Solving Norm Constrained Portfolio Optimization via Coordinate-Wise Descent Algorithms

YU-MIN YEN[a], TSO-JUNG YEN[b]

[a]*Institute of Economics, Academa Sinica, 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan.*

[b]*Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan.*

Abstract

A fast method based on coordinate-wise descent algorithms is developed to solve portfolio optimization problems in which asset weights are constrained by l_q norms for $1 \leq q \leq 2$. The method is first applied to solve a minimum variance portfolio (mvp) optimization problem in which asset weights are constrained by a weighted l_1 norm and a squared l_2 norm. Performances of the weighted norm penalized mvp are examined with two benchmark data sets. When the sample size is not large in comparison with the number of assets, the weighted norm penalized mvp tends to have a lower out-of-sample portfolio variance, lower turnover rate, fewer numbers of active constituents and shortsale positions, but higher Sharpe ratio than the one without such penalty. Several extensions of the proposed method are illustrated; in particular, an efficient algorithm for solving a portfolio optimization problem in which assets are allowed to be chosen grouply is derived.

Keywords: Minimum variance portfolio, Weighted norm constraint, Berhu penalty, Grouped portfolio selection

1. Introduction

How to select assets to form an optimal portfolio is one of the central issues in financial studies. To solve a portfolio selection problem such as the mean-variance portfolio optimization (Markowitz, 1952), we usually need to estimate mean vector and covariance matrix of the asset returns, and then plug the estimates into the optimization problem. If there are N assets, then the total number of parameters needed to be estimated for the mean vector and covariance matrix is $N + N(N+1)/2$. Accurate estimations for these

Email addresses: YMYEN@econ.sinica.edu.tw (YU-MIN YEN), tjyen@stat.sinica.edu.tw. (TSO-JUNG YEN)

parameters are necessary for successfully implementing a portfolio selection strategy; however, it is not an easy task, especially when N becomes large. If sample size n for estimating these parameters is not relatively large enough to N, cumulative estimation errors of these estimated parameters will become non-negligible, and the optimal mean-variance portfolio with these calibrations will fail to work. Empirical evidences on bad performances of the mean-variance portfolio strategy due to insufficient sample size can be found in Jagannathan and Ma (2003), DeMiguel et al. (2009a) and Kan and Zhou (2007). Kan and Smith (2008) showed that when the ratio N/n is not small enough, if simple sample estimations are used, they generally will cause an upward biasness on mean and downward biasness on variance of return of the optimal portfolio. Consequently the resulting in-sample estimation on Sharpe ratio will be too optimistic.

To reduce impacts from the estimation errors, we can choose a smaller number of assets, say $N' < N$, and at the same time optimize the objective function in the portfolio optimization. Selecting fewer assets for a portfolio means that the optimal portfolio weight vector should have some elements exactly equal to zero. Such a portfolio is termed sparse portfolio in Brodie et al. (2009). Ideally, the sparse portfolio may be obtained by solving the following l_0 norm constrained minimum variance portfolio (mvp) optimization problem:

$$\min_{\mathbf{w}} \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} \quad \text{subject to } \|\mathbf{w}\|_{l_0} \leq N', \mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1, \text{ possibly } \mathbf{w}^{\mathbf{T}}\boldsymbol{\mu} = \overline{\mu}, \tag{1}$$

where $\|\mathbf{w}\|_{l_0} = \sum_{i=1}^{N} \mathbb{I}\{w_i \neq 0\}$, and $\mathbb{I}\{A\}$ denotes the indicator function such that $\mathbb{I}\{A\} = 1$ if event A is true and $\mathbb{I}\{A\} = 0$ otherwise. The $N \times N$ matrix Σ is the covariance matrix of the N asset returns. Throughout the paper, we assume Σ is symmetric and positive semidefinite (psd). In practice, Σ can be any kind of psd covariance matrix estimation. The $N \times 1$ vector $\boldsymbol{\mu} = (\mu_{1,\ldots,}\mu_N)^{\mathbf{T}}$ is the vector of expected asset returns, and $\overline{\mu}$ is the investor's required return.

Solving (1) involves combinatorial optimization, and it becomes extremely difficult when N is large. Practically we can replace the l_0 norm $\|\mathbf{w}\|_{l_0}$ with the l_1 norm $\|\mathbf{w}\|_{l_1} := \Sigma_{i=1}^{N}|w_i|$ in (1). The l_1 norm constraint can facilitate zero components (sparsity) in the weight vector, and hence it can function as the l_0 norm constraint to restrict the number of assets in the portfolio. The l_1 norm constraint also is a convex function of $\mathbf{w}$, and such convex relaxation makes the modified portfolio optimization problem easily tractable even when N becomes very large.

In statistics, the l_1 norm penalty approach is frequently used in dealing with high dimensional estimation problems. For example, Tibshirani (1996) proposed the lasso, which aims to regularize OLS estimation with the l_1 norm penalty on regression coefficients. The l_1 norm penalty also has been widely used on estimating structures of networks (Meinshausen and Buhlmann, 2006; Vinciotti and Hashem, 2013). For portfolio optimization problems, imposing the l_1 norm constraint can improve portfolio performances when the number of assets becomes very large (Brodie et al., 2009; DeMiguel et al., 2009a; Fan et al., 2012; Welsch and Zhou, 2007). In addition, the portfolio optimization may be affected by changes of the estimated parameters, and imposing the l_1 norm constraint can help to mitigate the effects and stabilize the optimization (Brodie et al., 2009; Fan et al., 2012).

In this paper, we develop fast and easy-to-implement coordinate-wise descent algorithms to solve the norm constrained portfolio optimization problems. We first focus on an algorithm for solving the optimal minimum variance portfolio (mvp) with a weighted l_1 and squared l_2 norm penalty and linear constraints, and later we will show the proposed method can be extended to mvp optimization problems with various norm constraints. The algorithms previously used to solve such norm constrained portfolio optimization problems are either quadratic programming or the least angle regression (LARS) type algorithms (Efron et al., 2004). Recently, the coordinate-wise descent algorithms have been shown to be powerful tools for solving large dimensional variable selection problems in which the norm penalties are imposed on covariate coefficients (Friedman et al., 2007). We demonstrate that the coordinate-wise descent algorithms can also be used to solve various norm constrained portfolio optimization problems.

The norm constraints also have been adopted on the index tracking problem in which a portfolio is formulated to replicate a market index. For example, Giamouridis and Paterlini (2010) used the l_1 norm constraint on the problem and showed that it results in better out-of-sample tracking performances. Gotoh and Takeda (2011) discussed the relations between norm constraints and robust portfolio optimization problems and then applied an approach of norm constrained conditional value-at-risk (CVaR) portfolio optimization to the index tracking problem. Fastrich et al. (2013) proposed to use the nonconvex l_q norm penalty, $0 < q < 1$ on the index tracking problem. In Takeda et al. (2013), they proposed to use both the l_0 and squared l_2 norm penalties on tracking a stock index. The authors developed a greedy algorithm to solve the penalized portfolio optimization and then applied their method on tracking Nikkei 225 index.

In addition to the norm constraint approach, we can assign asset weights with some simple rules in order to avoid massive estimations. The value weighted and equally weighted ($1/N$) portfolios are such examples. DeMiguel et al. (2009b) showed how such simple strategies can outperform more sophisticated strategies. Another frequently used way is to construct more robust statistical estimators for the mean vector and covariance matrix of the asset returns, such as bias-adjusted or Bayesian shrinkage estimators (El Karoui, 2009; Jorion, 1986; Kan and Zhou, 2007; Ledoit and Wolf, 2003; Lai, Xing, and Chen, 2011), and use them in the portfolio optimization problems. We also can combine the improved portfolios to form a new portfolio; for example Frahm and Christoph (2010) and Tu and Zhou (2011) showed that a suitable linear combination of weights of a benchmark portfolio and a more sophisticated strategy often provides better performances than either only one of them is considered. It is natural to incorporate the latter two approaches with the norm constraint strategy.

The rest of the paper is organized as follows. In Section 2, we introduce a benchmark case of the mvp optimization in which the asset weights are constrained by the weighted l_1 and squared l_2 norm. We then describe a coordinate-wise descent algorithm for solving the benchmark case in Section 3. In Section 4, we use real data sets to examine empirical properties of the weighted norm mvp. In Section 5, we discuss some extensions, which include portfolio optimization problems with different convex norm penalties, and possible ways to use our method in portfolio optimization when nonconvex norm constraints are imposed (e.g., Fastrich et al., 2012) or even more general objective functions are considered. We also have a brief discussion on limitations of our method. Section 6 is conclusion.

2. Weighted Norm Minimum Variance Portfolio

To begin our analysis, we first consider the minimum variance portfolio (mvp) optimization in which the asset weights are constrained by the weighted l_1 and squared l_2 norm constraint:

$$\min_{\mathbf{w}} \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} \qquad \text{subject to } \alpha \left\|\mathbf{w}\right\|_{l_1} + (1-\alpha)\left\|\mathbf{w}\right\|_{l_2}^2 \leq c \text{ and } \mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1,$$

where $\|\mathbf{w}\|_{l_1} = \sum_{i=1}^N |w_i|$, $\|\mathbf{w}\|_{l_2}^2 = \sum_{i=1}^N w_i^2$, $c > 0$ is some constant and $\alpha \in [0,1]$ is a parameter for adjusting the relative weight of the l_1 and squared l_2 norms. The constraint $\mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1$ is the full investment constraint, and 1 is normalized investor's wealth which can be invested on the assets. The mvp optimization given above can be rewritten as

$$\min_{\mathbf{w}} \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} + \lambda\left(\alpha\left\|\mathbf{w}\right\|_{l_1} + (1-\alpha)\left\|\mathbf{w}\right\|_{l_2}^2\right) \qquad \text{subject to } \mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1, \tag{2}$$

where $\lambda \in \mathbb{R}^+$ is called the penalty parameter. That is, we minimize the portfolio variance plus a penalty function on the portfolio weights, subject to the full investment constraint. In the following, we will call the optimal portfolio obtained from solving (2) with and without such weighted norm constraint as the weighted norm mvp and unconstrained mvp respectively.

As mentioned, since the l_1 norm constraint can facilitate sparsity on the portfolio weight vector, it often causes the number of assets with non zero optimal weights fewer than the total number of assets N in the portfolio optimization. In order to simultaneously satisfy the full investment constraint, some of the asset weights may become relatively large to the others, and the problems of under diversifications and extreme portfolio weights will arise. The squared l_2 norm added here can be useful in mitigating the problems, since the squared l_2 norm does not cause any further sparsity and can also efficiently regularize the size of the weight vector.

In statistics, the penalty $\alpha \left\|\mathbf{w}\right\|_{l_1} + (1-\alpha)\left\|\mathbf{w}\right\|_{l_2}^2$ is called the elastic net constraint (Zou and Hastie, 2005) in regression based variable selection problems. In the portfolio optimization problems (2), when $\alpha = 0$, the asset weights are regularized by the squared l_2 norm $\left\|\mathbf{w}\right\|_{l_2}^2$ only, and the solution of (2) is the same as that of the unconstrained mvp optimization with Σ replaced by a regularized covariance matrix $\Sigma' = \Sigma + \lambda \mathbf{I}_{N\times N}$. In this situation, as λ goes to infinity, the optimal weights will converge to 1/N (DeMiguel et al., 2009a). Also when $\alpha = 0$, if we set $\lambda = (1-a)/a$, where $a \in (0,1)$, problem (2) becomes similar to the one that uses the Ledoit-Wolf type covariance matrix (Ledoit and Wolf, 2004) as the calibrated covariance matrix. When $\alpha = 1$, the asset weights are regularized by the l_1 norm $\left\|\mathbf{w}\right\|_{l_1}$ only. When $0 < \alpha < 1$, the solution to (2) is equivalent to the solution of the l_1 norm constrained mvp problem with the regularized covariance matrix Σ' and penalty parameter $\lambda\alpha$.

3. The Algorithm

In this section we derive a coordinate-wise descent algorithm to solve the weighted norm mvp optimization problem stated in (2), which, essentially, is the same as

$$\min_{\mathbf{w}} \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} + g(\mathbf{w}) + \lambda\left(\alpha\left\|\mathbf{w}\right\|_{l_1} + (1-\alpha)\left\|\mathbf{w}\right\|_{l_2}^2\right), \tag{3}$$

where

$$g(\mathbf{w}) = \begin{cases} 0 & \text{if } \mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1 \\ \infty & \text{otherwise.} \end{cases} \tag{4}$$

Friedman et al. (2007) demonstrated that coordinate-wise descent algorithms can be powerful tools in solving regression problems regularized by convex constraints. Let $f(\mathbf{w}) = f(w_1, \ldots, w_N)$ be an objective function and assume $f(\mathbf{w})$ is convex in $\mathbf{w}$. The coordinate-wise descent algorithm starts by fixing w_i for $i = 2, \ldots, N$ and then finds a value for w_1 to minimize $f(\mathbf{w})$. The iteration step is then carried out over $i = 2, 3, \cdots, N$ before going back to start again for $i = 1$, and the procedure is repeated until $\mathbf{w}$ has converged.

Suppose the objective function has the following form:

$$f(\mathbf{w}) = f_0(\mathbf{w}) + \sum_{i=1}^{N} f_i(w_i).$$

Tseng (2001) showed that minimization of $f(\mathbf{w})$ can be achieved via coordinate-wise descent algorithms if some regularity conditions hold for $f_0(\mathbf{w})$, and $f_i(w_i)$ is additively separable for $i = 1, 2, \cdots, N$. Now consider (3) and (4) and let $f_0(\mathbf{w}) = \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} + g(\mathbf{w})$ and $f_i(w_i) = \lambda\alpha|w_i| + \lambda(1-\alpha)w_i^2$, $i = 1, \ldots, N$. Note that $\sum_{i=1}^{N} f_i(w_i)$ and $g(\mathbf{w})$ are convex functions of $\mathbf{w}$ and $\sum_{i=1}^{N} f_i(w_i)$ satisfies the additive separability. In addition, it can be shown that the objective function $\mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} + g(\mathbf{w})$ also satisfies the sufficient conditions which are needed for applying Theorem 5.1 of Tseng (2001). Therefore solving (3) with a suitable coordinate-wise descent algorithm can yield the global minimum.

3.1. Updating the Weight Vector via the Coordinate-Wise Descent Algorithm

To solve (2), we apply the method of Lagrange multipliers. Let γ be the Lagrange multiplier of the full investment constraint $\mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1$. Our strategy is to fix γ first, applying the algorithm to update $\mathbf{w}$ only, and then using the updated $\mathbf{w}$ to update γ via the full investment constraint.

The Lagrangian corresponding to the optimization problem stated in (2) is

$$\begin{aligned} L(\mathbf{w}, \gamma; \Sigma, \lambda, \alpha) &= \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} + \lambda\alpha\|\mathbf{w}\|_{l_1} + \lambda(1-\alpha)\|\mathbf{w}\|_{l_2}^2 - \gamma(\mathbf{w}^{\mathbf{T}}\mathbf{1}_N - 1) \\ &= \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} + \sum_{i=1}^{N}(\lambda\alpha|w_i| + \lambda(1-\alpha)w_i^2 - \gamma w_i) + \gamma. \qquad (5) \end{aligned}$$

Let σ_{ij} denote the (i,j)th off-diagonal term of Σ, $i, j = 1, \ldots, N$, and $i \neq j$, and $\sigma_{ii} = \sigma_i^2$ denote the ith diagonal term, $i = 1, \ldots, N$. The KKT conditions for the Lagrangian (5) are

$$
\begin{aligned}
2w_i\sigma_i^2 + 2\sum_{j\neq i}^{N} w_j\sigma_{ij} + 2\lambda\left(1-\alpha\right)w_i - \gamma &= -\lambda\alpha \text{ if } w_i > 0,\\
2w_i\sigma_i^2 + 2\sum_{j\neq i}^{N} w_j\sigma_{ij} + 2\lambda\left(1-\alpha\right)w_i - \gamma &= \lambda\alpha \quad \text{if } w_i < 0,\\
\left|2\sum_{j\neq i}^{N} w_j\sigma_{ij} - \gamma\right| &\leq \lambda\alpha \quad \text{if } w_i = 0,\\
\mathbf{w}^{\mathbf{T}}\mathbf{1}_N &= 1.
\end{aligned}
\tag{6}
$$

Now assume that γ is fixed. In our algorithm, we use the following formula to update each w_i:

$$
w_i \leftarrow \frac{ST\left(\gamma - 2\sum_{j\neq i}^{N} w_j\sigma_{ij}, \lambda\alpha\right)}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}. \tag{7}
$$

Here $ST\left(x,y\right) = sign\left(x\right)\max(|x|-y,0)$ is the soft thresholding function. We can obtain the updating scheme (7) by solving the KKT conditions (6) with respect to w_i, given that all other parameters γ, Σ, λ, α and w_j, $j = 1, \cdots, N$, $j \neq i$ are fixed.

However, the updated portfolio weight (7) may not satisfy the full investment constraint $\mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1$. When solving the mvp optimization problem without the penalty ($\lambda = 0$), we can adjust the value of γ to make the constraint $\mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1$ hold. For the mvp optimization with the penalty ($\lambda > 0$), our strategy for updating γ is through the same way. To see this, let $z_i = 2\sum_{j\neq i}^{N} w_j\sigma_{ij}$. From the KKT conditions (6), it can be shown that when $w_i \neq 0$, at the stationary point

$$
\begin{aligned}
w_i &= \frac{\gamma - z_i - \lambda\alpha}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}, \text{ if } w_i > 0,\\
w_i &= \frac{\gamma - z_i + \lambda\alpha}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}, \text{ if } w_i < 0.
\end{aligned}
$$

Further let $S_+ = \{i : w_i > 0\}$ and $S_- = \{i : w_i < 0\}$. Then

$$
\begin{aligned}
\mathbf{w}^{\mathbf{T}}\mathbf{1}_N = \gamma\left(\sum_{i\in S_+\cup S_-}\frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right) - \sum_{i\in S_+\cup S_-}\frac{z_i}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} + \\
\lambda\alpha\left(\sum_{i\in S_-}\frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} - \sum_{i\in S_+}\frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right).
\end{aligned}
$$

We can solve for γ by using $\mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1$, and the proposed updating form for γ is then given by:

$$\begin{aligned}\gamma \;\leftarrow\; & \left[\sum_{i\in S_+\cup S_-}\frac{1}{2\left(\sigma_i^2+\lambda\left(1-\alpha\right)\right)}\right]^{-1}\times\\ & \left[1+\sum_{i\in S_+\cup S_-}\frac{z_i}{2\left(\sigma_i^2+\lambda\left(1-\alpha\right)\right)}-\right.\\ & \left.\lambda\alpha\left(\sum_{i\in S_-}\frac{1}{2\left(\sigma_i^2+\lambda\left(1-\alpha\right)\right)}-\sum_{i\in S_+}\frac{1}{2\left(\sigma_i^2+\lambda\left(1-\alpha\right)\right)}\right)\right].\end{aligned}$$

To implement the algorithm, we set the initial value of each weight $w_1^{(0)} = w_2^{(0)} = \cdots = w_p^{(0)} = N^{-1}$, and $\gamma^{(0)} > \lambda$. The algorithm starts from updating w_1, $w_2, \ldots,$ and w_N sequentially, and then use the updated vector $\mathbf{w}$ to update γ. The procedure is repeated until $\mathbf{w}$ and γ have converged. We summarize the algorithm as follows.

Algorithm 1. ***Coordinate-wise descent update for mvp penalized by the weighted l_1 and squared l_2 norm.***

1. *Fix λ and $\alpha \in [0,1]$ at some constant levels.*
2. *Initialize $\mathbf{w}^{(0)} = N^{-1}\mathbf{1}_N$ and $\gamma^{(0)} > \lambda$*
3. *For $i = 1,\ldots,N$, and $k > 0$,*

$$w_i^{(k)} \leftarrow \frac{ST\left(\gamma^{(k-1)} - z_i^{(k)}, \lambda\alpha\right)}{2\left(\sigma_i^2+\lambda\left(1-\alpha\right)\right)},$$

where

$$z_i^{(k)} = 2(\sum_{j<i} w_j^{(k)}\sigma_{ij} + \sum_{j>i} w_j^{(k-1)}\sigma_{ij}).$$

4. *For $k > 0$, update γ as*

$$\begin{aligned}\gamma^{(k)} \;\leftarrow\; & \left[\sum_{i\in S_+^{(k)}\cup S_-^{(k)}}\frac{1}{2\left(\sigma_i^2+\lambda\left(1-\alpha\right)\right)}\right]^{-1}\times\\ & \left[1+\sum_{i\in S_+^{(k)}\cup S_-^{(k)}}\frac{z_i^{(k)}}{2\left(\sigma_i^2+\lambda\left(1-\alpha\right)\right)}-\right.\\ & \left.\lambda\alpha\left(\sum_{i\in S_-^{(k)}}\frac{1}{2\left(\sigma_i^2+\lambda\left(1-\alpha\right)\right)}-\sum_{i\in S_+^{(k)}}\frac{1}{2\left(\sigma_i^2+\lambda\left(1-\alpha\right)\right)}\right)\right],\end{aligned}$$

where $S_+^{(k)} = \left\{i : w_i^{(k)} > 0\right\}$ and $S_- = \left\{i : w_i^{(k)} < 0\right\}$.

5. *Repeat 3 and 4 until* $\mathbf{w}^{(k)}$ *and* $\gamma^{(k)}$ *have converged.*

The approach for updating the weights and the Lagrangian multiplier given above can be generalized. For example, consider if a set of linear constraints $A\mathbf{w} = \mathbf{u}$ should be satisfied, where A is a $p \times N$ matrix, $\mathbf{u}$ is a $p \times 1$ vector, and $p \geq 1$. The right hand side of (5) becomes the objective function minus $\boldsymbol{\gamma}(A\mathbf{w} - \mathbf{u})$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ is a $1 \times p$ vector of the Lagrangian multipliers. We can follow procedures similar to the one given above to derive the updated forms for w_i and $\boldsymbol{\gamma}$, and then $\mathbf{w}$, γ_1, $\gamma_2, \ldots, \gamma_p$ can be updated sequentially. In Supplementary Materials S.2.2 we provide an example of the portfolio optimization in which the full investment and target return constraints are both considered, and a derivation of the coordinate-wise descent algorithm for solving it is given in Supplementary Materials S.3.3. Other possible extensions of the proposed approach will be discussed in Section 5.

In the following analysis, Algorithm 1 (as well as the algorithms introduced in Section 5) is implemented with codes written in R and the codes are available from the authors. We will compare speeds of Algorithm 1 and another optimization solver in Section 4.5. Note that the speed of the algorithm will be very fast when λ is large, and the resulting $\mathbf{w}$ in this situation will be sparse. A fast and stable convergence is important for the empirical studies conducted in Section 4.4 since we will rebalance the portfolio quite often over a long sampling period. However, when λ is small, the speed will become slow as the resulting $\mathbf{w}$ is expected to have only few or even no zero-valued elements. This is perhaps the reason why coordinate-wise descent algorithms are often ignored in solving optimization problems in the situation when solution vectors are expected to be dense.

3.2. The Upper Bound of the Penalty Parameter

As pointed out by Brodie et al. (2009), when $\alpha = 1$, if λ is beyond some threshold (say $\overline{\lambda}$), the optimal weight vectors of the weighted norm mvp and no-shortsale mvp will be identical. Hence using any $\lambda \geq \overline{\lambda}$ for solving (2) will only generate the optimal no-shortsale weight vector. The reason is intuitive. First, increasing λ is equivalent to decreasing c in the constraint $\|\mathbf{w}\|_{l_1} \leq c$. In order to satisfy the full investment constraint $\sum_{i=1}^{N} w_i = 1$, c cannot be less than 1, since

$$1 = \sum_{i=1}^{N} w_i \leq \|\mathbf{w}\|_{l_1} \leq c.$$

We therefore can view the upper bound $\overline{\lambda}$ as the value of λ corresponding to the lower bound $c = 1$. The maximum value of c can be set equal to $\|\mathbf{w}_{un}\|_{l_1}$, where $\mathbf{w}_{un}$ is the optimal weight vector of the unconstrained mvp. Obviously when $c > \|\mathbf{w}_{un}\|_{l_1}$, the l_1 norm constraint is inactive, which is equivalent to $\lambda = 0$. The above property of c is used in DeMiguel et al. (2009a) and Fan, Zhang, and Yu (2012) to determine an appropriate regularization on $\mathbf{w}$. When $\alpha = 1$, the range of λ can also be easily determined. Let

$$\widehat{\zeta} = \max_{i \notin S_{ns}} \sum_{j \in S_{ns}} w_{ns,j}\sigma_{ij} - \sigma_{ns}^2,$$

where $w_{ns,i}$ is the optimal weight of asset i from the no-shortsale mvp, $S_{ns} = \{i : w_{ns,i} > 0\}$ and σ_{ns}^2 is the in-sample portfolio variance of the optimal no-shortsale mvp. We propose to use

$$\widehat{\overline{\lambda}} = \max(0, \widehat{\zeta}) \tag{8}$$

as the estimate for the upper bound $\overline{\lambda}$. The derivation of (8) is shown in Appendix A. Practically, we can solve the no-shortsale mvp to obtain S_{ns} and relevant quantities, and (8) can be easily constructed.

4. Empirical Results

In this section we study how the weighted norm mvp performs with real data. We solve the weighted norm mvp with Algorithm 1. The data sets considered here are the monthly return data (in percentage) of the Fama and French 25 and 100 portfolios formed on size and book-to-market ratio (FF-25 and FF-100), and the 48 industry portfolios (FF-48). The first data set is used to depict the profiles of the optimal portfolio weights, proportion of active constituents and proportion of shortsale constituents. The latter two data sets are for examining empirical properties of the weighted norm mvp. The three data sets are publicly available and can be downloaded from the following website: `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

4.1. The Performance Measures

For estimating the covariance matrix of the asset returns Σ, we adopt a rolling window strategy with window length τ. Let $\widehat{\Sigma}_t$ denote the sample covariance matrix of the asset returns from previous τ periods (from period $t - \tau + 1$ to t). We use $\widehat{\Sigma}_t$ as the plug-in estimate for the covariance matrix of the asset returns Σ at the end of the tth period. Note that the frequency of our data sets is on a monthly basis. Suppose that we have T monthly return observations. To guarantee the sample covariance matrix $\widehat{\Sigma}_t$ is positive semidefinite for every t, we set $100 < \tau = 120 < T$ for all of the following analysis. We then plug the sample covariance matrix estimate $\widehat{\Sigma}_t$ into the portfolio optimization (2), and solve the optimal weights $\widehat{w}_{i,t}$, $i = 1, \ldots, N$ and the Lagrangian multiplier $\widehat{\gamma}_t$ for period $t+1$.

We solve the optimal weight vector and rebalance the portfolio for each period (month). We define the out-of-sample portfolio return at period $t+1$ by

$$\widehat{r}_{por,t+1} = \sum_{i=1}^{N} \widehat{w}_{i,t} r_{i,t+1}.$$

Denote sample mean and sample standard deviation of $\widehat{r}_{por,t+1}$, $t = \tau, \ldots, T-1$ by $\overline{\widehat{r}}_{por}$ and $\widehat{\sigma}_{por}$ respectively. We define the empirical Sharpe ratio by

$$\widehat{SR}_{por} = \frac{\overline{\widehat{r}}_{por}}{\widehat{\sigma}_{por}}.$$

We also calculate the turnover rate of the trading strategy. Suppose at the end of period $t-1$, the investor has wealth θ_{t-1} that can be invested on the assets. Given the optimized weight $\widehat{w}_{i,t-1}$, the value of holding asset i at the end of period t is $\theta_{t-1}\widehat{w}_{i,t-1}\left(1 + r_{i,t}\right)$. If there are N assets, the total wealth at period t will be $\theta_t = \theta_{t-1}\left(1 + \widehat{r}_{por,t}\right)$. Given the optimal weight of asset i at period $t+1$ $\widehat{w}_{i,t}$, the amount of wealth to invest on asset i is $\theta_t \widehat{w}_{i,t}$. We define the turnover rate of asset i between t and $t+1$ by

$$TOR_{i,t+1} = \left| \widehat{w}_{i,t} - \widehat{w}_{i,t-1} \frac{(1 + r_{i,t})}{(1 + \widehat{r}_{por,t})} \right|, \tag{9}$$

which is just proportion of the wealth at the end of period t needed to invest on asset i in order to satisfy the amount $\theta_t \widehat{w}_{i,t}$. We further define the portfolio turnover rate at period $t+1$ as the sum of the turnover rate (9) over the N assets,

$$TOR_{por,t+1} = \sum_{i=1}^{N} TOR_{i,t+1}.$$

For the last two measures, we define the proportion of active constituents at period t by

$$PAC_t = \frac{\left| \widehat{S}_t^+ \cup \widehat{S}_t^- \right|}{N},$$

where $\widehat{S}_t^+ = \{i : \widehat{w}_{i,t} > 0\}$ and $\widehat{S}_t^- = \{i : \widehat{w}_{i,t} < 0\}$, and we define the absolute position of a shortsale portfolio at period t as the sum of absolute values of all negative asset weights at period t,

$$APS_t = \sum_{i \in \widehat{S}_t^-} \left| \widehat{w}_{i,t} \right|.$$

In Section 4.4, we will see that PAC_t and APS_t are useful for us to know the relationship between the no-shortsale and the weighted norm mvp as well as how the penalty function affects components of the optimal weight vector.

4.2. Choosing the Sequence of λ

Note that all the above quantities are calculated given that λ and α in (2) are fixed. Our main interest lies in how these relevant quantities vary as λ and α change. Therefore it is necessary to solve the weighted norm mvp over different values of λ and α. Here we choose a sequence of α from interval $[0, 1]$. Note that the upper bound of λ can be obtained by using (8). Let $\widehat{\overline{\lambda}}_t$ denote such an estimated upper bound at period t. The sequence of λ is chosen from the interval $[\lambda_{max}, \lambda_{min}]$ where

$$\lambda_{max} = \max_t \widehat{\overline{\lambda}}_t \text{ and } \lambda_{min} = \frac{\lambda_{max}}{1000}.$$

4.3. Profiles of the Optimal Weights

Figure 1 shows profiles of the portfolio weights, proportion of active constituents (PAC) and proportion of shortsale constituents (POS, defined as $\left|\widehat{S}^{-}\right|/N$) for the FF-25 under different λ and α. The data for calculating the sample covariance matrix has sampling period from November 1986 to October 1996. When $\alpha = 1$, only the l_1 penalty is activated, and profiles of the asset weights are similar to those under the lasso estimation (Tibshirani, 1996) in statistics: some of the weights are exactly zero when $\lambda > 0$. Note that the asset weight profiles do not all vanish to zero when λ becomes very large, since the constraint $\mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1$ needs to hold. When $\alpha > 0$, the proportion of active constituents declines as λ increases. The l_1 penalty facilitates sparsity, so $|S^{+} \cup S^{-}|/N = 1$ when $\alpha = 0$. When $\alpha = 0.5$ and λ is large, it can be seen that no portfolio has negative weights, but the number of active constituents is different from the case of $\alpha = 1$. The result is expected, since mathematically the case of $\alpha = 0.5$ is equivalent to the one with Σ replaced by $\Sigma + 0.5\lambda\mathbf{I}_{N\times N}$ and asset weights are regularized by the l_1 penalty only. Therefore as λ goes large enough, the resulting mvp will also be a no-shortsale mvp.

4.4. Performances Over a Long Sampling Period

We then compare performances of the weighted norm mvp with three benchmark portfolios: a naively diversified portfolio with equal weights $1/N$, the unconstrained mvp (UN, corresponding to $\lambda = 0$) and the no-shortsale mvp (NS). Figure 2 and 3 illustrate $\widehat{\sigma}_{por}$, $\widehat{SR}_{por}$, sample averages of $TOR_{por,t}$, PAC_t, APS_t and $\widehat{\gamma}_t$ when monthly

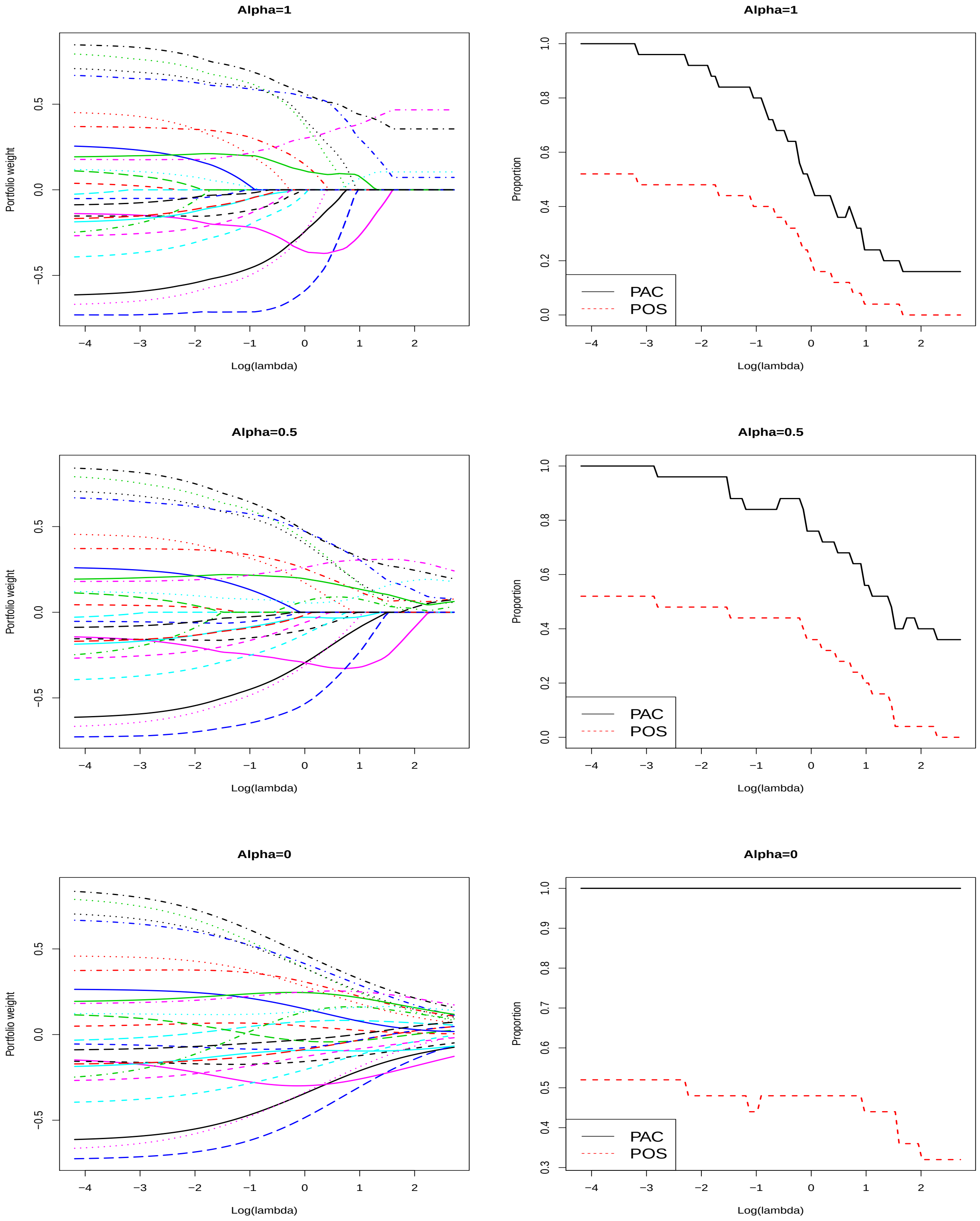

Figure 1: Profiles of portfolio weights, proportion of active constituents (PAC) and proportion of shortsale constituents (POS) from solving the optimal weighted norm mvp. The data for calculating the sample covariance matrix is the monthly return data of the FF 25 size and BM ratio portfolios from Nov-1986 to Oct-1996.

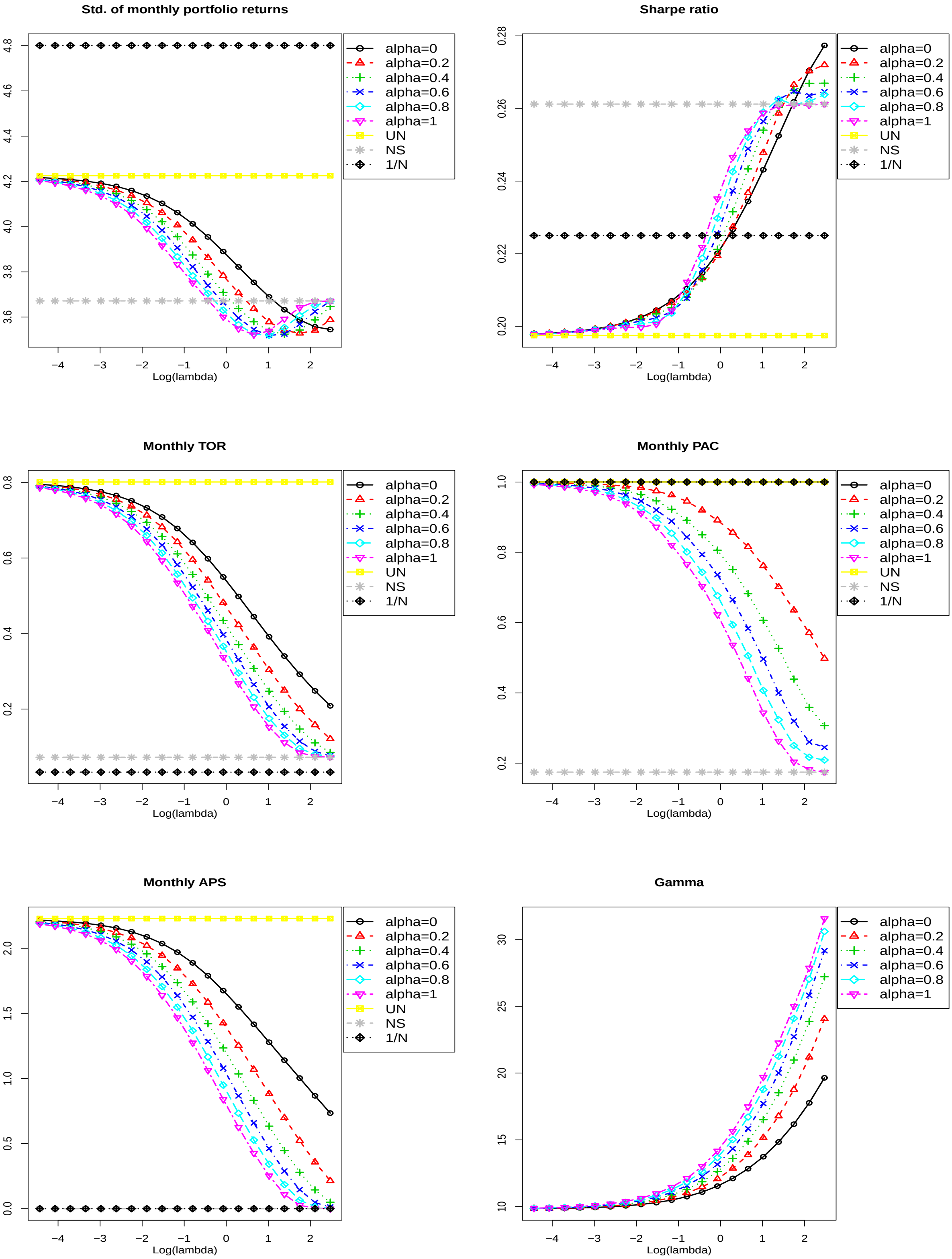

Figure 2: Standard deviation of out-of-sample portfolio returns, Sharpe ratio, average turnover rate (TOR), proportion of active constituents (PAC), absolute position of shortsales (APS) from the weighted norm mvp, no-shortsale mvp (NS), unconstrained mvp (UN) and 1/N portfolio, and optimal γ from the weighted norm mvp. The data used is the monthly return data of the FF 48 industry portfolios from July-1979 to Sep-2009.

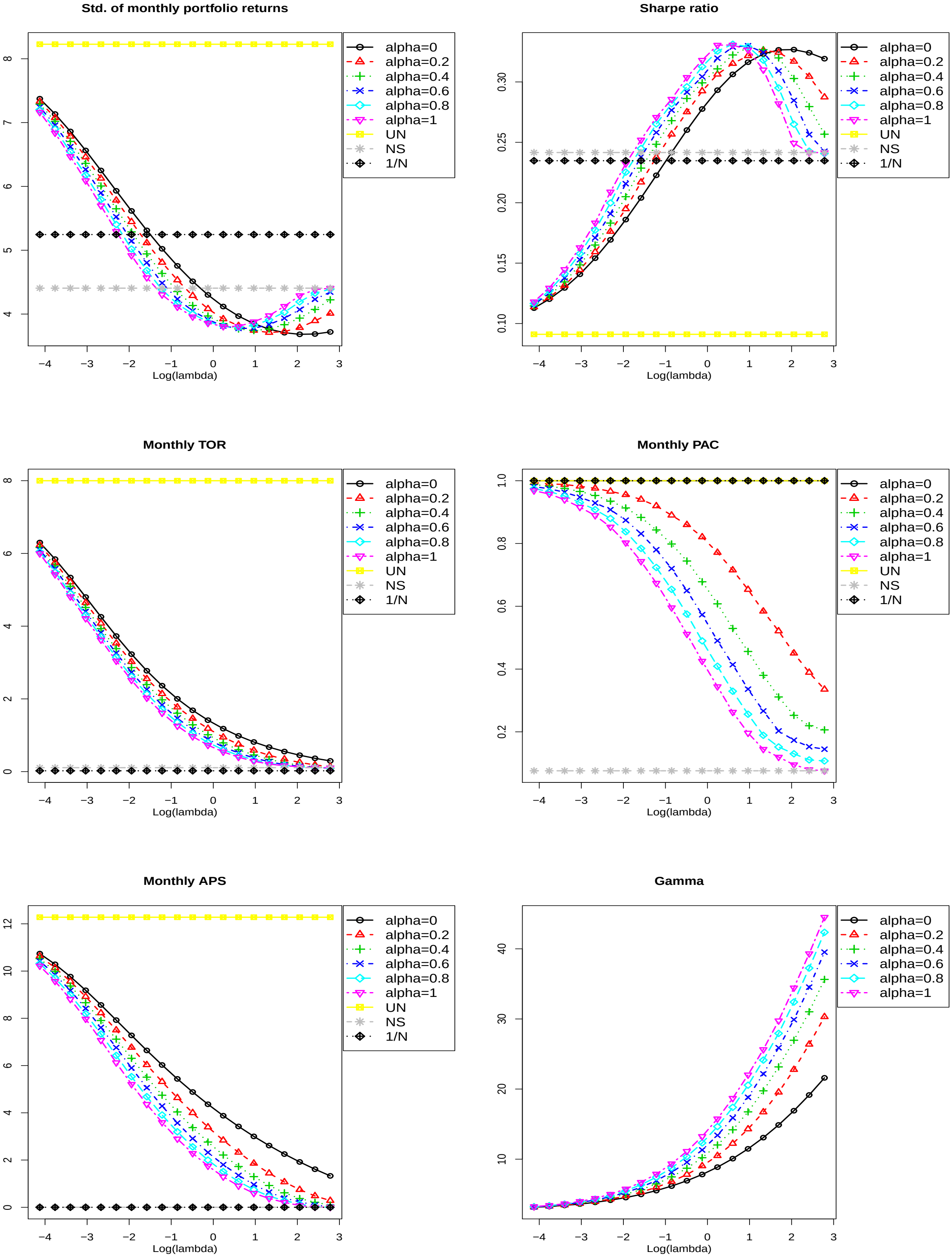

Figure 3: Standard deviation of out-of-sample portfolio returns, Sharpe ratio, average turnover rate (TOR), proportion of active constituents (PAC), absolute position of shortsales (APS) from the weighted norm mvp, no-shortsale mvp (NS), unconstrained mvp (UN) and 1/N portfolio, and optimal γ from the weighted norm mvp. The data used is the monthly return data of the FF 100 size and BM ratio portfolios from July-1973 to Sep-2009.

return data of the Fama and French 48 industry portfolios and the 100 size and book-to-market portfolios are used for the four different strategies. For the FF-48 data, the sampling period is from July 1979 to September 2009. For the FF-100 data, the sampling period is from July 1973 to September 2009.

We set $\alpha = 0, 0.2, 0.4, 0.6, 0.8$, and 1 and solve the weighted norm mvp optimization problem with different α values. The sequence of penalty parameter λ is chosen via the same way as in Section 4.3, but only with 20 different levels. When $\alpha = 1$, as λ increases, $\widehat{\sigma}_{por}$ of the weighted norm portfolio declines to a minimum and then converges to a level where the no-shortsale mvp holds. For the FF-48 case, the naive $1/N$ portfolio has the largest $\widehat{\sigma}_{por}$, whereas in the FF-100 case, the unconstrained mvp has the largest $\widehat{\sigma}_{por}$. These findings suggest that when the sample size is not relatively large enough to the number of assets N, the naive $1/N$ portfolio seems to be better than the unconstrained mvp in reducing portfolio volatility. However, $\widehat{\sigma}_{por}$ of the $1/N$ portfolio is still higher than that of the weighted norm mvp when λ becomes moderately large. For the case of $\alpha \neq 1$, all the above results are qualitatively similar to the case of $\alpha = 1$.

As for the Sharpe ratio, it increases monotonically with λ for the FF-48 case. But for the FF-100 case, $\widehat{SR}_{por}$ reaches its maximum when λ is moderately large, and then starts to decline to a level where the no-shortsale mvp is held.

The $1/N$ portfolio has the lowest turnover rate, and such property was well documented in previous studies. The turnover rate of the weighted norm mvp decreases monotonically as λ increases. This suggests that regularization on the asset weights can stabilize the weights over time. To further understand how turnover rate affects performances of different trading strategies, we calculate the terminal wealth net of transaction costs

$$TERW_t = \theta_0 \prod_{k=1}^{t} \left(1 + \widehat{r}_{por,k}\right)\left(1 - \eta \times to_{pot,k}\right),$$

where θ_0 is the initial investment and η is a fixed proportional transaction cost. Figure 4 shows time series plots of $TERW_t$ under different trading strategies. We set the initial investment $\theta_0 = 1$. In the left panel are plots for $\eta = 0$ and in the right panel are plots for $\eta = 0.0025$ (25 basis points). For each data set, when $\alpha = 0.6$, we found the average Sharpe ratio (over the 20 different λ's) of the weighted norm mvp has the highest value, and hence here we set $\alpha = 0.6$. For the penalty parameter, we set its value at two different levels: the $10th$ and $15th$ largest values among the 20 $\lambda' s$. Accordingly, for the FF-48 case, $\lambda = 0.31$ and 1.93, and for the FF-100 case, $\lambda = 0.43$ and 2.62.

From Figure 4 it can be seen that when $\eta = 0.0025$, the unconstrained mvp and the weighted norm mvp with a lower λ have lower $TERW_t$ than other portfolio strategies due to their higher turnover rates. For example, in the FF-100 case, at the end of our sampling period, $TERW_t$ for the unconstrained mvp is 5.9991 when $\eta = 0$ but drops to 0.0009 when $\eta = 0.0025$. It means that almost all gains of the unconstrained mvp are eroded by the transaction cost. The $1/N$ portfolio, which has the lowest turnover rate, suffers the least impacts from the transaction cost. As can be seen from the figure, at the end of our sampling period, the $1/N$ has the highest $TERW_t$ in the FF-48 case; but for the FF-100 case, it is the weighted norm mvp with $\lambda = 2.62$ has the highest $TERW_t$.

As for the PAC_t and APS_t, both decline as λ increases. For the PAC_t, it is due to the fact that the l_1 penalty facilitates sparsity. For the APS_t, the reason is that as λ increases, the optimal solution of (2) converges to the solution of the no-shortsale mvp. Consequently the absolute position of shortsales declines to zero. In general, given the same λ, higher α leads to lower PAC_t and APS_t. Finally, given λ and α, the Lagrangian multiplier $\widehat{\gamma}$ simply takes on the value to ensure the full investment constraint satisfied. From the KKT conditions, it can be shown that $\widehat{\gamma}$ is monotonically increasing with λ, and such relationship also can be seen from Figure 2 and 3.

4.5. Comparisons with Other Optimization Solvers

In this subsection, we compare performances of Algorithm 1 and other optimization routines which are available for solving the norm constrained mvp optimization problems. We first compare the optimal solutions produced by Algorithm 1 with those produced by optimization package `cvx` (Grant and Boyd, 2010). Let $\widehat{\mathbf{w}}_{cwd,t}$ and $\widehat{\mathbf{w}}_{cvx,t}$ be the weight vectors at period t produced by Algorithm 1 and `cvx`, respectively. We define the cumulative difference of the weights by

$$\sum_{t=\tau+1}^{T} \left\|\widehat{\mathbf{w}}_{cwd,t} - \widehat{\mathbf{w}}_{cvx,t}\right\|_{l_1} .$$

We use the FF-48 and FF-100 data to examine the differences. We fix $\alpha = 1$ and vary λ at six different levels. The two plots in the top panel of Figure 5 show the results. The values of the cumulative difference are overall small and decline with λ. We also use `cvx` to obtain the optimized no-shortsale weight vector. The dash and dot lines in Figure 5 are the cumulative difference between the optimized no-shortsale weights and $\widehat{\mathbf{w}}_{cwd,t}$ and the cumulative difference between the optimized no-shortsale weights and $\widehat{\mathbf{w}}_{cvx,t}$ with $\lambda = 30$, respectively. It can be seen that $\widehat{\mathbf{w}}_{cwd,t}$ is even closer to the no-shortsale weight vector produced by `cvx` than $\widehat{\mathbf{w}}_{cvx,t}$.

We then compare average CPU time of Algorithm 1 and another algorithm in solving the mvp optimization problems from simulations. Since our codes were implemented in

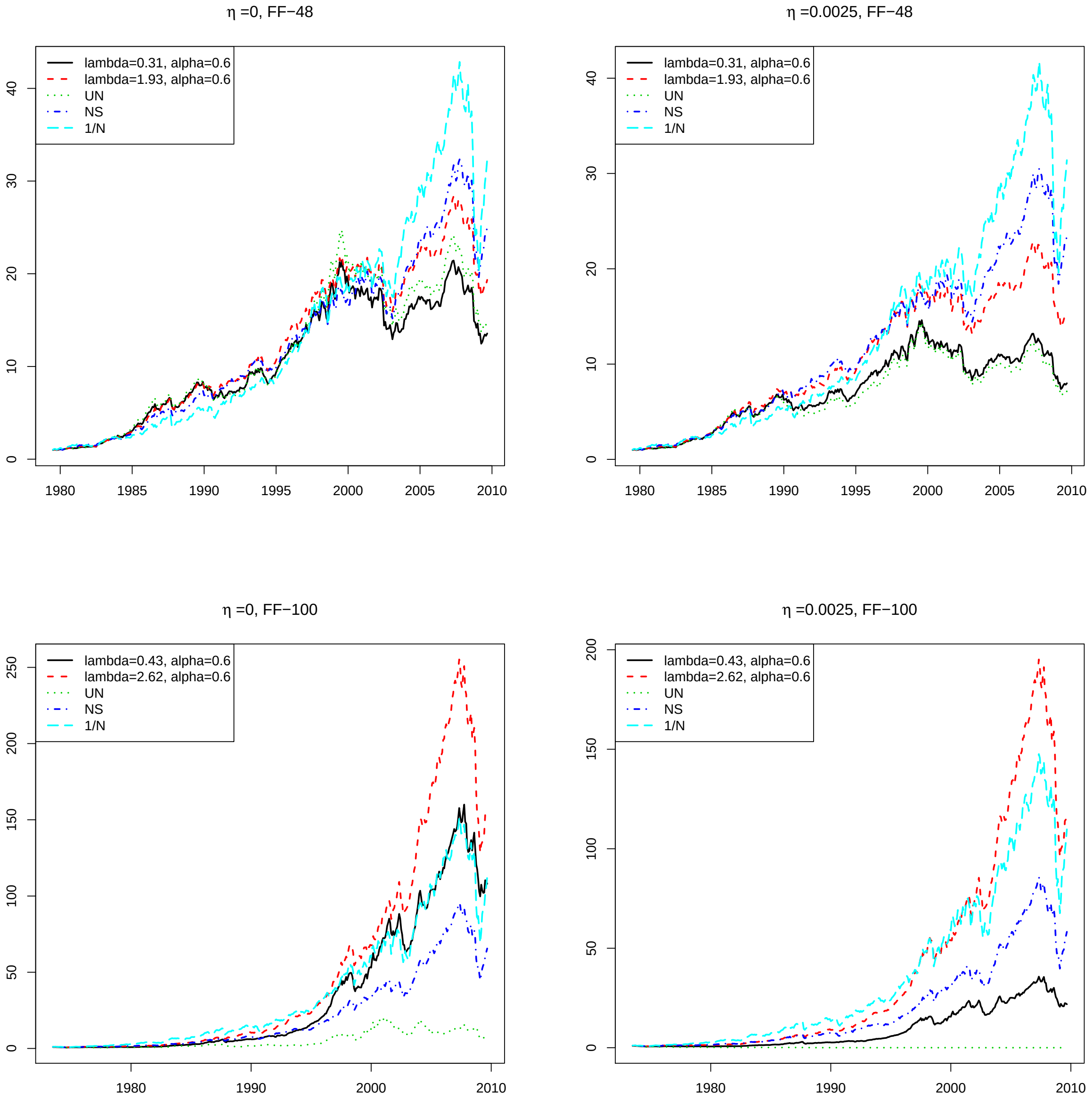

Figure 4: Time series plots of terminal wealth net of transaction costs $TERW_t$ of the weighted norm mvp (with $\alpha = 0.6$), no-shortsale mvp (NS), unconstrained mvp (UN) and 1/N portfolio.

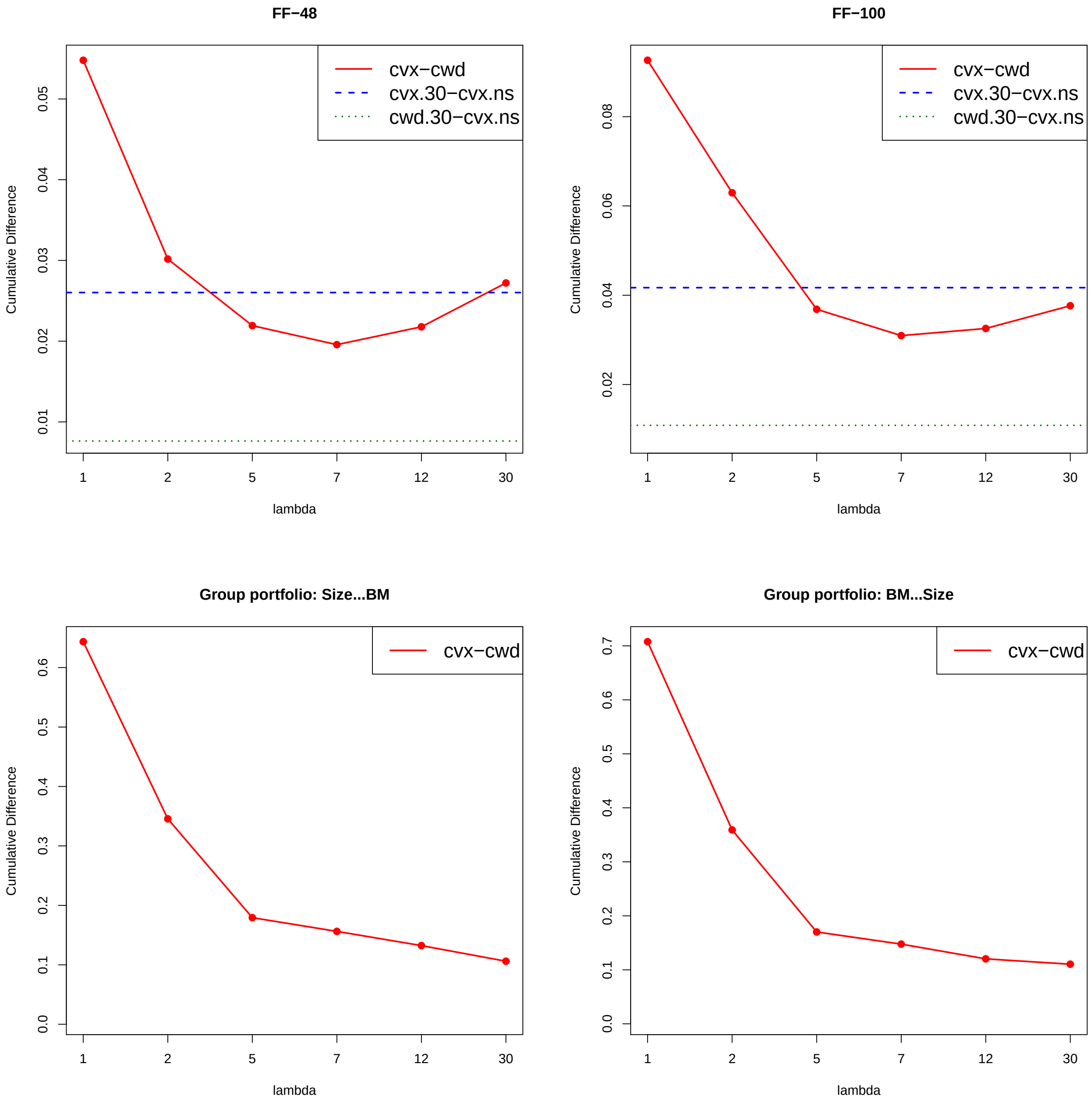

Figure 5: Cumulative differences between solutions from the coordinate-wise descent algorithms and `cvx` (Grant and Boyd, 2010) for solving (2) and (12).

Table 1: Average CPU time (seconds), proportion of active constituents (PAC) and l_1 distance of the optimal weight vectors from solving the weighted norm mvp optimization from Algorithm 1, no-shortsale (NS) and unconstrained (UN) mvp optimization problems from `solve.QP`. The l_1 distance between the solved optimal weight vectors shown here is only for (1) no-shortsale mvp and the weighted norm mvp with $\lambda = \widehat{\overline{\lambda}}$, and (2) unconstrained mvp and the weighted norm mvp with $\lambda = 0$. Each simulation is run 1000 times.

		$\Sigma = \mathbf{I}_{N\times N}$				
		$N = 50$	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
`solve.QP-NS`	time	0.0055	0.0377	0.2880	4.3413	34.1742
$\lambda = \widehat{\overline{\lambda}}$	time	0.0028	0.0033	0.0113	0.1736	0.8911
	PAC	0.7411	0.7339	0.7326	0.7292	0.7288
	l_1 dist.	1.63e-4	2.09e-6	2.89e-6	4.68e-6	5.98e-6
`solve.QP-UN`	time	0.0006	0.0020	0.0094	0.0870	0.5505
$\lambda = 0$	time	0.0055	0.0197	0.0673	1.0631	5.5748
	PAC	1.0000	1.0000	1.0000	1.0000	1.0000
	l_1 dist.	4.25e-7	4.82e-7	5.32e-7	5.65e-7	5.81e-7
$\lambda = 0.8 \times \widehat{\overline{\lambda}}$	time	0.0021	0.0034	0.0116	0.1761	0.9010
	PAC	0.7648	0.7488	0.7417	0.7349	0.7323
$\lambda = 0.6 \times \widehat{\overline{\lambda}}$	time	0.0014	0.0039	0.0122	0.1856	0.9400
	PAC	0.7847	0.7677	0.7581	0.7472	0.7422
$\lambda = 0.4 \times \widehat{\overline{\lambda}}$	time	0.0016	0.0044	0.0144	0.2129	1.0727
	PAC	0.8196	0.8026	0.7913	0.7771	0.7698
$\lambda = 0.2 \times \widehat{\overline{\lambda}}$	time	0.0022	0.0066	0.0208	0.3050	1.5206
	PAC	0.8821	0.8683	0.8572	0.8440	0.8357
		$\Sigma =$Toeplitz $\left(0.6^{\lvert i-j\rvert}\right)$				
		$N = 50$	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
`solve.QP-NS`	time	0.0082	0.0629	0.4889	7.4676	58.7770
$\lambda = \widehat{\overline{\lambda}}$	time	0.0008	0.0022	0.0078	0.1267	0.6526
	PAC	0.5081	0.4957	0.4916	0.4892	0.4894
	l_1 dist.	1.30e-6	1.76e-6	2.24e-6	2.92e-6	3.70e-6
`solve.QP-UN`	time	0.0006	0.0019	0.0090	0.0863	0.5510
$\lambda = 0$	time	0.0132	0.0441	0.1503	2.4064	12.5209
	PAC	1.0000	1.0000	1.0000	1.0000	1.0000
	l_1 dist.	9.62e-7	1.08e-6	1.18e-6	1.25e-6	1.28e-6
$\lambda = 0.8 \times \widehat{\overline{\lambda}}$	time	0.0009	0.0024	0.0082	0.1306	0.6674
	PAC	0.5415	0.5166	0.5064	0.4975	0.4946
$\lambda = 0.6 \times \widehat{\overline{\lambda}}$	time	0.0010	0.0029	0.0091	0.1427	0.7181
	PAC	0.5806	0.5515	0.5335	0.5206	0.5132
$\lambda = 0.4 \times \widehat{\overline{\lambda}}$	time	0.0016	0.0037	0.0125	0.1813	0.8873
	PAC	0.6492	0.6180	0.5990	0.5790	0.5671
$\lambda = 0.2 \times \widehat{\overline{\lambda}}$	time	0.0023	0.0073	0.0215	0.3157	1.5385
	PAC	0.7679	0.7438	0.7271	0.7062	0.6937

R, to make the comparison as fair as possible, the optimization routine we consider here is `solve.QP` in R package `quadprog`, which uses the dual method of Goldfarb and Idnani (1982, 1983) to solve quadratic programming problems. To demonstrate pros and cons on using the coordinate-wise descent algorithm, we focus the comparisons on solving two mvp optimization problems: the no-shortsale and unconstrained mvp, which are equivalent to two extreme cases of the weighted norm penalty mvp with $\alpha = 1$. The former is an mvp with a heavy norm penalty on the portfolio weights and the later is an mvp in which the portfolio weights are without any norm penalty. The two mvp optimization problems can be easily formulated and solved with `solve.QP`. To guarantee that the optimized no-shortsale weights can be obtained via Algorithm 1, we set the penalty parameter as its upper bound obtained via (8) ($\lambda = \widehat{\overline{\lambda}}$). For a more complete assessment of applicability of Algorithm 1, we also report the relevant results when different levels of regularizations are imposed. Here we still fix $\alpha = 1$ but vary λ according to its upper bound $\widehat{\overline{\lambda}}$. For the simulations, we generate data from a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ with two different types of covariance matrices: $\mathbf{I}_{N\times N}$ and Toeplitz type matrix with the (i, j)th element equal to $0.6^{|i-j|}$. Each time we generate $n = 1.2N$ samples for calculating the sample covariance matrix, which is then plugged in for solving the mvp optimization problem. Each simulation is run 1000 times.

Table 1 shows the average CPU time, proportion of active constituents (PAC), and l_1 distance between the solved optimal weight vectors (sum of absolute differences between the weights). The l_1 distances shown here are only for the weighted norm mvp with $\lambda = \widehat{\overline{\lambda}}$ and no-shortsale mvp, and the weighted norm mvp with $\lambda = 0$ and unconstrained mvp. From Table 1, it can be seen that Algorithm 1 is much faster than `solve.QP` in solving the optimal no-shortsale mvp. Without any norm penalty, however, `solve.QP` is instead more efficient in solving the unconstrained mvp optimization than Algorithm 1. The differences of computational times between the two optimization routines become even more obvious when the number of assets N becomes large. For example, as N grows from 50 to 1000, in solving the optimal no-shortsale mvp, the average CPU time (seconds) of `solve.QP` is 1.96 to 38.35 times longer than Algorithm 1 when $\Sigma = \mathbf{I}_{N\times N}$. For the case of the Toeplitz matrix, `solve.QP` takes 10.25 to 90.07 times longer than Algorithm 1 in solving the optimal no-shortsale mvp. In addition, except for $\lambda = 0$, the average computational times of solving the weighted norm mvp between different regularization levels only have mild differences. The result suggests that Algorithm 1 can work efficiently over different values of $\lambda > 0$. Finally, as can be seen from the negligible l_1 distances between the solved optimal weight vectors, the two methods generate almost identical optimal solutions for the no-shortsale and unconstrained mvp optimization problems.

5. Some Extensions

It would be interesting to see how the mvp performs when different norm constraints or penalty functions are imposed on the asset weights. In this section, we show that the coordinate-wise descent algorithm introduced in Section 3.1 can be extended to solve these modified mvp optimization problems, and empirical performances of these modified mvp's are demonstrated with the FF-48 and FF-100 portfolio data.

5.1. Berhu Penalty

The coordinate-wise descent algorithm can be applied to solve the mvp optimization problems regularized by other penalty functions, for example, the berhu penalty proposed by Owen (2007). The name "berhu" comes from the fact that it is the reverse of Huber's loss. In statistics, Huber's loss function is designed to mitigate effects from large error terms in regression estimations and has the following form:

$$\mathcal{H}(\varepsilon) = \begin{cases} \varepsilon^2 & |\varepsilon| < \delta \\ 2\delta|\varepsilon| - \delta^2 & |\varepsilon| \geq \delta \end{cases}, \tag{10}$$

where $\delta > 0$ is a constant. The Huber's loss (10) is a hybrid function of two parts: one is a quadratic function for the error having a relatively small magnitude ($|\varepsilon| < \delta$) and the other is a linear function for absolute value of the error having a relatively large magnitude ($|\varepsilon| \geq \delta$). By separating impacts from the errors having small and large magnitudes, the Huber's loss (10) can be used in robust regression estimations.

Owen (2007) showed that the reverse of Huber's loss, the berhu penalty (for the portfolio weights) can be expressed as

$$\lambda \sum_{i=1}^{N} \left(|w_i| \, \mathbb{I}\{|w_i| < \delta\} + \frac{w_i^2 + \delta^2}{2\delta} \mathbb{I}\{|w_i| \geq \delta\} \right), \tag{11}$$

where $\delta > 0$ is a constant and $\mathbb{I}\{A\}$ is the indicator function such that $\mathbb{I}\{A\} = 1$ if event A is true and $\mathbb{I}\{A\} = 0$ otherwise. The berhu penalty (11) is also a combination of the l_1 and squared l_2 norm penalties. However, unlike the weighted l_1 and squared l_2 norm penalty in (2), the berhu penalty adopts a different regularization rule. If $|w_i|$ is small (less than δ), then it will be regularized by the l_1 norm penalty; if $|w_i|$ is large (greater than or equal to δ), then it will be regularized by the squared l_2 norm penalty. By specifying the parameter δ, the berhu penalty can help us to regularize large and small portfolio weights separately, like the Huber loss dealing with error terms in the regression estimations. Such setting provides an alternative way to robustly deal with problem of extreme portfolio weight in the portfolio optimization problems.

The berhu penalty (11) is convex and satisfies the additive separability condition. The algorithm for solving the mvp constrained by the berhu penalty is summarized as follows.

Algorithm 2. ***Coordinate-wise descent update for mvp penalized by the Berhu penalty.***

1. *Fix λ and δ at some constant levels.*
2. *Initialize $\mathbf{w}^{(0)} = N^{-1}\mathbf{1}_N$ and $\gamma^{(0)} > \lambda$*
3. *For $i = 1, \ldots, N$, and $k > 0$,*

$$w_i^{(k)} \leftarrow \begin{cases} \frac{ST\left(\gamma^{(k-1)} - z_i^{(k)}, \lambda\right)}{2\sigma_i^2} & \textit{if } \left|\gamma^{(k-1)} - z_i^{(k)}\right| < 2\sigma_i^2\delta + \lambda \\ \frac{\gamma^{(k-1)} - z_i^{(k)}}{2\sigma_i^2 + \frac{\lambda}{\delta}} & \textit{otherwise,} \end{cases}$$

where

$$z_i^{(k)} = 2\Big(\sum_{j<i} w_j^{(k)}\sigma_{ij} + \sum_{j>i} w_j^{(k-1)}\sigma_{ij}\Big).$$

4. *For $k > 0$, update γ as*

$$\begin{aligned} \gamma^{(k)} \quad \leftarrow \quad & \left[\sum_{i\in\left(S_+^{(k)}\cap\mathbf{\Delta}_-^{(k)}\right)\cup\left(S_-^{(k)}\cap\mathbf{\Delta}_-^{(k)}\right)} \frac{1}{2\sigma_i^2} + \sum_{i\in\cap\mathbf{\Delta}_+^{(k)}} \frac{1}{2\sigma_i^2 + \delta^{-1}\lambda}\right]^{-1} \times \\ & \left[1 + \left(\sum_{i\in\left(S_+^{(k)}\cap\mathbf{\Delta}_-^{(k)}\right)\cup\left(S_-^{(k)}\cap\mathbf{\Delta}_-^{(k)}\right)} \frac{z_i^{(k)}}{2\sigma_i^2} + \sum_{i\in\cap\mathbf{\Delta}_+^{(k)}} \frac{z_i^{(k)}}{2\sigma_i^2 + \delta^{-1}\lambda}\right) - \right. \\ & \left. \lambda\left(\sum_{i\in S_-^{(k)}\cap\mathbf{\Delta}_-^{(k)}} \frac{1}{2\sigma_i^2} - \sum_{i\in S_+^{(k)}\cap\mathbf{\Delta}_-^{(k)}} \frac{1}{2\sigma_i^2}\right)\right], \end{aligned}$$

where

$$\begin{aligned} \mathbf{\Delta}_-^{(k)} &= \left\{i : \left|\gamma^{(k-1)} - z_i^{(k)}\right| < 2\sigma_i^2\delta + \lambda\right\}, \\ \mathbf{\Delta}_+^{(k)} &= \left\{i : \left|\gamma^{(k-1)} - z_i^{(k)}\right| \geq 2\sigma_i^2\delta + \lambda\right\}. \end{aligned}$$

5. *Repeat 3 and 4 until $\mathbf{w}^{(k)}$ and $\gamma^{(k)}$ have converged.*

The derivation of Algorithm 2 is shown in Supplementary Materials S.3.1.

To examine how the mvp with the berhu penalty performs on the FF-48 and FF-100 data, we solve the mvp optimization with $\delta = 0.5/\text{N}$, 1/N, 0.1, 0.2, and 1, and vary λ at 20 levels which are the same as in Section 4.4. Figure 6 and 7 show the results. Different δ's lead to very different patterns of $\widehat{\sigma}_{por}$ and average PAC_t. When δ is small (1/N and 0.5/N), a higher λ no longer yields a lower $\widehat{\sigma}_{por}$ and PAC_t. The Sharpe ratio is inversely related to $\widehat{\sigma}_{por}$, and its possible maximum value is similar to the weighted norm mvp case. On average, the TOR_t and APS_t decline as λ increases, and given the same λ, the lower the δ is, the higher the TOR_t, APS_t and γ are.

5.2. Grouped Portfolio Selection

Yuan and Lin (2006) proposed the group lasso that can be used to select covariates grouply in linear regression problems. The penalty for such variable selection is the Euclidean norm $\|.\|_{l_2}$. The Euclidean norm can only facilitate sparsity between groups, that is, either all covariates in a certain group are selected or all of them are dropped out. It cannot facilitate sparsity within a group. Friedman et al. (2010) proposed the group sparse lasso in which the penalty is a combination of the Euclidean norm and an l_1 norm. They showed that such setting can facilitate sparsity between groups as well as within a group.

Categorizing assets into different groups based on certain asset characteristics and then forming a portfolio according to these characteristics is a standard process used in almost all portfolio management settings. A strategic asset selection based on certain groups may be due to some practical reasons, e.g., financial regulations, investor's risk preferences or a fund's own objective. However, such a strategy often concentrates on certain groups of assets and ignores benefits from diversification. Therefore when minimizing the overall portfolio variance, the strategy may not be optimal. In the following, we apply the idea of group variable selection in minimizing the overall portfolio variance. To fairly achieve the goal, we will not specify which groups of assets should be more important than others, nor do we ditch boundaries of groups and purely minimize the portfolio variance. Instead we group these assets according to their common features, and then minimizing the portfolio variance under the group penalty.

Assume there are L groups of assets. Let $\mathbf{w}_l = (w_{l1}, \ldots, w_{lK})$ be the portfolio weight vector for assets in group l. Without loss of generality, for each group, we assume the number of assets is equal to K. The case of different numbers of assets in different groups can be easily modified in our algorithm. However, we do not allow different groups having the same assets. If different groups could have the same assets, the block coordinate-wise descent algorithm we use in solving the specified mvp problem might not guarantee a stable convergence.

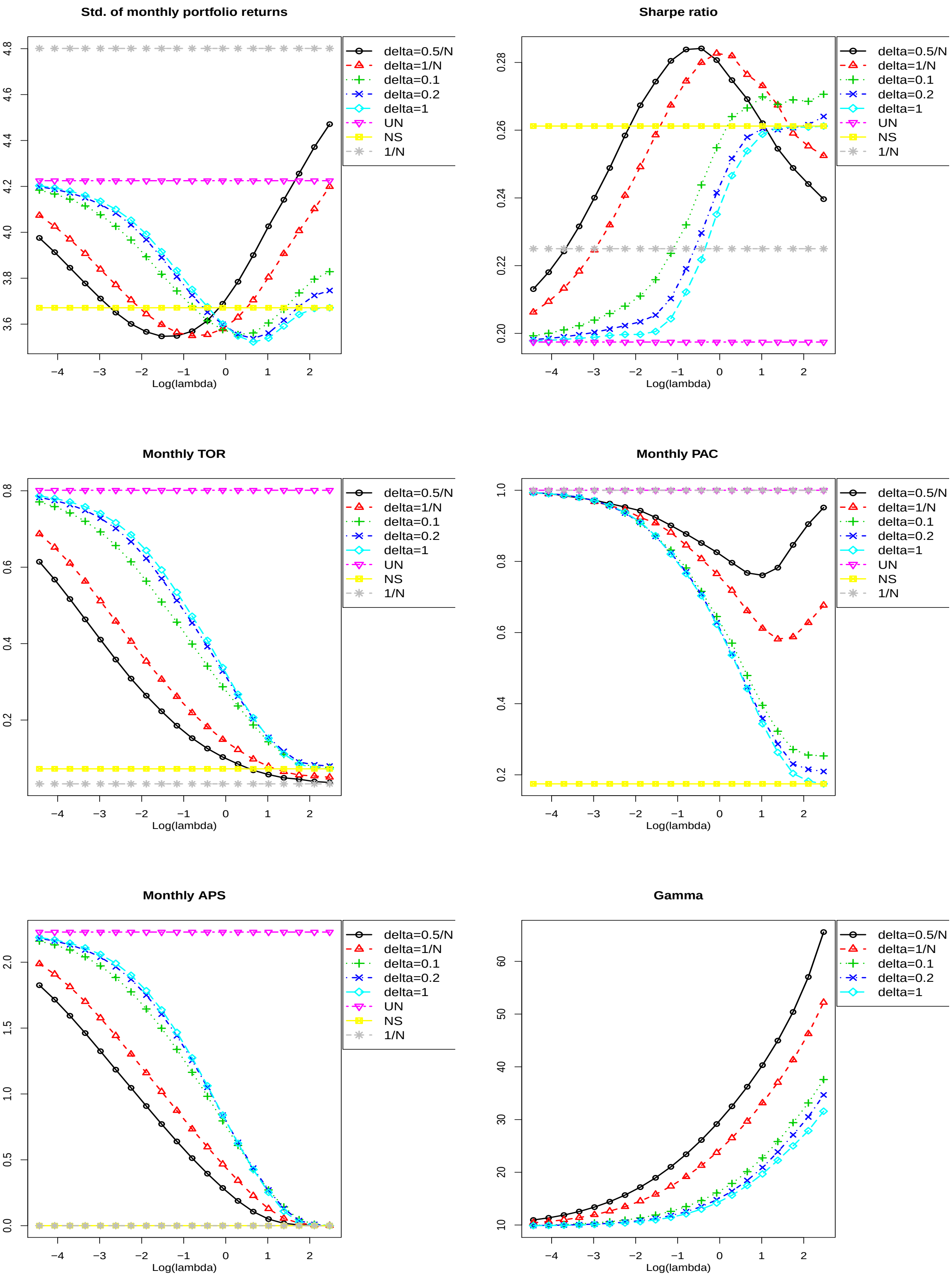

Figure 6: Standard deviation of out-of-sample portfolio returns, Sharpe ratio, average turnover rate (TOR), proportion of active constituents (PAC), absolute position of shortsales (APS) from the mvp with the berhu penalty, no-shortsale mvp (NS), unconstrained mvp (UN) and 1/N portfolio, and optimal γ from the mvp with the berhu penalty. The data used is the monthly return data of the FF 48 industry portfolios from July-1979 to Sep-2009.

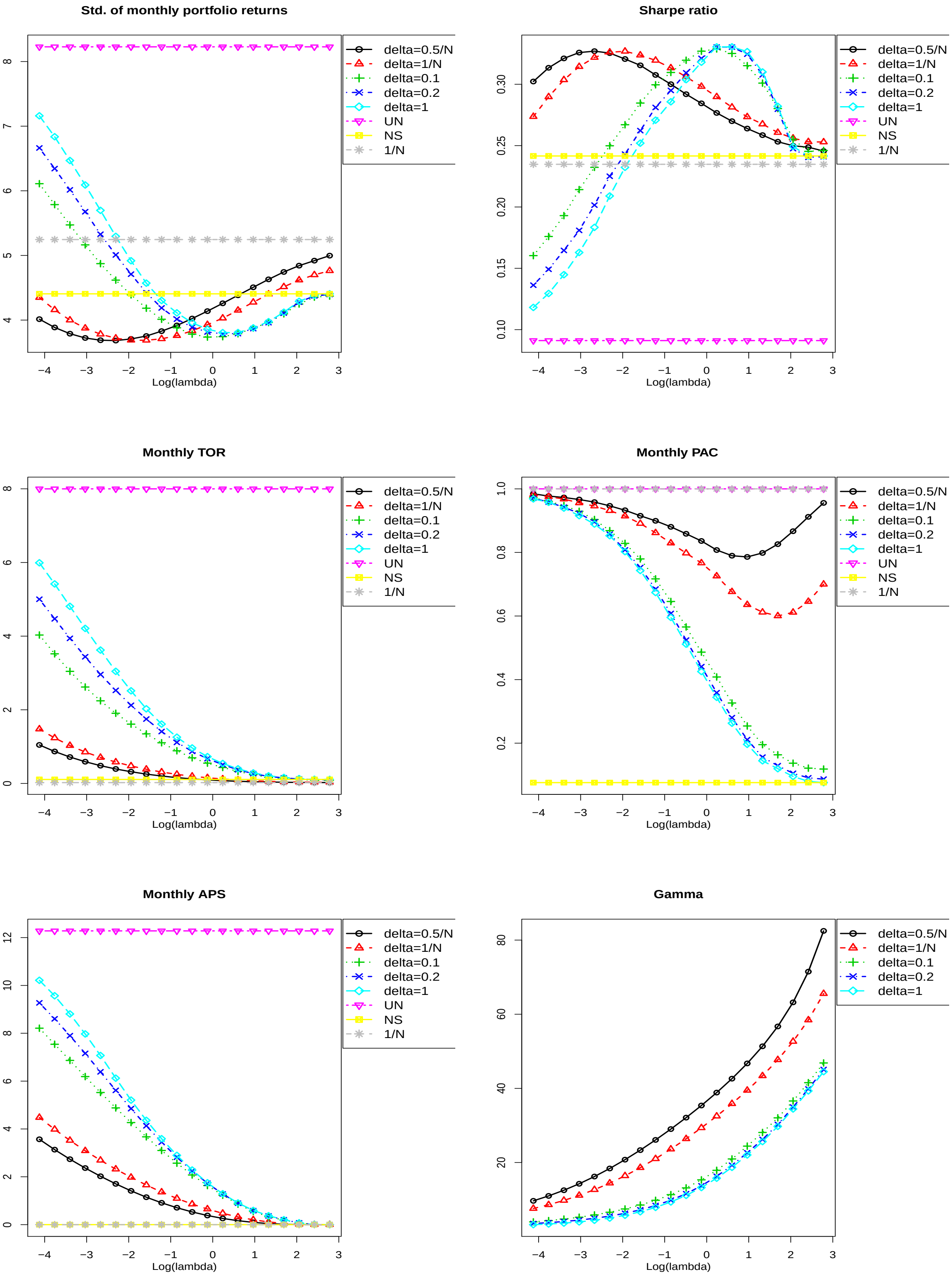

Figure 7: Standard deviation of out-of-sample portfolio returns, Sharpe ratio, average turnover rate (TOR), proportion of active constituents (PAC), absolute position of shortsales (APS) from the mvp with the berhu penalty, no-shortsale mvp (NS), unconstrained mvp (UN) and 1/N portfolio, and optimal γ from the mvp with the berhu penalty. The data used is the monthly return data of the FF 100 size and BM ratio portfolios from July-1973 to Sep-2009.

Let $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_L)$. The modified mvp problem can be stated as

$$\min_{\mathbf{w}} \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} + \lambda \sum_{l=1}^{L} \|\mathbf{w}_l\|_{l_2,\Omega_l} \qquad \text{subject to } \mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1, \tag{12}$$

where $\|\mathbf{w}_l\|_{l_2,\Omega_l} = \sqrt{\mathbf{w}_l^T \Omega_l \mathbf{w}_l}$, and Ω_l is a kernel matrix, which is required to be symmetric and positive definite. By definition, the Euclidean norm of $\mathbf{w}_l$ can then be expressed as

$$\|\mathbf{w}_l\|_{l_2} = \|\mathbf{w}_l\|_{l_2,\mathbf{I}_{K\times K}}.$$

Suppose we have N assets, then $N = L \times K$. We can re-express Σ as

$$\Sigma = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1L} \\ A_{21} & A_{22} & \cdots & A_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ A_{L1} & A_{L2} & \cdots & A_{LL} \end{pmatrix}.$$

For l, $l' = 1, \ldots, L$, if $l = l'$, $A_{ll'}$ is a $K \times K$ covariance matrix of asset returns in group l. If $l \neq l'$, $A_{ll'}$ is a $K \times K$ matrix for covariances of asset returns between group l and l'.

We consider a special case of (12) in which $\Omega_l = A_{ll}$. Under this setting, the update form for $\mathbf{w}_l$ and γ can be solved explicitly. We summarize the algorithm as follows.

Algorithm 3. ***Block coordinate-wise descent update for mvp penalized by weighted Euclidean norm penalty***

1. *Fix* λ *at some non-negative constant level.*
2. *Initialize* $\mathbf{w} = N^{-1}\mathbf{1}_N$ *and* $\gamma > \lambda\sqrt{\max_{i=1,\ldots,N} \sigma_i^2}$.
3. *For* $l = 1, \ldots, L$, *and* $k > 0$,

$$\mathbf{w}_l^{(k)} \leftarrow \frac{1}{2}\left(1 - \frac{\lambda}{\Lambda_l^{(k)}\left(\gamma^{(k-1)}\right)}\right)_+ A_{ll}^{-1}\left(\gamma\mathbf{1}_K - B_l^{(k)}\right),$$

where

$$B_l^{(k)} = 2\left(\sum_{j<l} A_{lj}\mathbf{w}_j^{(k)} + \sum_{j>l} A_{lj}\mathbf{w}_j^{(k-1)}\right),$$

and

$$\Lambda_l^{(k)}\left(\gamma^{(k-1)}\right) = \left\|A_{ll}^{-\frac{1}{2}}\gamma^{(k-1)}\mathbf{1}_K - A_{ll}^{-\frac{1}{2}}B_l^{(k)}\right\|_{l_2}.$$

4. *For $k > 0$, update γ as*

$$\gamma^{(k)} \leftarrow \arg\min_{\gamma} \left(1 - \sum_{l \in S_l^{(k)}} \left[\frac{1}{2}\left(1 - \frac{\lambda}{\Lambda_l^{(k)}(\gamma)}\right)_+ A_{ll}^{-1}\left(\gamma \mathbf{1}_K - B_l^{(k)}\right)\right]\right)^2,$$

where $S_l^{(k)} = \left\{l : \mathbf{w}_l^{(k)} \neq \mathbf{0}\right\}$.

5. *Repeat 3 and 4 until $\mathbf{w}$ and γ have converged.*

The derivation of Algorithm 3 is shown in Supplementary Materials S.3.2.

We use the Fama and French 100 size and BM ratio portfolio data as an example. The data set contains weighted returns for the intersections of 10 market cap portfolios and 10 BM ratio portfolios. We categorize the 100 portfolios via two different methods. The first method is to group them according to 10 market cap levels. In each group, we have 10 different BM ratio portfolios. The second one is opposite: we group them according to 10 BM ratio levels. In each group, we have 10 different market cap portfolios. Thus the two settings both have $L = K = 10$.

Let size-bm and bm-size denote the first and second methods for grouping the assets respectively. We first compare the optimal weights from solving (12) with Algorithm 3 and `cvx`. The plots in the bottom panel of Figure 5 show the cumulative difference of the optimal weights from the two methods. Similar to the case of the weighted norm mvp, the difference between the solutions declines as λ increases. But on average, the cumulative difference is larger than the weighted norm mvp case, especially when λ is small.

Figure 8 shows the six performance measures. It can be seen that different grouping methods lead to similar results in the standard deviation of the out-of-sample portfolio returns, turnover rate, PAC, APS and γ. But for the Sharp ratio, the size-bm group mvp performs much better than the bm-size group mvp, which indicates that different grouping methods still can have an impact on the performance of the group portfolio optimization.

5.3. Other Possible Extensions and Limitations of The Method

In this section we outline some other possible extensions of the proposed method. We focus our discussions on how the proposed method can be applied to portfolio optimization problems with more flexible forms and nonconvex norm penalties, and limitations of the proposed method when some general risk measures are used.

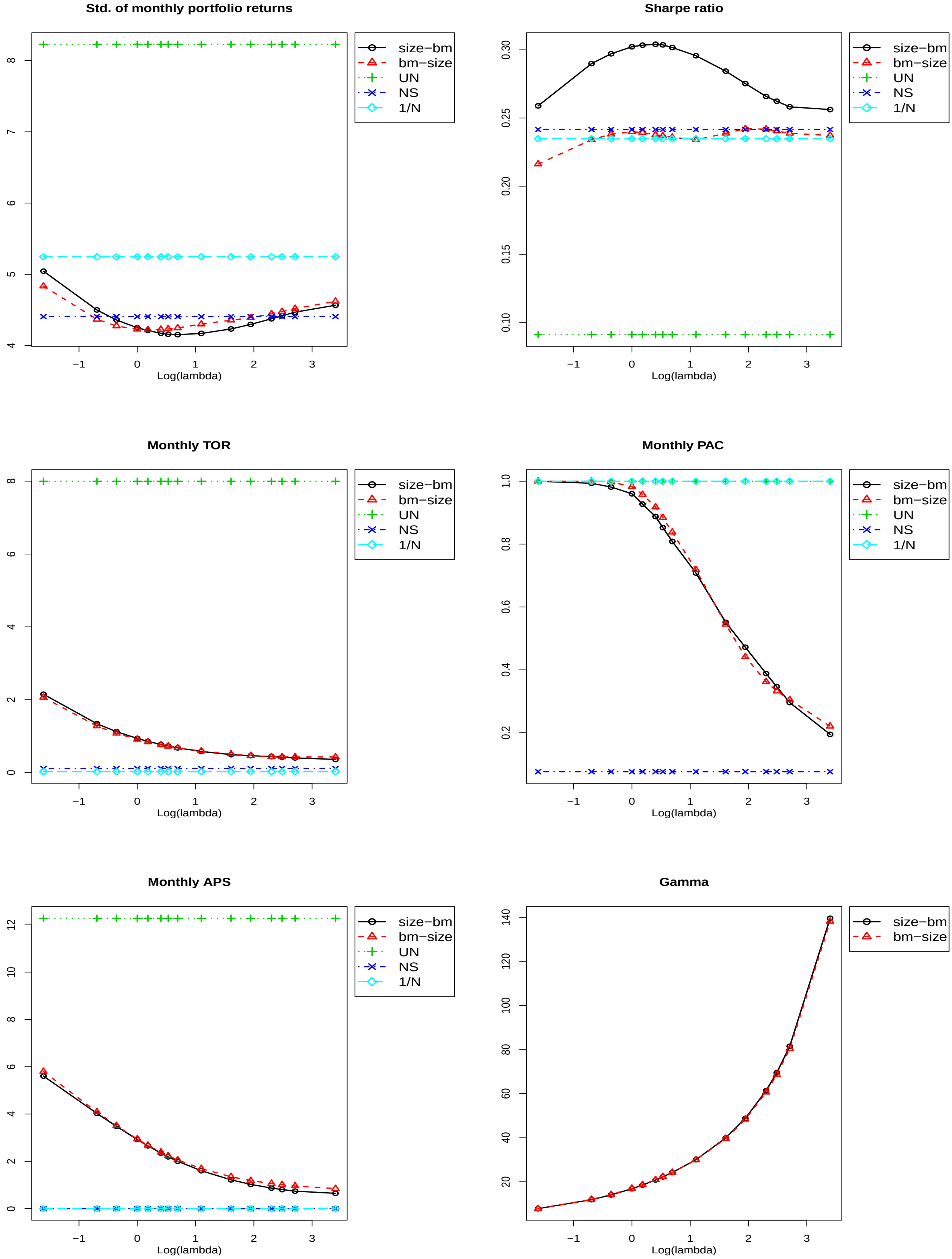

Figure 8: Standard deviation of out-of-sample portfolio returns, Sharpe ratio, average turnover rate (TOR), proportion of active constituents (PAC), absolute position of shortsales (APS) from the mvp with group selection method, no-shortsale mvp (NS), unconstrained mvp (UN) and 1/N portfolio, and optimal γ from the mvp with group selection method. The data used is the monthly return data of the FF 100 size and BM ratio portfolios from July-1973 to September-2009.

5.3.1. Choosing Optimal Portfolios via Utility Maximizations

So far we only apply the proposed method to the mvp optimization problems in which the objective function is the portfolio variance. It is not difficult to extend the method to the case when the objective function has an alternative form. For example, suppose the asset returns $\mathbf{R} \overset{i.i.d}{\sim} N(\mu, \Sigma)$, and the investor's objective function is an expectation of the exponential utility function:

$$\mathbb{E}\left(1 - \exp\left(-a\mathbf{w}^{\mathbf{T}}\mathbf{R}\right)\right) = 1 - \exp\left(-a\mathbf{w}^{\mathbf{T}}\mu + \frac{a^2}{2}\mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w}\right),$$

where a is a measure for the investor's risk aversion and is assumed to be positive. Maximizing the above expected utility is equivalent to minimizing the term inside the exponential. We may therefore formulate the penalized portfolio optimization as:

$$\min_{\mathbf{w}} -\tau\mathbf{w}^{\mathbf{T}}\mu + \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w} + \lambda Pen(\mathbf{w}) \qquad \text{subject to } \mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1, \tag{13}$$

where $\tau = 2a^{-1}$ is an inverse of the risk preference parameter (scaled by 2) and $Pen(\mathbf{w})$ is a convex penalty function of $\mathbf{w}$. The portfolio optimization (13) can be solved by the coordinate-wise descent algorithm if the penalty function $Pen(\mathbf{w})$ has one of the forms presented in the previous sections. For instance, if $Pen(\mathbf{w}) = \alpha\|\mathbf{w}\|_{l_1} + (1-\alpha)\|\mathbf{w}\|_{l_2}^2$, $\mathbf{w}$ and γ can be solved with the following updated form:

$$\begin{aligned}
w_i^{(k)} &\leftarrow \frac{ST\left(\gamma^{(k-1)} - \left(\tau\mu_i + z_i^{(k)}\right), \lambda\alpha\right)}{2\left(\sigma_i^2 + \lambda(1-\alpha)\right)}, \\
\gamma^{(k)} &\leftarrow \left[\sum_{i \in S_+^{(k)} \cup S_-^{(k)}} \frac{1}{2\left(\sigma_i^2 + \lambda(1-\alpha)\right)}\right]^{-1} \times \\
&\quad \left[1 + \sum_{i \in S_+^{(k)} \cup S_-^{(k)}} \frac{\tau\mu_i + z_i^{(k)}}{2\left(\sigma_i^2 + \lambda(1-\alpha)\right)} - \right. \\
&\quad \left. \lambda\alpha\left(\sum_{i \in S_-^{(k)}} \frac{1}{2\left(\sigma_i^2 + \lambda(1-\alpha)\right)} - \sum_{i \in S_+^{(k)}} \frac{1}{2\left(\sigma_i^2 + \lambda(1-\alpha)\right)}\right)\right].
\end{aligned}$$

Moreover, the proposed method can be applied to a more general case in which the investor's objective function is an expectation of a concave utility function $u(\cdot, \boldsymbol{\phi})$. Let $U(\mathbf{w}, \boldsymbol{\phi}) = \mathbb{E}\left(u\left(\mathbf{w}^{\mathbf{T}}\mathbf{R}, \boldsymbol{\phi}\right)\right)$, where $\boldsymbol{\phi}$ is a set of parameters of the investor's risk preferences. Note that $U(\cdot, \boldsymbol{\phi})$ is also concave since it is a linear combination of $u(\cdot, \boldsymbol{\phi})$ with nonnegative weights (probabilities). Suppose the investor chooses her optimal portfolio through the following penalized utility maximization:

$$\max_{\mathbf{w}} U\left(\mathbf{w}, \boldsymbol{\phi}\right) - \lambda Pen\left(\mathbf{w}\right) \qquad \text{subject to } \mathbf{w}^{\mathbf{T}}\mathbf{1} = 1. \tag{14}$$

By approximating the function $U\left(\mathbf{w}, \boldsymbol{\phi}\right)$ with the second order Taylor series expansion, the proposed method can be easily applied. Let $h\left(\mathbf{x}, \boldsymbol{\phi}\right)$ and $H\left(\mathbf{x}, \boldsymbol{\phi}\right)$ denote the gradient and Hessian of $U\left(\mathbf{w}, \boldsymbol{\phi}\right)$ with respect to $\mathbf{w}$ at point $\mathbf{x}$. The function $U\left(\mathbf{w}, \boldsymbol{\phi}\right)$ can be approximated by

$$\begin{aligned} U\left(\mathbf{w}, \boldsymbol{\phi}\right) &\approx U\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) + \left(\mathbf{w}^{\mathbf{T}} - \widetilde{\mathbf{w}}^{\mathbf{T}}\right) h\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) + \frac{\left(\mathbf{w}^{\mathbf{T}} - \widetilde{\mathbf{w}}^{\mathbf{T}}\right) H\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)\left(\mathbf{w} - \widetilde{\mathbf{w}}\right)}{2} \\ &= \frac{1}{2}\mathbf{w}^{\mathbf{T}} H\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)\mathbf{w} + \mathbf{w}^{\mathbf{T}} V_1\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) + V_2\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right), \end{aligned}$$

where $\widetilde{\mathbf{w}}$ is an $N \times 1$ deterministic vector such that $\widetilde{\mathbf{w}}^{\mathbf{T}}\mathbf{1} = 1$, and $V_1\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)$ and $V_2\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)$ are functions of $\widetilde{\mathbf{w}}$, which have the following expressions:

$$\begin{aligned} V_1\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) &= h\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) - H\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)\widetilde{\mathbf{w}}, \\ V_2\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) &= U\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) - \widetilde{\mathbf{w}}^{\mathbf{T}} h\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) + \frac{1}{2}\widetilde{\mathbf{w}}^{\mathbf{T}} H\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)\widetilde{\mathbf{w}}. \end{aligned}$$

Here $\dim\left(V_1\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)\right) = N \times 1$ and $V_2\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)$ is a scalar. To solve (14) we can iteratively solve the following modified optimization problem:

$$\max_{\mathbf{w}} \frac{1}{2}\mathbf{w}^{\mathbf{T}} H\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)\mathbf{w} + \mathbf{w}^{\mathbf{T}} V_1\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) - \lambda Pen\left(\mathbf{w}\right) \qquad \text{subject to } \mathbf{w}^{\mathbf{T}}\mathbf{1} = 1.$$

In each iteration, the vector $\widetilde{\mathbf{w}}$ used is the solved optimal weight vector from the previous iteration. The above optimization can be further reformulated as

$$\min_{\mathbf{w}} -\mathbf{w}^{\mathbf{T}} V_1\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right) + \frac{1}{2}\mathbf{w}^{\mathbf{T}} H_1\left(\widetilde{\mathbf{w}}, \boldsymbol{\phi}\right)\mathbf{w} + \lambda Pen\left(\mathbf{w}\right) \qquad \text{subject to } \mathbf{w}^{\mathbf{T}}\mathbf{1} = 1, \tag{15}$$

where $H_1\left(\cdot, \boldsymbol{\phi}\right) = -H\left(\cdot, \boldsymbol{\phi}\right)$. The optimization (15) has a similar form as (13): given $\widetilde{\mathbf{w}}$, the objective function (without the penalty) in (15) is a quadratic function of $\mathbf{w}$. Hence the optimization can be solved by using the coordinate-wise descent algorithm if the penalty function $Pen(\mathbf{w})$ has one of the forms presented in the previous sections.

5.3.2. Nonconvex Penalties

The penalty functions we consider in this paper are all convex functions of the portfolio weights. The main reason to require a penalty function convex is to keep the objective function plus the penalty convex. However, if with a nonconvex penalty the whole objective function is still convex, such requirement is not needed. Fan and Li (2001) and Zhang (2010) showed advantages of using nonconvex penalties on large dimensional regression estimation problems. Guan and Gray (2013) demonstrated how the nonconvex l_q norm penalty, $0 < q < 1$ can be used in the support vector machines (SVMs). As for portfolio optimization, Fastrich et al. (2012) demonstrated how imposing the nonconvex norm penalties can improve portfolio performances when the number of assets becomes large. Breheny and Huang (2011) and Mazumder et al. (2011) applied the coordinate-wise descent algorithms to solve large dimensional variable selection problems with the nonconvex norm penalties. Their idea is that, in order to keep the whole objective function convex, we can adjust the parameters inside the nonconvex penalties before or when running the coordinate-wise descent algorithms. Combining with such idea, the proposed method in this paper might also be applied to portfolio optimization problems with the nonconvex penalties, since the coordinate-wise descent algorithms used in the large dimensional variable selection and portfolio optimization problems bear some similarities. Detail procedures for such kinds of approaches are left for future research.

5.3.3. Limitations on Using More General Risk Measures or Utility Functions

When more general risk measures or utility functions are concerned in the portfolio optimization problems, the proposed method still have limitations in some situations. Among them, the worst one is that the risk measures are not convex or the utility functions are not concave, which may make the coordinate-wise descent algorithms fail to work. In addition, some issues of numerical evaluations for the risk measures or utility functions should also be addressed. For example, the expectation of the concave utility function $U(\mathbf{w})$ used in section 5.3.1 may not have a closed form. Hence the gradient $h(\mathbf{w})$ and Hessian $H(\mathbf{w})$ at $\widetilde{\mathbf{w}}$ are needed to be numerically evaluated, and the numerical evaluation errors may worsen qualities of the solutions when the number of assets becomes large.

The proposed method also meets difficulties when some of the constraints on the portfolio weights are not linear equalities and have complicated forms. Consider the following loss risk constraint:

$$P\left(\mathbf{w^T R} \le \zeta\right) \le \beta. \tag{16}$$

Normally the parameter $\zeta < 0$, and the constraint requires that probability of loss of a portfolio exceeding a certain level ζ should be less than a specified criterion β. Again suppose the asset returns $\mathbf{R} \overset{i.i.d}{\sim} N(\mu, \Sigma)$. The constraint (16) can be rewritten as

$$\mathbf{w^T}\mu + \Phi^{-1}(\beta)\sqrt{\mathbf{w^T}\Sigma\mathbf{w}} \ge \zeta, \tag{17}$$

where $\Phi^{-1}(.)$ is the inverse cumulative distribution function of the standard norm, and

$\beta \leq 0.5$ $(\Phi^{-1}(\beta) \leq 0)$. Note that (17) is an inequality and its left hand side is a nonlinear function of $\mathbf{w}$. Boyd and Vandenberghe (2004) showed that if the investor only considers maximizing the expected portfolio return, the portfolio optimization with constraint (17) and $\mathbf{w^T 1} = 1$ is a convex optimization and can be formulated as a second order cone programming, and adding a convex norm penalty in the portfolio choice problem will not change its nature of convex optimization. Nevertheless the method presented in this paper may not be able to directly solve it since the loss risk constraint (17) has a form of nonlinear inequality, which may cause difficulties on updating the Lagrange multipliers γ in the algorithms.

6. Conclusion

In this paper, we first develop the coordinate-wise descent algorithm for solving the minimum variance portfolio optimization with the weighted norm penalty. Our simulation results indicate that the proposed algorithm can perform well on solving the weighted norm constrained mvp optimization when the number of assets becomes large. We then show that the proposed method can be extended to portfolio optimization problems with different norm penalties. In addition, empirical properties of these norm constrained portfolios are investigated with two benchmark data sets.

In Section 4.5 we have compared Algorithm 1 with other optimization solvers, and shown that they generate almost identical optimal solutions when they are applied to solving the weighted norm mvp optimization. Similar results were also shown in Section 5.2 when solving the group portfolio optimization. However, properties of convergence rate of the proposed coordinate-wise descent algorithms are not explored in this paper. In fact, it is difficult to theoretically prove convergence and derive the convergence rate for a coordinate-wise descent algorithm based on a cyclic coordinate search strategy (Nesterov, 2012). In Supplementary Materials S.1 we show that when only the squared l_2 norm penalty is imposed, solving the penalized portfolio optimization with Algorithm 1 is equivalent to solving a system of linear equations with the Gauss-Seidel iteration. However, the result cannot be applied to more general cases such as the portfolio optimization problems that involve the l_1 norm penalty or other norm penalties. A more complete analysis on the issues of convergence rate of the coordinate-wise descent algorithms for the norm constrained portfolio optimization problems is left for future research.

There are other issues we have not addressed in this paper but may be worth for future research. The first one is how to choose the optimal penalty parameter λ and other tuning parameters. Yen (2012) showed that the optimal λ can be specified as a function of the sample size T and the number of available assets N. DeMiguel et al. (2009a) applied the cross validation method to minimize the portfolio variance or maximize the portfolio return to obtain the optimal λ. We may derive other optimal specifications for the penalty parameter through deeply investigating theoretical properties of the norm constrained portfolios. The other interesting issue is how a better estimated covariance matrix and return vector can improve the performances of the norm constrained portfolios. Fan, Zhang, and Yu (2012) showed that the covariance matrix estimated from the factor model seems to be better in reducing volatility of portfolio returns than the sample covariance matrix. It will be interesting to see how performances of the norm constrained portfolios can be improved when more sophisticated estimates for the covariance matrix or the return vector are plugged into the portfolio optimization problems.

Appendix: Derivations of the Upper Bound of the Penalty Parameter

Let $w_{ns,i}$ be the optimal weight of asset i from the no-shortsale mvp, and let $S_{ns} = \{i : w_{ns,i} > 0\}$. For the upper bound suppose $\lambda \geq \overline{\lambda}$, then the optimal weight vector from (2) is the same as the optimal no-shortsale weight vector. From (6), for $w_{ns,i} \geq 0$, $i = 1, \ldots, N$, we get

$$\begin{aligned} 2w_{ns,i}\sigma_i^2 + 2\sum_{j\neq i}^{N} w_{ns,j}\sigma_{ij} &= \gamma - \lambda, \text{if } w_{ns,i} > 0, \\ 2\left|\sum_{j\in S_{ns}} w_{ns,j}\sigma_{ij} - \gamma\right| &\leq \lambda, \qquad \text{if } w_{ns,i} = 0, \\ \mathbf{w}_{ns}^{\mathbf{T}}\mathbf{1}_N &= 1, \end{aligned}$$

where $\mathbf{w}_{ns}$ is the optimal weight vector of the no-shortsale mvp. Let $\mathbf{w}_{ns}^{+}$ denote the vector that only contains the non-zero components of $\mathbf{w}_{ns}$. Note that $\dim(\mathbf{w}_{ns}^{+}) = |S_{ns}| \times 1$ and $|S_{ns}|$ is the cardinality of the set S_{ns}. From the above system of equations, it can be shown that

$$2\sigma_{ns}^2 = \gamma - \lambda.$$

Here

$$\sigma_{ns}^2 = \mathbf{w}_{ns}^{+\mathbf{T}}\Sigma_{S_{ns}S_{ns}}\mathbf{w}_{ns}^{+},$$

is the in-sample variance of the no-shortsale mvp, and $\Sigma_{S_{ns}S_{ns}}$ is the $|S_{ns}| \times |S_{ns}|$ covariance matrix of the assets with non-zero weights. As for $w_{ns,i} = 0$, from the KKT conditions,

$$\gamma - \lambda \leq 2\sum_{j\in S_{ns}} w_{ns,j}\sigma_{ij} \leq \gamma + \lambda.$$

Since $2\sigma_{ns}^2 = \gamma - \lambda$, then $\gamma \geq \lambda$ and $2\sigma_{ns}^2 + 2\lambda = \gamma + \lambda$. It implies that

$$0 \leq \sum_{j \in S_{ns}} w_{ns,j}\sigma_{ij} - \sigma_{ns}^2 \leq \lambda.$$

Let

$$\zeta^* = \max_{i \notin S_{ns}} \sum_{j \in S_{ns}} w_{ns,j}\sigma_{ij} - \sigma_{ns}^2.$$

The quantity ζ^* may be used to estimate the upper bound $\overline{\lambda}$. To guarantee the estimated upper bound nonnegative, the following estimator can be used

$$\widehat{\overline{\lambda}} = \max(0, \zeta^*).$$

Practically, we can solve the no-shortsale mvp to obtain S_{ns} and relevant quantities, and $\widehat{\overline{\lambda}}$ can be easily constructed.

Acknowledgments

We thank editor, associate editor and two anonymous reviewers for their constructive and insightful suggestions. We also thank Yeutien Chou, Clifford Lam, Oliver Linton and Raman Uppal for comments on the initial draft of the paper. Tso-Jung Yen was supported by grants NSC 97-3112-B-001-020 and NSC 98-3112-B-001-027 in the National Research Program for Genomic Medicine.

References

Bai, J., Ng, S., October 2008. Forecasting economic time series using targeted predictors. Journal of Econometrics 146 (2), 304–317.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press, New York, NY, USA.

Breheny, P., Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Annals of Applied Statistics 5 (1), 232–253.

Brodie, J., Daubechies, I., Mol, C. D., Giannone, D., Loris, I., 2009. Sparse and stable markowitz portfolios. Proceedings of the National Academy of Sciences of the United States of America 106 (30), 12267–12272.

De Mol, C., Giannone, D., Reichlin, L., October 2008. Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? Journal of Econometrics 146 (2), 318–328.

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

DeMiguel, V., Garlappi, L., Uppal, R., 2009b. Optimal versus naive diversification: How inefficient is the *1/N* portfolio strategy? Review of Financial Studies 22 (5), 1915–1953.

DeMiguel, V., Nogales, F. J., Uppal, R., 2010. Stock return serial dependence and out-of-sample portfolio performance. SSRN eLibrary.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of Statistics 32 (2), 407–451.

El Karoui, N., 2009. High dimensional effects in the markowitz problem and other quadratic programs with linear equality constraints: risk underestimation. Technical Report 781, Department of Statistics, UC Berkeley.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96 (456), 1348–1360.

Fan, J., Zhang, J., Yu, K., 2012. Vast portfolio selection with gross-exposure constraints. Journal of the American Statistical Association 107 (498), 592–606.

Fastrich, B., Paterlini, S., Winker, P., 2012. Constructing optimal sparse portfolios using regularization methods. SSRN eLibrary.

Fastrich, B., Paterlini, S., Winker, P., 2013. Cardinality versus q-norm constraints for index tracking. Quantitative Finance, accepted.

Frahm, G., Christoph, M., 2010. Dominating estimators for minimum-variance portfolios. Journal of Econometrics 159 (2), 289–302.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. The Annals of Applied Statistics 1 (2), 302–332.

Friedman, J., Hastie, T., Tibshirani, R., Jan. 2010. A note on the group lasso and a sparse group lasso. ArXiv e-prints.

Gabaix, X., 2011. A sparsity-based model of bounded rationality. NBER Working Paper 16911.

Garleanu, N., Pedersen, L. H., 2011. Margin-based asset pricing and deviations from the law of one price. Review of Financial Studies 24 (6), 1980–2022.

Giamouridis, D., Paterlini, S., 2010. Regular(ized) hedge fund clones. Journal of Financial Research 33 (3), 223–247.

Goldfarb, D., Idnani, A., 1982. Dual and primal-dual methods for solving strictly convex quadratic programs. In: Hennart, J. (Ed.), Numerical Analysis. Vol. 909 of Lecture Notes in Mathematics. Springer Berlin Heidelberg, pp. 226–239.

Goldfarb, D., Idnani, A., 1983. A numerically stable dual method for solving strictly convex quadratic programs. Mathematical Programming 27, 1–33.

Gotoh, J.-Y., Takeda, A., November 2011. On the role of norm constraints in portfolio selection. Computational Management Science 8 (4), 323–353.

Grant, M., Boyd, S., May 2010. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx.

Guan, W., Gray, A., 2013. Sparse high-dimensional fractional-norm support vector machine via dc programming. Computational Statistics and Data Analysis 67 (0), 136 – 148.
URL `http://www.sciencedirect.com/science/article/pii/S0167947313000352`

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

Jorion, P., September 1986. Bayes-stein estimation for portfolio analysis. Journal of Financial and Quantitative Analysis 21 (03), 279–292.

Kan, R., Smith, D. R., 2008. The distribution of the sample minimum-variance frontier. Management Science 54 (7), 1364–1380.

Kan, R., Zhou, G. F., 2007. Optimal portfolio choice with parameter uncertainty. Journal of Financial and Quantitative Analysis 42, 621–656.

Lai, T. L., Xing, H., Chen, Z., 2011. Mean-variance portfolio optimization when means and covariances are unknow. The Annals of Applied Statistics 5 (2A), 798–823.

Ledoit, O., Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance 10 (5), 603–621.
URL `http://econpapers.repec.org/RePEc:eee:empfin:v:10:y:2003:i:5:p:603-621`

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88, 365–411.

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

Mazumder, R., Friedman, J. H., Hastie, T., 2011. Sparsenet : Coordinate descent with nonconvex penalties. Journal of the American Statistical Association 106 (495).

Meinshausen, N., Buhlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34 (3), 1436–1462.

Nesterov, Y., 2012. Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization 22 (2), 341–362.

Owen, A. B., 2007. A robust hybrid of lasso and ridge regression. Contemporary Mathematics 443, 59–72.

Saad, Y., 2003. Iterative Methods for Sparse Linear Systems, 2nd Edition. Society for Industrial and Applied Mathematics.

Takeda, A., Niranjan, M., Gotoh, J.-y., Kawahara, Y., 2013. Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios. Computational Management Science 10, 21–49.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58 (1), 267–288.

Tseng, P., June 2001. Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of Optimization Theory and Applications 109 (3), 475–494.

Tu, J., Zhou, G., 2011. Markowitz meets talmud: A combination of sophisticated and naive diversification strategies. Journal of Financial Economics 99, 204–215.

Vinciotti, V., Hashem, H., 2013. Robust methods for inferring sparse network structures. Computational Statistics and Data Analysis 67 (0), 84 – 94.
URL `http://www.sciencedirect.com/science/article/pii/S0167947313001655`

Welsch, R. E., Zhou, X., March 2007. Application of robust statistics to asset allocation models. Revstat 5 (1), 97–114.

Yen, Y. M., 2012. Sparse weighted norm minimum variance portfolio, phD. Thesis, London School of Economics and Political Science.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1), 49–67.

Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38 (2).

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2), 301–320.

Supplementary Materials

S.1. Solving the Squared l_2 Norm Penalized Portfolio with Algorithm 1 and its Relation to the Gauss-Seidel Iteration

Suppose $\mathbf{w}^* = (w_1^*, \ldots, w_N^*)$, and γ^* are the limits of convergence of $\mathbf{w}^{(k)}$ and $\gamma^{(k)}$. From Algorithm 1, at the limit,

$$\begin{aligned}
w_i^* &= \frac{ST\left(\gamma^* - z_i^*, \lambda\alpha\right)}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}, \\
\gamma^* &= \left[\sum_{i \in S_+^* \cup S_-^*} \frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right]^{-1} \left[1 + \sum_{i \in S_+^* \cup S_-^*} \frac{z_i^*}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} - \right. \\
&\qquad \left. \lambda\alpha\left(\sum_{i \in S_-^*} \frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} - \sum_{i \in S_+^*} \frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right)\right],
\end{aligned}$$

where

$$\begin{aligned}
z_i^* &= 2\left(\sum_{j<i} w_j^*\sigma_{ij} + \sum_{j>i} w_j^*\sigma_{ij}\right) = 2\sum_{j \neq i}^{N} w_j^*\sigma_{ij}, \\
S_+^* &= \left\{i : w_i^* > 0\right\}, \text{ and } S_-^* = \left\{i : w_i^* < 0\right\}.
\end{aligned}$$

Rearranging the above equations yields a system of equations

$$\begin{aligned}
2w_i^*\sigma_i^2 + 2\sum_{j \neq i, j \in S_+^* \cup S_-^*} w_j^*\sigma_{ij} + 2\lambda\left(1-\alpha\right)w_i^* - \gamma^* &= -\lambda\alpha, \text{ if } w_i^* > 0 \\
2w_i^*\sigma_i^2 + 2\sum_{j \neq i, j \in S_+^* \cup S_-^*} w_j^*\sigma_{ij} + 2\lambda\left(1-\alpha\right)w_i^* - \gamma^* &= \lambda\alpha, \quad \text{if } w_i^* > 0, \\
\sum_{i \in S_+^* \cup S_-^*} w_i &= 1,
\end{aligned}$$

and the condition $\left|2\sum_{j \neq i}^{N} w_j^*\sigma_{ij} - \gamma^*\right| \leq \lambda\alpha$, if $w_i^* = 0$ holds. These are just the KKT conditions (6) with $\mathbf{w}$ and γ replaced by $\mathbf{w}^*$ and γ^*. Therefore if Algorithm 1 converges, the convergence limit satisfies the KKT conditions, and hence it is the global minimum of (2).

We then investigate a special case of using Algorithm 1 when only the squared l_2 norm penalty is imposed. Our analysis is based on some well known results for the convergence properties of the Gauss-Seidel iteration for solving a *linear* system of equations. But before we proceed to the analysis, we should make a caution that our results can not be applied to more general cases such as the portfolio optimization problems have the

l_1 norm penalty or other norm penalties. The reason is that the KKT conditions for solving the portfolio optimization with the l_1 norm penalty (or other norm penalties) involve with the soft thresholding function, which makes the KKT conditions a nonlinear system of equations with sign-dependent structure, and the aforementioned convergence properties of the Gauss-Seidel iteration are not valid for such case.

When only the squared l_2 norm penalty is imposed, and the KKT conditions for the portfolio optimization can be expressed as $B^{\Sigma,\lambda}\mathbf{x} = \mathbf{d}$ where

$$B^{\Sigma,\lambda} = \begin{pmatrix} 2(\Sigma + \lambda \mathbf{I}) & \mathbf{1}_N \\ \mathbf{1}_N^{\mathbf{T}} & 0 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} \mathbf{w} \\ -\gamma \end{pmatrix}, \ \text{and } \mathbf{d} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix}.$$

Let $D^{\Sigma,\lambda}$, $E^{\Sigma,\lambda}$ and $F^{\Sigma,\lambda}$ be the matrices of the diagonal elements, strict lower and strict upper part of $B^{\Sigma,\lambda}$ respectively,

$$\begin{aligned} D^{\Sigma,\lambda} &= \begin{pmatrix} 2(\sigma_1^2 + \lambda) & 0 & \cdots & 0 \\ 0 & 2(\sigma_2^2 + \lambda) & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \\ E^{\Sigma,\lambda} &= \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 2\sigma_{21} & 0 & 0 & \cdots & 0 \\ 2\sigma_{31} & 2\sigma_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}, \text{ and } F^{\Sigma,\lambda} = \begin{pmatrix} 0 & 2\sigma_{12} & 2\sigma_{13} & \cdots & 1 \\ 0 & 0 & 2\sigma_{23} & \cdots & 1 \\ 0 & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}. \end{aligned}$$

Using Algorithm 1 is equivalent to updating $\mathbf{x}$ with

$$\mathbf{x}^{(k+1)} = \left(E^{\Sigma,\lambda} + D^{\Sigma,\lambda}\right)^{-1} \mathbf{d} - \left(E^{\Sigma,\lambda} + D^{\Sigma,\lambda}\right)^{-1} F^{\Sigma,\lambda}\mathbf{x}^{(k)}.$$

This is just the vector form of the Gauss-Seidel iteration for solving $B^{\Sigma,\lambda}\mathbf{x} = \mathbf{d}$. Under the Gauss-Seidel method, if the iteration converges, the limit will be guaranteed to be the solution of the system (Saad, 2003). As shown in Theorem 4.1 in Saad (2003), if $-\left(E^{\Sigma,\lambda} + D^{\Sigma,\lambda}\right)^{-1} F^{\Sigma,\lambda}$ is square and its the largest eigenvalue is less than 1, the iteration converges for any $\mathbf{d}$ and $\mathbf{x}^{(0)}$. One of the sufficient conditions to guarantee this is that

$$\left\| -\left(E^{\Sigma,\lambda} + D^{\Sigma,\lambda}\right)^{-1} F^{\Sigma,\lambda} \right\| < 1$$

for any matrix norm $\|.\|$. Another popular sufficient condition is that $B^{\Sigma,\lambda}$ is strictly diagonally dominant, i.e. $\left|b_{jj}^{\Sigma,\lambda}\right| > \sum_{i=1, i\neq j}^{N} \left|b_{ij}^{\Sigma,\lambda}\right|$. But note that the strictly diagonally dominant condition does not apply to $B^{\Sigma,\lambda}$ since its $(N+1, N+1)$th entity is 0. However, Algorithm 1 can still work well on convergence if a suitable choice of Σ and λ are used in the penalized squared l_2 norm portfolio optimization.

S.2. More Extensions

S.2.1. Weighted l_1 Norm Penalty

We can impose different penalties on the asset weights according to prior information on relative importances of the assets. Consider the following modified version of (2):

$$\min_{\mathbf{w}} \mathbf{w^T}\Sigma\mathbf{w} + \lambda\sum_{i=1}^{N}\nu_i|w_i| \qquad \text{subject to } \mathbf{w^T}\mathbf{1}_N = 1, \tag{18}$$

where ν_i is a nonnegative constant. The weighted l_1 penalty $\lambda\sum_{i=1}^{N}\nu_i|w_i|$ was initially used in the adaptive lasso estimation in Zou (2006). The updating forms used in solving (18) are given by

$$\begin{aligned} w_i^{(k)} &\leftarrow \frac{ST\left(\gamma^{(k-1)} - z_i^{(k)}, \lambda\nu_i\right)}{2\left(\sigma_i^2 + \lambda\nu_i\right)}, \\ \gamma^{(k)} &\leftarrow \left[\sum_{i\in S_+^{(k)}\cup S_-^{(k)}} \frac{1}{2\left(\sigma_i^2+\lambda\nu_i\right)}\right]^{-1} \times \\ &\quad \left[1 + \sum_{i\in S_+^{(k)}\cup S_-^{(k)}} \frac{z_i^{(k)}}{2\left(\sigma_i^2+\lambda\nu_i\right)} - \right. \\ &\quad \left. \lambda\left(\sum_{i\in S_-^{(k)}} \frac{\nu_i}{2\left(\sigma_i^2+\lambda\nu_i\right)} - \sum_{i\in S_+^{(k)}} \frac{\nu_i}{2\left(\sigma_i^2+\lambda\nu_i\right)}\right)\right]. \end{aligned}$$

As mentioned in Section S.4, ν_i can be viewed as a measure of transaction cost or the requirement of margin charged on asset i. We also can view ν_i as a function of the risk of asset i, such as its σ_i^2 or Beta. Imposing such ν_i on the penalty is then equivalent to incorporating more information about the investor's risk preference on the portfolio optimization.

S.2.2. Incorporating the Target Return Constraint

We also can incorporate the target return constraint $\mathbf{w}^T\mu = \overline{\mu}$ into (2), which makes the weighted norm mvp optimization become the weighted norm Markowitz portfolio optimization. When $\alpha = 1$, problem (2) with the additional target return constraint $\mathbf{w^T}\mu = \overline{\mu}$ is already considered by Brodie et al. (2009). The algorithm they proposed to solve the norm constrained Markowitz portfolio optimization is a LARS type algorithm.

Here we show that the coordinate-wise descent algorithm can also be easily applied to solve the same problem.

With the target return constraint $\mathbf{w^T}\mu = \overline{\mu}$, the Lagrangian for the portfolio optimization problem becomes

$$\begin{aligned} L(\mathbf{w}, \gamma; \Sigma, \lambda, \alpha) &= \mathbf{w^T}\Sigma\mathbf{w} + \lambda\alpha \|\mathbf{w}\|_{l_1} + \lambda(1-\alpha)\|\mathbf{w}\|_{l_2}^2 \\ &\quad -\gamma_1(\mathbf{w^T}\mathbf{1}_N - 1) - \gamma_2(\mathbf{w^T}\mu - \overline{\mu}) \end{aligned} \tag{19}$$

At the stationary point of (19), the followings should hold,

$$\begin{aligned} w_i &= \frac{\gamma_1 + \gamma_2\mu_i - z_i - \lambda\alpha}{2\left(\sigma_i^2 + \lambda(1-\alpha)\right)}, \text{ if } w_i > 0, \\ w_i &= \frac{\gamma_1 + \gamma_2\mu_i - z_i + \lambda\alpha}{2\left(\sigma_i^2 + \lambda(1-\alpha)\right)}, \text{ if } w_i < 0. \end{aligned}$$

Now we need to additionally update the Lagrangian parameter γ_2 of the target return constraint. Again we update $\mathbf{w}$, γ_1 and γ_2 sequentially. That is, the updated vector $\mathbf{w}$ is used to update γ_1, and then the updated $\mathbf{w}$ and γ_1 are used to update γ_2. We summarize the algorithm as follows.

Algorithm S 1. ***Coordinate-wise descent update for Markowitz portfolio penalized by a weighted l_1 and squared l_2 norm.***

1. *Fix λ, $\alpha \in [0, 1]$ and $\overline{\mu}$ at some constant levels.*
2. *Initialize $\mathbf{w}^{(0)} = N^{-1}\mathbf{1}_N$ and $\gamma_1^{(0)} + \gamma_2^{(0)} \times \max_i \mu_i > \lambda$*
3. *For $i = 1, \ldots, N$, and $k > 0$,*

$$w_i^{(k)} \leftarrow \frac{ST\left(\gamma_1^{(k-1)} + \gamma_2^{(k-1)}\mu_i - z_i^{(k)}, \lambda\alpha\right)}{2\left(\sigma_i^2 + \lambda(1-\alpha)\right)},$$

where

$$z_i^{(k)} = 2(\sum_{j<i} w_j^{(k)}\sigma_{ij} + \sum_{j>i} w_j^{(k-1)}\sigma_{ij}).$$

4. *For $k > 0$, update γ_1 as*

$$\gamma_1^{(k)} \leftarrow \left[\sum_{i \in S_+^{(k)} \cup S_-^{(k)}} \frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right]^{-1} \times$$
$$\left[1 - \sum_{i \in S_+^{(k)} \cup S_-^{(k)}} \frac{\gamma_2^{(k-1)}\mu_i - z_i^{(k)}}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} - \right.$$
$$\left.\lambda\alpha\left(\sum_{i \in S_-^{(k)}} \frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} - \sum_{i \in S_+^{(k)}} \frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right)\right],$$

where $S_+^{(k)} = \left\{i : w_i^{(k)} > 0\right\}$ and $S_-^{(k)} = \left\{i : w_i^{(k)} < 0\right\}$.

5. *For $k > 0$, update γ_2 as*

$$\gamma_2^{(k)} \leftarrow \left[\sum_{i \in S_+^{(k)} \cup S_-^{(k)}} \frac{\mu_i^2}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right]^{-1} \times$$
$$\left[\overline{\mu} - \sum_{i \in S_+^{(k)} \cup S_-^{(k)}} \frac{\left(\gamma_1^{(k)} - z_i^{(k)}\right)\mu_i}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} - \right.$$
$$\left.\lambda\alpha\left(\sum_{i \in S_-^{(k)}} \frac{\mu_i}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} - \sum_{i \in S_+^{(k)}} \frac{\mu_i}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right)\right].$$

6. *Repeat 3 to 5 until $\mathbf{w}^{(k)}$, $\gamma_1^{(k)}$ and $\gamma_2^{(k)}$ have converged.*

The derivations of the sequentially updating procedure and Algorithm S1 are shown in Supplementary materials S.3.3.

We then compare Algorithm S1 with `solve.QP` in R package `quadprog` for solving the optimal no-shortsale and unconstrained Markowitz portfolios. Following the same argument stated in Section 3.2, it can be shown that when $\alpha = 1$, if λ is no less than a certain threshold, the optimal weights from minimizing (19) and from the no-shortsale Markowitz portfolio optimization will be identical. Since here we incorporate an additional constraint, the estimated upper bound of λ from (8) will no longer guarantee that we can obtain the optimal no-shortsale weights when Algorithm S1 is used. Therefore the following alternative estimate for the upper bound is used:

$$\begin{aligned}\widehat{\zeta}_{mk} &= \max_{i \notin S_{nsmk}} \sum_{j \in S_{nsmk}} w_{nsmk,j}\sigma_{ij}, \\ \widehat{\widehat{\lambda}}_{mk} &= \max(0, \widehat{\zeta}_{mk}). \end{aligned} \quad (20)$$

Here $w_{nsmk,j}$ is the weight and S_{nsmk} is the set of active assets in the optimal no-shortsale Markowitz portfolio.

The data generated for the simulations are the same as those used in Section 4.5. Here μ_i is estimated with sample mean of the $1.2N$ return observations of asset i, and the target return is set $\overline{\mu} = N^{-1} \sum_{i=1}^{N} \mu_i$. Table S1 shows the results which indicate that the relative performances of Algorithm S1 and `solve.QP` for solving the no-shortsale Markowitz portfolio optimization have significant differences when the number of assets become large. Also except for $\lambda = 0$, the average computational time for different regularization levels only have mild differences (all within 1 second). The results shown here are qualitatively similar to those shown in Table 1.

To see how the norm constraint affects the portfolio performance, Figure S3 and S4 compare the sample averages of out-of-sample portfolio returns $\overline{\overline{r}}_{por}$ and the other five performance measures of the weighted norm Markowitz portfolio and the other three benchmark portfolios: 1/N, no-shortsale Markowitz and unconstrained Markowitz portfolios. μ is estimated with the rolling sample mean of the asset returns in previous 120 periods. We set the (monthly) target return $\overline{\mu} = a \times N^{-1} \sum_{i=1}^{N} \mu_i$. For the weighted norm Markowitz portfolio, we set $a = 1, 1.5$ and 2 and $\alpha = 1$ and 0.2. For the no-shortsale and unconstrained Markowitz portfolios, we keep $a = 1$.

On average, a higher target return (a higher a) leads to a higher $\overline{\overline{r}}_{por}$. But for different data sets, $\widehat{\sigma}_{por}$ and $\widehat{SR}_{por}$ exhibit different behaviors when the target return changes. Given λ fixed, in the FF-48 case, $\widehat{\sigma}_{por}$ is much higher when $a = 2$ than $a = 1$ and 1.5. In the FF-100 case, however, the portfolio volatility changes only mildly when the parameter a varies. Consequently as a increases, it tends to result in higher Sharpe ratio in the FF-100 case. Given a suitable penalty parameter, Sharpe ratio of the norm constrained Markowitz portfolio can be much higher than that of the benchmark cases, but the Sharpe ratio also varies dramatically across different a, especially in the FF-48 case. In addition, when λ is small, comparing with the weighted norm mvp, the turnover rate of the weighted norm Markowitz portfolio is obviously higher. In summary, performance of the weighted norm Markowitz portfolio is very sensitive to the choices of the target return and penalty parameter.

S.3. Derivations of the Algorithms

S.3.1. Coordinate-Wise Descent Algorithm on the MVP Constrained by the Berhu Penalty

With the berhu penalty, we can construct a Lagrangian for the corresponding penalized mvp optimization problem. The first order derivative of the Lagrangian with respect to w_i is given by

$$2w_i\sigma_i^2 + 2\sum_{j\neq i}^{N} w_j\sigma_{ij} - \gamma + \lambda sign\left(w_i\right)\mathbb{I}\left\{|w_i| < \delta\right\} + \lambda\frac{w_i}{\delta}\mathbb{I}\left\{|w_i| \geq \delta\right\} = 0 \qquad (21)$$

for $i = 1, 2, \cdots, N$. It can be shown that the following should hold when solving (21) with respect to w_i:

$$\begin{aligned}
2w_i\sigma_i^2 + 2\sum_{j\neq i}^{N} w_j\sigma_{ij} - \gamma &= -\lambda \;\text{ if } 0 < w_i < \delta,\\
2w_i\sigma_i^2 + 2\sum_{j\neq i}^{N} w_j\sigma_{ij} - \gamma &= \lambda \quad \text{if } -\delta < w_i < 0,\\
\left|2w_i\sigma_i^2 + 2\sum_{j\neq i}^{N} w_j\sigma_{ij} - \gamma\right| &\leq \lambda \quad \text{if } w_i = 0,\\
2w_i\sigma_i^2 + 2\sum_{j\neq i}^{N} w_j\sigma_{ij} - \gamma + \frac{\lambda w_i}{\delta} &= 0 \quad\; \text{if } \delta \leq |w_i|\,,\\
\mathbf{w^T1}_N &= 1.
\end{aligned}$$

Fixing w_j for $j = 1, \ldots, N$ and $i \neq j$, we can solve the above equation for w_i. Now let $z_i = 2\sum_{j\neq i}^{N} w_j\sigma_{ij}$. When $\delta \leq |w_i|\,,\ w_i = (2\sigma_i^2 + \delta^{-1}\lambda)^{-1}(\gamma - z_i)$. Since $2\sigma_i^2 + \delta^{-1}\lambda > 0$, $\delta \leq |w_i|$ implies that $|\gamma - z_i| \geq 2\sigma_i^2\delta + \lambda$. When $0 < w_i < \delta$, $w_i = (2\sigma_i^2)^{-1}(\gamma - z_i - \lambda)$, and this implies that $\gamma - z_i < 2\sigma_i^2\delta + \lambda$. When $-\delta < w_i < 0$, $w_i = (2\sigma_i^2)^{-1}(\gamma - z_i + \lambda)$, and it implies that $\gamma - z_i > -2\sigma_i^2\delta - \lambda$.

With the arguments given above, we propose the following form to update w_i:

$$w_i \leftarrow \begin{cases} \frac{ST(\gamma - z_i, \lambda)}{2\sigma_i^2}, & \text{if } |\gamma - z_i| < 2\sigma_i^2\delta + \lambda,\\ \frac{\gamma - z_i}{2\sigma_i^2 + \frac{\lambda}{\delta}}, & \text{if } |\gamma - z_i| \geq 2\sigma_i^2\delta + \lambda. \end{cases}$$

Since the updating form for w_i is also linear in γ, we can again solve for γ by using $\mathbf{w^T1}_N = 1$. Let $\mathbf{\Delta}^- = \{i : |\gamma - z_i| < 2\sigma_i^2\delta + \lambda\}$ and $\mathbf{\Delta}^+ = \{i : |\gamma - z_i| \geq 2\sigma_i^2\delta + \lambda\}$. We then have

$$\begin{aligned}\mathbf{w}^{\mathbf{T}}\mathbf{1}_N &= \gamma\left(\sum_{i\in(S^{+}\cap\boldsymbol{\Delta}^{-})\cup(S^{-}\cap\boldsymbol{\Delta}^{-})}\frac{1}{2\sigma_i^2}+\sum_{i\in\boldsymbol{\Delta}^{+}}\frac{1}{2\sigma_i^2+\delta^{-1}\lambda}\right)-\\ &\quad\left(\sum_{i\in(S^{+}\cap\boldsymbol{\Delta}^{-})\cup(S^{-}\cap\boldsymbol{\Delta}^{-})}\frac{z_i}{2\sigma_i^2}+\sum_{i\in\boldsymbol{\Delta}^{+}}\frac{z_i}{2\sigma_i^2+\delta^{-1}\lambda}\right)+\\ &\quad\lambda\left(\sum_{i\in S^{-}\cap\boldsymbol{\Delta}^{-}}\frac{1}{2\sigma_i^2}-\sum_{i\in S^{+}\cap\boldsymbol{\Delta}^{-}}\frac{1}{2\sigma_i^2}\right).\end{aligned}$$

With the constraint $\mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1$, we propose the following form to update γ,

$$\begin{aligned}\gamma &\leftarrow \left[\sum_{i\in(S_{+}\cap\boldsymbol{\Delta}_{-})\cup(S_{-}\cap\boldsymbol{\Delta}_{-})}\frac{1}{2\sigma_i^2}+\sum_{i\in\boldsymbol{\Delta}_{+}}\frac{1}{2\sigma_i^2+\delta^{-1}\lambda}\right]\times\\ &\quad\left[1+\left(\sum_{i\in(S_{+}\cap\boldsymbol{\Delta}_{-})\cup(S_{-}\cap\boldsymbol{\Delta}_{-})}\frac{z_i}{2\sigma_i^2}+\sum_{i\in\cap\boldsymbol{\Delta}_{+}}\frac{z_i}{2\sigma_i^2+\delta^{-1}\lambda}\right)-\right.\\ &\quad\left.\lambda\left(\sum_{i\in S_{-}\cap\boldsymbol{\Delta}_{-}}\frac{1}{2\sigma_i^2}-\sum_{i\in S_{+}\cap\boldsymbol{\Delta}_{-}}\frac{1}{2\sigma_i^2}\right)\right].\end{aligned}$$

Figure S1 shows profiles of portfolio weights, proportion of active constituents, and proportion of shortsales under different δ and $\log(\lambda)$ when the berhu penalty is imposed. The data used here is the same as the one used to generate Figure 1. When $\delta = 1$, the profiles are almost the same as those in the weighted mvp case when only the l_1 penalty is active. However, when δ is less than 1, the profiles look quite different from those of the weighted norm mvp.

S.3.2. Block Coordinate-Wise Descent Algorithm for the MVP Constrained by the Weighted Euclidean Norm Penalty

Under this restriction $\Omega_l = A_{ll}$, (12) becomes becomes

$$\min_{\mathbf{w}} \mathbf{w}^{\mathbf{T}}\Sigma\mathbf{w}+\lambda\sum_{l=1}^{L}\|\mathbf{w}_l\|_{l_2,A_{ll}} \qquad \text{subject to } \mathbf{w}^{\mathbf{T}}\mathbf{1}_N = 1. \tag{22}$$

We can reparameterize $\mathbf{w}_l$ as

$$\mathbf{w}_l = A_{ll}^{-\frac{1}{2}}\mathbf{x}_l,$$

where $\mathbf{x}_l$ is also a $K\times 1$ vector. Therefore $\|\mathbf{w}_l\|_{l_2,A_{ll}} = \|\mathbf{x}_l\|_{l_2}$.

Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$, then problem (22) becomes

$$\min_{\mathbf{x}} \mathbf{x}^{\mathbf{T}} \Sigma' \mathbf{x} + \lambda \sum_{l=1}^{L} \|\mathbf{x}_l\|_{l_2} \qquad \text{subject to} \quad \sum_{l=1}^{L} \mathbf{x}^{\mathbf{T}}{}_l A_{ll}^{-\frac{1}{2}} \mathbf{1}_K = 1, \tag{23}$$

where

$$\Sigma' = \begin{pmatrix} 1 & A_{11}^{-\frac{1}{2}} A_{12} A_{22}^{-\frac{1}{2}} & \cdots & A_{11}^{-\frac{1}{2}} A_{1L} A_{LL}^{-\frac{1}{2}} \\ A_{22}^{-\frac{1}{2}} A_{21} A_{11}^{-\frac{1}{2}} & 1 & \cdots & A_{22}^{-\frac{1}{2}} A_{2L} A_{LL}^{-\frac{1}{2}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{LL}^{-\frac{1}{2}} A_{L1} A_{11}^{-\frac{1}{2}} & A_{LL}^{-\frac{1}{2}} A_{L2} A_{22}^{-\frac{1}{2}} & \cdots & 1 \end{pmatrix}$$

It can be shown that, under the KKT conditions,

$$\begin{aligned} 2\mathbf{x}_l + 2A_{ll}^{-\frac{1}{2}} \left(\sum_{j \neq l}^{L} A_{lj} A_{jj}^{-\frac{1}{2}} \mathbf{x}_j \right) + \lambda \frac{\mathbf{x}_l}{\|\mathbf{x}_l\|_{l_2}} - \gamma A_{ll}^{-\frac{1}{2}} \mathbf{1}_K &= \mathbf{0}, \text{ If } \mathbf{x}_l \neq \mathbf{0}, \\ 2A_{ll}^{-\frac{1}{2}} \left(\sum_{j \neq l}^{L} A_{lj} A_{jj}^{-\frac{1}{2}} \mathbf{x}_j \right) + \lambda \mathbf{s}_l - \gamma A_{ll}^{-\frac{1}{2}} \mathbf{1}_K &= \mathbf{0}, \text{ If } \mathbf{x}_l = \mathbf{0}, \\ \sum_{l=1}^{L} \mathbf{x}_l^T A_{ll}^{-\frac{1}{2}} \mathbf{1}_K &= 1 \end{aligned} \tag{24}$$

for $l = 1, \ldots, L$, where $\mathbf{s}_l$ is a $K \times 1$ vector, and $\|\mathbf{s}_l\|_{l_2} \leq 1$. Let $B_l = 2 \sum_{j \neq l}^{L} A_{lj} A_{jj}^{-\frac{1}{2}} \mathbf{x}_j$ and $\Lambda_l(\gamma) = \left\| A_{ll}^{-\frac{1}{2}} \gamma \mathbf{1}_K - A_{ll}^{-\frac{1}{2}} B_l \right\|_{l_2}$. The necessary and sufficient condition for $\mathbf{x}_l = \mathbf{0}$ is

$$\Lambda_l(\gamma) \leq \lambda. \tag{25}$$

From the first line of (24), if $\mathbf{x}_l \neq \mathbf{0}$,, then

$$\mathbf{x}_l = \frac{A_{ll}^{-\frac{1}{2}} (\gamma \mathbf{1}_K - B_l)}{2 + \lambda \|\mathbf{x}_l\|_{l_2}^{-1}}.$$

As $\mathbf{x}_l \neq \mathbf{0}$ implies $\mathbf{w}_l \neq \mathbf{0}$, so

$$\mathbf{w}_l = A_{ll}^{-\frac{1}{2}} \mathbf{x}_l = \frac{A_{ll}^{-1} (\gamma \mathbf{1}_K - B_l)}{2 + \lambda \|\mathbf{w}_l\|_{l_2, A_{ll}}^{-1}}.$$

We now show $\|\mathbf{w}_l\|_{l_2, A_{ll}}$ can be solved as a function of γ. To see this, from (24) we know that

$$\left(2 + \frac{\lambda}{\|\mathbf{w}_l\|_{l_2, A_{ll}}} \right) A_{ll}^{\frac{1}{2}} \mathbf{w}_l = A_{ll}^{-\frac{1}{2}} (\gamma \mathbf{1}_K - B_l).$$

Then

$$\mathbf{w^T}_l A_{ll}^{\frac{1}{2}} \left(4 + \frac{4\lambda}{\|\mathbf{w}_l\|_{l_2,A_{ll}}} + \frac{\lambda^2}{\|\mathbf{w}_l\|_{l_2,A_{ll}}^2}\right) A_{ll}^{\frac{1}{2}} \mathbf{w}_l = \left(\gamma \mathbf{1}_K - B_l\right)^{\mathbf{T}} A_{ll}^{-1} \left(\gamma \mathbf{1}_K - B_l\right).$$

Thus we can obtain

$$\left(2\|\mathbf{w}_l\|_{l_2,A_{ll}} + \lambda\right)^2 = \Lambda_l^2(\gamma).$$

Since $\|\mathbf{w}_l\|_{l_2,A_{ll}} > 0$ when $\mathbf{w}_l \neq \mathbf{0}$ and $\lambda \geq 0$, if $\Lambda_l - \lambda > 0$, the solution for $\|\mathbf{w}_l\|_{l_2,A_{ll}}$ is given by

$$\|\mathbf{w}_l\|_{l_2,A_{ll}} = \frac{\Lambda_l(\gamma) - \lambda}{2}.$$

Therefore if $\mathbf{w}_l \neq \mathbf{0}$,

$$\mathbf{w}_l = \frac{1}{2}\left(1 - \frac{\lambda}{\Lambda_l(\gamma)}\right) A_{ll}^{-1} \left(\gamma \mathbf{1}_K - B_l\right).$$

Combining with the group-level test condition (25), we propose the following form to update the weight vector $\mathbf{w}_l$,

$$\mathbf{w}_l \leftarrow \frac{1}{2}\left(1 - \frac{\lambda}{\Lambda_l(\gamma)}\right)_+ A_{ll}^{-1} \left(\gamma \mathbf{1}_K - B_l\right),$$

where $(x)_+ := \max(x, 0)$. For updating γ, we can use the constraint $\mathbf{w^T 1}_N = 1$. However, since the weights are non-linear in γ, we cannot update γ as we did in the previous sections. Let $S^l = \{l : \mathbf{w}_l \neq \mathbf{0}\}$. Practically, we can update γ by

$$\gamma \leftarrow \arg\min_{\gamma} \left(1 - \sum_{l \in S^l} \left[\frac{1}{2}\left(1 - \frac{\lambda}{\Omega_l(\gamma)}\right)_+ A_{ll}^{-1} \left(\gamma \mathbf{1}_K - B_l\right)\right]\right)^2.$$

To implement the algorithm, we can set initial value of each weight $w_1^{(0)} = w_2^{(0)} = \cdots = w_N^{(0)} = N^{-1}$, and $\gamma^{(0)} > \lambda\sqrt{\max_{i=1,\ldots,N} \sigma_i^2}$. The algorithm starts from updating $\mathbf{w}_1$, $\mathbf{w}_2, \ldots$, and $\mathbf{w}_L$ sequentially, and then the updated vector $\mathbf{w}$ is used to update γ. The procedure is repeated until $\mathbf{w}$ and γ have converged.

Figure S2 illustrates profiles of portfolio weights, proportion of constituents, and proportion of shortsales of the group mvp. The portfolio weights shown here are aggregations of weights within a certain group. The samples used here span from February of 1980 to January of 1990. It can be seen that (within the sampling period) portfolios with large market cap values or high BM ratios are more likely to be selected.

S.3.3. Coordinate-Wise Descent Algorithm on the Weighted Norm Markowitz Portfolio

With the target return constraint $\mathbf{w^T}\mu = \overline{\mu}$, the Lagrangian corresponding to the optimization problem becomes

$$\begin{aligned} L(\mathbf{w}, \gamma; \Sigma, \lambda, \alpha) &= \mathbf{w^T}\Sigma\mathbf{w} + \lambda\alpha \left\|\mathbf{w}\right\|_{l_1} + \lambda\left(1-\alpha\right)\left\|\mathbf{w}\right\|_{l_2}^2 - \\ &\quad \gamma_1(\mathbf{w^T}\mathbf{1}_N - 1) - \gamma_2(\mathbf{w^T}\mu - \overline{\mu}) \end{aligned}$$

The KKT conditions corresponding to the above Lagrangian are

$$\begin{aligned} 2w_i\sigma_i^2 + 2\sum_{j\neq i}^{N} w_j\sigma_{ij} + 2\lambda\left(1-\alpha\right)w_i - \gamma_1 - \gamma_2\mu_i &= -\lambda\alpha \text{ if } w_i > 0, \\ 2w_i\sigma_i^2 + 2\sum_{j\neq i}^{N} w_j\sigma_{ij} + 2\lambda\left(1-\alpha\right)w_i - \gamma_1 - \gamma_2\mu_i &= \lambda\alpha \quad \text{if } w_i < 0, \\ \left|2\sum_{j\neq i}^{N} w_j\sigma_{ij} - \gamma_1 - \gamma_2\mu_i\right| &\leq \lambda\alpha \quad \text{if } w_i = 0, \end{aligned}$$

also with $\mathbf{w^T}\mathbf{1}_N = 1$ and $\mathbf{w^T}\mu = \overline{\mu}$ held. At the stationary point,

$$\begin{aligned} w_i &= \frac{\gamma_1 + \gamma_2\mu_i - z_i - \lambda\alpha}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}, \text{ if } w_i > 0, \\ w_i &= \frac{\gamma_1 + \gamma_2\mu_i - z_i + \lambda\alpha}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}, \text{ if } w_i < 0. \end{aligned}$$

Now we need to additionally update the Lagrangian parameter γ_2 of the target return constraint. Again we update $\mathbf{w}$, γ_1 and γ_2 sequentially. That is, the updated vector $\mathbf{w}$ is used to update γ_1, and then the updated $\mathbf{w}$ and γ_1 are used to update γ_2.

Following similar fashion as in Section 3.2, we know that

$$\begin{aligned} \mathbf{w^T}\mathbf{1}_N &= \gamma_1\left(\sum_{i\in S_+\cup S_-}\frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right) + \sum_{i\in S_+\cup S_-}\frac{\gamma_2\mu_i - z_i}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} + \\ &\quad \lambda\alpha\left(\sum_{i\in S_-}\frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} - \sum_{i\in S_+}\frac{1}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right), \\ \mathbf{w^T}\mu &= \gamma_1\left(\sum_{i\in S_+\cup S_-}\frac{\mu_i}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right) + \sum_{i\in S_+\cup S_-}\frac{\gamma_2\mu_i^2 - z_i\mu_i}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} + \\ &\quad \lambda\alpha\left(\sum_{i\in S_-}\frac{\mu_i}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)} - \sum_{i\in S_+}\frac{\mu_i}{2\left(\sigma_i^2 + \lambda\left(1-\alpha\right)\right)}\right) \end{aligned}$$

By $\mathbf{w^T 1}_N = 1$, given γ_2, the following form is proposed for updating the γ_1

$$\gamma_1 \leftarrow \frac{1 - \sum_{i\in S_+\cup S_-} \frac{\gamma_2\mu_i - z_i}{2\left(\sigma_i^2+\lambda(1-\alpha)\right)} - \lambda\alpha\left(\sum_{i\in S_-}\frac{1}{2\left(\sigma_i^2+\lambda(1-\alpha)\right)} - \sum_{i\in S_+}\frac{1}{2\left(\sigma_i^2+\lambda(1-\alpha)\right)}\right)}{\left(\sum_{i\in S_+\cup S_-}\frac{1}{2\left(\sigma_i^2+\lambda(1-\alpha)\right)}\right)}.$$

Then by $\mathbf{w^T}\mu = \overline{\mu}$, the following form is proposed for updating γ_2

$$\gamma_2 \leftarrow \frac{\overline{\mu} - \sum_{i\in S_+\cup S_-} \frac{(\gamma_1 - z_i)\mu_i}{2\left(\sigma_i^2+\lambda(1-\alpha)\right)} - \lambda\alpha\left(\sum_{i\in S_-}\frac{\mu_i}{2\left(\sigma_i^2+\lambda(1-\alpha)\right)} - \sum_{i\in S_+}\frac{\mu_i}{2\left(\sigma_i^2+\lambda(1-\alpha)\right)}\right)}{\left(\sum_{i\in S_+\cup S_-}\frac{\mu_i^2}{2\left(\sigma_i^2+\lambda(1-\alpha)\right)}\right)}$$

To implement the algorithm, we can set initial value of each weight $w_1^{(0)} = w_2^{(0)} = \cdots = w_p^{(0)} = N^{-1}$, and $\gamma_1^{(0)} + \gamma_2^{(0)} \times \max_i \mu_i > \lambda$. The algorithm starts from updating w_1, $w_2, \ldots,$ and w_N sequentially, and then the updated vector $\mathbf{w}$ is used to update γ_1, and then γ_2. The procedure terminates until $\mathbf{w}$, γ_1 and γ_2 have converged. 32. In lines 18 to 20 in the initial manuscript, the statement is not clear and misleading. The point should be addressed is that "how a (small) change of the optimal weight affects the (in-sample) portfolio variance" rather than "whether including an asset or not affects the (in-sample) portfolio variance". We have rewritten the statements more precisely as:

...At the stationary point, if the small increase of the asset's weight causes a large (small) enough increment of the portfolio variance, say $\gamma + \lambda_1$ ($\gamma - \lambda_1$), then the asset's weight should have a negative (positive) sign. If the increase of the asset's weight only causes a mild increase of the portfolio variance, say between $[\gamma - \lambda_1,\ \gamma + \lambda_1]$, then the asset will not be included in the portfolio. ...

We should address that the whole argument is based on the assumption that the investor view $\hat{\Sigma}' = \hat{\Sigma} + \lambda_2 \mathbf{I}_{NN}$ as the "valid" in-sample estimate for the covariance matrix and the portfolio optimization is only subject to the full investment constraint. As such, the weighted norm mvp portfolio optimization can be viewed as a penalized l_1 norm mvp portfolio optimization in which the portfolio variance used is $\mathbf{w^T}\hat{\Sigma}'\mathbf{w}$. We also notice that in the weighted norm portfolio optimization, at the stationary point, including the asset which was excluded from the optimal portfolio indeed might decrease the in-sample portfolio variance $\mathbf{w^T}\hat{\Sigma}'\mathbf{w}$, but it will definitely do no help on decreasing the whole objective function $(\mathbf{w^T}\hat{\Sigma}'\mathbf{w} + \lambda_1 \|\mathbf{w}\|_{l_1})$.

It can be proven that $\gamma - \lambda_1 > 0$ at the stationary point. To see this, let s denote the set of active assets. At the stationary point, the following holds for portfolio weights of the active assets:

$$2\left(\hat{\Sigma}_{ss}+\lambda_2\mathbf{I}_{ss}\right)\mathbf{w}_s=\gamma\mathbf{1}-\lambda_1 sign\left(\mathbf{w}_s\right),$$

where $\hat{\Sigma}_{ss}$ is the sample covariance matrix of the $|s|$ active assets, $\mathbf{I}_{ss}$ is a $|s|\times|s|$ diagonal matrix, $\mathbf{w}_s$ is the solved optimal portfolio weight vector for the active assets. Multiplying both side with $\mathbf{w}_s^{\mathbf{T}}$, then

$$2\mathbf{w}_s^{\mathbf{T}}\left(\hat{\Sigma}_{ss}+\lambda_2\mathbf{I}_{ss}\right)\mathbf{w}_s=\gamma\mathbf{w}_s^{\mathbf{T}}\mathbf{1}-\lambda_1\mathbf{w}_s^{\mathbf{T}} sign\left(\mathbf{w}_s\right). \tag{26}$$

Note that the left hand side of (26) is $2\mathbf{w}_s^{\mathbf{T}}\hat{\Sigma}_{ss}\mathbf{w}_s+\lambda_2\left\|\mathbf{w}_s\right\|_{l_2}^2>0$, and the right hand side is $\gamma-\lambda_1\left\|\mathbf{w}_s\right\|_{l_1}$ since $\mathbf{w}_s^{\mathbf{T}}\mathbf{1}=1$ and $\mathbf{w}_s^{\mathbf{T}} sign\left(\mathbf{w}_s\right)=\left\|\mathbf{w}_s\right\|_{l_1}$. Also $\left\|\mathbf{w}_s\right\|_{l_1}\geq\mathbf{w}_s^{\mathbf{T}}\mathbf{1}=1$, and therefore $\gamma-\lambda_1>0$.

66. The statements in lines 34-end in P.19 of the old version of the paper are problematic, we have re-written as

...Finally, given λ and α, the Lagrangian multiplier $\widehat{\gamma}$ simply takes on the value to ensure the full investment constraint satisfied. From the KKT conditions, it can be shown that $\widehat{\gamma}$ is increasing with λ. As can be seen in the figures, $\widehat{\gamma}$ indeed increases monotonically with λ for each case....

S.4. Some Explanations on the Weighted Norm MVP Optimization

S.4.1. Individual's Financing Constraint

When $\alpha=1$, the norm constraint in (2) becomes $\|\mathbf{w}\|_{l_1}\leq c$, which is called a gross exposure constraint in Fan, Zhang, and Yu (2012). It can be viewed as the investor wants to minimize the portfolio variance, but still trying to limit the investment positions exposed to the risky assets. The constant c is the maximum allowable amounts of investments on the risky assets, which reflects the investor's concern on parameter uncertainty due to statistical estimation errors.

Now consider a more general version of the l_1 norm constraint:

$$\sum_{i=1}^{N}\nu_i\left|w_i\right|\leq c,$$

where ν_i is a nonnegative constant. In Brodie et al. (2009), ν_i is viewed as a measure of transaction cost, such as bid-ask spread of asset i. We can also interpret ν_i as the requirement of margin on asset i (Garleanu and Pedersen, 2011). As for the case of (2) with $\alpha=1$, it is equivalent to treating all of such transaction costs being equal to one.

S.4.2. Decision Based on Marginal Increment of the Portfolio Variance

Now let $\lambda\alpha = \lambda_1$ and $\lambda(1-\alpha) = \lambda_2$. The weighted norm mvp portfolio optimization can be viewed as a penalized l_1 norm mvp portfolio optimization when the investor uses $\hat{\Sigma}' = \hat{\Sigma} + \lambda_2 \mathbf{I}_{NN}$ as the covariance matrix. Now suppose that due to fear of estimation errors, the investor believes $\hat{\Sigma}'$ is the valid in-sample estimate for the covariance matrix. Also to simplify the analysis, suppose the mvp optimization is only subject to the full investment constraint. At the stationary point, the marginal change of the (valid) in-sample portfolio variance due to a change of w_i is given by

$$\frac{\partial \mathbf{w}^{\mathbf{T}}\hat{\Sigma}'\mathbf{w}}{\partial w_i} = 2w_i\left(\hat{\sigma}_i^2 + \lambda_2\right) + 2\sum_{j\neq i}^{N} w_j\hat{\sigma}_{ij} = \gamma - \lambda_1 \tag{27}$$

if $w_i > 0$, and

$$\frac{\partial \mathbf{w}^{\mathbf{T}}\widehat{\Sigma}'\mathbf{w}}{\partial w_i} = 2w_i\left(\hat{\sigma}_i^2 + \lambda_2\right) + 2\sum_{j\neq i}^{N} w_j\hat{\sigma}_{ij} = \gamma + \lambda_1 \tag{28}$$

if $w_i < 0$. If $\lambda_1 = 0$, the marginal change is the Lagrangian multiplier γ, which is the shadow price to measure how the portfolio variance changes when the investor's wealth changes. The marginal change of $\mathbf{w}^{\mathbf{T}}\hat{\Sigma}'\mathbf{w}$ due to an increase of a positive weight is $\gamma - \lambda_1$, which is smaller than that due to an increase of a negative weight $(\gamma + \lambda_1)$. It can be shown that $\gamma > \lambda_1$ always holds at the stationary point, and hence the marginal changes $\gamma - \lambda_1$ and $\gamma + \lambda_1$ are always positive. For $w_i = 0$, from (6),

$$\gamma - \lambda_1 \leq \frac{\partial \mathbf{w}^{\mathbf{T}}\widehat{\Sigma}'\mathbf{w}}{\partial w_i} \leq \gamma + \lambda_1. \tag{29}$$

If λ_1 becomes large, it is more likely that $\partial \mathbf{w}^{\mathbf{T}}\widehat{\Sigma}'\mathbf{w}/\partial w_i$ will fall into the interval $[\gamma - \lambda_1,\ \gamma + \lambda_1]$, and then more assets will be excluded in the optimal portfolio. Meanwhile, it is less likely that (28) will still hold as λ_1 increases, since $\gamma+\lambda_1$ will also increase. As mentioned, if λ_1 is beyond some upper bound $\bar{\lambda}$, the portfolio only has no-shortsales positions, and in this extreme case, only (27) will hold. Thus the optimal no-shortsale solution is equivalent to the optimal solution of the mvp with a subset of all assets.

The weighted norm portfolio optimization can be viewed as a decision process in which the investor assigns the penalty parameter to decide whether to include an asset in the portfolio or not. The decision is based on how the (in sample) portfolio variance changes due to an increase of the asset's weight. At the stationary point, if the increase of the asset's weight causes a large (small) enough increment of the portfolio variance, say $\gamma + \lambda_1$ $(\gamma - \lambda_1)$, then the asset's weight should have a negative (positive) sign. If the increase of the asset's weight only causes a mild increase of the portfolio variance, say between $[\gamma - \lambda_1,\ \gamma + \lambda_1]$, then the asset will not be included in the portfolio. We also can interpret it as the investor is concerned with both the sign and magnitude of the asset weight. If increasing an asset's weight only makes the portfolio variance change at some level between $\gamma - \lambda_1$ and $\gamma + \lambda_1$, then the investor will view such information is not enough to determine the sign of the asset weight, and hence it had better not to have the asset in the portfolio. It reflects the investor's attitude to parameter uncertainty, and the penalty parameter λ_1 controls the degree of such attitude.

When λ_1 becomes extremely large, only a small number of assets will be included in the portfolio. Now the assets with negative weights can have much larger impacts on the portfolio variance than the assets with positive weights. However, including the assets with negative weights will be risky, since a small change in their weights may cause a large change in the portfolio variance. Thus the investor will try to avoid such assets. In this situation, the assets included in the portfolio will all have positive sign, and changes of their weights will only have small impacts on the portfolio variance.

S.4.3. Relation to Bounded Rationality and Psychological Phenomenons

Gabaix (2011) showed that adding the l_1 norm penalty to an individual's optimization problem can generate rich bounded rationality and psychological effects. In his model, the individual tries to simplify her optimal decision process by considering only a few important parameters in the utility function. To achieve this, at first the individual minimizes a cost function for choosing the important parameters in the utility function. The cost function has a quadratic form plus the l_1 norm penalty on the parameters. With the l_1 norm penalty, it is easy to achieve a sparse solution for the parameter vector. Such a decision process can be explained as the individual prefer to frame her view on the real world not too complex. It is a reasonable setting in reality, since no one can freely consider a large amount of relevant information for her decisions. If some of the relevant information is not so important, the individual had better to damp it. Then the individual can consider her optimal actions based on the simplified utility function.

To get more insights of Gabaix (2011), we can further generalize the l_1 norm penalty $\|\mathbf{w}\|_{l_1}$ as

$$\|\mathbf{w} - \mathbf{w}_0\|_{l_1}.$$

In portfolio optimization, $\mathbf{w}_0$ can be viewed as the investor's default decision on the asset allocation. With such a generalized l_1 norm penalty, we can image that some elements in the optimal weight vector will naturally be the default weights. It means that the investor's decision with respective to some of the assets will not be changed.

In practice, this property is helpful on reducing portfolio turnover rates, as shown in DeMiguel et al. (2010), who set $\mathbf{w}_0$ equal to the optimal weight vector of the minimum variance portfolio.

In economics, such sticking-to-default effect is called inattention, while in psychology, it is called an endowment effect. This effect often arises in real world when the investor faces too many assets to choose. The investor may fear that simultaneously making many decisions may deteriorate overall qualities of these decisions. To avoid such a situation, the investor can keep as many initial decisions as possible, and change some of them if it is really necessary.

S.4.4. The Maximum a Posteriori Probability (MAP) Estimator

Zou and Hastie (2005) showed that regression coefficients regularized by the weighted norm constraint (or the elastic net constraint termed in Zou and Hastie (2005)) can be viewed as having a compromised prior between the Gaussian and Laplace distributions. Based on this result, we can give the optimal weights in (2) a similar statistical explanation. Let $\mathbf{R}_t$ be a $N \times 1$ vector of asset returns at time t, $t = 1, \ldots, T$. Suppose that given $\mathbf{w}$, sample mean return $\overline{\mathbf{R}}$, and portfolio variance σ^2_{por}, the investor believes that the portfolio return $\mathbf{w^T R}_t$ follows

$$\mathbf{w^T R}_t \mid \mathbf{w},\, \overline{\mathbf{R}},\, \sigma^2_{por} \overset{iid}{\sim} \mathcal{N}\left(\mathbf{w^T}\overline{\mathbf{R}}, \sigma^2_{por}\right) \tag{30}$$

for all t. Also suppose that the investor has a prior belief that the weight vector $\mathbf{w}$ follows a distribution with the density proportional to

$$\pi(\mathbf{w}) = \exp\left(-\psi \times \left(\alpha \|\mathbf{w}\|_{l_1} + (1-\alpha)\|\mathbf{w}\|^2_{l_2}\right)\right) \mathbb{I}\{A\mathbf{w} = \mathbf{u}\},$$

where $\psi > 0$, $\alpha \in [0,1]$, and $A\mathbf{w} = \mathbf{u}$ is a set of linear constraints for the portfolio weights. Conditional on σ^2_{por}, $\{\mathbf{R}_t\}_{t=1}^n$ and $\overline{\mathbf{R}}$, the density of the posterior distribution of the portfolio weights $\mathbf{w}$ is given by

$$\begin{aligned} p\left(\mathbf{w}|\sigma^2_{por}, \{\mathbf{R}_t\}_{t=1}^n, \overline{\mathbf{R}}\right) &\propto \exp\left(-\frac{T-1}{2\sigma^2_{por}}\mathbf{w^T}\widehat{\Sigma}\mathbf{w}\right) \times \pi(\mathbf{w}) \\ &= \exp\left(-\frac{T-1}{2\sigma^2_{por}}\mathbf{w^T}\widehat{\Sigma}''\mathbf{w} - \psi\alpha\|\mathbf{w}\|_{l_1}\right) \times \mathbb{I}\{A\mathbf{w} = \mathbf{u}\} \quad (31) \end{aligned}$$

where $\widehat{\Sigma}$ is the sample covariance matrix of $\mathbf{R}_t$, and

$$\widehat{\Sigma}'' = \widehat{\Sigma} + \frac{2\sigma^2_{por}\psi(1-\alpha)}{T-1}\mathbf{I}_{N\times N}.$$

Thus maximizing log of (31) with respect to $\mathbf{w}$ is equivalent to solving problem (2) with

$\Sigma = \widehat{\Sigma}$ and $\lambda = (2\sigma_{por}^2\psi)/(T-1)$, and the optimal $\mathbf{w}$ is the maximum a posteriori probability (MAP) estimator for $\mathbf{w}$. The above result is related to proposition 8 and 9 in DeMiguel et al. (2009a), in which they stated the cases when $\alpha = 1$. Equation (31) also implies the investor has a prior belief that $\mathbf{w}$ follows a distribution with the density proportional to $\exp\left(-\psi\alpha \left\|\mathbf{w}\right\|_{l_1}\right) \mathbb{I}\left\{A\mathbf{w} = \mathbf{u}\right\}$ and the regularized covariance matrix estimator $\widehat{\Sigma}''$ is used as the covariance matrix estimation.

S.4.5. MVP Optimization as a Minimum Mean Square Deviation Problem

The mvp optimization bears some similar properties as the squared loss-based linear regression estimation. Consider the following minimum mean square deviation problem:

$$\min_{\mathbf{w}} \mathbb{E}(Y - \mathbf{s}^{\mathbf{T}}\mathbf{w})^2, \quad \text{subject to } A\mathbf{w} = \mathbf{u}, \tag{32}$$

where the expectation is taken with respect to Y, and $A\mathbf{w} = \mathbf{u}$ is a set of linear constraints on $\mathbf{w}$. Suppose Y is dependent variable, $\mathbf{s}$ is a $N \times 1$ vector of covariates, and $\mathbf{w}$ is a $N \times 1$ vector of coefficients. $\mathbb{E}(Y - \mathbf{s}^{\mathbf{T}}\mathbf{w})^2$ can be interpreted as the expected squared prediction error. If we set $Y = \mathbf{R}^{\mathbf{T}}\mathbf{w}$, $\mathbf{s} = \mathbb{E}(\mathbf{R})$, where $\mathbf{R}$ is a $N \times 1$ vector of asset returns, the optimization is equivalent to seeking the minimum mean square deviation between $(\mathbf{R} - \mathbb{E}(\mathbf{R}))^{\mathbf{T}}\mathbf{w}$ and zero, subject to $A\mathbf{w} = \mathbf{u}$. Under this setting, the objective function is just the portfolio variance $\mathbb{E}((\mathbf{R} - \mathbb{E}(\mathbf{R}))^{\mathbf{T}}\mathbf{w})^2$, and the optimization (32) becomes the minimum variance portfolio optimization.

When the number of covariates becomes relatively large to the sample size, recent research on large dimensional variable selections in the linear regression shows that regularization methods can work well not only for model selections but also for improving out-of-sample predictions. Bai and Ng (2008) and De Mol, Giannone, and Reichlin (2008) showed that linear regression models penalized by the l_1 norm penalty can perform at least equally well or better than other traditional methods on predicting important macroeconomic indicators when a large number of predictors are jointly considered. As shown above, the optimization problems for obtaining the optimal mvp and the linear regression estimation have a similar form in which the goal is to find an optimal coefficient vector $\mathbf{w}$ to minimize the mean square deviation between two points. Thus for the mvp optimization, when the number of assets becomes relatively large to the sample size, it is reasonable to expect that the regularization methods can do improvements on reducing the out-of-sample portfolio variance as they do on reducing the mean squared prediction error of the linear regression.

Table S1: Average CPU time (seconds), proportion of active constituents (PAC) and l_1 distance of the optimal weight vectors from solving the weighted norm Markowitz portfolio optimization from Algorithm S1, no-shortsales (NS) and unconstrained (UN) Markowitz portfolio optimizations from `solve.QP`. The target return $\overline{\mu} = N^{-1}\sum_{i=1}^{N}\mu_i$. Here μ_i is estimated with sample mean of $1.2N$ return observations of asset i. The l_1 distance between the solved optimal weight vectors shown here is only for (1) no-shortsales Markowitz portfolio and the weighted norm Markowitz portfolio with $\lambda = \widehat{\lambda}_{mk}$, and (2) unconstrained Markowitz portfolio and the weighted norm Markowitz portfolio with $\lambda = 0$. Each simulation is run 1000 times.

		$\Sigma = \mathbf{I}_{N\times N}$				
		$N = 50$	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
`solve.QP-NS`	time	0.0126	0.0933	0.7100	10.8916	86.4270
$\lambda = \widehat{\widehat{\lambda}}_{mk}$	time	0.0021	0.0065	0.0292	0.3817	2.5906
	PAC	0.7377	0.7319	0.7318	0.7291	0.7287
	l_1 dist.	4.06e-11	4.94e-11	5.13e-11	7.71e-11	7.82e-11
`solve.QP-UN`	time	0.0022	0.0096	0.0412	0.3625	2.3497
$\lambda = 0$	time	0.0071	0.0249	0.0955	1.3038	7.3739
	PAC	1.0000	1.0000	1.0000	1.0000	1.0000
	l_1 dist.	3.50e-7	4.39e-7	5.05e-7	5.52e-7	5.74e-07
$\lambda = 0.8 \times \widehat{\widehat{\lambda}}_{mk}$	time	0.0022	0.0067	0.0293	0.3819	2.5997
	PAC	0.7411	0.7340	0.7333	0.7300	0.7296
$\lambda = 0.6 \times \widehat{\widehat{\lambda}}_{mk}$	time	0.0023	0.0072	0.0295	0.3858	2.6168
	PAC	0.7647	0.7502	0.7445	0.7370	0.7345
$\lambda = 0.4 \times \widehat{\widehat{\lambda}}_{mk}$	time	0.0024	0.0076	0.0313	0.4045	2.6998
	PAC	0.7962	0.7787	0.7705	0.7594	0.7537
$\lambda = 0.2 \times \widehat{\widehat{\lambda}}_{mk}$	time	0.0026	0.0091	0.0379	0.4812	3.0732
	PAC	0.8559	0.8439	0.8342	0.8222	0.8153
		$\Sigma =$Toeplitz $\left(0.6^{\lvert i-j\rvert}\right)$				
		$N = 50$	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
`solve.QP-NS`	time	0.0199	0.1492	1.1676	17.9990	142.7230
$\lambda = \widehat{\widehat{\lambda}}_{mk}$	time	0.0013	0.0053	0.0235	0.3329	2.3066
	PAC	0.5057	0.4957	0.4910	0.4894	0.4894
	l_1 dist.	4.00e-11	4.01e-11	4.44e-11	4.85e-11	5.05e-11
`solve.QP-UN`	time	0.0025	0.0094	0.0416	0.3624	2.3141
$\lambda = 0$	time	0.0154	0.0529	0.1799	2.6946	14.2980
	PAC	1.0000	1.0000	1.0000	1.0000	1.0000
	l_1 dist.	8.04e-7	9.87e-7	1.12e-06	1.23-e6	1.27e-06
$\lambda = 0.8 \times \widehat{\widehat{\lambda}}_{mk}$	time	0.0016	0.0053	0.0238	0.3329	2.3067
	PAC	0.5057	0.4957	0.4910	0.4894	0.4894
$\lambda = 0.6 \times \widehat{\widehat{\lambda}}_{mk}$	time	0.0017	0.0058	0.0242	0.3333	2.3134
	PAC	0.5090	0.4979	0.4927	0.4906	0.4903
$\lambda = 0.4 \times \widehat{\widehat{\lambda}}_{mk}$	time	0.0018	0.0055	0.0247	0.3442	2.3565
	PAC	0.5506	0.5288	0.5189	0.5097	0.5060
$\lambda = 0.2 \times \widehat{\widehat{\lambda}}_{mk}$	time	0.0025	0.0072	0.0302	0.4180	2.6896
	PAC	0.6644	0.6392	0.6262	0.6122	0.6040

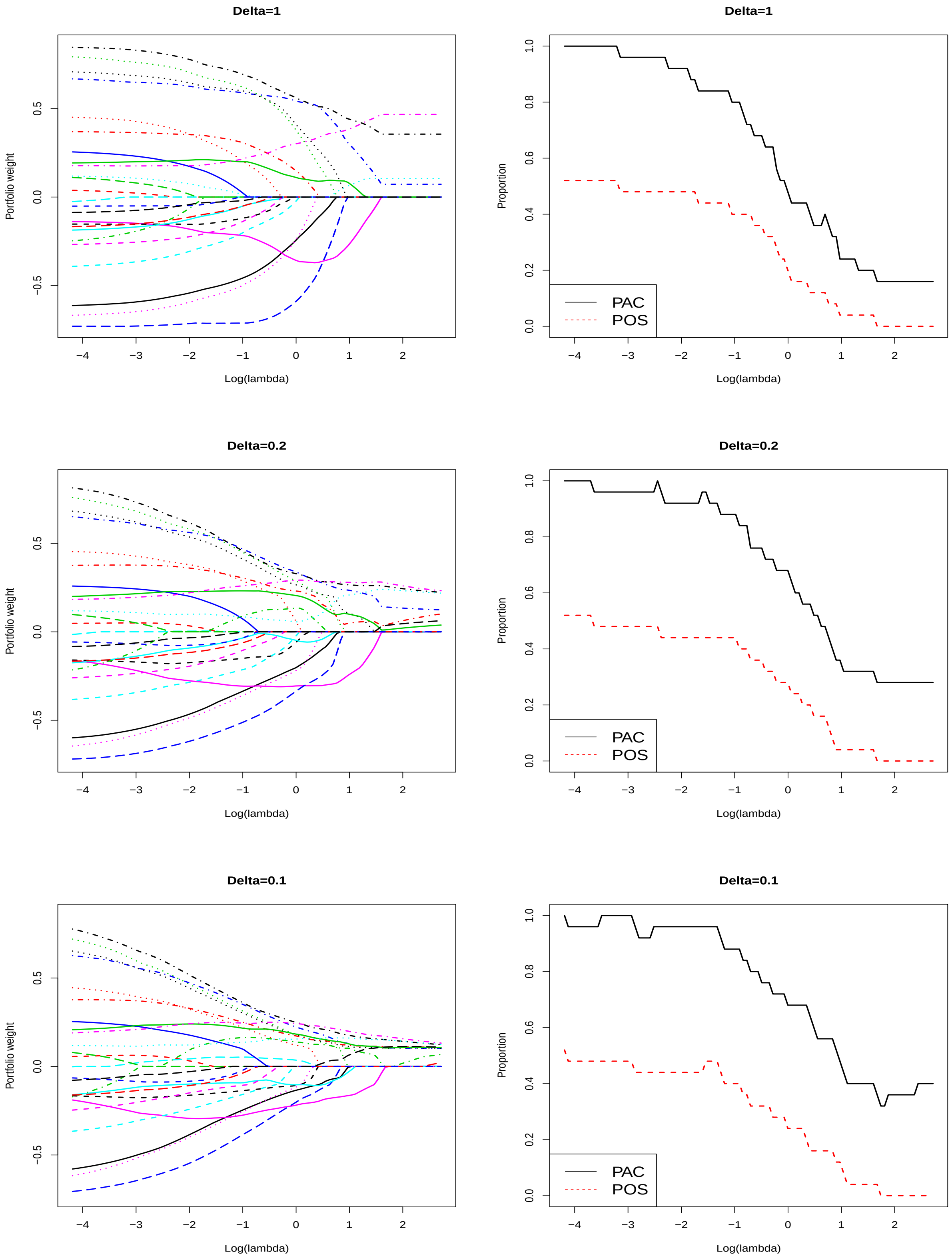

Figure S1: Profiles of portfolio weights, proportion of active constituents (PAC) and proportion of shortsale constituents (POS) from solving the mvp with the berhu penalty. The data for calculating the sample covariance matrix is the monthly return data of the FF 25 size and BM ratio portfolios from Nov-1986 to Oct-1996.

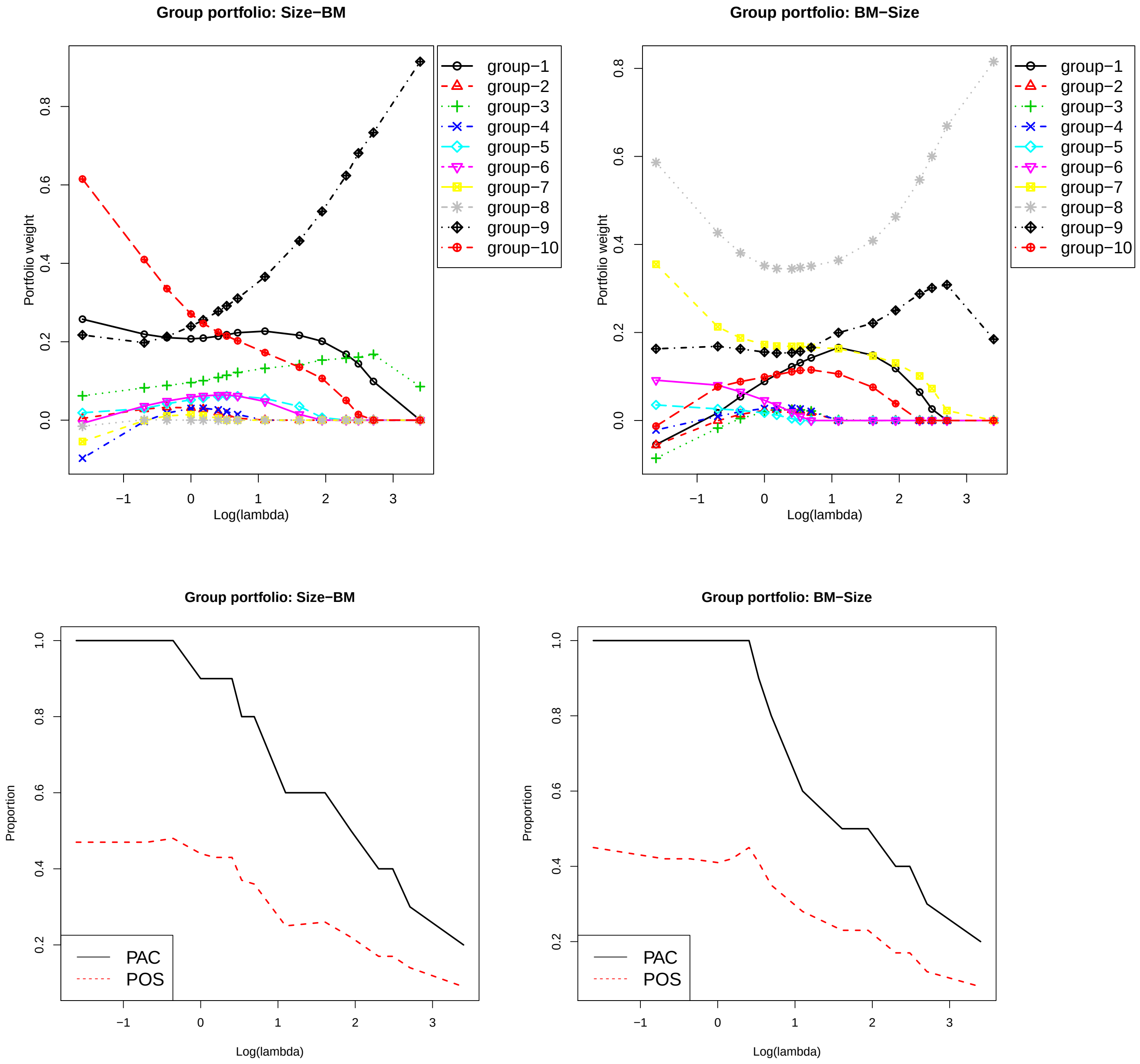

Figure S2: Profiles of portfolio weights, proportion of active constituents (PAC) and proportion of shortsale constituents (POS) from solving the mvp with the berhu penalty. The data for calculating the sample covariance matrix is the monthly return data of the FF 100 size and BM ratio portfolios from Feb-1980 to Jan-1990. We categorise these 100 portfolios via two different ways as described in Section 5.2.

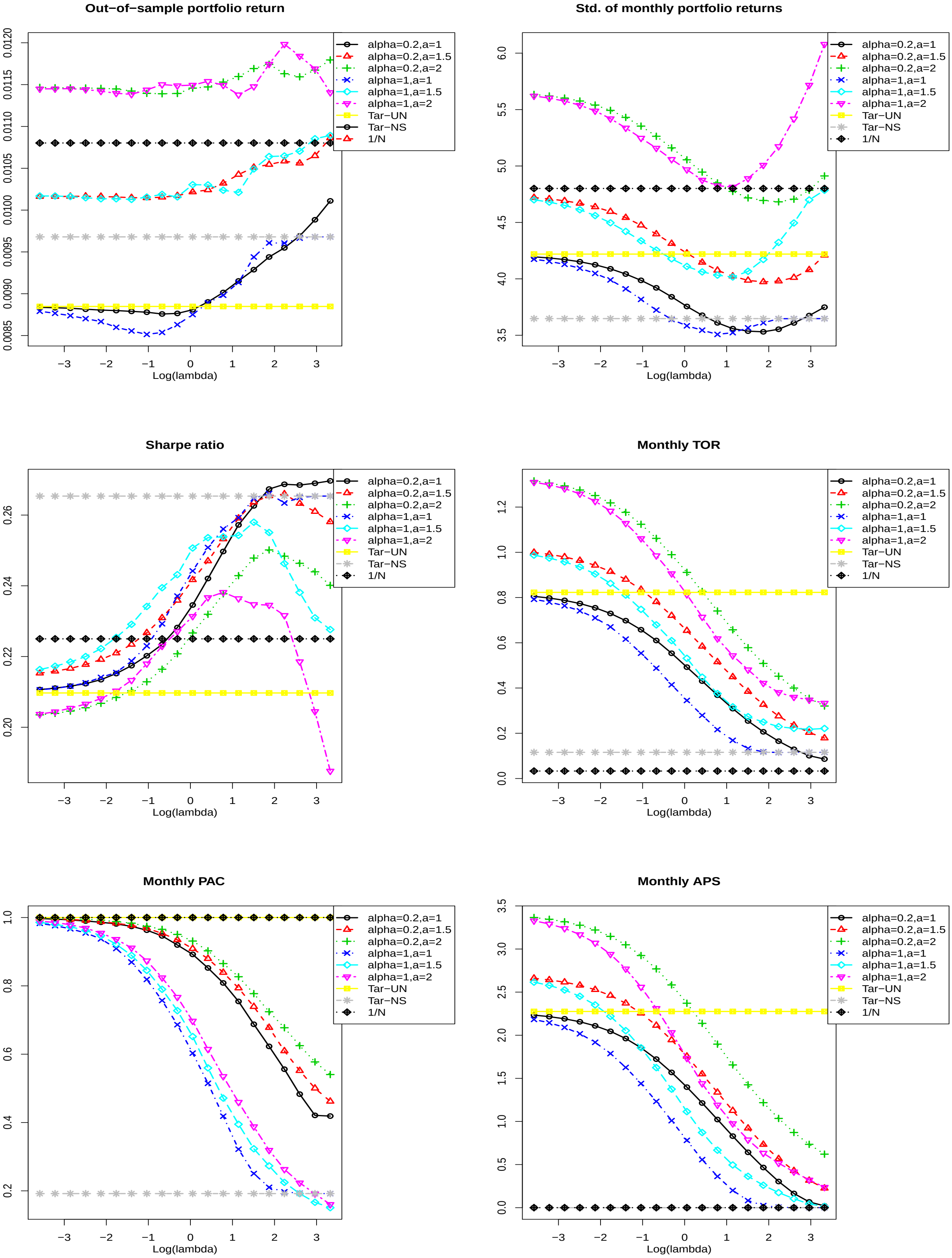

Figure S3: Standard deviation of out-of-sample portfolio returns, Sharpe ratio, average turnover rate (TOR), proportion of active constituents (PAC), absolute position of shortsales (APS) from the weighted norm Markowitz portfolio, no-shortsale Markowitz portfolio (Tar-NS), unconstrained Markowitz portfolio (Tar-UN) and 1/N portfolio, and optimal γ from the weighted norm Markowitz portfolio. The data used is the monthly return data of the FF 48 industry portfolios from July-1979 to Sep-2009.

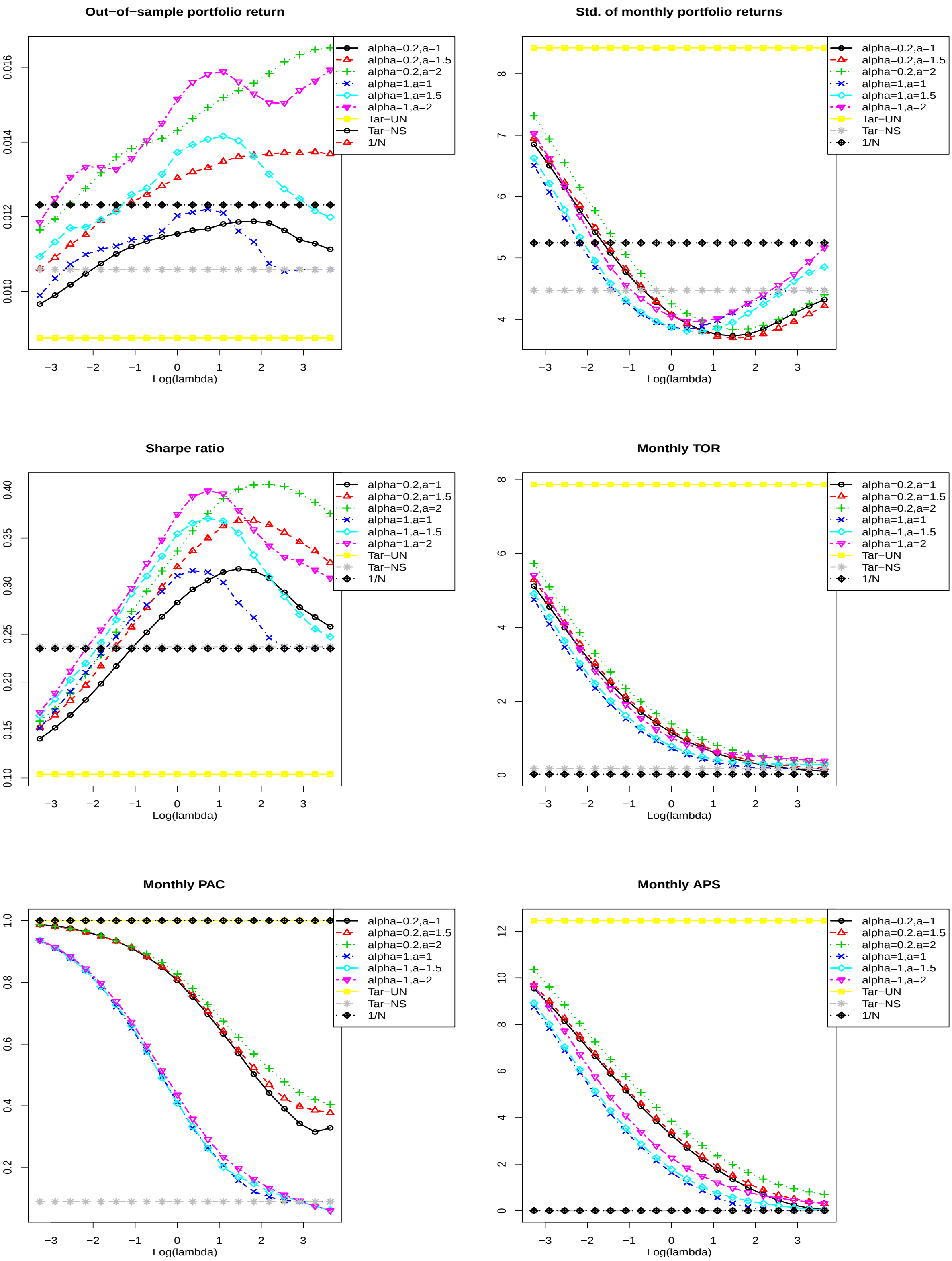

Figure S4: Standard deviation of out-of-sample portfolio returns, Sharpe ratio, average turnover rate (TOR), proportion of active constituents (PAC), absolute position of shortsales (APS) from the weighted norm Markowitz portfolio, no-shortsale Markowitz portfolio (Tar-NS), unconstrained Markowitz portfolio (Tar-UN) and 1/N portfolio, and optimal γ from the weighted norm Markowitz portfolio. The data used is the monthly return data of the FF 100 size and BM ratio portfolios from July-1973 to Sep-2009.