MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN Xxxx-Xxxx | EISSN Xxxx-Xxxx | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Sparse Portfolio Selection via Quasi-Norm Regularization

Caihua Chen
International Center of Management Science and Engineering, School of Management and Engineering, Nanjing Univeristy, China, chchen@nju.edu.cn

Xindan Li
International Center of Management Science and Engineering, School of Management and Engineering, Nanjing University, China, xdli@nju.edu.cn

Caleb Tolman
Department of Management Science and Engineering, School of Engineering, Stanford University, USA, calebj@stanford.edu

Suyang Wang
Shanghai Stock Exchange, Shanghai, China;
International Center of Management Science and Engineering, School of Management and Engineering, Nanjing University, China, suyangwang@sse.com.cn

Yinyu Ye
Department of Management Science and Engineering, School of Engineering, Stanford University, USA; and International Center of Management Science and Engineering, School of Management and Engineering, Nanjing Univeristy, China, yyye@stanford.edu

In this paper, we propose ℓ_p-norm regularized models to seek near-optimal sparse portfolios. These sparse solutions reduce the complexity of portfolio implementation and management. Theoretical results are established to guarantee the sparsity of the second-order KKT points of the ℓ_p-norm regularized models. Unlike the existing sparse portfolio strategies such as the convex ℓ_1-norm regularized strategy, our proposed strategies have the full flexibility to choose any level of portfolio sparsity. Based on the analysis of the optimization models, we introduce a new concept "*Portfolio Risk Substitution Variance* (PRSV)"—the variance of the cost-neutral stock trades, and develop a corresponding portfolio theory which relates the sparsity of our portfolio strategy with the PSRV and the portfolio Sharpe Ratio. We also design an interior point algorithm to obtain an approximate second-order KKT solution of the ℓ_p-norm models in polynomial time with a fixed error tolerance, and then test our ℓ_p-norm models on CRSP(1992-2013) and also S&P 500 (2008-2012) data. The empirical results illustrate that our ℓ_p-norm regularized models can generate portfolios of any desired sparsity with portfolio variance, portfolio return and Sharpe Ratio comparable to those of the cardinality constrained Markowitz model. Furthermore, we also combine our sparsity ℓ_p-norm portfolio model with the ℓ_2-norm regularizers; we find that the combined model is able to produce extremely high performing portfolios that exceed the 1/N strategy, ℓ_1-norm regularized portfolio and ℓ_2-norm regularized portfolio.

Key words: Sparse portfolio management, Markowitz Model,, ℓ_p-norm regularization, PRSV, Sharpe Ratio.

1. Introduction

The origin of modern portfolio theory can be traced back to the early 1950's, beginning with Markowitz's work (Markowitz 1952) on mean-variance formulation. Given a basket of securities, the Markowitz model seeks to find the optimal asset allocation of the portfolio by minimizing the estimated variance with an expected return above a specified level.

Although the Markowitz mean-variance model captures the most two essential aspects in portfolio management—risk and return, it is not trivial to implement the model directly in the real world. One of the most critical challenge is the overfitting problem. Overfitting arises from the inability to perfectly estimate the mean and covariance of real-world objects. In fact, due to high dimensionality and non-normal distribution of the unknown variable, these estimates are especially inaccurate for stock data. Indeed, Merton (1980) shows that most of the difficulty lies on the mean estimate. Furthermore, DeMiguel et al. (2009) show that in order to estimate the expected return of portfolio of 25 stocks with satisfactorily low error, one would need on the order of 3000 months of data, which is both extremely difficult to acquire and too long for the model to obey the time-invariance assumptions. The Markowitz model does nothing to prevent the overfitting that comes from misestimation, and thus performs poorly across most out-of-sample metrics. For example, DeMiguel et al. (2009b) evaluate the out-of-sample performance of the mean-variance model and find that none of algorithms to compute the solution of the Markowitz model consistently outperforms the naive $1/N$ (equal amounts of every stock) portfolio.

To alleviate the overfitting, several variants of the Markowitz model have been proposed in the literature. In light of the difficulty of estimating investment mean and variance, many scholars study the mean-variance portfolio selection in the context of robustness; —see for example, Ben-Tal and Nemirovski (1999), Goldfarb and Iyengar (2003) and DeMiguel and Nogales (2009). More information on the robust portfolio selection can be found in the monograph Fabozzi et al. (2007b). Alternatively, Green and Hollifield (1992) and Chan et al. (1999) propose to use the factor models to estimate the covariance matrix of the return, which reduces the number of parameters to be estimated and makes the Markowitz model much more well-posed. Another approach is to impose some prior beliefs of the investors on the true yet unknown return distributions, and then to construct the corresponding posterior optimal portfolios (called Bayesian portfolios in the literature); see, Frost and Savarino (1986), Jorion (1986), Black and Litterman (1992) and references therein. Finally, a popular approach closely related to our work is to add regularizers or additional constraints to the Markowitz model in order to stabilize it. The modifications can be interpreted as adding prior beliefs on the portfolio weights. In Jagannathan and Ma (2003), the authors impose a shortsale constraint to the mean-variance formulation despite the fact that leading theory speaks against this constraint. Surprisingly, the "wrong" constraint helps the model to find solution with better out-of-sample performance. More recently, Brodie et al. (2009) and Rosenbaum and Tsybakov (2010) succeed in applying the ℓ_1-norm technique to the Markowitz model to obtain sparse portfolios with higher Sharpe Ratio than the naive $1/N$ rule. By adding a norm ball constraint to the portfolio-weight vector (equivalent to adding a convex norm regularization to the objective), DeMiguel et al. (2009) provide a general framework for determining the optimal portfolio. The empirical results of the constrained portfolios demonstrate that the norm ball constrained portfolios typically achieve lower out-of-sample variance and higher out-of-sample Sharpe Ratio than the proposed strategies in Jagannathan and Ma (2003), the naive $1/N$ portfolio and many others in the literature. Similar ideas of adding norm constraints to stabilize the optimization model has been investigated extensively both in theory and in practice in the statistics literature; see Ridge regression (Hoerl and Kennard 1970) and LASSO (Tibshirani 1996) for statistical properties on this technique.

Meanwhile, the optimal portfolio of Markowitz's classical model —especially without a shortsale constraint—often holds a huge number of assets and some assets admit extremely small weights. Such a solution, however, is not attainable in most situations of the real market. Due to physical, political, psychological and economical constraints, investors would be willing to sacrifice a small degree of portfolio performance for a more manageable sparse portfolio (see Shefrin and Statman (2000), Boyle et al. (2012), Guidolin and Rinaldi (2013) and references therein). An illustrative example comes from the most successful investor of the 20th century, Warren Buffet, who advocates investing in a few familiar stocks, which is also supported by the early work of Keynes (see Moggridge (1983)).

A popular way to construct the sparse portfolio is via the cardinality constrained portfolio selection (CCPS) model (Bertsimas and Shioda (2009), Cesarone et al. (2009), Maringer and Kellerer (2003)) , i.e., a model that considers all portfolios of a a specified number of assets and selects an efficient portfolio. Unfortunately, the inherent combinatorial property makes the cardinality constrained problem NP-hard generally and hence computationally intractable. To deal with the hard cardinality constraint, many interesting methods such as Bienstock (1996), Chang et al. (2000) have been proposed to solve the CCPS alternatively. Very recently, by relaxing the objective function as some separable functions, Gao and Li (2013) obtain a cardinality constrained relaxation of CCPS with closed-form solution. The new relaxation combined with a branch-and-bound algorithm (Bnb) yields a highly efficient solver, which outperforms CPLEX significantly.

The main objective of our paper is to propose a novel and non-CCPS portfolio strategy with *complete flexibility* in choosing sparsity while still maintaining satisfactory out-of-sample performance; through the portfolio strategy, we investigate a three-dimensional portfolio theory which integrates *sparsity* into the classical mean-variance portfolio theory. To achieve this objective, we discuss a new regularization of Markowitz's portfolio construction both with and without the short-sale constraint. Indeed, we turn to the ℓ_p-norm $(0 < p < 1)$ regularization which recently attracts a growing interest from the optimization community due to its important role in inducing sparsity. Theoretical and empirical results indicate that the ℓ_p-norm regularization (Chartrand (2007), Xu et al. (2009), Ji et al. (2013), Saab et al. (2008)) could have better stability and sparsity than the traditional ℓ_1-norm regularization under certain conditions. More recently, Fastrich et al. (2012) employ the ℓ_p-norm constrained model to identify sparse portfolios replicating a given benchmark in the index-tracking problem. Due to the nonconvexity of the problem, the authors of Fastrich et al. (2012) provide a hybrid heuristic to solve the optimization model approximately but without theoretical guarantee. Therefore, we cannot expect any theoretical properties of the resulted portfolio. Our work differs from Fastrich et al. (2012) in that our goal is to study the theoretical and computational performance of the ℓ_p-norm regularized portfolio optimization and establish a mean-variance-sparsity portfolio theory in the framework of Markowitz model, while Fastrich et al. (2012) is an empirical application of ℓ_p-norm constraint models in the area of index tracking without any theoretical analysis.

The contributions of our paper include: (i) we introduce ℓ_p-norm regularized models to seek near-optimal sparse portfolios. Unlike the existing sparse portfolio strategies such as the convex ℓ_1-norm regularized strategy, our proposed strategies have the full flexibility to choose portfolio sparsity of any level. (ii) Based on the analysis of the ℓ_p-norm model, we define a new concept ''Portfolio Risk Substitution Variance"—the variance of cost-neutral stock trades, and develop a corresponding portfolio theory which relates the sparsity of the portfolio with the PSRV. Through the study of the PRSV, we define the *relative sparsity cost index*, which allows one to identify the stock to be removed from the given portfolio in order to increase portfolio sparsity while incurring minimum utility cost; (iii) we establish a three-dimensional mean-variance-sparsity portfolio theory that relates sparsity to the PRSV and the Shapre ratio and give an "efficient frontier" outlining the optimal tradeoff between sparsity and expected return and variance; (iv) we do a comprehensive study of the ℓ_p-norm involved regularized portfolio strategies including the pure ℓ_p-norm regularized portfolio strategy, and the $\ell_2 - \ell_p$-norm regularized strategy. Empirical results illustrate that our proposed portfolio strategies (especially the $\ell_2 - \ell_p$-norm regularized model) to produce 50%–98% more sparse portfolios with competitive out-of-sample performance compared with the Markowitz model and the ℓ_1-norm model; (v) we design a polynomial time interior point algorithm to compute the second-order KKT solutions of our ℓ_p-norm models which makes our strategy implementable in practice in a relative fast time.

The remainder of this paper is organized as follows. In Section 2, we review some relevant portfolio models in the literature and present our ℓ_p-norm regularized formulations for sparse portfolio selection with/without shortsale constraints. A Bayesian explanation of the ℓ_p-norm regularization is also given in this section. In Section 3, we develop the ℓ_p-norm regularization portfolio theory with financial interpretation, and design a fast interior point algorithm to compute the KKT points of our regularized models in polynomial time. In Section 4, based on the portfolio theory, we obtain a Markowitz and PRSV heuristic to obtain our ℓ_p-norm portfolios with transaction costs explicitly considered. We construct some toy examples to show the intuition of our portfolio theory. Section 5 is devoted to the computational results of the regularized models and comparison between different models, which show our portfolio strategies have high sparsity but still maintain out-of-sample performance. Section 6 concludes our work and provides a possible application of our research. All proofs of the propositions can be found in the Appendix I and the details of our interior point algorithm are described in the Appendix II.

2. The Related Models

2.1. Models

Given a portfolio consisting of n stocks. The Markowitz mean-variance portfolio is the solution of the following constrained optimization problem:

$$\begin{aligned} \min\ & \frac{1}{2} x^T Q x \\ \text{s.t. } & e^T x = 1, \\ & m^T x \geq m_0, \end{aligned} \tag{1}$$

where $Q \in \Re^{n\times n}$ is the estimated covariance matrix of the portfolio, $m \in \Re^n$ is the estimated return vector, $m_0 \in \Re$ is a specific return level, and e is the vector of all ones with a matching dimension. Note that, if the non-shortsale constraint $x \geq 0$ is added to (1), the resulting model is the formulation of the shorting-prohibited Markowitz model. Since m_0 might not be a clear setting in practice, we limit our focus into the following portfolio optimization model closely related to the Markowitz model:

$$\begin{aligned} \min\ & \frac{1}{2} x^T Q x - \phi m^T x \\ \text{s.t. } & e^T x = 1, \\ & (x \geq 0), \end{aligned} \tag{2}$$

where the opposite of the objective is called the utility score of the portfolio x which measures the investors' satisfaction with the investment, and the scale $1/\phi$ is the risk-aversion coefficient to measure the tradeoff between the risk and return of the portfolio; investors who are risk averse would set a smaller value of ϕ. When $\phi = 0$, the model (2) reduces to the minimum-variance portfolio model which often has better out-of-sample performance than the mean-variance portfolio, see Jagannathan and Ma (2003) for an example.

For the minimum-variance portfolio model, Brodie et al. (2009) discuss the following ℓ_1-norm regularization:

$$\begin{aligned} \min\ & \frac{1}{2} x^T Q x + \lambda \|x\|_1 \\ \text{s.t. } & e^T x = 1, \\ & m^T x = m_0. \end{aligned} \tag{3}$$

Here the ℓ_1-norm of a vector $x \in \Re^n$ is defined by $\|x\|_1 := \sum_{i=1}^n |x_i|$ and λ is a positive penalty parameter. Sparse portfolios can be obtained by solving (3) with increasing values of λ. The ℓ_1-norm, however, cannot be effective in conjunction with the shortsale constraint, and thus it cannot induce sparsity beyond the sparsity of the no-shorting Markowitz portfolio. This fact can be explained as follows: let x^+ and $-x^-$ denote the positive and negative entries of x, respectively. Then, in order to satisfy the budget constraint, we must have:

$$e^T x^+ = e^T x^- + 1.$$

Since $\|x\|_1 = e^T x^+ + e^T x^-$, we also have that $\|x\|_1 = 2e^T x^- + 1$. Thus, adding $\|x\|_1$ into the objective penalizes shorting activity—the sum of the absolute negative entries in x—and thus has less effect as a penalty on sparsity.

Such a gap motivates us to study the following concave ℓ_p-norm $(0 < p < 1)$ regularization of the no-shorting mean-variance model:

$$\begin{aligned} \min\ & \frac{1}{2} x^T Q x - \phi m^T x + \lambda \|x\|_p^p \\ \text{s.t. } & e^T x = 1, \\ & x \geq 0, \end{aligned} \tag{4}$$

where the ℓ_p-norm of $x \in \Re^n$ is defined as $\|x\|_p = \sqrt[p]{\sum_{j=1}^n |x_j|^p}$. And then when $x \geq 0$, $\|x\|_p^p = \sum_{j=1}^n x_j^p$. Here, we allow a tunable parameter λ in the model to control the sparsity of the portfolios. This ℓ_p model formulation also offers us the flexibility to study the mean-variance-sparsity portfolio theory.

Another idea to achieve the sparse portfolio is via solving the following cardinality constrained portfolio selection (CCPS) problem:

$$\begin{aligned} \min\ & \frac{1}{2} x^T Q x - \phi m^T x \\ \text{s.t. } & e^T x = 1, \\ & \|x\|_0 \leq K, \\ & x \geq 0, \end{aligned} \tag{5}$$

where $\|x\|_0$ represents the number of the nonzero entries of x and K is the chosen limit of stocks to be managed in the portfolio. Since both the ℓ_p-norm regularizer and the cardinality constraint $\|x\|_0 \le K$ play the same role in sparse portfolio selection, the ℓ_p-norm regularized problem (4) can be regarded as a continuous heuristic of the CCPS problem in the sense of inducing sparsity.

We also study the portfolio selection problem with the shortsale constraint removed. As an analogue to the model (4), we consider the following ℓ_p-norm model:

$$\begin{aligned} \min\ & \frac{1}{2}\, x^T Q x - \phi m^T x + \lambda \|x\|_p^p \\ \text{s.t. } & e^T x = 1. \end{aligned} \tag{6}$$

Moreover, DeMiguel et al. (2009) construct the optimal portfolio with high Sharpe Ratio via solving the following the minimum-variance problem subject to a norm ball constraint, i.e.,

$$\begin{aligned} \min\ & \frac{1}{2}\, x^T Q x \\ \text{s.t. } & e^T x = 1, \\ & \|x\| \le \delta, \end{aligned} \tag{7}$$

where δ is a given threshold. Moreover, the ℓ_2-norm constrained portfolio will, in general, remain relatively close to the portfolio $1/N$ and achieve a higher Sharpe Ratio than ℓ_1-norm constrained portfolios. Following this work and specifying the general norm as the ℓ_2-norm, we consider the ℓ_2-norm ball constrained the ℓ_p-norm regularized Markowitz model:

$$\begin{aligned} \min\ & \tfrac{1}{2} x^T Q x - \phi m^T x + \lambda \|x\|_p^p \\ \text{s.t. } & e^T x = 1, \\ & \|x\|_2 \le \delta, \end{aligned} \tag{8}$$

or the following $\ell_2 - \ell_p$-norm double regularization Markowitz model:

$$\begin{aligned} \min\ & \tfrac{1}{2} x^T Q x - \phi m^T x + \lambda \|x\|_p^p + \mu \|x\|_2^2 \\ \text{s.t. } & e^T x = 1, \end{aligned} \tag{9}$$

which can be seen as a Lagrangian form of (8). By splitting the vector $x := x^+ - x^-$, (9) can be equivalently written as

$$\begin{aligned} \min\ & \tfrac{1}{2}(x^+ - x^-)^T Q (x^+ - x^-) - c^T (x^+ - x^-) + \lambda \|x^+\|_p^p + \lambda \|x^-\|_p^p + \mu \|x^+ - x^-\|_2^2 \\ \text{s.t. } & e^T x^+ - e^T x^- = 1, \\ & x^+ \ge 0,\ x^- \ge 0. \end{aligned} \tag{10}$$

It can be shown later that the regularization model (10) always produce a complementary pair x^+ and x^-: that is, $x_j^+ x_j^- = 0$ for all j. In this paper, we shall see whether the combined model (10) with proper choices of λ and μ will yield a portfolio strategy with moderate-high sparsity but having a high Sharpe Ratio.

2.2. Bayesian Interpretation of the ℓ_p-norm Regularization

In DeMiguel et al. (2009), the authors provide a Bayesian interpretation of the ℓ_1- and ℓ_2 (more generally, A)-norm-constrained minimum-variance portfolios. Concretely, the ℓ_1-norm constrained portfolio is the posterior evaluation of portfolio weights for an investor with a prior belief that the portfolio weights are identically distributed as a multi-dimensional Laplacian distribution while the prior belief on the distribution of the ℓ_2-norm constrained portfolio weights represents a multivariate norm distribution. Following exactly the same idea, we present the following propositions without proof, which state that our ℓ_p-norm regularized portfolios with $\phi = 0$ (corresponding to the ℓ_p-norm regularized minimum-variance model) are also the posterior estimators of portfolio weights conforming to some exponential family distributions.

PROPOSITION 1. *Assume that asset returns are normally distributed. Moreover, assume that the investor believes a priori that each of the shorting-allowed Markowitz portfolios by* (2), x_i, *is independently and identically distributed as a special distribution in the exponential family with probability density function:*

$$\pi(x_i) = c(v)e^{-v|x_i|^p}$$

Furthermore, assume that the investor believes a priori that the variance of Markowitz portfolio return, denoted by σ^2 *has an independent prior distribution* $\pi(\sigma^2)$. *Then there exists a parameter* λ *such that the weights of the* ℓ_p*-norm regularized portfolio via* (6) *with* $\phi = 0$ *are the mode of the posterior distribution of the Markowitz portfolio weights.*

PROPOSITION 2. *Assume that asset returns are normally distributed. Moreover, assume that the investor believes a prior that each of the shorting-allowed Markowitz portfolios by* (2), x_i, *is independently and identically distributed as a special distribution in the exponential family with probability density function:*

$$\pi(x_i) = \begin{cases} c(v)e^{-vx_i^p}, & x_i \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, assume that the investor believes a priori that the variance of Markowitz portfolio return, denoted by σ^2 *has an independent prior distribution* $\pi(\sigma^2)$. *Then there exists a parameter* λ *such that the weights of the* ℓ_p*-norm regularized portfolio via* (6) *with* $\phi = 0$ *are the mode of the posterior distribution of the Markowitz portfolio weights.*

There is also a Bayesian interpretation of the general ℓ_p-norm regularizations (4) and (6): the ℓ_p-norm $(0 < p < 1)$ regularizer can, roughly, be viewed as an exponential family prior belief on the optimal portfolio weights. The objective function can be seen as maximizing the negative utility score. The ℓ_p-norm penalizes the coefficients (especially around 0) more so than the ℓ_1-norm regularizer (See the graph of $|x|$ and $\sqrt{|x|}$ in Figure 1), and thus these models are expected to generate sparser portfolios than the convex ℓ_1-norm regularized (or, ball constrained) portfolios.

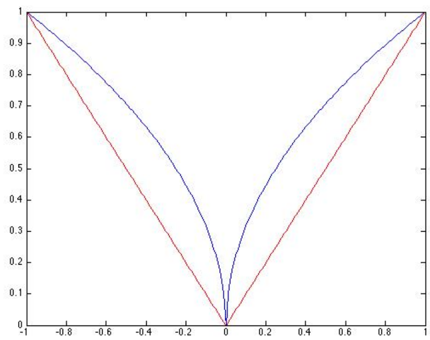

Figure 1 **Portfolio Sparsity**

3. ℓ_p-norm Regularized Portfolio Theory

In this section, we develop theoretical results on the sparsity of the ℓ_p-norm regularized models and also provide financial interpretation of the theory. Our approach to establish the theoretical results is motivated by the results (Chen et al 2010) in signal processing. For simplicity, hereafter we will fix $p = 1/2$.

3.1. Bounds of Nonzero Elements of KKT Points

First, we develop bounds on the non-zero entries of any KKT solution of the ℓ_p-norm regularized Markowitz model with the shortsale constraint $x \geq 0$.

THEOREM 1. *Let $\bar{x}$ be any second-order KKT solution of* (4)*, that is, a first-order KKT solution that also satisfies the second-order necessary condition, $\bar{P}$ be the support of $\bar{x}$, $\bar{c}$ and $\bar{Q}$ be the corresponding return sub-vector and covariances sub-matrix. Furthermore, let $K = |\bar{P}|$ and*

$$L_i = \bar{Q}_{ii} - \frac{2}{K}(\bar{Q}e)_i + \frac{1}{K^2}(e^T\bar{Q}e),\ i \in \bar{P},$$

which are the diagonal entries of the projection of $\bar{Q}$ onto the null space of vector e:

$$\left(I - \frac{1}{K}ee^T\right)\bar{Q}\left(I - \frac{1}{K}ee^T\right).$$

Then it holds that

(i)

$$(K-1)K^{3/2} \leq \frac{4\sum_{i\in\bar{P}} L_i}{\lambda} = \frac{4}{\lambda}\Big[\mathrm{tr}(\bar{Q}) - \frac{1}{K}e^T\bar{Q}e\Big].$$

(ii) *If $L_i = 0$ for some $i \in \bar{P}$, then $K = 1$ so that $\bar{x}_i = 1$; otherwise,*

$$\bar{x}_i \geq \left(\frac{\lambda(K-1)^2}{4L_iK^2}\right)^{2/3}$$

and hence

$$L_i \geq \frac{\lambda(K-1)^2}{4K^2}.$$

(iii)

$$\|(I-\frac{1}{k}ee^T)(\bar{Q}\bar{x}-\bar{c})\| \le \sqrt{K}\left(\frac{\lambda^2 \max_{i\in P} L_i K^2}{2(K-1)^2}\right)^{1/3}.$$

Note that if $\sum_{i\in\bar{P}} L_i = 0$, the first statement of our theorem implies that $K=1$. This can be explained as follows. $\sum_{i\in\bar{P}} L_i = 0$ implies the projected $\bar{Q}$ matrix

$$\left(I-\frac{1}{K}ee^T\right)\bar{Q}\left(I-\frac{1}{K}ee^T\right)=0.$$

Then, $\bar{Q}=\alpha ee^T$ for some $\alpha \ge 0$, in which case the portfolio variance $\bar{x}^T\bar{Q}\bar{x}=\alpha$ and it is a constant. Thus, the optimal solution of the regularized problem would allocate 100% into the stock with the highest c_i or highest return factor. Our theorem also implies that K is decreasing with the parameter λ. The quantity of $\sum_{i\in\bar{P}} L_i$ represents the total diversification coefficient of the set of stocks $i\in\bar{P}$: smaller quantities will result in smaller $\bar{P}$ – the set of selected stocks in the portfolio by the ℓ_p-norm regularized Markowitz model.

The second statement provides an even stronger notion: if any $L_i=0$, $i\in\bar{P}$, then $K=1$. Basically, it says that investing only into the ith stock suffices, since no diversification can help in this case. Note that L_i can be interpreted as other stocks' correlation to stock i. If $L_i=0$, then other stocks present no diversification to the ith stock.

The third statement provides a theoretical property on the marginal negative portfolio utility. Note that K is nearly proportional to $\lambda^{-2/5}$. Thus, the marginal negative portfolio utility increases with λ.

Next, we establish portfolio theories for the short-prohibited ℓ_p-norm regularized model (6) and the $\ell_2-\ell_p$ double regularized model (9). The following theorem characterize the bound of nonzero elements of any second-order KKT points of the ℓ_p-norm model without shortsale constraint ($\mu=0$) and the double regularized model ($\mu>0$).

THEOREM 2. *Let $\bar{x}=(\bar{x}^+,\bar{x}^-)$ be any second-order KKT solution of (10), $\bar{P}^+$ and $\bar{P}^-$ be the support of $\bar{x}^+$ and $\bar{x}^-$, and $\bar{P}=\bar{P}^+\cup\bar{P}^-$. Furthermore, let $\bar{c}$ and $\bar{Q}$ be the return sub-vector and covariance sub-matrices corresponding to $\bar{P}$, $K=|\bar{P}|$, and*

$$L_i=\bar{Q}_{ii}+2\mu-\frac{2}{K}(\bar{Q}e)_i+\frac{1}{(K)^2}(e^T\bar{Q}e),\ \ i\in\bar{P}.$$

Then it holds that

(i)

$$\bar{P}^+\cap\bar{P}^-=\emptyset.$$

(ii) *If $\|\bar{x}\|_2\le\delta$, then*

$$(K-1)K^{3/4}\le\frac{4\delta^{3/2}\sum_{i\in\bar{P}} L_i}{\lambda}=\frac{4\delta^{3/2}}{\lambda}\Big[\mathrm{tr}(\bar{Q})-\frac{1}{K}e^T\bar{Q}e\Big].$$

(ii) *If $L_i = 0$ for some $i \in \bar{P}$, then $K = 1$ so that $\bar{x}_i = 1$ and $i \in \bar{P}^+$; otherwise,*

$$\bar{x}_i^j \geq \left(\frac{\lambda(K-1)^2}{4L_iK^2}\right)^{2/3}, \; i \in \bar{P}.$$

If $\bar{x} \leq \delta$, then

$$L_i \geq \frac{\lambda(K-1)^2}{4K^2\delta^{\frac{3}{4}}}, \; i \in \bar{P}.$$

(iii)

$$\|(I - \frac{1}{k}ee^T)(\bar{Q}\bar{x} - \bar{c})\| \leq \sqrt{K}\left(\frac{\lambda^2 \max_{i\in P} L_i K^2}{2(K-1)^2}\right)^{1/3}.$$

The theories developed above indicate the importance of computing a second-order KKT solution, rather than just a first-order KKT solution, of the ℓ_p-norm regularized portfolio management problem (4). In this paper, we present an interior point algorithm to compute an approximate second KKT point in polynomial time with a fixed error tolerance; see details in the Appendix 1. The overall idea of using the interior-point algorithm is to start from a fully supported portfolio x that is, $x_i > 0$ of every stock i in consideration) and iteratively remove the stocks estimated to be least needed for an optimal solution.

3.2. The Portfolio Risk Substitution Variance - L_i

In the theory supporting our model (see Section 3.1), there arose several interesting facts and characteristics to note about the "Portfolio Risk Substitution Variance" — $\{L_i\}$ over the support set of a portfolio selected by the ℓ_p-norm regularized Markowitz models.

Given any stock portfolio, with the non-zero portion denoted as x, having support P of size K one can rewrite the quantity L_i in Theorem 1, as follows:

$$L_i = (e^i - e^0)^T\bar{Q}(e^i - e^0) = \mathrm{Var}\,[\eta^T(e^i - e^0)], \tag{11}$$

$e_i \in R^K$ is the vector of all zeros except 1 at the ith position and $e^0 = \frac{1}{K}e \in R^k$. Here e^i and e^0 are the respective distributions obtained by investing 100% in stock i and $\frac{1}{K}$ in each stock of the portfolio x, and η represents the random return vector. Note that L_i, $i = 1, ..., K$, is independent of the entry values of x.

The difference vector $(e^i - e^0)$ can be viewed as the "cost-neutral portfolio action" that sells an equal amount of everything in the current portfolio and uses all those funds to buy exactly one stock, stock i, within the current portfolio. Thus, L_i estimates the variance of this action. Let us now consider the feasible and optimal solutions of the Markowitz Model

$$\begin{aligned} \min\ & \tfrac{1}{2}x^TQx - \phi m^T x \\ \text{s.t.}\ & e^T x = 1, \\ & x \geq 0. \end{aligned} \tag{12}$$

For any portfolio—the non-zero portion denoted as x—one can plot the objective function of moving in a feasible exchange direction $e^i - e^0$:

$$f[x+\varepsilon(e^i-e^0)] = \frac{1}{2}[x+\varepsilon(e^i-e^0)]^T Q[x+\varepsilon(e^i-e^0)] - \phi m^T[x+\varepsilon(e^i-e^0)] \quad (13)$$
$$= f(x) + \varepsilon \,\mathrm{Cov}\,[x^T\eta, (e^i-e^0)^T\eta] - \varepsilon\phi(\bar{m}_i - \bar{m}_0) + \frac{1}{2}\varepsilon^2 L_i,$$

where $f = \frac{1}{2}x^T Q x - \phi m^T x$. We now consider which stock would increase the negative utility least when we remove it from that portfolio x. Suppose we remove stock i in the direction $e^i - e^0$, then we have a new portfolio support $P/\{i\}$ with distribution $[x]^i = x - \frac{Kx_i}{K-1}(e^i - e^0)$. Equation (13) would give us the

Marginal Costs of Sparsity (14)

$$MCS_i = -\frac{K}{K-1}x_i[\mathrm{Cov}(x^T\eta, (e^i-e^0)^T\eta) - \phi(\bar{m}_i - \bar{m}_0)] + \frac{1}{2}\Big(\frac{K}{K-1}\Big)^2 x_i^2 L_i.$$

These marginal costs are only *upper-bounds* on the true costs of sparsity. They do not consider any further improvement that could be made by re-balancing, and thus over-estimate costs.

When our current portfolio x is a near-KKT point or local minimizer, we know from the first-order conditions that the first part $[\mathrm{Cov}\,(x^T\eta, (e^i-e^0)^T\eta) - \phi(\bar{m}_i - \bar{m}_0)]$ must be near zero and thus the second order term will be a good approximation for the Marginal Cost by itself. Hence, at a near (locally) optimal portfolio x, the best candidate for removal can be found by searching for the smallest values of $x_i\sqrt{L_i}$.

Relative Sparsity Cost Index (15)

$$\mathrm{RSC}_i = x_i\sqrt{L_i}$$

Where the smallest non-zero RSC index is the cheapest (on the margin) to eliminate from x, and is likely to be the cheapest (absolutely) to remove. Thus the quantity L_i can be viewed as measures of elasticity: they indicate how sensitive the objective value is to small cost-neutral changes in x; small L_i values therefore indicate which stocks could be removed from the portfolio with lowest cost.

3.3. The Portfolio Substitution Sharpe Ratio

However, risk is not the only factor in the portfolio selection, we also care about the return that the portfolio can achieve. Therefore, we proceed to our discussion by considering the Sharpe Ratios of the new portfolios with one stock removed from the portfolio x. Suppose we remove stock i and assign the weight x_i to the other stocks with equal amount, we then have a new portfolio support $P/\{i\}$ with

$$[x]^i = x - \frac{Kx_i}{K-1}(e^i - e^0) \quad (16)$$

The Sharpe Ratio of $[x]^i$ with respect to the benchmark asset x is given by

$$S_i = \frac{E(\eta^T[x]^i - \eta^T x)}{\sqrt{\text{Var}(\eta^T[x]^i - \eta^T x)}}.$$

By substituting (16) into the above equality, we further have

$$S_i = \frac{\bar{m}_0 - \bar{m}_i}{\sqrt{L_i}}, \tag{17}$$

where $\bar{m}_i$ is the expected return of stock i, $\bar{m}_0$ is the average expected return of the portfolio x and L_i is the Portfolio Risk Substitution Variance of stock i. As the mean return divided by the standard deviation of taking the cost-neutral action $\{e^i - e^0\}$, S_i can be regarded as the Portfolio Substitution Sharpe Ratio of the ith stock, with higher S_i values indicating the stocks i with relatively higher performance.

Since all S_i's are computed with respect to the asset x, the portfolio with the largest S_i provides the best return for the same risk. In order to seek the best performance of portfolio in terms of mean deviation ratio, one would like to remove the stock having the lowest Portfolio Substitution Sharpe Ratio. According to (17), we know the natural candidates of stocks to be removed from the portfolio x are the stocks with their expected returns less than the average return and small PRSVs. This is also consistent with our discussion in Section 3.2.

4. Heuristic and Toy Examples

4.1. Near-Optimal Portfolios with Transaction Costs

We have shown that for any choice of ℓ_p-norm coefficient λ, the resulting portfolio x, with $K = \|x\|_0$ satisfies the second-order KKT condition of the regularized model. Moreover, in section 5, we see clearly that these portfolios perform near the cardinality-constrained portfolios empirically. These facts and the theory behind the Portfolio Risk Substitution Variance give us a basis to develop a near-optimality heuristic for our portfolio strategy even with the transaction costs. The heuristic servers as a complement to the interior point algorithm in pursuing the ℓ_p-norm portfolios, as it offers a simpler and faster way of approximating the desired sparse optimal portfolio especially in the case that only few stocks need to be removed.

To model the transaction costs, we use the transaction cost formula introduced in Fastrich et al. (2012), where two kinds of transaction costs — fixed and linear transaction costs are explicitly considered. The importance of the two costs varies for different type of investors: for large investors, fixed cost (which is associated with the frequency of buy and sell actions) can be neglected and the linear transaction cost induced by transaction volume is the main concern; while for small investors, the situation is nearly opposite. In (18), the third and forth term in the objective function represent fixed cost and linear transaction cost respectively :

$$\min \ \frac{1}{2}x^TQx - \phi m^T x + \underbrace{\sum_j t_j I_{(x_j \neq 0)} + l_j x_j}_{\text{Transaction Cost}} \qquad (18)$$
$$\text{s.t.} \quad e^T x = 1, \ x \geq 0,$$

where t and l are some non-negative vectors.

4.2. Markowitz and PRSV Heuristic

Consider a nonzero portfolio x with K elements. Suppose we decide to remove stock i and distribute x_i equally to the other stocks in this portfolio. Then the total cost of this action- $TS(x)_i$ for $x_i \neq 0$ is given approximately by

$$TS(x)_i \approx \frac{1}{2}\Big(\frac{K}{K-1}\Big)^2 x_i^2 L_i - t_i + (l_i - \frac{1}{K-1}\sum_{j\neq i} l_j)x_i,$$

where the first term is the marginal cost of sparsity and the second term is the incremental transaction cost by this action. With this preparation, we provide the following heuristic, which relies on computing the Markowitz model and PRSV:

Step 1: Start with the full set of stocks X;

Step 2: Compute the unregularized Markowitz portfolio x;

Step 3: Calculate L_i for each stock of current portfolio (based on this basis);

Step 4: For some fixed number n, remove from X the n stocks that minimize the $TS(x)_i$;

Step 5: Repeat Step 2-4 until the stopping criterion is satisfied.

A typical stopping criterion in Step 5 is to ask $TS(x)_i > \delta$ for all i, where δ is a threshold defined by the user. The condition means taking any cost neutral action will increase the objective function more than the threshold. Thus we may not pursue portfolios with more sparsity. To increase the speed of the heuristic, one could choose n to equal the number of stocks that do not satisfy the needed criteria $TS(x)_i > \delta$. We see from Theorem 1 that the quantities L_i are the sum of three averages (cell, row, and entire matrix) on the covariance matrix Q. Inasmuch as this matrix has small row and matrix differences, the values of the L_i will change little as the basis is reduced.

4.3. Toy Examples

4.3.1. Toy Examples A In this subsection, we illustrate the previous sensitivity analysis by some dummy examples without transaction cost. In this case, the total cost of eliminating a stock is merely the loss in expected performance: the marginal sparsity cost. Thus we shall remove the stocks with small RSC.

Consider the first example in Table 1, where the portfolios include three stocks with identical variances yet different expected returns. The lower returning stocks admit a smaller percentage in the optimal portfolio due to the large reward for the expected return ($\phi = 0.5$). Since the RSC of stock 1 is the smallest, according to our sensitive analysis, the investor would remove the first stock from the basis to form a sparse portfolio. We predict this stock would be the first to leave the basis in an ℓ_p norm penalized optimization, with the increasing of the coefficient λ. Direct calculation also shows that this is the stock of the lowest cost to remove.

Table 1 Toy example 1

$$\text{Mean}\ \begin{bmatrix}1.00001\\1.00002\\1.00003\end{bmatrix}\quad \text{Variance}\ \begin{bmatrix}0.0002 & 0.0001 & 0.0001\\0.0001 & 0.0002 & 0.0001\\0.0001 & 0.0001 & 0.0002\end{bmatrix}\quad x^*(\phi=0.5)\ \begin{bmatrix}0.2833\\0.3333\\0.3833\end{bmatrix}\quad L_i\ \begin{bmatrix}0.00007\\0.00007\\0.00007\end{bmatrix}\quad \text{OK to drop}\ \begin{bmatrix}Yes\\No\\No\end{bmatrix}\quad \text{RSC}\ \begin{bmatrix}0.000019\\0.000023\\0.000026\end{bmatrix}$$

As a second example, consider Table 2, where we have a set of stocks that include two with high variance and covariance with most other stocks, yet highly negative correlation with each other. These two stocks alone would make an excellent portfolio of size two. Here, we see that the smallest investments in the Markowitz portfolio are not necessarily the stocks to remove. To achieve the best portfolio of size 1, we include only stock 1 in the portfolio. For the portfolio of size 2, the best choice is a combination of stocks 3 and 4. The best portfolio of size 3 is the portfolio containing all the stocks but excluding stock 1. The Relative Sparsity Costs seem to hint at many of those choices.

Table 2 Toy example 2

$$\text{mean}\ \begin{bmatrix}1.0\\1.0\\1.0\\1.0\end{bmatrix}\quad \text{Variance}\ \begin{bmatrix}0.0008 & 0.0007 & 0.0006 & 0.0006\\0.0007 & 0.0026 & 0.0006 & 0.0000\\0.0006 & 0.0006 & 0.0096 & -0.0068\\0.0006 & 0.0000 & -0.0068 & 0.0073\end{bmatrix}\quad x^*(\phi=0.5)\ \begin{bmatrix}0.2913\\0.1166\\0.2714\\0.3207\end{bmatrix}\quad L_i\ \begin{bmatrix}0.000181\\0.001381\\0.008331\\0.007481\end{bmatrix}\quad \text{OK to drop}\ \begin{bmatrix}Yes\\No\\No\\No\end{bmatrix}\quad \text{RSC}\ \begin{bmatrix}0.00392\\0.00433\\0.02477\\0.02773\end{bmatrix}$$

4.3.2. Toy Examples B Next, we proceed to our empirical study on the Markowitz heuristic by an example using real-world data. Consider optimizing some balance of variance, mean and transaction costs, using the set of 8 stocks as shown below:

$$\min\ \frac{1}{2}x^TQx - \frac{1}{2}m^Tx + \frac{1}{10000}e^Tx + \frac{1}{10000}e^TI_{(x>0)} \tag{19}$$

Here, we use the Markowitz and PRSV heuristic developed in Section 4.2 to quickly find a near-optimal solution.

Table 3 Toy Example 3

$$\begin{array}{ccccc} \text{Mean} & \text{Variance} & L_i & x_i & TS(x)_i \\ \begin{bmatrix} 0.0038 \\ 0.0039 \\ 0.0037 \\ 0.0025 \\ 0.0040 \\ 0.0038 \\ 0.0038 \\ 0.0038 \end{bmatrix} & \begin{bmatrix} 0.007 & 0.001 & 0.001 & 0.000 & 0.001 & 0.001 & 0.000 & 0.001 \\ 0.001 & 0.009 & 0.001 & 0.000 & 0.002 & 0.003 & 0.000 & 0.000 \\ 0.001 & 0.001 & 0.006 & 0.000 & 0.001 & 0.001 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.003 & 0.001 & 0.001 & 0.000 & 0.000 \\ 0.001 & 0.002 & 0.001 & 0.001 & 0.007 & 0.001 & 0.000 & 0.002 \\ 0.001 & 0.003 & 0.001 & 0.001 & 0.001 & 0.008 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.020 & -0.01 \\ 0.001 & 0.000 & 0.000 & 0.000 & 0.002 & 0.000 & -0.01 & 0.017 \end{bmatrix} & \begin{bmatrix} 0.0077 \\ 0.0095 \\ 0.0069 \\ 0.0042 \\ 0.0094 \\ 0.0085 \\ 0.0209 \\ 0.0179 \end{bmatrix} & \begin{bmatrix} 0.1386 \\ 0.1374 \\ 0.0993 \\ 0.0000 \\ 0.1912 \\ 0.1031 \\ 0.1657 \\ 0.1648 \end{bmatrix} & \begin{bmatrix} 0.0007 \\ 0.0221 \\ -0.0537 \\ 0 \\ 0.1339 \\ -0.0385 \\ 0.2905 \\ 0.2308 \end{bmatrix} \times 10^{-3} \end{array}$$

Table 4 Toy Example 4

$$\begin{array}{ccccc} \text{Mean} & \text{Variance} & L_i & x & TS(x)_i \\ \begin{bmatrix} 0.0038 \\ 0.0039 \\ 0.0040 \\ 0.0038 \\ 0.0038 \end{bmatrix} & \begin{bmatrix} 0.007 & 0.001 & 0.001 & 0.000 & 0.001 \\ 0.001 & 0.009 & 0.002 & 0.000 & 0.000 \\ 0.001 & 0.002 & 0.007 & 0.000 & 0.002 \\ 0.000 & 0.000 & 0.000 & 0.020 & -0.01 \\ 0.001 & 0.000 & 0.002 & -0.01 & 0.017 \end{bmatrix} & \begin{bmatrix} 0.0082 \\ 0.0100 \\ 0.0098 \\ 0.0212 \\ 0.0182 \end{bmatrix} & \begin{bmatrix} 0.1918 \\ 0.2029 \\ 0.2195 \\ 0.1939 \\ 0.1920 \end{bmatrix} & \begin{bmatrix} 0.1357 \\ 0.2216 \\ 0.2689 \\ 0.5227 \\ 0.4242 \end{bmatrix} \times 10^{-3} \end{array}$$

Table 3 and Table 4 list the Markowitz portfolios without transaction costs, PRSVs and the total costs $TS(x)_i$ in the first and second rounds of the Markowitz and PRSV heuristic, respectively. Clearly from Table 3, we observe that stock 3, 4, and 6 take non-positive values on the $TS(x)_i$ and thus should be removed from the portfolio in the first round of the heuristic by setting $\delta = 0$ in the heuristic. The resulted portfolio is a near optimal solution with objective $-8.7e^{-4}$ near the optimal solution objective $-9.0e^{-4}$. From table 4, we see $TS(x)_i > 0$ for all i, and thus we stop removing stocks any further. Also, we see that the last two stocks perhaps could make a near-optimal portfolio of size two, as they both have a high $TS(x)_i$ (indeed they do: $-4.3e^{-4}$ versus $-4.2e^{-4}$).

5. Computational Results

5.1. Data, Parameters and Models

To examine the out-of-sample performance of ℓ_p-norm regularized models, we use historical daily returns of several exchange-traded assets to test—the list of the data used is described in Table 5.[1] The first dataset we use is similar to the dataset used in Jagannathan and Ma (2003). In each April from 1992 to 2013, we randomly select 500 stocks from all stocks in CRSP database. We also use another datasets to test the robustness of our results, the one we use is daily stock price data of S&P 500 index which spans from 31/12/2007 to 31/12/2012.[2].

[1] We choose this short time-interval due to the need for a large number of intervals and the common belief that the distribution of stock prices fundamentally change shape over decades.

[2] We do not include any company unless it is traded on the market at least 90% of the trading days during the data period, nor do any company not listed on the market for the entire timescale

Table 5 List of Datasets.

No.	Dataset	Number of stock	Time Period	Estimation Window	Prediction Window
1	500 CRSP	500	04/1992-04/2013	651 days	5 days
2	S& P	461	12/2007-12/2012	500 days	21 days

Note that to solve the ℓ_p-norm Markowitz models, proper ϕ values should be chosen. Here the inverse of ϕ is the risk-aversion parameter, the smaller ϕ values indicating the more risk averse investors. In our empirical study, we try different risk aversion parameters, such as $\phi = 0$, $\phi = 0.05$, $\phi = 0.2$, $\phi = 0.5$ and so on, which makes the risk aversion parameters range from 2 to infinite. Due to the constraint of the paper length, we only report part of results but we restrict our insights to those only valid for all experiments performed.

The rolling-window method is used to evaluate the out-of-sample performance. Different length of the estimation window and rolling window is chosen to test the robustness of our results. We use 20 rolling-windows for 500 CRSP datasets, with 651 days for estimation, 5 days for prediction and weekly rebalancing; 36 rolling-windows for S&P datasets, with 500 days for estimation, 21 days for prediction, monthly rebalancing. We use the portfolio mean, variance and Sharpe Ratio to evaluate the out-of-sample performance, where the Sharpe Ratio here is computed by the same method as in DeMiguel et al. (2009).

5.2. No-shorting Constraint Case

In DeMiguel et al. (2009b), the authors apply the ℓ_1-norm technique to seek sparse portfolios. The ℓ_1-norm, however, plays no role in the Markowitz model with no-shorting constraints. Since no-shorting environments and investors exist extensively in the real market, we turn to the ℓ_p-norm regularization to seek portfolios with desired sparsity in this situation. As we will see later, our ℓ_p-norm regularized model (4) with no-shorting constraints produces extremely sparse portfolios as compared to the already sparse Markowitz no-shorting portfolios.

The ℓ_p-norm regularized model is compared with two benchmarks in the framework of Markowitz model with shortsale constraints. The first one is the Markowitz model without regularization (λ=0) and the second is the cardinality constrained portfolio selection (CCPS) model. The global optimal cardinality-constrained portfolios are found by solving the following integer formulation of problem (5):

$$\begin{aligned} \min \ & \frac{1}{2}x^T Qx - \phi m^T x \\ \text{s.t. } & e^T x = 1, \\ & 0 \le x \le d, \\ & e^T d \le K, \\ & d \in \{0,1\}^n. \end{aligned}$$

Table 6 **Mean, Variance and Sparsity of no-shorting ℓ_p-norm regularized Model for 500 CRSP Dataset**

	$\phi = 0$			$\phi = 0.05$			$\phi = 0.2$		
	Mean	Variance	Sparsity	Mean	Variance	Sparsity	Mean	Variance	Sparsity
0	1.97e-4	2.78e-6	148.22	2.52e-3	7.70e-5	34.19	3.62e-3	3.11e-4	11.29
5.0e-8	1.84e-4	2.83e-6	66.90	2.53e-3	7.71e-5	28.82	3.62e-3	3.11e-4	10.82
1.5e-7	1.71e-4	3.04e-6	45.66	2.53e-3	7.74e-5	24.81	3.62e-3	3.11e-4	10.10
2.0e-7	1.65e-4	3.18e-6	39.59	2.53e-3	7.76e-5	23.43	3.62e-3	3.11e-4	9.89
3.5e-7	1.48e-4	3.64e-6	28.03	2.53e-3	7.81e-5	21.19	3.62e-3	3.11e-4	9.47

Table 6 reports the in-sample performance of the portfolio mean, variance and sparsity of the Markowitz portfolios with the specified ϕ ranging from 0 to 0.2. The numbers of the investing stocks in the regularized portfolios range from 9 to 148 stocks, which are about 2%-30% of the full set. The in-sample expected return of portfolios vary for different ϕ. Basically, the mean and variance increase with the increasing of ϕ, this phenomenon is consistent with the investor's higher risk tolerance with larger ϕ; we also find that when $\phi < 0.05$, increasing λ will worse the in-sample performance, but when $\phi \geq 0.05$, increasing λ will not sacrifice the in-sample performance any longer.

Moreover, we observe clearly that, for most cases especially when λ and ϕ are not large , the ℓ_p-norm regularized portfolios have means and variances comparable with the Markowitiz portfolios yet invest in fewer stocks. Thus the ℓ_p-norm regularized portfolios with proper λs outperform the Markowize portfolios, in the sense that they have higher overall return minus overall cost.

We also compare the results of ℓ_p-norm regularized model and CCPS model by testing the data of S& P, and find that the ℓ_p-norm model performs almost as well as theoretically possible (Table 10). Detailed results are reported in Appendix III.

5.2.1. Out-of-Sample Performance Brodie et al. (2009) show that sparse portfolios are often more robust and thus outperform the larger portfolios in terms of out-of-sample performance. In their analysis, the no-shorting constraint ($x \geq 0$) is taken as the *most* extreme sparsity inducing measure. We continue this investigation by taking the no-shorting constraint as the *least* extreme measure and adding the ℓ_p-norm regularizer onto the objective function. It is interesting to ask whether the sparsest portfolios will continue to outperform.

Table 7 reports the results of out-of-sample performance for the no-shorting case using CRSP dataset. In the case that λ is small, we can get return comparable to the shortsale constrained minimum-variance model ($\lambda = 0$ and $\phi = 0$) yet with higher sparsity. For example when $\lambda = 5e-8$ and $\phi = 0.05$, the Sharpe Ratio is 0.19 which is similar to the shortsale constrained minimum-variance model (can be verified by t test), yet only has 1/3 as many stocks. If we consider the transaction cost, these moderate portfolios will perform better.

Table 7 **Sharpe Ratio and Sparsity of no-shorting ℓ_p-norm regularized Model for 500 CRSP Dataset**

	$\phi=0$		$\phi=0.05$		$\phi=0.2$		$\phi=0.5$	
λ	Spa	SRatio	Spa	SRatio	Spa	SRatio	Spa	SRatio
0	148.22	0.24	34.19	0.19	11.29	0.14	5.64	0.13
t_{diff}	–	–	$(-8.58)^{***}$	(-0.75)	$(-10.85)^{***}$	(-1.19)	$(-11.42)^{***}$	(-1.26)
5.00e-8	66.90	0.20	28.82	0.19	10.82	0.14	5.56	0.13
t_{diff}	$(-5.00)^{***}$	(-1.45)	$(-9.13)^{***}$	(-0.75)	$(-10.90)^{***}$	(-1.19)	$(-11.43)^{***}$	(-1.26)
1.50e-7	45.66	0.21	24.81	0.18	10.10	0.14	5.38	0.13
t_{diff}	$(-6.90)^{***}$	$(-2.12)^{**}$	$(-9.55)^{***}$	(-0.75)	$(-10.97)^{***}$	(-1.19)	$(-11.45)^{***}$	(-1.26)
2.00e-7	39.59	0.20	23.43	0.18	9.89	0.14	5.36	0.13
t_{diff}	$(-7.51)^{***}$	$(-1.98)^{*}$	$(-9.70)^{***}$	(-0.75)	$(-10.99)^{***}$	(-1.19)	$(-11.45)^{***}$	(-1.26)
3.50e-7	28.03	0.17	21.19	0.18	9.47	0.14	5.24	0.13
t_{diff}	$(-8.70)^{***}$	$(-2.58)^{**}$	$(-9.93)^{***}$	(-0.74)	$(-11.04)^{***}$	(-1.19)	$(-11.46)^{***}$	(-1.26)

It is not surprising to see that the largest values of λ produce extremely sparse poorly performing portfolios. Yet, for some values of ϕ, we can find that the ℓ_p-norm regularized model with moderate λ also produces very sparse portfolios—although the Sharpe Ratio is smaller for lager ϕ. This result is consistent with the concept that passive investment styles outperform active ones. Also we can see that, consistent with the previous studies, the minimum-variance model (with $\phi=0$) outperforms the mean-variance model.

Moreover, a statistical t-test is run to measure the level of significance of the difference between our ℓ_p-norm regularized model and the minimum-variance model. The statistically significant difference with confidence level 10%, 5% and 1% is marked with one, two and three asterisks, respectively. Clearly from Table 7, we observe that, for all cases, the difference in sparsity is statistically significant with the confidence level 1% while the Sharpe Ratio difference is not statistically significant with the same confidence level. Even setting the confidence level as 5%, only a few cases ($\phi=0$ and $\lambda=1.5e-7$, $2.0e-7$, $3.5e-7$) have significant differences in Sharpe Ratio. This clearly presents the evidence that our ℓ_p-norm models can produce portfolios with the Sharpe Ratio comparable to that of the minimum-variance model with shortsale constraint, but with much better sparsity.

5.3. Shorting-Allowed Extension

Next we relax our constraint to allow the short-selling of stocks, and study the out-of-sample performance of the ℓ_p-norm regularized model and combined $\ell_2-\ell_p$ double regularized model.

5.3.1. ℓ_p-norm Models V.S. ℓ_1-norm Models In this subsection, we investigate and compare the empirical performances of the ℓ_p-norm and ℓ_1-norm regularized portfolios when the parameter $\phi=0$. [3] Figure 2 shows that, both the ℓ_1-norm regularizer and the ℓ_p-norm regularizer are

[3] We also run the experiments when $\phi\neq 0$; the regularized portfolios in this case often have much lower Sharpe Ratio than the minimum-variance regularized portfolios. Thus we limit our attentions the regularized portfolio with $\phi=0$ for the shorting-allowed models in order to produce highly performing portfolios in the real market.

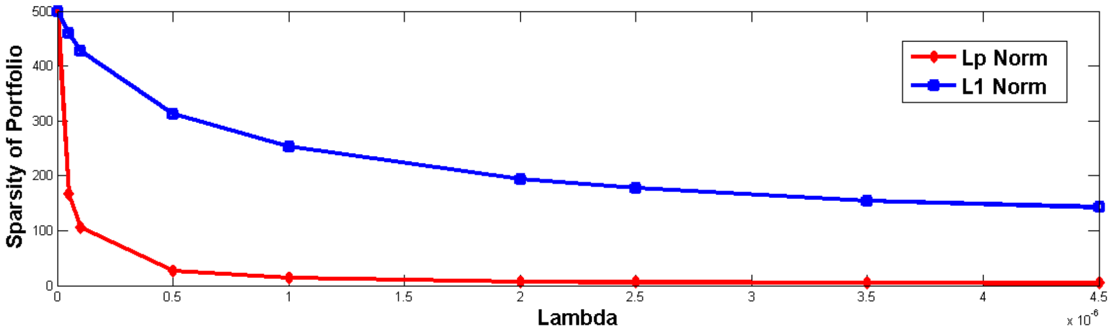

Figure 2 Portfolio Sparsity of shorting-allowed ℓ_p norm and ℓ_1 norm Model for 500 CRSP Dataset

capable of inducing sparsity in the shorting-allowed Markowitz model. Note that the parameter λ can be regarded as a server to control the portfolio sparsity. With the increasing of λ, our ℓ_p-norm regularized model (6) reduces the number of investing stocks drastically while the the number of stocks in the ℓ_1-norm regularized portfolio decreases in a much more smooth manner. For example, only 27 stocks are involved in the Markowitz regularized portfolio for $\lambda = 5e-7$, and thus there is a 95% reduction of the portfolio size. But for the ℓ_1-norm regularized model, the resulting portfolio contain more than 300 stocks which is only 40% reduction of the portfolio size.

Similar to the shorting-prohibited case, there is a clear tradeoff between the portfolio sparsity and performance out-of-sample. The Sharpe Ratio attains its maximum when λ is not large, and then decreases as the parameter λ goes up. When $\lambda \geq 2e-6$, the resulting portfolios are extremely sparse yet also have extremely poor Sharpe Ratios. However, with λ proper chosen say $\lambda \leq 1e-6$, the ℓ_p-norm regularized models produce portfolios with out-of-sample performance competitive with the Markowitz portfolios and the ℓ_1-norm regularized portfolios but with much better sparsity. This is clear from the statistical t-test [4] on the portfolio sparsity and Sharpe Ratio: when $\lambda \leq 1e-6$, the difference in the Sharpe Ratios of ℓ_1-norm regularized portfolio and ℓ_p-norm regularized portfolio is not statistically significant even with a confidence level of 10%, but the difference between sparsity is significant at the 1% level. It is clear from the results that when λ is small, our ℓ_p-norm model can achieve performance similar to the ℓ_1-norm model but with much more sparsity, and therefore our ℓ_p-norm model is expected to perform better than the ℓ_1-norm when the transaction costs are considered. Moreover, by comparing the performance of ℓ_p-norm and ℓ_1-norm model with 1/N model, we find the difference is also not significant when λ is small.

5.4. $\ell_2 - \ell_p$-norm Double Regularized Model

As mentioned in DeMiguel et al. (2009), the ℓ_2-norm constraint can be viewed as placing a prior belief that the optimal strategy is the 1/N strategy, thus it is reasonable to expect the results

[4] In order to keep the Table clear, we do not report the result of t-test in the Table as we did in no-shorting case, but choose to report the results in the paragraph.

Table 8 Sharpe Ratio and Sparsity of shorting-allowed ℓ_p and ℓ_1 Model for 500 CRSP Dataset

	$\ell_p(\phi=0)$		ℓ_1	
λ	Spa	SRatio	Spa	SRatio
0	499.24	0.18	–	–
5e-8	165.79	0.16	459.22	0.19
1e-7	105.89	0.16	427.76	0.19
5e-7	26.67	0.18	311.94	0.22
1e-6	14.08	0.15	253.39	0.22
2e-6	6.69	0.09	193.65	0.23
2.5e-6	5.55	0.03	177.32	0.23

to be close to the 1/N strategy. In the world markets today, it is common sense that the passive investment style can achieve better results than the active style. Yet, for most individual investors who aim at seeking portfolios with a comparable performance to the 1/N strategy, it is not possible to invest in a portfolio with such a huge number of stocks. Given that individual investors often dominate the less mature markets of the world, we are motivated to determine whether a sparse portfolio can embody the 1/N strategy in theory and also remain competitive in out-of-sample performance. For this purpose, it is natural to consider the ℓ_p-norm regularization of the ℓ_2-norm constrained Markowitz model:

$$\begin{aligned} \min\ & \tfrac{1}{2}x^TQx - \phi m^T x + \lambda\|x\|_p^p \\ \text{s.t. } & e^T x = 1, \\ & \|x\|_2^2 \le \delta^2, \end{aligned}$$

Here we focus on those portfolios with $\phi = 0$ [5], which has Lagrangian form (9), to see if we can obtain a portfolio that balances sparsity and uniform prior.

The out-of-sample results of the $\ell_2 - \ell_p$-norm regularized models are shown in Table 9, with different choices of λ and δ, and $\phi = 0$. From Table 9, we see that in the case $\delta = 0.075$ and $\lambda = 1e-6$, the $\ell_2 - \ell_p$ double regularization can produce the highest performing portfolio with a Sharpe Ratio much higher than the $1/N$ strategy (with a Sharpe Ratio 0.276) and a similar Sharpe Ratio to the ℓ_2-norm regularized portfolio [6] Changing δ but maintaining a small λ will worsen the Sharpe Ratio slightly, but the resulting portfolios still perform better than the 1/N strategy. This means the strategy to invest in all stocks does not always perform best out-of-sample, and it is

[5] We also run the experiments when $\phi \neq 0$, but in this case it is very often that the double regularized portfolios have lower Sharpe Ratios than the $1/N$ portfolios. To produce highly performing portfolios, we focus on the case $\phi = 0$ for the $\ell_2 - \ell_p$ double regularized models.

[6] We do a t test on the difference between our $\ell_2 - \ell_p$ portfolios and the minimum-variance portfolios and also the 1/N portfolio. For most cases, the difference of sparsity and Sharpe Ratio is statistically significant at 1% level. Also, in order to keep the table clear, we do not report the result of t-test in the table.

Table 9 Sparsity and Sharpe Ratio of the $\ell_2 - \ell_p$-Norm for 500 CRSP Dataset

		$\delta = 0.05$		$\delta = 0.075$		$\delta = 0.1$	
	λ	Spar	SRatio	Spar	SRatio	Spar	SRatio
$\phi = 0$	1.0e-7	488.8	0.324	370.9	0.382	267.9	0.327
	5.e-7	463.4	0.322	242.5	0.396	128.3	0.325
	1.0e-6	444.5	0.323	193.5	0.401	88.7	0.293
	2.0e-6	414.9	0.325	148.2	0.383	59.0	0.252
	4.5e-6	359.3	0.338	100.2	0.357	34.6	0.197

possible to obtain a portfolio (such as our $\ell_2 - \ell_p$-norm double regularized portfolio) with better performance but with sparse composition.

Clearly, the portfolios becomes more sparse with the increasing of λ and fixed δ. This trend shows a tradeoff between ℓ_p-norm regularization and ℓ_2-norm regularization. For any given δ, the small λ setting often produces high performing portfolios, suggesting that the presence of a strong uniform prior on all stocks helps mitigate overfitting due to poor variance/covariance estimates. The out-of-sample performance is *increasing* in λ when λ is not too large. These moderately sparse, highly ℓ_2-norm constricted portfolios perform excellently (nearly all have Sharpe Ratio near or above 0.35). Thus the ℓ_2 and ℓ_p norms appear to exhibit *synergy* in reducing overfitting.

6. Discussions and Conclusions

6.1. ℓ_p-norm regularized Dynamic Portfolios

A closely related application to our model is the dynamic portfolio selection. Instead of seeking a sparse portfolio, we are looking for a sparse *adjustment* to an already existing portfolio. Consider the following cardinality constrained optimization model:

$$\begin{aligned} \min\ & \tfrac{1}{2}x^T Q x - c^T x \\ \text{s.t. } & e^T x = 1, \\ & x \geq 0, \\ & \|x - a\|_0 \leq K. \end{aligned} \tag{20}$$

Here the a-vector is a feasible portfolio ($e^T a = 1$ and $a \geq 0$), representing the current state of our dynamic portfolio. Similar to the Markowitz model, the dynamic portfolio has found many applications. One is the situation where implementing the portfolio takes a significant amount of time (perhaps we must execute our orders sequentially with long delays in-between) and we wish our first orders to constitute an near-optimal portfolio. Another is the situation where our estimates Q and $c = \phi m$ are themselves varying over time, enough to warrant a re-balancing, yet we still have limits on trading—either due to transaction costs or structural limitations.

This model has a non-differentiable point in the middle of the feasible region ($x = a$), but can be reformulated (by substitution: $y = x - a$) to achieve a model very similar to the non-dynamic sparse portfolio model:

$$\begin{aligned} \min\ & \tfrac{1}{2} y^T Q y + Q a^T y - c^T y \\ \text{s.t. } & e^T y = 0 \\ & y \geq -a \\ & \|y\|_0 \leq K, \end{aligned} \tag{21}$$

We note that the objective function is still a quadratic function, and that the constraints are also of the same shape. Instead of solving the original model (21), we consider the following p norm regularized dynamic Markowitz model.

$$\begin{aligned} \min\ & \tfrac{1}{2} y^T Q y + (a^T Q - c^T) y + \lambda \|y\|_p^p \\ \text{s.t. } & e^T y = 0, \\ & y \geq -a. \end{aligned} \tag{22}$$

By letting $y = y^+ - y^-$ and using the concavity of $\|\cdot\|_p^p$, we know the regularized model (22) can be equivalently written as

$$\begin{aligned} \min\ & \tfrac{1}{2} (y^+ - y^-)^T Q (y^+ - y^-) + (a^T Q - c^T)(y^+ - y^-) + \lambda \|y^+\|_p^p + \lambda \|y^-\|_p^p \\ \text{s.t. } & e^T y^+ - e^T y^- = 0, \\ & y^+ - y^- \geq -a, \\ & y^+ \geq 0,\ y^- \geq 0, \end{aligned} \tag{23}$$

which can be further simplified to the following model

$$\begin{aligned} \min\ & \tfrac{1}{2} (y^+ - y^-)^T Q (y^+ - y^-) + (a^T Q - c^T)(y^+ - y^-) + \lambda \|y^+\|_p^p + \lambda \|y^-\|_p^p \\ \text{s.t. } & e^T y^+ - e^T y^- = 0, \\ & y^+ \geq 0,\ \ 0 \leq y^- \leq a, \end{aligned} \tag{24}$$

Similar as the non-dynamic ℓ_p- norm portfolio model, this resulting ℓ_p-norm model can also be solved by the second order interior point method.

6.2. Conclusions

In this paper, we propose ℓ_p-norm regularized models with/without shortsale constraints to seek near-optimal sparse portfolios to reduce the complexity of portfolio implementation and management. Theoretical results are established to guarantee the sparsity of the novel portfolio strategy. Computational evidence also clearly shows that the ℓ_p-norm regularized portfolio is able to choose sparsity with completely flexibility while still maintaining satisfactory out-of-sample performance—comparable to that of the cardinality-constrained portfolios.

The ℓ_2-norm can be viewed as a prior on the estimated covariances; we find that a large ℓ_2-penalty can greatly improve performance. It also could improve tractability by bounding the feasible region. And ℓ_2-norm and the ℓ_p-norm have positive cross-effects on performance—the combined model consistently portfolios outperformed all others.

Generally, when we do not pursue the most sparse portfolio, then the cost of sparsity is low—especially when the original portfolio of stocks is diverse. Our research provides a tool to evaluate the tradeoffs between sparsity and out-of-sample performance. In this framework, sparsity can be studied in relation to Sharpe-Ratio and financial theory, where both practical bounds and qualitative insights can be made.

7. Appendix

7.1. Appendix I: Proofs of the Propositions

Proof of Theorem 1. Since the second-order necessary condition of (4) holds at the point $\bar{x}$, the sub-Hessian matrix of the objective function corresponding to the indices $\bar{P}$

$$\bar{Q} - \frac{\lambda}{4}\bar{X}^{-3/2} \succeq 0$$

on the null space of e. This means the projected Hessian matrix

$$\left(I - \frac{1}{K}ee^T\right)\left(\bar{Q} - \frac{\lambda}{4}\bar{X}^{-3/2}\right)\left(I - \frac{1}{K}ee^T\right)$$

is positive semidefinite. By direct calculation, we know that the ith diagonal entry of the projected Hessian matrix is given by

$$L_i - \frac{\lambda}{4}\left((\bar{x}_i)^{-3/2}\left(1 - \frac{2}{K}\right) + \frac{\sum_{j\in\bar{P}}(\bar{x}_j)^{-3/2}}{K^2}\right) \geq 0, \tag{25}$$

and also the trace of projected Hessian matrix

$$\sum_{i\in\bar{P}} L_i - \frac{\lambda}{4}\frac{K-1}{K}\sum_{i\in\bar{P}}(\bar{x}_i)^{-3/2} \geq 0.$$

The quantity $\sum_{i\in\bar{P}}(\bar{x}_i)^{-3/2}$, with $\sum_{i\in\bar{P}}\bar{x}_i = 1$, achieves its minimum at $\bar{x}_i = 1/K$ for all $i \in \bar{P}$ with the minimum value $K\cdot K^{3/2}$. Thus,

$$\frac{\lambda}{4}(K-1)K^{3/2} \leq \sum_{i\in\bar{P}} L_i,$$

or

$$(K-1)K^{3/2} \leq \frac{4\sum_{i\in\bar{P}} L_i}{\lambda},$$

which complete the proof of the first claim. Moreover, from (25) we have

$$\frac{\lambda}{4}\left((\bar{x}_i)^{-3/2}\left(1 - \frac{2}{K}\right) + \frac{\sum_{j\in\bar{P}}(\bar{x}_j)^{-3/2}}{K^2}\right) \leq L_i.$$

Or

$$\frac{\lambda}{4}\left((\bar{x}_i)^{-3/2}\left(1 - \frac{1}{K}\right)^2 + \frac{\sum_{j\in\bar{P},j\neq i}(\bar{x}_j)^{-3/2}}{K^2}\right) \leq L_i,$$

which implies

$$\frac{\lambda}{4}(\bar{x}_i)^{-3/2}\left(1-\frac{1}{K}\right)^2 \le L_i. \tag{26}$$

Hence, if $L_i = 0$, we must have $K = 1$ so that $\bar{x}_i$ is the only non-zero entry in $\bar{x}$ and $\bar{x}_i = 1$. Otherwise, from (26), we have the desired second statement in the theorem. To prove the third assertion, we write the first order KKT condition of (4):

$$\begin{cases} (\bar{Q}\bar{x} - \bar{c} + \frac{\lambda}{2\sqrt{\bar{x}}} - \lambda_1 e = 0, \\ e^T\bar{x} = 1, \end{cases}. \tag{27}$$

where λ_1 is the Lagrangian multiplier corresponding to the equality constraint. It therefore holds that

$$\|(I - \frac{1}{k}ee^T)(\bar{Q}\bar{x} - \bar{c})\| = \|(I - \frac{1}{k}ee^T)(\frac{\lambda}{2\sqrt{\bar{x}}} - \lambda_1 e)\| \tag{28}$$
$$= \|(I - \frac{1}{k}ee^T)\frac{\lambda}{2\sqrt{\bar{x}}}\| \le \|(I - \frac{1}{k}ee^T\|_2\|\frac{\lambda}{2\sqrt{\bar{x}}}\| = \|\frac{\lambda}{2\sqrt{\bar{x}}}\|$$

which together with (26) completes the proof of this theorem.

Proof of Theorem 2. (i) Assume the contrary that $\bar{P}^+ \cap \bar{P}^- \neq \emptyset$. Then there exists an index j such that $\bar{x}_j^+ > 0$ and $\bar{x}_j^- > 0$. Let λ_1 and $\lambda_2\,(\le 0)$ be the optimal Lagrangian multiplier associated with the constraints of (10). Since (x^+, x^-) is a KKT point of (10), it holds that

$$\begin{cases} \left[(Q+\mu I)(\bar{x}^+ - \bar{x}^-)\right]_i - c_i + \frac{\lambda}{2\sqrt{(\bar{x}^+)_i}} - \lambda_1 - \lambda_2 = 0 \\ \left[(Q+\mu I)(\bar{x}^- - \bar{x}^+)\right]_i + c_i + \frac{\lambda}{2\sqrt{(\bar{x}^-)_i}} + \lambda_1 - \lambda_2 = 0 \end{cases}. \tag{29}$$

By adding the two equalities above, we have

$$\frac{\lambda}{2\sqrt{(\bar{x}^+)_i}} + \frac{\lambda}{2\sqrt{(\bar{x}^-)_i}} - 2\lambda_2 = 0. \tag{30}$$

However, since $(\bar{x}^+)_i > 0$, $(\bar{x}^-)_i > 0$ and $\lambda_2 \le 0$, the equality (30) cannot hold. This contradiction shows that $\bar{P}^+ \cap \bar{P}^- \neq \emptyset$. (ii,iii) Since the proof of the remainder parts of this theorem is similar to that of Theorem 1, we omit the details.

7.2. Appendix II: Polynomial Time Interior Point Algorithms

Most nonlinear optimization solvers can only guarantee to compute a first-order KKT solution. In this section, we extend the interior-point algorithm described in Bian et al. (2012) to solve the following generally ℓ_p-norm regularized model

$$\begin{aligned} \min\ & f(x) := \frac{1}{2}\,x^TQx - c^Tx + \lambda\|x\|_p^p \\ \text{s.t. } & Ax = b, \\ & x \ge 0, \end{aligned} \tag{31}$$

where A is a matrix in $\Re^{p\times n}$, b is a vector in $\Re^p$ and the feasible region is strictly feasible. For simplicity, we fix $p=\frac{1}{2}$.

Naturally, we would start from an interior-point feasible solution such as the analytical of the feasible set, and let the iterative algorithm to decide which entry goes to zero. This is the basic idea of affine scaling algorithm developed in Bian et al. (2012) for regularized nonconvex programming. The algorithm starts from an initial interior-point solution, then follows an interior feasible path and finally converges to either a global minimizer or a second-order KKT solution. At each step, it chooses a new interior point which produces a reduction to the objective function by an affine-scaling trust-region iteration.

Specifically, give an interior point x^k of the feasible region, the algorithm looks for an objective reduction by a update from x^k to x^{k+1}. Let d^k be a vector in $\Re^p$ satisfying $Ad^k=0$ and $x^{k+1} := x^k+d^k>0$. Using the second Taylor expansion of $f(\cdot)$, we know

$$f(x^{k+1})\approx f(x^k)+\frac{1}{2}(d^k)^T\big(Q-\frac{\lambda}{4}(X^k)^{-3/2}\big)d^k+\big(Qx^k-c+\frac{\lambda}{2\sqrt{x^k}}\big)^T d^k,$$

where $X^k=\text{Diag}(x^k)$. For given $\varepsilon\in(0,1]$, we solve the ellipsoidal trust-region constrained problem

$$\begin{aligned}\min\ & \frac{1}{2}(d^k)^T\big(Q-\frac{\lambda}{4}(X^k)^{-3/2}\big)d^k+\big(Qx^k-c+\frac{\lambda}{2\sqrt{x^k}}\big)^T d^k\\ \text{s.t. } & Ad^k=0,\\ & \|X_k^{-1}d^k\|^2\le\beta^2\varepsilon<1,\end{aligned}$$

to obtain the direction d^k. By letting $\tilde{d}^k=X_k^{-1}d^k$, we can recast the above ellipsoidal trust-region constrained problem above as a ball-constrained quadratic problem

$$\begin{aligned}\min\ & \frac{1}{2}(\tilde{d}^k)^T X^k\big(Q-\frac{\lambda}{4}(X^k)^{-3/2}\big)X^k\tilde{d}^k+\big(Qx^k-c+\frac{\lambda}{2\sqrt{x_k}}\big)^T X^k\tilde{d}^k,\\ \text{s.t. } & AX_k\tilde{d}^k=0,\\ & \|\tilde{d}^k\|^2\le\beta^2\varepsilon.\end{aligned}\tag{32}$$

Note that problem (32) can be solved efficiently even when it is nonconvex (see Bian et al. 2012).

Let $\widetilde{Q}^k=X_kQX_k-\frac{\lambda}{4}\sqrt{X^k}$ and $\tilde{c}^k=X_k(Qx^k-c)+\frac{\lambda}{2}\sqrt{x^k}$. If $\widetilde{Q}_k$ is semidefinite, the solution $\tilde{d}^k$ of problem (32) satisfies the following necessary and sufficient conditions:

$$\begin{cases}(\widetilde{Q}^k+\mu_kI)\tilde{d}^k-(AX^k)^Ty_k=-\tilde{c}^k,\\ AX^k\tilde{d}^k=0,\\ \mu_k\ge 0,\ \|\tilde{d}^k\|^2\le\beta^2\varepsilon,\ \mu_k(\|\tilde{d}^k\|^2-\beta^2\varepsilon)=0.\end{cases}\tag{33}$$

In the case that $\widetilde{Q}^k$ is indefinite, it holds that

$$\begin{cases}(\widetilde{Q}^k+\mu_kI)\tilde{d}^k-(AX^k)^Ty_k=-\tilde{c}^k,\\ AX_k\tilde{d}^k=0,\\ \mu_k\ge 0,\ N_k^T\widetilde{Q}^kN_k+\mu_kI\succeq 0,\\ \|\tilde{d}^k\|=\beta\sqrt{\varepsilon},\end{cases}\tag{34}$$

where N_k is an orthogonal basis spanning the space of $X^k A^T$.

To evaluate the performance of the affine scaling method, we need the definitions of ε scaled first-order and second-order KKT solutions. x^* is said to be an ϵ scaled first-order KKT solution of (31) if there exists a $y^* \in \Re^p$ such that

$$\begin{cases} \|X^*(Qx^* - c) + \dfrac{\lambda}{2}\sqrt{x^*} - X^* A^T y^*\| \leq \epsilon, \\ Ax^* = b, \\ x^* \geq 0. \end{cases} \tag{35}$$

Furthermore, if $X^*QX^* - \dfrac{\lambda}{4}\sqrt{X^*} + \sqrt{\epsilon} I$ is also semidefinite on the null space of $X^* A^T$, we call x^* an ϵ scaled second-order KKT solution. If $\varepsilon = 0$, the ε scaled first-order KKT solution reduces to

$$X^*(Qx^* - c) + \frac{\lambda}{2}\sqrt{x^*} - X^* A^T y^* = 0,$$

which is exactly the first-order condition of (31). In this case, the ε scaled second-order condition collapses to

$$N^T X^* Q X^* N - \frac{\lambda}{4} N^T \sqrt{X^*} N \succeq 0 \tag{36}$$

where N is an orthogonal basis spanning the space of $X^* A^T$. By direct computation, we know (36) recovers exactly the second-order optimality condition of problem (31).

For the convergence analysis of our proposed interior-point algorithm, we make the following standard assumption.

Assumption 1. *For any given $x^0 \geq 0$ such that $Ax = b$, there exists $R \geq 1$ such that*

$$\sup\{\|x\|_\infty : f(x) \leq f(x_0), Ax = b, x \geq 0\} \leq R.$$

Under the assumption above, we are able to establish the next theorem showing that the affine scaling is able to obtain either an ε-scaled second-order KKT solution or an ε global minimizer in polynomial time.

Theorem 3. *Let $\varepsilon \in (0, 1]$. There exists a positive number τ such that the proposed second-order interior point obtains either an ε scaled second-order KKT solution or ε global minimizer of* (31) *in no more than $O(\varepsilon^{-3/2})$ iterations provided that $\beta \in (0, \tau)$.*

PROOF. With loss of generality, we assume the radius $R = 1$ in the assumption. To proceed the proof of this theorem, we first introduce the following Lemma.

Lemma 1. *If $\mu_k > \lambda/6\|\tilde{d}^k\|$ holds for all $k = 0, 1, 2, \ldots$, then the second-order interior point algorithm produces an ε global minimizer of* (31) *in at most $O(\varepsilon^{-3/2})$ iterations.*

PROOF. By the Taylor expansion of $\sqrt{\cdot}$, it is easily to show that

$$f(x^{k+1}) - f(x^k) \le \frac{1}{2}\langle \tilde{d}^k, \widetilde{Q}^k \tilde{d}^k \rangle + \langle \tilde{c}^k, \tilde{d}^k \rangle + \frac{3\lambda}{48}\|\tilde{d}_k\|^3.$$

From (33) and (34), then

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\le \frac{1}{2}\langle \tilde{d}^k, \widetilde{Q}^k \tilde{d}^k \rangle + \langle -\widetilde{Q}^k \tilde{d}^k - \mu_k \tilde{d}^k + (AX_k)^T y_k, \tilde{d}^k \rangle + \frac{3\lambda}{48}\|\tilde{d}^k\|^3 \\ &= -\frac{1}{2}\tilde{d}^k \widetilde{Q}^k \tilde{d}^k - \mu_k \|\tilde{d}^k\|^2 + \frac{3\lambda}{48}\|\tilde{d}^k\|^3 \\ &= -\frac{1}{2}(v^k)^T (N_k)^T \widetilde{Q}^k N_k v^k - \mu_k \|\tilde{d}^k\|^2 + \frac{3\lambda}{48}\|\tilde{d}^k\|^3 \\ &\le \frac{\mu_k}{2}\|v^k\|^2 - \mu_k \|\tilde{d}^k\|^2 + \frac{\lambda}{48}\|\tilde{d}^k\|^3 \\ &= -\frac{\mu_k}{2}\|\tilde{d}^k\|^2 + \frac{3\lambda}{48}\|\tilde{d}^k\|^3 \\ &\le -\frac{1}{8}\mu_k \|\tilde{d}_k\|^2, \end{aligned} \tag{37}$$

where the second inequality follows from the semidefiniteness of $(N_k)^T(\widetilde{Q}^k)N_k + \mu_k I$ and the last inequality comes from the relationship that $\|\tilde{d}_k\| < 6\mu_k/\lambda$. Combining (37) with the fact that $\|\tilde{d}^k\| = \beta\sqrt{\varepsilon}$ due to $\mu_k > 0$, we further have

$$f(x^k) - f(x^0) \le -\frac{1}{8}\sum_{j=0}^{k-1} \mu_j \|\tilde{d}_j\|^2 \le -\frac{\lambda}{48} k \left(\beta^2 \varepsilon\right)^{3/2}$$

and hence the interior-point algorithm produces an ε global minimizer in $O(\varepsilon^{-\frac{3}{2}})$ iterations. □

In what follows, we analyze to the case where $\mu_k \le \lambda/6\|\tilde{d}^k\|$ for some k.

LEMMA 2. *Let $\beta \le \min\{\frac{1}{2}, \sqrt{\frac{2}{\lambda}}, \frac{3}{(18\sqrt{2}+2)\lambda}\}$. If there exists some k such that $\mu_k \le \frac{\lambda}{6}\|\tilde{d}^k\|$, then x^{k+1} is an ε second-order KKT solution of* (31).

PROOF. (i) We firstly show x^{k+1} is an ε scaled first order KKT solution when β is restricted into the special range. From (33) and (34), it follows that

$$-\mu_k \tilde{d}^k = X^k(Qx^{k+1} - c) - \frac{\lambda}{4}\sqrt{X^k}\tilde{d}^k + \frac{\lambda\sqrt{x^k}}{2} - X^k A^T y^k,$$

which implies that

$$Qx^{k+1} - c - A^T y^k = \frac{\lambda}{4}(X^k)^{-1/2}\tilde{d}^k - \frac{\lambda}{2}(x^k)^{-1/2} - \mu_k (X^k)^{-1}\tilde{d}^k.$$

Therefore, we have

$$\begin{aligned} &\|X^{k+1}(Qx^{k+1} - c) + \frac{\lambda\sqrt{x^{k+1}}}{2} - X^{k+1}A^T y^k\| \\ =\ &\|\frac{\lambda\sqrt{x^{k+1}}}{2} - \frac{\lambda}{2}X^{k+1}(x^k)^{(-1/2)} + \frac{\lambda}{4}X^{k+1}(X^k)^{-1/2}\tilde{d}^k - \mu_k X^{k+1}(X^k)^{-1}\tilde{d}^k\| \\ \le\ &\|\frac{\lambda\sqrt{x^{k+1}}}{2} - \frac{\lambda}{2}X^{k+1}(x^k)^{(-1/2)} + \frac{\lambda}{4}X^{k+1}(X^k)^{-1/2}\tilde{d}^k\| + \mu_k \|X^{k+1}(X^k)^{-1}\tilde{d}^k\| \\ \le\ &\frac{\lambda}{2}\|\sqrt{X^k}\|_\infty \|\sqrt{\tilde{d}^k + e} - e - \frac{1}{2}\tilde{d}^k + \frac{1}{2}(\tilde{d}^k)^2\| + \mu_k \|\tilde{d}^k\|(1 + \|\tilde{d}^k\|) \end{aligned} \tag{38}$$

Since the condition $\mu_j > \frac{\lambda}{6}\|\tilde{d}^j\|$ holds for $j = 0,1,2\ldots,k-1$, by the proof of Lemma 1, we have $f(x^k) \le f(x^0)$, which together with Assumption 1 implies $\|x^k\|_\infty \le 1$. Moreover, we know from the proof of Lemma 4 in Bian et al. (2012) that

$$\|\sqrt{\tilde{d}^k + e} - e - \frac{1}{2}\tilde{d}^k + \frac{1}{2}(\tilde{d}^k)^2\| \le \frac{1}{2}\|\tilde{d}^k\|^2$$

and hence

$$\begin{aligned} &\|(X^{k+1})(Qx^{k+1} - c) + \frac{\lambda\sqrt{x^{k+1}}}{2} - X^{k+1}A^T y^k\| \\ \le\ & \frac{\lambda}{4}\|\tilde{d}^k\|^2 + \frac{3}{2}\mu_k\|\tilde{d}^k\| \le \frac{\lambda}{2}\|\tilde{d}^k\|^2 \le \varepsilon, \end{aligned}$$

which means x^{k+1} is an ε scaled first-order KKT solution.

(ii) Again from (33) and (34), we know that

$$X^k Q X^k - \frac{\lambda}{4}\sqrt{X^k} + \mu_k I$$

is positive semidefinite on the null space that $X^k A^T$. Let N_k be the orthogonal basis of this null space and it therefore holds

$$N_k^T(X^k Q X^k - \frac{\lambda}{4}\sqrt{X^k})N_k \succeq -\mu_k I \succeq -\frac{\lambda}{6}\beta\sqrt{\varepsilon}I. \tag{39}$$

Clearly, $N_{k+1} := (X^{k+1})^{-1}X^k N_k$ is a basis of the null space of $X^{k+1}A^T$. By simple algebraic computation, we can easily obtain that

$$\begin{aligned} & N_{k+1}^T\left[X^{k+1}QX^{k+1} - \frac{\lambda}{4}\sqrt{X^{k+1}} + \sqrt{\varepsilon}I\right]N_{k+1} \\ =\ & N_k^T(X^k Q X^k - \frac{\lambda}{4}\sqrt{X^k})N_k + \sqrt{\varepsilon}N_k^T\left[X_{k+1}^{-2}(X^k)^2\right]N_k \\ & + \frac{\lambda}{4}N_k^T\sqrt{X^k}\left[I - (X^k)^{3/2}(X^{k+1})^{-3/2}\right]N_k \\ \succeq\ & -\tfrac{\lambda}{6}\beta\sqrt{\varepsilon}I + \sqrt{\varepsilon}N_k^T(I + D_k)^{-2}N_K + \tfrac{\lambda}{4}N_k^T\sqrt{X^k}\left[I - (I + D_k)^{-3/2}\right]N_k \end{aligned} \tag{40}$$

where $D_k = \mathrm{Diag}(\tilde{d}^k)$. Since $\|\tilde{d}^k\| \le \beta\sqrt{\varepsilon} \le \frac{1}{2} < 1$, we know

$$(I + D_k)^{-2} \succeq (1 + \beta\sqrt{\varepsilon})^{-2}I \succeq \frac{1}{4}I \tag{41}$$

and

$$I - (I + D_k)^{-3/2} \succeq \left[1 - (1 - \beta\sqrt{\varepsilon})^{-3/2}\right]I. \tag{42}$$

Moreover, the mean-value theorem applied to the function $x^{-3/2}$ yields that

$$1 - (1 - \beta\sqrt{\varepsilon})^{-3/2} = -\frac{3}{2}\beta\sqrt{\varepsilon}\theta^{-5/2},$$

where θ is in the open interval $(1-\beta\sqrt{\varepsilon}, 1)$. Note that $\beta\sqrt{\varepsilon} \le \frac{1}{2}$, then it holds that

$$1-(1-\beta\sqrt{\varepsilon})^{-3/2} \ge -6\sqrt{2}\beta\sqrt{\varepsilon}. \tag{43}$$

By substituting (41), (42) and (43) into (40), we immediately get that

$$N_{k+1}^T\left[X^{k+1}QX^{k+1}-\frac{\lambda}{4}\sqrt{X^{k+1}}+\sqrt{\varepsilon}I\right]N_{k+1} \succeq (\frac{1}{4}-\frac{3\sqrt{2}\beta\lambda}{2}-\frac{\lambda}{6}\beta)\sqrt{\varepsilon}I \succeq 0.$$

Thus x^{k+1} is an ε scaled second-order KKT solution. □

According to the above two lemmas, we know the proposed second order interior point obtains either an ε scaled second KKT solution or ε global minimizer in no more than $O(\varepsilon^{-3/2})$ iterations provided that $\beta_k \le \min\{\frac{1}{2}, \sqrt{\frac{2}{\lambda}}, \frac{3}{(18\sqrt{2}+2)\lambda}\}$. This completes the proof of this Theorem.

7.3. Appendix III: ℓ_p-norm Portfolios VS CCPS Portfolios

A comprehensive comparison of in sample computational results between our ℓ_p-norm regularized model and the cardinality constrained portfolio selection (CCPS) model using S&P dataset are reported in the Table 10. Here we use a different way to calculate proper ϕ to test the robustness of the model. Specifically, we consider the following mean-variance Markowitz model

$$\begin{aligned} \min\ & \frac{1}{2}x^TQx \\ \text{s.t. } & e^Tx=1, \\ & m^Tx \ge m_0, \\ & x \ge 0, \end{aligned} \tag{44}$$

and calculate the ϕ-values from the Lagrangian multipliers associated with the return constraint by setting reasonable values for the minimum target return m_0. As can be seen in the table, our regularized ℓ_p-norm performs almost as well as theoretically possible—the differences of the variance estimation between the two models are within 0.2% in all cases and the differences of the mean estimation are within 0.02%. Therefore, compared to the computational intractable cardinality constrained portfolio optimization, our ℓ_p-norm regularized portfolio, which can be obtained in polynomial time, performs almost as well and seeks near optimal sparse portfolios.

Figure 3 and Figure 4 show the out-of-sample portfolio returns and variances obtained by the ℓ_p-norm regularized Markowitz model with λ ranging from 5.0e-7 to 5.5e-6 and the CCPS, respectively. From Figure 3, we observe clearly that most of the plots go up slightly and then achieve its maximum, indicating that the portfolios with moderate sparsity (around 10) perform very well, even better than the Markowitz portfolio. However, with the continuously increasing of sparsity, the mean will go down dramatically and thus the regularized portfolios with extreme sparsity perform poorly in the sense of portfolio mean. Figure 4 shows that the variance of the regularized portfolios is increasing with a incremental rate with the increasing sparsity of the portfolios. However, though

Table 10 Comparison of Sparsity, Mean and Variance between the ℓ_p-norm Model and CCPS (S&P Dataset).

	ℓ_p-norm				CCPS		
	λ	Sparsity	Mean	Variance	Sparsity	Mean	Variance
$m_0 = 0.02\%$	5.0e-7	9	0.05%	4.45%	9	0.05%	4.46%
	1.0e-6	7	0.03%	3.68%	7	0.04%	3.66%
	2.0e-6	5	0.03%	3.90%	5	0.04%	3.75%
	3.5e-6	4	0.02%	4.08%	4	0.04%	3.89%
	4.5e-6	3	0.02%	4.12%	3	0.04%	4.06%
$m_0 = 0.1\%$	5.0e-7	10	0.06%	4.57%	10	0.05%	4.58%
	1.0e-6	7	0.10%	4.90%	7	0.10%	4.86%
	2.0e-6	7	0.09%	5.27%	5	0.09%	5.37%
	3.5e-6	6	0.09%	5.18%	6	0.09%	5.28%
	4.5e-6	6	0.09%	5.18%	6	0.09%	5.28%

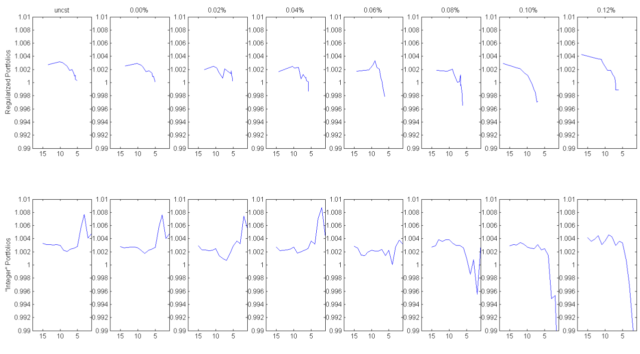

Figure 3 Portfolio Returns(S & P Dataset)

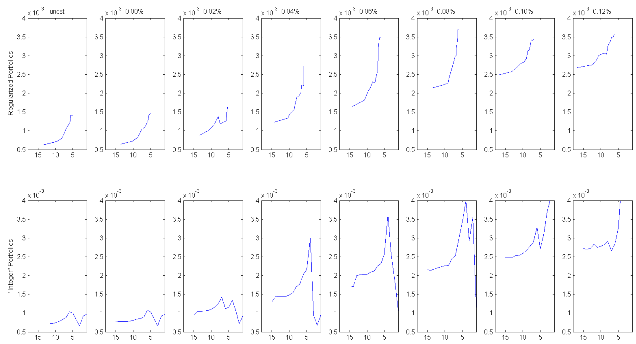

Figure 4 Portfolio Variances(S & P Dataset)

the highly sparse portfolios performs poorly in the sense of portfolio variance, the intermediate portfolios with about 10 companies suffered a 15-25% increase in variance which is also comparable to the CCPS integer portfolios.

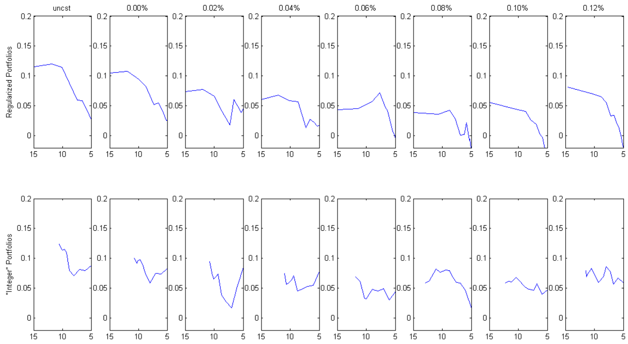

Figure 5 Portfolio Sharpe Ratios(S & P Dataset)

Figure 5 shows the out-of-performance Sharpe Ratios of our ℓ_p-norm regularized portfolio and the CCPS integer portfolio. Although the Markowitz portfolio (with $\lambda = 0$) outperforms our ℓ_p-norm regularized model in terms of the out-of-sample Sharpe Ratio, the sparse portfolios may be more implementable due to the transaction costs or logistical limitations reasons. Our results indicate that an intermediate sparse portfolio may get a comparable or at most only 10-20% cost in Sharpe Ratio while reducing more construction costs. Also, the ℓ_p-norm regularized approach is competitive with the computationally gigantic integer approach in the sense of out-of-sample performance.

References

Ben-Tal, A., A. Nemirovski. 1999. Robust solutions of uncertain linear programs. Operations research letters. *Oper. Res. Lett.* **25**(1), 1-13.

Bertsimas, D., R. Shioda. 2009. Algorithm for cardinality-constrained quadratic optimization. *Comput. Optim. Appl.* **43**(1) 1–22.

Bian, W., X. J. Chen, Y. Y Ye. 2012. Complexity analysis of interior point algorithms for non-lipschitz and nonconvex minimization. *Math. Programming*.

Bienstock, D. 1996. Computational study of a family of mixed-integer quadratic programming problems. *Math. Programming* **74**(2) 121–140.

Black, F., R. Litterman. 1992. Global portfolio optimization. *Financial Analysts Journal* **48**(5) 28–43.

Boyle, P., L. Garlappi, R. Uppal, T. Wang. 2012. Keynes meets markowitz: The trade-off between familiarity and diversification. *Management Sci.* **58**(2) 253–272.

Brodie, J., I. Daubechies, C. De Mol, D. Giannone, I. Loris. 2009. Sparse and stable markowitz portfolios. *Proc. Natl. Acad. Sci.* **106**(30) 12267–12272.

Cesarone, F., A. Scozzari, F. Tardella. 2009. Efficient algorithms for mean-variance portfolio optimization with hard real-world constraints. *Giornale dellIstituto Italiano degli Attuari* **72** 37–56.

Chan, L. K. C., J. Karceski, J. Lakonishok. 1999. On portfolios optimization: Forecasting covariance and choosing the risk Model. *Rev. Financial Stud.* **12** 937–974.

Chang, T. J., N. Meade, J. E. Beasley, Y. M. Sharaiha. 2000. Heuristics for cardinality constrained portfolio optimisation. *Computers & Operations Research* **27**(13) 1271–1302.

Chartrand, R. 2007. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Singal Proc. Let.* **14**(10) 707–710.

Chen, X. J., F. M. Xu, Y. Y. Ye. 2010. Lower Bound Theory of Nonzero Entries in Solutions of $\ell_2 - \ell_p$ Minimization. *SIAM J. Sci. Comput.* **32**(5) 2832–2852.

DeMiguel, F. J. Nogales 2009. Portfolio selection with robust estimation. *Oper. Res.* **55**(5) 798–812.

DeMiguel, V., L. Garlappi, F. J. Nogales, R. Uppal. 2009a. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Sci.* **55**(5) 798–812.

DeMiguel, V., L. Garlappi, R. Uppal. 2009b. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Rev. Financ. Stud.* **22**(5) 1915–1953.

Fabozzi, F. J., L. Garlappi, F. J. Nogales, R. Uppal. 2009a. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Sci.* **55**(5) 798–812.

Fabozzi, F. J., P. N. Kolm, D. Pachamanova, S. M. Focardi. 2007. Robust portfolio optimization and management. *John Wiley & Sons.*

Fastrich, B., S. Paterlini, P. Winker. 2012. Cardinality versus q-norm constraints for index tracking. *Quantitative Finance.* **ahead of print** 1–14.

Frost, P. A., Savarino, J. E. 1986. An empirical Bayes approach to efficient portfolio selection. Journal of Financial and Quantitative Analysis. *Jour. Fina. Quan. Anal.* **21**(3), 293-305.

Green, R., B. Hollifield. 1992. DeMiguel, F. J. Nogales 2009. When will mean-variance efficient portfolios be well diversified? *J. Finance* **47**(5) 1785–1809..

Gao, J. J., D. Li. 2013. Optimal cardinality constrained portfolio selection. *Oper. Res.* **61**(3) 745–761.

Goldfarb, D., G. Iyengar. 2003. Robust portfolio selection problems. *Math. Oper. Res.* **28**(1) 1–38.

Guidolin, M., F. Rinaldi. 2013. Ambiguity in asset pricing and portfolio choice: a review of the literature. *Theory and Decision* **74**(2) 183–217.

Hoerl, A. E., R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **8** 27–51.

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

Ji, S., K. F. Sze, Z. Zhou, A. M. C. So, Y. Y. Ye. 2013. Beyond convex relaxation: A polynomial–time non–convex optimization approach to network localization. *NFOCOM* .

Jorion. 1986. Bayes–Stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis* **21**(3) 279–292..

Maringer, D., H. Kellerer. 2003. Optimization of cardinality constrained portfolios with a hybrid local search algorithm. *Or Spectrum* **25**(4) 481–495.

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

Merton, R. C. 1980. On estimating the expected return on the market: An exploratory investigation. *J. Financ. Econ.* **8**(4) 323–361.

Moggridge, D. E. 1983. Keynes as an investor. *The Collected Works of John Maynard Keynes* **12** 1–113.

Rosenbaum, M., A. B. Tsybakov. 2010. Sparse recovery under matrix uncertainty. *Ann. Stat.* **38**(5) 2620–2651.

Saab, R., R. Chartrand, O. Yilmaz. 2008. Stable sparse approximations via nonconvex optimization. *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008.*. IEEE, 3885–3888.

Shefrin, H., M. Statman. 2000. Behavioral portfolio theory. *J. Financial Quant. Anal.* **35**(02) 127–151.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. J. Roy. Statist. Soc. B* **58** 267–288.

Xu, Z. B., H. Zhang, Y. Wang, X. Y. Chang. 2009. $\ell_{\frac{1}{2}}$ regularizer. *Science in China Series F-Inf Sci.* **52** 1–9.