Copyright
by
Long Zhao
2019

The Dissertation Committee for Long Zhao
certifies that this is the approved version of the following dissertation:

Essays on Data-Driven Optimization

Committee:

Kumar Muthuraman, Supervisor

Deepayan Chakrabarti, Co-Supervisor

Efstathios Tompaidis

Constantine Caramanis

Essays on Data-Driven Optimization

by

Long Zhao

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2019

Dedicated to my parents, Gongqing Zhao and Xiaomei Ji.

Acknowledgments

I am forever indebted to my advisors for introducing me the joy of doing research. Kumar, thank you for mentoring and guiding me through the Ph.D. journey. I not only learned how to be a good researcher and teacher, but also how to be a good person. Deepayan, thanks for spending so many hours with me discussing preliminary ideas. Your timely feedbacks made those explorations like playing video games. Stathis, thank you for the support and care for the past several years.

I am also grateful for having wonderful Ph.D. colleagues. David, I am lucky to fight with you through the hardships and thank you for being my American culture advisor. Yuxin, thank you for always being there for my best and worst days. Yixuan, thanks for frankly sharing different world views which dragged me out of my naivety.

Finally, I want to thank my parents. I could not have completed this degree without your love. You are my first and best audience. You made it possible for me to have a story-worthy life everyday.

Essays on Data-Driven Optimization

Publication No. _____

Long Zhao, Ph.D.
The University of Texas at Austin, 2019

Supervisor: Kumar Muthuraman
Co-Supervisor: Deepayan Chakrabarti

The estimation of a data matrix contains two parts: the well estimated and the poorly estimated. The latter is usually throwing away because the estimations are off. As argued in this paper, ignoring is the wrong thing to do as the poorly estimated part is orthogonal to the well estimated. I will show how to use such orthogonality information via robust optimization and provide application in portfolio optimization, least-square regression, and dimension reduction. Across a large number of experiments, utilizing the orthogonality information consistently improves the performance.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Two-Step Apporach

The most direct way to make decisions based on analytics of data involves two steps. First, I find the best estimations using historical data. Next, I use these estimations as inputs to an optimization problem and solve it to obtain decisions. However, as shown in DeMiguel et al. (2009b), such procedure can be seriously problematic: the most straightforward equally-weighted portfolio, which divides any investment equally amongst the risky assets, has an embarrassingly good out-of-sample performance compared to 14 sophisticated portfolios that belong to the two-step family. The classical portfolio (Markowitz, 1952) is among them.

As will be shown later, the primary cause is the compounding effect of the optimization-driven error amplification on the initial estimation errors. Specifically speaking, the bottom eigenvalues and corresponding eigenvectors tend to be hard to estimate than the others. Unfortunately, the optimization procedure puts too much weights the former resulting in amplifying the error which leads to an unacceptable performance.

A common way to mitigate this effect is to directly ignore the bottom

eigenpairs. Indeed, the popular principal components analysis (PCA) related ideas take this approach. However, I want to argue there is still information in the bottom eigenpairs and they should not be thrown away. The intuition is that the space spanned by the bottom eigenvectors are well estimated because they are orthogonal to the other well-estimated eigenvectors. I call this orthogonality the forgotten information.

I propose a conservative way to utilize the forgotten orthogonality information and demonstrate its value in portfolio optimization, linear regression, and dimension reduction. That is to say, by modifying the second-step optimization based on the characteristics of the first-step estimation error, lots of improvements can be achieved.

1.2 Structure of the Thesis

Chapter 2 uses portfolio optimization to introduce the structure of estimation errors, namely the well estimated and the poorly estimated part, and how the errors from the poorly estimated are amplified through the optimization. I propose a way of portfolio construction by mitigating the amplification issue in Chapter 3. This solution involves using the poorly estimated part via the robust optimization. Chapter 4 focuses on generalizing the idea towards least-squares regression problems. Finally, in Chapter 5, I explore the possibility of utilizing the orthogonality information in dimension reduction applications.

The thesis closely depends on Zhao et al. (2019a) and Zhao et al.

(2019b) which are my collaboration works with Prof. Chakrabarti and Prof. Muthuraman.

Chapter 2

Error Amplification in Portfolio Optimization

2.1 Introduction

The seminal mean-variance portfolio framework (Markowitz, 1952) initiates the modern era of finance by constructing a portfolio by solving an intuitive optimization problem. Here the target is to minimize the variance of a portfolio given its expected return is larger than a target level. Unfortunately, the optimizer does not know either the true expected return and the true covariance matrix. It is natural to use the corresponding sample estimates are used instead. Namely, creating a two-step approach: first estimate then optimize. However, the resulting portfolio has an unacceptable out-of-sample performance (Jobson and Korkie, 1981; Frost and Savarino, 1986, 1988; Jorion, 1986; Michaud, 1989). Even the simpler sample-variance minimizing portfolio, denoted as the estimated Min-Var portfolio, often has a similarly unacceptable performance (Jagannathan and Ma, 2003; DeMiguel et al., 2009b).

A plethora of research papers suggest ways to address this poor out-of-sample performance. However, DeMiguel et al. (2009b) examine 14 popu-

This Chapter closely follows Long Zhao, Deepayan Chakrabarti, and Kumar Muthurman, 'Portfolio construction by mitigating error amplification: The bounded-noise portfolio'. Operations Research, 2019. The method is fine tuned by all authors while I implement all the experiments.

lar methods in terms of their Sharpe ratio, certainty-equivalent return, and turnover, and find that none of the methods consistently outperforms the naïve equally-weighted portfolio which assigns the same weight across all risky assets.

Recently, some papers start to focus on improving the estimated Min-Var portfolio which seems to be an easier problem than the mean-variance one. Among them, some manage to obtain better performance than the naïve portfolio. Next, I will present more details about these methods.

2.2 Literature Review

I divide the literature into three groups. The first category tries to develop methods that provide better covariance estimates than the sample covariance matrix. Namely, it focuses on the first estimation step. In the second category, the estimated Min-Var portfolio is combined with the equally-weighted portfolio to maximize a utility measure other than variance. The third category includes modification of the optimization problem itself with the hope of improving performance. That is to say, the second optimization step is the battlefield.

1. Improving covariance estimation: A lot of research exists on the estimation of the covariance matrix in the context of portfolio optimization.[1] One common approach is to shrink the sample covariance. Ledoit and Wolf

[1]For a more detailed discussion, please see Ledoit and Wolf (2012, 2017) and the references therein.

(2003) shrink the sample covariance matrix toward the single-index covariance matrix. One can also shrink the eigenvalues of the sample covariance matrix linearly (Ledoit and Wolf, 2004) or nonlinearly (Ledoit and Wolf, 2012, 2017). The former is equivalent to shrinking the sample covariance matrix toward identity matrix. The shrinkage level is chosen such that it is asymptotically optimal under the Frobenius norm. The shrinkage methods have been shown to dominate the multi-factor models on the real-world data (Ledoit and Wolf, 2003). A second approach is to use robust statistics to counteract sudden movements in the stock price. DeMiguel and Nogales (2009) provide a careful evaluation on both simulated and real-world datasets and show that the robust statistics can indeed improve performance. A third approach is to use the information from the option price documented in DeMiguel et al. (2013b). They indicate that using option-implied volatility can reduce the out-of-sample standard deviation by more than 10% for various modified Min-Var portfolios on two real-world datasets.

Estimation errors might be reduced by the these methods, but they cannot be eliminated, and I will show that this error is amplified by the solver of the portfolio optimization.

2. Combining with the equally-weighted portfolio: The second category is inspired by the good performance of the equally-weighted portfolio (Jobson and Korkie, 1980; DeMiguel et al., 2009b; Duchin and Levy, 2009). With five reasonable assumptions, Frahm and Memmel (2010) prove that the portfolio constructed by carefully combining the estimated Min-Var portfolio

with any reference portfolio dominates the former. They use a loss function that is closely related to out-of-sample variance. In the extensive simulation test and a small real-world dataset evaluation, they take the equally-weighted portfolio as the reference portfolio and demonstrate the benefit of the combination. By minimizing the expected utility loss, Tu and Zhou (2011) estimate the combination level of each of four different portfolios and the equally-weighted portfolio. Using an exhaustive assessment of both the simulated and the real-world datasets, they show that the new portfolios perform better than the equally-weighted portfolio. DeMiguel et al. (2013a) use different criteria and calibration methods to decide the combination level and show that the combined portfolios can achieve good performance across several real-world datasets.

I will provide theoretical reasons for the good performance of the equally-weighted portfolio and propose an intuitive way to combine it with the estimated Min-Var portfolio.

3. Modifying the optimization: In the third category, the portfolio optimization is modified by penalizing portfolios with some predefined characteristics (or, equivalently, by adding extra constraints based on these characteristics). The most common modification is to avoid aggressive short positions. An extreme case is the no-shorting portfolio, which avoids shorting altogether. This approach is analyzed in Jagannathan and Ma (2003), who argue that the "wrong" no-shorting constraint helps because it reduces the effects of the estimation error. They give evidence for better performance us-

ing both simulated and real-world data. A weaker version of the no-shorting constraint involves penalizing a norm of the portfolio weights,

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\Sigma\boldsymbol{w} + \eta\|\boldsymbol{w}\|_p^p \qquad \text{subject to } \boldsymbol{w}'\mathbf{1} = 1. \tag{2.1}$$

Two common norms are the $\mathbb{L}_1$ norm (Welsch and Zhou, 2007; Brodie et al., 2009; Fan et al., 2012) and the $\mathbb{L}_2$ norm (Lauprête, 2001; DeMiguel et al., 2009a). Among these studies, Fan et al. (2012) is the only one that uses both simulated and real-world data to show better performance and that also provides a mathematical justification. Lauprête (2001) takes the view that norm-constrained portfolios are regularizations that counteract the deviations from the normality of the distribution of returns. Empirical evidence is provided via simulations, but only one real-world dataset is used. DeMiguel et al. (2009a) provide more comprehensive empirical results. They show that the norm-constrained portfolios dominate the equally-weighted portfolio and the estimated Min-Var portfolio, in terms of both the out-of-sample variance and the Sharpe ratio. They also show the relation between norm-constrained portfolios and Bayesian priors on the sample covariance matrix. Gotoh and Takeda (2011) find that the norm constraints are equivalent to the robust constraints associated with the return vector, and Olivares-Nadal and DeMiguel (2018) point out that the norm constraints can be interpreted as the transaction costs. These relations indicate that the same basic idea underpins many seemingly disparate models.

However, the norm-constrained approach presents several problems,

stemming primarily from the ad hoc nature of merely modifying the objective to keep the portfolio weights low. First, Green and Hollifield (1992) argue that the optimal portfolio can have sizeable asset weights. Hence, although norm constraints might help, they also might be wrong because they exclude the optimal solution, which involves large portfolio weights. Second, the choice of the norm is arbitrary. Third, the performance of the norm-constrained portfolios depends on the selection of a parameter that captures the importance of keeping the portfolio weights low; that is, the coefficient of the norm. The best parameter value depends on the particular financial dataset and the amount of training data, and it even changes over the time horizon of a dataset. This makes parameter tuning particularly important.

I will show why the norm-constrained portfolio can achieve good out-of-sample performance and how to construct an even better portfolio endogenously.

2.3 Estimation Errors

In this section, I describe the errors from estimating covariance matrix. Though there are different estimates based on different criterions, surprisingly, their errors all share a similar structure.

The following proposition shows that the relative errors (percentage deviations from the true values) in estimating the large eigenvalues of the true covariance matrix are small while the relative errors in estimating the small eigenvalues are large. I represent the true covariance matrix as Σ and its

estimate as $\hat{\Sigma}$. Here, $||\cdot||_{op}$ denotes the operator norm. The sample size is n, and the number of assets is p.

Proposition 2.3.1 (Eigenvalue Concentration)**.** *Let λ_i and $\hat{\lambda}_i$ represent the i^{th} largest eigenvalues of Σ and $\hat{\Sigma}$, respectively. Then I have:*

$$\frac{|\lambda_i - \hat{\lambda}_i|}{\lambda_i} \leq \frac{\|\Sigma - \hat{\Sigma}\|_{op}}{\lambda_i}.$$

Proof. By Weyl's inequality, $|\lambda_i - \hat{\lambda}_i| \leq \|\Sigma - \hat{\Sigma}\|_{op}$. Dividing both sides by λ_i proves the proposition. □

Estimation errors for the eigenvectors are a bit more complicated to characterize. The following Lemma shows that the estimation error not only depends on $\|\Sigma - \hat{\Sigma}\|_{op}$, but also how separated the eigenvalues are.

Lemma 2.3.2 (Concentration of Eigenvectors (Yu et al., 2015))**.** *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$, respectively. Fix $1 \leq r \leq s \leq p$, and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) > 0$, where $\lambda_0 = \infty$ and $\lambda_{p+1} = -\infty$. Let $d = s - r + 1$. Let $V = (\boldsymbol{v}_r, \boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{\boldsymbol{v}}_r, \hat{\boldsymbol{v}}_{r+1}, \ldots, \hat{\boldsymbol{v}}_s) \in \mathbb{R}^{p\times d}$ have orthogonal columns satisfying $\Sigma\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j$ and $\hat{\Sigma}\hat{\boldsymbol{v}}_j = \hat{\lambda}_j \hat{\boldsymbol{v}}_j$; then there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{d\times d}$, such that:*

$$\|\hat{V}\hat{O} - V\|_F \leq \frac{2^{3/2} d^{1/2} \|\Sigma - \hat{\Sigma}\|_{op}}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}.$$

It is worth to notice that given data, $\|\Sigma - \hat{\Sigma}\|_{op}$ is a constant. That is to say, different estimates might affect the tightness of the bounds but won't change the structure. If $\hat{\Sigma}$ happens to be the sample covariance, Vershynin

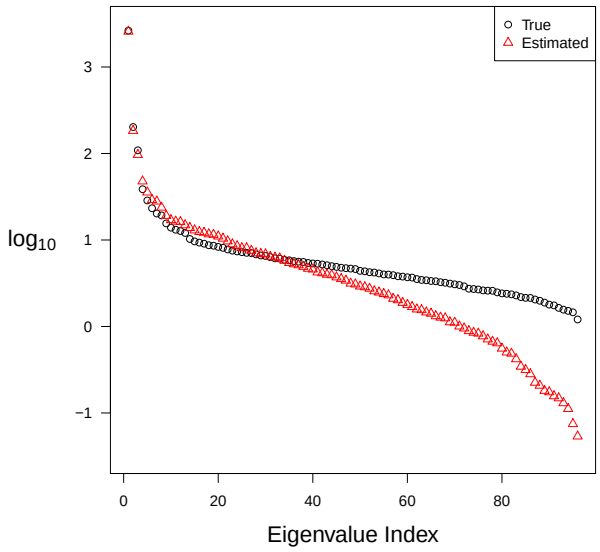
(2011) provides a description of $||\Sigma - \hat{\Sigma}||_{op}$ in terms of n and p: under mild conditions, a high-probability upper bound of $||\Sigma - \hat{\Sigma}||_{op}$ is roughly of order $(p/n)^{\frac{1}{2}-\frac{2}{q}}$, where the qth moment of the data is bounded. Thus, for a given number of assets p, the difference decays when more observations are available, as expected.

Previous work on financial datasets shows that a few factors can explain a significant portion of the variance of asset returns (Fama and French, 2015). This finding suggests that Σ has only a few large eigenvalues (whose corresponding eigenvectors mirror the relevant factors) while the most of the eigenvalues are small (so their eigenvectors have just a small contribution to the variance of asset returns).

This intuition is supported by my observations from a historical covariance matrix constructed from the monthly returns of the Fama-French value-weighted dataset with 96 instruments, aggregated over 625 months. Figure 2.1 shows the eigenvalues of this "true" covariance matrix, as well as those of a sample covariance matrix simulated from the covariance matrix (both of which are ordered from largest to smallest eigenvalue). Observe that the largest eigenvalues are well separated, but the smallest ones are densely packed (note that I scale the y-axis logarithmically). Note also that the relative difference between the estimated and the true eigenvalues is small for the largest eigenvalues, implying that they are relatively well estimated. In addition to these simulation results and the arguments from the finance literature, I see widespread evidence of similar phenomena in the eigenvalue spectra of many

real-world networks (Mihail and Papadimitriou, 2002; Chakrabarti and Faloutsos, 2006).

Figure 2.1: Distribution of True and Estimated Eigenvalues



2.4 Error Amplification

The previous discussion shows that the largest eigenvalues and related eigenvectors in the covariance estimate $\hat{\Sigma}$ are relatively good estimates of the corresponding eigenvalues and eigenvectors of the true covariance matrix Σ. The smaller eigenvalues and the corresponding eigenvectors are poor estimates. Hence, I separate the true eigenvectors $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ into two sets: from index 1 to k, and from $k+1$ to p. When the split index k is chosen appropriately, I expect the first set to be better estimated than the second set. I will show that

the first set of estimated eigenvalues and eigenvectors are also more reliable for portfolio construction, while the remaining ones are not.

For now, given a k, denote the space spanned by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ as $\mathcal{S}$ and the space spanned by the other eigenvectors as $\mathcal{N}$. To understand how these two parts influence portfolio optimization, I first provide a new characterization of the true Min-Var portfolio.

Lemma 2.4.1 (Portfolio Decomposition)**.** *For any separation ($\mathcal{S}$, $\mathcal{N}$), the optimal portfolio $\boldsymbol{w}^*$ can be expressed as:*

$$\boldsymbol{w}^* = \alpha \boldsymbol{w}_S^* + (1-\alpha)\boldsymbol{w}_N^*, \tag{2.2}$$

$$\alpha = \frac{1/RV(\boldsymbol{w}_S^*)}{1/RV(\boldsymbol{w}_S^*) + 1/RV(\boldsymbol{w}_N^*)}. \tag{2.3}$$

Here $\boldsymbol{w}_S^*$ and $\boldsymbol{w}_N^*$ are defined as the solution to the following optimization problems,

$$\begin{aligned} \boldsymbol{w}_S^* = \arg\min_{\boldsymbol{w}} \quad & \boldsymbol{w}'\Sigma\boldsymbol{w}, \\ \textit{subject to} \quad & \boldsymbol{w}'\mathbf{1} = 1, \\ & \boldsymbol{w} \in \mathcal{S}, \end{aligned} \qquad \Bigg| \qquad \begin{aligned} \boldsymbol{w}_N^* = \arg\min_{\boldsymbol{w}} \quad & \boldsymbol{w}'\Sigma\boldsymbol{w}, \\ \textit{subject to} \quad & \boldsymbol{w}'\mathbf{1} = 1, \\ & \boldsymbol{w} \in \mathcal{N}. \end{aligned}$$

That is, $\boldsymbol{w}_S^*$ is the solution to the Min-Var problem given the restriction of being a linear combination of the first k eigenvectors (the vectors that span $\mathcal{S}$) and $\boldsymbol{w}_N^*$ the solution with the restriction of being a linear combination of the other eigenvectors. In the above, $RV(\boldsymbol{w})$ is the out-of-sample variance (henceforth, the realized variance[2]) of $\boldsymbol{w}$, namely,

$$RV(\boldsymbol{w}) = \boldsymbol{w}'\Sigma\boldsymbol{w}.$$

[2]Our definition of realized variance is slightly different from some in the literature. For

Proof. Using the Lagrangian multiplier method, I can easily find:

$$\boldsymbol{w}^* = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} = \frac{\sum_i \frac{\boldsymbol{v}_i'\mathbf{1}}{\lambda_i}\boldsymbol{v}_i}{\sum_i \frac{(\boldsymbol{v}_i'\mathbf{1})^2}{\lambda_i}},$$

where I use $\Sigma^{-1} = \sum_i (1/\lambda_i)\boldsymbol{v}_i\boldsymbol{v}_i'$. Similarly, I have:

$$\boldsymbol{w}_S^* = \frac{\sum_{j=1}^k \frac{\boldsymbol{v}_j'\mathbf{1}}{\lambda_j}\boldsymbol{v}_j}{\sum_{j=1}^k \frac{(\boldsymbol{v}_j'\mathbf{1})^2}{\lambda_j}}, \quad RV(\boldsymbol{w}_S^*) = \frac{1}{\sum_{j=1}^k \frac{(\boldsymbol{v}_j'\mathbf{1})^2}{\lambda_j}}, \quad \frac{1}{RV(\boldsymbol{w}_S^*)}\boldsymbol{w}_S^* = \sum_{j=1}^k \frac{\boldsymbol{v}_j'\mathbf{1}}{\lambda_j}. \quad (2.4)$$

Repeat this process for $\boldsymbol{w}_N^*$, and some algebraic manipulations yield Equation (2.2). □

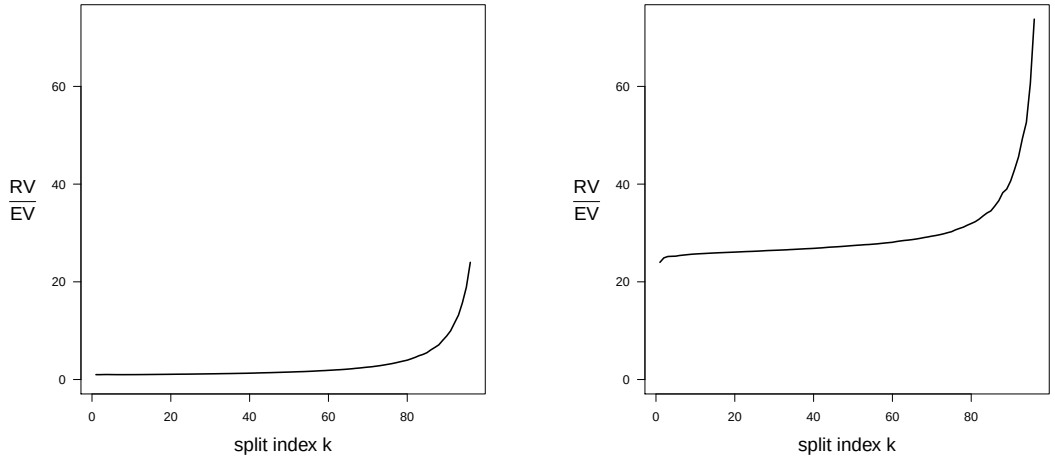
Thus, the true Min-Var portfolio can be seen as a convex combination of two portfolios: one restricted to space $\mathcal{S}$ and the other confined to space $\mathcal{N}$. The weight of each portfolio is proportional to the inverse of its realized variance.

Now consider $\hat{\boldsymbol{w}}^*$. It can be expressed in the same form as in Lemma 2.4.1, but with the true parameters replaced by their estimated counterparts. In particular, the eigenspace $\mathcal{S}$ is replaced by $\hat{\mathcal{S}} = \text{span}(\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_k)$; $\mathcal{N}$ is replaced by $\hat{\mathcal{N}} = \text{span}(\hat{\boldsymbol{v}}_{k+1}, \ldots, \hat{\boldsymbol{v}}_p)$; the portfolios $\boldsymbol{w}_S^*$ and $\boldsymbol{w}_N^*$ are replaced by $\hat{\boldsymbol{w}}_S^*$ and $\hat{\boldsymbol{w}}_N^*$. I use $\hat{\boldsymbol{w}}_S^*$ instead of $\hat{\boldsymbol{w}}_{\hat{S}}^*$ solely to simplify notation. Also, crucially, the realized variance $RV(\boldsymbol{w}) = \boldsymbol{w}'\Sigma\boldsymbol{w}$ is replaced by the estimated variance

example, Hansen and Lunde (2006) directly use the square of returns without subtracting the sample mean. This definition is reasonable when the sample mean is close to 0 and much smaller than the sample variance. This argument is validated in studies that use daily data. However, I use monthly data, and the sample mean is not negligible.

Figure 2.2: The Ratio between RV and EV

The top k eigen portfolio $\hat{\boldsymbol{w}}_S^*$ The bottom $(p-k)$ eigen portfolio $\hat{\boldsymbol{w}}_N^*$



$EV(\boldsymbol{w}) = \boldsymbol{w}'\hat{\Sigma}\boldsymbol{w}$. Thus, the relative weight of $\hat{\boldsymbol{w}}_S^*$ to $\hat{\boldsymbol{w}}_N^*$ in the overall portfolio $\hat{\boldsymbol{w}}^*$ (Equation 2.3) is now driven by the estimated variance instead of the realized variance.

To further illustrate the differences between the realized variance and the estimated variance, **I set $\hat{\Sigma}$ to be the sample covariance matrix** and perform simulations on the Fama-French value-weighted dataset comprising 96 risky assets. In the simulation, I assume that the true covariance matrix Σ and the true expected return $\boldsymbol{\mu}$ are the sample covariance matrix and the sample mean using all monthly data from July 1963 to July 2015 (625 observations). I also assume that the returns follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ, and I draw 120 observations (10-year monthly data)

from this distribution.

I calculate the realized and the estimated variances for various split indices k. I repeat this experiment 100 times and calculate related averages. Figure 2.4 shows the ratio of realized variance to estimated variance for $\hat{\boldsymbol{w}}_S^*$ and $\hat{\boldsymbol{w}}_N^*$. The realized variance of $\hat{\boldsymbol{w}}_S^*$ is similar to its estimated variance when k is small (Figure 2.4 left). However, for $\hat{\boldsymbol{w}}_N^*$, the realized variance is much larger than its estimated variance (Figure 2.4 right). Indeed, it is at least 20 times larger for any k. This underestimation means that $\hat{\boldsymbol{w}}_N^*$, which uses the poorly-estimated parameters, gets overweighted significantly when $\hat{\boldsymbol{w}}_S^*$ and $\hat{\boldsymbol{w}}_N^*$ are combined to construct $\hat{\boldsymbol{w}}^*$.

2.5 Why Do Norm Constraints Work?

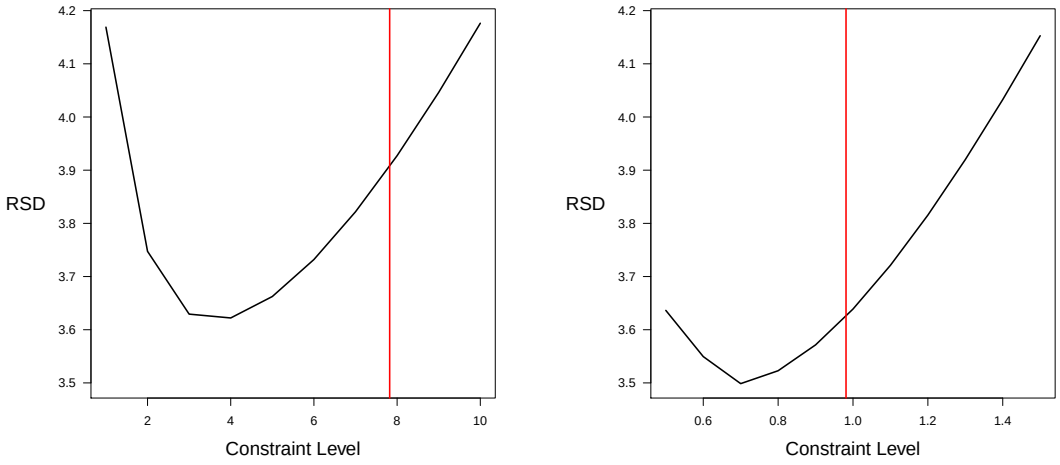
In this section, I will use simulations to show that norm-constrained portfolio perform well not because they propose the right constraints but because they limit the error amplification phenomenon which is the primary cause of unacceptable performance of the estimated Min-Var portfolio.

2.5.1 Imposing the Wrong Constraints

A penalty on the p-norm of portfolio weights, $\|\boldsymbol{w}\|_p$, is equivalent to a constraint of the form $\|\boldsymbol{w}\|_p \leq \delta$ for some $\delta > 0$. Such a constraint can be justified if it renders infeasible a large set of poorly performing portfolios that might otherwise be selected because of estimation errors. However, the constraint must not be so restrictive that even the optimal portfolio $\boldsymbol{w}^*$ becomes

Figure 2.3: Realized Standard Deviation (RSD) with Respect to Different Norm-Constraint Levels

L_1-norm, the vertical line: $\delta = ||\boldsymbol{w}^*||_1$ L_2-norm, the vertical line: $\delta = ||\boldsymbol{w}^*||_2$



infeasible.

Figure 2.3 shows how the realized standard deviation (RSD) varies with different constraint levels, δ, for the $\mathbb{L}_1$ and $\mathbb{L}_2$ norm-constrained portfolios under the simulations using the Fama-French value-weighted dataset with 96 assets. In both cases, as expected, the RSD is too high at the extremes, because the constraints become either too strict or too weak. However, the optimum RSD is achieved for a constraint level at which the optimal is infeasible; indeed, the optimum δ is about half of the norm of the optimal portfolio $\|\boldsymbol{w}^*\|_p$. This agrees with Green and Hollifield (1992), who show that the optimal portfolio could have large weights. Thus, the norm-constrained methods can achieve a

low RSD only by imposing the wrong constraints, and they cannot be justified simply as a means of capping the estimation error effects.

2.5.2 Wrong Constraints Combat Error Amplification Indirectly

The next lemma shows how, for a given k, any portfolio can be split into two unique "projection" portfolios on the top-k eigenspace and the others, and a specific mixing proportion.

Lemma 2.5.1 (Projection Portfolios)**.** *Denote the eigenvectors of* $\hat{\Sigma}$ *by* $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_p$. *For any integer* k *between* 1 *and* p, *let* $\hat{\mathcal{S}} = span(\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_k)$ *and* $\hat{\mathcal{N}} = span(\hat{\boldsymbol{v}}_{k+1}, \ldots, \hat{\boldsymbol{v}}_p)$. *Also introduce matrix* $\hat{S} = (\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_k)$, *and matrix* $\hat{N} = (\hat{\boldsymbol{v}}_{k+1}, \ldots, \hat{\boldsymbol{v}}_p)$. *For any weight* $\boldsymbol{w}$ *that satisfies* $\boldsymbol{w}'\mathbf{1} = 1$, *there is a unique decomposition,*

$$\boldsymbol{w} = \theta \boldsymbol{w}_S + (1-\theta)\boldsymbol{w}_N, \tag{2.5}$$

such that $\boldsymbol{w}_S \in \hat{\mathcal{S}}$, $\boldsymbol{w}_S'\mathbf{1} = 1$, *and* $\boldsymbol{w}_N \in \hat{\mathcal{N}}$, $\boldsymbol{w}_N'\mathbf{1} = 1$. *These "projection portfolios"* $\boldsymbol{w}_S$ *and* $\boldsymbol{w}_N$, *and the inferred mixing proportion* θ, *are given by:*

$$\theta = \boldsymbol{w}'\hat{S}\hat{S}'\mathbf{1}, \quad \boldsymbol{w}_S = \frac{\hat{S}\hat{S}'\boldsymbol{w}}{\boldsymbol{w}'\hat{S}\hat{S}'\mathbf{1}}, \quad \boldsymbol{w}_N = \frac{\hat{N}\hat{N}'\boldsymbol{w}}{\boldsymbol{w}'\hat{N}\hat{N}'\mathbf{1}}. \tag{2.6}$$

Also, as discussed in Section 2.4, the mixing proportion for the estimated Min-Var portfolio is

$$\frac{1/EV(\hat{\boldsymbol{w}}_S^*)}{1/EV(\hat{\boldsymbol{w}}_S^*) + 1/EV(\hat{\boldsymbol{w}}_N^*)}. \tag{2.7}$$

Proof. Clearly, $\boldsymbol{w}_S$ and $\boldsymbol{w}_N$ as defined in Equation (2.6) satisfy $\boldsymbol{w}_S \in \hat{\mathcal{S}}$, $\boldsymbol{w}_S'\mathbf{1} = 1$ and $\boldsymbol{w}_N \in \hat{\mathcal{N}}$, $\boldsymbol{w}_N'\mathbf{1} = 1$. Combining $\hat{S}\hat{S}' + \hat{N}\hat{N}' = I$ with $\boldsymbol{w}'\mathbf{1} = 1$, I have

$$1 = \boldsymbol{w}'\mathbf{1} = \boldsymbol{w}'(\hat{S}\hat{S}' + \hat{N}\hat{N}')\mathbf{1} = \theta + \boldsymbol{w}'\hat{N}\hat{N}'\mathbf{1},$$

which implies $1 - \theta = \boldsymbol{w}'\hat{N}\hat{N}'\mathbf{1}$. Plugging this equation into the right-hand side of Equation (2.5),

$$RHS = \hat{S}\hat{S}'\boldsymbol{w} + \hat{N}\hat{N}'\boldsymbol{w} = \boldsymbol{w} = LHS.$$

In this way, I prove that Equation (2.6) gives one solution. Assume that there is another solution,

$$\boldsymbol{w} = \tilde{\theta}\tilde{\boldsymbol{w}}_S + (1 - \tilde{\theta})\tilde{\boldsymbol{w}}_N.$$

Then, I have

$$\theta\boldsymbol{w}_S - \tilde{\theta}\tilde{\boldsymbol{w}}_S = -(1 - \theta)\boldsymbol{w}_N + (1 - \tilde{\theta})\tilde{\boldsymbol{w}}_N.$$

The left-hand side belongs to $\hat{\mathbb{S}}$ while the right-hand side belongs to $\hat{\mathbb{N}}$. Because $\hat{\mathbb{S}} \cap \hat{\mathbb{N}} = \mathbf{0}$, both sides are $\mathbf{0}$. However, $\boldsymbol{w}_S'\mathbf{1} = \tilde{\boldsymbol{w}}_S'\mathbf{1} = 1$. Therefore, the following holds:

$$0 = \mathbf{0}'\mathbf{1} = (\theta\boldsymbol{w}_S - \tilde{\theta}\tilde{\boldsymbol{w}}_S)'\mathbf{1} = \theta - \tilde{\theta}.$$

The equation implies that $\boldsymbol{w}_S = \tilde{\boldsymbol{w}}_S$ and $\boldsymbol{w}_N = \tilde{\boldsymbol{w}}_N$. □

The strong performance of norm-constrained portfolios could be because they have better 'projection portfolios than the estimated Min-Var portfolio, or because they use a better mixing proportion than relying on the estimated variance (Equation 2.7). I explore this by simulating sample returns from a multivariate normal distribution (with μ and Σ from the Fama-French value-weighted dataset) and constructing portfolios from these samples. I then

calculate the RSD of the corresponding projection portfolios. All results are averaged over 100 iterations.

For brevity, I will call the $\mathbb{L}_1$-norm constrained portfolio the $\mathbb{L}_1$ portfolio with weight vector $\hat{w}^{\mathbb{L}_1}$; the $\mathbb{L}_2$ portfolio with weight vector $\hat{w}^{\mathbb{L}_2}$ is defined accordingly[3]. Here, k is chosen to be the largest number that satisfies the bootstrapped estimated ratio of $RV(\hat{\boldsymbol{w}}_S)/EV(\hat{\boldsymbol{w}}_S)$ is smaller than 1.25. I also tested different thresholds, such as 1.15 and 1.4. The results are similar.

Table 2.1: RSD of Projection Portfolios

Portfolio	$\hat{\boldsymbol{w}}_S^*$	$\hat{\boldsymbol{w}}_S^{\mathbb{L}_1}$	$\hat{\boldsymbol{w}}_S^{\mathbb{L}_2}$	$\boldsymbol{w}_S^{EW}$	$\hat{\boldsymbol{w}}_N^*$	$\hat{\boldsymbol{w}}_N^{\mathbb{L}_1}$	$\hat{\boldsymbol{w}}_N^{\mathbb{L}_2}$	$\boldsymbol{w}_N^{EW}$
Mean RSD(%)	3.696	3.753	3.719	5.168	7.687	6.008	4.928	4.948

Table 2.1 compares the RSD of the projection portfolios for the $\mathbb{L}_1$ and $\mathbb{L}_2$ portfolios, as well as the equally-weighted portfolio (EW). Except the EW portfolio, the signal-space projections of all portfolios the have similar RSD. Thus, even though the $\mathbb{L}_1$ and $\mathbb{L}_2$ portfolios do not explicitly construct a split, they indirectly use the top-k eigenpairs just as effectively.

The $\mathcal{N}$-space projections of the $\mathbb{L}_1$ and $\mathbb{L}_2$ portfolios achieve a much lower RSD than the aggressive noise-only portfolio $\hat{\boldsymbol{w}}_N^*$. Thus, norm-based penalties indirectly lead to improved $\mathcal{N}$-space portfolios. Also, the $\mathcal{N}$-space projection of the EW portfolio is as good as the projection of the $\mathbb{L}_2$ portfolio and much better than the $\mathbb{L}_1$ portfolio.

[3]The penalty parameter is chosen by leave-one-out crossvalidation, as in DeMiguel et al. (2009a). I do a bisection search within the interval $[10^{-4}, 10^4]$ to find the parameter with the lowest cross-validated standard deviation. This "best" parameter is then used to build a portfolio using the entire 120 monthly returns.

Table 2.2: Effects of Mixing Proportion on RSD

Portfolio	$\hat{\boldsymbol{w}}^{\mathbb{L}_1}$	$\hat{\boldsymbol{w}}^{\tilde{\mathbb{L}}_1}$	$\hat{\boldsymbol{w}}^{\mathbb{L}_2}$	$\hat{\boldsymbol{w}}^{\tilde{\mathbb{L}}_2}$	$\hat{\boldsymbol{w}}_S^*$
Mean RSD(%)	3.700	4.215	3.531	3.979	3.696

To investigate the effect of the mixing proportion, I create new portfolios $\tilde{\mathbb{L}}_1$ and $\tilde{\mathbb{L}}_2$ that have the same projection portfolios as the $\mathbb{L}_1$ and $\mathbb{L}_2$ portfolios respectively, but where the mixing proportion is calculated using estimated variances (Equation 2.7).

Table 2.2 shows that the $\tilde{\mathbb{L}}_1$ and $\tilde{\mathbb{L}}_2$ portfolios are much worse than the $\mathbb{L}_1$ and $\mathbb{L}_2$ portfolios, respectively. In fact, they are even worse than the signal-only portfolio, $\hat{\boldsymbol{w}}_S^*$. This indicates that even with improved noise-space projection portfolios, finding the right mixing proportion is essential.

The inferred mixing proportion θ (from Lemma 2.5.1) for the $\mathbb{L}_1$ portfolio is, on average, 1.65 times as large as it for the $\tilde{\mathbb{L}}_1$ portfolio. The corresponding ratio is 2.09 for the $\mathbb{L}_2$ portfolio versus the $\tilde{\mathbb{L}}_2$ portfolio. This shows that norm-constrained portfolios avoid overweighting the noise-space projection portfolios, and hence escape the error amplification trap.

Chapter 3

The Unified Portfolio

3.1 Intuitions

Based on Section 2.3, I know the top-k eigenpairs tend to be much better estimates than other eigenpairs. Thus, it is reasonable to trust the former and obtain the corresponding optimal portfolio. This portfolio happens to be $\hat{\boldsymbol{w}}_S^*$, the $\mathcal{S}$-space projection portfolio of the estimated Min-Var portfolio. For the other eigenpairs, it is tempting to throw them away since they tend to be bad estimates. However, "poorly estimated" does not imply unimportant: the $\mathcal{N}$ space is well estimated because it is orthogonal to the well-estimated $\mathcal{S}$ space while each eigenvector from the $\mathcal{N}$ space is poorly estimated. The orthogonality implies that a portfolio from the noise space has the potential to improve performance when combined with $\hat{\boldsymbol{w}}_S^*$. This space-level information motives a $\mathcal{N}$-space projection portfolio which might only utilizes the space-level information. A perfect candidate is $\boldsymbol{w}_N^{EW}$, the $\mathcal{N}$-space projection portfolio of the equally-weighed portfolio. Not only $\boldsymbol{w}_N^{EW}$ only depends on the $\mathcal{N}$ space but also the good out-of-sample performance of the equally-weighted portfolio has

This Chapter closely follows Long Zhao, Deepayan Chakrabarti, and Kumar Muthurman, 'Portfolio construction by mitigating error amplification: The bounded-noise portfolio'. Operations Research, 2019. The method is fine tuned by all authors while I implement all the experiments.

been widely documented in the literature (Jobson and Korkie, 1980; DeMiguel et al., 2009b; Duchin and Levy, 2009).

Table 2.1 provides evidence for the previous intuition. Among four competing $\mathcal{S}$-space projection portfolios, $\hat{\boldsymbol{w}}_S^*$ achieves the best performance. Meanwhile, $\boldsymbol{w}_N^{EW}$ is a close runner-up among four $\mathcal{N}$-space projection portfolios. Moreover, $\boldsymbol{w}_N^{EW}$ just has a slightly higher RSD than $\hat{\boldsymbol{w}}_S^*$ implying that throwing away other eigenpairs might be wasteful. What left is to choose k endogenously and figure out a smart way to combine $\hat{\boldsymbol{w}}_S^*$ and $\boldsymbol{w}_N^{EW}$.

3.2 Construction of the Unified Portfolio

Justifications for $\hat{\boldsymbol{w}}_S^*$. By construction, the $\mathcal{S}$ space consists of sample eigenvectors whose estimated variance is a reliable indicator of their realized variance. Thus, $\hat{\boldsymbol{w}}_S^*$, constructed from these sample eigenvectors should also be reliable. Mathematically speaking, this portfolio is equivalent to a PCA-based portfolio that ignores a certain number of the low eigenvalues of $\hat{\Sigma}$ and corresponding eigenvectors.

Justifications for $\boldsymbol{w}_N^{EW}$. Because the eigenpairs which belong to the $\mathcal{N}$ space are poorly estimated, one way to robustness is to pick a portfolio that has the best "worst-case" realized variance (i.e., the portfolio that is robust against all possible configurations of eigenvectors from $\hat{\mathcal{N}}$ and is also robust against the corresponding eigenvalues). I could achieve this solution by solving

the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w}} \max_{\Psi \in \mathcal{U}} \quad & \boldsymbol{w}' \Psi \boldsymbol{w}, \\ \text{subject to} \quad & \boldsymbol{w}' \mathbf{1} = 1 \\ & \boldsymbol{w} \in \hat{\mathcal{N}}, \end{aligned} \tag{3.1}$$

where $\mathcal{U}$ is the uncertainty set of all possible covariance matrices Ψ that have the same signal eigenvectors and eigenvalues as $\hat{\Sigma}$. Since Equation (3.1) considers only $\boldsymbol{w} \in \hat{\mathcal{N}}$, I can use the following uncertainty set:

$$\mathcal{U} = \{\Psi \mid \hat{N}' \Psi \hat{N} \preceq b I_{n-\hat{k}+1}\}, \tag{3.2}$$

where b is a constant and $I_{n-\hat{k}+1}$ is a $(n-\hat{k}+1) \times (n-\hat{k}+1)$ identity matrix.

The idea of a robust portfolio has been expressed previously in the literature in the form of the equally-weighted portfolio. This strategy is the right one in the extreme case where no historical data is available. Otherwise, applying this idea just to the noise space is reasonable. Indeed, the projection of the equally-weighted portfolio on the noise space yields precisely the portfolio of Equation (3.1), as shown in Lemma 3.2.1.

Lemma 3.2.1 (The Solution to the Robust Optimization)**.** *The solution to the robust optimization problem Equation (3.1) with the uncertainty set defined in Equation (3.2) is the projection portfolio of the equally-weighted portfolio on* $\hat{\mathcal{N}}$*.*

Proof. Because $\boldsymbol{w} \in \hat{\mathcal{N}}$, I have $\boldsymbol{w} = \hat{N}\boldsymbol{a}$. Thus,

$$\max_{\Psi \in \mathcal{U}} \boldsymbol{w}' \Psi \boldsymbol{w} = \max_{\Psi \in \mathcal{U}} \boldsymbol{a}' \hat{N}' \Psi \hat{N} \boldsymbol{a} = b \boldsymbol{a}' I_{n-k+1} \boldsymbol{a}.$$

The last equality holds because of the definition of the uncertainty set. Then Equation (3.1) becomes

$$\begin{aligned} \min_{\boldsymbol{a}} \quad & b\boldsymbol{a}' I_{n-k+1}\boldsymbol{a}, \\ \text{subject to} \quad & \boldsymbol{a}'(\hat{N}'\mathbf{1}) = 1. \end{aligned}$$

Its solution is

$$\boldsymbol{a}^* = \frac{\hat{N}'\mathbf{1}}{\mathbf{1}'\hat{N}\hat{N}'\mathbf{1}},$$

which implies that the solution to the robust optimization is

$$\hat{N}\boldsymbol{a}^* = \frac{\hat{N}\hat{N}'\mathbf{1}}{\mathbf{1}'\hat{N}\hat{N}'\mathbf{1}}.$$

From Equation (2.6), the projection portfolio of the equal-weighted portfolio on $\hat{\mathcal{N}}$ is:

$$\boldsymbol{w}_N^{EW} = \frac{\hat{N}\hat{N}'(\mathbf{1}/p)}{(\mathbf{1}/p)'\hat{N}\hat{N}'\mathbf{1}} = \frac{\hat{N}\hat{N}'\mathbf{1}}{\mathbf{1}'\hat{N}\hat{N}'\mathbf{1}} = \hat{N}\boldsymbol{a}^*.$$

□

Combine $\hat{\boldsymbol{w}}_S^*$ and $\boldsymbol{w}_N^{EW}$. For each possible split k, I use cross-validation to estimate the variance of $\hat{\boldsymbol{w}}_S^*$ and $\boldsymbol{w}_N^{EW}$, and the covariance between them. Then the combined portfolio is set to be

$$\boldsymbol{w}_k^{Comb} = a_k\hat{\boldsymbol{w}}_S^* + (1-a_k)\boldsymbol{w}_N^{EW}, \tag{3.3}$$

where

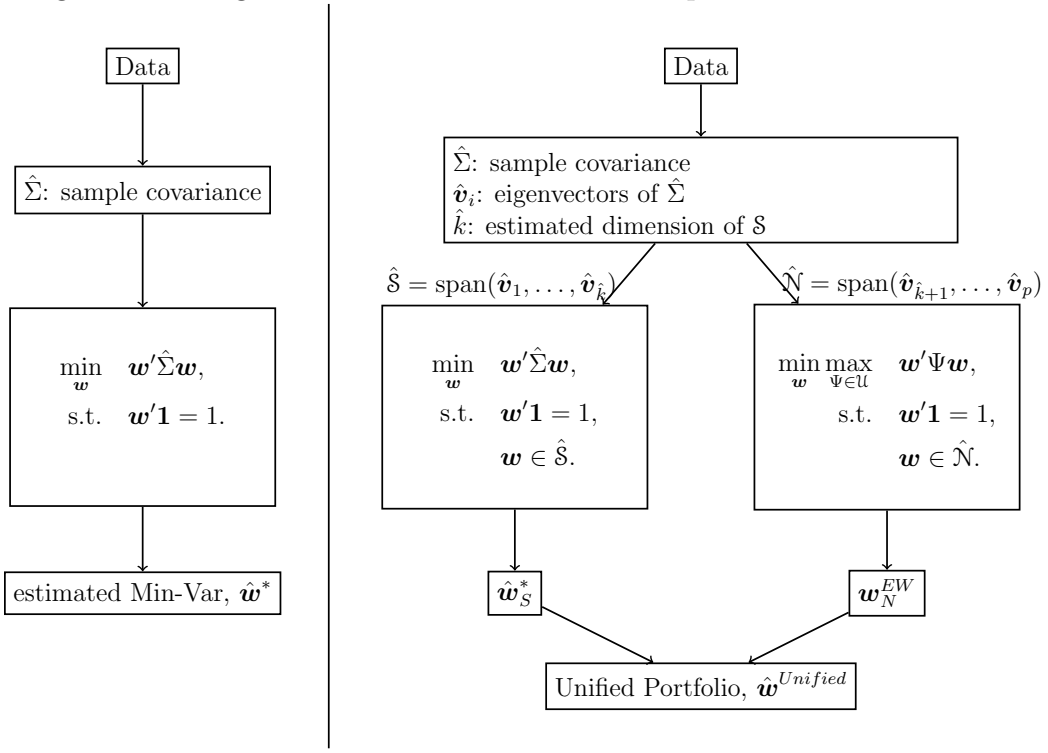
$$a_k = \frac{\hat{Var}(\boldsymbol{w}_N^{EW}) - \hat{Cov}(\hat{\boldsymbol{w}}_S^*, \boldsymbol{w}_N^{EW})}{\hat{Var}(\boldsymbol{w}_N^{EW}) - 2\hat{Cov}(\hat{\boldsymbol{w}}_S^*, \boldsymbol{w}_N^{EW}) + \hat{Var}(\hat{\boldsymbol{w}}_S^*)}. \tag{3.4}$$

If the estimations of variance and covariance are correct, a_k is the optimal level to combine $\hat{\boldsymbol{w}}_S^*$ and $\boldsymbol{w}_N^{EW}$.

Estimating k. Because the eigenpap gradually decreases as the eigenvalue index grows, I also want to create a gradual change from space $\mathcal{S}$ to space $\mathcal{N}$. For each validation set from the cross-validation procedure, I obtain the index of $\boldsymbol{w}_k^{Comb}$ with the lowest variance. Then I take a probabilistic view that k can be the any of the previously selected indexes with equal probability. In this way, a gradual change is achieved.

Figure 3.1 contrasts the classical approach with the unified procedure.

Figure 3.1: Diagram of Estimated Min-Var Compared to Unified Portfolio



3.3 Empirical Results

In this section, I compare the out-of-sample performance of the Unified portfolio to eight other portfolios from the literature (Table 3.2) across twelve different datasets (Table 3.1). The time period for all datasets is July 1963 to July 2015 which shares the same starting point as DeMiguel et al. (2009a). All datasets except the ones for individual stocks come from Kenneth French's website.[1] For the one hundred Fama and French (1992) dataset, because there are missing values for four risky assets for an extended period, I deleted them, leaving 96 of the original 100 portfolios. The individual stocks datasets come from CRSP. There is a challenge in creating the stocks datasets due to market issues like mergers, acquisitions, delistings, IPOs, etc. Ledoit and Wolf (2017) use a procedure that provides a more stable collection of stocks than random selections (Jagannathan and Ma, 2003; DeMiguel et al., 2009a). I use this procedure annually and update my list by choosing the largest 100 or 500 stocks[2], as measured by their market value.[3] Updating the stock list selection annually facilitates my turnover investigations as well (Section 3.3.2).

Competing methods. I consider two naïve portfolios, the equally-weighted (EW) and the value-weighted (VW) portfolio, as my benchmarks. Every asset in the EW portfolio is given equal weight when it is rebalanced. For the VW

[1] `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

[2] I only include the stocks whose returns are available for the past ten years and the future one year.

[3] The number of asset changes for each update is 2.5 and 50 on average for the 100 and 500 stock dataset, respectively.

Table 3.1: List of Datasets Considered

Dataset	*Abbreviation*	p
Six Fama and French (1992) portfolios of firms sorted by size and book-to-market	6FFEW, 6FFVW	6
Ten industry portfolios representing U.S. stock market	10IndEW, 10IndVW	10
Twenty-five Fama and French (1992) portfolios of firms sorted by size and book-to-market	25FFEW, 25FFVW	25
Forty-eight industry portfolios representing U.S. stock market	48IndEW, 48IndVW	48
One hundred Fama and French (1992) portfolios of firms sorted by size and book-to-market	96FFEW, 96FFVW	96
Top 100 market-value individual stocks with annual updates	100	100
Top 500 market-value individual stocks with annual updates	500	500

I use EW (equally-weigthed) and VW (value-weighted) to indicate the corresponding weighting type in the abbreviation.

portfolio, the fraction of the market capitalization is assigned to each asset as its portfolio weight. DeMiguel et al. (2009b) provide a thorough analysis for both portfolios. The EstMinVar portfolio, is the estimated Min-Var portfolio formulated in Markowitz (1952).

In addition to these standard benchmarks, I consider three others that add additional constraints or penalties to the Min-Var portfolio optimization problem. The first one is the shortsale-constrained portfolio (Jagannathan and Ma, 2003, Section 1), which has a non-negativity constraint on the portfolio weights. I call it the NoShorting portfolio. The remaining two are norm-constrained portfolios, with parameters set via cross-validation over standard deviation. These portfolios are detailed in DeMiguel et al. (2009a, Section 3.1 and 3.2). The L_1-norm constrained portfolio is labeled as $\mathbb{L}_1$, and the L_2-norm constrained portfolio is labeled as $\mathbb{L}_2$.

Finally, I also include two relatively recent and well-performing benchmarks. The partial Min-Var portfolio, whose parameter is calibrated by maximizing the portfolio return in the previous period, is labeled as PARR and is detailed in DeMiguel et al. (2009a, Section 3.3). Ledoit and Wolf (2017, Section 3.4) introduce the nonlinear shrinkage method, which provides an ex-

Table 3.2: List of Portfolios Considered in Empirical Experiments

Model	*Abbreviation*
Unified Portfolio	Unified
Equally-weighted portfolio	EW
Value-weighted portfolio	VW
Min-Var portfolio with sample covariance	EstMinVar
Min-Var portfolio with sample covariance and shortsale constrained	NoShorting
L_1-norm-constrained Min-Var portfolio	$\mathbb{L}_1$
L_2-norm-constrained Min-Var portfolio	$\mathbb{L}_2$
Partial Min-Var portfolio with parameter calibrated by maximizing portfolio return in previous period	PARR
Min-Var portfolio with nonlinear shrunk covariance	NonLin

The penalty parameter for the norm-constrained portfolios is chosen by cross-validation over standard deviation.

cellent estimation of the covariance matrix. I call the corresponding portfolio the NonLin portfolio.

Evaluation method. I report two performance measures: the out-of-sample standard deviation and the out-of-sample Sharpe ratio. The turnover discussion can be seen in Section 3.3.2. Following the convention of Brodie et al. (2009); DeMiguel et al. (2009a), and Fan et al. (2012), I use the "rolling-horizon" procedure, which uses a fixed-length training period to estimate. I denote the length of training period as $n < T$, where T is the total number of observations in the dataset. As in DeMiguel et al. (2009a), I use $n = 120$ (10-year monthly return data). I construct various portfolios using the same training data. Then, I roll over to the next month, dropping the earliest month from the previous training window. This procedure yields $T - n$ portfolio-weight vectors for each portfolio. I denote the weight vector as $\boldsymbol{w}_t^i$ for $t = n, \ldots, T-1$ and for each portfolio i.

Following DeMiguel et al. (2009a), I hold the portfolio weight $\boldsymbol{w}_t^i$ for one month. This approach generates the out-of-sample return for time $t+1$: $r_{t+1}^i = (\boldsymbol{w}_t^i)'\boldsymbol{r}_{t+1}$, where $\boldsymbol{r}_{t+1}$ denotes the asset returns at time $t+1$. I use the time series of returns and weights to calculate the out-of-sample standard deviation:

$$(\hat{\sigma}^i)^2 = \frac{1}{T-n-1}\sum_{t=n}^{T-1}\left((\boldsymbol{w}_t^i)'\boldsymbol{r}_{t+1} - \hat{\mu}^i\right)^2, \quad \text{where}$$

I use Levene's test (Levene, 1960) to calculate the statistical significance of the difference in the standard deviation. This test, with the sample median as an estimation of the location parameter, is favored in the literature because of its power and robustness against non-normality (Brown and Forsythe, 1974; Conover et al., 1981; Lim and Loh, 1996).

3.3.1 Out-of-Sample Standard Deviation

Table 3.3: Out-of-Sample Monthly Standard Deviation in Percentage

Portfolio	6FFEW	6FFVW	10IndEW	10IndVW	25FFEW	25FFVW	48IndEW	48IndVW	96FFEW	96FFVW	100	500
Unified	4.475	4.068	3.573	3.644	**3.672**	**3.684**	3.658	**3.576**	**3.685**	**3.657**	3.467	3.152
EW	5.418	4.916	5.732	4.308	5.348	5.107	5.712	4.900	5.414	5.204	4.624	4.795
VW	5.133	4.453	5.817	4.031	4.814	4.409	5.321	4.347	4.746	4.424	4.388	4.386
EstMinVar	4.474	4.059	3.559	3.609	3.858	3.878	5.984	9.978	7.172	7.077	6.499	*NA*
NoShorting	4.870	4.377	3.605	3.615	4.614	4.293	**3.597**	3.694	4.506	4.267	3.482	3.332
$\mathbb{L}_1$	**4.415**	4.058	3.720	3.680	3.758	3.790	3.754	3.605	3.902	3.757	3.602	3.487
$\mathbb{L}_2$	4.468	4.066	**3.514**	**3.574**	3.703	3.697	3.697	3.588	3.723	3.651	**3.410**	3.133
PARR	4.652	4.154	4.518	3.792	4.101	3.981	4.783	4.291	5.244	5.186	5.157	3.546
NonLin	4.469	**4.044**	3.545	3.583	3.690	3.717	3.662	3.651	3.732	3.666	3.435	**3.047**

Notes. This table reports the monthly out-of-sample standard deviation as a percentage. The number in **bold** is the smallest standard deviation for one dataset. The p-value is calculated between the **Unified** portfolio and other portfolios.
One underline, two underlines, and three underlines indicate that the related p-value is smaller than .1, .05, and .01, respectively.
Because the sample covariance is degenerate, there is an NA of the estimated Min-Var portfolio portfolio.

Table 3.3 shows that the Unified portfolio achieves the best out-of-sample standard deviation on five out of the six large[4] portfolio datasets and

[4]I use the phrase large datasets when the number of assets, p, is larger than ten.

is second-best on the 48IndEW dataset. For all datasets, the Unified portfolio is always significantly[5] better than the EW portfolio. The results for the stock portfolios should be interpreted with caution as these are aggregates over not perfectly comparable stock datasets.

For the small datasets, the out-of-sample standard deviation of the EstMinVar portfolio is only about 1% larger than the best portfolio. This relationship indicates that 120 observations are enough for the small datasets to have the whole eigenspace as the signal space. Hence, the BN portfolio should not differ much from the EstMinVar portfolio, and indeed the correlation between their returns is more than 0.99. For the same reason, I expect cross-validation to determine very loose norm constraints for all the norm-constrained methods. Thus, their corresponding portfolios should be essentially the same as the EstMinVar portfolio. This result is again supported by the high correlation (about 0.99) between the returns of the norm-constrained portfolios and the EstMinVar portfolio. Meanwhile, the NoShorting portfolio's constraint cannot be relaxed, and as expected its performance suffers because its constraint interferes with portfolio selection using a well-estimated covariance matrix. However, it does better on some big datasets, where its constraint helps to avoid the effects of covariance estimation errors.

[5] p-value is less than .05 (Levene's test).

3.3.2 Robustness of Holding Length and Turnover

To get a sense of how portfolio performance depends on turnover, I compare the performance of the earlier monthly-rebalanced portfolios with the annually-rebalanced portfolios (Brodie et al., 2009). This allows us to evaluate the effects of turnover without making the results sensitive to either the type or the magnitude of transaction costs. The primary benefit here is that the performance measure now coincides with the objective, making it a fair comparison. The secondary benefit is that, from a taxation perspective, holding a portfolio one year also reduces the taxation rate from short term to long term. Olivares-Nadal and DeMiguel (2018) show that by penalizing the turnover in the portfolio construction procedure, it is possible to sharply reduce the turnover without sacrificing much in performance.

Table 3.4: Hold for One Year, Out-of-Sample Monthly Standard Deviation in Percentage

Portfolio	6FFEW	6FFVW	10IndEW	10IndVW	25FFEW	25FFVW	48IndEW	48IndVW	96FFEW	96FFVW	100	500
Unified	4.845	4.583	4.473	3.565	3.933	**3.833**	4.912	**3.593**	**3.994**	**3.799**	3.545	3.290
EW	5.388	4.911	5.695	4.276	5.320	5.109	5.661	4.843	5.372	5.203	4.501	4.633
VW	5.128	**4.450**	5.788	4.040	4.796	4.404	5.300	4.340	4.740	4.443	4.380	4.379
EstMinVar	4.835	4.606	4.513	3.577	4.130	3.950	27.439	11.896	7.397	7.417	7.232	NA
NoShorting	4.908	4.469	**3.628**	3.630	4.653	4.353	**3.634**	3.761	4.597	4.364	**3.522**	3.382
$\mathbb{L}_1$	4.860	4.607	3.746	3.642	4.034	3.935	4.372	3.682	4.126	4.006	3.789	3.357
$\mathbb{L}_2$	4.835	4.613	4.198	**3.540**	**3.922**	3.824	4.835	3.664	4.027	3.864	3.523	3.243
PARR	4.985	4.821	4.427	3.738	4.291	4.473	4.833	4.255	5.505	6.292	5.463	3.511
NonLin	**4.796**	4.561	4.411	3.560	3.970	3.839	4.847	3.705	4.034	3.825	3.573	**3.228**

Notes. This table reports the monthly out-of-sample standard deviation as a percentage. The number in **bold** is the smallest standard deviation for one dataset. The p-value is calculated between the **Unified** portfolio and other portfolios.
One underline, two underlines, and three underlines indicate that the related p-value is smaller than .1, .05, and .01, respectively.
Because the sample covariance is degenerate, there is an NA of the estimated Min-Var portfolio portfolio.

Compared to Tables 3.3, Tables 3.4 show that the performance of the low turnover portfolios (EW, VW, and NoShorting) remains similar.

3.3.3 Robustness of Training Length

In this subsection, following Brodie et al. (2009), I show the results using the same datasets but with only 60 (5-year monthly data) observations as training data. When the length of rolling window n is not larger than the number of assets p, the sample covariance matrix is singular.[6] Especially since the portfolio construction problem assumes stationarity over n periods, small values of n are common. Hence, assessing the performance of portfolio optimization in the degenerate case (i.e., $n \leq p$) is important. By using 60 observations, the sample covariance matrix for datasets 96FFEW, 96FFVW, 100, and 500 are singular.

Table 3.5: Out-of-Sample Monthly Standard Deviation in Percentage Using 60 Observations

Portfolio	6FFEW	6FFVW	10IndEW	10IndVW	25FFEW	25FFVW	48IndEW	48IndVW	96FFEW	96FFVW	100	500
Unified	**4.268**	3.979	**3.483**	**3.602**	**3.740**	**3.712**	3.683	**3.553**	**3.813**	**3.703**	3.516	3.233
EW	5.418	4.916	5.732	4.308	5.348	5.107	5.712	4.900	5.414	5.204	4.624	4.795
VW	5.133	4.453	5.817	4.031	4.814	4.409	5.321	4.347	4.746	4.424	4.388	4.386
EstMinVar	4.292	3.992	3.611	3.719	4.447	4.381	7.489	11.168	NA	NA	NA	*NA*
NoShorting	4.741	4.296	3.565	3.610	4.518	4.262	3.665	3.615	4.453	4.202	3.553	3.341
$\mathbb{L}_1$	4.399	4.121	3.800	3.723	3.912	3.942	3.900	4.031	4.286	4.418	3.928	3.462
$\mathbb{L}_2$	4.278	3.973	3.505	3.635	3.775	3.726	3.836	3.742	4.047	3.955	3.669	3.119
PARR	4.572	4.129	4.286	3.773	4.345	4.167	5.213	5.209	4.549	4.722	4.177	3.538
NonLin	4.278	**3.947**	3.518	3.616	3.742	3.770	**3.607**	3.590	3.822	3.782	**3.485**	**3.078**

Notes. This table reports the monthly out-of-sample standard deviation as a percentage. The number in **bold** is the smallest standard deviation for one dataset. The p-value is calculated between the **Unified** portfolio and other portfolios.
One underline, two underlines, and three underlines indicate that the related p-value is smaller than .1, .05, and .01, respectively.
To allow for a fair comparison with the 120-observation case, I truncate the return to the same period.
Because the sample covariance is degenerate, there are NAs of the estimated Min-Var portfolio portfolio.

The results in Table 3.5 show that the Unified portfolio is the best on eight out of ten portfolio datasets, including five (of six) large portfolio datasets, and the second-best for the sixth. Comparing Table 3.3 to Table 3.5,

[6] In the calculation of the sample covariance matrix, the sample mean is subtracted. Thus, when $n \leq p$, the rank of the sample covariance matrix is at most $n - 1$, which is smaller than p.

I find that the out-of-sample standard deviation of the BN portfolio is robust to the choice of training length. The reason for the robustness is that both the Unified portfolio becomes more cautious when training length becomes smaller. Indeed, the $\mathbb{S}$ space becomes smaller when fewer observations are available.

As shown in Table 3.5, the out-of-sample standard deviations of the $\mathbb{L}_1$ portfolio and $\mathbb{L}_2$ portfolio increase significantly compared to those in Table 3.3. This change increases the margin between the standard deviations of the Unified portfolio and other portfolios. For example, for the dataset 96FFVW, the standard deviation of the BN portfolio is 6% better than that of the $\mathbb{L}_2$ portfolio and 11% better than that of the $\mathbb{L}_1$ portfolio. The intuitive reason is that, the cross-validation for them is unable to generate a more conservative portfolio when there are fewer data. In fact, in about 36% of the time periods, the penalty parameter (Equation 2.1) with 60 observations η_{60} is smaller than η_{120}.

Chapter 4

Unified Classical and Robust Optimization

4.1 Introduction

Regression analysis and its variants have become the primary workhorse of statistical and machine learning techniques in quantitative social sciences. The classical ordinary least-squares (LS) regression problem is to find a p-vector $\boldsymbol{b}$, given an $n \times p$ data matrix X and a n-vector of observations $\boldsymbol{y}$, such that $||\boldsymbol{y} - X\boldsymbol{b}||_2$ is minimized. The result of this L_2-norm minimization of the residuals has favorable properties if the underlying assumptions on $(X, \boldsymbol{y})$ are true. Some of these assumptions include linearity, homo-scedasticity, no-autocorrelation, normality of residuals and the error-free observation of X. If these assumptions are violated the results can be very misleading (Eldén, 1980; Björck, 1991; Van Huffel and Vandewalle, 1991; Higham and Higham, 1992; Fierro and Bunch, 1994; Golub and Van Loan, 2012). Moreover, the presence of unusual observations (data that do not belong to the same data generating process) could severely distort the LS estimates even when the data sets are large (Andersen, 2008). Apart from the risk of providing misleading ex-

This Chapter closely follows Long Zhao, Deepayan Chakrabarti, and Kumar Muthurman, 'Unified classical and robust optimization for least squares'. Submitted to Operations Research. The method is fine tuned by all authors while I implement all the experiments.

planatory variables and coefficients, this sensitivity and the over-fitting nature of ordinary LS regression also ends up degrading the out-of-sample predictive performance (Eldén, 1980; Higham and Higham, 1992; Fierro and Bunch, 1994; Golub and Van Loan, 2012), which is the primary objective in several application settings.

Ordinary least squares (OLS) method is hence said to be not robust, especially for small data sizes. Several different ways of addressing this sensitivity have been proposed. Most popular among these are the regularization methods like ridge regression (Tikhonov, 1943), LASSO regression (Tibshirani, 1996), principal components regression (PCR) (Hotelling, 1957), and partial least squares regression (PLS) (Wold, 1966). While ridge regression adds the L_2 norm of $\boldsymbol{b}$ as a regularization term to the LS objective, LASSO uses the L_1 norm and is hence sparsity-inducing. If the corresponding underlying specifications are correct, these methods will provide a better estimation of $\boldsymbol{b}$ than OLS (Hoerl and Kennard, 1970; Tibshirani, 1996; Park, 1981). As pointed out in Golub and Van Loan (2012), the choice of weights (or regularization parameter) is usually not obvious and application dependent. Criteria for optimizing the regularization parameter have been proposed, but are however chosen using some additional information (see El Ghaoui and Lebret (1997) and references therein). Extensive surveys of regularizations include Nashed (1981); Demoment (1989); Hanke and Hansen (1993).

Various other alternatives have also been proposed to address this sensitivity and are commonly referred to as *Robust regression methods* (Andersen,

2008). Robust regression methods are designed to be not overly affected by violations of assumptions by the underlying data-generating process. Some LS alternatives include the least absolute deviations method, the M-estimator ("M-" standing for "maximum likelihood type"), least trimmed squares, Theil-Sen estimator, S-estimator and the MM-estimator. Each of these estimators has their pros and cons and are usually robust towards specific types of outliers. See Andersen (2008) and Rousseeuw and Leroy (2005) for a detailed discussion of these methods and their sensitivities. Another approach is to replace the normal distribution assumption of residuals with a heavy-tailed distribution like the t-distribution (Lange et al., 1989) or a mixture model of normal distributions. Bayesian robust regression relies heavily on such distributions (Gelman et al., 2003). Nevertheless, such models still assume that the assumptions they make on the distribution of residuals are true. The method of unit weights is also considered a robust method. However, Bobko et al. (2007) conclude that decades of empirical studies show that unit weights perform similarly to ordinary regression weights on cross-validation.

More specifically though, the term "Robust regression" is used for the many robust optimization counterparts (Xu et al., 2009) of the LS problem. The LS problem being fundamentally an optimization problem, the robust counterparts seek in their modified objectives, a certain measure of robustness against uncertainty in the data. In other words, they deal with the problem of error in variables. Robust Optimization problems only require knowledge of the support of the uncertain data, rather than the full distribution itself (Ben-

Tal et al., 2009; Bertsimas et al., 2011). These robust optimization counterparts have shown to provide robustness against assumptions, perturbations, and outliers and also help increase out-of-sample predictive power (Ben-Tal et al., 2009). The robust counterparts begin by defining an uncertainty set that contains the unknown but bounded disturbance in the data. Given an uncertainty set that captures the ambiguity in the data, the robust counterpart minimizes the largest possible L_2 norm of the residuals for all probable cases belongs to the uncertainty set. The uncertainty sets that have been considered include bounded total perturbation errors in data matrices (El Ghaoui and Lebret, 1997) and bounded individual disturbance in independent variables (Xu et al., 2009). It has been shown that for the former is equivalent to ridge (El Ghaoui and Lebret, 1997) while the latter is the same as LASSO (Xu et al., 2009), thereby allowing the interpretation that the LASSO and ridge techniques are the robust versions of the fundamental LS optimization problem.

In general, the practical impact of robust optimization methods has been limited primarily due to three reasons. Firstly, by design, robust optimization focuses on worst-case performance as the primary way of making the results robust. Hence, the robust solution is sometimes too conservative. Secondly, to alleviate the first problem, additional knowledge and assumptions are required to obtain a smaller or more reasonable uncertainty set. Ellipsoidal approximations of the true uncertainty set (El Ghaoui and Lebret, 1997; El Ghaoui et al., 1998; Ben-Tal and Nemirovski, 1998), the assumption

that only a few of the parameters are uncertain (Bertsimas and Sim, 2004), or the assumption that the distribution belongs to a tractable family of distributions (Delage and Ye, 2010) have all been used. However, the choice is often driven by the desire for mathematical convenience: it is unclear when these assumptions and approximations are reasonable, or how one should pick the right distribution family. Finally, the solution to robust optimization problems can be sensitive to seemingly minor differences even in the size of the uncertainty set. That is to say, one might end up trading one type of sensitivity for another type.

4.1.1 Contribution and Outline

It is understandable that for small-size data the robust versions of the LS problem is more reliable than the classical methodology while the classical is more trustable for massive data. Unfortunately, most problems have data sizes that cannot be characterized as very large or very small. Hence, by recognizing the advantages and disadvantages of both the classical LS problem and the robust optimization counterparts, in this paper, I construct a new method that strikes a balance between these two approaches. More specifically, I first present a new robust version of the LS problem that facilitates my methodology. In this robust optimization, the size of the uncertainty set does not affect the solution. Namely, I are not trading one type of sensitivity for another. I then construct a sequence of problems from the classical LS on one end to my Robust LS on the other end, by parameterizing them. The

parametrization is in terms of the number of eigenvalue-eigenvector pairs that are well estimated, which I obtain from the data itself. my method, called ULS (Unified Least Squares), essentially splits the feasible space into two: the well-estimated subspace and the not-well-estimated. In the former, I solve the classical LS problem, and in the latter, I solve my robust variant of the LS problem. Finally, I combine these two to yield a prediction.

Eigenpairs are the natural basis for the data and hence using eigenpairs or their estimates to aid prediction has been shown to add value in several contexts, such as finance (Chen et al., 2014; Chen and Yuan, 2016; Zhao et al., 2019a), clustering (Ng et al., 2002), and low-rank models (Blei et al., 2003; Airoldi et al., 2008; Mao et al., 2018). However, eigenpairs have not been used as the building block for a sequence of problems spanning between classical and robust variations of a problem.

Using a simulation test I demonstrate that the ULS tends to have a better and more stable on-average performance than other methods. I also consider 68 experiments based on 17 different real-world datasets. The results show that the ULS consistently outperforms methods, like PLS and PCR, that ignore the not-well-estimated subspace. This shows that the robust optimization part of ULS is very valuable. The ULS also outperforms both ridge and LASSO regression by a big margin for more than 20 experiments.

The rest of the Chapter is structured as follows. Section 4.2 provides an overview and the rationale behind my methodology while Section 4.3 describes the method in detail. Section 4.4 provides several insights into the method-

ology while drawing connections between my methodology and other popular methods. Section 4.5 collects the results from my simulation experiments and Section 4.6 collects those from empirical experiments. Concluding remarks are presented in Section 4.7.

4.2 Overview

Consider the regression model, where the data are independent and identically distributed draws from some unknown distribution $q(.)$. The goal is to minimize the expected squared prediction error:

$$\min_{\boldsymbol{b}\in\mathbb{R}^p} E_{(y,\boldsymbol{x}^T)\sim q}(y-\boldsymbol{x}^T\boldsymbol{b})^2. \tag{4.1}$$

Without losing generality I can assume that $Ey=0,\ E\boldsymbol{x}=\mathbf{0}$ and that there is no intercept term. By introducing $\boldsymbol{z}=\begin{pmatrix}1\\ \boldsymbol{b}\end{pmatrix}$ the in Eq. 4.1 above becomes

$$f_\Theta(\boldsymbol{z})=\boldsymbol{z}^T\Theta\boldsymbol{z},\quad \text{where } \Theta=E_{(y,\boldsymbol{x}^T)\sim q}\begin{bmatrix} y^2 & -y\boldsymbol{x}^T\\ -y\boldsymbol{x} & \boldsymbol{x}\boldsymbol{x}^T\end{bmatrix}.$$

Rewriting Eq. 4.1,

$$\min_{\boldsymbol{z}\in\mathcal{Z}} f_\Theta(\boldsymbol{z}),\quad \text{where } \mathcal{Z}=\{\boldsymbol{z}\in\mathbb{R}^{p+1}\mid \boldsymbol{z}^T\boldsymbol{e}_1=1\}. \tag{4.2}$$

Here, $\boldsymbol{e}_1$ is the unit vector along the first dimension with the solution to Eq. 4.1 being the last p components of the solution to Eq. 4.2. Note that the (out-of-sample) R^2, a common measure of success in regression, can be written as $OR^2(\boldsymbol{z})=1-\frac{f_\Theta(\boldsymbol{z})}{Ey^2}=1-\frac{f_\Theta(\boldsymbol{z})}{f_\Theta(\boldsymbol{e}_1)}$, where $\boldsymbol{z}\in\mathcal{Z}$. Thus, minimizing $f_\Theta(\boldsymbol{z})$ is the same as maximizing the out-of-sample R^2.

Since Θ is unknown, it is impossible to find the optimal solution of Eq. 4.2. Instead, I desire a feasible solution with a good out-of-sample performance (like the expected squared prediction error) from n observations. A common approach is to minimize the in-sample sum of squared errors:

$$\min_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{b})^2, \tag{4.3}$$

where $(y_i, \boldsymbol{x}_i^T)$ is the i^{th} observation. Obviously, Eq. 4.3 is equivalent to

$$\min_{\boldsymbol{z}\in\mathcal{Z}} f_{\hat{\Theta}}(\boldsymbol{z}), \quad \text{where } \hat{\Theta} = \frac{1}{n} \begin{pmatrix} \boldsymbol{y}^T\boldsymbol{y} & -\boldsymbol{y}^T X \\ -X^T\boldsymbol{y} & X^T X \end{pmatrix}, \tag{$\hat{\boldsymbol{\mathcal{P}}}$}$$

where $\boldsymbol{y}$ is a column vector of y_i, and X is a matrix whose i^{th} row is $\boldsymbol{x}_i$. I will call the solution $\hat{\boldsymbol{z}}^\star$ of ($\hat{\boldsymbol{\mathcal{P}}}$) the ordinary least squares solution, or the "classical" solution.

Clearly, the estimation error $\hat{\Theta} - \Theta$ affects the out-of-sample performance of $\hat{\boldsymbol{z}}^\star$. Under mild conditions, the operator norm $\|\hat{\Theta} - \Theta\|$ of the error decays as $O((log_2 log_2 p)^2 (p/n)^{1/2-2/q})$ with high probability, where the qth ($q > 4$) moment of the data is bounded (Vershynin, 2011). Thus, for a fixed number of covariates p, as the number of observation n grows to infinity, the out-of-sample performance of $\hat{\boldsymbol{z}}^\star$ converges to the optimal. However, when data are limited, the estimation of Θ can be so poor that $\hat{\boldsymbol{z}}^\star$ can be inferior to the solution $\tilde{\boldsymbol{z}}^\star$ of the following robust optimization problem:

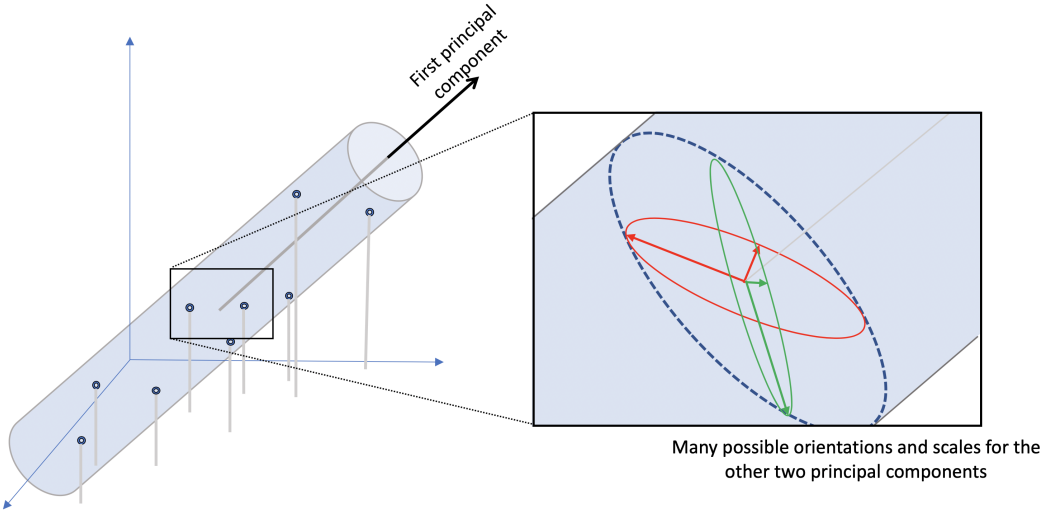
$$\min_{\boldsymbol{z}\in\mathcal{Z}} \max_{\Theta\in\mathcal{U}} f_{\Theta}(\boldsymbol{z}), \tag{$\tilde{\boldsymbol{\mathcal{P}}}$}$$

where $\mathcal{U}$ is an appropriate uncertainty set determined from the data. This phenomenon is remarkable because $\tilde{\boldsymbol{z}}^\star$ is determined by worst-case performance but might have a better on-average performance than $\hat{\boldsymbol{z}}^\star$.

It would be nice if I could combine these two approaches to get good out-of-sample performance regardless of the p/n ratio. The most straightforward idea would be to interpolate between the solutions of ($\hat{\boldsymbol{\mathcal{P}}}$) and ($\tilde{\boldsymbol{\mathcal{P}}}$), but both solutions could be poor, and it is not apparent that their combination would be much better. A better approach is to develop a variant of ($\hat{\boldsymbol{\mathcal{P}}}$) for model parameters that can be estimated confidently, and another variant of ($\tilde{\boldsymbol{\mathcal{P}}}$) for use with parameters that are poorly estimated. A combination of these two could span the spectrum from "classical" solutions (all estimates are accurate) to robust solutions (all estimates are inaccurate), with the proper interpolation being inferred from the data itself. This would capture the strengths of both the classical and robust approaches without being as sensitive as the former or as conservative as the latter.

The first few eigenpairs are easier to estimate than the others (Yu et al., 2015). Methods like the principal component regression leverage on this idea, with a parameter choice, K, that picks the top K eigenvectors to regress y against. However, even when each of the lower eigenvectors is not well estimated, the subspace spanned by these are well estimated. This idea is illustrated for the simple case of regress y against two covariates in Figure 4.1. Consider the few data points that are roughly scattered within a cylindrical region potted with y on the vertical axis and the two covariates on the lower plane. The first principal component of the data, which is also the first eigenvector of $\hat{\Theta}$ is easy to estimate, while the second and third eigenvectors are not. However, space spanned by the second and third eigenvectors is the subspace

Figure 4.1: The Orthogonal Subspace is Well Estimated



Given a few data points, only the first eigenvector of $\hat{\Theta}$ (or equivalently, the principal component of the data) is easy to estimate. The remaining two eigenvectors and eigenvalues are hard to estimate, as shown by many possible orientations and scales (on the zoomed plot). However, the subspace spanned by these two eigenvectors is just the orthogonal plane to the first eigenvector, and hence the subspace is well-estimated.

that is orthogonal to the first eigenvector and is hence easy to estimate.

This motivates splitting the eigenpairs of $\hat{\Theta}$ into the top eigenpairs and the remaining eigenpairs and handling them separately. Given the well-estimated eigenpairs, I use a variant of $\hat{\boldsymbol{\mathcal{P}}}$ to seek a solution belonging to the span of the top eigenvectors. Similarly, a variant of $\tilde{\boldsymbol{\mathcal{P}}}$ will be used to search for a solution belonging to the span of the remaining poorly-estimated eigenvectors. I show that this robust solution is insensitive to the specific orientation of these eigenvectors or their associated eigenvalues. Instead, it only depends on the subspace spanned by these eigenvectors. That is to say, the robust solution only uses information about their support. This is precisely the setting

in which robust optimization excels. The right split of the eigenpairs is chosen from the data so that the combined solution is the best possible.

This approach effectively bridges classical and robust solutions to the regression problem. However, it has one critical pitfall. It ignores the fact that the overarching goal is to predict y. More precisely, splitting based solely on estimation errors in $\hat{\Theta}$ and their effect on the objective $f_\Theta(\boldsymbol{z})$ ignores the constraint $\boldsymbol{z}^T\boldsymbol{e}_1 = 1$, which emphasizes that the first element of $\boldsymbol{z}$ (corresponding to y) is very special. Indeed, the top eigenpairs, even without any estimation errors, could yield poor predictors of y. This mismatch between estimation and prediction only increases for datasets with many covariates. Left unchecked, this mismatch can lead to a useless "classical" solution, which severely limits the benefits of combining the classical and robust solutions.

This leads to the next key component of my approach, the *objective matching* tweak that transforms the data with two specific goals. First, I must ensure that the classical solution based on the top eigenvectors can predict y at least as well as a baseline predictor. This solves the mismatch problem. Second, the transformed data should still keep the pattern of top eigenpairs being estimated better than the bottom eigenpairs. This will allow justifiable combinations of classical and robust solutions, as discussed above. In the context of regression, objective matching simply amounts to transforming the data $(\boldsymbol{y}, X)$ to $(\boldsymbol{y}, X/c)$ for some $c \gg 0$. The corresponding matrix $\hat{\Theta}(c)$ clearly underweights X as compared to y, which ensures that the top eigenvectors of $\hat{\Theta}(c)$ capture the variation in y as against focusing on the variation in $\boldsymbol{x}$. I show

Algorithm 1 Unified Least-Squares Algorithm (with implicit objective matching)

function ULS($X, \boldsymbol{y}, M$) ▷ M is a cross-validation parameter
  Split the data into M (training set, holdout set) pairs as in cross-validation
  for the s^{th} (training set, holdout set) pair **do**
    $\hat{\Phi} \leftarrow \frac{1}{n}\left(X^TX - \frac{X^T\boldsymbol{y}\boldsymbol{y}X^T}{\boldsymbol{y}^T\boldsymbol{y}}\right)$ from training set
    $\Theta_{H_s} \leftarrow \frac{1}{n}\begin{pmatrix} \boldsymbol{y}^T\boldsymbol{y} & -\boldsymbol{y}^TX \\ -X^T\boldsymbol{y} & X^TX \end{pmatrix}$ computed from the holdout set
    Define $g_{\Theta_{H_s}}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \boldsymbol{z}_1^T\Theta_{H_s}\boldsymbol{z}_2$, and $f_{\Theta_{H_s}}(\boldsymbol{z}) = \boldsymbol{z}^T\Theta_{H_s}\boldsymbol{z}$
    $\{(\lambda_i(\hat{\Phi}), \boldsymbol{v}_i(\hat{\Phi}))\} \leftarrow$ eigenvalue-eigenvector pairs of $\hat{\Phi}$
    $\boldsymbol{w}_i \leftarrow \begin{pmatrix} \boldsymbol{v}_i(\hat{\Phi})^TX^T\boldsymbol{y}/\|\boldsymbol{y}\|^2 \\ \boldsymbol{v}_i(\hat{\Phi}) \end{pmatrix}$ for $i = 1, \ldots, p$
    for $k = 1, \ldots, p+1$ **do**
      $\hat{\boldsymbol{z}}^\star_{1:k} \leftarrow \frac{\boldsymbol{e}_1 + \sum_{i=1}^{k-1} \frac{\boldsymbol{v}_i(\hat{\Phi})^TX^T\boldsymbol{y}}{n\lambda_i(\hat{\Phi})}\boldsymbol{w}_i}{1 + \sum_{i=1}^{k-1} \frac{(\boldsymbol{v}_i(\hat{\Phi})^TX^T\boldsymbol{y})^2}{n\lambda_i(\hat{\Phi})\cdot\|\boldsymbol{y}\|^2}}$ ▷ Classical solution from top eigenpairs
      $\tilde{\boldsymbol{z}}^\star_{k+1:p+1} \leftarrow \frac{\sum_{i=k}^{p}(\boldsymbol{v}_i(\hat{\Phi})^TX^T\boldsymbol{y})\boldsymbol{w}_i}{\sum_{i=k}^{p}\frac{(\boldsymbol{v}_i(\hat{\Phi})^TX^T\boldsymbol{y})^2}{\|\boldsymbol{y}\|^2}}$ if $k \le p$ ▷ Robust solution from other eigenpairs
      $C_k(s) \leftarrow f_{\Theta_{H_s}}(\hat{\boldsymbol{z}}^\star_{1:k})$ ▷ Error of classical solution
      $R_k(s) \leftarrow f_{\Theta_{H_s}}(\tilde{\boldsymbol{z}}^\star_{k+1:p+1})$ if $k \le p$ ▷ Error of robust solution
      $CR_k(s) \leftarrow g_{\Theta_{H_s}}(\hat{\boldsymbol{z}}^\star_{1:k}, \tilde{\boldsymbol{z}}^\star_{k+1:p+1})$ if $k \le p$ ▷ Cross-product term
    end for
  end for
  $C_k \leftarrow \frac{\sum_s C_k(s)}{M}$, $R_k \leftarrow \frac{\sum_s R_k(s)}{M}$, $CR_k \leftarrow \frac{\sum_s CR_k(s)}{M}$ for each k
  $a^\star_k \leftarrow \frac{R_k - CR_k}{C_k + R_k - 2\cdot CR_k}$ if $k \le p$; $a^\star_{p+1} = 1$ ▷ Mixing proportions
  $k^\star(s) \leftarrow \arg\max_k \{(a^\star_k)^2 \cdot C_k(s) + (1-a^\star_k)^2 \cdot R_k(s) + 2 \cdot a^\star_k \cdot (1-a^\star_k) \cdot CR_k(s)\}$ for the s^{th} (training set, holdout set) pair
  Calculate $\hat{\boldsymbol{z}}^\star_{1:k}$ and $\tilde{\boldsymbol{z}}^\star_{k+1:p+1}$ as in Steps 10 and 11 using the entire dataset
  Calculate $\hat{\boldsymbol{z}}_k = a^\star_k \cdot \hat{\boldsymbol{z}}^\star_{1:k} + (1-a^\star_k)\cdot\tilde{\boldsymbol{z}}^\star_{k+1:p+1}$ for each k ▷ Combined solution for each k
  $\boldsymbol{z}^{ULS} \leftarrow \frac{\sum_s \hat{\boldsymbol{z}}_{k^\star(s)}}{M}$
  return $\boldsymbol{z}^{ULS}$
end function

that the pattern of estimation errors in $\hat{\Theta}(c)$ is broadly similar to that of $\hat{\Theta}$, with the top eigenpairs being estimated better than the remaining eigenpairs. Finally, I show that these classical and robust solutions have well-defined limit points for large c. This lets us derive an algorithm that yields the results of objective matching while avoiding numerical instabilities associated with large c (Algorithm 1).

4.3 The details

I will now present detailed explanations for objective matching, the splitting of eigenpairs, and the combination of solutions generated from the splits. I will assume that $\boldsymbol{y} \neq \mathbf{0}$ (so $Ey^2 \neq 0$), otherwise one can just predict $y = 0$ always.

4.3.1 Objective Matching

Well-estimated eigenpairs of Θ do not necessarily imply that they are useful for prediction. Indeed, the opposite is likely to be true. Since there are p rows/columns for $\boldsymbol{x}$ but only one for y in the matrix Θ, the top eigenvectors are more likely to capture the variation in $\boldsymbol{x}$ than in y. Formally, let $\lambda_1(\Theta) \geq \lambda_2(\Theta) \geq \ldots \geq \lambda_{p+1}(\Theta)$ be the eigenvalues of Θ, and with corresponding eigenvectors $\boldsymbol{v}_i(\Theta)$. Then, $\boldsymbol{w} := (\boldsymbol{v}_1(\Theta)^T \boldsymbol{e}_1)^{-1} \boldsymbol{v}_1(\Theta)$ is a feasible point[1] for $(\hat{\boldsymbol{\mathcal{P}}})$, and I have the following upper-bound on the out-of-sample R^2 of $\boldsymbol{w}$ (and hence a lower bound on the objective $f_\Theta(\boldsymbol{w})$).

[1]It is easy to show that $\boldsymbol{v}_1(\Theta)^T \boldsymbol{e}_1 \neq 0$ when $Ey^2 \neq 0$.

Theorem 4.3.1 (Upper Bound of $OR^2(\frac{\boldsymbol{v}_1(\Theta)}{\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1})$)**.**

$$OR^2\left(\frac{\boldsymbol{v}_1(\Theta)}{\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1}\right) \leq \min\left(0,\ 1-2^{-3}\frac{\lambda_1(\Theta)\big(\lambda_1(\Theta)-\lambda_2(\Theta)\big)^2 Ey^2}{\big((Ey^2)^2+(Ey\boldsymbol{x}^T)(E\boldsymbol{x}y)\big)^2}\right).$$

Observing that $OR^2(\boldsymbol{e}_1) = 0$, the above theorem states that a feasible solution constructed only from the first eigenvector, even if it is perfectly estimated, can be no better than the baseline solution $\boldsymbol{e}_1$. In fact, it can be much worse, as shown by the following simulation based on the `Diabetes1` dataset. There are $n = 442$ observations regarding $p = 10$ covariates. For $k = 1, \ldots, p+1$, I construct the optimal solution to Eq. 4.1 under the additional constraint that the solution must be a linear combination of the top k eigenvectors. Figure 4.2 left shows the out-of-sample R^2 of these solutions as a function of k, compared against the baseline solution $\boldsymbol{e}_1$. I see that $OR^2((\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1)^{-1}\boldsymbol{v}_1(\Theta)) = -40$, which is much worse than the baseline. Indeed, the best solution using 8 perfectly estimated eigenpairs is still worse than the naive baseline. Thus, even without estimation errors, the top eigenpairs need not have good predictive power.

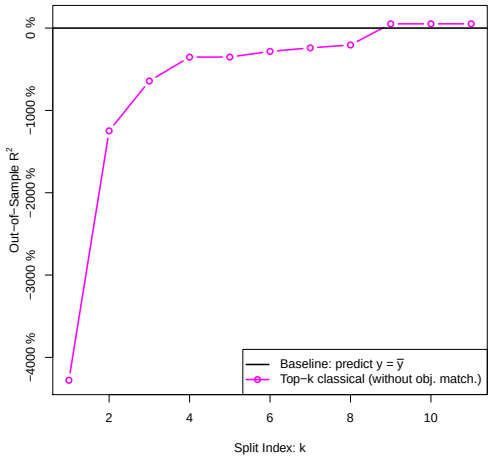
At first sight, I appear to have hit a dead end. However, I do have an extra degree of freedom that can be exploited.

Theorem 4.3.2 (Possible ways to Change Θ)**.** *For any invertible $M \in \mathbb{R}^{p\times p}$ such that $M_{11} = 1$, and $M_{i1} = M_{1i} = 0$ for all $i = 2, \ldots, p$, I have*

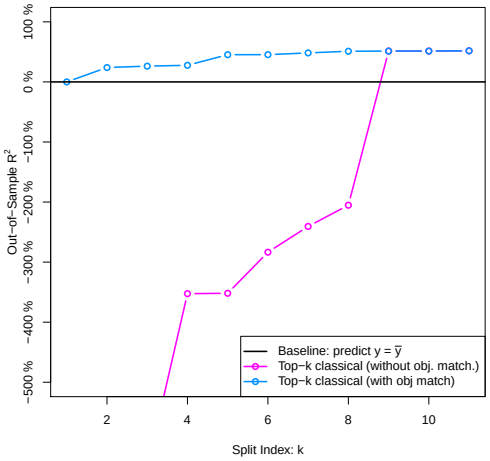
$$\min_{\boldsymbol{z}|\boldsymbol{z}^T\boldsymbol{e}_1=1} \boldsymbol{z}^T\Theta\boldsymbol{z} = \min_{\boldsymbol{z}|\boldsymbol{z}^T\boldsymbol{e}_1=1} \boldsymbol{z}^T M^T\Theta M\boldsymbol{z}.$$

Figure 4.2: Out-of-sample R^2 of the classical solution using the top k eigenpairs, as a function of k

OR^2 without objective matching



OR^2 with objective matching



(Left) Without objective matching, the top-k solution is worse than baseline for $k \leq 8$. (Right) With objective matching, it is always at least as good as baseline. The two yield the same solution (the ordinary least squares solution) when $k = p + 1 = 11$.

Thus, I can choose to minimize $f_{M^T\Theta M}(\boldsymbol{z})$ instead of $f_\Theta(\boldsymbol{z})$, under the same constraints on $\boldsymbol{z}$. The choice of M should satisfy two desiderata. First, it should solve the mismatch problem by aligning estimation and prediction. The out-of-sample R^2 of the easily-estimated top eigenvector is upper-bounded by 0, as seen from Theorem 4.3.1 with Θ replaced by $M^T\Theta M$. I should select an M that achieves this upper bound. Second, the robust solution constructed from the bottom eigenpairs should achieve an out-of-sample R^2 of the same order as the solution constructed from the top eigenpairs. Only then will the combination of the classical and robust solutions yield a significant benefit.

I can achieve these desiderata by a diagonal matrix $M(c) = diag(1, 1/c, 1/c, \ldots, 1/c)$ for some constant $c \gg 0$. Since $\Theta(c) := M(c)^T\Theta M(c) = E\begin{pmatrix} y^2 & -\frac{y\boldsymbol{x}^T}{c} \\ -\frac{\boldsymbol{x}y}{c} & \frac{\boldsymbol{x}\boldsymbol{x}^T}{c^2} \end{pmatrix}$, this is equivalent to replacing $(y, \boldsymbol{x}^T)$ by $(y, c^{-1} \cdot \boldsymbol{x}^T)$. This suggests that for a large c, the first eigenvector $\boldsymbol{v}_1(c)$ of $\Theta(c)$ is nearly aligned with the $\boldsymbol{e}_1$ direction, which is also the baseline solution. Hence, in contrast to the first eigenvector of Θ (Theorem 4.3.1), a solution constructed from the first eigenvector of $\Theta(c)$ achieves at least the baseline out-of-sample R^2 of 0.

Theorem 4.3.3 (Best Starting Point)**.** *I have*

$$\lambda_1(c) = Ey^2 + O(1/c), \quad \boldsymbol{v}_1(c) = \boldsymbol{e}_1 + O(1/c), \quad OR^2\left(\frac{\boldsymbol{v}_1(c)}{\boldsymbol{v}_1(c)^T\boldsymbol{e}_1}\right) = O(1/c),$$

where $(\lambda_i(c), \boldsymbol{v}_i(c))$ *represent eigenpairs of* $\Theta(c)$*, with* $\lambda_1(c) \geq \lambda_2(c) \geq \ldots \geq \lambda_{p+1}(c)$.

The solution using more than one top eigenpairs will only improve upon the baseline. Figure 4.2 right confirms this in simulation: the out-of-sample

R^2 now starts from 0 (for split index $k = 1$), which is the best starting point from Theorem 4.3.1, and keeps improving with increasing k. This resolves the mismatch between good estimation of top eigenpairs of Θ and their poor prediction accuracy. With my choice of $M(c)$, I expect the top eigenpairs of $\Theta(c)$ to be predictive of y, which is the overall objective of regression. This motivates the name "objective matching."

Our second desired property was that the bottom eigenpairs should also have predictive power. Otherwise, the robust solution constructed from them will be useless and should be ignored. It may seem that this property is unlikely; any solution vector $\boldsymbol{z}$ must satisfy $\boldsymbol{z}^T\boldsymbol{e}_1 = 1$, but the eigenvectors $\boldsymbol{v}_i(c)$ for $i \in [2, p+1]$ are nearly orthogonal to $\boldsymbol{e}_1$ (from Theorem 4.3.3 and the orthogonality of eigenvectors). The following theorem guarantees that these eigenvectors still have predictive power.

Theorem 4.3.4 (All Eigenvectors are Useful)**.** *For all $i = 2, \ldots, p+1$, I have*

$$OR^2\left(\frac{\boldsymbol{v}_i(c)}{\boldsymbol{v}_i(c)^T\boldsymbol{e}_1}\right) = O(1).$$

Thus, the out-of-sample R^2 for a solution vector constructed from any one eigenvector does not decay to 0 for large c. This is why the solutions obtained for $k > 1$ in Figure 4.2 right improved upon the baseline. It also suggests that the robust solution, constructed from the bottom $p - k + 1$ eigenvectors, will achieve an out-of-sample R^2 of $O(1)$ regardless of the values of c or the split index k.

4.3.2 Splitting eigenpairs

Objective matching gives us a matrix $\Theta(c)$ whose top eigenvector corresponds to the baseline, and every other eigenvector offers non-trivial predictive power in terms of out-of-sample R^2. It can be argued, as in section 4.2, that because the top eigenpairs of $\hat{\Theta}$ are good estimations of their counterparts of Θ, the classical solution constructed from the former is a close approximation to the one built based on the latter. However, this argument regarding the ability to estimate eigenvectors relies on the presence of large eigengaps, or differences, between successive top eigenvalues of Θ (Yu et al., 2015). This no longer holds for $\Theta(c)$: while $\lambda_1(c) = Ey^2 + O(1/c)$, all other eigenvalue are close to zero, and the eigengap between $\lambda_i(c)$ and $\lambda_{i+1}(c)$ is negligible for all $i > 1$. This necessitates a very different theoretical justification for splitting the eigenpairs of $\Theta(c)$. I will provide this justification next and then discuss how I devise two solutions from the two sets of eigenpairs.

Theoretical justification for splitting eigenpairs of $\Theta(c)$. I will now prove that the eigenvalues $\lambda_2(c), \ldots, \lambda_{p+1}(c)$, appropriately normalized, are close to the eigenvalues of the matrix $\Phi = E\boldsymbol{x}\boldsymbol{x}^T - \frac{(E\boldsymbol{x}y)(Ey\boldsymbol{x}^T)}{Ey^2}$, which is independent of c. The same holds for the corresponding eigenvectors as well.

Theorem 4.3.5 (The Connection between $\Theta(c)$ and Φ)**.** *For all $i = 2, \ldots, p+1$, I have*

$$c^2\lambda_i(c) = \lambda_{i-1}(\Phi) + O(1/c). \tag{4.4}$$

Moreover, if I assume that $\lambda_i(\Phi)$ is strictly monotone[2]*, then I have*

$$\left\| \boldsymbol{v}_i(c) - \begin{pmatrix} 0 \\ \boldsymbol{v}_{i-1}(\Phi) \end{pmatrix} \right\|_2 = O(1/c).$$

Similar statements link the empirical eigenpairs $(\hat{\lambda}_i(c), \hat{\boldsymbol{v}}_i(c))$ of $\hat{\Theta}(c)$ to the eigenpairs of $\hat{\Phi} = \frac{1}{n}\left(\sum_{j=1}^n \boldsymbol{x}_j \boldsymbol{x}_j^T - \frac{\left(\sum_{\ell=1}^n y_i \boldsymbol{x}_i\right)\left(\sum_{\ell=1}^n y_i \boldsymbol{x}_i^T\right)}{\sum_{\ell=1}^n y_i^2}\right)$.

Corollary 1. Under the conditions of Theorem 4.3.5, for all $i = 2, \ldots, p+1$, I have

$$\frac{|\lambda_i(c) - \hat{\lambda}_i(c)|}{\lambda_i(c)} = \frac{|\lambda_{i-1}(\Phi) - \lambda_{i-1}(\hat{\Phi})|}{\lambda_{i-1}(\Phi)} + O(1/c),$$
$$\|\boldsymbol{v}_i(c) - \hat{\boldsymbol{v}}_i(c)\| = \|\boldsymbol{v}_{i-1}(\Phi) - \boldsymbol{v}_{i-1}(\hat{\Phi})\| + O(1/c).$$

Corollary 1 shows that the relative estimation errors of eigenpairs of $\Theta(c)$ are equivalent to those for Φ. Thus, the estimability of eigenpairs, and hence the location of the best split k, are driven by differences between consecutive eigenvalues (i.e., eigengaps) in Φ, instead of the eigengaps in Θ. The following theorem characterizes the eigenvalues of Φ.

Theorem 4.3.6 (Eigenvalues of Φ)**.** *For any $i = 1, \ldots, p$, I have*

$$\lambda_i(\Theta) \geq \lambda_i(E\boldsymbol{x}\boldsymbol{x}^T) \geq \lambda_i(\Phi) \geq \lambda_{i+1}(\Theta) \geq \lambda_{i+1}(E\boldsymbol{x}\boldsymbol{x}^T),$$

[2]This assumption is just for the simplicity of the result. If there exists $1 \leq r < s \leq p$ that $\lambda_r(\Phi) = \lambda_{r+1}(\Phi) = \cdots = \lambda_s(\Phi)$, the result will be the about the matrix $(\boldsymbol{v}_r(\Phi), \ldots, \boldsymbol{v}_s(\Phi))$ instead of individual eigenvectors.

where $\lambda_{p+1}(E\boldsymbol{x}\boldsymbol{x}^T)$ is taken to be 0. Moreover,

$$\sum_{i=1}^{p} \big(\lambda_i(\Phi) - \lambda_{i+1}(E\boldsymbol{x}\boldsymbol{x}^T)\big) = \lambda_1(E\boldsymbol{x}\boldsymbol{x}^T) - \frac{(Ey\boldsymbol{x}^T)(E\boldsymbol{x}y)}{Ey^2} \geq 0.$$

Thus, the eigenvalues of Φ are interlaced between those of Θ and $E\boldsymbol{x}\boldsymbol{x}^T$, with a bounded total deviation from $\{\lambda_{i+1}(E\boldsymbol{x}\boldsymbol{x}^T) \mid i = 1, \ldots, p\}$. If y is uncorrelated with $\boldsymbol{x}$, namely $Ey\boldsymbol{x} = 0$, then $\lambda_i(\Phi) = \lambda_i(E\boldsymbol{x}\boldsymbol{x}^T)$. At the other extreme, if y is maximally correlated with $\boldsymbol{x}$, namely y is along the direction of the first eigenvector of $E\boldsymbol{x}\boldsymbol{x}^T$, then $\lambda_i(\Phi) = \lambda_{i+1}(E\boldsymbol{x}\boldsymbol{x}^T)$. For both these extremes, the eigengaps of Φ and $E\boldsymbol{x}\boldsymbol{x}^T$ are provably related. This suggests that, even in general, the pattern of eigengaps in Φ is similar to that of Θ or $E\boldsymbol{x}\boldsymbol{x}^T$. Hence, it is still reasonable to split eigenpairs into well-estimated and poorly estimated parts.

The top eigenpairs. Let S_k represent the subspace spanned by the top k empirical eigenvectors $\{\hat{\boldsymbol{v}}_1(c), \ldots, \hat{\boldsymbol{v}}_k(c)\}$, and N_k represent the subspace spanned by the remaining eigenvectors. Then, for any k, I can solve $(\hat{\boldsymbol{\mathcal{P}}})$ under the restriction that the solution $\boldsymbol{z}$ should be a linear combination of the top k eigenvectors.

$$\min_{\substack{\boldsymbol{z}^T\boldsymbol{e}_1 = 1,\\ \boldsymbol{z} \in S_k}} f_{\hat{\Theta}(c)}(\boldsymbol{z}), \quad \text{whose solution is} \qquad \hat{\boldsymbol{z}}^\star_{1:k}(c) = \frac{\sum_{i=1}^{k} \frac{\hat{\boldsymbol{v}}_i(c)^T\boldsymbol{e}_1}{\hat{\lambda}_i(c)} \hat{\boldsymbol{v}}_i(c)}{\sum_{i=1}^{k} \frac{(\hat{\boldsymbol{v}}_i(c)^T\boldsymbol{e}_1)^2}{\hat{\lambda}_i(c)}}. \tag{4.5}$$

Observe that the solution $\hat{\boldsymbol{z}}^\star_{1:k}(c)$ is a function of only the top k eigenpairs. When k is properly chosen, all k eigenpairs are well-estimated, and $\hat{\boldsymbol{z}}^\star_{1:k}$ is

reliable, i.e., it's in-sample R^2 is close to its out-of-sample R^2. This justifies solving the "classical" optimization ($\hat{\boldsymbol{\mathcal{P}}}$) under the restriction $\boldsymbol{z} \in S_k$.

The remaining eigenpairs. In contrast to the top eigenpairs, the remaining eigenpairs are likely to be poorly estimated. Hence, the solution of a "classical" optimization restricted to $\boldsymbol{z} \in N_k$ is not reliable; its out-of-sample R^2 may be much worse than its in-sample R^2. Instead, observe that the subspace N_k spanned by these remaining eigenvectors is, in fact, well-estimated, since it is the subspace that is orthogonal to the well-estimated S_k. This motivates using robust optimization to find a solution $\boldsymbol{z} \in N_k$ that is agnostic to the bottom eigenpairs $\{(\hat{\lambda}_i(c), \hat{\boldsymbol{v}}_i(c)) \mid i > k\}$ but respects the subspace N_k. To achieve this, I propose the following uncertainty set:

$$\mathcal{U}(N_k) = \{\Theta(c) \mid \Theta(c) \mid_{N_k} \preceq m I_{p-k+1}\}, \tag{4.6}$$

where $\Theta(c) \mid_{N_k}$ is the projection of $\Theta(c)$ on N_k and m is a constant. With this uncertainty set, I propose the following robust optimization solution derived from the bottom eigenpairs:

$$\min_{\boldsymbol{z} \in N_k \cap \mathcal{Z}} \max_{\Theta(c) \in \mathcal{U}(N_k)} \boldsymbol{z}^T \Theta(c) \boldsymbol{z}, \quad \text{whose solution is} \qquad \tilde{\boldsymbol{z}}^{\star}_{k+1:p+1}(c) = \frac{P_{N_k} \boldsymbol{e}_1}{\boldsymbol{e}_1^T P_{N_k} \boldsymbol{e}_1}, \tag{4.7}$$

where P_{N_k} is the projection matrix on the subspace N_k. Note that the solution $\tilde{\boldsymbol{z}}^{\star}_{k+1:p+1}(c)$ is independent of m. That is to say, given k, the size of the uncertainty set does not matter. Moreover, the solution remains the same for any rotation of the eigenvectors from N_k. In other words, the solution only uses

information about the subspace N_k, and not about the bottom eigenvalues or the orientations of these eigenvectors.

$\hat{\boldsymbol{z}}^\star_{1:k}(c)$ and $\tilde{\boldsymbol{z}}^\star_{k+1:p+1}(c)$ are solutions under $\hat{\Theta}(c)$. Using Theorem 4.3.2, I can easily convert them to the corresponding solutions under $\hat{\Theta}$ by multiplying them with $M(c) = diag(1, 1/c, 1/c, \ldots, 1/c)$. I now show that these solutions are well-defined as $c \to \infty$.

Theorem 4.3.7 (Solutions converge for large c)**.** *For any k, as $c \to \infty$, the classical and robust solutions converge to the following:*

$$\hat{\boldsymbol{z}}^\star_{1:k} := \lim_{c\to\infty} M(c)\hat{\boldsymbol{z}}^\star_{1:k}(c) = \frac{\boldsymbol{e}_1 + \sum_{i=1}^{k-1} \frac{\boldsymbol{v}_i(\hat{\Phi})^T X^T \boldsymbol{y}}{n\lambda_i(\hat{\Phi})} \boldsymbol{w}_i}{1 + \sum_{i=1}^{k-1} \frac{\left(\boldsymbol{v}_i(\hat{\Phi})^T X^T \boldsymbol{y}\right)^2}{n\lambda_i(\hat{\Phi}) \cdot \boldsymbol{y}^T\boldsymbol{y}}} \quad \textit{for } k = 1, \ldots, p+1$$

$$\tilde{\boldsymbol{z}}^\star_{k+1:p+1} := \lim_{c\to\infty} M(c)\tilde{\boldsymbol{z}}^\star_{k+1:p+1}(c) = \frac{\sum_{i=k}^{p} (\boldsymbol{v}_i(\hat{\Phi})^T X^T \boldsymbol{y}) \boldsymbol{w}_i}{\sum_{i=k}^{p} \frac{(\boldsymbol{v}_i(\hat{\Phi})^T X^T \boldsymbol{y})^2}{\boldsymbol{y}^T\boldsymbol{y}}} \qquad \textit{for } k = 1, \ldots, p$$

$$\textit{where } \boldsymbol{w}_i = \begin{pmatrix} \boldsymbol{v}_i(\hat{\Phi})^T X^T \boldsymbol{y} / \boldsymbol{y}^T\boldsymbol{y} \\ \boldsymbol{v}_i(\hat{\Phi}) \end{pmatrix} \qquad \textit{for } i = 1, \ldots, p$$

4.3.3 Combining solutions from the two splits

For a particular k, I can get a classical solution $\hat{\boldsymbol{z}}^\star_{1:k}$ and a robust solution $\tilde{\boldsymbol{z}}^\star_{k+1:p+1}$. Now, I seek a combined solution of the form $\hat{\boldsymbol{z}}^{ULS}_k := a_k \cdot \hat{\boldsymbol{z}}^\star_{1:k} + (1 - a_k) \cdot \tilde{\boldsymbol{z}}^\star_{k+1:p+1}$. The following $a^\star_k$ minimizes $f_\Theta(\hat{\boldsymbol{z}}^{ULS}_k)$,

$$a^\star_k = \frac{f_\Theta(\tilde{\boldsymbol{z}}^\star_{k+1:p+1}) - g_\Theta(\hat{\boldsymbol{z}}^\star_{1:k}, \tilde{\boldsymbol{z}}^\star_{k+1:p+1})}{f_\Theta(\hat{\boldsymbol{z}}^\star_{1:k}) + f_\Theta(\tilde{\boldsymbol{z}}^\star_{k+1:p+1}) - 2g_\Theta(\hat{\boldsymbol{z}}^\star_{1:k}, \tilde{\boldsymbol{z}}^\star_{k+1:p+1})}, \tag{4.8}$$

$$\text{where } g_\Theta(\boldsymbol{z}_1, \boldsymbol{z}_2) = \boldsymbol{z}_1^T \Theta \boldsymbol{z}_2, \tag{4.9}$$

$$f_\Theta(\boldsymbol{z}) = \boldsymbol{z}^T \Theta \boldsymbol{z}.$$

4.3.4 Selection of the split index k

There are two problems left. First, in Eq. 4.8, $f_\Theta(.)$ and $g_\Theta(.)$ cannot be directly computed, since I only have access to $\hat{\Theta}$. Second, I need to pick k. I solve both of these problems via cross-validation. I first split the training data into M parts. I then group $M-1$ of these parts into a proto-training data and set the last part aside as a holdout set. Then, for each k, I construct $\hat{\boldsymbol{z}}^\star_{1:k}$ and $\tilde{\boldsymbol{z}}^\star_{k+1:p+1}$ from the proto-training data. I compute the corresponding $a^\star_k$ using the holdout sets as ground truth, i.e., I compute Eq. 4.8 with Θ replaced by the $\Theta_H := \hat{\Theta}$ computed over all the holdout sets. This gives a solution $\boldsymbol{z}^{ULS}_k$. The best $k^\star$ for one holdout set is then chosen by picking the k with the smallest $f_{\Theta_{H_s}}(\boldsymbol{z}^{ULS}_k)$, with the Θ_{H_s} from the holdout set standing in for Θ. This process creates a probabilistic view of $k^\star$ which is consistent to the gradual change from well-estimated eigenpairs to the not-well-estimated. Finally, the solution vectors $\boldsymbol{z}^{ULS}_{k^\star}$ obtained from these iterations are averaged, and that is returned as the final answer.

I use $M = 3$ for all methods across this paper. The results are similar for $M = 5$ and $M = 10$. Because there are two parameters a and k to estimate for ULS, I find that using 10 permutations, namely introducing a total of 30 groups, tends to increase the performance. I also tried 10 permutations for other methods where only one parameter needs to be estimated. The corresponding improvement is much smaller. Because other methods traditionally do not involve more permutations, I decide to stick to the conventional version.

4.4 Relation to Other Methods

In this section, I will describe the relations between ULS and ordinary least squares regression (OLS), principal components regression (PCR), partial least squares regression (PLS), and robust optimization. I will also draw parallels between regression and portfolio optimization, and the relations between ULS and existing methods for portfolio optimization.

4.4.1 Relation to OLS

It is easy to show

$$\boldsymbol{b}_{OLS} = \frac{\hat{\Phi}^{-1} X^T \boldsymbol{y}}{n + \frac{\boldsymbol{y}^T X \hat{\Phi}^{-1} X^T \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{y}}}.$$

This is precisely the solution if $k = p + 1$ in Theorem 4.3.7 because $\boldsymbol{z} = \begin{pmatrix} 1 \\ \boldsymbol{b} \end{pmatrix}$.

4.4.2 Relation to PCR and PLS

PCR predicts $\boldsymbol{y}$ using the top k eigenvectors of the matrix $X^T X$. The next theorem shows that these eigenvectors are close to the top eigenvectors of $\hat{\Phi}$.

Theorem 4.4.1 (Eigenvectors of $X^T X$ and $\hat{\Phi}$)**.** *For all $i = 1, \ldots, p$, I have*

$$\left\| \boldsymbol{v}_i(X^T X) - \boldsymbol{v}_i(\hat{\Phi}) \right\|_2 \leq \frac{2^{3/2} \frac{\|X^T \boldsymbol{y}\|_2^2}{\boldsymbol{y}^T \boldsymbol{y}}}{\min(\lambda_{i-1}(X^T X) - \lambda_i(X^T X), \lambda_i(X^T X) - \lambda_{i+1}(X^T X))},$$

where $(\lambda_i, \boldsymbol{v}_i)$ represent eigenvalue-eigenvector pairs in descending order of eigenvalues, and I assume that $X^T X$ has no repeated eigenvalues[3].

[3]Once again, this assumption is just for the simplicity of the result.

If the top k eigenvalues of X^TX are separated by large eigengaps, the denominator of the right side is large for each $i \leq k$. Then the top eigenvectors $\boldsymbol{v}_i(X^TX)$ and $\boldsymbol{v}_i(\hat{\Phi})$ are close to each other, and so are the the subspace spanned by the top k eigenvectors. However, PCR's solution is not necessarily close to that of ULS, because PCR ignores the bottom eigenvectors completely. This leaves PCR vulnerable to the mismatch problem; the top eigenvectors of X^TX could be well-estimated, but still not predictive of y. Indeed, this was observed by Jolliffe (1982). Thanks to objective matching, ULS avoids this problem.

PLS predicts $\boldsymbol{y}$ using the k-dimensional subspace spanned by the vectors $\left\{(X^TX)^{i-1}X^T\boldsymbol{y}, i \leq k\right\}$. In contrast to PCR, this subspace depends on both $\boldsymbol{y}$ and X. This can be interpreted as an ad hoc fix for the mismatch problem. Still, like PCR, PLS throws away the orthogonal subspace which is well estimated as a space. The ULS method utilizes the orthogonal subspace conservatively using robust optimization. This is particularly useful in low-data settings, where the orthogonal subspace is of high dimension.

4.4.3 Connection with Robust Optimization

Robust optimization approaches typically assume uncertainty sets that are mathematically convenient to analyze. In the LS context, the uncertainty is about the matrix Θ. A common approach for such covariance matrices is to use the uncertainty set $\mathcal{U}(\Theta) = \{\Theta \mid \Theta \preceq h\hat{\Theta}\}$ for some constant $h \geq 1$ (Delage and Ye, 2010). However, this is easily shown to yield precisely the OLS solution

for all h. Clearly, this uncertainty set is not useful.

The ULS method formally defines an uncertainty set only for the poorly estimated bottom eigenpairs of $\Theta(c)$. This can nonetheless be interpreted as the following uncertainty set on $\Theta(c)$: $\mathcal{U}(\Theta(c)) = \{\Theta(c) \mid \Theta(c)|_{S_k} \preceq h\hat{\Theta}(c)|_{S_k}, \Theta(c)|_{N_k} \preceq mI_{p-k+1}\}$, where S_k and N_k are the subspaces corresponding to the top k and bottom $p-k+1$ eigenvectors of $\hat{\Theta}(c)$, $R|_{S_k}$ and $R|_{N_k}$ denote the restriction of R to these subspaces, and $h \geq 1$ and $m > 0$ are constants. However, this interpretation cannot match the full flexibility of the ULS method. ULS picks k via cross-validation, and indeed, the final result is an average of solutions for several values of k. This effect is difficult to achieve under one uncertainty set based on one value of k. Also, the ULS solution is a combination of $\hat{\boldsymbol{z}}^\star_{1:k}$ from S_k and $\tilde{\boldsymbol{z}}^\star_{k+1:p+1}$ from N_k, with the combination level determined by cross-validation to maximize average-case performance. The uncertainty-set understanding should choose the combination level based on h and m. To be consistent with the objective of robust optimization, these quantities should be set to optimize the worst-case performance in the cross-validation. Focusing on the worst case can lead to solutions that are too conservative. Finally, it is difficult to justify why the uncertainty set should be this specific form.

Thus, I believe that the reasons for the strong performance of the ULS method lie in requiring robustness only where it is needed, i.e., for the poorly estimated eigenpairs. By using the well-estimated eigenpairs directly, and combining the classical and robust solutions based on average-case performance,

ULS avoids the trap of being too conservative.

4.4.4 Relation to Portfolio Optimization

The minimum-variance portfolio is a combination of p assets that has the least risk (as defined by variance) among all possible portfolios:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}^T \Sigma \boldsymbol{w} \\ \text{subject to} \quad & \boldsymbol{w}^T \mathbf{1} = 1, \end{aligned} \tag{4.10}$$

where $\Sigma \in \mathbb{R}^{p \times p}$ is the covariance matrix of asset returns, and $\mathbf{1}$ is a vector with all elements being 1. The constraint specifies that the portfolio must invest all available wealth. Here, both "long" and "short" positions are allowed (i.e., the components of $\boldsymbol{w}$ can be positive or negative). This optimization encapsulates a basic problem regarding risk-minimization that has applications far beyond finance, and so it has been widely studied (Jagannathan and Ma, 2003; DeMiguel et al., 2009a; Brodie et al., 2009; Fan et al., 2012; Zhao et al., 2019a).

The above optimization shares obvious similarities with the LS problem ($\hat{\boldsymbol{\mathcal{P}}}$). In fact, it can be formally cast as a LS problem:

$$\min_{z^T \boldsymbol{e}_1 = 1} \boldsymbol{z}^T (F^T \Sigma F) \boldsymbol{z}, \quad \text{where} \quad F = \left(\begin{array}{c|ccc} 1/p & -1 & \cdots & -1 \\ \hline 1/p & & & \\ \vdots & & I_{p-1} & \\ 1/p & & & \end{array} \right) \tag{4.11}$$

with $\boldsymbol{w} = F\boldsymbol{z}$ being the link between the two formulations. Hence, the $\boldsymbol{z} = \boldsymbol{e}_1$ baseline solution yields the equal-weighted portfolio, which is a robust portfolio that invests equally in each asset. The goal is to find the optimal deviations

from this baseline portfolio[4]. In practice, Σ is unknown, and only the estimated covariance $\hat{\Sigma}$ is available.

Given the close link between the minimum-variance and LS problems, it is not surprising that many of the proposed methods share some similarities. These include the norm-based penalties as in ridge (Lauprête, 2001; DeMiguel et al., 2009a) and LASSO regression (Welsch and Zhou, 2007; Brodie et al., 2009; Fan et al., 2012). Motivated by the observation that the top eigenpairs of the sample covariance matrix are well estimated, DeMiguel et al. (2009a) proposed a PCA-based portfolio that is constructed using only the top eigenpairs. Zhao et al. (2019a) try to use other eigenpairs as well by constructing a conservative portfolio by computing the portfolio variance bounds. They then improve the PCA-based portfolio by combining it with this conservative portfolio.

Here is the critical issue with the implications of the method in Zhao et al. (2019a). When minimal data are available, the corresponding bound will be so large that the combined portfolio is essentially the PCA-based portfolio. This is mainly because it puts little faith in the conservative portfolio. That is to say, the conservative portfolio is abandoned when it is needed the most. Apart from this, their method cannot adjust itself to the objective of interest. This is unnecessary in portfolio optimization since there is no special stock a priori. However, in the LS problems there is the particular dependent variable

[4]Other reductions to the LS problem also exist, but do not have this interpretation.

around which the measure of success revolves. Though not that meaningful, the direct use of Zhao et al. (2019a) for the LS problem hence results in disastrous performance.

4.5 Simulation Experiments

In this section, I will compare ULS against other methods on simulations based on the `Diabetes1` dataset. This dataset has 442 observations and $p = 10$ covariates. I first estimate the Θ matrix using all observations. Then I simulate data by generating $(y_i, \boldsymbol{x}_i)$ from a multivariate normal distribution with zero mean and covariance Θ. Since the goal is to investigate the accuracy of the LS algorithms in a limited-data setting, I simulate $n = 20 = 2p$ observations as the training data. I run all methods on the simulated data and then calculate the out-of-sample R^2 for each method using the true Θ. The results are aggregated over 1000 repetitions of this process.

I compare ULS against ordinary least-squares regression (OLS), principal components regression (PCR), partial least-squares regression (PLS), ridge regression (L2), LASSO regression (L1), and non-linear shrinkage (NonLin) (Hotelling, 1957; Wold, 1966; Tikhonov, 1943; Tibshirani, 1996; Ledoit and Wolf, 2017). PCR, PLS, L2, and L1 traditionally transform the data to have mean 0 and unit variance before computing their solutions, and then apply the inverse transform when making predictions (Friedman et al., 2001). I use this standardization step for ULS too.

The optimal (OPT) can only be achieved by knowing Θ. OPT has the

Table 4.1: Statistics Of Out-of-Sample R^2

Methods	ULS	L2	PLS	PCR	L1	NonLin	OLS
Mean	0.361	0.308	0.283	0.267	0.267	0.239	-0.087
SD	0.102	0.174	0.197	0.168	0.197	0.229	0.441

best possible out-of-sample R^2, 0.528. This is unachievable but serves as a bound.

Table 4.3 shows that the ULS method has the best average out-of-sample R^2 among all competing algorithms. ULS is 21% closer to OPT than the next-best algorithm, L2. I also see that OLS performs worse than the baseline that predicts the mean of $\boldsymbol{y}$ resulting in an out-of-sample R^2 of 0. OLS is by definition the "classical" solution that assumes that all eigenpairs are well-estimated. The baseline is precisely the robust solution when the uncertainty set encompasses all eigenpairs. Thus, the "classical" solution, that optimizes the in-sample R^2, actually has lower out-of-sample R^2 than the robust solution that only considers the worst-case.

Another interesting finding is that NonLin performs worse than L2, even though nonlinear shrinkage (NonLin) provides a better estimation than linear shrinkage (L2) (Ledoit and Wolf, 2012). The reason is the mismatch between estimation and prediction: a better estimate of Θ does not guarantee a better predictor. This emphasizes the need for objective matching.

Table 4.3 also shows that ULS also has the lowest variation in out-of-sample R^2 over the 1000 repetitions. This is due to the splitting mechanism of ULS. The well-estimated top eigenpairs are similar in most repetitions, so the

classical solution constructed from them does not vary much. The remaining poorly-estimated eigenpairs do vary a lot, but the robust solution that ULS builds for them depends only on the subspace spanned by these eigenvectors. This again does not vary much. Together, these lead to the observed stability and robustness of ULS.

4.6 Experiments on Real-world Datasets

This section compares ULS against competing methods using real-world datasets. The first set of results are on seven classic regression datasets. Then, I demonstrate an application of ULS to portfolio optimization, by converting the well-known minimum-variance optimization into a regression problem and testing it on ten financial datasets. All datasets are listed in Table 4.2.

Table 4.2: List of Datasets

Classic Regression Datasets	Number of covariates (p)
Diabetes1	10
Community	99
Protein	88
Diabetes2	64
Crime	15
Supernova	10
Prostate	9
Financial Datasets	**Number of assets**
Six Fama and French (1992) portfolios of firms sorted by size and book-to-market	6
Ten industry portfolios representing U.S. stock market	10
Twenty-five Fama and French (1992) portfolios of firms sorted by size and book-to-market	25
Forty-eight industry portfolios representing U.S. stock market	48
One hundred Fama and French (1992) portfolios of firms sorted by size and book-to-market	96

Each financial dataset has an "equal-weighted" and "value-weighted" version, for a total of 10 financial datasets.
For the last financial dataset, there are missing values for four risky assets for an extended period. Thus, I deleted them, leaving 96 of the original 100 assets.
The regression datasets are from (Hahn and Carvalho, 2015; Efron and Hastie, 2016; Dheeru and Karra Taniskidou, 2017). The financial datasets are available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

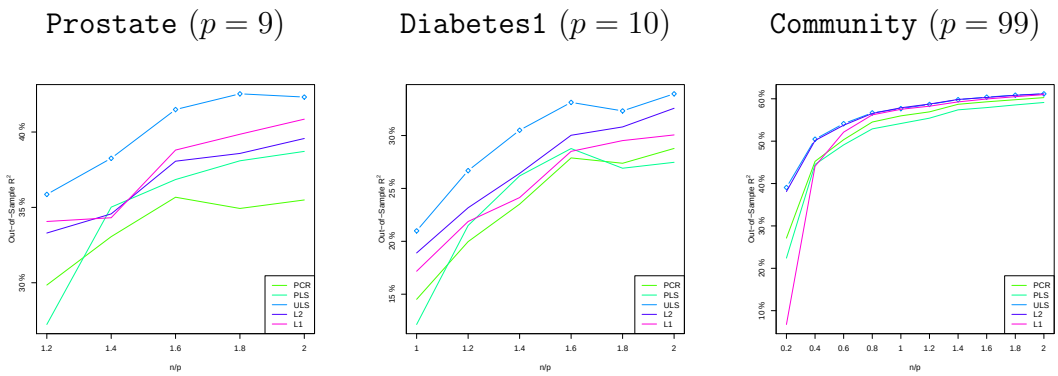
4.6.1 Classic Regression Datasets

Our goal is to study the performance of ULS under varying levels of data insufficiency. So, for each dataset, I construct smaller training sets by varying the n/p ratio from 20% to 200% with 20% increments. For example, for the `Protein` dataset with $p = 88$ covariates, I construct 100 training sets with $n = 18$ randomly chosen data points, another 100 sets with $n = 36$ data points, and so on. I ignore instances with $n < 10$ which is too little data for any method. Each competing algorithm is trained on these subsets with standardization, and then its out-of-sample R^2 is measured on all remaining data points. I report the average out-of-sample R^2 over the 100 repetitions. The list of competing algorithms is the same as in Section 4.5 except for the NonLin method, which sometimes did not yield any answer[5] and hence its performance could not be measured reliably.

Accuracy of ULS. Figure 4.3 shows the results on the `Prostate`, `Diabetes1`, and `Community` datasets, ranging from small to large p. As expected, all methods improve as the n/p ratio increases. ULS clearly dominates in the `Diabetes1` and `Prostate` datasets, and shares the honors with ridge regression (L2) in the `Community` dataset. OLS is not shown in the plots because it performs worse than even the baseline that predicts the mean of $\boldsymbol{y}$ value irrespective of $\boldsymbol{x}$. Results for the other datasets show a similar pattern and are gathered in Figure 4.4.

[5]The `nlshrink` R package occasionally broke down.

Figure 4.3: Out-of-Sample R^2 for three classic regression datasets.

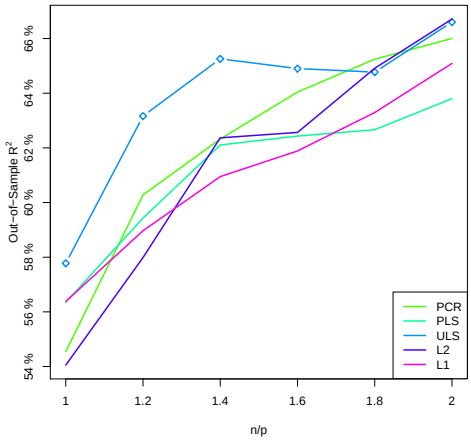


Importance of the robust solution. Observe that ULS always outperforms PCR and PLS. This is because both PCR and PLS throw away the subspace information. However, ULS uses the fact that the subspace spanned by the lower eigenvectors is estimated well, even though the individual eigenpairs are not. The robust solution constructed from this subspace lets ULS consistently improve upon PCR and PLS.

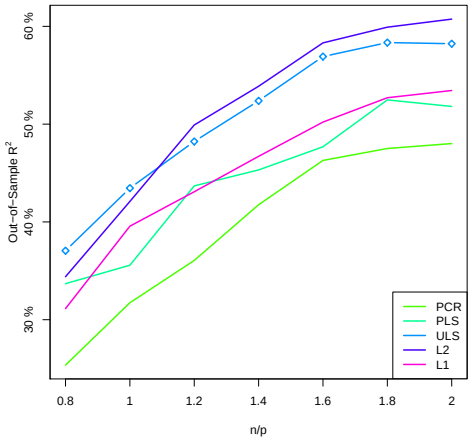
I also observe that ridge regression (L2) also achieves a similar effect as ULS, though it does not specifically split eigenpairs. Ridge regression minimizes $\boldsymbol{z}^T\Theta\boldsymbol{z} + \tau \cdot \|\boldsymbol{z}\|_2^2$ subject to $\boldsymbol{z}^T\boldsymbol{e}_1 = 1$, or equivalently, it minimizes $\boldsymbol{z}^T(\Theta + \tau I)\boldsymbol{z} =: \boldsymbol{z}^T\Theta_{L2}\boldsymbol{z}$ subject to that constraint. This is the same as adding τ to all eigenvalues of Θ and then computing the OLS solution. If $\lambda_i(\Theta) \ll \tau$ (typically the lower eigenvalues), then $\lambda_i(\Theta_{L2}) \approx \tau$, so the lower eigenvectors of Θ_{L2} are all indistinguishable in terms of their eigenvalues. This is similar in spirit to the uncertainty set of ULS, where the eigenvalues of the

Figure 4.4: Out-of-Sample R^2 for four classic regression datasets.

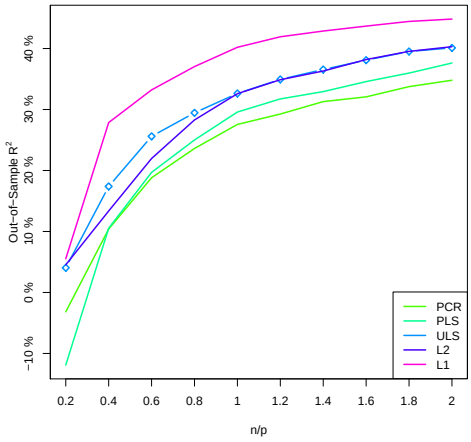
`Supernova` ($p = 10$)



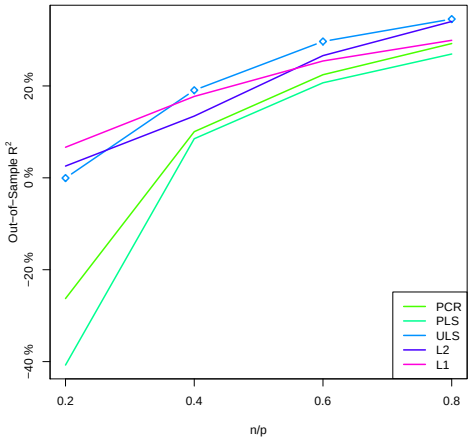
`Crime` ($p = 15$)



`Diabetes2` ($p = 64$)



`Protein` ($p = 88$)



lower eigenvectors are ignored, and only the subspace spanned by these eigenvectors is used. However, note that Ridge regression does not do objective matching, so it is still affected by the mismatch problem.

Table 4.3: Comparison of ULS over ULS without objective matching

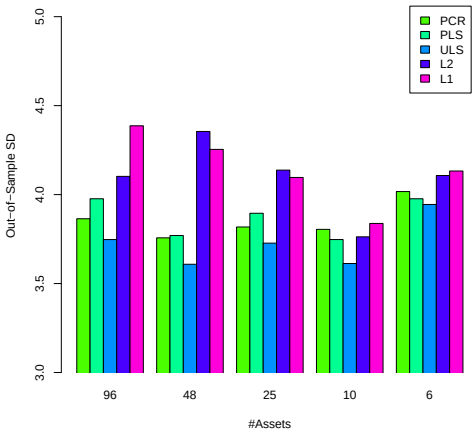
n/p	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
`Prostate`	-	-	-	-	-	14.4	16.4	5.2	5.4	0.8
`Crime`	-	-	-	11.7	6.6	3.0	2.5	2.2	1.4	1.8
`Diabetes1`	-	-	-	-	42.9	36.7	3.2	2.8	7.3	2.3
`Diabetes2`	117.5*	41.1	11.9	3.4	3.6	4.4	3.3	3.4	3.3	3.1
`Protein`	98.8*	50.9	8.7	3.2	-	-	-	-	-	
`Supernova`	-	-	-	-	10.3	2.7	2.7	1.1	1.1	0.9
`Community`	7.3	1.1	1.4	0.6	0.5	0.2	0.1	0.4	0.3	0.3

The table reports $\frac{OR^2(ULS)-OR^2(ULSno)}{|OR^2(ULSno)|}$ as percentages where ULSno stands for ULS without objective matching. The stars represent settings where $OR^2(ULSno) < 0$, so ULSno was worse than the baseline.

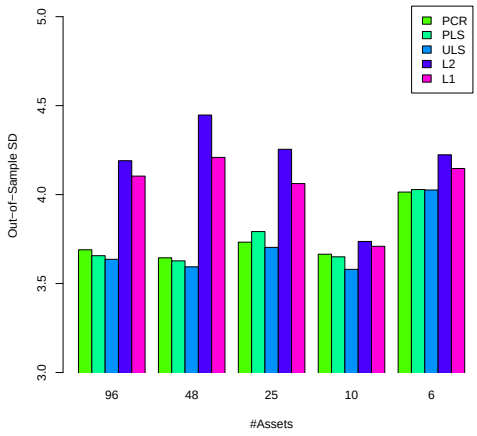
Importance of objective matching. To quantify the effect of objective matching, I compare ULS against a variant of ULS (called ULSno) that does not use objective matching (i.e., it uses $c = 1$, so $\Theta(c) = \Theta$). Table 4.3 shows the relative difference in the out-of-sample R^2 of ULS versus ULSno, averaged over 100 repetitions. I find that ULS outperforms ULSno in every case, with the greatest differences appearing when n/p is small. These are precisely the cases where prediction is difficult. As $n/p \to \infty$, I expect all eigenpairs to be well estimated, so both ULS and ULSno converge to the "classical" OLS solution. This emphasizes the need for objective matching in low-data settings.

Figure 4.5: Out-of-Sample Standard Deviation using $n = 60$ and $n = 120$ observations for five financial datasets

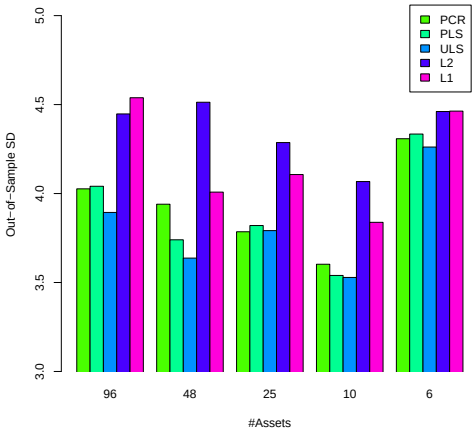
Value-Weighted Datasets ($n = 60$)



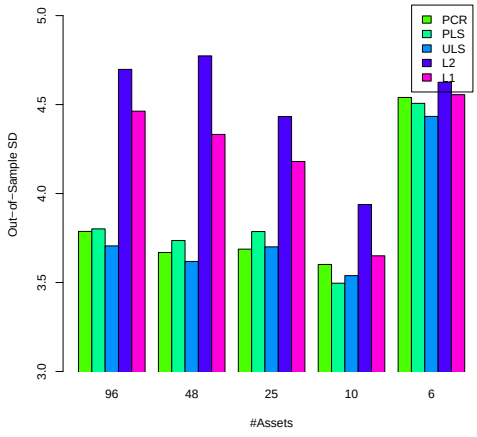
Value-Weighted Datasets ($n = 120$)



Equal-Weighted Datasets ($n = 60$)



Equal-Weighted Datasets ($n = 120$)



4.6.2 Financial Datasets

For an application outside the classic regression datasets, I turn to the minimum-variance optimization. I convert this into a LS problem (Eq. 4.11) and compare ULS against competing regression methods. The setup of the experiments follows Zhao et al. (2019a). I present results on five value-weighted and five equally-weigthed Fama and French (1992) datasets mentioned in Table 4.2, starting from July 1963 and ending in July 2015. I evaluate algorithms using a rolling window: at time t, the past n monthly returns are used to construct a portfolio, which is held for the next month. The return of this portfolio is recorded. Then, the starting and ending period are both shifted forward by one month, a new portfolio is built from this set of n monthly returns. The process is repeated until the end of the dataset. This yields a sequence of returns, and I report the standard deviation of these returns. Since the number of stocks p is fixed for each dataset, and portfolios are always constructed using returns from the previous n months, the n/p ratio is fixed. Following DeMiguel et al. (2009a); Brodie et al. (2009); Zhao et al. (2019a), I run experiments with $n = 60$ months and $n = 120$ months.

Figures 4.5 shows the out-of-sample standard deviation of returns for $n = 60$ and $n = 120$. The ULS method consistently beats PCR and PLS which, once again, supports the usefulness of the robust optimization. Note that LASSO (L1) and ridge regression (L2) perform worse than PCR and PLS. This is the opposite of what I observed for the datasets shown in Figure 4.3. Thus, while ULS is the best or second-best method for most datasets and

settings, no other method is as consistent.

I also compared ULS against the Bounded-Noise (BN) algorithm, which is specifically designed for portfolio optimization and has been shown to be competitive or better than state of the art on this problem (Zhao et al., 2019a). Both performed similarly across all datasets, with a relative difference of 0.15% on average for $n = 120$ and 0.79% for $n = 60$. There was no clear winner. Note that BN requires a parameter while ULS is fully automatic, and BN is much slower because intensive bootstrapping is required. Thus, ULS provides a competitive algorithm for portfolio optimization even though it is designed for the LS regression problem.

4.6.3 Appendix

4.7 Proofs of Lemmas and Theorems

Lemma 4.7.1 (Concentration of Eigenvectors (Yu et al., 2015))**.** *Let $A, B \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1(A) \geq \ldots \geq \lambda_p(A)$ and $\lambda_1(B) \geq \ldots \geq \lambda_p(B)$, respectively. Fix $1 \leq r \leq s \leq p$, and assume that $\min(\lambda_{r-1}(A) - \lambda_r(A), \lambda_s(A) - \lambda_{s+1}(A)) > 0$, where I define $\lambda_0(A) = \infty$ and $\lambda_{p+1}(A) = -\infty$. Let $d = s - r + 1$. Let $V(A) = (\boldsymbol{v}_r(A), \boldsymbol{v}_{r+1}(A), \ldots, \boldsymbol{v}_s(A)) \in \mathbb{R}^{p\times d}$ and $V(B) = (\boldsymbol{v}_r(B), \boldsymbol{v}_{r+1}(B), \ldots, \boldsymbol{v}_s(B)) \in \mathbb{R}^{p\times d}$. Then there exists an orthogonal matrix $O \in \mathbb{R}^{d\times d}$ such that*

$$\|V(A) - V(B)O\|_F \leq \frac{2^{3/2} d^{1/2} \|A - B\|_{op}}{\min\left(\lambda_{r-1}(A) - \lambda_r(A), \lambda_s(A) - \lambda_{s+1}(A)\right)}.$$

Theorem 4.7.2. *Let $\lambda_i(c)$ and $\boldsymbol{v}_i(c)$ represent eigenvalues and the corresponding eigenvectors of $\Theta(c)$, with $\lambda_1(c) \geq \lambda_2(c) \geq \ldots \geq \lambda_{p+1}(c)$. Define $\lambda_i(\Phi)$*

and $\boldsymbol{v}_i(\Phi)$ *accordingly. Let* $\boldsymbol{v}_i^{[1]} := \boldsymbol{v}_i(c)^T \boldsymbol{e}_1$ *denote the first element of* $\boldsymbol{v}_i$*, and* $\boldsymbol{v}_i^{[-1]}$ *be* $\boldsymbol{v}_i$ *without the first element. Assume* $Ey^2 > 0$*, and* Φ *has no repeated eigenvalues*[6]*. Then, I have Then,*

$$\lambda_1(c) = Ey^2 + O(1/c^2), \quad \left\| \boldsymbol{v}_1(c) - \frac{\boldsymbol{\ell}}{\|\boldsymbol{\ell}\|} \right\| = O(1/c^2), \quad \textit{where } \boldsymbol{\ell} = \begin{pmatrix} \sqrt{Ey^2} \\ -\frac{1}{c}\frac{E\boldsymbol{x}y}{\sqrt{Ey^2}} \end{pmatrix}.$$

Moreover, for all $i \in [2, p+1]$*, I have*

$$\begin{aligned} \lambda_i(c) &= \frac{1}{c^2}\lambda_{i-1}(\Phi) + O(1/c^3), \\ \|\boldsymbol{v}_i^{[-1]}(c) - \boldsymbol{v}_{i-1}(\Phi)\| &= O(1/c), \\ \boldsymbol{v}_i^{[1]}(c) &= \frac{1}{c}\frac{\boldsymbol{v}_{i-1}(\Phi)^T E[\boldsymbol{x}y]}{Ey^2} + O(1/c^2). \end{aligned}$$

Further, the corresponding statements hold if $\Theta(c)$ *and* Φ *are replaced by their empirical counterparts.*

Proof. Proof. By Weyl's inequality,

$$|\lambda_1(c) - \lambda_1(\boldsymbol{\ell}\boldsymbol{\ell}^T)| \leq \|\Theta(c) - \boldsymbol{\ell}\boldsymbol{\ell}^T\|_{op} = \left\| \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \frac{E\boldsymbol{x}\boldsymbol{x}^T - \frac{(E\boldsymbol{x}y)(Ey\boldsymbol{x}^T)}{Ey^2}}{c^2} \end{pmatrix} \right\|_{op} = O(1/c^2).$$

By applying Lemma 4.7.1 with $A = \boldsymbol{\ell}\boldsymbol{\ell}^T$, $B = \Theta(c)$, and $r = s = 1$, I have

$$\left\| \boldsymbol{v}_1(c) - \frac{\boldsymbol{\ell}}{\|\boldsymbol{\ell}\|_2} \right\|_2 \leq \frac{2^{3/2}\|\Theta(c) - \boldsymbol{\ell}\boldsymbol{\ell}^T\|_{op}}{\|\boldsymbol{\ell}\|_2^2} = O(1/c^2),$$

since $\boldsymbol{\ell}\boldsymbol{\ell}^T$ is a rank-one matrix with eigenvalue $\|\boldsymbol{\ell}\|_2^2$. Note that this also implies the looser bound $\|\boldsymbol{v}_1(c) - \boldsymbol{e}_1\| = O(1/c)$. From the fact that $\boldsymbol{v}_1(c)^T \boldsymbol{v}_i(c) = 0$

[6]This assumption is just for the simplicity of the result. If there exists $1 \leq r < s \leq p$ that $\lambda_r(\Phi) = \lambda_{r+1}(\Phi) = \cdots = \lambda_s(\Phi)$, the result will be the about the matrix $(\boldsymbol{v}_r(\Phi), \ldots, \boldsymbol{v}_s(\Phi))$ instead of individual eigenvectors.

for all $i > 1$, I have

$$\frac{\sqrt{Ey^2}}{\|\boldsymbol{\ell}\|}\boldsymbol{v}_i^{[1]}(c) - \frac{1}{c\|\boldsymbol{\ell}\|\sqrt{Ey^2}}E[y\boldsymbol{x}^T]\boldsymbol{v}_i^{[-1]}(c) = O(1/c^2)$$

$$\Rightarrow \quad \boldsymbol{v}_i^{[1]}(c) - \frac{1}{c \cdot Ey^2}E[y\boldsymbol{x}^T]\boldsymbol{v}_i^{[-1]}(c) = O(1/c^2). \tag{4.12}$$

Now, using $\Theta(c)\boldsymbol{v}_i(c) = \lambda_i(c)\boldsymbol{v}_i(c)$, I have

$$\Theta(c)\boldsymbol{v}_i(c) = \begin{pmatrix} Ey^2 & -\frac{Ey\boldsymbol{x}^T}{c} \\ -\frac{E\boldsymbol{x}y}{c} & \frac{E\boldsymbol{x}\boldsymbol{x}^T}{c^2} \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_i^{[1]}(c) \\ \boldsymbol{v}_i^{[-1]}(c) \end{pmatrix} = \lambda_i(c) \begin{pmatrix} \boldsymbol{v}_i^{[1]}(c) \\ \boldsymbol{v}_i^{[-1]}(c) \end{pmatrix}$$

$$\Rightarrow \quad -\frac{E\boldsymbol{x}y}{c}\boldsymbol{v}_i^{[1]}(c) + \frac{E\boldsymbol{x}\boldsymbol{x}^T}{c^2}\boldsymbol{v}_i^{[-1]}(c) = \lambda_i(c)\boldsymbol{v}_i^{[-1]}(c)$$

$$\Rightarrow \quad \frac{1}{c^2}\left(E\boldsymbol{x}\boldsymbol{x}^T - \frac{(E\boldsymbol{x}y)(Ey\boldsymbol{x}^T)}{Ey^2}\right)\boldsymbol{v}_i^{[-1]}(c) + O(1/c^3)E\boldsymbol{x}y = \lambda_i(c)\boldsymbol{v}_i^{[-1]}(c)$$

(from Eq. 4.12)

$$\Rightarrow \quad c^2\lambda_i(c)\boldsymbol{v}_i^{[-1]}(c) = \left(\Phi + \frac{O(1/c)}{\|\boldsymbol{v}_i^{[-1]}(c)\|^2}E[\boldsymbol{x}y]\boldsymbol{v}_i^{[-1]}(c)^T\right)\boldsymbol{v}_i^{[-1]}(c) =: (\Phi + \Delta)\boldsymbol{v}_i^{[-1]}(c). \tag{4.13}$$

The last statement used the fact that $\|\boldsymbol{v}_i^{[-1]}(c)\| \neq 0$ (otherwise $\boldsymbol{v}_i^{[1]}(c) = O(1/c^2)$ from Eq. 4.12, but I need $\|\boldsymbol{v}_i(c)\| = 1$). Hence, $c^2\lambda_i(c)$ and $\boldsymbol{v}_i^{[-1]}(c)$ are the $(i-1)^{th}$ eigenvalue and corresponding eigenvector of the matrix $\Phi + \Delta$, with $\|\Delta\|_{op} = O(1/c)$. By Weyl's inequality, I have

$$|c^2\lambda_i(c) - \lambda_{i-1}(\Phi)| \leq \|\Phi + \Delta - \Phi\|_{op} = O(1/c).$$

Similarly, by Lemma 4.7.1,

$$\|\boldsymbol{v}_i^{[-1]}(c) - \boldsymbol{v}_{i-1}(\Phi)\|_2 \leq \frac{2^{3/2}\|\Phi + \Delta - \Phi\|_{op}}{\min\left(\lambda_{i-1}(\Phi) - \lambda_i(\Phi), \lambda_i(\Phi) - \lambda_{i+1}(\Phi)\right)} = O(1/c). \tag{4.14}$$

Applying Eq. 4.14 to Eq. 4.12, I get

$$\boldsymbol{v}_i^{[1]}(c) = \frac{1}{c}\frac{\boldsymbol{v}_{i-1}(\Phi)^T E[\boldsymbol{x}y]}{Ey^2} + O(1/c^2).$$

Similar arguments hold for the empirical matrices $\hat{\Theta}(c)$ and $\hat{\Phi}$. □

Theorem 4.3.1(Upper Bound of $OR^2(\frac{\boldsymbol{v}_1(\Theta)}{\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1})$)

Proof. Proof. Apply Lemma 4.7.1 with $A = \Theta$ and $B = \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \Phi \end{pmatrix}$, where $\Phi = E(\boldsymbol{x}\boldsymbol{x}^T) - \frac{(E\boldsymbol{x}y)(Ey\boldsymbol{x}^T)}{Ey^2}$. Since $A - B = \boldsymbol{\ell}\boldsymbol{\ell}^T$, where $\boldsymbol{\ell} = \begin{pmatrix} \sqrt{Ey^2} \\ -\frac{E\boldsymbol{x}y}{\sqrt{Ey^2}} \end{pmatrix}$, I have

$$\left\| \boldsymbol{v}_1(\Theta) - \begin{pmatrix} 0 \\ \boldsymbol{v}_1(\Phi) \end{pmatrix} \right\|_2 \leq 2^{3/2} \frac{Ey^2 + \frac{(Ey\boldsymbol{x}^T)(E\boldsymbol{x}y)}{Ey^2}}{\lambda_1(\Theta) - \lambda_2(\Theta)}.$$

Thus, I have the following bound for $(\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1)^2$:

$$\begin{aligned}(\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1)^2 &= \left(\left(\boldsymbol{v}_1(\Theta) - \begin{pmatrix} 0 \\ \boldsymbol{v}_1(\Phi) \end{pmatrix} \right)^T \boldsymbol{e}_1 \right)^2 \\ &\leq \left\| \boldsymbol{v}_1(\Theta) - \begin{pmatrix} 0 \\ \boldsymbol{v}_1(\Phi) \end{pmatrix} \right\|_2^2 \leq 2^3 \frac{\big((Ey^2)^2 + (Ey\boldsymbol{x}^T)(E\boldsymbol{x}y)\big)^2}{(Ey^2)^2\big(\lambda_1(\Theta) - \lambda_2(\Theta)\big)^2}.\end{aligned}$$

Hence,

$$OR^2\left(\frac{\boldsymbol{v}_1(\Theta)}{\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1}\right) = 1 - \frac{\lambda_1(\Theta)/(\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1)^2}{Ey^2} \leq 1 - 2^{-3}\frac{\lambda_1(\Theta)\big(\lambda_1(\Theta) - \lambda_2(\Theta)\big)^2 Ey^2}{\big((Ey^2)^2 + (Ey\boldsymbol{x}^T)(E\boldsymbol{x}y)\big)^2} \tag{4.15}$$

Noting that $\lambda_1(\Theta) \geq \boldsymbol{e}_1^T\Theta\boldsymbol{e}_1 = Ey^2$ and $\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1 \leq \|\boldsymbol{v}_1(\Theta)\|_2 = 1$, Eq. 4.15 also shows that

$$OR^2\left(\frac{\boldsymbol{v}_1(\Theta)}{\boldsymbol{v}_1(\Theta)^T\boldsymbol{e}_1}\right) \leq 1 - \frac{Ey^2}{Ey^2} = 0.$$

□

Theorem 4.3.2(Possible Ways to Change Θ)

Proof. Proof. Because $\Theta = (M^{-1})^T M^T \Theta M M^{-1}$ holds for any invertible M, I have

$$\min_{\boldsymbol{z}|\boldsymbol{z}^T\boldsymbol{e}_1=1} \boldsymbol{z}^T\Theta\boldsymbol{z} = \min_{\boldsymbol{z}|\boldsymbol{z}^T\boldsymbol{e}_1=1} (M^{-1}\boldsymbol{z})^T M^T\Theta M(M^{-1}\boldsymbol{z}).$$

Moreover,

$$(M^{-1}\boldsymbol{z})^T\boldsymbol{e}_1 = (M^{-1}\boldsymbol{z})^T M\boldsymbol{e}_1 = \boldsymbol{z}^T\boldsymbol{e}_1 = 1,$$

where the first equality holds because $M\boldsymbol{e}_1 = \boldsymbol{e}_1$. Thus, I have

$$\min_{\boldsymbol{z}|\boldsymbol{z}^T\boldsymbol{e}_1=1} \boldsymbol{z}^T\Theta\boldsymbol{z} = \min_{\boldsymbol{z}|\boldsymbol{z}^T\boldsymbol{e}_1=1} \boldsymbol{z}^T M^T\Theta M\boldsymbol{z}.$$

□

Theorem 4.3.3(Best Starting Point)

Proof. Proof. Theorem 4.7.2 shows that $\boldsymbol{v}_1(c) = \boldsymbol{e}_1 + O(1/c)$ and $\lambda_1(c) = Ey^2 + O(1/c)$. Thus,

$$f_{\Theta(c)}\left(\frac{\boldsymbol{v}_1(c)}{\boldsymbol{v}_1(c)^T\boldsymbol{e}_1}\right) = \frac{\lambda_1(c)}{(\boldsymbol{v}_1(c)^T\boldsymbol{e}_1)^2} = \frac{Ey^2 + O(1/c)}{1 + O(1/c)} = Ey^2 + O(1/c),$$

which means that

$$OR^2\left(\frac{\boldsymbol{v}_1(c)}{\boldsymbol{v}_1(c)^T\boldsymbol{e}_1}\right) = 1 - \frac{f_{\Theta(c)}\left(\frac{\boldsymbol{v}_1(c)}{\boldsymbol{v}_1(c)^T\boldsymbol{e}_1}\right)}{Ey^2} = O(1/c).$$

□

Theorem 4.3.4(All Eigenvectors are Useful)

Proof. Proof. For $i \in [2, p+1]$, $f_{\Theta(c)}\left(\frac{\boldsymbol{v}_i(c)}{\boldsymbol{v}_i(c)^T \boldsymbol{e}_1}\right) = \frac{\lambda_i(c)}{(\boldsymbol{v}_i(c)^T \boldsymbol{e}_1)^2}$. From Theorem 4.7.2, both the numerator and denominator are of the order $O(1/c^2)$. Thus, $f_{\Theta(c)}(\frac{\boldsymbol{v}_i(c)}{\boldsymbol{v}_i(c)^T \boldsymbol{e}_1}) = O(1)$, so $OR^2(\frac{\boldsymbol{v}_i(c)}{\boldsymbol{v}_i(c)^T \boldsymbol{e}_1}) = O(1)$. □

Theorem 4.3.5(The Connection between $\Theta(c)$ and Φ)

Proof. Proof. All statements are derived from Theorem 4.7.2. □

Theorem 4.3.6(Eigenvalues of Φ)

Proof. Proof. Since $E\boldsymbol{x}\boldsymbol{x}^T$ is a submatrix of Θ, I have $\lambda_i(\Theta) \geq \lambda_i(E\boldsymbol{x}\boldsymbol{x}^T)$ by the Cauchy interlacing theorem. Since $E\boldsymbol{x}\boldsymbol{x}^T = \Phi + \frac{1}{Ey^2} E[y\boldsymbol{x}]E[y\boldsymbol{x}^T]$, I have $\lambda_i(E\boldsymbol{x}\boldsymbol{x}^T) \geq \lambda_i(\Phi)$ by Weyl's inequality. Finally, observe that Φ is the Schur complement of the top-left block of Θ (containing Ey^2), so $\lambda_i(\Phi) \geq \lambda_{i+1}(\Theta)$ (see Theorem 5 of **?**). This proves $\lambda_i(\Theta) \geq \lambda_i(E\boldsymbol{x}\boldsymbol{x}^T) \geq \lambda_i(\Phi) \geq \lambda_{i+1}(\Theta) \geq \lambda_{i+1}(E\boldsymbol{x}\boldsymbol{x}^T)$ for all i.

Because $E(\boldsymbol{x}\boldsymbol{x}^T) = \Phi + \frac{(E\boldsymbol{x}y)(Ey\boldsymbol{x}^T)}{Ey^2}$, I have

$$tr(E(\boldsymbol{x}\boldsymbol{x}^T)) = tr(\Phi) + tr(\frac{(E\boldsymbol{x}y)(Ey\boldsymbol{x}^T)}{Ey^2})$$

$$\Rightarrow \quad \sum_{i=1}^{p} \lambda_i(E\boldsymbol{x}\boldsymbol{x}^T) = \sum_{i=1}^{p} \lambda_i(\Phi) + \frac{(Ey\boldsymbol{x}^T)(E\boldsymbol{x}y)}{Ey^2}.$$

Here $tr(A)$ indicates the trace of matrix A. By rearranging, the second equality of Theorem 4.3.6 is proved.

□

Theorem 4.3.7(Solutions converge for large c)

Proof. Proof. I will apply Theorem 4.7.2 to Eq. 4.5. The denominator of $\hat{\boldsymbol{z}}^{\star}_{1:k}(c)$ is given by

$$\sum_{i=1}^{k} \frac{(\hat{\boldsymbol{v}}_i(c)^T \boldsymbol{e}_1)^2}{\hat{\lambda}_i(c)} = \frac{n}{\boldsymbol{y}^T\boldsymbol{y}} + O(1/c) + \sum_{i=2}^{k} \frac{1/c^2 (\boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T \boldsymbol{y})^2/(\boldsymbol{y}^T\boldsymbol{y})^2 + O(1/c^3)}{1/c^2 \lambda_{i-1}(\hat{\Phi}) + O(1/c^3)}$$
$$\rightarrow \frac{n}{\boldsymbol{y}^T\boldsymbol{y}} + \sum_{i=2}^{k} \frac{(\boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T \boldsymbol{y})^2}{\lambda_{i-1}(\hat{\Phi}) \cdot (\boldsymbol{y}^T\boldsymbol{y})^2}$$

as $c \rightarrow \infty$. The numerator of $M(c)\hat{\boldsymbol{z}}^{\star}_{1:k}(c)$ is given by

$$\sum_{i=1}^{k} \frac{\hat{\boldsymbol{v}}_i(c)^T \boldsymbol{e}_1}{\hat{\lambda}_i(c)} M(c)\hat{\boldsymbol{v}}_i(c)$$
$$= \frac{n}{\boldsymbol{y}^T\boldsymbol{y}}\boldsymbol{e}_1 + O(1/c) + \sum_{i=2}^{k} \frac{1/c \cdot \boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T \boldsymbol{y}/(\boldsymbol{y}^T\boldsymbol{y}) + O(1/c^2)}{1/c^2 \cdot \lambda_{i-1}(\hat{\Phi}) + O(1/c^3)} \begin{pmatrix} 1/c \cdot \boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T \boldsymbol{y}/(\boldsymbol{y}^T\boldsymbol{y}) \\ 1/c \cdot \boldsymbol{v}_{i-1}(\hat{\Phi}) \end{pmatrix}$$
$$\rightarrow \frac{n}{\boldsymbol{y}^T\boldsymbol{y}}\boldsymbol{e}_1 + \sum_{i=2}^{k} \frac{\boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T \boldsymbol{y}}{\lambda_{i-1}(\hat{\Phi}) \cdot (\boldsymbol{y}^T\boldsymbol{y})} \boldsymbol{w}_{i-1},$$

where the statement $\boldsymbol{a} = \boldsymbol{b} + O(.)$ is taken to imply $\|\boldsymbol{a} - \boldsymbol{b}\| = O(.)$. Together, these yield the limit point $\hat{\boldsymbol{z}}^{\star}_{1:k}$.

Similarly, I find

$$
\begin{aligned}
&M(c)\tilde{\boldsymbol{z}}^{\star}_{k+1:p+1}(c) \\
=&\frac{M(c)P_{N_k}\boldsymbol{e}_1}{\boldsymbol{e}_1^T P_{N_k}\boldsymbol{e}_1} = \frac{\sum_{k+1}^{p+1}(\hat{\boldsymbol{v}}_i(c)^T\boldsymbol{e}_1)M(c)\hat{\boldsymbol{v}}_i(c)}{\sum_{k+1}^{p+1}(\hat{\boldsymbol{v}}_i(c)^T\boldsymbol{e}_1)^2} \\
=&\frac{\sum_{k+1}^{p+1}\left(1/c\cdot\boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T\boldsymbol{y}/(\boldsymbol{y}^T\boldsymbol{y})+O(1/c^2)\right)\begin{pmatrix}1/c\cdot\boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T\boldsymbol{y}/(\boldsymbol{y}^T\boldsymbol{y}) \\ 1/c\cdot\boldsymbol{v}_{i-1}(\hat{\Phi})\end{pmatrix}}{\sum_{k+1}^{p+1}1/c^2\cdot(\boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T\boldsymbol{y})^2/(\boldsymbol{y}^T\boldsymbol{y})^2+O(1/c^3)} \\
\to&\frac{\sum_{i=k+1}^{p+1}(\boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T\boldsymbol{y})\boldsymbol{w}_{i-1}}{\sum_{i=k+1}^{p+1}(\boldsymbol{v}_{i-1}(\hat{\Phi})^T X^T\boldsymbol{y})^2/(\boldsymbol{y}^T\boldsymbol{y})},
\end{aligned}
$$

which yields the limit point $\tilde{\boldsymbol{z}}^{\star}_{k+1:p+1}$.

□

Theorem 4.4.1(The Connection between $\hat{\boldsymbol{v}}_i(X^TX)$ and $\hat{\boldsymbol{v}}_i(\hat{\Phi})$)

Proof. Proof. Apply Lemma 4.7.1 with $A = X^TX$ and $B = \hat{\Phi}$ to prove the inequality. □

Chapter 5

Enhanced Principle Component Analysis

For the convenience of reading, I try to make this chapter self-contained. Namely, I will reiterate some definitions and assumptions. This chapter loosely depends on Zhao et al. (2019a) and Zhao et al. (2019b).

5.1 Introduction

As data become more complex, dimensionality reduction plays a more and more critical role. Such reduction can reduce the time and storage required, generate good visualization of data, and avoid the curse of dimensionality. Principle component analysis (PCA) (Pearson, 1901; Hotelling, 1933) might be the most popular linear technique to reduce dimensionality and has applications through science and engineering (Jolliffe, 2011). It transforms data from possibly correlated variables to orthogonal principal components (PCs).

Though widely used, PCA still has several disadvantages. As an illustration, I will list the three main weaknesses. Firstly, PCs tend to be a linear combination of all variables. This characteristic restricts the power of interpretation. d'Aspremont et al. (2005) and Zou et al. (2006) among others

add sparsity to the PCs. Secondly, PCA is sensitive to outliers because its objective is variance. Devlin et al. (1981), Xu and Yuille (1995), and Xu et al. (2010) among others propose ways to address this issue. Thirdly, the decomposition of PCA does not take the objective into account. For example, in a regression setting, one applies PCA on the covariates without considering the dependent variable whose prediction is the target. Bair et al. (2006) proposes the supervised PCA to incorporate the objective. Because the top PCs might not be useful in predicting (Jolliffe, 1982), the supervised PCA might not take the top PCs.

There is still one issue of PCA that is ignored in the literature. No matter which PCs are selected, the subspace that is orthogonal to the chosen PCs is always ignored. However, if the selected PCs are well-estimated, the subspace is also well-estimated because of orthogonality. Ignoring it might lead to a loss of information.

In this chapter, I will propose a way to not only use the orthogonality information but also take the objective into account. I call this new methodology PCA+. To achieve this goal, I will first introduce the classical PCA in Section 5.2. I will illustrate the idea of PCA+ in portfolio optimization in Section 5.3. This is the most natural case to enhance PCA because the objectives of the optimization and PCA coincide. In Section 5.4, I apply the idea of PCA+ in the linear regression setting.

5.2 PCA

PCA tries to linearly transform the data to a new coordinate system such that the greatest variance along the first coordinate is largest, along the second coordinate is the second largest, and so on. The new coordinates are called the principal components (PCs). Mathematically speaking, the first PC in the original system is the solution to

$$\boldsymbol{v}_1 = \arg \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}' X' X \boldsymbol{v},$$

where X is the n-by-p data matrix with each row representing one observation. That is to say, there are n observations and p variables. Thus, the first PC is the first eigenvector of matrix $X'X$ which explains the choice of using the same notation $\boldsymbol{v}_1$. Also, the objective, $\boldsymbol{v}_1' X' X \boldsymbol{v}_1$, is the first eigenvalue of $X'X$, namely λ_1.

To calculate the $k+1$th PC, the previous k PCs are taken out from the data matrix X: $X - \sum_{i=1}^{k} X \boldsymbol{v}_i \boldsymbol{v}_i'$. Because $\boldsymbol{v}_i \boldsymbol{v}_i'$ is equivalent to an operator that projects values along $\boldsymbol{v}_i$, the subtraction means to taken out all the variations along the top k PCs. Since $\sum_{i=1}^{p} \boldsymbol{v}_i \boldsymbol{v}_i'$ is an identity matrix, $X - \sum_{i=1}^{k} X \boldsymbol{v}_i \boldsymbol{v}_i' = X \left(\sum_{i=k+1}^{p} \boldsymbol{v}_i \boldsymbol{v}_i' \right)$. For the simplicity, I denote $\sum_{i=k+1}^{p} \boldsymbol{v}_i \boldsymbol{v}_i'$ as $P_\perp$. That is to say, the new data matrix is $XP_\perp$.

To obtain the $k+1$th PC, one applies the first PC idea to $XP_\perp$.

$$\begin{aligned} \boldsymbol{v}_k &= \arg \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}' \left[(XP_\perp)'(XP_\perp') \right] \boldsymbol{v}, \\ &= \arg \max_{\|\boldsymbol{v}\|_2=1} (P_\perp \boldsymbol{v})' X' X (P_\perp \boldsymbol{v}). \end{aligned}$$

$P_\perp$ is an operator that projects values on the orthogonal subspace of the top k PCs, the solution to the above optimization is the k+1th eigenvector of the matrix $X'X$ which is orthogonal to the top k PCs. All in all, PCs are the eigenvectors of the matrix $X'X$.

PCA is closely related to the singular value decomposition (SVD) which decomposes data matrix X directly as

$$X = UDV' = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q) diag(d_1, \ldots, d_q) \begin{pmatrix} \boldsymbol{v}_1' \\ \vdots \\ \boldsymbol{v}_q' \end{pmatrix} = \sum_{i=1}^{q} d_i \boldsymbol{u}_i \boldsymbol{v}_i',$$

where $q = \min(n, p)$ and $diag(d_1, \ldots, d_q)$ is a diagonal matrix with $d_1, \ldots, d_q$ as the diagonal values. The n-by-q matrix U contains the left-singular vectors which are the eigenvectors of XX'. Similarly, the p-by-q matrix V contains the right-singular vectors which are the eigenvectors of $X'X$, namely the PCs of X. Moreover, it is easy to prove $d_i = \sqrt{\lambda_i}$.

Let $\mathbf{1}$ and $\bar{\boldsymbol{x}}$ be both length-p column vectors with all ones and the mean of each column of X as elements. Because $X'X = (X - \mathbf{1}\bar{\boldsymbol{x}}')'(X - \mathbf{1}\bar{\boldsymbol{x}}') + (\mathbf{1}\bar{\boldsymbol{x}}')'(\mathbf{1}\bar{\boldsymbol{x}}')$, $\bar{\boldsymbol{x}}$ affects the PCs of $X'X$. To eliminate such influence, it is conventional to demean the data matrix X, namely making $\bar{\boldsymbol{x}} = \mathbf{0}$ before using PCA. With the demeaned data matrix X, $X'X = (n-1)\hat{\Sigma}$, where $\hat{\Sigma}$ is the sample covariance matrix. For convenience, I assume the data is demeaned before PCA for all experiments from now on.

5.3 PCA+ in Portfolio Optimization

The minimum-variance optimization tries to minimize the variance of a portfolio $\boldsymbol{w}$. Mathematically speaking,

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}'\Sigma\boldsymbol{w}, \\ \text{subject to} \quad & \boldsymbol{w}'\mathbf{1} = 1, \end{aligned}$$

where Σ is the true covariance matrix and $\mathbf{1}$ is a length-p vector with all elements being 1. In reality, because Σ is unknown, one might use the sample covariance matrix $\hat{\Sigma}$ in place of it. Since $\hat{\Sigma} = \frac{1}{n-1}X'X$, and the optimization is invariant to scaling, replacing Σ with $\hat{\Sigma}$ in the above optimization is equivalent to solving

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}'(X'X)\boldsymbol{w}, \\ \text{subject to} \quad & \boldsymbol{w}'\mathbf{1} = 1. \end{aligned}$$

Interestingly, it shares the same objective with applying PCA on X. Because of the error amplification phenomenon mentioned in Chapter 2, the resulting portfolio might a terrible out-of-sample performance. PCA, which reduces the dimensionality, might come to rescue. Indeed, the $\hat{\boldsymbol{w}}_S^*$ portfolio mentioned in Chapter 2 is the solution when only the top PCs of X are considered in the optimization. Thus, I will call $\hat{\boldsymbol{w}}_S^*$ the PCA portfolio. As shown in DeMiguel et al. (2009a), the PCA portfolio does improve out-of-sample performance. However, as argued in Chapter 2, though the bottom PCs are individually poorly estimated, the subspace spanned by them are well estimated because of the orthogonality to the top PCs. Only keeping top PCs seems to be wasting the orthogonality information. I propose to utilize such information using

$\boldsymbol{w}_N^{EW}$ portfolio earlier. If one uses the perspective of dimensionality reduction, it is natural to use the following directions

$$V_{PCA+} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, P_\perp \mathbf{1}),$$

where $P_\perp = \sum_{i=k+1}^{p} \boldsymbol{v}_i \boldsymbol{v}_i'$ is the projection operator for the subspace that is orthogonal to the top k PCs. It is worth to notice that the L2 norm of $P_\perp \mathbf{1}$ is not 1. Because only the direction matters for the final solution, for simplicity, I keep it this way. It is tempting to reduce dimension by using XV_{PCA+}, but such operation will destroy the robustness of $P_\perp \mathbf{1}$ because the projection of X on the orthogonal subspace, namely $XP_\perp$, is poorly estimated. Instead, I propose to replace X with $X(\eta)$ which shares the same U, V matrix, and $d_1, \ldots, d_k$ with X, but all the remaining singular values are η which is a constant. In this way, the projection of $X(\eta)$ on the orthogonal subspace only depends on the well-estimated subspace. The reduced data become

$$X_{PCA+} = (X\boldsymbol{v}_1, \ldots, X\boldsymbol{v}_k, X(\eta)P_\perp \mathbf{1}).$$

Because $\boldsymbol{w} = V_{PCA+}\boldsymbol{l}$, the minimum-variance optimization now becomes

$$\begin{aligned} \min_{\boldsymbol{l}} \quad & \boldsymbol{l}'(X_{PCA+}'X_{PCA+})\boldsymbol{l}, \\ \text{subject to} \quad & (V_{PCA+}\boldsymbol{l})'\mathbf{1} = 1. \end{aligned}$$

It generates the following portfolio,

$$\boldsymbol{w}_{PCA+}(\eta) = \frac{\sum_{i=1}^{k} \frac{\boldsymbol{v}_i'\mathbf{1}}{\lambda_i}\boldsymbol{v}_i + \frac{1}{\eta^2}P_\perp \mathbf{1}}{\sum_{i=1}^{k} \frac{(\boldsymbol{v}_i'\mathbf{1})^2}{\lambda_i} + \frac{\mathbf{1}'P_\perp \mathbf{1}}{\eta^2}} = (1 - a_1(\eta))\hat{\boldsymbol{w}}_S^* + a_1(\eta)\boldsymbol{w}_N^{EW}, \qquad (5.1)$$

where $a_1(\eta) = \left(\sum_{i=1}^{k} \frac{\mathbf{1}'P_\perp \mathbf{1}}{\eta^2}\right) / \left(\sum_{i=1}^{k} \frac{(\boldsymbol{v}_i'\mathbf{1})^2}{\lambda_i} + \frac{\mathbf{1}'P_\perp \mathbf{1}}{\eta^2}\right)$. Because $\hat{\boldsymbol{w}}_S^\star$ is the PCA solution, the PCA+ solution is a combination of the PCA solution and the

projection of the equally-weighted portfolio. Clearly, the choice of η is essential because it determines the combination level between $\hat{\boldsymbol{w}}_S^*$ and $\boldsymbol{w}_N^{EW}$. I will document the way to choose η in Section 5.3.2. Instead of simply pursuing the best out-of-sample performance as the Unified portfolio, the PCA+ portfolio focuses on achieving a good performance with a descent reduction of dimensionality. That is to say, a small k is preferred.

5.3.1 Exploration Using Simulation

In this subsection, I would like to use the following simulation example to explore the possibility of obtaining good performance with a small k.

The simulation is based on the Fama-French value-weighted dataset comprising 96 risky assets. I assume that the true covariance matrix Σ and the true expected return $\boldsymbol{\mu}$ are the sample covariance matrix and the sample mean using all monthly data from July 1963 to July 2015 (625 observations). I also assume that the returns follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ, and I draw 120 observations (10-year monthly data) from this distribution.

Given one set of the simulated data, for each possible number of PCs, I calculate the PCA portfolio and the PCA+ portfolio with the optimal η that minimizes its realized variance[1]. For readability, I call the latter the oracle PCA+ portfolio because it needs the true covariance matrix Σ in the

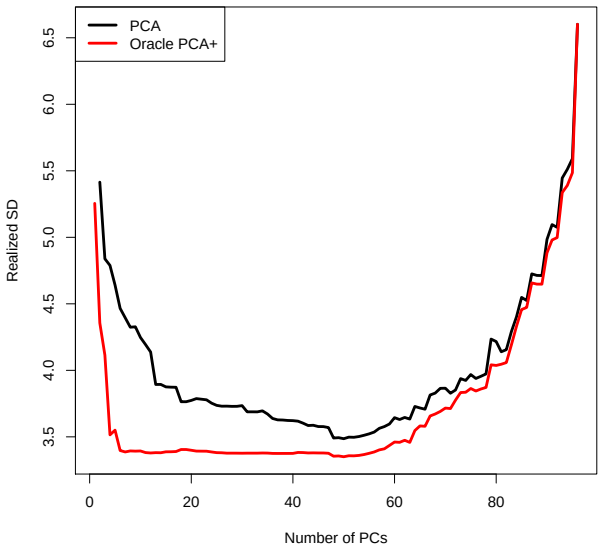
[1]The realized variance of portfolio $\boldsymbol{w}$ is defined as $RV(\boldsymbol{w}) = \boldsymbol{w}'\Sigma\boldsymbol{w}$.

combination procedure. Though not achievable in reality, it helps to show the potential of the PCA+ portfolio.

From Figure 5.1, the oracle PCA+ portfolio sees a sharp decrease in realized standard deviation (SD) for the several top PCs where the most significant difference between it and the PCA portfolio exists. The improvement afterward is minuscule. Moreover, the oracle PCA+ portfolio can achieve a lower realized SD using only several PCs than the PCA portfolio involving almost 50 PCs.

Here, I will explain the phenomenon mentioned above. From Figure 2.1, the eigengaps of Σ for the several top eigenvalues are much larger than others. Based on Lemma 2.3.2, the estimation errors of several top PCs should be much smaller than other PCs. Namely, the subspace that is orthogonal to the several top PCs is extremely well estimated. Thus, even though the PCA portfolio performs better as it includes more PCs, the corresponding orthogonal subspace not only becomes smaller but also contains more error. That is to say, for the oracle PCA+ portfolio, there is a tradeoff when a new PC is added. For the several top PCs, the benefit of PCA dominates the penalty of smaller and worst orthogonality subspace, and a sharp drop in realized SD occurs. Then they become similar, and the realized SD remains almost the same for a long time. Finally, the penalty grows more prominent than the benefit, and the realized SD starts to increase. The most significant difference between them happens for the top PCs because the orthogonal subspace is both extensive and exceptionally well estimated.

Figure 5.1: Realized SD for PCA and oracle PCA+ Portfolio



5.3.2 Parameter Choices of k and η

In this subsection, I will propose a way to approximate the oracle PCA+ portfolio by choosing k and η without knowing the true covariance matrix. The presented way is ad-hoc but serves the purpose of demonstrating the possibility to achieve excellent performance with a small dimension.

Across all portfolio experiments, I choose $k = 4$. Namely, I reduce the dimensionality of data to $k + 1 = 5$. With such a small k, the estimated variance[2] of $\hat{\boldsymbol{w}}_S^*$ should be a great estimation of its realized variance, and the realized covariance[3] between $\hat{\boldsymbol{w}}_S^*$ and $\boldsymbol{w}_N^{EW}$ should be small. Based on Lemma

[2]The estimated variance of a portfolio $\boldsymbol{w}$ is defined as $EV(\boldsymbol{w}) = \boldsymbol{w}'\hat{\Sigma}\boldsymbol{w}$.
[3]The realized covariance between portfolio $\boldsymbol{w}_A$ and $\boldsymbol{w}_B$ is defined as $\boldsymbol{w}_A'\Sigma\boldsymbol{w}_B$

3.2.1, the optimal combination level should be close to

$$\frac{RV(\boldsymbol{w}_N^{EW})}{RV(\hat{\boldsymbol{w}}_S^*) + RV(\boldsymbol{w}_N^{EW})} \approx \frac{RV(\boldsymbol{w}_N^{EW})}{EV(\hat{\boldsymbol{w}}_S^*) + RV(\boldsymbol{w}_N^{EW})}. \tag{5.2}$$

Thus, a good choice of η should be able to estimate $RV(\boldsymbol{w}_N^{EW})$. Because λ_k serves a bound for all eigenvalues corresponding to the thrown away PCs, $\eta = \sqrt{\lambda_k}$ will generate a bound for $RV(\boldsymbol{w}_N^{EW})$. Because k is small, this bound should be loose. I find that $\sqrt{\lambda_k/1.5}$, $\sqrt{\lambda_k/2}$, and $\sqrt{\lambda_k/2.5}$ all generate similar results. For simplicity, I decided to show the results with $\eta = \sqrt{\lambda_k/2}$.

5.3.3 Empirical Results

In this subsection, I will compare the PCA+ portfolio with the following PCA related portfolios.

Competing methods. The oracle PCA portfolio chooses the number of PCs **after** observing the out-of-sample returns. It is unachievable in reality but serves as an upper bound for PCA related methods. Because its existence, PCA with cross-validation is not presented. To see the improvement of the PCA+ portfolio, I also include the PCA4 and PCA5 portfolio which uses the top 4 and 5 PCs, respectively. The PCA90% (PCA95%) portfolio chooses the number of PCs such that 90% (95%) of the total variance is included. These two ways are very commonly used across the applications of PCA.

The setup of the experiments. I present results on five value-weighted and five equally-weigthed Fama and French (1992) datasets mentioned in Table 4.2, starting from July 1963 and ending in July 2015. I evalu-

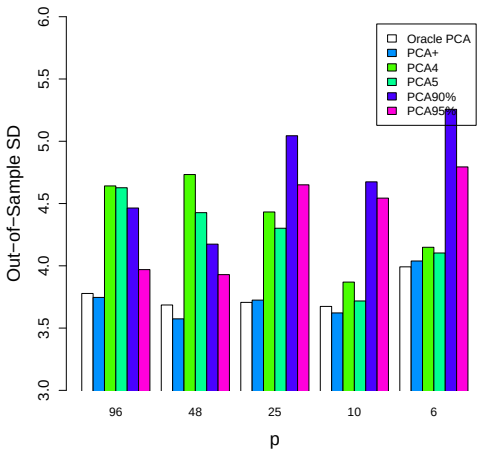
ate algorithms using a rolling window: at time t, the past n monthly returns are used to construct a portfolio, which is held for the next month. The return of this portfolio is recorded. Then, the starting and ending period are both shifted forward by one month, a new portfolio is built from this set of n monthly returns. The process is repeated until the end of the dataset. This yields a sequence of returns, and I report the standard deviation of these returns. Since the number of stocks p is fixed for each dataset, and portfolios are always constructed using returns from the previous n months, the n/p ratio is fixed. Following DeMiguel et al. (2009a); Brodie et al. (2009); Zhao et al. (2019a), I run experiments with $n = 60$ months and $n = 120$ months.

Figure 5.2 presents the out-of-sample standard deviation (SD) for all six portfolios. Across all 20 experiments, the PCA+ portfolio is almost as good as the oracle PCA portfolio. For the datasets with more than six assets, the oracle PCA portfolio never chooses all PCs indicating that the estimation errors are large enough to affect performance negatively. For these datasets, the PCA+ portfolio achieves its performance by handling the estimation errors effectively. For the datasets with six assets, the oracle PCA portfolio uses all PCs. Thus, the good performance of the PCA+ portfolio comes from the fact that the fifth and sixth eigenvalues are similar. That is to say, PCA+ works well because the assumptions are approximately correct.

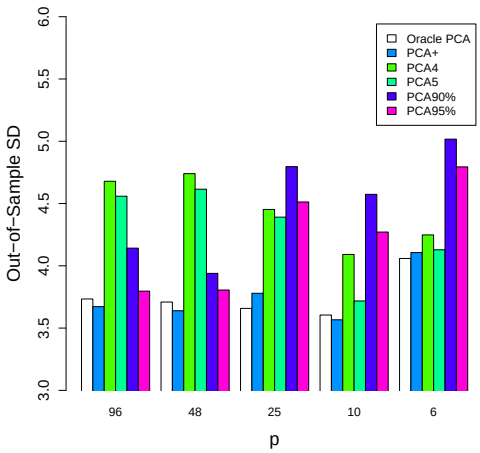
The PCA+ portfolio dominates the PCA4 portfolio for all experiments. The gap tends to be more prominent as the number of assets increases. This makes sense because the orthogonality subspace becomes more and more im-

Figure 5.2: Out-of-Sample SD of PCA+ and PCA related Portfolios

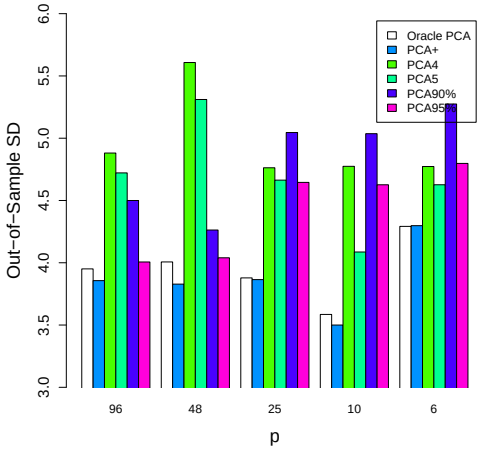
Value-Weighted Datasets ($n = 60$)



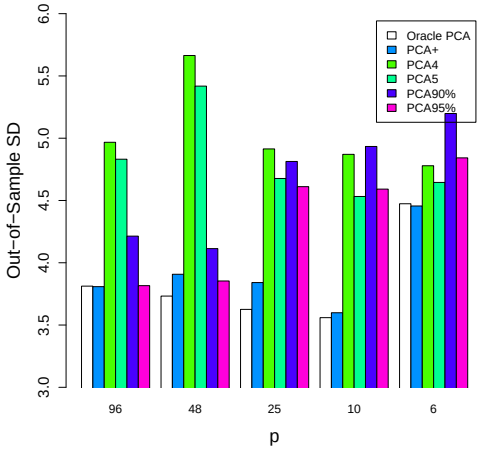
Value-Weighted Datasets ($n = 120$)



Equal-Weighted Datasets ($n = 60$)



Equal-Weighted Datasets ($n = 120$)



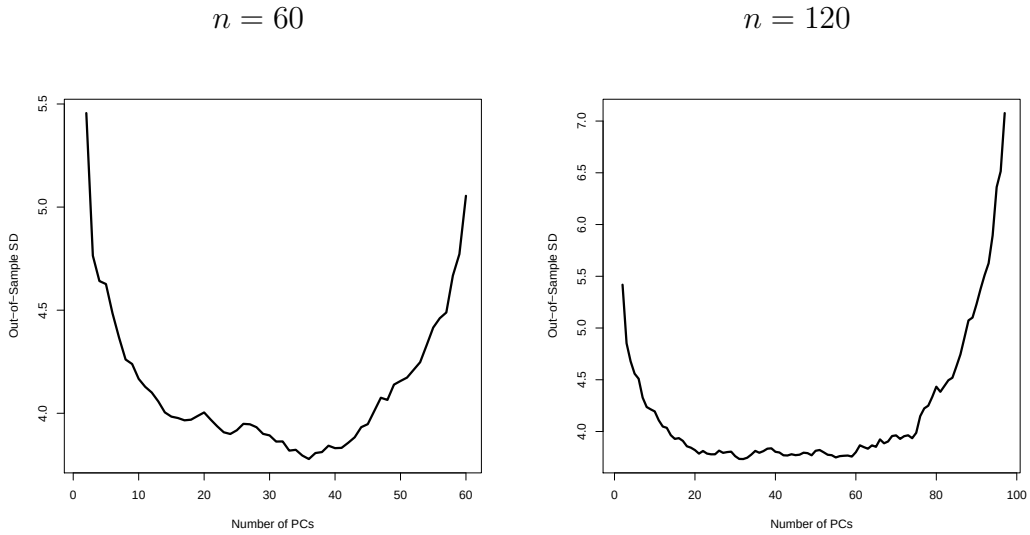
portant as its dimensionality grows. This shows the power of enhancing PCA using orthogonality information.

The PCA5 portfolio is always inferior to the PCA+ portfolio. Because they have the same number of dimensions, this shows that PCA+ does provide a higher quality reduction of dimensionality. For most cases, the PCA5 portfolio has a similar level of performance with the PCA4 portfolio indicating a gradual change as more PCs are included.

Both PCA90% and PCA95% portfolios perform much worse than the oracle PCA portfolio most of the times. Moreover, PCA90% consistently performs worse than PCA95%. This seems to be counterintuitive for the large datasets because PCA95% chooses more PCs than PCA90% which should be problematic for big datasets. This mystery is resolved in Figure 5.3 which shows the out-of-sample SD with respect to the number of PCs used for the 96FFVW dataset. Because PCA90% chooses only about 10 PCs while PCA95% chooses about 20 PCs, based on Figure 5.3, PCA95% should have a lower out-of-sample SD.

The PCA+ portfolio only uses 5 dimensions while the oracle PCA portfolio chooses more than 30 PCs for both $n = 60$ and $n = 120$ cases. This coincides with the expectation motivated by the simulation exploration: PCA+ provides a much efficient reduction of dimensionality.

Figure 5.3: Out-of-Sample SD of PCA portfolios for 96FFVW dataset



5.4 PCA+ in Linear Regression

In this section, I first discuss the challenge of extending the PCA+ portfolio idea to linear regression. Then, just like the portfolio optimization case, I will use simulation to explore the possibility to obtain good prediction results with just a few dimensions. Finally, I will present experiments on real-world datasets.

5.4.1 Inconsistent Objectives

Linear regression brings the issue of inconsistent objectives: the objective for PCA is the variance of X while the objective for linear regression is the prediction of $\boldsymbol{y}$, the dependent variable. Technically speaking, the new

direction added in the PCA+ portfolio, namely $P_\perp \mathbf{1}$, no longer works, because it is independent of the dependent variable $\boldsymbol{y}$. To fix this issue, one needs to understand the reasoning behind the choice of $P_\perp \mathbf{1}$ for portfolio optimization: the good performance and robustness of the equally-weighted portfolio (Jobson and Korkie, 1980; DeMiguel et al., 2009b; Duchin and Levy, 2009). Thus, I need to find a robust solution, $\boldsymbol{\beta}_{rob}$ for the linear regression that has good performance, and then replace $P_\perp \mathbf{1}$ with $P_\perp \boldsymbol{\beta}_{rob}$.

Because the derivation of the PCA+ solution goes through with any robust solution, I will present my choice of $\boldsymbol{\beta}_{rob}$ later.

For the readability, I still use V_{PCA+} and X_{PCA+} here representing

$$V_{PCA+} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, P_\perp \boldsymbol{\beta}_{rob}),$$

$$X_{PCA+} = (X\boldsymbol{v}_1, \ldots, X\boldsymbol{v}_k, X(\eta) P_\perp \boldsymbol{\beta}_{rob}).$$

Assuming $\boldsymbol{\beta} = V_{PCA+}\boldsymbol{l}$, the linear regression becomes

$$\min_{\boldsymbol{l}} (\boldsymbol{y} - X_{PCA+}\boldsymbol{l})'(\boldsymbol{y} - X_{PCA+}\boldsymbol{l}).$$

It generates the following solution,

$$\boldsymbol{\beta}_{PCA+} = \sum_{i=1}^{k} \frac{\boldsymbol{v}_i'(X'\boldsymbol{y})}{\lambda_i}\boldsymbol{v}_i + \frac{(P_\perp \boldsymbol{\beta}_{rob})'(X'(\eta)\boldsymbol{y})}{\eta^2 \sum_{i=k+1}^{p} (\boldsymbol{v}_i'\boldsymbol{\beta}_{rob})^2} P_\perp \boldsymbol{\beta}_{rob}.$$

Noticing that the first term is the PCA solution, $\boldsymbol{\beta}_{PCA}$, I can rewrite it as

$$\boldsymbol{\beta}_{PCA+} = \boldsymbol{\beta}_{PCA} + a_2(\eta) P_\perp \boldsymbol{\beta}_{rob}, \tag{5.3}$$

where $a_2(\eta) = (P_\perp \boldsymbol{\beta}_{rob})'(X'(\eta)\boldsymbol{y})/(\eta^2 \sum_{i=k+1}^{p} (\boldsymbol{v}_i'\boldsymbol{\beta}_{rob})^2)$. Just as the portfolio case, the PCA+ solution is closely related the PCA solution and the projection

of a robust solution. Instead of a linear combination of two in the minimum-variance optimization, it is adding a proportion of the latter to the former for the linear regression.

What left is the choice of $\boldsymbol{\beta}_{rob}$ The most robust solution is $\boldsymbol{\beta} = \mathbf{0}$. However, it is useless here because $P_{\perp}\mathbf{0} = \mathbf{0}$ which leads to the traditional PCA. Recall that the OLS solution, $\boldsymbol{\beta}_{OLS} = (X'X)^{-1}X'\boldsymbol{y}$, suffers from the collinearity because of the existence of $(X'X)^{-1}$. One way to fix it is to replace $(X'X)^{-1}$ with a matrix which is proportional to the identity matrix. This leads to a solution that is proportional to $\boldsymbol{\beta}_{rob} \propto X'\boldsymbol{y}$. The vector $X'\boldsymbol{y}$ also plays an important role in PLS (Wold, 1966) and ULS (Zhao et al., 2019b).

5.4.2 Exploration Using Simulation

In this subsection, I use simulation to explore the possibility of obtaining a good prediction using a small number of dimensions.

The simulation is based on dataset `Diabetes2` which has 442 observations and 64 covariates from Table 4.2. I generate the independent variables, X, using a multi-normal distribution where the mean and covariance are the sample counterparts using all observations. Then I use $N(X\boldsymbol{\beta}, \sigma)$ to generate the dependent variable, $\boldsymbol{y}$, where $\boldsymbol{\beta}$ and σ are obtained by a linear regression of $\boldsymbol{y}$ on X using all observations. I draw $n = 128 = 2p$ from this procedure.

Given one set of simulated data, for each possible split, I calculate the

PCA solution[4] and the PCA+ solution with the optimal η that maximizes the out-of-sample R^2. As in the portfolio simulation, I call the latter the oracle PCA+ solution. It is unachievable in reality but serves an upper bound for the PCA+ method.

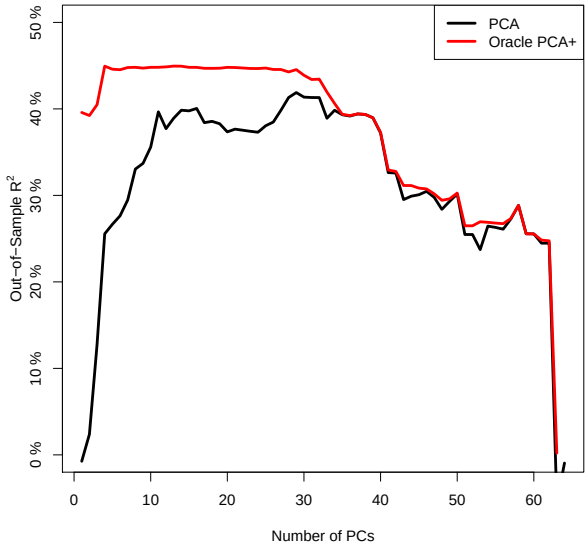
Figure 5.4 shows the out-of-sample R^2 for PCA and the oracle PCA+ solution. Same as in the portfolio case, the PCA solution needs about 30 PCs to achieve the best performance while the oracle PCA+ can achieve a similar result with just several top PCs. Also, the difference between PCA and the oracle PCA+ is biggest when k is small. Same explanation also applies here. The top PCs are extremely well estimated which means the orthogonal subspace is also a great estimation. Moreover, this subspace is also of high dimension which leads to high potential. Once again, the simulation results show the power of the forgotten orthogonality information.

5.4.3 Parameter Choices of k and η

Unlike the portfolio experiments, I use different ks for different datasets. This decision is based on the design of empirical experiments which have the same low-data settings regardless of the number of covariates, p. To have PCA+ method work well for dataset with a small p, the dimensionality of the orthogonal subspace needs to be large enough. That is to say a large $p - k$ is needed. Indeed, the choice of k grows as p grows.

[4]It is usually called the principal components regression (PCR), but for consistency, I still use the name PCA.

Figure 5.4: Out-of-Sample R^2 for PCA and oracle PCA+ Solution



I consider not using $\eta = \sqrt{\lambda/2}$ as in portfolio optimization because the structures of enhancement are different based on Eq. 5.1 and Eq. 5.3. In fact, I find that this choice of η no longer works well while $\sqrt{\lambda_k/1}$, $\sqrt{\lambda_k/1.1}$, and $\sqrt{\lambda_k/1.2}$ all generate great performance. For simplicity, I choose $\eta = \sqrt{\lambda_k}$.

With $\eta = \sqrt{\lambda_k}$, it is tempting to think what PCA+ does is a simple modification of $X'X$: all its eigenvalues are clipping to λ_k except the top k, and the new matrix is used to replace the $X'X$ matrix in the OLS solution. However, it can be proved to be not true based on Eq. 5.3.

5.4.4 Empirical Results

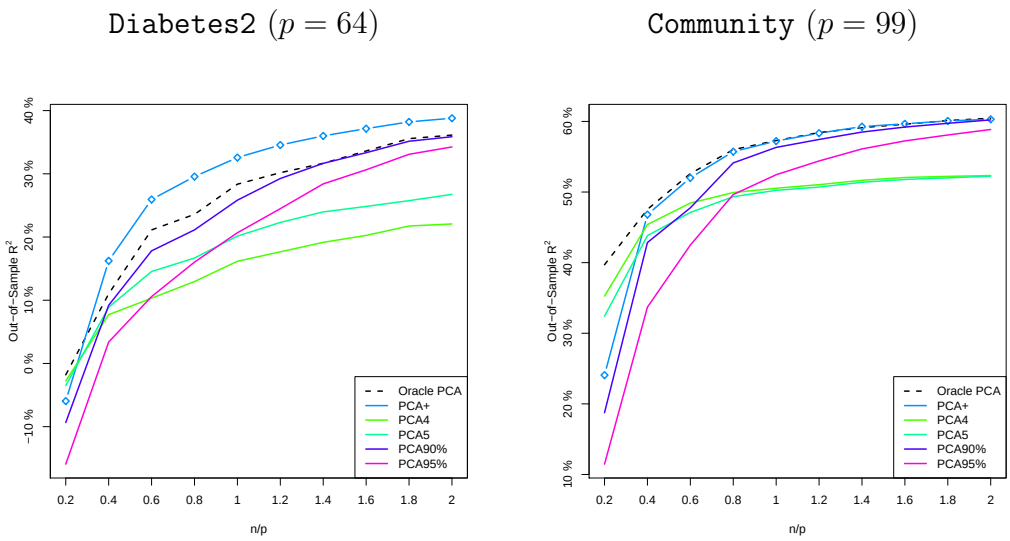
In this subsection, I will compare the PCA+ method using different ks and $\eta = \lambda_k$ with five PCA related methods defined in Section 5.3.3.

I decide to focus on dataset `Diabetes2` ($p = 64$) and dataset `Community` ($p = 99$) from Table 4.2 because they are the largest datasets. Dataset `Protein` ($p = 88$) is ignored because there are only 96 observations resulting in overfitting for the oracle PCA solution. For both datasets, I use $k = 4$.

My goal is to study the performance of PCA+ under varying levels of data insufficiency. Thus, for both datasets, I construct smaller training sets by varying the n/p ratio from 20% to 200% with 20% increments. For example, for the `Community` dataset with $p = 99$ covariates, I construct 100 training sets with $n = 20$ chosen observations, another 100 sets with $n = 40$ chosen observations, and so on. I ignore instances with $n < 10$ which is too little data for any method. Each competing method is trained on these datasets with standardization, and then its out-of-sample R^2 is measured on the remaining data points. I report the average out-of-sample R^2 over the 100 repetitions.

Figure 5.5 shows the results for all six methods. As expected, all methods improve as the n/p ratio increases. For all 20 experiments, the oracle PCA method is the best among PCA methods. It is expected since the oracle PCA method chooses the number of PCs after observing all testing data. Unlike the portfolio tests, PCA90% works well and even matches the performance of the oracle PCA on some datasets. Thus, the conventional wisdom of cutting

Figure 5.5: Out-of-Sample R^2 for Two Classic Regression Datasets.

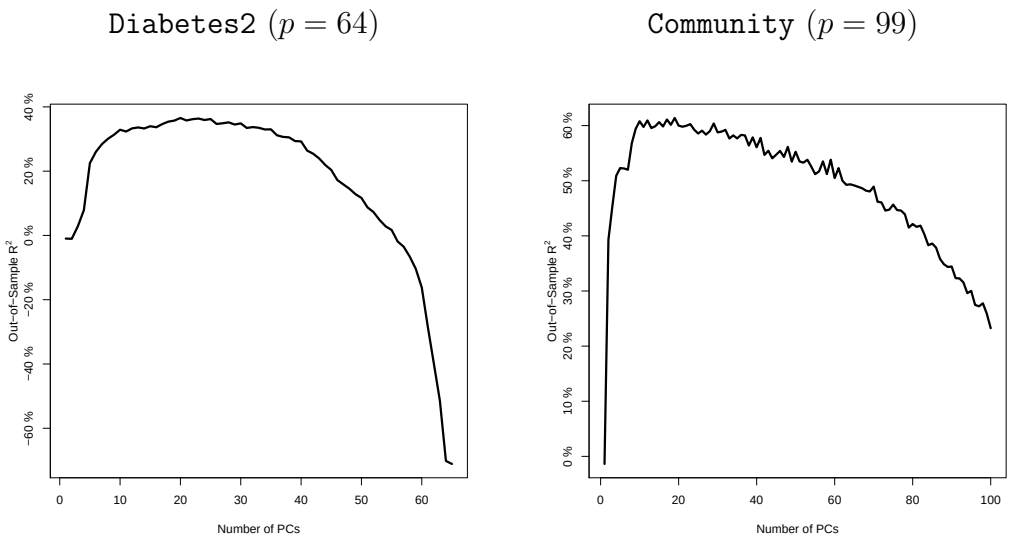


at 90% does have some merit.

The PCA+ method works extremely well on the `Diabetes2` dataset: it is even better than the oracle PCA for 9 out of 10 tests. For the `Community` dataset, most of the time, it matches the performance of the oracle PCA. However, PCA+ doesn't work well when $n/p = 20\%$. This happens because fewer than 4 PCs should be used. Indeed, PCA4 and PCA5 are also better than the PCA+ method in such a low data scenario. This motivates choosing k based on n/p ratio which requires future research. For cases with $n/p > 40\%$, the PCA+ method is much better than both PCA4 and PCA5 indicating the power of the orthogonality information.

Figure 5.6 presents the out-of-sample R^2 with respect to the number of

Figure 5.6: Out-of-Sample R^2 Regarding #PCs When $n/p = 2$



PCs selected for both datasets. The oracle PCA selects 20 PCs for `Diabetes2` and 19 PCs for `Community`. This is consistent to the previous simulation study. Meanwhile, PCA90% (PCA95%) chooses about 25 (30) and 20 (30), respectively. For the `Community` case, though PCA+ and PCA90% have similar performance, PCA+ only utilizes 1/5 of the PCs that PCA90% uses.

For dataset `Prostate` ($p = 9$), `Diabetes1` ($p = 10$), and `Supernova` ($p = 10$), I use $k = 1$ while for dataset `Crime` ($p = 15$), I use $k = 3$. The corresponding out-of-sampler R^2 is presented in Figure 5.7 and 5.8. Similar as the previous results, PCA+ can achieve at least as good as the oracle PCA.

Figure 5.7: Out-of-Sample R^2 with $k = 1$

`Prostate` ($p = 9$) `Diabetes1` ($p = 10$) `Supernova` ($p = 10$)

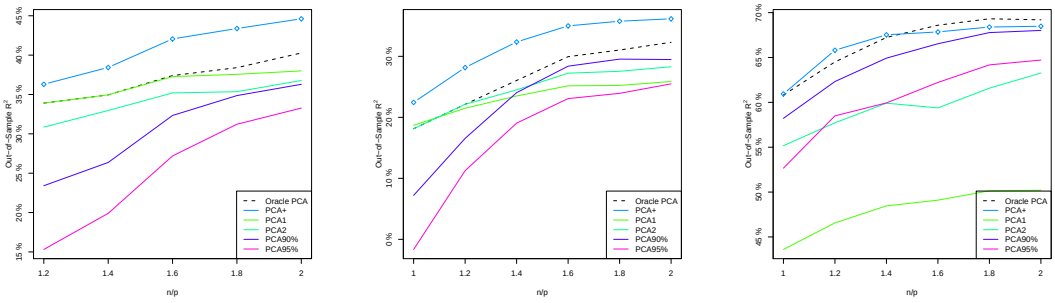
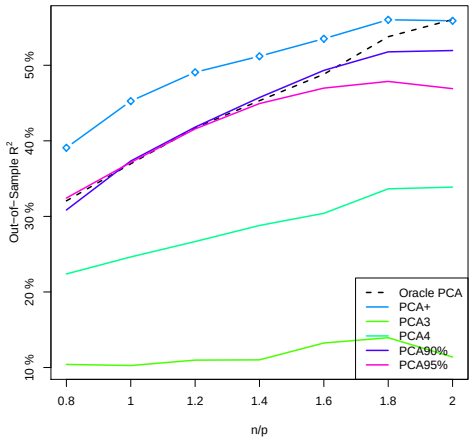


Figure 5.8: Out-of-Sample R^2 for `Crime` ($p = 15$) with $k = 3$



Chapter 6

Concluding Remarks

The essence of the thesis lies in recognizing the value of the orthogonality information: the poorly estimated part of a data matrix should not be ignored because they are orthogonal to the well-estimated. I propose two ways to use orthogonality information. The first way tries to build a robust solution from the poorly-estimated via robust optimization. The second way is to reduce dimensionality by projecting a robust solution on the poorly-estimated. They are mathematically equivalent for the minimum-variance portfolio optimization while different for the least-squares regression. Across a large number of experiments, both ways consistently improve the performance showing the importance of the orthogonality information.

For the enhancing principal component analysis part, several aspects could benefit from further investigations. First of all, extend the dimension reduction to additional applications including max-Sharpe portfolio optimization, quantile regression, and logistic regression. Secondly, for different applications, find a way to choose parameters and the robust direction endogenously. Thirdly, explore the idea in a large dataset setting. Finally, explore the possibility to use a similar approach in a dynamic programming problem

which suffers from the curse of dimensionality matters. The last one is the hardest and the most exciting direction.

Bibliography

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. Journal of machine learning research, 9(Sep):1981–2014, 2008.

Robert Andersen. Modern methods for robust regression. Sage, 2008.

Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. Journal of the American Statistical Association, 101(473):119–137, 2006.

Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. Mathematics of operations research, 23(4):769–805, 1998.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. Robust optimization. Princeton University Press, 2009.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. Operations research, 52(1):35–53, 2004.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. SIAM review, 53(3):464–501, 2011.

Åke Björck. Component-wise perturbation analysis and error bounds for linear least squares solutions. BIT Numerical Mathematics, 31(2):237–244, 1991.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.

Philip Bobko, Philip L Roth, and Maury A Buster. The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. Organizational Research Methods, 10 (4):689–709, 2007.

Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable Markowitz portfolios. Proceedings of the National Academy of Sciences, 106(30):12267–12272, 2009.

Morton B Brown and Alan B Forsythe. The ANOVA and multiple comparisons for data with heterogeneous variances. Biometrics, pages 719–724, 1974.

Deepayan Chakrabarti and Christos Faloutsos. Graph mining: Laws, generators, and algorithms. ACM Computing Surveys (CSUR), 38(1):2, 2006.

Cheng-Wen Chang, David A Laird, Maurice J Mausbach, and Charles R Hurburgh. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Science Society of America Journal, 65 (2):480–490, 2001.

Chen Chen, Garud Iyengar, and Ciamac C Moallemi. Asset-based contagion models for systemic risk. Operations Research, 2014.

Jiaqin Chen and Ming Yuan. Efficient portfolio selection in a large market. Journal of Financial Econometrics, 14(3):496–524, 2016.

William J Conover, Mark E Johnson, and Myrle M Johnson. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics, 23(4):351–361, 1981.

Christophe Croux and Gentiane Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika, 87(3):603–618, 2000.

Alexandre d'Aspremont, Laurent E Ghaoui, Michael I Jordan, and Gert R Lanckriet. A direct formulation for sparse pca using semidefinite programming. In Advances in neural information processing systems, pages 41–48, 2005.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. SIAM Journal on Numerical Analysis, 7(1):1–46, 1970.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations research, 58(3):595–612, 2010.

Victor DeMiguel and Francisco J Nogales. Portfolio selection with robust estimation. Operations Research, 57(3):560–577, 2009.

Victor DeMiguel, Lorenzo Garlappi, Francisco J Nogales, and Raman Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. Management Science, 55(5):798–812, 2009a.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naïve diversification: How inefficient is the 1/n portfolio strategy? Review of Financial Studies, 22(5):1915–1953, 2009b.

Victor DeMiguel, Alberto Martin-Utrera, and Francisco J Nogales. Size matters: Optimal calibration of shrinkage estimators for portfolio selection. Journal of Banking & Finance, 37(8):3018–3034, 2013a.

Victor DeMiguel, Yuliya Plyakha, Raman Uppal, and Grigory Vilkov. Improving portfolio selection using option-implied volatility and skewness. Journal of Financial and Quantitative Analysis, 48(6):1813–1845, 2013b.

Guy Demoment. Image reconstruction and restoration: Overview of common estimation structures and problems. IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(12):2024–2036, 1989.

Susan J Devlin, Ramanathan Gnanadesikan, and Jon R Kettenring. Robust estimation of dispersion matrices and principal components. Journal of the American Statistical Association, 76(374):354–362, 1981.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Ran Duchin and Haim Levy. Markowitz versus the Talmudic portfolio diversification strategies. Journal of Portfolio Management, 35(2):71, 2009.

Bradley Efron and Trevor Hastie. Computer age statistical inference, volume 5. Cambridge University Press, 2016.

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. SIAM Journal on matrix analysis and applications, 18(4):1035–1064, 1997.

Laurent El Ghaoui, Francois Oustry, and Hervé Lebret. Robust solutions to uncertain semidefinite programs. SIAM Journal on Optimization, 9(1): 33–52, 1998.

Lars Eldén. Perturbation theory for the least squares problem with linear equality constraints. SIAM Journal on Numerical Analysis, 17(3):338–350, 1980.

Eugene F Fama and Kenneth R French. The cross-section of expected stock returns. The Journal of Finance, 47(2):427–465, 1992.

Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. Journal of Financial Economics, 116(1):1 – 22, 2015. ISSN 0304-405X.

Jianqing Fan, Jingjin Zhang, and Ke Yu. Vast portfolio selection with gross-exposure constraints. Journal of the American Statistical Association, 107 (498):592–606, 2012.

Ricardo D Fierro and James R Bunch. Collinearity and total least squares. SIAM Journal on Matrix Analysis and Applications, 15(4):1167–1181, 1994.

Gabriel Frahm and Christoph Memmel. Dominating estimators for minimum-variance portfolios. Journal of Econometrics, 159(2):289–302, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning, volume 1. Springer series in statistics New York, 2001.

Peter A Frost and James E Savarino. An empirical Bayes approach to efficient portfolio selection. Journal of Financial and Quantitative Analysis, 21(03): 293–305, 1986.

Peter A Frost and James E Savarino. For better performance: Constrain portfolio weights. The Journal of Portfolio Management, 15(1):29–34, 1988.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. Bayesian Data Analysis. Chapman & Hall/CRC, 2003.

Gene H Golub and Charles F Van Loan. Matrix computations, volume 3. JHU Press, 2012.

Jun-ya Gotoh and Akiko Takeda. On the role of norm constraints in portfolio selection. Computational Management Science, 8(4):323–353, 2011.

Richard Green and Burton Hollifield. When will mean-variance efficient portfolios be well diversified? The Journal of Finance, 47(5):1785–1809, 1992.

P Richard Hahn and Carlos M Carvalho. Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. Journal of the American Statistical Association, 110(509):435–448, 2015.

Martin Hanke and Per Christian Hansen. Regularization methods for large-scale problems. Surv. Math. Ind, 3(4):253–315, 1993.

Peter R Hansen and Asger Lunde. Realized variance and market microstructure noise. Journal of Business & Economic Statistics, 24(2):127–161, 2006.

Desmond J Higham and Nicholas J Higham. Backward error and condition of structured linear systems. SIAM Journal on Matrix Analysis and Applications, 13(1):162–175, 1992.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970.

Philip Holmes, John L Lumley, Gahl Berkooz, and Clarence W Rowley. Turbulence, coherent structures, dynamical systems and symmetry. Cambridge university press, 2012.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6):417, 1933.

Harold Hotelling. The relations of the newer multivariate statistical methods to factor analysis. British Journal of Statistical Psychology, 10(2):69–79, 1957.

Ravi Jagannathan and Tongshu Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. The Journal of Finance, 58(4):1651–1684, 2003.

J David Jobson and Bob Korkie. Estimation for Markowitz efficient portfolios. Journal of the American Statistical Association, 75(371):544–554, 1980.

J David Jobson and Robert M Korkie. Putting Markowitz theory to work. The Journal of Portfolio Management, 7(4):70–74, 1981.

Ian Jolliffe. Principal component analysis. Springer, 2011.

Ian T Jolliffe. A note on the use of principal components in regression. Applied Statistics, pages 300–303, 1982.

Philippe Jorion. Bayes-Stein estimation for portfolio analysis. Journal of Financial and Quantitative Analysis, 21(03):279–292, 1986.

DD Kosambi. Dd kosambi, j. indian math. soc. 7, 76 (1943). J. Indian Math. Soc., 7:76, 1943.

Kenneth L Lange, Roderick JA Little, and Jeremy MG Taylor. Robust statistical modeling using the t distribution. Journal of the American Statistical Association, 84(408):881–896, 1989.

Geoffrey Jean Lauprête. Portfolio risk minimization under departures from normality. PhD thesis, Massachusetts Institute of Technology, 2001.

Oliver Ledoit and Michael Wolf. Robust performance hypothesis testing with the Sharpe ratio. Journal of Empirical Finance, 15(5):850–859, 2008.

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance, 10(5):603–621, 2003.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2): 365–411, 2004.

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics, 40(2):1024–1060, 04 2012.

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. The Review of Financial Studies, 30(12):4349–4388, 2017.

Howard Levene. Robust tests for equality of variances1. Contributions to probability and statistics: Essays in honor of Harold Hotelling, 2:278–292, 1960.

Tjen-Sien Lim and Wei-Yin Loh. A comparison of tests of equality of variances. Computational Statistics & Data Analysis, 22(3):287–301, 1996.

Robert Litterman and Jose Scheinkman. Common factors affecting bond returns. Journal of fixed income, 1(1):54–61, 1991.

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Overlapping clustering models, and one (class) svm to bind them all. In Advances in Neural Information Processing Systems, pages 2126–2136, 2018.

Harry Markowitz. Portfolio selection. The Journal of Finance, 7(1):77–91, 1952.

Richard O Michaud. The Markowitz optimization enigma: is 'optimized' optimal? Financial Analysts Journal, 45(1):31–42, 1989.

Milena Mihail and Christos Papadimitriou. On the eigenvalue power law. In International Workshop on Randomization and Approximation Techniques in Computer Science, pages 254–262. 2002.

M Nashed. Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory. IEEE Transactions on Antennas and Propagation, 29(2):220–231, 1981.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems, pages 849–856, 2002.

Alba V. Olivares-Nadal and Victor DeMiguel. Technical note: A robust perspective on transaction costs in portfolio optimization. Operations Research, 66(3):733–739, 2018.

Sung H Park. Collinearity and optimal restrictions on regression parameters for estimating responses. Technometrics, 23(3):289–295, 1981.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11):559–572, 1901.

Peter J Rousseeuw and Annick M Leroy. Robust regression and outlier detection, volume 589. John wiley & sons, 2005.

Mervyn Stone and Rodney J Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. Journal of the Royal Statistical Society: Series B (Methodological), 52(2):237–258, 1990.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

Andrey Nikolayevich Tikhonov. On the stability of inverse problems. In Dokl. Akad. Nauk SSSR, volume 39, pages 195–198, 1943.

Jun Tu and Guofu Zhou. Markowitz meets Talmud: A combination of sophisticated and naïve diversification strategies. Journal of Financial Economics, 99(1):204–215, 2011.

Sabine Van Huffel and Joos Vandewalle. The total least squares problem: computational aspects and analysis, volume 9. Siam, 1991.

Roman Vershynin. How Close is the Sample Covariance Matrix to the Actual Covariance Matrix? Journal of Theoretical Probability, 25(3):655–686, January 2011.

Roy E Welsch and Xinfeng Zhou. Application of robust statistics to asset allocation models. REVSTAT–Statistical Journal, 5(1):97–114, 2007.

Herman Wold. Estimation of principal components and related models by iterative least squares. Multivariate analysis, pages 391–420, 1966.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. In Advances in Neural Information Processing Systems, pages 1801–1808, 2009.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In Advances in Neural Information Processing Systems, pages 2496–2504, 2010.

Lei Xu and Alan L Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. IEEE Transactions on Neural Networks, 6(1):131–143, 1995.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. Biometrika, 102(2):315–323, June 2015.

Long Zhao, Deepayan Chakrabarti, and Kumar Muthuraman. Portfolio construction by mitigating error amplification: The bounded-noise portfolio. Operations Research, 2019a.

Long Zhao, Deepayan Chakrabarti, and Kumar Muthuraman. Unified classical and robust optimization for least squares. 2019b.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. Journal of computational and graphical statistics, 15(2):265–286, 2006.