following property:

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

where $\|\hat{\Sigma} - \Sigma\|_\infty := \max_{1\leq i,j\leq p}|\hat{\Sigma}_{i,j} - \Sigma_{i,j}|$. Using lemma 14.12 in Buhlmann and Van de Geer (2011), we have $\max_{1\leq i,j\leq p}|\hat{\Sigma}_{i,j} - \Sigma_{i,j}| = O_p(\sqrt{\log p/n})$. Together with $\|\hat{\beta} - \beta^0\|_1 = O_p(s\sqrt{\log p/n})$ obtained in (2.A.23), this implies that (2.A.24) can be bounded as

$$\begin{aligned} \frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 &= (\hat{\beta} - \beta^0)'\Sigma(\hat{\beta} - \beta^0) + (\hat{\beta} - \beta^0)'(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0) \\ &\geq \Lambda_{min}^2\|\hat{\beta} - \beta^0\|_2^2 - \|\hat{\Sigma} - \Sigma^0\|_\infty\|\hat{\beta} - \beta^0\|_1^2 \\ &\geq \Lambda_{min}^2\|\hat{\beta} - \beta^0\|_2^2 - O_p\left(s^2\left(\frac{\log p}{n}\right)^{3/2}\right). \end{aligned} \tag{2.A.25}$$

Note that $\lambda_0 \asymp \sqrt{\log p/n}$. So by (2.4) we obtain

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 = O_p\left(\frac{s\log p}{n}\right). \tag{2.A.26}$$

Plugging (2.A.26) into (2.A.25) and rearranging (2.A.25), we obtain

$$\|\hat{\beta}-\beta^0\|_2^2 \leq \frac{1}{\Lambda_{min}^2} O_p\left(\frac{s\log p}{n}\right) + \frac{1}{\Lambda_{min}^2} O_p\left(s^2\left(\frac{\log p}{n}\right)^{3/2}\right),$$

where $O_p\left(s^2\left(\frac{\log p}{n}\right)^{3/2}\right) = O_p\left(\frac{s\log p}{n}\right) O_p\left(s\sqrt{\frac{\log p}{n}}\right)$. Note that $\Lambda_{min} \geq a > 0$ where a is a constant, hence $\frac{1}{\Lambda_{min}^2} = O(1)$. Then by Assumption 2.2.3, we obtain

$$\|\hat{\beta}-\beta^0\|_2^2 = o_p(1), \tag{2.A.27}$$

which proves the first claim of (2.7). Also, by Theorem 2.2.1 part one, (2.A.23) and (2.A.27) hold with probability tending to one. This completes the proof. □

2.A.1.3 Proof of Theorem 2.2.2

Proof. By the definition of $\hat{b}$ in (2.14) and by extracting $\sqrt{n}$ from (2.12), it is easy to show that

$$\sqrt{n}(\hat{b}-\beta^0) = \hat{\Theta}X'\epsilon/\sqrt{n} - \Delta,$$

where Δ is defined in (2.13). Then to prove (2.19), it suffices to show that

$$\Delta = o_p(1). \tag{2.A.28}$$

Let X_i be a $1 \times p$ vector and denote

$$\hat{\Sigma}_{X\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i'\hat{\epsilon}_i)(X_i'\hat{\epsilon}_i)'\right]. \tag{2.A.29}$$

To show (2.20) and (2.A.28), it suffices to prove that for any $l = 1, 2, \cdots, p$ such that

$$t = \frac{\sqrt{n}(\hat{b}_l-\beta_l^0)}{\sqrt{\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l}} = \frac{\hat{\Theta}_l X'\epsilon/\sqrt{n}}{\sqrt{\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l}} + \frac{-\Delta}{\sqrt{\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l}} := t_1 + t_2,$$

where t_1 is asymptotically normal and $t_2 = o_p(1)$.

Step 1: we will show that t_1 is asymptotically normal. Let

$$t_1^* = \frac{\Theta_l' X'\epsilon/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}} = \frac{\Theta_l'\sum_{i=1}^{n} X_i'\epsilon_i/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}},$$

where $\Sigma_{X\epsilon} = E[\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)']$. We assume in Theorem 2.2.2 that $X_i'\epsilon_i$ is a stationary sequence, then $\Sigma_{X\epsilon} = \mathrm{E}[(X_1'\epsilon_1)(X_1'\epsilon_1)'] = \mathrm{Var}(X_1'\epsilon_1) > 0$. By Assumption 2.2.1 and the definition of $\Sigma_{X\epsilon}$, we have

$$E\left[\frac{\Theta_l' X'\epsilon/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}}\right] = E\left[\frac{\Theta_l'\sum_{i=1}^{n}X_i'\epsilon_i/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}}\right] = 0,$$

and

$$E\left[\frac{\Theta_l' X'\epsilon/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}}\right]^2 = E\left[\frac{\Theta_l'\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)'\Theta_l}{\Theta_l'\Sigma_{X\epsilon}\Theta_l}\right] = 1,$$

where $\Theta_l'\Sigma_{X\epsilon}\Theta_l$ is bounded away from zero. Indeed, since $\Sigma_{X\epsilon}$ is a symmetric positive definite matrix, it can be decomposed such that

$$\Theta_l'\Sigma_{X\epsilon}\Theta_l = \Theta_l'P'\mathrm{eig}(\Sigma_{X\epsilon})P\Theta_l \geq \Lambda_{min}(\Sigma_{X\epsilon})\|\Theta_l\|_2^2 > 0, \tag{2.A.30}$$

where $\mathrm{eig}(\Sigma_{X\epsilon})$ is the diagonal matrix that collects the eigenvalues of $\Sigma_{X\epsilon}$, and P is an orthonormal matrix. Because $\Lambda_{min}(\Sigma_{X\epsilon}) \geq a > 0$ where a is a constant and $\|\Theta_l\|_2^2 > 0$, so $\Theta_l'\Sigma_{X\epsilon}\Theta_l > 0$. Then by Theorem 24.6 and Corollary 24.7 in Davidson (1994), $\Theta_l'X'\epsilon/\sqrt{n} \to \mathbb{N}(0, \Theta_l\Sigma_{X\epsilon}\Theta_l')$, or $t_1^* \to \mathbb{N}(0,1)$.

Next we will show that

$$|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| = o_p(1). \tag{2.A.31}$$

Set

$$\tilde{\Sigma}_{X\epsilon} = \frac{1}{n}\sum_{i=1}^{n}[(X_i'\epsilon_i)(X_i'\epsilon_i)']\,. \tag{2.A.32}$$

Then

$$\begin{aligned}|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| &\leq |\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l| + |\hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \\ &\leq |\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l| + |\hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l| \\ &\quad + |\hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \\ &= (I) + (II) + (III).\end{aligned} \tag{2.A.33}$$

For (I), we have

$$|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l| \leq \|\hat{\Sigma}_{X\epsilon} - \tilde{\Sigma}_{X\epsilon}\|_\infty \|\hat{\Theta}_l\|_1^2.$$

Note that $\hat{\epsilon}_i = \epsilon_i + X_i(\beta^0 - \hat{\beta})$. Plugging $\hat{\epsilon}_i$ into $\hat{\Sigma}_{X\epsilon} - \tilde{\Sigma}_{X\epsilon}$, we obtain

$$\begin{aligned}
\hat{\Sigma}_{X\epsilon} - \tilde{\Sigma}_{X\epsilon} =& \frac{1}{n}\sum_{i=1}^n \Big[[X_i'(\epsilon_i + X_i(\beta^0 - \hat{\beta})][X_i'(\epsilon_i + X_i(\beta^0 - \hat{\beta})]'\Big] - \frac{1}{n}\sum_{i=1}^n [(X_i'\epsilon_i)(X_i'\epsilon_i)']\\
=& \frac{1}{n}\sum_{i=1}^n X_i'X_i(\beta^0 - \hat{\beta})[X_i'X_i(\beta^0 - \hat{\beta})]' + \frac{1}{n}\sum_{i=1}^n X_i'\epsilon_i[X_i'X_i(\beta^0 - \hat{\beta})]'\\
&+ \frac{1}{n}\sum_{i=1}^n [X_i'X_i(\beta^0 - \hat{\beta})](X_i'\epsilon_i)'\\
=&(i) + (ii) + (iii).
\end{aligned}$$

Next, we will show that $\|(i)\|_\infty = O_p(s\sqrt{\log p/n})$, $\|(ii)\|_\infty = O_p(\sqrt{s\log p/n})$ and $\|(iii)\|_\infty = O_p(\sqrt{s\log p/n})$. First of all, for (i), we have

$$\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^n X_i'X_i(\hat{\beta} - \beta^0)[X_i'X_i(\hat{\beta} - \beta^0)]'\|_\infty =& \|\frac{1}{n}\sum_{i=1}^n X_i'X_i(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)'X_i'X_i\|_\infty\\
\leq& \frac{1}{n}\sum_{i=1}^n \|X_i'X_iX_i'X_i\|_\infty \|\hat{\beta} - \beta^0\|_1^2\\
\leq& \max_j \frac{1}{n}\sum_{i=1}^n X_{i,j}^4 \,\|\hat{\beta} - \beta^0\|_1^2,
\end{aligned} \tag{2.A.34}$$

where $j = 1, \cdots, p$. By Assumption 2.2.1, $\mathbb{P}(|X_{i,j}| > a) \leq c_1 \exp(-c_2 a^q)$. Set $Y_{i,j} = X_{i,j}^4$, then $\mathbb{P}(|Y_{i,j}| > a) = \mathbb{P}(|X_{i,j}| > a^{1/4}) \leq c_1 \exp(-c_2 a^{q/4})$. So $X_{i,j}^4$ also has exponential tail bound (with a different parameter). Then by (2.A.17) and (2.A.21), for all $j = 1, \cdots, p$, we have $\mathbb{P}(|n^{-1}\sum_{i=1}^n X_{i,j}^4 - E(X_{i,j}^4)| > \lambda_0/2) \leq c_0' p^{-\epsilon} \to 0$ as $n, p \to \infty$, where c_0' is a positive constant and $\epsilon > 0$ is a small number. Note that $\lambda_0 \asymp \sqrt{\log p/n}$. Hence $|n^{-1}\sum_{i=1}^n X_{i,j}^4 - E(X_{i,j}^4)| = O_p(\sqrt{\log p/n})$ with probability tending to one. Then by Assumption 2.2.1, $E(X_{i,j}^4) < \infty$, and by Assumption 2.2.3, $\sqrt{\log p/n} = o_p(1)$, so we have $|n^{-1}\sum_{i=1}^n X_i^4| \leq |n^{-1}\sum_{i=1}^n X_{i,j}^4 - E(X_{i,j}^4)| + |E(X_{i,j}^4)| = o_p(1) + O_p(1) = O_p(1)$. By (2.7) we have $\|\hat{\beta} - \beta^0\|_1 = O_p(s\sqrt{\log p/n})$ with probability tending to one. Therefore, we have $\|(i)\|_\infty = O_p(s\sqrt{\log p/n})$ with probability tending to one.

For (ii), we have

$$\begin{aligned}\|\frac{1}{n}\sum_{i=1}^{n} X_i'\epsilon_i[X_i'X_i(\beta^0-\hat{\beta})]'\|_\infty =& \|\frac{1}{n}\sum_{i=1}^{n} X_i'\epsilon_i X_i\|_\infty[X_i(\beta^0-\hat{\beta})]' \\ \leq& \frac{1}{n}\|\sum_{i=1}^{n} X_i'X_iX_i'X_i\epsilon_i^2\|_\infty^{1/2}(\sum_{i=1}^{n}[X_i(\beta^0-\hat{\beta})]^2)^{1/2} \\ \leq& (\frac{1}{n}\max_j\sum_{i=1}^{n} X_{i,j}^4\epsilon_i^2)^{1/2}(\frac{1}{n}\|X_i(\beta^0-\hat{\beta})\|_2^2)^{1/2} \\ \leq& (\frac{1}{n}\max_j\sum_{i=1}^{n} X_{i,j}^4\epsilon_i^2)^{1/2}(\frac{1}{n}\|X(\beta^0-\hat{\beta})\|_2).\end{aligned}$$

By (2.4) and Assumption 2.2.3, we have $\|n^{-1}X(\hat{\beta}-\beta^0)\|_2 = O_p(\sqrt{(s\log p)/n})$. Since both $X_{i,j}$ and ϵ_i are $\alpha-$mixing and have exponential tail distributions, then following a similar argument as in (i) and using (2.A.17) and (2.A.21), for any $j = 1, \cdots, p$, we have $n^{-1}\sum_{i=1}^{n} X_{i,j}^4\epsilon_i^2 = O_p(\sqrt{\log p/n}) + O_p(1) = o_p(1) + O_p(1) = O_p(1)$. Therefore, $\|(ii)\|_\infty = O_p(\sqrt{s\log p/n})$. For (iii), it is easy to show that $(iii) = (ii)'$, so $\|(iii)\|_\infty = O_p(\sqrt{s\log p/n})$.

Then by Lemma 2.A.3 below, $\|\hat{\Theta}_l\|_1 = O_p(\sqrt{s_l})$ and by Assumption 2.2.3, we obtain

$$(I) = O_p\left(s\sqrt{\frac{\log p}{n}}\right)O_p(s_l) + O_p\left(\sqrt{\frac{s\log p}{n}}\right)O_p(s_l) = o_p(1).$$

For (II), we have

$$|\hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l| \leq \|\tilde{\Sigma}_{X\epsilon}-\Sigma_{X\epsilon}\|_\infty\|\hat{\Theta}_l\|_1^2,$$

where

$$\|\tilde{\Sigma}_{X\epsilon}-\Sigma_{X\epsilon}\|_\infty = \left\|\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)' - E[\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)']\right\|_\infty.$$

Since $X_i'\epsilon_i$ is $\alpha-$mixing and has exponential tail distribution, by (2.A.17) and (2.A.21), $\|\tilde{\Sigma}_{X\epsilon}-\Sigma_{X\epsilon}\|_\infty = O_p(\sqrt{\log p/n})$ with probability tending to one. Therefore, by assumption 2.2.3, we obtain $(II) = O_p(\sqrt{\log p/n})\ O_p(s_l) = o_p(1)$.

For (III), by Lemma 3.1 in the supplement material of Van De Geer et al. (2014),

$$|\hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \leq \|\Sigma_{X\epsilon}\|_\infty\|\hat{\Theta}_l-\Theta_l\|_1^2 + 2\|\Sigma_{X\epsilon}\Theta_l\|_2\|\hat{\Theta}_l-\Theta_l\|_2,$$

where, by Lemma 2.A.3, $\|\hat{\Theta}_l-\Theta_l\|_1 = O_p(s_l\sqrt{\log p/n})$ and $\|\hat{\Theta}_l-\Theta_l\|_2 = O_p(\sqrt{s_l\log p/n})$.

Furthermore, note that $\Sigma_{X\epsilon}$ and $\Theta := \Sigma^{-1}$ are symmetric positive definite matrices, and their smallest eigenvalues are strictly greater than zero and their largest eigenvalues are bounded above. Therefore,

$$\begin{aligned}\|\Sigma_{X\epsilon}\|_\infty \leq& \|\Sigma_{X\epsilon}\|_2 = \Lambda_{\max}(\Sigma_{X\epsilon}) = O_p(1),\\ \|\Sigma_{X\epsilon}\Theta_l\|_\infty \leq& \|\Sigma_{X\epsilon}\|_\infty \|\Theta_l\|_\infty \leq O_p(1)\|\Theta\|_2 = O_p(1)\Lambda_{max}(\Theta) = O_p(1)/\Lambda_{min}(\Sigma) = O_p(1).\end{aligned}$$

Thus, by Assumption 2.2.3, we obtain that $(III) = O_p(s_l^2 \log p/n) + O_p(\sqrt{s_l \log p/n}) = o_p(1)$. Therefore, in equation (2.A.33) we have

$$|\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \Theta_l' \Sigma_{X\epsilon}\Theta_l| \leq (I) + (II) + (III) = o_p(1).$$

Next, we will show that

$$|\hat{\Theta}_l' X'\epsilon/\sqrt{n} - \Theta_l' X'\epsilon/\sqrt{n}| = o_p(1). \tag{2.A.35}$$

By Lemma 2.A.3, $\|\hat{\Theta}_l - \Theta_l\|_1 = O_p(s_l\sqrt{\log p/n})$ and by (2.A.17) and (2.A.21), $\|X'\epsilon/n\|_\infty = O_p(\sqrt{\log p/n})$. Then by Assumption 2.2.3, equation (2.A.35) can be written as

$$\begin{aligned}|\hat{\Theta}_l' X'\epsilon/\sqrt{n} - \Theta_l' X'\epsilon/\sqrt{n}| \leq& \|\hat{\Theta}_l - \Theta_l\|_1 \|\frac{X'\epsilon}{n}\|_\infty \sqrt{n}\\ =& O_p(s_l\sqrt{\frac{\log p}{n}}) O_p(\sqrt{\frac{\log p}{n}})\sqrt{n} = O_p(\frac{s_l \log p}{\sqrt{n}}) = o_p(1),\end{aligned}$$

with probability tending to one, which completes the proof of (2.20).

Step 2: now we will show that $t_2 = o_p(1)$. Note that for any $l = 1, \cdots, p$,

$$\|\Delta\|_\infty = \|\sqrt{n}(\hat{\Theta}\hat{\Sigma} - I)(\hat{\beta} - \beta^0)\|_\infty \leq \sqrt{n}\max_l \|\hat{\Sigma}\hat{\Theta}_l - e_l\|_\infty \|\hat{\beta} - \beta^0\|_1,$$

where $\hat{\Theta}_l$ is the l^{th} row of $\hat{\Theta}$ written as a column vector and e_l is a $p \times 1$ column vector where the l^{th} element is one, while elsewhere being zeros. By Lemma 5.3 in Van De Geer et al. (2014), $1/\hat{\delta}_l^2 = O_p(1)$ where $\hat{\delta}_l^2$ is defined as in (2.17), and by (2.A.49), we obtain

$$\|\Delta\|_\infty \leq \sqrt{n}\frac{\lambda_l}{\hat{\delta}_l^2} O_p(s\sqrt{\frac{\log p}{n}}) = \sqrt{n} O_p(\sqrt{\frac{\log p}{n}}) O_p(s\sqrt{\frac{\log p}{n}}) = O_p(\frac{s\log p}{\sqrt{n}}) = o_p(1).$$

We have shown that $|\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \Theta_l' \Sigma_{X\epsilon}\Theta_l| = o_p(1)$ and by (2.A.30), $|\Theta_l' \Sigma_{X\epsilon}\Theta_l| \geq a > 0$

where a is a constant. Using triangle inequality $|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \geq |\Theta_l'\Sigma_{X\epsilon}\Theta_l| - |\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l|$, we obtain $|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l| \geq a - o_p(1) > 0$. Therefore,

$$t_2 = \frac{-\Delta}{\sqrt{\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l}} = \frac{-\sqrt{n}(\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon} - e_l)(\hat{\beta} - \beta^0)}{\sqrt{\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l}} = o_p(1),$$

which proves (2.A.28). □

2.A.1.4 Proof of Lemma 2.2.1

Proof. The restricted eigenvalue condition for $\hat{\Sigma}$ in (2.3) implies that

$$0 < \phi_0^2 \leq \frac{\beta'\hat{\Sigma}\beta}{\|\beta_{s_0}\|_2^2} \leq \frac{\beta'\hat{\Sigma}\beta s}{\|\beta_{s_0}\|_1^2},$$

where for the second inequality we utilize the norm inequality $\sqrt{s}\|\beta_{s_0}\|_2 \geq \|\beta_{s_0}\|_1$. Rearranging the above inequality, we have

$$\|\beta_{s_0}\|_1^2 \leq (\beta'\hat{\Sigma}\beta)s/\phi_0^2,$$

which completes the proof. □

2.A.1.5 Proof of Proposition 2.2.1

Proof. We utilize the self-normalized sum properties in Lemma 2.A.2 under weak dependence to bound tuning parameters λ_1 and λ_2. To choose appropriate values for tuning parameters such that the penalty level is large enough to cancel out noises from estimation errors, we need to ensure that $\mathbb{P}(\|X'\epsilon\|_\infty/n \leq \lambda_0/2)$ is close to one. Or equivalently we want to show that

$$\mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2) \leq \alpha, \tag{2.A.36}$$

where α is a small positive number. First, suppose that all $X_{i,j}'s$ are normalized, such that for all $j = 1, \ldots, p$, $\frac{1}{n}\sum_{i=1}^{n} X_{i,j}^2\epsilon_i^2 \to \sigma^2$ as $n \to \infty$. Let G denote an event such that $G = \left\{\max_{j\leq p}|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}^2\epsilon_i^2 - \sigma^2| \leq \frac{\sigma^2}{\log n}\right\}$. Suppose that when $n \to \infty$, $\mathbb{P}(G) \to 1$, and on G, $\frac{1}{n}\sum_{i=1}^{n} X_{i,j}^2\epsilon_i^2 \leq (1 + 1/\log n)\sigma^2$. The definition of G ensures that $\frac{1}{n}\sum_{i=1}^{n} X_i^2\epsilon_i^2$ converges to σ^2 at the rate of $\log n$. Then, utilizing the union bound in (2.A.36), we have

$$\begin{aligned}\mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2) &\leq \mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2, G) + \mathbb{P}(G^C) && (2.A.37)\\ &= \mathbb{P}(\max_j |\frac{1}{n}\sum_{i=1}^n X_{i,j}\epsilon_i| > \frac{\lambda_0}{2}, G) + \mathbb{P}(G^C) && (2.A.38)\\ &\leq p\ \mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^n X_{i,j}\epsilon_i| > \lambda_0/2, G\right) + \mathbb{P}(G^C) \leq \alpha. && (2.A.39)\end{aligned}$$

Note that on G, we have $\left(\frac{1}{n}\sum\limits_{i=1}^n X_i^2\epsilon_i^2\right)^{1/2} \geq (1+1/\log n)^{1/2}\sigma$. So (2.A.39) can be written as

$$\mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2) \leq p\ \mathbb{P}\left(\left\{\frac{|\frac{1}{n}\sum\limits_{i=1}^n X_{i,j}\epsilon_i|}{\left(\frac{1}{n}\sum\limits_{i=1}^n X_{i,j}^2\epsilon_i^2\right)^{1/2}} > \frac{\lambda_0}{2\sigma(1+1/\log n)^{1/2}}\right\}\cap G\right) + \mathbb{P}(G^C) \qquad (2.A.40)$$

$$\leq 2p\ \mathbb{P}\left(\frac{\sum\limits_{i=1}^n X_{i,j}\epsilon_i/\sqrt{n}}{\left(\sum\limits_{i=1}^n X_{i,j}^2\epsilon_i^2/n\right)^{1/2}} > \frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right) + o(1) \qquad (2.A.41)$$

$$\leq \alpha. \qquad (2.A.42)$$

Applying the self-normalization theorem of Chen et al. (2016) given in Lemma 2.A.2 below on (2.A.41) gives

$$\mathbb{P}\left(\frac{\sum\limits_{i=1}^n X_{i,j}\epsilon_i/\sqrt{n}}{\left(\sum\limits_{i=1}^n X_{i,j}^2\epsilon_i^2/n\right)^{1/2}} > \frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right) \to 1-\Phi\left(\frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right).$$

Together with (2.A.42), this implies

$$\begin{aligned}2p\left[1-\Phi\left(\frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right)\right] &\leq \alpha - o(1),\\ \lambda_0 &\geq \frac{2\sigma}{\sqrt{n}}(1+\frac{1}{\log n})^{1/2}\Phi^{-1}(1-\frac{\alpha}{2p}).\end{aligned} \qquad (2.A.43)$$

Since $\lambda_1/n = 2\lambda_0$, we obtain the first part of (2.8). Also, since $\lambda_2/n = O_p(s\log p/(np))$ and $\frac{\sqrt{\log p}}{\sqrt{n}} \asymp \lambda_1/n$, and we assume that s is small relative to p, so we approximate

$\frac{s\sqrt{\log p}}{\sqrt{n}} \approx \lambda_1/n$, which gives the second part of (2.8). However, σ is unknown. We implement a recursive procedure to evaluate the unknown variance following Algorithm A.1 in Belloni et al. (2012). In particular, we first set $\sigma = 1$ to evaluate the penalized regression and get a preliminary empirical variance $\hat{\sigma}^2$. Then we refine the estimation result using the updated empirical variance for σ. We repeat this exercise K times to get the final estimate.[12] □

2.A.1.6 Auxiliary lemmas

Lemma 2.A.1 (Dendramis et al. (2019), Lemma 1)**.** *Let $\{X\}_n$ be a sequence that satisfies Assumption 2.2.1. Then*

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} X_i\right| \geq \xi\right) \leq c_0\left[\exp(-c_1\xi^2) + \exp\left(-c_2\left(\frac{\xi\sqrt{n}}{\log^2 n}\right)^s\right)\right],$$

where $s = q/(q+1)$, and constants c_0, c_1, c_2 do not depend on ξ and i.

Proof: see Dendramis et al. (2019).

Lemma 2.A.2 (Chen et al. (2016), Theorem 4.1)**.** *Let weekly dependent random variable X_i be zero-mean, i.e. $E(X_i) = 0$. Write $S_{k,m} = \sum_{i=k+1}^{k+m} X_i$. Suppose for a positive constant c, $E(S_{k,m}^2) \geq c^2 m$ for any $k \geq 0$, $m \geq 1$. Let $m_1 > m_2 > 0$, $m^* = m_1 + m_2$, $k = \lfloor n/m^* \rfloor$.* [13] *For $1 \leq j \leq k$, denote $H_{j,1} = \{i : (j-1)m^* + 1 \leq i \leq (j-1)m^* + m_1\}$ and $H_{j,2} = \{i : (j-1)m^* + m_1 + 1 \leq i \leq jm^*\}$. Define $Y_j := \sum_{i \in H_{j,1}} X_i$ and $W_n := \sum_{j=1}^{k} Y_j / (\sum_{j=1}^{k} Y_j^2)^{1/2}$. Then*

$$\frac{\mathbb{P}(W_n \geq t)}{1 - \Phi(t)} \to 1,$$

uniformly in $0 \leq t \leq o(n^{1/8})$.

Proof: see Chen et al. (2016).

[12] For instance in Belloni et al. (2012), $K = 15$.

[13] We use $\lfloor \cdot \rfloor$ to denote the integer part of a floating number.

2.A.2 $\hat{\Theta}$ as approximation of Σ^{-1}

In this section, we closely follow Van De Geer et al. (2014) and Kock (2016) to check whether $\hat{\Theta}$ is a good approximation of Σ^{-1}. The first order condition of (2.15) implies

$$X'_{-j}(X_j - X_{-j}\hat{\gamma}_j)/n = \lambda_j \hat{\tau}_j. \tag{2.A.44}$$

Note that $\hat{\gamma}'_j \lambda_j \hat{\tau}_j = \lambda_j \|\hat{\gamma}_j\|_1$. Then left-multiplying $\hat{\gamma}'_j$ on both sides of (2.A.44) implies

$$\hat{\gamma}'_j X'_{-j}(X_j - X_{-j}\hat{\gamma}_j)/n = \lambda_j \|\hat{\gamma}_j\|_1. \tag{2.A.45}$$

Therefore, plugging the above equation into (2.17), we have

$$\begin{aligned}\hat{\delta}_j^2 &= \frac{1}{n}(X_j - X_{-j}\hat{\gamma}_j)'(X_j - X_{-j}\hat{\gamma}_j) + \frac{1}{n}\hat{\gamma}'_j X'_{-j}(X_j - X_{-j}\hat{\gamma}_j)\\ &= \frac{1}{n}[(X_j - X_{-j}\hat{\gamma}_j)' + \hat{\gamma}'_j X'_{-j}](X_j - X_{-j}\hat{\gamma}_j)\\ &= \frac{1}{n}X'_j(X_j - X_{-j}\hat{\gamma}_j).\end{aligned} \tag{2.A.46}$$

By definition of $\hat{C}_j$ (j^{th} row of matrix $\hat{C}$) in (2.16), we have $X_j - X_{-j}\hat{\gamma}_j = X\hat{C}_j$, and by the definition of $\hat{\Theta}_j = \hat{C}_j/\hat{\delta}_j^2$ in (2.18), equation (2.A.46) becomes

$$\hat{\delta}_j^2 = \frac{1}{n}X'_j X\hat{C}_j, \qquad \text{or} \qquad \frac{1}{n}X'_j X\hat{\Theta}_j = 1. \tag{2.A.47}$$

where $\hat{\Theta}_j$ is the j^{th} row of $\hat{\Theta}$ written as a column vector. Thus we can see that $\hat{\Theta}$ is a good approximation of the inverse of the Gram matrix $\hat{\Sigma} := X'X/n$.

Next, we look into the approximation error $\|\hat{\Theta}\hat{\Sigma} - I\|_\infty$, or specifically the j^{th} column of the approximation error, which is $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty$ for all $j = 1, \cdots, p$, where e_j is the j^{th} column of the identity matrix. By the definition of $\hat{\tau}$ in (2.10), $\|\hat{\tau}\|_\infty \leq 1$. Taking the norm on both sides of (2.A.44) and using $\hat{\Theta}_j = \hat{C}_j/\hat{\delta}_j^2$, we obtain

$$\begin{aligned}\|X'_{-j}(X_j - X_{-j}\hat{\gamma}_j)\|_\infty/n = \|X'_{-j}X\hat{C}_j\|_\infty = \|\lambda_j\hat{\tau}_j\|_\infty,\\ \|X'_{-j}X\hat{\Theta}_j\|_\infty/n = \lambda_j\|\hat{\tau}_j\|_\infty/\hat{\delta}^2 \leq \lambda_j/\hat{\delta}_j^2.\end{aligned} \tag{2.A.48}$$

By the definition of X_{-j} and $\hat{\Sigma} := X'X/n$ and by (2.A.47), we have $\|X'_{-j}X\hat{\Theta}_j\|_\infty = \|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty$. Thus (2.A.48) can be written as

$$\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \leq \lambda_j/\hat{\delta}_j^2. \tag{2.A.49}$$

Next, we formally investigate the asymptotic properties of $\hat{\Theta}$.

Asymptotic properties of $\hat{\Theta}$

Let Θ denote the population value of $\hat{\Theta}$ such that $\Theta := E(\hat{\Theta}) := \Sigma^{-1}$. First, partitioning Σ^{-1} into the first element and the remaining ones gives

$$\begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix}^{-1} = \begin{pmatrix} \overbrace{(\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})^{-1}}^{\Theta_{1,1}} & \overbrace{-\Theta_{1,1}\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}}^{\Theta_{1,-1}} \\ -\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\Theta_{1,1} & (\Sigma_{-1,-1} - \Sigma_{-1,1}\Sigma_{1,1}^{-1}\Sigma_{1,-1})^{-1} \end{pmatrix},$$

where '-1' indicates all the rows (columns) excluding the first row (column). More generally, for the j^{th} row and column of Θ, we can write

$$\Theta_{j,j} = (\Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j})^{-1}, \qquad \Theta_{j,-j} = -\Theta_{j,j}\Sigma_{j,-j}\Sigma_{-j,-j}^{-1}. \tag{2.A.50}$$

Denote γ_j the population value of $\hat{\gamma}_j$. Then

$$\gamma_j := \arg\min_{\gamma} \frac{1}{n}\sum_{i=1}^{n} E(X_{i,j} - X_{i,-j}'\gamma)^2.$$

Then the first order condition of the above equation implies,

$$\gamma_j = [\frac{1}{n}\sum_{i=1}^{n} E(X_{i,-j}'X_{i,-j})]^{-1}[\frac{1}{n}\sum_{i=1}^{n} E(X_{i,-j}'X_{i,j})] = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j}. \tag{2.A.51}$$

Thus, (2.A.50) and (2.A.51) implies that $\Theta_{j,-j} = -\Theta_{j,j}\gamma_j'$. Denoting δ_j^2 the population value of $\hat{\delta}_j^2$ and utilizing (2.A.51), we obtain

$$\begin{aligned} \delta_j^2 &= E[\frac{1}{n}\sum_{i=1}^{n} E(X_{i,j} - X_{i,-j}'\gamma_j)^2] \\ &= \Sigma_{j,j} + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j} - 2\Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j} \\ &= \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j} = \frac{1}{\Theta_{j,j}}, \end{aligned}$$

where the last equality comes from (2.A.50). Therefore, $\Theta_{j,j} = 1/\delta_j^2$ and $\Theta_{j,-j} = -\gamma_j'/\hat{\delta}_j{}^2$. Then it follows that $\Theta = T^{-2}C$, where C is the population value of $\hat{C}$ in (2.16) (by replacing $\hat{\gamma}_j$ with γ_j) and T^2 is the population value of $\hat{T}^2$ in (2.16) (by replacing $\hat{\delta}_j^2$ with δ_j^2).

Formally, the following lemma derives the rate of the approximation $\hat{\Theta}_j$ and the true value Θ_j.

Lemma 2.A.3. *Suppose Assumption 2.2.1 and 2.2.2 hold, then*

$$\|\hat{\Theta}_j - \Theta_j\|_1 = O_p(s_j\sqrt{\frac{\log p}{n}}),$$
$$\|\hat{\Theta}_j - \Theta_j\|_2 = O_p(\sqrt{\frac{s_j \log p}{n}}),$$
$$\|\Theta_j\|_1 = O(\sqrt{s_j}),$$
$$\|\hat{\Theta}_j\|_1 = O_p(\sqrt{s_j}).$$

Proof of Lemma 2.A.3. First, we consider $|\hat{\delta}_j^2 - \delta_j^2|$. From (2.A.46) we have $\hat{\delta}_j^2 = X_j'(X_j - X_{-j}\hat{\gamma}_j)/n$. Suppose $X_j = X_{-j}\gamma_j + \eta_j$ and $X_j = X_{-j}\hat{\gamma}_j + \hat{\eta}_j$, where η_j and $\hat{\eta}_j$ are residuals. Then we obtain that $\hat{\delta}_j^2 = X_j'\hat{\eta}_j/n$ and $\hat{\eta}_j = X_{-j}(\gamma_j - \hat{\gamma}_j) + \eta_j$. Plugging X_j and $\hat{\eta}_j$ into $\hat{\delta}_j^2$ gives

$$\begin{aligned}\hat{\delta}_j^2 &= \frac{1}{n}(X_{-j}\hat{\gamma}_j + \hat{\eta}_j)'[X_{-j}(\gamma_j - \hat{\gamma}_j) + \eta_j] \\ &= \frac{1}{n}\gamma_j'X_{-j}'X_{-j}(\gamma_j - \hat{\gamma}_j) + \frac{1}{n}\gamma_j'X_{-j}\eta_j + \frac{1}{n}\eta_j'X_{-j}'(\gamma_j - \hat{\gamma}_j) + \frac{1}{n}\eta_j'\eta_j.\end{aligned} \tag{2.A.52}$$

Therefore, we obtain

$$\begin{aligned}|\hat{\delta}_j^2 - \delta_j^2| \le |\frac{1}{n}\eta_j'\eta_j - \delta_j^2| + |\frac{1}{n}\eta_j'X_{-j}'(\gamma_j - \hat{\gamma}_j)| + |\frac{1}{n}\gamma_j'X_{-j}\eta_j| + |\frac{1}{n}\gamma_j'X_{-j}'X_{-j}(\gamma_j - \hat{\gamma}_j)| \\ := I + II + III + IV.\end{aligned} \tag{2.A.53}$$

For (I), note that $\delta_j = \mathrm{E}(X_j - X_{-j}\gamma_j) = \mathrm{E}(\eta_j)$. We assume η_j^2 is $\alpha-$mixing with exponential decaying mixing coefficients as in Assumption 2.2.1. Then by (2.A.17) and (2.A.21), we obtain $|\frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_{i,j}^2 - \mathrm{E}\eta_{i,j}^2| = O_p(1)$. Therefore,

$$|\frac{1}{n}\eta_j'\eta_j - \delta_j^2| = |\frac{1}{n}\sum_{i=1}^n \eta_{i,j}^2 - \mathrm{E}\eta_{i,j}^2| = O_p(\frac{1}{\sqrt{n}}). \tag{2.A.54}$$

For (II), we have

$$|\frac{1}{n}\eta_j'X_{-j}'(\gamma_j - \hat{\gamma}_j)| \le \frac{1}{n}\|\eta_j'X_{-j}\|_\infty\|\gamma_j - \hat{\gamma}_j\|_1, \tag{2.A.55}$$

where $\frac{1}{n}\|\eta_j'X_{-j}\|_\infty = \max_{k\in\{1,\cdots,p\}\setminus\{j\}}|\frac{1}{n}\sum_{i=1}^n X_{i,k}\eta_{i,j}|$. Note that $X_{i,k}\eta_{i,j}$ is $\alpha-$mixing with exponential decaying tail distribution. Then by (2.A.17) and (2.A.21), we obtain

$$\frac{1}{n}\|\eta_j' X_{-j}\|_\infty = O_p(\sqrt{\log p/n}). \tag{2.A.56}$$

Together with $\|\gamma_j - \hat{\gamma}_j\|_1 = O_p(s_j\sqrt{\log p/n})$, (2.A.55) can be bounded

$$|\frac{1}{n}\eta_j' X_{-j}'(\gamma_j - \hat{\gamma}_j)| = O_p(\sqrt{\frac{\log p}{n}})O_p(s_j\sqrt{\frac{\log p}{n}}) = O_p(\frac{s_j \log p}{n}). \tag{2.A.57}$$

For (III), we have

$$|\frac{1}{n}\gamma_j' X_{-j}\eta_j| \leq \|\frac{1}{n}X_{-j}'\eta_j\|_\infty \|\gamma_j\|_1. \tag{2.A.58}$$

Note that $X_j = X_{-j}\gamma_j + \eta_j$. we can bound $\Sigma_{j,j}$ as

$$E(X_j' X_j) = \Sigma_{j,j} \geq E[(X_{-j}\gamma_j)' X_{-j}\gamma_j] = \gamma_j' \Sigma_{-j,-j}\gamma_j \geq \Lambda_{min}^2 \|\gamma_j\|_2^2, \tag{2.A.59}$$

where Λ_{min} is the smallest eigenvalue of $\Sigma_{-j,-j}$ (i.e., removing j^{th} row and column from Σ gives $\Sigma_{-j,-j}$). Since Σ is a symmetric positive definite matrix, so $\Lambda_{min} \geq a > 0$, thus $1/\Lambda_{min}^2 = O(1)$. Then the above inequality implies that $\|\gamma_j\|_2 \leq \sqrt{\Sigma_{j,j}}/\Lambda_{min}$. Further utilizing the norm inequality $\|\gamma_j\|_1 \leq \sqrt{s_j}\|\gamma_j\|_2$, we obtain $\|\gamma_j\|_1 \leq \sqrt{s_j\Sigma_{j,j}}/\Lambda_{min}$. Therefore, by (2.A.56), inequality (2.A.58) can be bounded as

$$|\frac{1}{n}\gamma_j' X_{-j}\eta_j| = O_p(\sqrt{\frac{\log p}{n}})O_p(\sqrt{s_j}) = O_p(\sqrt{\frac{s_j \log p}{n}}).$$

For (IV), the first order condition of nodewise LASSO in (2.A.44) implies

$$\lambda_j \hat{\tau}_j + \frac{1}{n}X_{-j}' X_{-j}\hat{\gamma}_j - \frac{1}{n}X_{-j}' X_j = 0.$$

Plugging $X_j = X_{-j}\gamma_j + \eta_j$ into the above equation gives

$$\frac{1}{n}X_{-j}' X_{-j}(\gamma_j - \hat{\gamma}_j) = \lambda_j \hat{\tau}_j - \frac{1}{n}X_{-j}'\eta_j.$$

By (2.A.56) and $\lambda_j \asymp \sqrt{\log p/n}$, $\|\hat{\tau}_j\|_\infty \leq 1$, we obtain

$$\|\frac{1}{n}X_{-j}' X_{-j}(\gamma_j - \hat{\gamma}_j)\|_\infty \leq \|\frac{1}{n}X_{-j}'\eta_j\|_\infty + \lambda_j\|\hat{\tau}_j\|_\infty = O_p(\sqrt{\frac{\log p}{n}}).$$

Note that by (2.A.59), $\|\gamma_j\|_2 = O(1)$. Then using the norm inequality, we have

$\|\gamma_j\|_1 \leq \sqrt{s_j}\|\gamma_j\|_2 = O(\sqrt{s_j})$. Therefore, (IV) can be bounded as

$$|\frac{1}{n}\gamma_j' X_{-j}' X_{-j}(\gamma_j - \hat{\gamma}_j)| \leq \|\frac{1}{n} X_{-j}' X_{-j}(\gamma_j - \hat{\gamma}_j)\|_\infty \|\gamma_j\|_1 = O_p(\sqrt{\frac{s_j \log p}{n}}). \quad (2.A.60)$$

Note that $\max_j(s_j \log p/n) = o(1)$, thus for any $j = 1, \cdots, p$, $s_j \log p/n \leq \sqrt{s_j \log p/n}$. Therefore, we have

$$|\hat{\delta}_j^2 - \delta_j^2| = O_p(\sqrt{\frac{s_j \log p}{n}}).$$

By Lemma 5.3 in Van De Geer et al. (2014), we have $\dfrac{1}{\hat{\delta}_j^2} = O_p(1)$ and $\dfrac{1}{\delta_j^2} = O(1)$. Then it follows

$$\left|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\right| \leq \frac{|\hat{\delta}_j^2 - \delta_j^2|}{\hat{\delta}_j^2 \delta_j^2} = O_p(\sqrt{\frac{s_j \log p}{n}}).$$

Then, by the definition of $\hat{\Theta}$ and $\hat{C}$ in (2.18) and (2.16), we obtain

$$\begin{aligned}
\|\hat{\Theta}_j - \Theta_j\|_1 &= \|\frac{\hat{C}_j}{\hat{\delta}_j^2} - \frac{C_j}{\delta_j^2}\|_1 = \|\frac{1 - \hat{\gamma}_j}{\hat{\delta}_j^2} - \frac{1 - \gamma_j}{\delta_j^2}\|_1 \\
&\leq \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1 + \|\frac{\hat{\gamma}_j}{\hat{\delta}_j^2} - \frac{\gamma_j}{\hat{\delta}_j^2} + \frac{\gamma_j}{\hat{\delta}_j^2} - \frac{\gamma_j}{\delta_j^2}\|_1 \\
&\leq \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1 + \|\frac{1}{\hat{\delta}_j^2}\|_1 \|\hat{\gamma}_j - \gamma_j\|_1 + \|\gamma_j\|_1 \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1 \\
&= O_p(\sqrt{\frac{s_j \log p}{n}}) + O_p(1) O_p(s_j \sqrt{\frac{\log p}{n}}) + O_p(\sqrt{s_j}) O_p(\sqrt{\frac{s_j \log p}{n}}) \\
&= O_p(s_j \sqrt{\frac{\log p}{n}}).
\end{aligned} \quad (2.A.61)$$

Next, we will bound $\|\hat{\Theta}_j - \Theta_j\|_2$. Note that $\|\hat{\gamma}_j - \gamma_j\|_2 = O_p(\sqrt{s_j \log p/n})$ and $\|\dfrac{1}{\hat{\delta}_j^2} - \dfrac{1}{\delta_j^2}\|_2 = \|\dfrac{1}{\hat{\delta}_j^2} - \dfrac{1}{\delta_j^2}\|_1$ and $\|\dfrac{1}{\hat{\delta}_j^2}\|_2 = \|\dfrac{1}{\hat{\delta}_j^2}\|_1$ since they are both scalars. Similarly to (2.A.61) we have

$$\begin{aligned}
\|\hat{\Theta}_j - \Theta_j\|_2 &\leq \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_2 + \|\frac{1}{\hat{\delta}_j^2}\|_2 \|\hat{\gamma}_j - \gamma_j\|_2 + \|\gamma_j\|_2 \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_2 \\
&= O_p(\sqrt{\frac{s_j \log p}{n}}) + O_p(1) O_p(\sqrt{\frac{s_j \log p}{n}}) + O_p(1) O_p(\sqrt{\frac{s_j \log p}{n}}) \\
&= O_p(\sqrt{\frac{s_j \log p}{n}}).
\end{aligned}$$

Next, by the definition of Θ and $\sqrt{\log p/n} = o_p(1)$, we obtain

$$\|\Theta_j\|_1 \le \|\frac{1}{\delta_j^2}\|_1 \|C_j\|_1 \le \|\frac{1}{\delta_j^2}\|_1 + \|\frac{1}{\delta_j^2}\|_1 \|\gamma_j\|_1 = O(\sqrt{s_j}),$$

$$\|\hat{\Theta}_j\|_1 \le \|\hat{\Theta}_j - \Theta_j\|_1 + \|\Theta_j\|_1 = O_p(s_j\sqrt{\frac{\log p}{n}}) + O(\sqrt{s_j}) = O_p(\sqrt{s_j}),$$

which completes the proof of Lemma 2.A.3. □

Chapter 3

Portfolio Selection with Machine Learning: Sparsity, Correlation and Constraints

3.1 Introduction

The mean variance efficient portfolio theory (MVE), put forward by Markowitz (1952), is a milestone in finance literature. However, despite its theoretical elegance, MVE performs poorly out-of-sample due to difficulties in precisely estimating two important ingredients in the optimization problem: the expected asset returns and covariances. Jagannathan and Ma (2003) argue that *[... the estimation error in the sample mean is so large that nothing much is lost in ignoring the mean altogether when no further information about the population mean. (pp.1652-1653)]*, and Michaud (1989) suggests that the MVE portfolio optimization problem is actually "error maximization" in practice. Although many efforts have been made to improve the performance of optimized portfolio strategies,[1] DeMiguel et al. (2009b) demonstrate that the simple 1/N strategy (i.e. all assets are equally weighted) outperforms 14 other optimized portfolio strategies, making this non-optimized naive diversification strategy a competitive benchmark in comparing portfolio selection strategies. This paper builds upon and extends the norm constrained optimization strategy by DeMiguel et al. (2009a) and Fastrich et al. (2015) while incorporating stock correlation considered

[1] See Section 3.2 for a detailed discussion.

in DeMiguel et al. (2014). However, our method differs in several ways. First, we introduce a newly developed machine learning tool - the Ordered-Weighted-LASSO (OWL) estimation method, which encompasses the LASSO norm constraint considered in DeMiguel et al. (2009a). The OWL estimator is tailored for estimating and identifying correlated variables while the LASSO estimator is often criticized for lacking robustness in correlated data, see Figueiredo and Nowak (2016) and Kozak et al. (2020) for example. Second, DeMiguel et al. (2014) find that asset correlation is important for portfolio performance and they implement a VAR(1) model to harness the lagged correlations between asset returns. However, contemporaneous correlations, which are important for cross-sectional asset allocations (i.e. 'stock picking'), are left unexploited. On the other hand, the OWL estimation method specifically exploits contemporaneous correlations between assets. Third, our optimization problem embeds a novel norm constraint as well as allowing investors to impose additional weight constraints based on their beliefs. For example, investors can set up an upper and/or lower bound of the investment weight for each asset based on their prior beliefs. We develop efficient algorithms that achieve fast convergence for such optimization problem.

Empirically, we test our strategies in five asset classes. First, we consider the Fama-French 25 (FF25) portfolios because of their popularity as test assets in the finance literature and because sorted portfolios are less prone to large variations in returns compared to individual stocks. Second, we consider the S&P 500 (SP500) stocks with *daily* return frequency and we rebalance our hedge portfolio either weekly or monthly. Third, we also consider the S&P 100 (SP100) stocks with *monthly* return frequency and we rebalance our hedge portfolio monthly.[2] S&P 500 and S&P 100 stocks are usually the largest stocks in the market. To test our strategies on small and medium stocks, we adopt the randomly selected stocks approach, similar to Jagannathan and Ma (2003) and DeMiguel et al. (2009b): in April each year, we randomly select 500 stocks for daily return series (and 100 stocks for monthly return series) which have no missing data in the past 3 years (10 years for monthly return series) and in the next 1 year. The randomly selected 500 stocks with *daily* returns and 100 stocks with

[2] We require an invertible sample covariance matrix as an input in our optimization problem, thus we need the time series dimension larger than the cross-sectional dimension to obtain a non-singular covariance matrix.

monthly returns consist of our fourth and fifth test assets classes.

We adopt an out-of-sample procedure to compare and test each strategy. At each point of time, we use a rolling window to estimate each stock's weight (portfolio's weight in the FF25 universe) to invest for the next period, then we roll the training sample forward until the next rebalancing point and compute hedge portfolio return and turnovers. Turnover is the change in portfolio weights right before and after rebalancing. In the end, we obtain a sequence of out-of-sample returns and portfolio weights, from which we can compute the out-of-sample Sharpe ratio and turnovers. Notably, transaction cost, which is a monotonically increasing function of turnovers, is an important consideration for investors. So we also consider a transaction cost adjusted Sharpe ratio (TCadjSR), which will be one of our main comparison criteria. We also introduce the model confidence set (MCS) method of Hansen et al. (2011) to compare each strategy. Sharpe ratio, formulated as the ratio between mean return and portfolio risk (i.e. standard deviation of portfolio returns), is often dominated by the portfolio risk component when mean returns are small. The MCS method answers this question: which strategy offers *statistically* the best out-of-sample returns, while portfolio risk controls the confidence band?

Our empirical findings complement some existing literature and shed light on new perspectives of portfolio selection theory. First and foremost, DeMiguel et al. (2009b) show that the naive 1/N strategy outperforms 14 optimization-based strategies. Our method bridges the gap between the naive diversification strategy and a well-defined optimization framework. We demonstrate that asset correlations play important role in our optimization method. Together with hyper parameters λ_1 and λ_2 they jointly determine portfolio positions: λ_1 controls the overall sparsity of the asset selection, which is a similar role to the LASSO norm constraint in DeMiguel et al. (2009a); λ_2 controls the sensitivity to correlations. Large λ_2 encourages weights-clustering (i.e., stocks will be assigned with similar weights), thus portfolio weight will be close to the equal weighted position. We rigorously compare the 1/N strategy and our OWL related strategies and find that OWL related strategies consistently outperform the 1/N strategy in all asset classes, in terms of Sharpe ratio, turnover and mean returns. It is a remarkable discovery, since in the related literature, we can find hardly any strategy that can outperform the 1/N strategy by *both* the Sharpe ratio *and* the turnover criteria, see DeMiguel et al. (2009b) for example. Therefore, we argue that the OWL optimization method serves as an enhancement to the naive 1/N strategy: OWL related strategies obtain near-equal-weighted positions within an optimization framework, where dedicated parameters control model sensitivity to asset correlations and sparsity.

Second, OWL related strategies perform better in large stocks than small stocks and in monthly returns than daily returns. Using Fama-French 25 portfolios and the Standard & Poor 100 stocks with monthly returns, we find that OWL related strategies outperform other candidate strategies in terms of the Sharpe ratio and transaction cost adjusted returns. However, we also find the performance of OWL related strategies is less effective using randomly selected 500 stocks. A stylized fact is that large stocks with monthly returns exhibit higher cross-sectional correlation than smaller or daily returns. This empirical finding confirms that the crucial ingredient in our optimization method is stock correlation: OWL related strategies do well when assets are correlated.

Third, the Mean Variance Efficient portfolio (MVE) performs poorly due to excessive estimation errors in expected returns and asset covariance matrix. However, the OWL embedded MVE (MVE-OWL) strategy together with weight constraints produces sizable Sharpe ratio and low turnovers. The MCS test confirms that the MVE-OWL strategy produces *significantly* larger portfolio returns than all other strategies using FF25 as test assets. This finding challenges some common stances in the existing literature. Because the sample mean return is such a noisy approximation of the expected return, many researchers find it is better just to focus on the portfolio risk and optimizing the minimum variance portfolio. Our finding suggests that it is possible to optimize the MVE portfolio with sample mean if we further impose the OWL norm constraint and additional weight constraints.

The rest of this paper is organized as follows. Section 3.2 reviews related literature. Section 3.3.1 outlines some popular methodologies employed for portfolio optimization problems, which are also used as benchmarks in our empirical analysis. Sections 3.3.2 - 3.3.3 introduce the OWL shrinkage method and discuss its statistical properties before devising an ADMM algorithm to solve the OWL optimization problem with multiple constraints. Section 3.4 applies OWL based portfolio strategies on five different asset classes and compares them with other benchmarks.

3.2 Literature review

This paper naturally builds on a strand of literature devoted to exploring the portfolio optimization theory. Since the groundbreaking work of Markowitz (1952), modern portfolio theory has evolved rapidly. However, Markowitz's portfolio theory has long been criticized for working poorly empirically, because one needs to obtain the *ex-ante* expected returns and covariances of stock returns, which are difficult to be estimated with precision. Michaud (1989) looks into the "Markowitz optimization enigma" and finds that the mean variance optimization is in fact "error maximization". DeMiguel et al. (2009b) study the simple equal weighted strategy and find it outperforms 14 other optimization based strategies. They argue that estimation error in the expected asset return or covariances erodes any gains from optimization. Kan and Zhou (2007) show that using the sample analogs for the expected returns and covariance matrix can lead to very poor out-of-sample performance due to parameter uncertainty. They suggest that holding the tangent portfolio and the risk free asset is not optimal, though holding some other risky portfolios will help reduce the portfolio risk caused by parameter uncertainty. Ledoit and Wolf (2003) propose a shrinkage based estimation method for the covariance matrix. They suggest that shrinking the sample covariance matrix linearly towards a target matrix (for example the identity matrix) will improve the out-of-sample performance of the minimum variance portfolio. Ledoit and Wolf (2017) propose a non-linear version of shrinkage estimator for the covariance matrix which shows better performance than the linear version. Jagannathan and Ma (2003) suggest no-short-sale constraint on all stocks and find significant gains in out-of-sample Sharpe ratio for the minimum variance portfolio. They argue that such constraint helps reduce the upward biased estimation errors in the variance-covariance matrix. DeMiguel et al. (2014) show that stocks are correlated, and by implementing a VAR(1) model they obtain substantial gains in portfolio performance. DeMiguel et al. (2020) consider a portfolio optimization problem by selecting a large number of firm characteristics while embedding the transaction cost in their object function.

This paper is also related to a new and fast growing field which uses machine learning techniques for portfolio optimization problems. DeMiguel et al. (2009a) propose norm constraints on portfolio weights for the minimum variance portfolio. In particular, they consider separately the LASSO and Ridge penalties on portfolio weights and find significant improvement on the out-of-sample Sharpe ratio of the minimum variance portfolio. Ao et al. (2018) combine the unconstrained regression with LASSO penalty and achieve superior portfolio performance. Inspired by the adaptive LASSO in Zou (2006), Fastrich et al. (2015) incorporate the financial information into the adaptive weights to determine the portfolio composition. Figueiredo and Nowak (2016) study the ordered and weighted LASSO estimator and show that it has appealing property of clustering correlated variables by assigning them with similar coefficients.

This paper is closely related to DeMiguel et al. (2014) and DeMiguel et al. (2009a). However, our portfolio optimization method differs in several ways. First, we endeavor to exploit the *contemporaneous* correlation between stocks instead of serial correlations considered in DeMiguel et al. (2014). Therefore, our strategies are more relevant to “stock-picking” investors. Second, we propose a novel norm constraint on the portfolio weights which encompasses the LASSO norm constraint in DeMiguel et al. (2009a), and our novel norm constraint can specifically exploit contemporaneous stock correlations. Third, we devise efficient algorithms that enable investors to incorporate their prior beliefs into the optimization problem (i.e., investors can set upper/lower bound on the weight of each individual asset based on prior beliefs).

3.3 Methodology

We first consider a simple case of the OWL shrinkage method in the portfolio optimization problem where no constraints are imposed on the weights of individual stocks. Then we further impose constraints on portfolio weights for the portfolio optimization problem.

3.3.1 Setup

Consider N assets in the investment universe. Denote by R_t the returns of N assets in the excess of risk-free rate at time t. Denote by μ $(N \times 1)$ and Σ $(N \times N)$ the population mean and population variance-covariance matrix of N asset returns, while $\hat{\mu}$ and $\hat{\Sigma}$ are their sample estimates. An investor, according to Markowitz (1952)'s classical portfolio theory, aims to maximize the risk-adjusted portfolio returns, or equivalently:

$$\begin{aligned} &\min_{w} \quad \left(\frac{\gamma}{2} w'\Sigma w - \mu' w\right) \\ s.t. \quad &w'e = 1 \end{aligned} \tag{3.1}$$

where γ is a scalar that represents the investor's absolute risk aversion, w is the $N \times 1$ weighting vector of N assets, also referred to as positions, and e is a column vector of ones. The closed-form solution of the above optimization problem is $w = \frac{1}{\gamma}\Sigma^{-1}\mu$. However, Σ and μ are unobservable. Typically, we use the sample analogs $\hat{\mu}$ and $\hat{\Sigma}$ in the above equation, which gives

$$w_{MVE} = \frac{1}{\gamma}\hat{\Sigma}^{-1}\hat{\mu}. \tag{3.2}$$

Michaud (1989), DeMiguel et al. (2009b) and Jagannathan and Ma (2003) have pointed out that the sample analogs of μ and Σ are subject to large estimation errors. In particular, the estimation of the expected asset returns (μ) proves to be extra challenging. In fact, the estimation error (of μ) is so large that it offsets all gains from optimization (DeMiguel et al., 2009b). So in practice, focusing on the minimum variance (minVar) portfolio proves to have better out-of-sample results than the mean-variance efficient (MVE) portfolio. In this regard, the optimal weights for the minimum variance portfolio are obtained by optimizing (3.1) while setting $\mu = 0$, that is

$$w_{minVar} = \frac{\hat{\Sigma}^{-1}e}{e'\hat{\Sigma}^{-1}e}, \tag{3.3}$$

where we use the sample analog for Σ. For the estimation error in the sample covariance matrix, shrinkage estimator proves to be a useful remedy. For instance, Ledoit and Wolf (2003) propose to shrink the sample covariance matrix towards a target matrix. They suggest the following estimator

$$\hat{\Sigma}_{LW03} = \delta\hat{\Sigma} + (1-\delta)\hat{\Sigma}_{target}, \tag{3.4}$$

where $\delta \in (0,1)$ is a shrinkage intensity parameter and $\hat{\Sigma}_{target}$ is a target estimator, which can be, for example, the identify matrix.

DeMiguel et al. (2009a) show that imposing norm constraints on portfolio weights to shrink them towards zeros substantially improves the out-of-sample Sharpe ratio of the hedged portfolios.[3] They suggest the following norm shrinkage methods

$$\min_{w} \quad \left(\frac{\gamma}{2} w'\Sigma w - \mu' w + \lambda ||w||_1\right), \tag{3.5}$$

or

$$\min_{w} \quad \left(\frac{\gamma}{2} w'\Sigma w - \mu' w + \lambda ||w||_2^2\right), \tag{3.6}$$

where

$$||w||_1 = \sum_{i=1}^{N} |w_i|, \qquad ||w||_2^2 = \sum_{i=1}^{N} w_i^2,$$

and λ is a shrinkage intensity parameter. Shrinkage method in (3.5) is broadly known as LASSO shrinkage, which produces sparse estimator for w, while (3.6) is referred to as Ridge shrinkage, which shrinks all elements in w towards zero.

Jagannathan and Ma (2003) find that imposing no-short-sale constraints on portfolio weights helps to improve out-of-sample Sharpe ratio. They propose the following constraint in addition to (3.1):

$$w_i \geq 0, \quad \text{for all} \;\; i \in \{1, ..., N\}. \tag{3.7}$$

They argue that imposing this constraint leads to substantial reduction of extreme negative positions of stocks which are caused by upward biased estimation of variances.

In addition, DeMiguel et al. (2014) reveal that stock correlations matter for portfolio construction. They propose a vector-autoregressive (VAR) model to capture stocks' serial correlations and find that VAR-based portfolios outperform traditional unconditional portfolios. To fix ideas, let us assume that the vector of asset return R_t follows a VAR(1) process,

$$R_{t+1} = a + BR_t + \epsilon_{t+1}, \tag{3.8}$$

where a is a $N \times 1$ vector of intercepts, B is a $N \times N$ matrix of parameters, and

[3] We set $\mu = 0$ for the minimum variance portfolio. Otherwise, we are optimizing the mean-variance efficient portfolio.

ϵ_t is the *i.i.d.* error term. However, equation (3.8) is a reduced model, which suggests that tomorrow's expected stock returns depend linearly on today's return. The linear dependence is characterized by the coefficient matrix B, which describes the lagged cross-sectional and serial dependence. On the other hand, contemporaneous correlations between stocks are left unexplained.

This paper builds on and extends DeMiguel et al. (2009a) and DeMiguel et al. (2014). We introduce a newly developed machine learning tool, the ordered and weighted LASSO (OWL), which (1) encompasses the LASSO shrinkage method in DeMiguel et al. (2009a); (2) exploits *contemporaneous* correlations between stocks (not reduced model), drawing a distinctive line between our portfolio optimization approach and that in DeMiguel et al. (2014); (3) enables adopting bespoke constraints on portfolio weights if investors have prior beliefs.[4] We devise efficient algorithms to solve the OWL optimization problem with/without additional constraints on portfolio weights.

3.3.2 The OWL shrinkage method

We follow the idea of DeMiguel et al. (2009a) to add a penalty term in the object function $f(w)$, which measures the loss given portfolio weight w. The optimization problem can be written as

$$\hat{w} = \arg\min_{w} f(w) + \Omega_\omega(w), \qquad \Omega_\omega(w) = \omega' |w|_\downarrow, \tag{3.9}$$

where ω is a pre-specified weighting vector which will be specified in (3.11) below. w is a vector of stock weights (positions) and $|w|_\downarrow$ is the vector that stores the absolute value of stock weights, decreasingly ordered by its magnitude. Both ω and $|w|_\downarrow$ take values in a monotone non-negative cone κ, which is defined as $\kappa := \{x \in R^N : x_1 \geq x_2 \geq ... \geq x_N \geq 0\}$. $f(w)$ can be any continuously differentiable function of w. However, in this paper we focus on the mean-variance efficient portfolio (or the minimum variance portfolio if we set $\mu = 0$), which corresponds to

$$f(w) = \frac{\gamma}{2} w' \Sigma w - \mu' w. \tag{3.10}$$

[4] Prior beliefs could come from investors' exclusive information, or an existing trading strategy. In that respect, it can be used as an improvement/refinement of any existing strategies.

We further specify ω to have a linear weighting structure

$$\omega_i = \lambda_1 + (N - i)\lambda_2, \tag{3.11}$$

where λ_1 and λ_2 are two hyper parameters which pin down ω, and the values of λ_1 and λ_2 are determined through cross validation.[5] In order to solve the optimization problem in (3.9) - (3.11), we use the proximal descent algorithm, more details about this algorithm can be found in Sun (2019).

Next, we discuss some econometric properties of the OWL shrinkage method.

Lemma 3.3.1. *Suppose that the pre-specified weighting vector ω of the OWL shrinkage method is defined in (3.11). If λ_2 is set to be zero, then the OWL shrinkage method is equivalent to the LASSO shrinkage as in (3.5), or equivalently*

$$\lambda \|w\|_1 = \Omega_\omega(w).$$

Proof: see Appendix 3.A.2.3.

Lemma 3.3.1 shows that the OWL shrinkage method encompasses the LASSO shrinkage method used by DeMiguel et al. (2009a). Furthermore, once we adopt a linear weighting scheme for ω as in (3.11), the OWL shrinkage method is linked to the OSCAR regularization introduced in Bondell and Reich (2008), which has appealing properties of clustering correlated features. The OSCAR regularization unit is defined as

$$\Omega_{OSCAR}(w) = \lambda_1 ||w||_1 + \lambda_2 \sum_{i<j} \max\{|w_i|, |w_j|\}, \tag{3.12}$$

which is a combination of the LASSO regularization (ℓ_1 norm) unit and a pair-wise ℓ_∞ norm unit.

Lemma 3.3.2. *Suppose that the pre-specified weighting vector ω of the OWL shrinkage method is defined in (3.11). Then the OWL shrinkage method is equivalent to the*

[5] In particular, we set a grid value of λ_1 and λ_2. Then, at each point on the grid, we split the sample into 5 folds, using 4 folds to evaluate the model and obtain the estimated parameters. Then we use the other 1 fold as out-of-sample to evaluate the Mean Square Forecast Errors (MSE). We rotate each fold as the out-of-sample fold, and compute the average MSE. We repeat these procedures on each grid, and compare the average MSE for each point on the grid. The one receiving the smallest average MSE will determine the hyper parameter values.

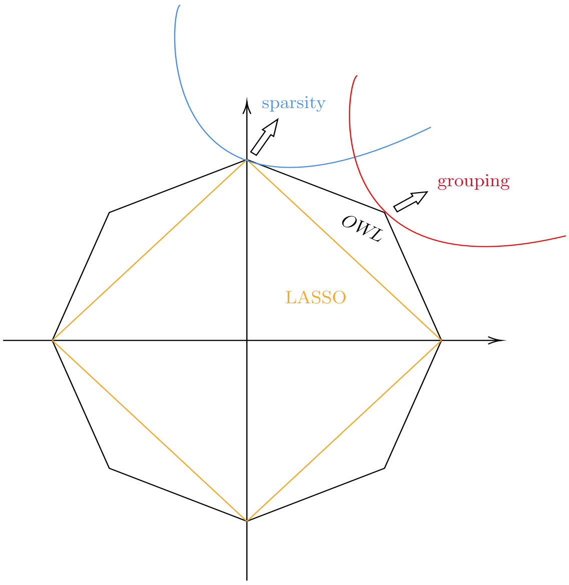


Figure 3.1. Geometric interpretation of the atomic norm of LASSO and OWL regularization

OSCAR shrinkage method as in (3.12), *or equivalently*

$$\Omega_{OSCAR}(w) = \Omega_\omega(w).$$

Proof: see Appendix 3.A.2.4.

Lemma 3.3.2 shows that by adopting a linear decreasing weighting scheme for ω as in (3.11), the OWL shrinkage method is equivalent to the OSCAR regularization, which has property of clustering correlated variables. However, the OWL shrinkage is a more general method than the OSCAR regularization. For instance, by adopting a non-linear (for instance, the inverse of the normal cumulative distribution function) weighting scheme for ω, the OWL shrinkage method is equivalent to the SLOPE estimator proposed by Bogdan et al. (2015), which is widely used in multiple testing. In the scope of this paper, we restrict the weighting vector ω as defined in (3.11) because of the clustering property offered by the OSCAR regularization and our objective of exploiting the contemporaneous correlations between stocks. To gain some impression of how the OWL shrinkage method achieves sparse selection and correlation identification simultaneously, we first look at the geometric interpretation of the atomic norm of $\Omega_\omega(w)$ in Figure 3.1.

Figure 3.1 depicts the atomic norm of OWL and LASSO regularization in a two-dimensional space. We can see that the atomic norm of LASSO has all vertices on axes, which encourages sparse selection of variables. On the other hand, the atomic norm of the OWL regularization is octagonally shaped, having vertices on both axes and the ±45 degree lines. The former (vertices on axes) encourages sparse selection and the latter (vertices on the ±45 degree lines) encourages variable grouping.[6]

The geometric interpretation offers a ballpark explanation of how the OWL shrinkage achieves both sparse selection and correlation identification (grouping) simultaneously. Next, we formally investigate some econometric properties for the OWL shrinkage method. There is a rich literature in finance focusing on the sparse selection property offered by LASSO type of estimators, see DeMiguel et al. (2020, 2009a), Fastrich et al. (2015) and Chinco et al. (2019) for example. The OWL shrinkage method encompasses and shares similar sparse-selection-properties of the LASSO estimator, and see Sun (2019) for a formal investigation of the asymptotic property of the OWL estimator. For this reason we focus on investigating the grouping property in this paper. Theorem 3.3.1 and 3.3.2 below state the conditions that need to be satisfied to have the grouping property.

Theorem 3.3.1. *Let $\Sigma_{i.}$ and $\Sigma_{j.}$ denote the i^{th} and j^{th} columns of the variance-covariance matrix and λ_2 be the parameter defined as in (3.11). Suppose the loss function is defined as in (3.10) while setting $\mu = 0$ (i.e. minimum variance portfolio). If*

$$\|\Sigma_{i.} - \Sigma_{j.}\|_2 < \lambda_2,$$

then $\hat{w}_i = \hat{w}_j$, where $\hat{w}_i$ and $\hat{w}_j$ are obtained by optimizing (3.9).
Proof: see Appendix 3.A.2.1.

Theorem 3.3.1 shows that in the minimum variance portfolio optimization problem, if two assets are highly correlated, i.e. $\|\Sigma_{i.} - \Sigma_{j.}\|_2$ is small, then they will receive the same positions $\hat{w}_i = \hat{w}_j$. We regard this property as grouping. The tuning parameter λ_2 plays an active role in influencing the grouping property: a larger λ_2 means the atomic norm of the OWL regularization in Figure 3.1 is more pointy, i.e. the

[6] Sparse selection means the vertices on axes will assign one variable zero coefficient and another non-zero (in this 2-dimensional space), thus performing sparse selection. The variable assigned with zero coefficient is shrunk off. Variable grouping means variables exhibiting high correlations will be assigned with the same or similar coefficients. The vertices on the ±45 degree lines will dictate the tangent point with the contour from the un-regularized solutions, which give the same or similar coefficients to both variables.

vertices on the ±45 degree lines extend further out, and hence it encourages more grouping.

Theorem 3.3.2. *Let $\Sigma_{i.}$ and $\Sigma_{j.}$ be defined as in Theorem 3.3.1. Denote by μ_i, μ_j the expected returns of the i^{th} and j^{th} asset. Let γ represent investor's risk aversion level. Suppose the loss function is defined as in (3.10) (i.e. mean-variance efficient portfolio). If*

$$\gamma \|\Sigma_{i.} - \Sigma_{j.}\|_2 + |\mu_i - \mu_j| < \lambda_2,$$

then $\hat{w}_i = \hat{w}_j$, where $\hat{w}_i$ and $\hat{w}_j$ are obtained by optimizing (3.9).
Proof: see Appendix 3.A.2.2

Theorem 3.3.2 extends Theorem 3.3.1 into a mean-variance efficient portfolio optimization problem where investors care about both risks and expected returns. We find that, compared to the minimum variance portfolio, both the difference in the expected returns $|\mu_i - \mu_j|$ and correlation with other assets $\|\Sigma_{i.} - \Sigma_{j.}\|_2$ influence the grouping property: if two assets have similar expected returns and are similarly correlated with other assets (i.e. $|\mu_i - \mu_j|$ and $\|\Sigma_{i.} - \Sigma_{j.}\|_2$ are small), then they are likely to be grouped together (i.e. $\hat{w}_i \approx \hat{w}_j$). The risk aversion parameter γ can be viewed as a scaling parameter adjusting weights between the risk component and the expected return component. Also, large λ_2 encourages grouping.

It is worth stressing that we derive the grouping property using the population values of the variance-covariance matrix Σ and the expected returns μ, which are unobservable and difficult to estimate with precision. However, the proof of Theorem 3.3.1 and 3.3.2 does not depend on the asymptotic properties of Σ or μ. In other words, we arrive at those results only using properties of the OWL regularization, and those results are also applicable to $\hat{\Sigma}$ and $\hat{\mu}$, which are sample analogs of Σ and μ. It is well known that large estimation errors in those sample analogs erode any gains in optimization. In the next subsection, we set out to constrain portfolio weights while using these sample analogs $\hat{\mu}$ and $\hat{\sigma}$ to mitigate estimation errors.

3.3.3 The OWL optimization problem and the ADMM algorithm

Let us consider a more common problem, where investors have some prior information on stocks. For instance, an investor may hold positive opinions on some specific stocks while negative on others, thus she may want to impose some bounds constraints on those stocks. To generalize those constraints, we impose the following inequality

$$\mathbf{lb} \preceq w \preceq \mathbf{ub}, \tag{3.13}$$

where **lb** (**ub**) is a lower (upper) bound for the vector of portfolio weights w. For any $x, y \in R^N$, $x \preceq y$ implies $x_i \leq y_i$, for all $i \in \{1, ..., N\}$. However, the optimization problem of (3.9) with constraint (3.13) is challenging to solve with gradient descent algorithm, which is commonly used in machine learning algorithms. Hence, we introduce a newly developed ADMM (Alternating Direction Method of Multiplier, Boyd et al. (2010)) algorithm to solve this constrained optimization problem. The outline of the algorithm is the following: equation (3.9) consists of two components, one is $f(w)$ which is differentiable with respect to w, another is $\Omega_\omega(w)$ which is not differentiable with respect to w. In order to make computation easier and tangible, we introduce a new variable v, and replace it with w in the undifferentiable component $\Omega_\omega(w)$, so that we can optimize each component separately. In addition, we impose an extra constraint that these two random variables are equal $w = v$. For this reason, this algorithm is named "alternating direction method of multiplier". Therefore, the constrained OWL optimization problem can be written as

$$\min_{w,v \in R^N} \left[\frac{\gamma}{2} w' \hat{\Sigma} w - \hat{\mu}' w + \Omega_\omega(v) \right], \tag{3.14}$$

$$s.t. \quad w = v, \tag{3.15}$$

$$w'e = 1, \tag{3.16}$$

$$w \succeq \mathbf{lb}, \tag{3.17}$$

$$w \preceq \mathbf{ub}, \tag{3.18}$$

where $\Omega_\omega(v) = \omega' |v|_\downarrow$ defined similarly as in (3.9), $\hat{\Sigma}$ and $\hat{\mu}$ are sample analogs of Σ and μ, and e is a column vector of ones. For the technical details of the ADMM algorithm, see Appendix 3.A.1.

3.4 Empirical Analysis

In this section, we apply the OWL shrinkage method with or without additional constraints on portfolio weights and compare them with other portfolio strategies in the literature. We consider five different asset classes with daily and/or monthly returns. The variety in characteristics of these asset classes summarizes the pros and cons of each strategy which we will discuss in details later.

In Section 3.4.1, we first introduce the data (five asset classes) and all candidate strategies. Section 3.4.2 explains the empirical method we employ to compare strategies and Section 3.4.3 outlines the comparison criterion we use for ranking strategies. To test the significance of difference in Sharpe ratios, we employ a bootstrap based Sharpe ratio test by Ledoit and Wolf (2008). To test which strategies statistically produce higher returns, we use the model confidence set (MCS) test by Hansen et al. (2011).

3.4.1 Data

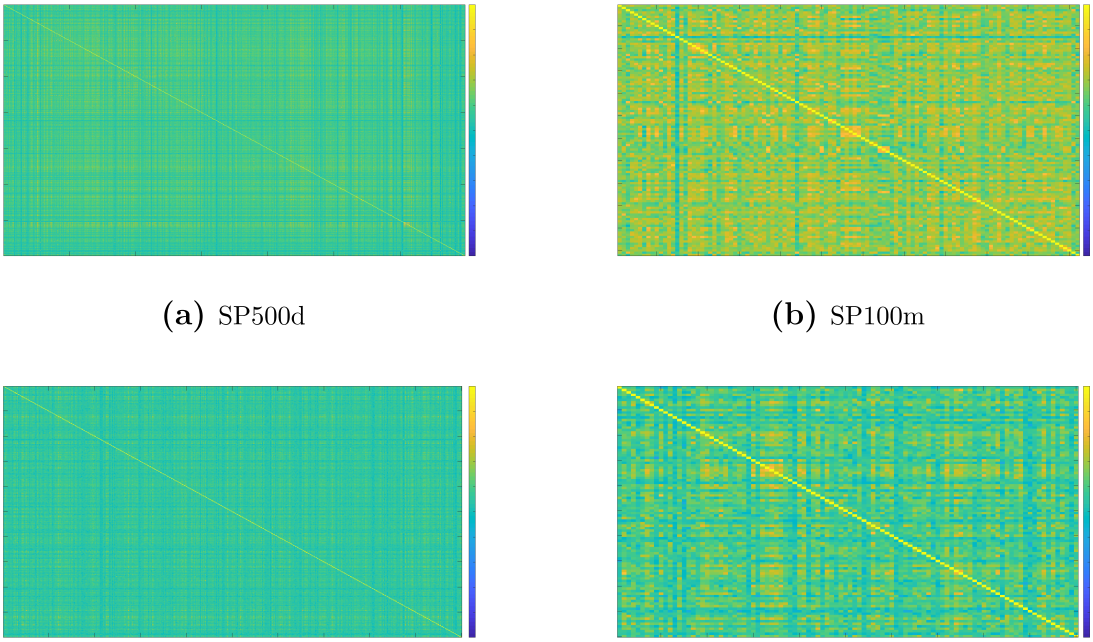
We first consider the Fama French 25 portfolios (FF25) from July 1927 to December 2017.[7] FF25 is obtained by sorting stocks into five by five tranches according to their size and book-to-market ratio. The return of each tranche is the average returns of a large number of stocks allocated to the tranche sharing similar characteristics (size and book-to-market ratio in this case). Since returns of these tranches are averaged returns of many stocks, they are less prone to large variations caused by idiosyncratic risks.

We then consider the S&P 500 stocks with *daily* returns (SP500d), from 1st January 1978 to 31st December 2017 and the S&P 100 stocks with *monthly* returns (SP100m) from January 1978 to December 2017. Stock return data are obtained from the CRSP stock return files from Wharton Research Data Services.

While S&P 500 and S&P 100 stocks are typically the largest stocks in the market,

[7] Data is downloaded from Kenneth French's website at
$http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data\_library.html$

to investigate stocks with medium or small sizes, we follow DeMiguel et al. (2009a) and Jagannathan and Ma (2003) to consider the randomly selected 500 stocks from the CRSP *daily* return file (CRSP500d). We also conduct a similar procedure to randomly select 100 stocks from the CRSP *monthly* return file (CRSP100m).



(a) SP500d

(b) SP100m

(c) CRSP500d

(d) CRSP100m

Figure 3.2. Correlation coefficient matrices of four asset classes

Note: yellow and deep blue indicate high correlation, while green indicating low correlation.

Figure 3.2 shows the correlation coefficient matrices of SP500d, SP100m, CRSP500d and CRSP100m stocks, respectively. We observe that SP500d returns exhibit higher correlations than the randomly selected CRSP500d returns. Similarly, we observe that the SP100m exhibit higher correlations than CRSP100m returns. These patterns may reflect the fact that large stocks are less prone to idiosyncratic noises and more affected by market-wide common factors than small stocks, so large stocks exhibit higher correlations than small stocks. Meanwhile, we also observe that, by comparing the left panels and right panels in Figure 3.2, monthly returns shows higher correlations than daily returns. This can be characterized as the Epps effect (Epps, 1979): the sample correlation tends to be biased towards zero when the sampling frequency progressively shrinks.

Next, Table 3.1 lists all considered candidate strategies. First, we consider the

Table 3.1. **Candidate strategies**

Abbreviation	Strategies	Source
EW (1/N)	equal weighted	DeMiguel et al. (2009b)
minVar	minimum variance portfolio	N/A
minVar-JM	minVar with no-short-sale constraint	Jagannathan and Ma (2003)
minVar-LW	minVar with Ledoit-Wolf shrinkage	Ledoit and Wolf (2003)
minVa-OWL	OWL shrinkage on minVar	New Proposal
minVar-OWL-Pos	OWL shrinkage with no-short-sale constraint	New Proposal
minVar-OWL-bounds	OWL shrinkage with bounds constraints	New Proposal
minVar-LW-OWL	OWL with LW shrinkage on Cov matrix	New Proposal
minVar-hard-OWL	OWL with hard-thresholding for Cov matrix	New Proposal
MVE-OWL-Pos	OWL shrinkage with no-short-sale constraint on MVE	New Proposal
MVE-OWL-bounds	OWL shrinkage with bounds constraints on MVE	New Proposal

equal weighted (EW, also known as 1/N) strategy which has attracted great attention after DeMiguel et al. (2009b) showing that this non-optimized naive diversification strategy achieves superior out-of-sample performance against other optimized ones. We also consider the no-short-sale constraint on the minimum variance portfolio by Jagannathan and Ma (2003) and the linear shrinkage method by Ledoit and Wolf (2003) in our candidate strategies. In the newly proposed OWL shrinkage methods, we focus on the "minVar-OWL" method which implements the OWL shrinkage method on the minimum variance portfolio without additional constraints. The rest are some enhanced OWL strategies. For instance, "Pos" indicates that we further impose a no-short-sale constraint on stock weights. "Bounds" indicates that we impose upper and lower bounds for each stock. In this case, since we do not hold any additional information about each stock in our exercise, we blindly impose a bound constraint between -5% and 30% for all stocks. "Hard" indicates a hard-thresholding method for estimating the covariance matrix as in Bickel and Levina (2008) and Dendramis et al. (2019).

Next, we set out to conduct out-of-sample based empirical methods to implement each strategy and compare their performances using various criteria.

3.4.2 Empirical methods

For the FF25 asset class, since returns are sorted portfolio returns, we have balanced panel data, which is convenient for our analysis. We choose a rolling window size, say five years (60 months). At time t, we use the recent 60 months (from $t-59$ to t) data to estimate the model with each strategy and obtain the weighting vector for the next period. At the beginning of $t+1$ we invest in each 25 portfolios according to the weighting vector we obtained at time t. Then, at the end of $t + 1$, returns will be realized, so we can compute the returns for the hedge portfolio. Next, we roll the window one month forward (from $t - 58$ to $t + 1$) to estimate the weight for next month's investment.

For SP500d, we first find all stocks that have been in SP500 index at least once between January 1978 and December 2017, total 1439 stocks. Then we implement a rolling window scheme, with rolling window size equal to 750 working days, approximately 3 years. In each rolling window, we remove stocks having missing data, which typically leaves 500 to 700 stocks in the investment universe in each rolling window. We then perform various portfolio selection strategies, and get weights for stocks which constitute next period's investment amount. We consider two rebalance frequencies, weekly or monthly. When the rebalance period is met, we compute the portfolio's return and turnover. Then we move forward to the next rolling window and repeat these steps until the end of out-of-sample period. We follow a similar procedure for the SP100m dataset except we rebalance only monthly and use a rolling window size of 10 years (120 months).

For CRSP500d, we follow DeMiguel et al. (2009a)'s procedure. In April each year, we randomly choose 500 stocks that have no missing data for the past 10 years as well as the following one year. Then in each rolling window, with window size equal to 750 working days, we estimate weights using various strategies. We also consider rebalancing portfolios weekly or monthly. At each rebalance point, we compute OOS portfolio returns and turnovers. At the end of the out-of-sample period we can compute out-of-sample returns, standard deviation, Sharpe ratio and turnovers. We follow a similar procedure for CRSP100m, except the rolling window size is 10 years (120 months), and rebalance monthly only.

3.4.3 Out-of-sample comparison

To compare our OWL based strategies with other existing ones in the literature, we consider the following criteria: (1) the out-of-sample Sharpe ratio (SR), (2) portfolio turnovers (transaction cost), (3) transaction cost adjusted Sharpe ratios (TCadjSR), (4) transaction cost adjusted portfolio returns (TCadjR). We follow the methodology of DeMiguel et al. (2009a) to construct the first two criteria. We add the third criterion because the first two criteria look at Sharpe ratio and turnover separately, which leads to (on many occasions) contradictory preferences: one strategy that delivers higher SR usually entails higher turnover (transaction cost), and vice verse. TCadjSR allows one to look into and compare strategies in a complete fashion. In addition, Sharpe ratio comparison is usually dominated by its variance component and, by definition, the minVar portfolio typically delivers much lower portfolio risk than the MVE portfolio. Hence, we introduce the fourth criterion, the model confidence set (MCS) of Hansen et al. (2011). MCS compares TCadjR and puts all strategies that *significantly* produce the highest returns in a set while portfolio risk controls the confidence level. More details about the MCS method are included in Section 3.4.5. We argue that looking at the criteria three and four together gives a more completed profile of portfolio performance.

To fix ideas, let $r_t = (r_{1,t}, r_{2,t}, \cdots, r_{N,t})'$ denote the vector of N asset returns in excess of the risk-free rate $r_{f,t}$ at time t and w_t denote the vector storing portfolio weights of N assets at time t. For the monthly dataset, we choose a rolling window size $\tau = 120$ months, where $\tau \ll T$ and T is the total number of time series observations. We set the rebalancing frequency as monthly ($q = 1$ month). For the daily data set, we choose a rolling window size of three calendar years (about 756 time series observations) and rebalance either weekly ($q = 5$ days) or monthly ($q = 21$ days). At time t, we estimate portfolio weights w_t using data from $t - \tau + 1$ to t for each strategy. w_t will be the investment amount at the beginning of time $t+1$ and we hold this position until the next rebalancing point $t + q$. At time $t + q$, before rebalancing we need to compute the *weight before rebalance*. The weight of each asset changes between the beginning and the end of time t due to the price fluctuation. We re-calculate the weight at the end of time t using new stock prices and call it the "weight before rebalance" (w_{t^+}). Then at the beginning of time $t + 1$ we invest according to the new weight (w_{t+1}) obtained at the end of time t. The difference between them ($|w_{t+1} - w_{t^+}|$) is the *turnover*. More specifically, we compute the summation of absolute value of this difference at each point of time, then take the average across time. Then, we consider the following comparison criteria. For strategy i, the standard deviation of the out-of-sample return is

$$\hat{\sigma}^i = \sqrt{\frac{1}{|\Upsilon|}\sum_{t\in\Upsilon}\left(w_t^{i\prime} r_{t+q} - \hat{\mu}^i\right)^2}, \tag{3.19}$$

where $\hat{\mu}^i = \frac{1}{|\Upsilon|}\sum_{t\in\Upsilon} w_t^{i\prime} r_{t+q}$ is the mean of OOS returns and $t \in \Upsilon := \{x \mid \{x = \tau, \tau+q, \tau+2q, \cdots\} \cap \{x \leq T-q\}\}$. $|\Upsilon|$ denotes the cardinality of the set Υ which represents the number of times rebalancing portfolio. The Sharpe ratio of portfolio strategy i is

$$SR^i = \frac{\hat{\mu}^i}{\hat{\sigma}^i}, \tag{3.20}$$

and the turnover of portfolio strategy i is defined as

$$TO^i = \frac{1}{|\Upsilon|}\sum_{t\in\Upsilon}\sum_{j=1}^{N}\left(|w_{j,t+q}^i - w_{j,t^+}^i|\right), \tag{3.21}$$

where w_{t^+} indicates the weight before rebalancing due to price fluctuation between two consecutive rebalancing points, and w_{j,t^+}^i is the j^{th} element in $w_{t^+}^i$ for portfolio strategy i, where $j \in \{1, 2, \cdots, N\}$. Hence, $|w_{j,t+q}^i - w_{j,t^+}^i|$ measures the change in portfolio weight for asset j at rebalancing point $t+q$, and TO measures the average weight change for all $t \in \Upsilon$. Therefore, the transaction cost adjusted Sharpe ratio for portfolio strategy i is defined as

$$TCadjSR^i = \frac{\hat{\mu}^i - TC^i}{\hat{\sigma}^i}, \tag{3.22}$$

where

$$TC^i = TO^i * |\Upsilon| * cost\_per\_transaction \tag{3.23}$$

denotes the transaction cost. Recall that $|\Upsilon|$ is the cardinality of the set Υ, measuring the total number of times investors rebalance their portfolios. Rebalancing frequency q directly affects $|\Upsilon|$: higher frequency of rebalancing (i.e. smaller q) will result in larger $|\Upsilon|$, thus greater transaction cost. $cost\_per\_transaction$ can be interpreted as per US dollar transaction cost to trade stocks. In line with DeMiguel et al. (2013), we set $cost\_per\_transaction = 50$ basis points. Finally, the transaction cost adjusted returns for portfolio strategy i are defined as

$$TCadjR_t^i = w_t^{i\prime} r_{t+q} - TC^i. \tag{3.24}$$

Note that $t \in \Upsilon$ is a subscript indicating each rebalancing point. For Sharpe ratio comparison, we further utilize a Sharpe ratio test devised by Ledoit and Wolf (2008) to reveal whether Sharpe ratios are statistically different between portfolio strategies by pair-wise comparison. Specifically, we implement a circular block bootstrap method, which is robust to correlated returns.[8] Let μ_i (μ_j) denote the mean (excess) return and σ_i (σ_j) denote the standard deviation for strategy i (j). The null hypothesis is

$$H_0 : \quad \frac{\mu_i}{\sigma_i} - \frac{\mu_j}{\sigma_j} = 0. \tag{3.25}$$

3.4.4 Empirical results

In this subsection, we compare the criteria described above for all strategies listed in Table 3.1 and discuss their implications. We also provide robustness check for the OWL-ADMM algorithm and we illustrate detailed weight distributions for each asset class using various portfolio strategies in Appendix 3.A.4.

3.4.4.1 Fama-French 25 portfolios

Table 3.2. OOS scores using FF25

	SR	$\hat{\sigma}$	TO	$\hat{\mu}$(annualized)	TC	$TCadjSR$
EW	0.7271	0.0549	0.0172	0.1384	0.0010	0.7216
minVar	0.9785	0.0401	0.7558	0.1358	0.0453	0.6518
minVar-JM	0.8020	0.0425	0.0678	0.1182	0.0041	0.7744
minVar-LW	0.8613	0.0425	0.2952	0.1268	0.0177	0.7410
minVar-OWL	0.7727	0.0484	0.0278	0.1295	0.0017	0.7627
minVar-OWL-Pos	0.7323	0.0544	0.0172	0.1379	0.0010	0.7268
minVar-OWL-bounds	0.7299	0.0549	0.0178	0.1389	0.0011	0.7243
minVar-hard-OWL	0.0141	0.4128	9.8011	0.0202	0.5881	-0.3971
minVar-LW-OWL	0.7739	0.0483	0.0295	0.1295	0.0018	0.7633
MVE-OWL-Pos	0.7344	0.0547	0.0147	0.1391	0.0009	0.7298
MVE-OWL-bounds	0.7311	0.0551	0.0163	0.1395	0.0010	0.7260

Note: this table reports performance scores for various strategies using Fam-French 25 portfolios. The transaction cost is calibrated to be 50 base points for trading 1 US dollar.

Table 3.2 reports the out-of-sample performance scores using various criteria including the Sharpe ratio, standard deviation, turnover, (annualized) mean returns, transaction cost and transaction cost adjusted Sharpe ratios for each trading strategy. We find that the plain minVar strategy achieves highest OOS Sharpe ratio and

[8] We download the code from $https : //www.econ.uzh.ch/en/people/faculty/wolf/publications$ and we compute the two-sides p-values with 1000 (B=1000) bootstrap random draws and block size sets to be 5 (b=5).

lowest standard deviation. This may be because the FF25 portfolios are less prone to idiosyncratic noises and hence less prone to estimation errors in the sample covariance matrix compared to other asset classes. Moreover, the FF25 portfolios have relatively small cross-sectional dimension but have large time-series dimension. This helps to obtain a relatively precise estimate of the covariance matrix, which is crucial in our optimization problems. Meanwhile, the minVar-JM and minVar-LW also do well in achieving high Sharpe ratios. On the other hand, the minVar strategy, although it produces high Sharpe ratio, suffers from high turnovers. The EW strategy produces the smallest turnovers, closely followed by some OWL related strategies. Note that we calibrate the cost of trading stocks worth one US dollar to be 50 basis points, and this can be viewed as a scaling parameter to tilting weights between Sharpe ratio and the transaction cost; a higher value on this parameter will favor strategies with low transaction cost. By looking at the transaction cost adjusted Sharpe ratio, we find that the plain minVar strategy is outperformed by many other competitors. Notably, the mean-variance efficient (MVE) portfolio typically performs poorly, but once regularized by OWL and further imposing (no-short-sale or bounds) constraints, the MVE portfolio achieves sizeable Sharpe ratio and low transaction cost. It is worth stressing that some of those transaction cost adjusted Sharpe ratios are very similar. To see the significance in their performance, we run a bootstrap based test outlined in Section 3.4.3.

Table 3.3. Pairwise Sharpe ratio test using FF25

		1	2	3	4	5	6	7	8	9	10	11
EW	1	N/A										
minVar	2	0.0190	N/A									
minVar-JM	3	0.1089	0.0470	N/A								
minVar-LW	4	0.2537	0.1029	0.5135	N/A							
minVar-OWL	5	0.2298	0.0470	0.5554	0.4306	N/A						
minVar-OWL-Pos	6	0.0150	0.0220	0.1399	0.2957	0.2498	N/A					
minVar-OWL-bounds	7	0.0160	0.0190	0.1239	0.2488	0.2358	0.0140	N/A				
minVar-hard-OWL	8	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	N/A			
minVar-LW-OWL	9	0.2398	0.0490	0.5704	0.4605	0.8641	0.2787	0.2627	0.0010	N/A		
MVE-OWL-Pos	10	0.0010	0.0300	0.1518	0.2567	0.2907	0.2368	0.0050	0.0010	0.3227	N/A	
MVE-OWL-bounds	11	0.0010	0.0230	0.1129	0.2587	0.2687	0.4346	0.2777	0.0010	0.2767	0.0010	N/A

Note: this table reports the p-values of the pairwise Sharpe ratio tests in Ledoit and Wolf (2008) using the Fama-French 25 portfolios. If p-value is great than 5%, then we do not reject the hypothesis that these two strategies yield the same (TC adjusted) Sharpe ratio.

Table 3.3 reports the p-values of the pair-wise comparison of Sharpe ratios between any two strategies using the Fama-French 25 portfolios. First of all, we find that the minVar-OWL strategy is not statistically different from the Equal weighted, minVar-JM, and minVar-LW strategies which exhibit high Sharpe ratios in Table 3.2. Similarly, this insignificance also appears after comparing minVar-JM, minVar-LW, and equal weighted strategies, indicating these strategies are not significantly different in producing Sharpe ratios.

3.4.4.2 SP500 daily returns

Table 3.4. OOS scores using SP500

Panel A: SP500 daily returns with weekly rebalancing

	SR	$\hat{\sigma}$	TO	$\hat{\mu}$(annualized)	TC	$TCadjSR$
EW	1.0046	0.0244	0.0314	0.1771	0.0082	0.9584
minVar	0.5338	0.0540	12.9184	0.2079	3.3588	-8.0889
minVar-JM	1.5831	0.0143	0.0849	0.1629	0.0221	1.3684
minVar-LW	1.3914	0.0130	0.6057	0.1306	0.1575	-0.2866
minVar-OWL	1.0568	0.0227	0.0306	0.1728	0.0080	1.0082
minVar-OWL-Pos	1.0216	0.0238	0.0310	0.1756	0.0080	0.9748
minVar-OWL-bounds	1.0128	0.0240	0.0319	0.1751	0.0083	0.9649
minVar-hard-OWL	1.0656	0.0223	0.0309	0.1717	0.0080	1.0158
minVar-LW-OWL	1.0513	0.0227	0.0305	0.1722	0.0079	1.0030
MVE-OWL-Pos	0.9811	0.0257	0.0561	0.1815	0.0146	0.9023
MVE-OWL-bounds	0.9913	0.0244	0.0266	0.1742	0.0069	0.9520

Panel B: SP500 daily returns with monthly rebalancing

	SR	$\hat{\sigma}$	TO	$\hat{\mu}$(annualized)	TC	$TCadjSR$
EW	1.0399	0.0484	0.0684	0.1745	0.0041	1.0154
minVar	0.6107	0.0831	22.3297	0.1758	1.3398	-4.0433
minVar-JM	1.5013	0.0318	0.2021	0.1654	0.0121	1.3912
minVar-LW	1.3267	0.0288	1.3261	0.1326	0.0796	0.5306
minVar-OWL	1.0850	0.0455	0.0673	0.1711	0.0040	1.0594
minVar-OWL-Pos	1.0552	0.0474	0.0676	0.1732	0.0041	1.0305
minVar-OWL-bounds	1.0472	0.0476	0.0681	0.1726	0.0041	1.0223
minVar-hard-OWL	1.0888	0.0450	0.0684	0.1699	0.0041	1.0624
minVar-LW-OWL	1.0798	0.0456	0.0669	0.1705	0.0040	1.0544
MVE-OWL-Pos	1.0694	0.0494	0.1138	0.1829	0.0068	1.0294
MVE-OWL-bounds	1.0568	0.0473	0.0560	0.1733	0.0034	1.0363

Note: this table reports performance scores for various strategies using the Standard & Poor 500 stocks daily returns with weekly or monthly rebalancing frequency. The transaction cost is calibrated to be 50 base points for trading 1 US dollar.

Table 3.4 reports performance scores for various strategies using the S&P 500 daily returns with weekly or monthly rebalancing frequency. The transaction cost is calibrated to be 50 basis points for trading one US dollar of stocks. First of all, we find when using individual stocks as test assets, particularly if using daily return series, estimation errors in the sample covariance matrix become more evident: by looking into SR and $\hat{\sigma}$, the plain minVar strategy becomes inferior to many competitors, resulting from elevated estimation error in the sample covariance matrix. Then, by looking into turnovers, we find that the equal weighted strategy and some OWL related strategies (particularly for the MVE-OWL-bounds strategy) produce the smallest turnovers. It is worth stressing that we find that the "minVar-OWL" outperforms the equal weighted strategy in both Sharpe ratio and turnovers, with either weekly or monthly rebalancing frequency, which is a remarkable finding as it is difficult to find a strategy that outperforms the equal weighted strategy in both Sharpe ratio and turnovers. Next, we compare Sharpe ratios between strategies pairwisely and test for significance.

Table 3.5. Sharpe ratio test using SP500d

Panel A: SP500 daily returns with weekly rebalancing

		1	2	3	4	5	6	7	8	9	10	11
EW	1	N/A										
minVar	2	0.0330	N/A									
minVar-JM	3	0.0010	0.0010	N/A								
minVar-LW	4	0.0709	0.0010	0.2308	N/A							
minVar-OWL	5	0.0150	0.0250	0.0010	0.1059	N/A						
minVar-OWL-Pos	6	0.0030	0.0320	0.0010	0.0899	0.0160	N/A					
minVar-OWL-bounds	7	0.0100	0.0230	0.0020	0.0729	0.0110	0.0030	N/A				
minVar-hard-OWL	8	0.0110	0.0140	0.0010	0.1079	0.0989	0.0150	0.0140	N/A			
minVar-LW-OWL	9	0.0160	0.0180	0.0030	0.0929	0.0010	0.0340	0.0230	0.0290	N/A		
MVE-OWL-Pos	10	0.6174	0.0500	0.0010	0.0679	0.1678	0.3956	0.5095	0.1049	0.1848	N/A	
MVE-OWL-bounds	11	0.5884	0.0370	0.0010	0.0559	0.0340	0.2118	0.3556	0.0240	0.0440	0.7193	N/A

Panel B: SP500 daily reurns with monthly rebalancing

		1	2	3	4	5	6	7	8	9	10	11
EW	1	N/A										
minVar	2	0.1059	N/A									
minVar-JM	3	0.0080	0.0070	N/A								
minVar-LW	4	0.2757	0.0030	0.2977	N/A							
minVar-OWL	5	0.0899	0.0939	0.0110	0.3387	N/A						
minVar-OWL-Pos	6	0.0829	0.0989	0.0090	0.2977	0.1119	N/A					
minVar-OWL-bounds	7	0.0999	0.0909	0.0090	0.2817	0.1009	0.0669	N/A				
minVar-hard-OWL	8	0.1149	0.0929	0.0160	0.3107	0.4835	0.1439	0.1059	N/A			
minVar-LW-OWL	9	0.1189	0.0839	0.0060	0.3347	0.0020	0.1578	0.1149	0.1159	N/A		
MVE-OWL-Pos	10	0.7033	0.0919	0.0180	0.3427	0.8861	0.8771	0.7612	0.8332	0.8881	N/A	
MVE-OWL-bounds	11	0.6484	0.1089	0.0080	0.3057	0.5604	0.9700	0.7972	0.5235	0.6374	0.7722	N/A

Note: this table reports p-values of pair-wise Sharpe ratio test according to Ledoit and Wolf (2008) using SP500d returns. The rebalancing frequency in Panel A is weekly, and in Panel B is monthly.

Table 3.5 reports the p-values of the Sharpe ratio test outlined in Section 3.4.3 using SP500 stocks with daily returns. The pairwise comparison suggests that the minVar-OWL strategy is statistically outperforming the equal weighted strategy with weekly rebalancing frequency, while insignificant for the monthly rebalancing frequency. Meanwhile, we find that for the SP500d returns, the best performing strategies in terms of Sharpe ratios are minVar-JM and minVar-LW and their superior performance against other strategies is significant suggested by the Sharpe ratio test. Next, we set out to test the SP100 stocks with monthly returns, which exhibit higher correlations between stock returns compared to SP500d stocks.

3.4.4.3 SP100 monthly returns

Table 3.6. OOS scores using SP100m

	SR	$\hat{\sigma}$	TO	$\hat{\mu}$(annualized)	TC	$TCadjSR$
EW	0.9233	0.0460	0.0571	0.1472	0.0034	0.9018
minVar	0.0624	0.0851	4.4201	0.0184	0.2652	-0.8371
minVar-JM	1.0065	0.0345	0.1140	0.1203	0.0068	0.9493
minVar-LW	0.8537	0.0359	0.2974	0.1061	0.0178	0.7102
minVar-OWL	0.9811	0.0420	0.0555	0.1426	0.0033	0.9582
minVar-OWL-Pos	0.9418	0.0444	0.0566	0.1448	0.0034	0.9197
minVar-OWL-bounds	0.9257	0.0461	0.0572	0.1477	0.0034	0.9042
minVar-hard-OWL	0.9862	0.0415	0.0565	0.1419	0.0034	0.9626
minVar-LW-OWL	0.9633	0.0425	0.0546	0.1417	0.0033	0.9411
MVE-OWL-Pos	0.9302	0.0451	0.0581	0.1452	0.0035	0.9079
MVE-OWL-bounds	0.9293	0.0458	0.0516	0.1475	0.0031	0.9098

Note: this table reports performance scores for various strategies using the Standard & Poor 100 stocks with monthly returns and rebalanced monthly. The transaction cost is calibrated to be 50 base points for trading 1 US dollar.

Table 3.6 reports performance scores using Standard & Poor 100 stocks with monthly returns, and we rebalance the portfolio monthly. First of all, we find that minVar-OWL and some other OWL related strategies consistently outperform the equal weighted strategy in terms of the Sharpe ratio and turnovers. The MVE-OWL-bounds strategy yields the smallest turnover while the turnover of the minVar-JM strategy doubles that of OWL related strategies. Second, the raw Sharpe ratio (i.e. not adjusted by transaction cost) of the minVar-JM strategy tops the ranking, and it is closely followed by OWL related strategies. However, after being adjusted by transaction cost, the minVar-OWL and minVar-hard-OWL strategies top the ranking. Third, by comparing Table 3.6 and Table 3.4, we find that the performance of OWL related strategies has improved, and we reckon that is because SP100 stocks with monthly returns exhibit higher correlation between stocks which is a desirable property for the OWL shrinkage method to work well. Next, we apply the Sharpe ratio test in Section 3.4.3 to check the significance between strategies.

Table 3.7 reveals that the performance between the minVar-OWL strategy and the minVar-JM, minVar-LW strategies is not significantly different. However, it has

Table 3.7. Sharpe ratio test using SP100m

		1	2	3	4	5	6	7	8	9	10	11
EW	1	N/A										
minVar	2	0.0020	N/A									
minVar-JM	3	0.5195	0.0010	N/A								
minVar-LW	4	0.7982	0.0010	0.3526	N/A							
minVar-OWL	5	0.0030	0.0010	0.8202	0.6074	N/A						
minVar-OWL-Pos	6	0.0050	0.0020	0.6084	0.7123	0.0010	N/A					
minVar-OWL-bounds	7	0.0040	0.0030	0.5105	0.7822	0.0040	0.0060	N/A				
minVar-hard-OWL	8	0.0040	0.0010	0.8551	0.5844	0.2577	0.0090	0.0060	N/A			
minVar-LW-OWL	9	0.0320	0.0020	0.7233	0.6454	0.0010	0.1019	0.0280	0.0020	N/A		
MVE-OWL-Pos	10	0.8911	0.0030	0.5634	0.7802	0.2997	0.8062	0.9141	0.2687	0.0030	N/A	
MVE-OWL-bounds	11	0.4575	0.0050	0.5205	0.7862	0.0110	0.1998	0.6204	0.0170	0.1099	0.9740	N/A

Note: this table reports the p-values of the Sharpe ratio test according to Ledoit and Wolf (2008) using the SP100 monthly returns.

statistically higher Sharpe ratios than that of the equal weighted strategy.

It is worth stressing that our main target is to draw attention to the comparison between the minVar-OWL strategy and the equal weighted strategy. They receive very similar weight distributions, but we show that the minVar-OWL strategy outperforms the equal weighted strategy in both Sharpe ratio and turnover. In appendix 3.A.4.3, we show (3-dimensional) graphs that illustrate the weight distribution for some strategies and find that the minVar-OWL strategy has a very similar distribution to the equal weighted strategy. This near-equal-weighted weight distribution is caused by the grouping property as discussed in Section 3.3.2. We further find the superior performance in Sharpe ratio against the equal weighted strategy is indeed statistically significant after applying a bootstrap based test outlined in Section 3.4.3. Similar exercises are applied and tested on the CRSP500d stock returns and CRSP100m stock returns. We put those empirical results in Appendix 3.A.4.2.

So far, we have focused our comparison criteria on Sharpe ratios and turnovers (transaction cost). We find that the minVar-JM strategy delivers impressive Sharpe ratios, although in the SP100m and FF25 asset classes its Sharpe ratio is not significantly different from the minVar-OWL strategy after running a Sharpe ratio test. On the other hand, the minVar-OWL strategy and other OWL related strategies consistently yield the smallest turnovers.

In addition, we stress that we developed a flexible algorithm that can incorporate bespoke weight constraints on individual stocks in the optimization problem. However, in our empirical analysis, we applied (blindly) a -5% to 30% bound for all stocks, since we do not hold any further information on individual stocks. Thus, the bound-constrained OWL strategies can potentially do better if more information about individual stocks becomes available.

Although the Sharpe ratio incorporates both the mean portfolio returns and portfolio risk in its formula, it is often dominated by the portfolio risk component when portfolio returns are small. Alternatively, we use the model confidence set (MCS) method to compare strategies: it includes all the best performing strategies in a set where the average portfolio returns are the highest, while using the portfolio risk to control the confidence band of this set.

3.4.5 Model confidence set for comparing transaction cost adjusted returns

Hansen et al. (2011) propose the model confidence set (MCS) to compare loss sequences of candidate models and put the best candidates in a "confidence set". MCS avoids comparing models pairwisely, which often leads to inconclusive decisions. Instead, MCS enables us to compare multiple models while returning a set that includes all (single or multiple) best performing models. In our application, we want to compare out-of-sample portfolio returns. We want to answer this question: which strategies produce the highest returns while taking account of transaction cost, where portfolio risk controls the confidence band for including the best candidates in a set?

To fix ideas, let M^0 denote a set of finite candidate models (i.e. M^0 collects all candidate models) and M be the active model confidence set with size m[9]. Denote by $L_{i,t}$ the loss function of model i at time t.[10] Then,

$$d_{ij,t} = L_{i,t} - L_{j,t}, \qquad \forall i, j \in M^0, \tag{3.26}$$

is the loss difference function between model i and j at time t. Then, we denote

$$\begin{aligned} \mu_{ij} &= \mathrm{E}(d_{ij,t}), \\ \bar{d}_{ij} &= n^{-1} \sum_{t=1}^{n} d_{ij,t}, \end{aligned} \tag{3.27}$$

[9] We set $M = M^0$ at the beginning, then run a series of tests to remove inferior models from the active set M. In the end, what are left in the active set M will be the final model confidence set.

[10] The MCS compares models using a loss function of each model. Since returns are "gains" rather than "losses", we use the negative values of returns to measure each strategy's "loss".

where μ_{ij} is the expected value of the loss difference between model i and j, and $\bar{d}_{ij}$ is the sample analogy of μ_{ij}. Denote

$$\bar{d}_{i\cdot} \equiv m^{-1} \sum_{j \in M} \bar{d}_{ij}, \tag{3.28}$$

where m is the cardinality of set M, and M is the active set which collects models that need to be tested. $\bar{d}_{i\cdot}$ is the average loss difference sequence of model i with all models left in the active set M. Then, the model confidence set is defined as

$$M^* \equiv \{i \in M^0 : \mu_{ij} \leq 0 \quad \forall j \in M^0\}. \tag{3.29}$$

A detailed testing procedure of MCS is included in Appendix 3.A.3.

Table 3.8. MCS test for transaction cost adjusted returns

	FF25	SP500d, w	SP500d, m	SP100m	CRSP500d, w	CRSP500d, m	CRSP100m
EW	0.0020	0.4210	0.4650	0.0040	0.0000	0.8620	0.4170
minVar	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
minVar-JM	0.0000	0.0050	0.0010	0.0040	0.0000	0.0000	0.4070
minVar-hard-OWL	0.0000	0.0130	0.0080	0.0040	0.0000	0.0070	1.0000
minVar-LW	0.0000	0.0000	0.0000	0.0040	0.8270	1.0000	0.4070
minVar-LW-OWL	0.0000	1.0000	1.0000	0.8930	0.0000	0.0040	0.3430
minVar-OWL	0.0000	0.0130	0.0080	0.0040	0.0000	0.0000	0.4070
minVar-OWL-bounds	0.0000	0.0050	0.0010	0.0040	0.0000	0.0000	0.4170
minVar-OWL-Pos	0.0000	0.4210	0.4650	0.0040	1.0000	0.8860	0.4170
MVE-OWL-bounds	1.0000	0.0000	0.5520	1.0000	0.0000	0.0000	0.0000
MVE-OWL-Pos	0.0840	0.0050	0.0000	0.0040	0.0000	0.0000	0.4070

Note: this table reports the p-values of the MCS test. It compares transaction cost adjusted returns using various strategies and using different asset classes. We consider both weekly ('w') and monthly ('m') rebalancing frequencies for daily returns. If p-value is greater than 5%, then the corresponding strategy will be included in the MCS.

Table 3.8 reports the p-values of the MCS test. It compares transaction cost adjusted returns of various strategies within each asset class. We consider both weekly and monthly rebalancing frequencies for daily returns. If p-value is greater than 5%, then the corresponding strategy will be included in the MCS.

First of all, we notice that the equal weighted strategy has been included in the MCS four times, confirming that the naive 1/N strategy performs well in terms of producing sizeable returns. On the other hand, we find that the minVar-OWL-Pos strategy (OWL regularized minimum variance portfolio with no-short-sale constraint) has been included in the MCS fives times, which makes it the only strategy that has been included in the MCS more often than the equal weighted strategy.

We also notice that the MVE-OWL-bound and MVE-OWL-Pos strategies performs particularly well with the Fama-French portfolios. The OWL shrinkage method with further constraints on portfolio weights helps to utilize the optimization gains from the mean variance efficient portfolio. We reckon this is because sorted portfolio returns are less prone to idiosyncratic noises and thus the sample estimate of expected asset return and asset covariances are less biased compared to individual stock returns. Also, for the FF25 asset class, large T and small N (i.e. large in time series dimension and small in cross sectional dimension compared to other asset classes) help to improve the precision of the sample estimate of the covariance matrix.

Meanwhile, the minVar-JM strategy, which performs well when using Sharpe ratio as comparison criterion, performs poorly if we use MCS to compare portfolio returns: the minVar-JM strategy has been included in MCS only once, indicating that the minVar-JM strategy produces significantly lower returns than other strategies using various test assets. We find that the MCS for CRSP100m asset class (which consists of 100 randomly selected (usually small) stocks from the CRSP dataset) includes many (9 out of 11) candidate strategies, which is caused by large variations in out-of-sample returns for each strategy using this asset class.

3.5 Conclusion

In this paper, we introduce the OWL shrinkage method for efficient portfolio construction problems. The OWL shrinkage method encompasses the LASSO shrinkage setup and exploits contemporaneous correlations between stocks, thereby extending the LASSO shrinkage method in DeMiguel et al. (2009a) and the VAR(1) model in DeMiguel et al. (2014). We develop an efficient algorithm that incorporates the OWL shrinkage method together with bespoke constraints on individual stocks if prior information is available. We apply our OWL portfolio strategies on five asset classes and find that the OWL shrinkage method outperforms other benchmarks when stocks exhibit high correlations. DeMiguel et al. (2009b) compare the naive 1/N portfolio strategy with the other 14 optimization-based strategies, finding superb out-of-sample performance in the naive 1/N portfolio. In this paper, we bridge the gap between the naive 1/N portfolio strategy and an optimization based method: our OWL optimization problem yields similar portfolio weights to the 1/N portfolio strategy due to the grouping property, yet our OWL based portfolio strategies outperform the 1/N strategy in terms of Sharpe ratios and turnovers. A bootstrap based Sharpe ratio test by Ledoit and Wolf (2008) also confirms that this difference in Sharpe ratio is significant.

3.A Appendix

3.A.1 ADMM algorithm to solve the constrained OWL optimization problem

Boyd et al. (2010) proposed a general optimization algorithm which utilizes the augmented Lagrangian function and can decompose a complex optimization problem into two parts which share different characteristics in computational complexity. This algorithm optimizes these two parts separately and in an orderly fashion, hence gains the name of "alternating directions".

3.A.1.1 Augmented Lagrangian

First, define the augmented Lagrangian of the optimization problem (3.14) - (3.18) as

$$\begin{aligned}\ell_\rho(w, v, \alpha, \beta, \theta, \xi) = \ & \frac{\gamma}{2} w'\hat{\Sigma} w - \hat{\mu}' w + \Omega_\omega(v) + \alpha'(w - v) + \beta(w'e - 1) \\ & + \theta'(lb - w) + \xi'(w - ub) + \frac{\rho}{2}(||w - v||_2^2 + (w'e - 1)^2 \\ & + ||lb - w||_2^2 + ||w - ub||_2^2),\end{aligned} \tag{3.A.1}$$

where α, β, θ and ξ are Lagrangian multipliers, ρ is a parameter to control penalty and e is a column vector of ones. The ADMM algorithm consists of these updates for each step:

$$w^{k+1} = \arg\min_{w} \ \ell_\rho(w, v^k, \alpha^k, \beta^k, \theta^k, \xi^k), \tag{3.A.2}$$

$$w_i^{k+1} = lb_i; \quad if \quad w_i^{k+1} < lb_i \quad \forall \quad i = 1, 2, ..., N, \tag{3.A.3}$$

$$w_i^{k+1} = ub_i; \quad if \quad w_i^{k+1} > ub_i \quad \forall \quad i = 1, 2, ..., N, \tag{3.A.4}$$

$$v^{k+1} = \arg\min_{v} \ \ell_\rho(w^{k+1}, v, \alpha^k, \beta^k, \theta^k, \xi^k), \tag{3.A.5}$$

$$\alpha^{k+1} = \alpha^k + \rho(w^{k+1} - v^{k+1}), \tag{3.A.6}$$

$$\beta^{k+1} = \beta^k + \rho(e'w^{k+1} - 1), \tag{3.A.7}$$

$$\theta^{k+1} = \theta^k + \rho(lb - w^{k+1}), \tag{3.A.8}$$

$$\theta_i^{k+1} = 0; \quad if \quad lb_i - w_i^{k+1} < 0 \quad \forall \quad i = 1, 2, ..., N, \tag{3.A.9}$$

$$\xi^{k+1} = \xi^k + \rho(w^{k+1} - ub), \tag{3.A.10}$$

$$\xi_i^{k+1} = 0; \quad if \quad w_i^{k+1} - ub_i < 0 \quad \forall \quad i = 1, 2, ..., N, \tag{3.A.11}$$

where k is a superscript indicating the step number. We refer to equations (3.A.3) and (3.A.4) as *primal feasibility* conditions, and equations (3.A.9) and (3.A.11) as *complementary slackness* conditions. Moreover, (3.A.2) can be simplified as

$$\begin{aligned}
w^{k+1} &= \arg\min_w \quad \ell_\rho(w, v^k, \alpha^k, \beta^k, \theta^k, \xi^k) \\
&= \arg\min_w \quad \Big[\frac{\gamma}{2} w'\hat{\Sigma}w - \hat{\mu}'w + {\alpha^k}'(w - v^k) + \beta^k(e'w - 1) + {\theta^k}'(lb - w) + {\xi^k}'(w - ub) \\
&\qquad + \frac{\rho}{2}(||w - v^k||_2^2 + (w'e - 1)^2 + ||lb - w||_2^2 + ||w - ub||_2^2)\Big] \\
&= \arg\min_w \quad \Big[\frac{\gamma}{2} w'\hat{\Sigma}w - (\hat{\mu} - \alpha^k - \beta^k e + \theta^k - \xi^k)'w \\
&\qquad + \frac{\rho}{2}(||w - v^k||_2^2 + (w'e - 1)^2 + ||lb - w||_2^2 + ||w - ub||_2^2))\Big] \\
&= \arg\min_w \quad \Big[\frac{1}{2} w'(\gamma\hat{\Sigma} + \rho(3I + ee'))w - (\hat{\mu} - \alpha^k - \beta^k e + \theta^k - \xi^k \\
&\qquad + \rho(v^k + e - lb - ub))'w\Big] \\
&= \ (\gamma\hat{\Sigma} + \rho(3I + ee'))^{-1}(\hat{\mu} - \alpha^k - \beta^k e + \theta^k - \xi^k + \rho(v^k + e - lb - ub)).
\end{aligned}$$

Meanwhile, equation (3.A.5) can be simplified as

$$\begin{aligned}
v^{k+1} &= \arg\min_v \quad \Big[\frac{\rho}{2}||v - w^{k+1} - \frac{1}{\rho}\alpha^k||_2^2 + \Omega_\omega(v)\Big] \\
&= \mathrm{prox}_\Omega(w^{k+1} + \frac{1}{\rho}\alpha^k),
\end{aligned}$$

where $\mathrm{prox}_\Omega(.)$ is a proximal function for the OWL shrinkage method. Discussion of how to find a minimizer of the proximal function $\mathrm{prox}_\Omega(.)$ can be found in Appendix 1.A.2.

3.A.1.2 Optimality conditions

Suppose w^* and v^* are optimizers of the optimization problem (3.14) - (3.18). Then, the optimality conditions of (3.A.1) consist of the *primal feasibility* and the *dual* *feasibility* conditions. The *primal feasibility* concerns the following conditions

$$w^* - v^* = \mathbf{0}, \tag{3.A.12}$$

$$w^{*\prime} e - 1 = 0, \tag{3.A.13}$$

$$w^* \succeq \mathbf{lb}, \tag{3.A.14}$$

$$w^* \preceq \mathbf{ub}. \tag{3.A.15}$$

Equations (3.A.2) and (3.A.5) command the *dual feasibility* condition, which requires

$$\bigtriangledown f(w^*) + \alpha^* + \beta^* e - \theta^* + \xi^* = 0, \tag{3.A.16}$$

$$\bigtriangledown \Omega(v^*) - \alpha^* = 0, \tag{3.A.17}$$

where $f(w) = \frac{\gamma}{2} w' \hat{\Sigma} w - \hat{\mu}' w$. By equation (3.A.5), v^{k+1} minimizes the function $\ell_\rho(w^{k+1}, v, \alpha^k, \beta^k, \theta^k, \xi^k)$ w.r.t v, so we have

$$0 = \bigtriangledown \ell_\rho(v^{k+1}) = \bigtriangledown \Omega(v^{k+1}) - \alpha^k - \rho(w^{k+1} - v^{k+1}) = \bigtriangledown \Omega(v^{k+1}) - \alpha^{k+1},$$

which makes (3.A.17) hold automatically. Similarly, by (3.A.2), w^{k+1} minimizes the function $\ell_\rho(w, v^k, \alpha^k, \beta^k, \theta^k, \xi^k)$ w.r.t w, so we obtain

$$\begin{aligned} 0 = & \bigtriangledown f(w^{k+1}) + \alpha^k + \beta^k e - \theta^k + \xi^k + \rho(w^{k+1} - v^k) \\ & + \rho(w^{k+1\prime} e - 1)e - \rho(lb - w^{k+1}) + \rho(w^{k+1} - ub) \\ = & \bigtriangledown f(w^{k+1}) + \alpha^{k+1} + \beta^{k+1} e - \theta^{k+1} + \xi^{k+1} + \rho(v^{k+1} - v^k). \end{aligned}$$

Rearranging the above equation gives

$$\bigtriangledown f(w^{k+1}) + \alpha^{k+1} + \beta^{k+1} e - \theta^{k+1} + \xi^{k+1} = -\rho(v^{k+1} - v^k) := s^{k+1},$$

where we denote $s^{k+1} := -\rho(v^{k+1} - v^k)$ as the *dual residual* at step $k+1$, because s^{k+1} is the deviation from a dual feasibility condition in (3.A.16). Similarly, the *primal residual* at step k w.r.t the primal feasibility conditions in (3.A.12) and (3.A.13) is defined as

$$||r^k||_2 = \sqrt{||w^k - v^k||^2 + (w^{k\prime} e - 1)^2}.$$

3.A.1.3 Stopping criterion and the penalty parameter ρ

The stopping criterion for k suggested by Boyd et al. (2010) is such that k satisfies

$$\begin{aligned} ||r^k||_2 \leq \epsilon^{pri} \quad \text{and} \quad ||s^k||_2 \leq \epsilon^{dual}, \\ \epsilon^{pri} = \sqrt{N}\epsilon^{abs} + \epsilon^{rel} \max\{||w^k||_2, ||v^k||_2\}, \\ \epsilon^{dual} = \sqrt{T}\epsilon^{abs} + \epsilon^{rel}||\alpha^k + \beta^k e||_2, \end{aligned}$$

where ϵ^{rel} and ϵ^{abs} are calibrated to be 0.001.

Boyd et al. (2010) also argues that allowing ρ to change along steps makes computation more efficient and suggests the following scheme for the values of ρ:

$$\rho^{k+1} = \begin{cases} \tau\rho^k & \text{if } ||r^k||_2 > \eta||s^k||_2 \ , \\ \rho^k/\tau & \text{if } ||s^k||_2 > \eta||r^k||_2 \ , \\ \rho^k & \text{otherwise} \ , \end{cases}$$

where $\eta, \tau > 1$ are two tuning parameters, which are calibrated such that $\eta = 10$, $\tau = 2$ in our exercise.

3.A.2 Technical proofs

3.A.2.1 Proof of Theorem 3.3.1

Proof. The proof of Theorem 3.3.1 relies on the Pigou-Dalton transfer principle and the directional derivative lemma at the minimum of a convex function. It follows using a similar argument as in Figueiredo and Nowak (2016), except that we are dealing with different loss functions.

Lemma 3.A.1 (Pigou-Dalton transfer principle). *Let be given vector $x \in R_+^p$, and its two components x_i, x_j are such that $x_i > x_j$. Let $\epsilon \in (0, (x_i - x_j)/2)$, $z_i = x_i - \epsilon$, $z_j = x_j + \epsilon$, and $z_k = x_k$, for $k \neq i, j$. Set $\Omega_\omega(x) = \omega' x$, where $\omega \in R_+^p$, and $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_p$. Then it holds*

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

Lemma 3.A.2 (Directional derivative). *The directional derivative of function f :*

$R^K \to R$ at $x \in dom(f)$, in the direction $\xi \in R^K$ is given by

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

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

Denote the objective function as $Q(w) = \frac{1}{2}w'\Sigma w + \Omega_\omega(w)$. By definition, if $\hat{w}$ is the minimizer, then $Q(\hat{w}) \leq Q(w)$ for all w. Thus by Lemma 3.A.2, for any $\xi \in R^N$,

$$Q'(\hat{w},\xi) \geq 0. \tag{3.A.18}$$

Recall that $\Sigma_{i.}$ and $\Sigma_{j.}$ denote the i^{th} and j^{th} columns of the $N \times N$ variance-covariance matrix Σ. Suppose

$$\|\Sigma_{i.} - \Sigma_{j.}\|_2 < \lambda_2, \tag{3.A.19}$$

and assume $\hat{w}_i \neq \hat{w}_j$. We will show contradiction between assumption $\hat{w}_i \neq \hat{w}_j$ and (3.A.18). Without loss of generality, assume $\hat{w}_i > \hat{w}_j$, $i < j$. First we define a special directional vector $\xi = (\xi_1, \xi_2, \cdots, \xi_N)'$. Set $\xi_i = 1, \xi_j = -1$ and $\xi_k = 0$ for all $k \neq i, j$. The directional derivative of Q at $\hat{w}$ with such ξ is

$$Q'(\hat{w},\xi) = \lim_{\alpha \to 0^+} \left(QL_\alpha(\hat{w},\xi) + RP_\alpha(\hat{w},\xi)\right), \tag{3.A.20}$$

where

$$\begin{aligned} QL_\alpha(\hat{w},\xi) &= \frac{(\hat{w}+\alpha\xi)'\Sigma(\hat{w}+\alpha\xi) - \hat{w}\Sigma\hat{w}}{2\alpha} \\ &= \frac{\alpha\hat{w}'\Sigma\xi + \alpha\xi'\Sigma\hat{w} + \alpha^2\xi'\Sigma\xi}{2\alpha}, \end{aligned} \tag{3.A.21}$$

and

$$RP_\alpha(\hat{w},\xi) = \frac{\Omega_\omega(\hat{w}+\alpha\xi) - \Omega_\omega(\hat{w})}{\alpha}. \tag{3.A.22}$$

Note that $\Sigma' = \Sigma$ and $\hat{w}'\Sigma\xi$ is a scaler, so we have $\hat{w}'\Sigma\xi = \xi'\Sigma\hat{w}$. Then it follows

$$\lim_{\alpha \to 0^+} QL_\alpha(\hat{w},\xi) = \hat{w}'\Sigma\xi = \text{trace}(\hat{w}'\Sigma\xi) = \text{trace}(\xi\hat{w}'\Sigma). \tag{3.A.23}$$

Observe that $\xi\hat{w}'$ is a $N \times N$ matrix with i^{th} row as $\hat{w}'$, j^{th} row as $-\hat{w}'$ and the remaining rows are filled with zeros. Then we have

$$\lim_{\alpha \to 0^+} QL_\alpha(\hat{w}, \xi) = \text{trace}(\xi \hat{w}' \Sigma) = \hat{w}'(\Sigma_{i.} - \Sigma_{j.}), \tag{3.A.24}$$

where $\Sigma_{i.}$ and $\Sigma_{j.}$ are the i^{th} and j^{th} columns of Σ.

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

$$-RP_\alpha(\hat{w}, \xi)\alpha = \Omega_\omega(\hat{w}) - \Omega_\omega(\hat{w} + \alpha\xi) \geq \Delta_\omega \alpha. \tag{3.A.25}$$

So for any α and ξ,

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

By the definition of ω in (3.11), $\Delta_\omega = \lambda_2$. Therefore, applying the above bound in (3.A.20), we obtain

$$\begin{aligned} Q'(\hat{w}, \xi) &\leq \hat{w}'(\Sigma_{i.} - \Sigma_{j.}) - \Delta_\omega \\ &= \hat{w}'(\Sigma_{i.} - \Sigma_{j.}) - \lambda_2. \end{aligned} \tag{3.A.26}$$

Using Cauchy-Schwarz inequality, we have

$$\hat{w}'(\Sigma_{i.} - \Sigma_{j.}) \leq \|\hat{w}\|_2 \|\Sigma_{i.} - \Sigma_{j.}\|_2 \leq \|\hat{w}\|_1 \|\Sigma_{i.} - \Sigma_{j.}\|_2 = \|\Sigma_{i.} - \Sigma_{j.}\|_2,$$

so (3.A.26) becomes

$$Q'(\hat{w}, \xi) \leq \|\Sigma_{i.} - \Sigma_{j.}\|_2 - \lambda_2. \tag{3.A.27}$$

Then, (3.A.27) together with (3.A.19) implies

$$Q'(\hat{w}, \xi) < 0,$$

which violates (3.A.18). Hence, there is a contradiction between $\hat{w}_i \neq \hat{w}_j$ and (3.A.19). So we must have

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

which completes the proof. □

3.A.2.2 Proof of Theorem 3.3.2

Proof. The proof of Theorem 3.3.2 is similar to Theorem 3.3.1, except we have a different objective function, that is

$$Q(w) = \frac{\gamma}{2} w'\Sigma w - \mu' w + \Omega_\omega(w),$$

where γ is investors' risk aversion level and μ is the vector of expected returns. Following similar procedures as in the proof of Theorem 3.3.1 and by Lemma 3.A.2, we have that for any $\xi \in R^N$,

$$Q'(\hat{w}, \xi) \geq 0. \tag{3.A.28}$$

Suppose

$$\gamma \|\Sigma_{i.} - \Sigma_{j.}\|_2 + |\mu_i - \mu_j| < \lambda_2 \tag{3.A.29}$$

and assume $\hat{w}_i \neq \hat{w}_j$. Without loss of generality, assume $\hat{w}_i > \hat{w}_j$, $i < j$. Define a special directional vector $\xi = (\xi_1, \xi_2, \cdots, \xi_N)'$. Set $\xi_i = 1, \xi_j = -1$ and $\xi_k = 0$ for all $k \neq i, j$. The directional derivative of Q at $\hat{w}$ with such ξ is

$$Q'(\hat{w}, \xi) = \lim_{\alpha \to 0^+} \left(QL_\alpha(\hat{w}, \xi) + RP_\alpha(\hat{w}, \xi) \right), \tag{3.A.30}$$

where

$$\begin{aligned} QL_\alpha(\hat{w}, \xi) &= \frac{\frac{\gamma}{2}[(\hat{w} + \alpha\xi)'\Sigma(\hat{w} + \alpha\xi) - \hat{w}\Sigma\hat{w}] - \mu'(\hat{w} + \alpha\xi) + \mu'\hat{w}}{\alpha} \\ &= \frac{\frac{\gamma}{2}(\alpha\hat{w}'\Sigma\xi + \alpha\xi'\Sigma\hat{w} + \alpha^2\xi'\Sigma\xi) - \alpha\mu'\xi}{\alpha}, \end{aligned} \tag{3.A.31}$$

and

$$RP_\alpha(\hat{w}, \xi) = \frac{\Omega_\omega(\hat{w} + \alpha\xi) - \Omega_\omega(\hat{w})}{\alpha}. \tag{3.A.32}$$

Note that $\Sigma' = \Sigma$ and $\hat{w}'\Sigma\xi$ is a scalar, so we have $\hat{w}'\Sigma\xi = \xi'\Sigma\hat{w}$. Then it follows

$$\lim_{\alpha \to 0^+} QL_\alpha(\hat{w}, \xi) = \gamma\hat{w}'\Sigma\xi - \mu'\xi, \tag{3.A.33}$$

where $\gamma\hat{w}'\Sigma\xi = \text{trace}(\gamma\hat{w}'\Sigma\xi) = \gamma\ \text{trace}(\xi\hat{w}'\Sigma)$. Observe that $\xi\hat{w}'$ is a $N \times N$ matrix with i^{th} row as $\hat{w}$, j^{th} row as $-\hat{w}'$ and the remaining rows are filled with zeros. Then we have

$$\lim_{\alpha \to 0^+} QL_\alpha(\hat{w}, \xi) = \gamma\ \text{trace}(\xi\hat{w}'\Sigma) - \mu'\xi = \gamma\hat{w}'(\Sigma_{i.} - \Sigma_{j.}) - \mu'\xi, \tag{3.A.34}$$

where $\Sigma_{i.}$ and $\Sigma_{j.}$ are the i^{th} and j^{th} columns of Σ.

Similarly to the procedures we used to handle $RP_\alpha(\hat{w}, \xi)$ in the proof of Theorem 3.3.1, we can obtain

$$RP_\alpha(\hat{w}, \xi) \leq -\lambda_2.$$

Therefore,

$$Q'(\hat{w}, \xi) \leq \gamma \hat{w}'(\Sigma_{i.} - \Sigma_{j.}) - \mu'\xi - \lambda_2. \tag{3.A.35}$$

Using Cauchy-Schwarz inequality, we have

$$\hat{w}'(\Sigma_{i.} - \Sigma_{j.}) \leq \|\hat{w}\|_2 \|\Sigma_{i.} - \Sigma_{j.}\|_2 \leq \|\hat{w}\|_1 \|\Sigma_{i.} - \Sigma_{j.}\|_2 = \|\Sigma_{i.} - \Sigma_{j.}\|_2,$$

Observe that $\mu'\xi = \mu_i - \mu_j \geq -|\mu_i - \mu_j|$, so (3.A.35) becomes

$$Q'(\hat{w}, \xi) \leq \gamma \|\Sigma_{i.} - \Sigma_{j.}\|_2 + |\mu_i - \mu_j| - \lambda_2. \tag{3.A.36}$$

Then (3.A.36) together with (3.A.29) implies

$$Q'(\hat{w}, \xi) < 0,$$

which violates (3.A.28). Hence, there is a contradiction between $\hat{w}_i \neq \hat{w}_j$ and (3.A.29). So we must have

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

which completes the proof. □

3.A.2.3 Proof of Lemma 3.3.1

Proof. If $\lambda_2 = 0$, then $\omega = (\lambda_1, \lambda_1, ..., \lambda_1)' \in R^N$. So we have

$$\Omega_\omega(w) = \omega'|w|_\downarrow = \lambda_1 e'|w|_\downarrow = \lambda_1 \|w\|_1,$$

where e is a column vector of ones. If we set $\lambda_1 = \lambda$, then $\Omega_\omega(w) = \lambda \|w\|_1$ which completes the proof. □

3.A.2.4 Proof of Lemma 3.3.2

Proof. Note that $|w|_\downarrow = (|w|_{[1]}, |w|_{[2]}, \cdots, |w|_{[N]})'$ reorders the elements in vector $|w| = (|w_1|, |w_2|, \cdots, |w_N|)'$ decreasingly according to the absolute value of each element.

Denote by $|w_j|$ and $|w|_{[j]}$ the j^{th} element of $|w|$ and $|w|_\downarrow$, respectively. So we have $|w|_{[1]} > |w|_{[2]} > \cdots > |w|_{[N]}$. Then by the definition of the OSCAR penalty term, we obtain

$$\begin{aligned}\Omega_{OSCAR}(w) &= \lambda_1 ||w||_1 + \lambda_2 \sum_{1 \le i < j \le N} \max\{|w_i|, |w_j|\} \\ &= \sum_{i=1}^{N} [\lambda_1 + \lambda_2(N-i)]\ |w|_{[i]} \\ &= \omega' |w|_\downarrow = \Omega_\omega(w),\end{aligned}$$

which completes the proof. □

3.A.3 MCS testing procedure

The MCS (model confidence set, Hansen et al. (2011)) testing procedure consists of the following steps:

1. Initialize the active model confidence set $M \leftarrow M^0$, where M^0 contains all candidate models.

2. Compute the t-statistics for any pairwise loss difference sequences and the average t-statistics for each model:

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\widehat{\text{Var}}(\bar{d}_{ij})}} \quad \text{and} \quad t_{i\cdot} = \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{\text{Var}}(\bar{d}_{i\cdot})}}, \quad \text{for all } i, j \in M. \tag{3.A.37}$$

3. Find the model with the largest t-statistic

$$T_{max,M} = \max_{i \in M} \quad t_{i\cdot} \tag{3.A.38}$$

   and test whether $T_{max,M}$ is significantly different from zero.

4. If $T_{max,M}$ is statistically different (greater) than zero, eliminate this model from the model confidence set, and go back to step 1 while removing this model from set M. Repeat this procedure until $T_{max,M}$ is not significantly different from zero. What remains in M will be the model confidence set.

3.A.4 Robustness check

3.A.4.1 Convergence of OWL-ADMM algorithm

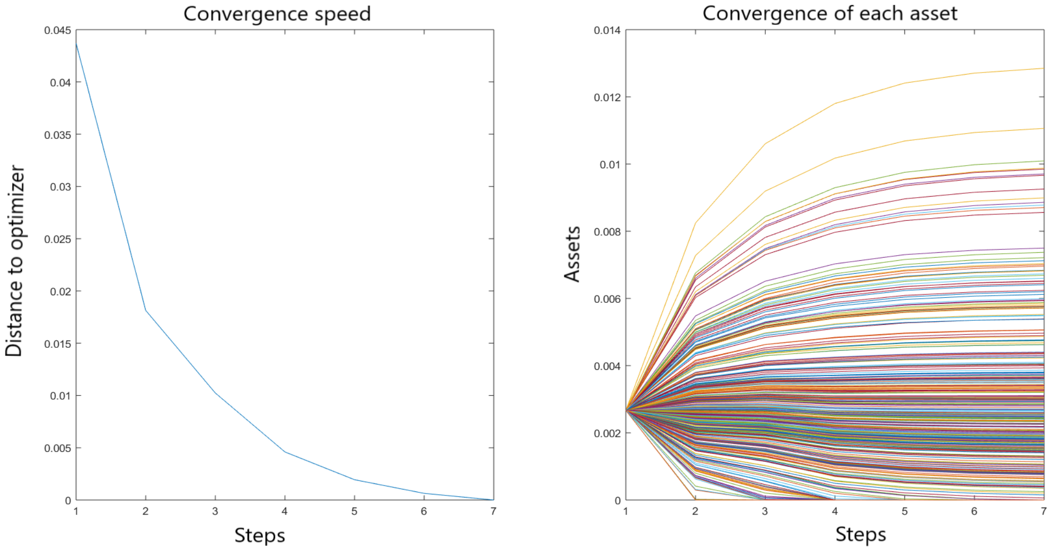


Figure 3.3. Convergence check for ADMM algorithm using SP500 stocks

Figure 3.3 shows the convergence diagram for the OWL-ADMM algorithm used to solve the optimization problem in (3.14) - (3.18) using SP500 daily returns. Left panel shows the distance (i.e. the ℓ_2 norm of two vectors) of the estimated portfolio weights at each step to the final optimizer. Compared to other algorithms, such as gradient descent, ADMM offers a much faster convergence speed. The right panel shows the individual stock's weights at each step until convergence. Each colored line represents one stock, and note that we initialize the portfolio weights as equal weighted at the beginning. We find that the ADMM algorithm is fast to find the optimizer, typically requiring less than 10 steps.

3.A.4.2 Empirical application using randomly selected stocks from CRSP dataset

Panels A and B in Table 3.9 report the OOS performance scores using 500 randomly selected stocks with daily returns from the CRSP dataset and Panels A and B in Table 3.10 report the p-values of the Sharpe ratio test by comparing portfolio strategies pairwisely. We find that the OWL related strategies consistently offer smaller turnovers (transaction costs) than minVar-JM and minVar-LW strategies, and the minVar-OWL

Table 3.9. OOS score using CRSP500d and CRSP100m

Panel A: CRSP500 daily returns with weekly rebalancing

	SR	$\hat{\sigma}$	TO	$\hat{\mu}$(annualized)	TC	$TCadjSR$
EW	1.0491	0.0244	0.0000	0.1844	0.0000	1.0491
minVar	1.4801	0.0125	0.6168	0.1337	0.1604	-0.2954
minVar-JM	1.8328	0.0114	0.0699	0.1512	0.0182	1.6124
minVar-LW	1.5108	0.0108	0.2054	0.1182	0.0534	0.8280
minVar-OWL	1.1281	0.0229	0.0027	0.1860	0.0007	1.1239
minVar-OWL-Pos	1.0687	0.0238	0.0018	0.1838	0.0005	1.0661
minVar-OWL-bounds	1.0588	0.0239	0.0107	0.1822	0.0028	1.0425
minVar-hard-OWL	1.1203	0.0228	0.0026	0.1843	0.0007	1.1161
minVar-LW-OWL	1.1049	0.0230	0.0023	0.1830	0.0006	1.1013
MVE-OWL-Pos	0.9150	0.0256	0.0644	0.1690	0.0167	0.8244
MVE-OWL-bounds	0.9797	0.0241	0.0280	0.1706	0.0073	0.9379

Panel B: CRSP500 daily return with monthly rebalancing

	SR	$\hat{\sigma}$	TO	$\hat{\mu}$(annualized)	TC	$TCadjSR$
EW	0.8525	0.0547	0.0000	0.1616	0.0000	0.8525
minVar	1.3888	0.0295	1.2631	0.1417	0.0758	0.6463
minVar-JM	1.4227	0.0302	0.1576	0.1488	0.0095	1.3322
minVar-LW	1.4136	0.0270	0.4481	0.1323	0.0269	1.1262
minVar-OWL	0.8887	0.0528	0.0071	0.1627	0.0004	0.8864
minVar-OWL-Pos	0.8664	0.0538	0.0026	0.1615	0.0002	0.8655
minVar-OWL-bounds	0.8591	0.0537	0.0108	0.1599	0.0006	0.8556
minVar-hard-OWL	0.8857	0.0527	0.0070	0.1616	0.0004	0.8834
minVar-LW-OWL	0.8778	0.0528	0.0061	0.1605	0.0004	0.8757
MVE-OWL-Pos	0.7724	0.0586	0.1240	0.1569	0.0074	0.7358
MVE-OWL-bounds	0.8136	0.0549	0.0529	0.1548	0.0032	0.7969

Panel C: CRSP100 monthly return with monthly rebalancing

	SR	$\hat{\sigma}$	TO	$\hat{\mu}$(annualized)	TC	$TCadjSR$
EW	0.8533	0.0505	0.0794	0.1492	0.0048	0.8261
minVar	0.4439	0.0640	1.6118	0.0985	0.0967	0.0079
minVar-JM	1.2616	0.0334	0.1102	0.1458	0.0066	1.2044
minVar-LW	1.0283	0.0346	0.1623	0.1232	0.0097	0.9470
minVar-OWL	0.8954	0.0435	0.0801	0.1348	0.0048	0.8634
minVar-OWL-Pos	0.8819	0.0480	0.0769	0.1467	0.0046	0.8542
minVar-OWL-bounds	0.8698	0.0490	0.0778	0.1476	0.0047	0.8423
minVar-hard-OWL	0.9094	0.0429	0.0861	0.1350	0.0052	0.8746
minVar-LW-OWL	0.8808	0.0444	0.0793	0.1354	0.0048	0.8499
MVE-OWL-Pos	0.7168	0.0542	0.0814	0.1346	0.0049	0.6908
MVE-OWL-bounds	0.7720	0.0513	0.0661	0.1371	0.0040	0.7497

Note: this table reports performance scores for various strategies using randomly selected 500 stocks (Panel A and B, daily returns) and 100 stocks (Panel C, monthly returns) from CRSP and rebalanced weekly or monthly. The transaction cost is calibrated to be 50 base points for trading 1 US dollar.

strategy consistently and significantly outperforms the equal weighted strategy in terms of Sharpe ratio. Panel C in Table 3.9 and 3.10 report the OOS scores and p-values of Sharpe ratio tests using 100 randomly selected stocks with monthly returns from the CRSP data-set. These results confirm the previous findings that OWL

Table 3.10. Sharpe ratio test using CRSP500d and CRSP100m

Panel A: CRSP 500 daily returns, weekly rebalancing

		1	2	3	4	5	6	7	8	9	10	11
EW	1	N/A										
minVar	2	0.0240	N/A									
minVar-JM	3	0.0010	0.0370	N/A								
minVar-LW	4	0.0400	0.8192	0.0559	N/A							
minVar-OWL	5	0.0150	0.0699	0.0010	0.0859	N/A						
minVar-OWL-Pos	6	0.0010	0.0310	0.0010	0.0519	0.0300	N/A					
minVar-OWL-bounds	7	0.0010	0.0280	0.0010	0.0400	0.0160	0.0010	N/A				
minVar-hard-OWL	8	0.0240	0.0450	0.0010	0.0959	0.0010	0.0619	0.0230	N/A			
minVar-LW-OWL	9	0.0509	0.0559	0.0010	0.0719	0.0010	0.1868	0.1049	0.0010	N/A		
MVE-OWL-Pos	10	0.0020	0.0070	0.0010	0.0120	0.0010	0.0010	0.0030	0.0020	0.0020	N/A	
MVE-OWL-bounds	11	0.0030	0.0080	0.0010	0.0250	0.0010	0.0010	0.0010	0.0010	0.0010	0.0090	N/A

Panel B: CRSP 500 daily returns, monthly rebalancing

		1	2	3	4	5	6	7	8	9	10	11
EW	1	N/A										
minVar	2	0.0160	N/A									
minVar-JM	3	0.0140	0.7782	N/A								
minVar-LW	4	0.0549	0.8352	0.9590	N/A							
minVar-OWL	5	0.2488	0.0320	0.0120	0.0819	N/A						
minVar-OWL-Pos	6	0.0100	0.0210	0.0120	0.0679	0.4266	N/A					
minVar-OWL-bounds	7	0.0210	0.0180	0.0090	0.0679	0.2957	0.0040	N/A				
minVar-hard-OWL	8	0.2488	0.0310	0.0070	0.0679	0.0330	0.4745	0.3467	N/A			
minVar-LW-OWL	9	0.3906	0.0220	0.0140	0.0689	0.0030	0.6623	0.4745	0.0050	N/A		
MVE-OWL-Pos	10	0.1259	0.0110	0.0050	0.0390	0.0779	0.0929	0.0959	0.1069	0.1099	N/A	
MVE-OWL-bounds	11	0.1309	0.0210	0.0040	0.0629	0.0849	0.0719	0.0999	0.1009	0.1259	0.1459	N/A

Panel C: CRSP 100 monthly returns, monthly rebalancing

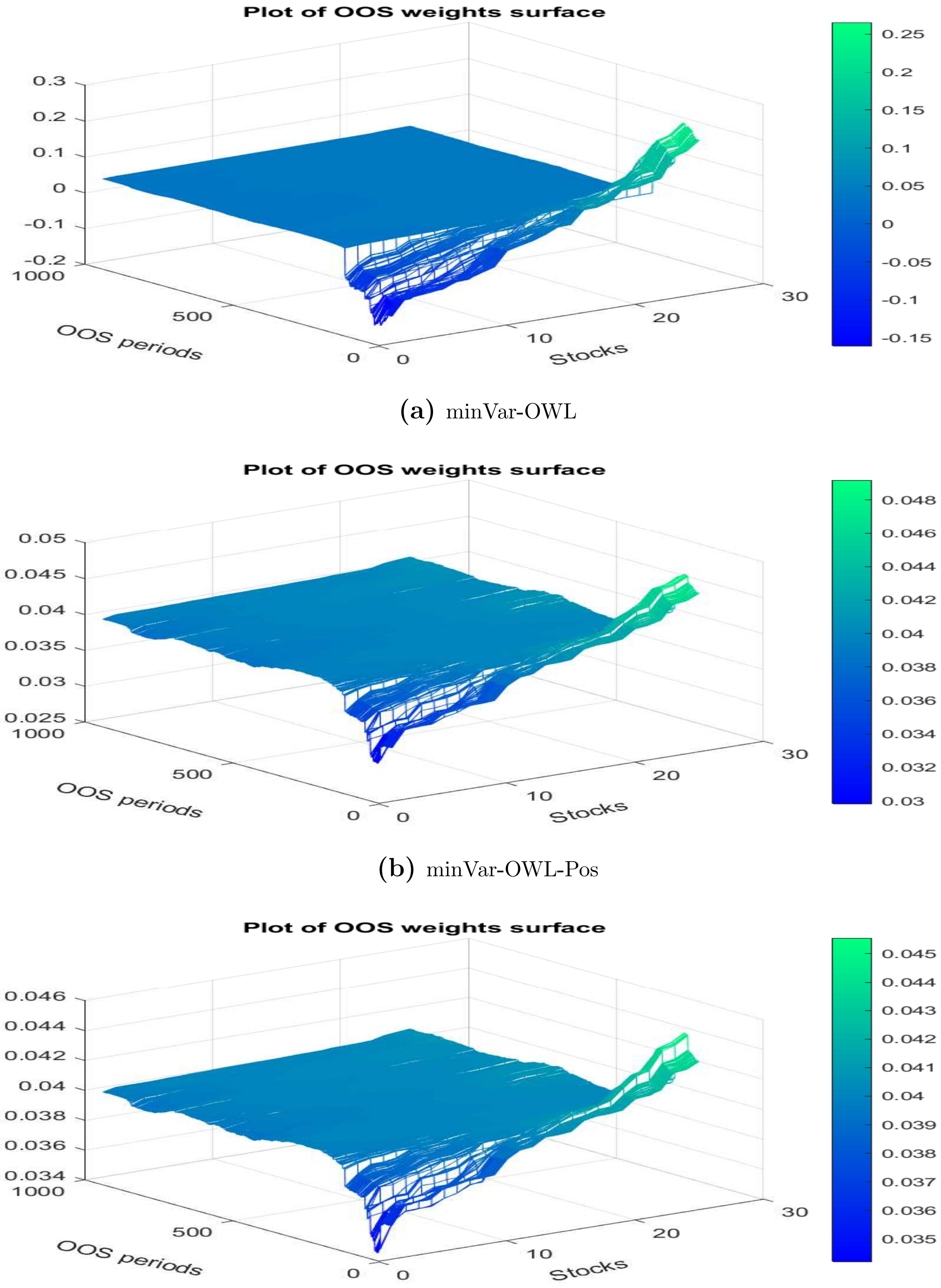
		1	2	3	4	5	6	7	8	9	10	11
EW	1	N/A										
minVar	2	0.0689	N/A									
minVar-JM	3	0.0030	0.0010	N/A								
minVar-LW	4	0.4226	0.0060	0.0889	N/A							
minVar-OWL	5	0.4915	0.0519	0.0090	0.5345	N/A						
minVar-OWL-Pos	6	0.0260	0.0719	0.0030	0.4955	0.8012	N/A					
minVar-OWL-bounds	7	0.0140	0.0829	0.0070	0.4675	0.7023	0.0210	N/A				
minVar-hard-OWL	8	0.4226	0.0420	0.0110	0.5534	0.2567	0.6683	0.5834	N/A			
minVar-LW-OWL	9	0.6583	0.0569	0.0030	0.4655	0.4266	0.9910	0.8611	0.1429	N/A		
MVE-OWL-Pos	10	0.0060	0.2098	0.0010	0.1768	0.0440	0.0070	0.0030	0.0350	0.0500	N/A	
MVE-OWL-bounds	11	0.0160	0.1648	0.0020	0.2498	0.0859	0.0050	0.0080	0.0599	0.1189	0.0110	N/A

Note: this table reports the p-values of Sharpe ratio test according to Ledoit and Wolf (2008) using randomly selected 500 stocks (with daily returns) and 100 stocks (with monthly returns) from CRSP data-set.

related strategies consistently and *significantly* outperform equal weighted strategy.

3.A.4.3 Plot of the out-of-sample portfolio weights distribution

We provide a diagnostic check on out-of-sample stock positions of each portfolio strategy. Figures 3.4 and 3.5 plot the 3-dimensional weights distribution in the OOS period for each asset. Restricted to limited space here, we only list a few strategies and only for the Fama-French 25 portfolios. More results are available upon request.

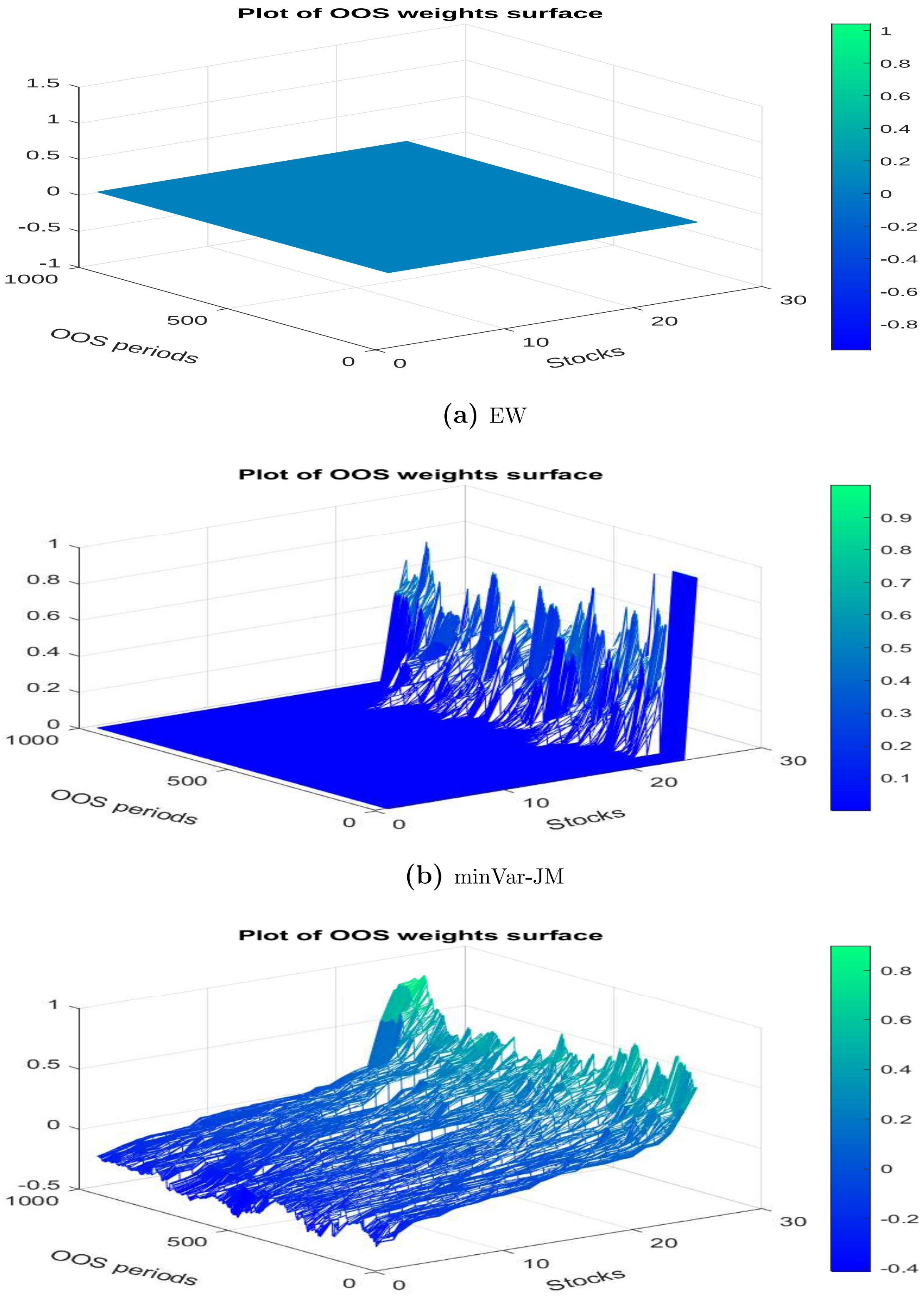


(a) minVar-OWL

(b) minVar-OWL-Pos

(c) minVar-OWL-bounds

Figure 3.4. Weight distribution of various strategies using FF25



(a) EW

(b) minVar-JM

(c) minVar-LW

Figure 3.5. Weight distribution of various strategies using FF25

Bibliography

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.

Ao, M., Y. Li, and X. Zheng (2018): "Approaching Mean-Variance Efficiency for Large Portfolios," *The Review of Financial Studies*, forthcoming.

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.

Babii, A., E. Ghysels, and J. Striaukas (2019): "Estimation and HAC-based Inference for Machine Learning Time Series Regressions," *SSRN Electronic Journal.*

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

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

Belloni, A. and V. Chernozhukov (2012): "High Dimensional Sparse Econometric Models: An Introduction," *SSRN Electronic Journal.*

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects After Selection Among High-dimensional Controls," *Review of Economic Studies*, 81, 608–650.

Bickel, P. J. and E. Levina (2008): "Covariance Regularization by Thresholding," *The Annals of statistics*, 36, 2577–2604.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *Annals of Statistics*, 37, 1705–1732.

Bogdan, M., E. van den Berg, C. Sabatti, W. Su, and E. J. Candès (2015): "SLOPE - Adaptive Variable Selection via Convex Optimization," *Annals of Applied Statistics*, 9, 1103–1140.

Bondell, H. D. and B. J. Reich (2008): "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR," *Biometrics*, 64, 115–123.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2010): "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 3, 1–122.

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

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

Caner, M. and A. B. Kock (2018): "Asymptotically Honest Confidence Regions for High Dimensional Parameters by the Desparsified Conservative Lasso," *Journal of Econometrics*, 203, 143–168.

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

Chen, X., Q. M. Shao, W. B. Wu, and L. Xu (2016): "Self-normalized Cramér-type Moderate Deviations Under Dependence," *Annals of Statistics*, 44, 1593–1617.

Chinco, A., A. D. Clark-Joseph, and M. Ye (2019): "Sparse Signals in the Cross-Section of Returns," *Journal of Finance*, 74, 449–492.

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.

Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*, Oxford University Press.

de Leeuw, J., K. Hornik, and P. Mair (2009): "Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods," *Journal of Statistical Software*, 32.

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal (2009a): "A Generalized Approach to Portfolio Optimization Improving Performance by Constraining Portfolio Norms," *Management Science*, 55, 798–812.

DeMiguel, V., L. Garlappi, and R. Uppal (2009b): "Optimal Versus Naive Diversification : How Inefficient is the 1 / N Portfolio Strategy ?" *Review of Financial Studies*, 22, 1915–1953.

DeMiguel, V., A. Martin-Utrera, F. J. Nogales, and R. Uppal (2020): "A Transaction-Cost Perspective on the Multitude of Firm Characteristics," *The Review Of Financial Studies*, 33, 2180–2222.

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.

DeMiguel, V., Y. Plyakha, R. Uppal, and G. Vilkov (2013): "Improving Portfolio Selection Using Option-Implied Volatility and Skewness," *The Journal of Financial and Quantitative Analysis*, 48, 1813–1845.

Dendramis, Y., L. Giraitis, and G. Kapetanios (2019): "Estimation of Time-Varying Covariance Matrices for Large Datasets," *QMUL working paper.*

Epps, T. W. (1979): "Comovements in Stock Prices in the Very Short Run," *Journal of the American Statistical Association*, 74, 291–298.

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

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

——— (2015): "A Five-factor Asset Pricing Model," *Journal of Financial Economics*, 116, 1–22.

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

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

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

Fastrich, B., S. Paterlini, and P. Winker (2015): "Constructing Optimal Sparse Portfolios Using Regularization Methods," *Computational Management Science*, 12, 417–434.

Feng, G., S. Giglio, and D. Xiu (2020): "Taming the Factor Zoo: A Test of New Factors," *The Journal of Finance*, 1–76.

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 (2020): "Dissecting Characteristics Nonparametrically," *Review of Financial Studies*, 33, 2326–2377.

Gospodinov, N., R. Kan, and C. Robotti (2014): "Misspecification-Robust Inference in Linear Asset-Pricing Models with Irrelevant Risk Factors," *Review of Financial Studies*, 27, 2139–2170.

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

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

Hansen, P. R., A. LUNDE, and J. M. NASON (2011): "The Model Confidence Set," *Econometrica*, 79, 453–497.

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

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

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

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

——— (2018b): "Replicating Anomalies," *The Review of Financial Studies.*

Jagannathan, R. and T. Ma (2003): "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraint Helps," *Journal of Finance*, 58, 1651–1684.

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

Kan, R. and G. Zhou (2007): "Optimal Portfolio Choice with Parameter Uncertainty," *The Journal of Financial and Quantitative Analysis*, 42, 621–656.

Kelly, B. T., S. Pruitt, and Y. Su (2019): "Characteristics Are Covariances : A Unified Model of Risk and Return," *Journal of Financial Economics*, 134, 501–524.

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

Kock, A. B. (2016): "Oracle Inequalities, Variable Selection and Uniform Inference in High-dimensional Correlated Random Effects Panel Data Models," *Journal of Econometrics*, 195, 71–85.

Kock, A. B. and L. Callot (2015): "Oracle Inequalities for High Dimensional Vector Autoregressions," *Journal of Econometrics*, 186, 325–344.

Kock, A. B. and H. Tang (2019): "Uniform Inference in High-dimensional Dynamic Panel Data Models With Approximately Sparse Fixed Effects," *Econometric Theory*, 35, 295–359.

——— (2020): "Shrinking the cross-section," *Journal of Financial Economics*, 135, 271–292.

Ledoit, O. and M. Wolf (2003): "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection," *Journal of Empirical Finance*, 10, 603–621.

——— (2008): "Robust Performance Hypothesis Testing with the Sharpe Ratio," *Journal of Empirical Finance*, 15, 850–859.

——— (2017): "Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz meets Goldilocks," *Review of Financial Studies*, 30, 4349–4388.

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.

Markowitz, H. (1952): "Portfolio Selection," *The Journal of Finance*, 7, 77–91.

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

Michaud, R. O. (1989): "The Markowitz Optimization Enigma: Is 'Optimized' Optimal?" *Financial Analysts Journal*, 45, 31–42.

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.

Sun, C. (2019): "Dissecting the Factor Zoo : A Correlation-Robust Approach," *Chapter One of this Thesis.*

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

Van De Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014): "On Asymptotically Optimal Confidence Regions and Tests for High-dimensional Models," *Annals of Statistics*, 42, 1166–1202.

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.