Weighted Elastic Net Penalized Mean-Variance Portfolio Design and Computation

Michael Ho[†]   Zheng Sun[‡]   Jack Xin[§]

October 13, 2015

Abstract

It is well known that the out-of-sample performance of Markowitz's mean-variance portfolio criterion can be negatively affected by estimation errors in the mean and covariance. In this paper we address the problem by regularizing the mean-variance objective function with a weighted elastic net penalty. We show that the use of this penalty can be motivated by a robust reformulation of the mean-variance criterion that directly accounts for parameter uncertainty. With this interpretation of the weighted elastic net penalty we derive data driven techniques for calibrating the weighting parameters based on the level of uncertainty in the parameter estimates. We test our proposed technique on US stock return data and our results show that the calibrated weighted elastic net penalized portfolio outperforms both the unpenalized portfolio and uniformly weighted elastic net penalized portfolio.

This paper also introduces a novel Adaptive Support Split-Bregman approach which leverages the sparse nature of ℓ_1 penalized portfolios to efficiently compute a solution of our proposed portfolio criterion. Numerical results show that this modification to the Split-Bregman algorithm results in significant improvements in computational speed compared with other techniques.

This article is to appear in SIAM J. Financial Mathematics.

[†]Department of Mathematics, UC Irvine, Irvine, CA 92697, USA. Email: mtho1@uci.edu.

[‡]The Paul Merage School of Business, UC Irvine, Irvine, CA 92697. Email: zsun@merage.uci.edu.

[§]Department of Mathematics, UC Irvine, Irvine, CA 92697, USA. Email: jxin@math.uci.edu.

1 Introduction

The birth of modern portfolio theory occurred in 1952 with the seminal publication of Harry Markowitz's criterion [29] for optimal single period portfolio construction that balances a portfolio's risk with return potential. A key assumption in modern portfolio theory is that given two portfolios with the same expected return an investor will always choose the portfolio with minimal risk. Markowitz proposed using the variance of portfolio's return as the measure of the portfolio risk. Thus Markowitz formulated the portfolio selection problem as minimizing portfolio return variance subject to a minimum expected value of return. Mathematically the Markowitz formulation can be written as a quadratic programming problem and the optimal portfolio can be computed using a variety of quadratic programming methods [3, 33].

One shortcoming of the Markowitz criterion for portfolio optimization is that it requires the practitioner to specify the expected return of each asset and the covariance of the returns of different assets. This presents a problem to an investor because the future mean and covariance matrix are not known. If incorrect parameter values are used then the portfolio performance will be sub-optimal [31, 8]. This additional risk due to parameter uncertainty is commonly referred to as estimation risk.

An intuitive technique that can be utilized when the mean and covariance are unknown is to estimate the mean and covariance matrix from historical return data [27] and to plug-in the estimated parameters in place of the truth. One approach to estimating the unknown parameters is to use sample averaging which is maximum likelihood (ML) optimal when the returns are i.i.d and normally distributed. This approach can be very accurate when the data is normally distributed and sufficient training data is available. For data that is not normally distributed robust estimation techniques for the covariance matrix can be considered [35, 5].

Although the sample average and plug-in approach is intuitive, there are difficulties in effectively implementing it. The primary difficulty is that there is often a limited amount of relevant historical financial return data available to estimate the mean and covariance. One reason for the lack of relevant data is that the investments' return statistics can be time-varying. Thus only a limited amount of past data is relevant in estimating the current mean and covariance. Since the volatility of assets returns can be large, the parameter estimates obtained from averaging only a small number of samples can be large. Further complicating the problem is that the covariance matrix can be ill-conditioned. This makes the portfolio weights extremely sensitive to

small parameter errors. The effect of these estimation errors is risk return performance that departs significantly from the optimal performance under known statistics [8, 1, 22].

As an alternative to sample average estimates, Bayesian estimators for both mean and covariance have been proposed [16, 23, 25]. These estimators effectively "shrink" the sample average estimates towards a more structured estimate (via a convex combination) which takes into account prior knowledge. Prior knowledge can take the form of structured data models such as a single factor model [34] or the Fama- French three-factor model [12]. Shrinking the sample average estimates towards the more structured model reduces the variability in the parameter estimates and can improve out-of-sample portfolio performance.

Another approach that has been shown to improve out-of-sample performance involves regularizing the portfolio selection criterion by adding penalties to the objective function [7, 4, 15, 39] such as portfolio norm penalties. In [7] ℓ_1 and squared ℓ_2 norm constraints are proposed for the minimum variance criterion and a cross-validation procedure is suggested to calibrate the constraints. In [39, 38] an elastic net penalty is proposed in the context of constrained minimum variance portfolio optimization. The authors also derive a method to calibrate the elastic net penalty which is designed to ensure that the variance of the resulting portfolio will not exceed the unpenalized portfolio variance (asymptotically). In [15] the authors study convex penalties such as a weighted LASSO approach [40] and a non-convex SCAD penalty [13] with application to minimum-variance portfolios. For the weighted LASSO approach the authors propose a calibration scheme where the weights are selected according to the variability in the volatility of each asset.

The above norm constrained and penalty approaches for portfolio optimization primarily focused on the minimum variance approach. Consequently the calibration and justification of the norm penalties above were derived by considering only the portfolio variance. The mean return is ignored in the calibration of the penalty. In this paper we propose a method which can be applied to the mean-variance criterion where both portfolio mean and variance are considered. In this setting we propose regularizing the objective function with a weighted elastic net penalty. A weighted elastic net penalty is a linear combination of a portfolio's weighted ℓ_1 norm and the square of a portfolio's weighted ℓ_2 norm. We show that the use of the weighted elastic net penalty can be justified by reformulating the mean-variance criterion as a robust optimization problem [17, 36] where the mean and volatilities of the asset returns belong to a known uncertainty

set. With this robust optimization interpretation, data driven techniques for calibrating the weight parameters in the weighted elastic net penalty are derived.

Our proposed penalized criterion which is equivalent to a special case of a robust optimization problem has two advantages over the general robust portfolio optimization problem. First our method can be solved using fast and well established algorithms for ℓ_1 penalized optimization problems such as the Split-Bregman algorithm [18] and the FISTA algorithm [2]. In the more general case, solving the robust portfolio optimization problem requires using semi-definite programming techniques [19] which are intractable for large portfolios. Finally, our formulation of the problem results in sparse portfolios which can contribute to reducing portfolio turnover and transaction costs. The general robust optimization problem does not necessarily result in a sparse portfolio.

This paper also addresses computational aspects of computing weighted elastic net penalized portfolios. In particular, we propose a novel Adaptive Support Split-Bregman approach to computing weighted elastic-net penalized portfolios. This new algorithm exploits the sparse nature of elastic net penalized solutions to minimize computational requirements. We show that this results in significant improvements in convergence speed versus other solvers.

The remainder of this paper's body is organized as follows: Section 2 introduces the weighted elastic net penalty and provides a justification for its use. In Section 3 we discuss the Adaptive Support Split-Bregman approach for computing the optimal portfolio. Section 4 presents experimental results using US equity data that demonstrate the benefit of our proposed approach. Finally in Section 5 we state our conclusions and a path forward for future work. The appendix contains proofs of some technical results presented in Section 3.

2 Portfolio Selection Criteria

In this section we first review the mean-variance portfolio selection criterion. We then present the weighted elastic-net penalized portfolio selection criterion and motivate its use.

2.1 Mean-Variance Portfolio selection criterion

Suppose that there exists a set of N risky assets and let $\{s_n(k)\}_{n=1}^N$ be the prices of each asset at time k. Then the excess return of the nth asset for

time period k is defined as

$$r_n(k) = \frac{s_n(k+1) - s_n(k)}{s_n(k)} - r^{(F)}(k) \tag{1}$$

where $r^{(F)}(k)$ is the return of a risk-free asset at time k. We model $\{r_n\}_{n=1}^N$ as random variables with finite mean and covariance. A portfolio is defined to be a set of weights $\{w_n\}_{n=1}^N \subset \mathbb{R}$. If $w_i > 0$ a long position has been taken in the ith asset whereas $w_i < 0$ indicates a short position.

The mean-variance criterion proposed by Markowitz [29] addresses single period portfolio selection. A portfolio of risky assets, w is mean-variance optimal if it is a solution to the following optimization problem

$$\min_w \ \varphi w^T \Gamma w - \mu^T w \tag{2}$$

where Γ and μ are the covariance and mean of r for the time period of interest and where $\varphi > 0$ is a risk aversion coefficient (since φ will only affect the portfolio weights up to a positive scalar multiple we shall set $\varphi = 1$) . Assuming that Γ is symmetric and positive definite we have that (2) is a convex quadratic program whose solution, w^*, satisfies

$$\Gamma w^* = \mu \tag{3}$$

Estimation of parameters is necessary to implement the mean-variance criteria. It has been recognized that estimation of mean return is more difficult than covariance [30] and thus a minimum variance criterion is often advocated for in recent literature [20, 7, 15]. In the minimum variance criterion the mean of asset returns are ignored and the following criterion is used for portfolio selection

$$\min_w \ w^T \Gamma w$$
$$\text{s.t.} \ \sum_{i=1}^N w_i = 1. \tag{4}$$

Despite ignoring all information on the mean return, the minimum variance criterion often outperforms the mean-variance criterion when judged by out-of-sample Sharpe ratio [7, 20].

2.2 Norm Penalized Portfolio optimization

As was stated in the introduction mean-variance portfolio optimization is sensitive to parameter estimation error. To address these concerns a number

of norm penalized criterions have been proposed, primarily in the context of minimum variance optimization. Commonly used convex norm penalties include the ℓ_1 norm , squared ℓ_2 norm and elastic net penalties [39]. The ℓ_1 and squared ℓ_2 norm penalties are given as

$$\lambda \sum_{i=1}^{N} |w_i| \tag{5}$$

and

$$\lambda \sum_{i=1}^{N} w_i^2 \tag{6}$$

respectively where $\lambda > 0$ is a weighting factor. The elastic net penalty is a weighted sum of the ℓ_1 and squared ℓ_2 norm penalties

$$\lambda_1 \sum_{i=1}^{N} |w_i| + \lambda_2 \sum_{i=1}^{N} w_i^2 \tag{7}$$

where $\lambda_1, \lambda_2 > 0$. Another convex penalty is the adaptive LASSO penalty [40] which was applied to minimum variance optimization in [15]. The adaptive LASSO penalty is a weighted ℓ_1 norm given by

$$\|w\|_{\vec{\beta},\ell_1} = \sum_{k=1}^{N} \beta_k |w_k| \tag{8}$$

where $\beta_k \geq 0$. Calibration of the weighting parameters for the above penalties has primarily been studied with the goal of improving the portfolio return variance [39, 15].

Several justifications for using ℓ_1 and squared ℓ_2 norms as penalties and constraints have been given in the literature. For example in [4] it is stated that the use of an uniformly weighted ℓ_1 penalty can be motivated by the desire to obtain sparse portfolios and to regularize the mean-variance problem when the covariance is ill-conditioned. In [14] the authors show that estimation risk in the mean-variance setting due to errors in the mean return estimation is bounded above by

$$\|\mu - \hat{\mu}\|_\infty \|w\|_{\ell_1} \tag{9}$$

and use that upper bound as a rationale for promoting small $\|w\|_{\ell_1}$. In [26] it is mentioned that a benefit of using a uniformly weighted ℓ_2 norm penalty is to stabilize the inverse covariance matrix which is often ill-conditioned in financial applications.

Non-convex penalized minimum-variance portfolio criterions were studied in [15]. One such penalty examined in [15] is the Softly Clipped Absolute Deviation (SCAD) penalty [13]. The SCAD penalty is defined as follows

$$\sum_{i=1}^{N} p_\lambda(w_i) \tag{10}$$

where

$$p_\lambda(x) = \begin{cases} \lambda|x| & \text{if } |x| \le \lambda \\ -\frac{x^2 - 2a_{SCAD}\lambda|x| + \lambda^2}{2(a_{SCAD}-1)} & \text{if } \lambda < |x| \le a_{SCAD}\lambda \\ \frac{(a_{SCAD}+1)\lambda^2}{2} & \text{if } |x| > a_{SCAD}\lambda \end{cases} \tag{11}$$

and where $a_{SCAD} > 2$. This penalty is similar to the ℓ_1 penalty and was initially proposed in context of variable selection. Calibration of the parameters a_{SCAD} and λ in (11) for portfolio optimization has not been fully addressed in the literature.

2.3 Weighted Elastic Net Penalized Portfolio

The preceding norm penalties are derived and calibrated primarily from a minimum variance perspective. In this section we extend the above methods for minimum variance portfolio design to mean-variance portfolios. Here we propose augmenting the mean-variance criterion with the sum of a weighted ℓ_1 and the square of a weighted ℓ_2 penalty. The penalty terms in the new portfolio selection criterion will be referred to as a weighted elastic net which was studied in the context of variable selection in [41].

Let $\{\alpha_i\}_{i=1}^{N}$ and $\{\beta_i\}_{i=1}^{N}$ be positive real numbers. Then the weighted elastic net penalty for a portfolio w is

$$\|w\|_{\vec{\beta},\ell_1} + \|w\|_{\vec{\alpha},\ell_2}^2 \tag{12}$$

where

$$\|w\|_{\vec{\beta},\ell_1} = \sum_{k=1}^{N} \beta_k |w_k| \tag{13}$$

and

$$\|w\|_{\vec{\alpha},\ell_2}^2 = \sum_{k=1}^{N} \alpha_k |w_k|^2. \tag{14}$$

Thus the weighted elastic net penalized mean-variance criterion may be written as

$$\min_{w} \quad w^T \hat{\Gamma} w - w^T \hat{\mu} + \|w\|_{\vec{\beta},\ell_1} + \|w\|_{\vec{\alpha},\ell_2}^2 \tag{15}$$

where $\hat{\Gamma}$ and $\hat{\mu}$ are estimates of Γ and μ respectively.

2.4 Motivation

A rationale for augmenting the mean-variance criterion with a weighted elastic net penalty can be obtained by reformulating the mean-variance criterion as a robust optimization problem. As was stated in the introduction it is well-known that the out-of-sample performance of mean-variance portfolio can degrade significantly when there are errors in the estimate of mean and covariance. The risk due to estimation errors can be reduced by accounting for them in the optimization criterion.

One way to model the parameter estimation risk is to assume the true covariance and mean belong to the following uncertainty sets

$$A = \left\{ R : R_{i,j} = \hat{\Gamma}_{i,j} + e_{i,j}; |e_{i,j}| \leq \Delta_{i,j}; R \geq 0 \right\}$$

$$B = \{ v : v_i = \hat{\mu}_i + c_i; |c_i| \leq \beta_i \}$$

where the matrix Δ is symmetric and diagonally dominant with $\Delta_{i,j} \geq 0$ for all i, j. This condition on Δ ensures that a matrix, R, of the form

$$
\begin{aligned}
R_{i,j} &= \begin{cases} \hat{\Gamma}_{i,i} + \Delta_{i,i} \text{ if } i = j \\ \hat{\Gamma}_{i,j} \pm \Delta_{i,j} \text{ if } i \neq j \end{cases} \\
R_{i,j} &= R_{j,i}
\end{aligned}
$$

is positive semi-definite (i.e. $R \in A$).

Since the mean and covariance are unknown, a conservative approach to selecting a portfolio is to optimize the worse case performance over the uncertainty sets. This can be written as the following robust optimization problem [17]

$$\min_{w} \max_{R \in A, v \in B} w^T R w - v^T w. \tag{16}$$

Note that for a fixed R and v this problem is convex in w. Since the pointwise maximum of a family of convex functions remains convex we have that

$$\max_{R \in A, v \in B} w^T R w - v^T w \tag{17}$$

is convex in w. Performing the inner maximization with respect to μ reduces the problem to

$$\min_{w} \max_{R \in A} w^T R w + \sum_{i=1}^{N} \left(-\hat{\mu}_i + \beta_i \text{sgn}(w_i) \right) w_i \tag{18}$$

where

$$\text{sgn}(w_i) = \begin{cases} \frac{w_i}{|w_i|} & \text{if } w_i \neq 0 \\ 0 & \text{else.} \end{cases}$$

This can be re-written as

$$\min_w \max_{R \in A} \text{tr}(Rww^T) - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1},$$

and the inner maximization with respect to R can be solved in closed form. Performing this final maximization gives us the following convex optimization problem

$$\min_w w^T \hat{\Gamma} w - w^T \hat{\mu} + |w|^T \Delta |w| + \|w\|_{\vec{\beta}, \ell_1} \tag{19}$$

where the $N \times 1$ vector $|w|$ is defined as

$$|w|_i = |w_i|. \tag{20}$$

Thus we see that problem (16) is equivalent to augmenting the mean-variance criterion with a weighted pairwise elastic net penalty [28].

When Δ equals the diagonal matrix D_α where

$$D_\alpha = \begin{pmatrix} \alpha_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \alpha_N \end{pmatrix} \tag{21}$$

the criterion simplifies to the weighted elastic net penalized problem defined in problem (15)

$$\min_w w^T \hat{\Gamma} w - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1} + \|w\|_{\vec{\alpha}, \ell_2}^2 \tag{22}$$

where $\alpha_i = \Delta_{i,i}$. This observation is summarized in the following theorem:

Theorem 1 *The weighted elastic net penalized problem in* (15) *is equivalent to the robust optimization problem in* (16), *when* $\Delta = D_\alpha$.

2.5   Data Driven Calibration of Weighting Parameters

We now address the problem of selecting the weighting parameters α and β. Recall that Theorem 1 states that problems (15) and (16) are equivalent. This implies that α and β represent the level of uncertainty in the mean and

variance of each asset. Thus we propose setting α and β to be commensurate with the amount of error in our parameter estimates.

Since the amount of error in the parameter estimates are unknown, we need to estimate them. One approach to estimate the amount of error is the bootstrap method [11]. Bootstrapping is a non-parametric approach that has been applied to portfolio optimization [32] and calibration of robust portfolio optimization problems [36]. One advantage of bootstrapping is that it does not require specification of a distribution of the return data.

The first step of bootstrapping is to measure T_{train} time samples of past training data to estimate μ_i and $\Gamma_{i,i}$, using estimators f_{μ_i} and $f_{\Gamma_{i,i}}$ respectively. Common choices for f_{μ_i} and $f_{\Gamma_{i,i}}$ are sample averages or shrinkage estimators. Once the parameter estimates are obtained, the training data is resampled with replacement and additional estimates of μ_i and $\Gamma_{i,i}$ are formed using the resampled data. The resampling can be described by independent uniformly distributed integer valued random variables, $v_{k,m}$, taking values between 1 and T_{train}. Here $k \in \{1, \ldots, K\}$ and $m \in \{1, \ldots, T_{train}\}$. Under the condition that the estimators f_{μ_i} and $f_{\Gamma_{i,i}}$ are invariant to the ordering of the training data, the bootstrap estimates of the estimation errors may be defined as

$$\mu_{i,err}(k) = \left| f_{\mu_i}(r_i(v_{k,1}), \ldots, r_i(v_{k,T_{train}})) - \hat{\mu}_i \right|$$

and

$$\Gamma_{i,err}(k) = \left| f_{\Gamma_{i,i}}(r_i(v_{k,1}), \ldots, r_i(v_{k,T_{train}})) - \hat{\Gamma}_{i,i} \right|$$

respectively. Here $r_i(t)$ is the return of the ith asset in the tth training sample. The percentiles of the empirical distributions of $\{\Gamma_{i,err}(k) | k = 1 \ldots K\}$ and $\{\mu_{i,err}(k) | k = 1 \ldots K\}$ can then be referenced to derive α_i and β_i. For example, suppose $0 \le p_1, p_2 \le 1$. Then the values for α_i and β_i can be defined as

$$\alpha_i = \min \left\{ x : \left| \{n : \Gamma_{i,err}(n) \le x\} \right| \le p_1 K \right\} \tag{23}$$

and

$$\beta_i = \min \left\{ x : \left| \{n : \mu_{i,err}(n) \le x\} \right| \le p_2 K \right\} \tag{24}$$

where K is the number of bootstrap estimates.

An economic interpretation of the percentile parameters p_1 and p_2 is that of model estimation risk aversion factors. Here p_1 represents the aversion to squared volatility estimation risk and p_2 is the aversion to mean estimation risk. A percentile value of 0 corresponds to no aversion to estimation risk whereas a value of 1 corresponds to a high aversion to estimation risk. Note that a higher aversion to estimation risk will increase the weights in the elastic net.

3 Numerical methods

In this section we review some numerical algorithms for determining solutions of (15). First we review an application of the Split-Bregman algorithm [18] for solving (15). Then we propose a novel Adaptive Support Split-Bregman approach which solves (15) faster than the Split-Bregman algorithm by exploiting the sparse nature of the portfolio weights.

3.1 Optimality and Approximate Optimality Conditions

In this section we derive approximate optimality conditions for (15). These conditions are then used to a design a numerical algorithm for determining the solution of (15).

Let $\Psi(w)$ denote the objective function for the weighted elastic net portfolio problem in equation (15)

$$
\begin{aligned}
\Psi(w) &= w^T \hat{\Gamma} w - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1} + \|w\|_{\vec{\alpha}, \ell_2}^2 \\
&= w^T R w - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1}
\end{aligned}
\tag{25}
$$

where $R = \hat{\Gamma} + D_\alpha$. Since Ψ is convex, w^* minimizes Ψ if and only if

$$
0 \in \partial\Psi(w^*)
\tag{26}
$$

where $\partial\Psi(w)$ is the sub-gradient of Ψ evaluated at w [3]. Note that since R is positive definite, Ψ is strictly convex and thus there is a unique solution to (15).

In most cases we are only interested in portfolios that are approximately optimal. Thus we can relax our optimality conditions to derive a stopping criterion for any iterative solver of (15). Before introducing our relaxed conditions we define the support of a portfolio w as

$$
\text{supp}(w) = \{i : |w_i| > 0\}
$$

and define the smallest variance uncertainty as

$$
\alpha_o = \min\{\alpha_i : 0 \le i \le N\}.
\tag{27}
$$

With the above definitions we have the following theorem which establishes an approximate optimality condition.

Theorem 2 *Let w^* be the solution of* (15). *Suppose that $\tilde{w}$ satisfies*

$$
\sum_{i \in \text{supp}(\tilde{w})} \left(\frac{\partial}{\partial w_i} \left(w^T R w - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1} \right) \Big|_{w=\tilde{w}} \right)^2 \le 2\epsilon\alpha_o
\tag{28}
$$

and

$$-\beta_i \le \frac{\partial}{\partial w_i}\left(w^T R w - w^T \hat{\mu}\right)\Big|_{w=\tilde{w}} \le \beta_i \qquad (29)$$

for all $i \notin \operatorname{supp}(\tilde{w})$. *Then*

$$\Psi(\tilde{w}) \le \Psi(w^*) + \epsilon \qquad (30)$$

Proof 1 *See Appendix.*

In a numerical algorithm it may happen that none of the portfolio weights are exactly 0, although they may be extremely close to zero. Thus the above theorem may not be very practical for use as a stopping criterion. For this reason let us separate the small portfolio weights from the larger portfolio weights. To do this we define

$$\operatorname{supp}_\epsilon(w) = \{i \in \operatorname{supp}(w) : |w_i| < \epsilon\}.$$

With this definition we have the following corollary which suggests a more practical stopping rule than Theorem 2.

Theorem 3 *Let* $M \ge 2\|R\|_{\ell_2}$ *and let* $\epsilon > 0$ *be given. Choose* $\eta < \frac{\epsilon \wedge \sqrt{\epsilon \alpha_o}}{\sqrt{N} M}$. *Let* w^* *be the solution the of* (15). *Suppose that* $\tilde{w}$ *satisfies*

$$\sum_{i \in \operatorname{supp}(\tilde{w}) \smallsetminus \operatorname{supp}_\eta(\tilde{w})} \left(\frac{\partial}{\partial w_i}\left(w^T R w - w^T \hat{\mu} + \|w\|_{\tilde{\beta},\ell_1}\right)\Big|_{w=\tilde{w}}\right)^2 \le 2\epsilon \alpha_o \qquad (31)$$

and

$$-\beta_i + \epsilon \le \frac{\partial}{\partial w_i}\left(w^T R w - w^T \hat{\mu}\right)\Big|_{w=\tilde{w}} \le \beta_i - \epsilon \qquad (32)$$

for $i \in \operatorname{supp}_\eta(\tilde{w}) \cup \overline{\operatorname{supp}(\tilde{w})}$. *Then*

$$\Psi(\zeta) \le \Psi(w^*) + \frac{(\sqrt{2}+1)^2}{2}\epsilon \qquad (33)$$

where

$$\zeta_i = \begin{cases} 0 \ \textit{if } i \in \operatorname{supp}_\eta(\tilde{w}) \\ \tilde{w}_i \ \textit{else} \ . \end{cases}$$

Proof 2 *See Appendix.*

3.2 Split-Bregman Algorithm

The weighted elastic net problem can be reformulated as a quadratic program and solved using general purpose solvers. However the reformulation involves adding an additional N primal variables as well as $2N$ dual variables. Thus this approach may not be applicable to large scale problems.

An algorithm better suited to handle problems like (15) is the Split-Bregman algorithm. The Split-Bregman algorithm was introduced in [18] for problems involving ℓ_1 regularization such as (15). When using the Split-Bregman method to solve (15) we solve an equivalent problem

$$\min_{w,d} \ w^T R w - w^T \hat{\mu} + \|d\|_{\ell_1}$$

$$\text{s.t.} \ \ d = \psi(w) \tag{34}$$

where $R = \rho \hat{\Gamma} + D_\alpha$ and where $\psi(w) = (\beta_1 w_1, \dots, \beta_N w_N)$. The Split-Bregman algorithm applied to (34) is

---

Algorithm 1 Split Bregman Algorithm for solving (34)

Initialize: $k = 1$, $b^k = 0, w^k = 0, d^k = 0$
while $\|w^k - w^{k-1}\|_{\ell_2} > tol$ **do**
   $w^{k+1} = \arg\min_w w^T R w - w^T \hat{\mu} + \frac{\lambda}{2}\|d^k - \psi(w) - b^k\|_{\ell_2}^2$
   $d^{k+1} = \arg\min_d \frac{\lambda}{2}\|d - \psi(w^{k+1}) - b^k\|_{\ell_2}^2 + \|d\|_{\ell_1}$
   $b_i^{k+1} = b_i^k + \beta_i w_i^{k+1} - d_i^{k+1}$
   $k = k + 1$
end while

Both inner optimization problems in Algorithm 1 have closed form solutions. The first problem is an unconstrained strictly convex quadratic program and the second problem can be solved using the shrinkage operator

$$d_j^{k+1} = shrink(\beta_j w_j^{k+1} + b_j^k, \frac{1}{\lambda})$$

where

$$shrink(x, \gamma) = \frac{x}{|x|} \cdot \max(|x| - \gamma, 0).$$

The stopping criterion in Algorithm 1 does not ensure that the objective value is within a desired tolerance. A modification to the algorithm can be made to ensure that this occurs. One such modification uses Theorem 3 to derive a stopping criterion.

Algorithm 2 Modified Split Bregman Algorithm for solving (34)

Initialize: $k = 0$, $b^k, w^k, d^k = |w^k|, tol > 0$
while w^k does not satisfy conditions of Theorem 3 for $\epsilon = \frac{2}{(\sqrt{2}+1)^2} tol$ and $\tilde{w} = w^k$ **do**
$\quad w^{k+1} = \arg\min_w w^T R w - w^T \hat{\mu} + \frac{\lambda}{2} \|d^k - \psi(w) - b^k\|_{\ell_2}^2$
$\quad d^{k+1} = \arg\min_d \frac{\lambda}{2} \|d - \psi(w^{k+1}) + b^k\|_{\ell_2}^2 + \|d\|_{\ell_1}$
$\quad b_i^{k+1} = b_i^k + \beta_i w_i^{k+1} - d_i^{k+1}$
$\quad k = k + 1$
end while
Output ζ and d^k where ζ is defined as in Theorem 3 using $\epsilon = \frac{2}{(\sqrt{2}+1)^2} tol$ and $\tilde{w} = w^k$.

By Theorem 3 this algorithm ensures that the objective value is within *tol* of the optimal value.

3.3 Adaptive Support Split Bregman

The first sub-problem in Algorithms 1 and 2 involves solving a $N \times N$ system of equations. When the number of assets is large completing this step becomes computational expensive. This is especially true for financial data where the covariance matrix is ill-conditioned and dense. Thus Algorithms 1 and 2 may be impractical in applications where real-time results are required or computational performance is limited.

It is well known [4] that portfolio optimization problems with an ℓ_1 regularization term can result in sparse portfolios i.e. the solution of (15) is only non-zero in a small number of indices. Figure 1 illustrates this behavior by showing the portfolio weights for 1600 assets obtained using the criterion in 15. For this example less than 11% of the assets have a non-zero weight.

Sparsity of the portfolio weights can be exploited to reduce computational complexity. To see this suppose w^* solves (15) and $I = \text{supp}(w^*)$ is known a priori (before computing the solution). Then the problem (15) can be relaxed to the equivalent problem

$$\min_w \ w^T R_{|I} w - w^T \hat{\mu}_{|I} + \|w\|_{\tilde{\beta}, \ell_1}$$

where $R_{|I}$ and $\mu_{|I}$ represent the covariance and mean restricted to I. This problem is of dimension $|I|$ and requires fewer operations to compute per iteration. This suggests that an Adaptive Support Split-Bregman Algorithm
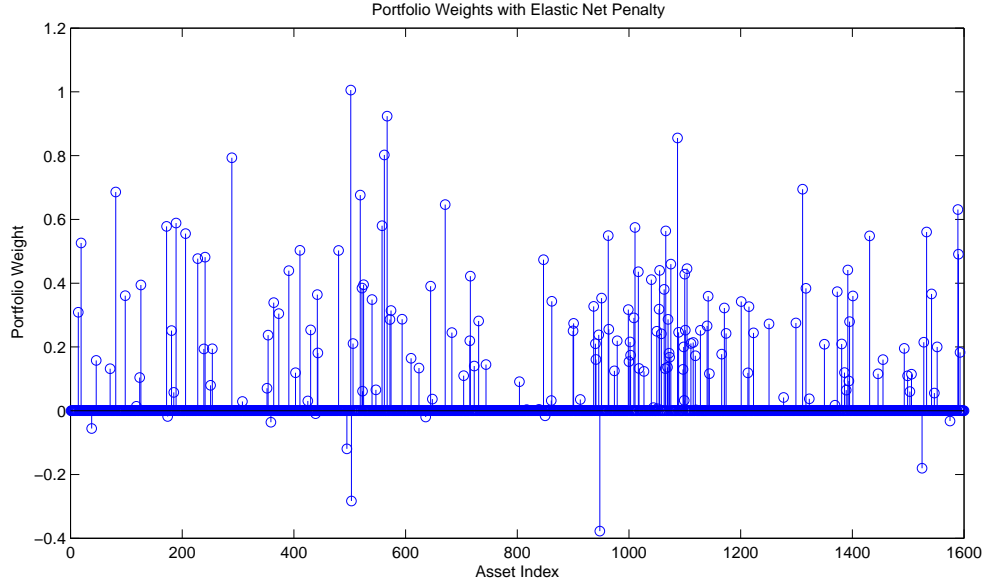
Figure 1: Elastic net penalty promotes sparsity in the portfolio weights

which attempts to solve (15) on smaller subspaces, I, where $\operatorname{supp}(w^*) \subset I$ can save computational time.

To develop an effective algorithm we first derive an optimality condition which can be used as a stopping criterion.

Lemma 4 w^* *solves* (15) *if and only if* $|(2Rw^*)_i - \hat{\mu}_i| \leq \beta_i$ *for all* $i \notin \operatorname{supp}(w^*)$ *and* $(2Rw^*)_i - \hat{\mu}_i + \beta_i \operatorname{sgn}(w_i^*) = 0$ *for all* $i \in \operatorname{supp}(w^*)$.

Proof 3 *Suppose* w^* *solves* (15) *and let* $i \in \operatorname{supp}(w^*)$. *Then since* w^* *is optimal and* $w_i^* \neq 0$ *the partial derivative of the objective function with respect to* w_i *exists and is equal to 0. Thus*

$$
\begin{aligned}
0 &= \left. \frac{\partial}{\partial w_i} \Psi(w) \right|_{w=w^*} \\
&= 2(Rw^*)_i - \hat{\mu}_i + \beta_i \operatorname{sgn}(w_i^*).
\end{aligned}
$$

Now suppose $i \notin \operatorname{supp}(w^*)$. *Now the partial derivative of the objective function does not exist. However by optimality we have*

$$
0 \in \partial \Psi(w^*)
$$

Thus

$$\lim_{h \downarrow 0} \frac{\Psi(w^* + h\delta_i) - \Psi(w^*)}{h} \geq 0$$

and

$$\lim_{h \uparrow 0} \frac{\Psi(w^* + h\delta_i) - \Psi(w^*)}{h} \leq 0$$

which imply

$$(2Rw^*)_i - \hat{\mu}_i \geq -\beta_i$$

and

$$(2Rw^*)_i - \hat{\mu}_i \leq \beta_i.$$

For the converse suppose that $|(2Rw^*)_i - \hat{\mu}_i| \leq \beta_i$ for all $i \notin \mathrm{supp}(w^*)$ and $(2Rw^*)_i - \hat{\mu}_i + \beta_i \mathrm{sgn}(w_i^*) = 0$ for all $i \in \mathrm{supp}(w^*)$. Choose $\epsilon = \min\{|w_i| : i \in \mathrm{supp}(w)\}$. Then for any w such that $\|w - w^*\|_\infty < \epsilon$

$$
\begin{aligned}
\Psi(w) - \Psi(w^*) \quad \geq \quad & \sum_{i \in \mathrm{supp}(w^*)} \left((2Rw^*)_i - \hat{\mu}_i + \beta_i \mathrm{sgn}(w_i^*)\right)(w_i - w_i^*) + \\
& + \sum_{i \notin \mathrm{supp}(w^*)} \left((2Rw^*)_i - \hat{\mu}_i\right) w_i + \beta_i |w_i| \\
\geq \quad & 0.
\end{aligned}
$$

Thus w^* is locally optimal which implies global optimality.

Lemma 4 can be used to derive a criterion for determining which indices in a portfolio, x, belong in the support. For example, suppose that $i \notin \mathrm{supp}(x)$, and $|(2Rx)_i - \hat{\mu}_i| > \beta_i$. Then the objective function in (15) can be reduced by adding i into $\mathrm{supp}(x)$. Thus x is not optimal and we should incorporate i into $\mathrm{supp}(x)$.

Next we look at how to prolongate the Split Bregman variables (w, d, b) from a lower dimensional space to a higher dimensional space. Prolongation of w and d can be achieved through simple zero filling. Prolongation of b is more delicate. The following Lemma suggests an effective prolongation.

Lemma 5 *Suppose (w^*, d^*) is the solution of (34) obtained with Algorithm 1. Then*

$$\lim_{k \to \infty} b_i^k = -(2Rw^* - \hat{\mu})_i / (\beta_i \lambda). \tag{35}$$

Proof 4 *By Algorithm 1 we have for all k*

$$2(Rw^{k+1})_i - \hat{\mu}_i - \lambda(d^k - \psi(w^{k+1}) - b^k)_i \beta_i = 0.$$

Since $\lim_{k\to\infty} w^k = w^*$ *and* $\lim_{k\to\infty} d^k = d^*$ *and* $d^* = \psi(w_i^*)$ *we have*

$$\lim_{k\to\infty} 2(Rw^{k+1})_i - \hat{\mu}_i + \lambda(b^k)_i\beta_i = 0$$

which implies that

$$\lim_{k\to\infty} (b^k)_i = \frac{\hat{\mu}_i - 2(Rw^*)_i}{\beta_i\lambda}.$$

This suggests that the prolongation of b can be defined from equation (35). For example suppose $(\tilde{w}, \tilde{d}, \tilde{b})$ solves (34) on a restricted domain $I \subset \{1, 2, \dots N\}$ and let w and d represent the prolongation of $\tilde{w}$ and $\tilde{d}$ to a set $J \supset I$ i.e.

$$w_j = \begin{cases} \tilde{w}_j & \text{if } j \in I \\ 0 & \text{if } j \in J - I \end{cases} \tag{36}$$

$$d_j = \begin{cases} \tilde{d}_j & \text{if } j \in I \\ 0 & \text{if } j \in J - I \ . \end{cases} \tag{37}$$

Then taking a cue from equation (35) the prolongation of $\tilde{b}$ may be defined as

$$b_i = (-2R_{|J}w + \hat{\mu}_{|J})_i / (\beta_i\lambda). \tag{38}$$

The Adaptive Support Split Bregman Algorithm for solving (34) is given below.

Algorithm 3 Adaptive Support Split Bregman Algorithm for solving 34

Initialize: $k = 0, w^0 = 0, d^0 = 0, b^0 = 0, \epsilon > 0, M > 0$
Define $D^0 = 2Rw^0 - \hat{\mu}$
while $|D_i^k| > \beta_i$ for any $i \notin \text{supp}(w^k)$ AND $k < N$ **do**
    Define the set $J^k = \{D_i^k : i \notin \text{supp}(w^k)\}$
    Set $K = M \vee (k + 1 - |\text{supp}(w^k)|)$
    Set $\tilde{J}^k$ equal to the largest K elements in J^k
    Set $I^k = \tilde{J}^k \cup \text{supp}(w^k)$
    Run Algorithm 2 on I^k with initialization $w_{|I^k}^k, b_{|I^k}^k, d_{I^k}^k$ and tolerance ϵ
    Set (w^{k+1}, d^{k+1}) to the prolongation of output of previous step
    Set $b_i^{k+1} = -2(Rw^{k+1} - \hat{\mu})_i / (\beta_i\lambda),$
    Set $D^{k+1} = 2Rw^{k+1} - \hat{\mu}$
    $k = k + 1$
end while

The next theorem shows that Algorithm 3 converges.

Theorem 6 *Let w^* be the optimal solution to (15) and let w' be a solution produced by Algorithm 3 for $\epsilon = tol$. Then*

$$\Psi(w') \le \Psi(w^*) + tol. \tag{39}$$

Proof 5 *By design the algorithm terminates after at most N iterations. Suppose the algorithm terminates in $k < N$ iterations. Let $I^{(k)}$ be the support in iteration k of the Adaptive Support Split-Bregman algorithm. Then by the proof of Theorem 3, w' satisfies the conditions of Theorem 2 with $\epsilon = tol$. Thus by Theorem 2 $\Psi(w') < \Psi(w^*) + tol$. Now suppose the algorithm terminates in N iterations. Since $I^{(N-1)}$ contains all asset indices it follows by the design of Algorithm 2 that $\Psi(w') < \Psi(w^*) + tol$.*

To evaluate the execution speed of Adaptive Support Split-Bregman algorithm we performed a comparison with the following fast algorithms described in the literature: Split-Bregman algorithm (Algorithm 2), FISTA [2] and Multilevel Iterated-Shrinkage [37]. To the best of our knowledge these algorithms are considered state of the art for large-scale ℓ_1-penalized quadratic programs. For the multi-level algorithm proposed in [37] we use the FISTA [2] algorithm for all relaxations and lowest level solvers. To make a fair comparison we have used the same error tolerance of 10^{-6} for each algorithm.

Tables 1 and 2 presents MATLAB run times for solving (15) for a large and small basket of US stocks. The machine running the simulation has the Windows 7 operating system and an Intel i7-3740 processor with 32.0 GB of RAM.

Table 1: Adaptive Support Split-Bregman converges quickly to a solution for sparse portfolios

Dimension	Sparsity Level	Adaptive Support Split-Bregman	Split-Bregman	FISTA	Multi-level FISTA [37]
2000	88	0.1 sec	20.6 sec	0.4 sec	0.2 sec
2000	142	0.2 sec	14.5 sec	0.8 sec	0.2 sec
2000	450	0.9 sec	14.6 sec	3.6 sec	1.5 sec
2000	853	4.8 sec	23.0 sec	8.8 sec	9.2 sec
2000	1692	10.4 sec	38.0 sec	21.4 sec	22.7 sec
3000	237	0.3 sec	48.2 sec	12.9 sec	2.7 sec
3000	805	1.3 sec	49.9 sec	55.7 sec	24.6 sec
4000	234	0.5 sec	107.6 sec	24.6 sec	2.2 sec

In Table 1 we see that the Adaptive Support Split-Bregman Algorithm converges much faster than both Split-Bregman, FISTA and Multi-Level

FISTA for sparse portfolios taken from a large set of assets. On the other hand Tables 1 and 2 show that the advantage of the Adaptive Support Split Bregman algorithm decreases when the cardinality of the asset set is small or when the support of the portfolio is large.

Table 2: Benefit of Adaptive Support Split-Bregman decreases when dimensionality is small

Dimension	Sparsity Level	Adaptive Support Split-Bregman	Split-Bregman	FISTA	Multi-level FISTA [37]
500	53	0.03 sec	0.8 sec	0.02 sec	0.02 sec
500	150	0.09 sec	0.6 sec	0.04 sec	0.03 sec
500	261	0.2 sec	0.5 sec	0.2 sec	0.2 sec

4   Experimental Results

In this section we quantify the performance benefit of using a weighted elastic net penalty by testing our criterion in (15) on daily return data from 630 U.S. stocks collected between January 1, 2001 and July 1, 2014 with market capitalization greater than 4 billion US dollars. The results are then compared with other portfolio selection criteria described in Section 2 and the naive equal-weighted portfolio.

In our experiments we compute new portfolios every 63 trading days using daily returns from the prior 252 trading days as training data for parameter estimation and calibration of the elastic net weights. Our criteria for evaluating the portfolio performance is the out-of-sample Sharpe ratio of the daily portfolio returns. Sharpe ratio is defined as the portfolio's excess return divided by its standard deviation. The formula used for computing the Sharpe ratio is given below

$$SR = \frac{\frac{1}{\tau} \sum_{i=1}^{\tau} w(t_i)^T r(t_i)}{\sqrt{\frac{1}{\tau} \sum_{i=1}^{\tau} \left(w(t_i)^T r(t_i) - \frac{1}{\tau} \left(\sum_{j=1}^{\tau} w(t_j)^T r(t_j) \right) \right)^2}} \tag{40}$$

where τ is the total number of trading days in our 13.5 year data set. Here $w(t_i)$ is the portfolio on day t_i, which is computed from the previous set of training data and remains fixed over intervals of 63 trading days.

4.1   Parameter Estimation and Calibration

Due to the large number of assets and small amount of training data, estimation of the covariance and mean in our experiments is performed using

shrinkage techniques [10]. We estimate the covariance matrix using the technique described in [24]. In that paper the following shrinkage estimator for Γ is proposed

$$\hat{\Gamma} = \rho_1 \Gamma_S + \rho_2 I \tag{41}$$

where Γ_S is the sample average covariance obtained from the training data and where ρ_1, ρ_2 are > 0. In our experiments we use the optimal values of $\rho_1 > 0$ and $\rho_2 > 0$ which are derived in [24]. Note that this choice of shrinkage target guarantees that $\hat{\Gamma}$ will be positive definite.

Since the weighted elastic net penalty consists of a squared weighted ℓ_2 norm, the shrinkage in (41) may appear to be redundant when applied with the weighted elastic net regularization in Section 2.3. However, this is not the case since the weights on the weighted elastic net and the shrinkage parameters in (41) are adaptively selected according to different criteria. Thus the covariance shrinkage target becomes a combination of the bootstrap derived target and the target derived according to [24]. One benefit of this approach is that there will always be some level of ℓ_2 regularization regardless of what the bootstrap criterion derives.

For estimation of the mean we employ a James-Stein estimator [9, 21] which was proposed for portfolio optimization in [23]. When applying the James-Stein approach we compute the estimate of μ using the equation

$$\hat{\mu} = (1 - \rho)\mu_S + \rho \eta \vec{1}. \tag{42}$$

Here μ_S is the sample mean vector and η is the maximum of average of the sample means and the daily historical return of the US stock market between 1928 and 2000 [6]

$$\eta = \left(\frac{1}{N} \sum_{i=1}^{N} \mu_{S,i} \right) \vee 0.0004. \tag{43}$$

The value of ρ is set according to [23] as

$$\rho = \min \left\{ 1, \frac{(N-2)}{T_{train}(\mu_S - \eta\vec{1})^T \hat{\Gamma}^{-1}(\mu_S - \eta\vec{1})} \right\}. \tag{44}$$

The weights for the weighted elastic net penalty are calibrated using the bootstrap technique described in Section 2.5 with identical estimation risk aversion factors for mean and squared volatility i.e. $p_1 = p_2$. Calibration of the weighted LASSO penalty is performed using the technique described in [15]. Since the weighted LASSO calibration in [15] is only defined up to a constant we perform a parametric study for various constants. Calibration

of the elastic net penalty is handled using the technique described in section 1.6.2 of [38]. The calibration method in [38] only determines the sum $\lambda_1 + \lambda_2$ in (7), the relative weighting of λ_1 and λ_2 is not addressed. Thus we perform a parametric analysis over the relative weighting between the parameters λ_1 and λ_2 in the elastic net. For SCAD there are no known calibration methods. Hence for SCAD we perform a parametric study for various λ values and a fixed a_{SCAD} parameter of 3.7 as suggested in [13].

4.2 Sharpe Ratio performance

In this section we present performance results for the following 5 mean-variance criteria: 1) unpenalized 2) weighted elastic net penalized, 3) weighted LASSO penalized [15], 4) elastic net penalized [39], and 5) SCAD penalized. As a comparison case we also tested the $1/N$ equal weighted portfolio.
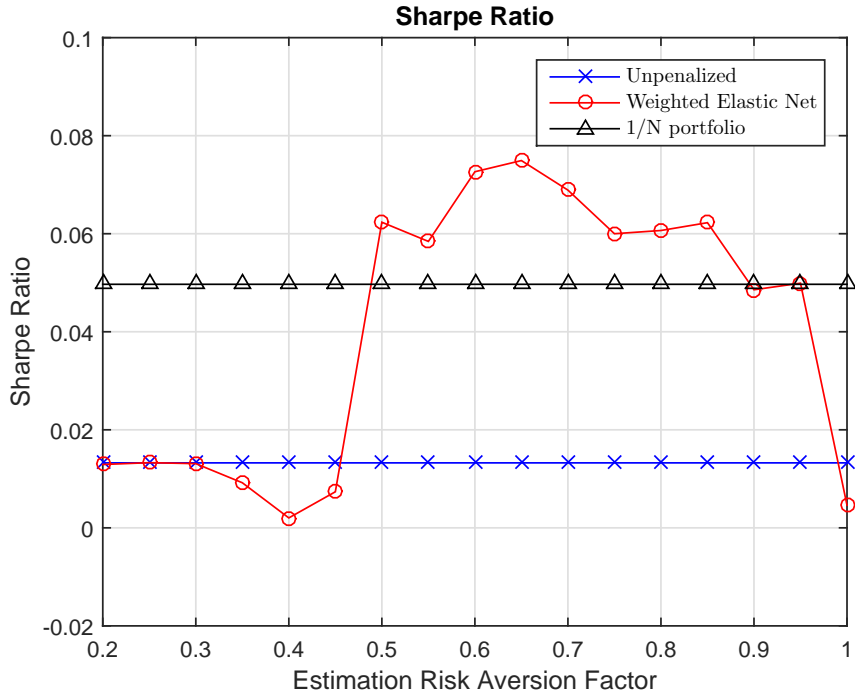

Figure 2: Weighted Elastic Net performance as a function of the estimation risk aversion factor

In Figure 2 we examine the Sharpe ratios of the weighted elastic net penalty as a function of estimation risk aversion factor, i.e. bootstrap per-

centile. As a comparison the performance of the $1/N$ and unpenalized portfolio are also shown. The figure demonstrates that the weighted elastic net penalized criterion and bootstrap calibration improves Sharpe ratio performance over the $1/N$ and unpenalized portfolio when the estimation risk aversion factor is between 0.5 and 0.95. Outside of this interval the weighted elastic net penalty did not improve performance, which suggests that a moderate amount of estimation risk aversion is optimal.
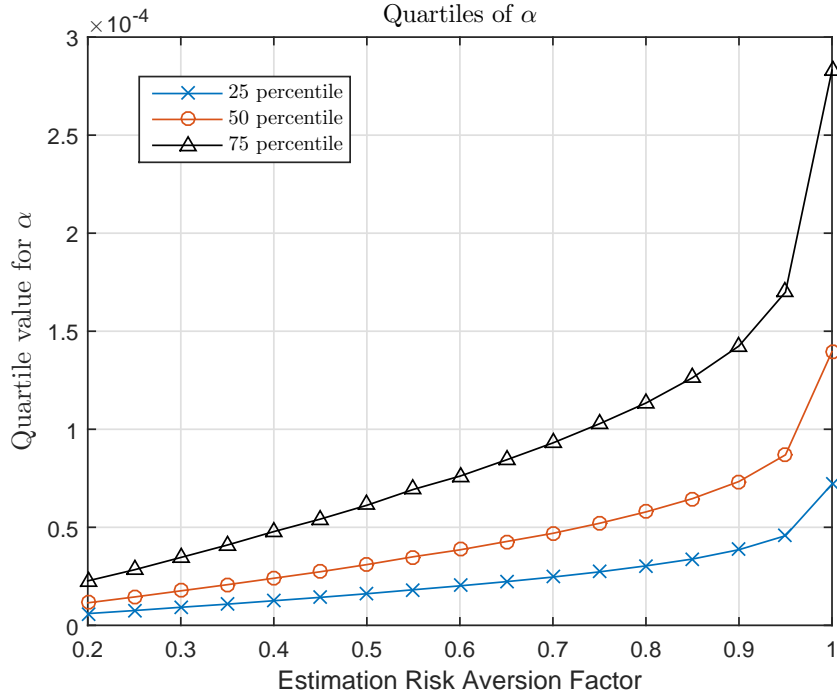

Figure 3: Quartiles for α weights as a function of the estimation risk aversion factor

In Figures 3 and 4 we present the quartiles of the α and β parameters obtained from our bootstrap technique as a function of the estimation risk aversion factors. The values increase sharply when moving from an aversion factor of 0.95 to 1.0. This may explain the dramatic loss in performance from 0.95 to 1.0 in Figure 2.

For comparison purposes the Sharpe ratio of the weighted LASSO, elastic net and SCAD penalized portfolios are shown in Figures 5, 6, and 7 as a function of their respective penalty scaling parameter. We see that both
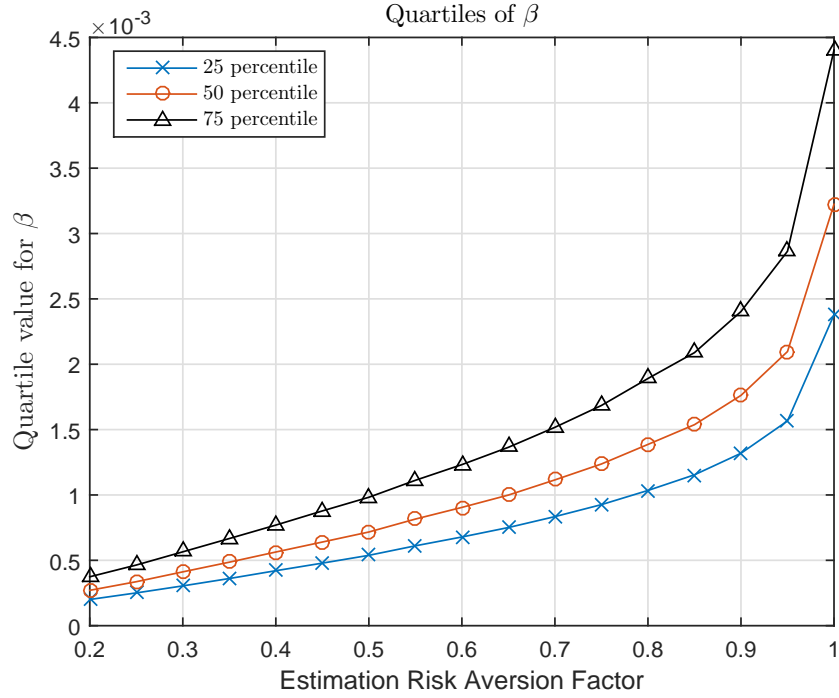
Figure 4: Quartiles for β weights as a function of the estimation risk aversion factor

weighted LASSO and the elastic net do not perform as well as the weighted elastic net penalty. This could be a consequence of their calibration being derived from a minimum variance perspective. The SCAD penalized portfolio performs comparable to the weighted elastic net penalty if the λ parameter is chosen correctly. However, it is still an open question on how to automate the selection of an optimal λ in the SCAD penalty for portfolio optimization problems.

5   Conclusions and Generalizations

In this paper the addition of a weighted elastic net penalty to mean-variance objective function has been proposed in order to improve out-of-sample portfolio performance when parameter estimates are uncertain. We have shown that this approach can be motivated by reformulating the mean-variance criterion as a robust optimization problem. With this view we develop
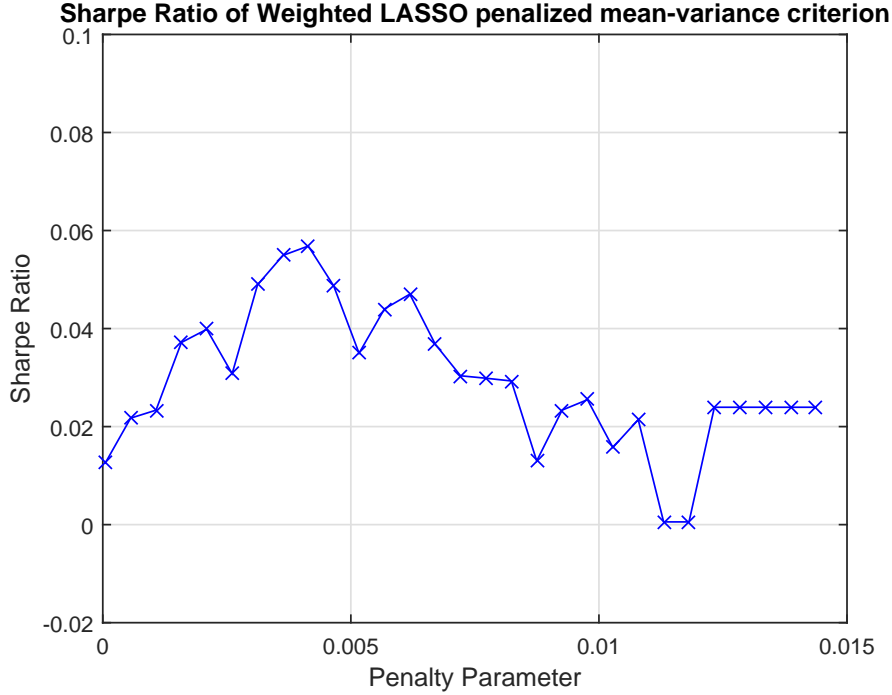
Figure 5: Weighted LASSO performance as a function of penalty normalization factor. Calibration of relative weight values performed using technique in [15].

a data-driven criterion for calibration of the elastic net weights based on bootstrapping and an investor's aversion to model estimation risk. To compute the portfolio weights efficiently we proposed a novel Adaptive Support Split-Bregman algorithm for solving our proposed optimization criterion. This technique exploits the sparsity promoting properties of the weighted elastic net penalty to reduce computational requirements.

Our experimental results demonstrate that using the weighted elastic net penalty and calibration approach can result in higher out-of-sample Sharpe ratio than the other norm penalization techniques designed for minimum variance portfolios. In addition, our MATLAB run-time results indicate that the proposed Adaptive Support Split-Bregman algorithm significantly reduces computation time compared with other algorithms such as Split-Bregman and FISTA.

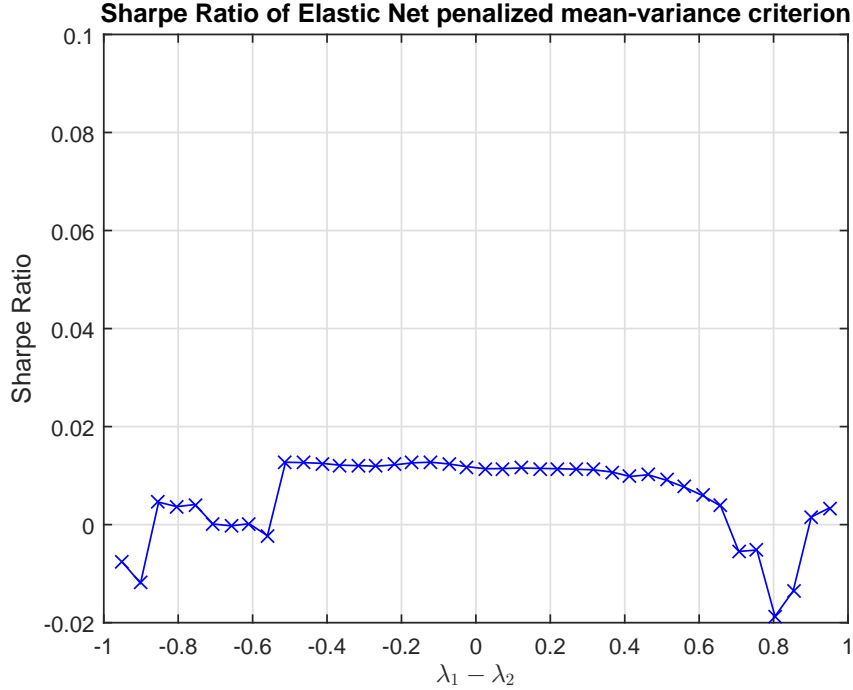An interesting question raised by this paper is whether the more gen-

Figure 6: Elastic Net performance as a function of difference of the ℓ_1 and squared ℓ_2 weights. Calibration performed using the technique in [38]

eral pairwise elastic net penalty in (19) will provide further performance enhancement than the weighted elastic net penalty. The pairwise penalty appears promising since it is derived from a more flexible model where uncertainty in the off-diagonal of Γ is allowed. However the pairwise elastic net requires specification of up to $\frac{N(N-1)}{2}$ more uncertainty parameters than the weighted elastic net. In addition numerical algorithms for computing solutions to (19) have not been extensively reported on in the literature. We plan to investigate these questions in future work.

6   Acknowledgements

We would like to thank the reviewers for their valuable comments on our paper. Their comments led us to improve the quality of our work.
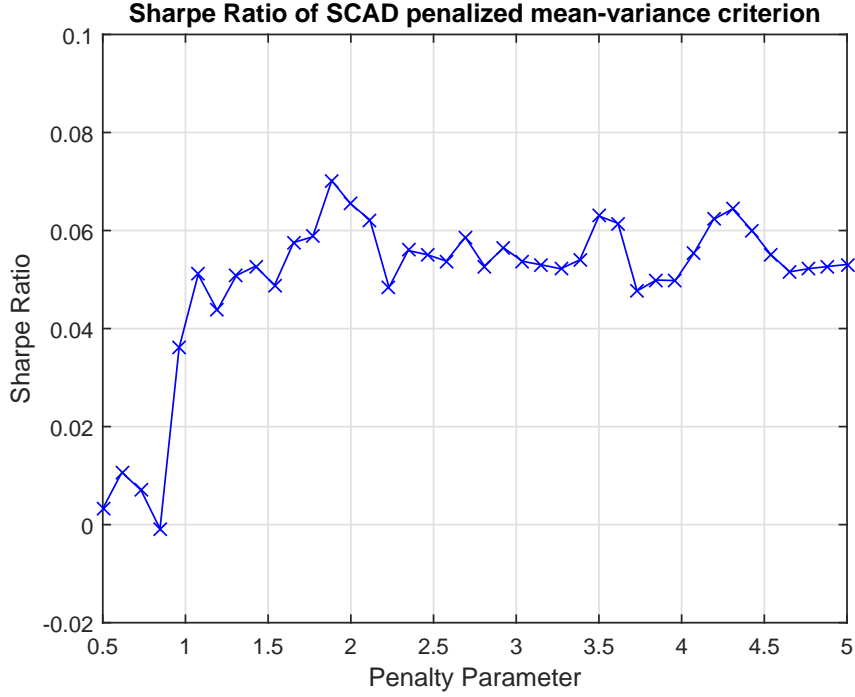
Figure 7: SCAD performance as a function of λ parameter. a_{SCAD} parameter fixed to 3.7

A   Proofs of Theorems 2 and 3

In this section we provide proofs for Theorems 2 and 3. To facilitate the proof we will first reformulate the criterion in (15) as a quadratic program.

A.1   Quadratic Program Reformulation

Problem (15) can be reformulated as a quadratic program with linear inequality constraints by introducing an auxiliary variable d,

$$
\begin{aligned}
\min_{w,d} \quad & \Phi(w,d) && (45) \\
\text{s.t.} \quad & -d_i \leq w_i \\
& -d_i \leq -w_i
\end{aligned}
$$

where $\Phi(w,d) = w^T R w - w^T \hat{\mu} + \sum_{i=1}^{N} \beta_i d_i$ and where $R = \hat{\Gamma} + D_\alpha$. The Lagrangian for this problem

$$L(w,d,\lambda) = w^T R w - w^T \hat{\mu} + \sum_{i=1}^{N} \beta_i d_i + \sum_{i=1}^{N} \lambda_i(-d_i - w_i) + \sum_{i=1}^{N} \lambda_{i+N}(-d_i + w_i) \quad (46)$$

plays an important role in our subsequent analysis in the next section.

A.2 Approximate Optimality Proofs

Here we prove Theorems 2 and 3 using the quadratic program reformulation (45). Our first task is to derive a lower bound on the Lagrangian for a fixed λ and when $d = |w|$. First note that R is symmetric positive definite whose smallest eigenvalue is $\geq \alpha_o$ where

$$\alpha_o = \min\{\alpha_i : 1 \leq i \leq N\}.$$

Thus for $d_i = |w_i|, \tilde{d}_i = |\tilde{w}_i|$ and $\lambda > 0$ we have

$$
\begin{aligned}
\Phi(w,d) \quad &\geq \quad L(w,d,\lambda) \\
&= \quad L(\tilde{w},\tilde{d},\lambda) + \nabla_w L(\tilde{w},\tilde{d},\lambda)^T(w - \tilde{w}) \\
&\qquad + \nabla_d L(\tilde{w},\tilde{d},\lambda)^T(d - \tilde{d}) + (w - \tilde{w})^T H_w(\tilde{w},\tilde{d},\lambda)(w - \tilde{w}) \\
&\geq \quad L(\tilde{w},\tilde{d},\lambda) + \nabla_w L(\tilde{w},\tilde{d},\lambda)^T(w - \tilde{w}) + \nabla_d L(\tilde{w},\tilde{d},\lambda)^T(d - \tilde{d}) \\
&\qquad + \alpha_o \|w - \tilde{w}\|_{\ell_2}^2 \\
&\geq \quad L(\tilde{w},\tilde{d},\lambda) + \nabla_w L(\tilde{w},\tilde{d},\lambda)^T(w - \tilde{w}) + \nabla_d L(\tilde{w},\tilde{d},\lambda)^T(d - \tilde{d}) \\
&\qquad + \frac{1}{2}\alpha_o \|w - \tilde{w}\|_{\ell_2}^2 + \frac{1}{2}\alpha_o \|d - \tilde{d}\|_{\ell_2}^2 \quad (47)
\end{aligned}
$$

where H_w is the Hessian of L w.r.t to the w variables.

We now present two lemmas which will be useful in deriving a stopping criterion. Our first lemma gives an upper bound for L when the gradient of L is small.

Lemma 7 *Suppose $d_i = |w_i|$ for all i and $\|\nabla_{w,d} L(\tilde{w},\tilde{d},\lambda)\|_{\ell_2} \leq \sqrt{2\epsilon\alpha_o}$. Then $L(\tilde{w},\tilde{d},\lambda) \leq \Phi(w^*,d^*) + \epsilon$ where w^* solves (15) and $d_i^* = |w_i^*|$ for all i.*

Proof 6 *By equation (47) we have*

$$
\begin{aligned}
\Phi(w^*,d^*) \geq L(w^*,d^*,\lambda) \quad &\geq \quad L(\tilde{w},\tilde{d},\lambda) + \nabla_w L(\tilde{w},\tilde{d})^T(w^* - \tilde{w}) + \nabla_d L(\tilde{w},\tilde{d})^T(d^* - \tilde{d}) \\
&\qquad + \frac{1}{2}\alpha_o \|w^* - \tilde{w}\|_{\ell_2}^2 + \frac{1}{2}\alpha_o \|d^* - \tilde{d}\|_{\ell_2}^2.
\end{aligned}
$$

The righthand side is minimized by substituting $-\frac{1}{\alpha_o}\nabla_d L(\tilde{w}, \tilde{d}, \lambda)$ in for $(d^* - \tilde{d})$ and $-\frac{1}{\alpha_o}\nabla_w L(\tilde{w}, \tilde{d}, \lambda)$ in for $(w^* - \tilde{w})$. With these substitutions we obtain

$$\Phi(w^*, d^*) \geq L(\tilde{w}, \tilde{d}, \lambda) - \frac{1}{2\alpha_o}\|\nabla_{w,d}L(\tilde{w}, \tilde{d}, \lambda)\|_{\ell_2}^2$$
$$\geq L(\tilde{w}, \tilde{d}, \lambda) - \epsilon.$$

The next lemma can be verified easily.

Lemma 8 *Suppose $|a| \leq b$. Then there exist $x_1, x_2 \geq 0$ such that*

$$x_1 + x_2 = b$$
$$-x_1 + x_2 = a.$$

A.2.1 Proof of Theorem 2

We are now ready to prove Theorem 2 which establishes a condition for approximate optimality of a portfolio under the weighted elastic net criterion (15).

Proof 7 of Theorem 2
Choose d^* and $\tilde{d}$ such that $d_i^* = |w_i^*|$ and $\tilde{d}_i = |\tilde{w}_i|$. For $i \in \mathrm{supp}(\tilde{w})$ define λ such that

$$\lambda_i = \begin{cases} 0 \text{ if } w_i > 0, i \in \mathrm{supp}(\tilde{w}) \\ \beta_i \text{ if } w_i < 0, i \in \mathrm{supp}(\tilde{w}) \end{cases}$$

and for $i \in \mathrm{supp}(\tilde{w})$, define $\lambda_{i+N} = \beta_i - \lambda_i$.

For $i \notin \mathrm{supp}(\tilde{w})$ we want to define λ_i and λ_{i+N} such that $\lambda_i \geq 0$, $\lambda_{i+N} \geq 0$,

$$\lambda_i + \lambda_{i+N} = \beta_i \tag{48}$$

and

$$-\lambda_i + \lambda_{i+N} = -\frac{\partial}{\partial w_i}\left(w^T R w - w^T \hat{\mu}\right)\Big|_{w=\tilde{w}}. \tag{49}$$

By Lemma 8, equation (29) implies that such a λ_i, λ_{i+N} exists.

Let us form the Lagrangian $L(w, d, \lambda)$ as in equation (46). Then for $i \in \mathrm{supp}(\tilde{w})$

$$\frac{\partial}{\partial w_i}L(w, d, \lambda)_{|(\tilde{w}, \tilde{d})} = \frac{\partial}{\partial w_i}\left(w^T R w - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1}\right)\Big|_{w=\tilde{w}}$$

and

$$\frac{\partial}{\partial d_i}L(w, d, \lambda)_{|(\tilde{w}, \tilde{d})} = 0.$$

For $i \notin \operatorname{supp}(\tilde{w})$ we have by equation (49)

$$\frac{\partial}{\partial w_i} L(w, d, \lambda)_{|(\tilde{w}, \tilde{d})} = 0$$

and by equation (48)

$$\frac{\partial}{\partial d_i} L(w, d, \lambda)_{|(\tilde{w}, \tilde{d})} = 0.$$

It then follows from equation (28) that

$$\|\nabla_{w,d} L(\tilde{w}, \tilde{d}, \lambda)\|_{\ell_2} \le \sqrt{2\epsilon\alpha_o}$$

and so by Lemma 7 and our choice of λ we have that

$$\begin{aligned} \Phi(\tilde{w}, \tilde{d}) &= L(\tilde{w}, \tilde{d}, \lambda) \\ &\le \Phi(w^*, d^*) + \epsilon. \end{aligned}$$

This clearly implies that

$$\Psi(\tilde{w}) \le \Psi(w^*) + \epsilon.$$

A.2.2 Proof of Theorem 3

Now we prove Theorem 3 which can be used to establish a more practical convergence criterion than Theorem 2.

Proof 8 of Theorem 3
By construction $\|\zeta - \tilde{w}\|_{\ell_\infty} \le \|\zeta - \tilde{w}\|_{\ell_2} \le \frac{\epsilon \wedge \sqrt{\epsilon\alpha_o}}{M}$. It follows that

$$\sum_{i \in \operatorname{supp}(\zeta)} \left(\frac{\partial}{\partial w_i} \left(w^T R w - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1} \right) \Big|_{w=\zeta} \right)^2 \le (\sqrt{2} + 1)^2 \alpha_o \epsilon$$

and

$$-\beta_i \le \frac{\partial}{\partial w_i} \left(w^T R w - w^T \hat{\mu} \right) \Big|_{w=\zeta} \le \beta_i$$

for all $i \notin \operatorname{supp}(\zeta)$. So by Theorem 2 we have that ζ satisfies (33).

References

[1] C. Barry, *Portfolio Analysis under Uncertain Means,Variances and Covariances*, Journal of Finance, 29 (1974), pp. 515–522.

[2] A. BECK AND M. TEBOULLE, *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, SIAM Imaging Scienses, 2 (2009), pp. 183–202.

[3] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge Univ. Press, 2004.

[4] J. BRODIE, I. DAUBECHIES, C. GUNTURK, Y. WANG, AND Y. O., *Sparse and Stable Markowitz Portfolios*, PNAS, 106 (2009), pp. 12267–12272.

[5] Y. CHEN, A. WIESEL, AND A. HERO, *Robust Shrinkage Estimation of High-dimensional Covariance Matrices*, IEEE Transactions on Signal Processing, 59 (2011), pp. 4097–4107.

[6] A. DAMODARAN, *Annual Returns on Stock, T.Bonds and T.Bills: 1928 - Current.* http://www.stern.nyu.edu/ adamodar/pc/datasets/histretSP.xls.

[7] V. DEMIGUEL, L. GARLAPPI, F. NOGALES, AND R. UPPAL, *A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms*, Management Science, (2009), pp. 798–812.

[8] V. DEMIGUEL, L. GARLAPPI, AND R. UPPAL, *Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?*, The Review of Financial Studies, 22 (2009), pp. 1915–1953.

[9] B. EFRON AND C. MORRIS, *Stein's Estimation Rule and its Competitors- An Empirical Bayes Approach*, Journal of American Statistical Association, 68 (1973), pp. 117–130.

[10] B. EFRON AND C. MORRIS, *Stein's Paradox in Statistics*, Scientific American, 237 (1977), pp. 119–127.

[11] B. EFRON AND R. TIBSHIRANI, *Bootstrap methods for standard errors,confidence intervals, and other measures of statistical accuracy*, Statistical Science, 1 (1986), pp. 54–75.

[12] E. FAMA AND K. FRENCH, *Common Risk Factors in the Returns on Stocks and Bonds*, Journal of Financial Economics, 33 (1993), pp. 3–56.

[13] J. FAN AND R. LI, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association, 96 (2001), pp. 1348–1360.

[14] J. FAN, J. ZHANG, AND K. YU, *Vast Portfolio Selection With Gross-Exposure Constraints*, Journal of the American Statistical Association, 107 (2012), pp. 592–607.

[15] B. FASTRICH, S. PATERLINI, AND P. WINKER, *Constructing optimal sparse portfolios using regularization methods*, Computational Management Science, 12 (2014), pp. 417–434.

[16] P. FROST AND J. SAVARINO, *An Empirical Bayes Approach to Efficient Portfolio Selection*, The Journal of Financial and Quantitative Analysis, 21 (1986), pp. 293–305.

[17] D. GOLDFARB AND G. IYENGAR, *Robust Portfolio Selection Problems*, Mathematics of Operations Research, 28 (2003), pp. 1–38.

[18] T. GOLDSTEIN AND S. OSHER, *The Split Bregman Method for ℓ_1 Regularized problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 323–343.

[19] B. HALLDORSSON AND R. TUTUNCU, *An Interior Point Method for a Class of Saddle Point Problems*, Journal of Optimization Theory and Applications, 116 (2003), pp. 559–590.

[20] R. JAGANNATHAN AND T. MA, *Risk reduction in large portfolios: Why imposing the wrong constraints helps*, Journal of Finance, 58 (2003), pp. 1651–1684.

[21] W. JAMES AND C. STEIN, *Estimation with quadratic loss*, in Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, 1961, pp. 361–379.

[22] J. JOBSON AND B. KORKIE, *Estimation for Markowitz Efficient Portfolios*, Journal of the American Statistical Association, 75 (1980), pp. 544–554.

[23] P. JORION, *Bayes-Stein Estimation for Portfolio Analyis*, The Journal of Financial and Quantitative Analysis, 21 (1986), pp. 279–292.

[24] O. LEDOIT AND M. WOLF, *A well-conditioned estimator for large-dimensional covariance matrices* , Journal of Multivariate Analysis, 88 (2004), pp. 365–411.

[25] ——, *Honey, I Shrunk the Sample Covariance Matrix*, The Journal of Portfolio Management, 30 (2004), pp. 110–119.

[26] J. Li, *Sparse and stable portfolio selection with parameter uncertainty*, Journal of Business and Economic Statistics, To Appear.

[27] R. Litterman and K. Winkelmann, *Estimating Covariance Matrices*, Risk Management Series, (1998). Goldman Sachs.

[28] A. Lorert, D. Eis, V. Kostina, and P. Ramadge, *Exploiting covariate similarity in sparse regression via the pairwise elastic net*, in Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, vol. 9, Chia Laguna Resort, Sardinia, Italy., 2010.

[29] H. Markowitz, *Portfolio selection*, Journal of Finance, 7 (1952), pp. 77–91.

[30] R. Merton, *On estimating the expected return on the market: An exploratory investigation*, Journal of Financial Economics, 8 (1980), pp. 323–361.

[31] R. Michaud, *The Markowitz optimization enigma: Is optimized optimal?*, Finan, Anal. J., 45 (1989), pp. 31–42.

[32] ——, *Efficient Assset Management: A Practial Guide to Stock Portfolio Optimization*, Oxford Univ. Press, New York, 1993.

[33] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, 1999.

[34] W. Sharpe, *A Simplified Model for Portfolio Analysis*, Management Science, 9 (1963), pp. 277–293.

[35] Y. Sun, P. Babu, and D. Palomar, *Regularized tylers scatter estimator: Existence, uniqueness, and algorithms*, IEEE Transactions on Signal Processing, 62 (2014), pp. 5143–5156.

[36] R. Tütüncü and M. Koenig, *Robust asset allocation*, Annals of Operations Research, 132 (2004), pp. 157–187.

[37] E. Treister and I. Yavneh, *A Multilevel Iterated-Shrinkage Approach to ℓ_1 Penalized Least-Squares Minimization*, IEEE Trans. Signal Proc., 60 (2012), pp. 6319–6329.

[38] Y. Yen, *Three essays in financial econometrics*, PhD thesis, University of London, 2012.

[39] Y. YEN AND T. YEN, *Solving norm constrained portfolio optimization via coordinate-wise descent algorithms*, Computational Statistics and Data Analysis, 76 (2014), pp. 737–759.

[40] H. ZOU, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association, 101 (2006), pp. 1418–1429.

[41] H. ZOU AND H. ZHANG, *On the Adaptive Elastic-Net with a Diverging Number of Parameters*, The Annals of Statistics, 37 (2009), pp. 1733–1751.