Weighted Elastic Net Penalized Mean-Variance Portfolio Design and Computation

Michael Ho† Zheng Sun‡ Jack Xin§

February 9, 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 $\ell_1$ 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.

1 Introduction

The birth of modern portfolio theory occurred in 1952 with the seminal publication of Harry Markowitz’s criterion \[25\] 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 \[3, 28\].

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

†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.
¶The work was partially supported by NSF grant DMS-1211179.
not known. If incorrect parameter values are used then the portfolio performance will be sub-optimal [26, 7]. 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 [23] 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 [30, 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 [7, 1, 18].

As an alternative to sample average estimates, Bayesian estimators for both mean and covariance have been proposed [14, 19, 21]. 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 models [29] or the Fama-French three-factor model [11]. 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 mean-variance criterion by adding portfolio norm constraints or penalties to the objective function [6, 4, 13, 22, 33]. The most commonly used norm constraints and penalties are the uniformly weighted $\ell_1$ and the squared $\ell_2$ norms. Financial rationales for using norm penalties are discussed in [6]. For example in [6] the use of an $\ell_1$ norm constraint is shown to limit the budget for large short sale positions. In [22] the use of norm penalties is motivated by reformulating the portfolio selection problem as a norm regularized linear regression problem. In [4] the authors use portfolio sparsity and transaction costs as rationales for using $\ell_1$ norm constraints.

In this paper 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 $\ell_1$ norm and the square of a portfolio’s weighted $\ell_2$ 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 [15, 31] 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.

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 will 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. Finally in Section 4 we present experimental results using US equity data that demonstrate the benefit of our proposed approach. 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 \( n^{th} \) asset for time period \( k \) is defined as

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

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 \( i^{th} \) asset whereas \( w_i < 0 \) indicates a short position.

The mean-variance criterion proposed by Markowitz [25] 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

\[
    \min_w w^T \Gamma w - \mu^T w
\]

where \( \Gamma \) and \( \mu \) are the covariance and mean of \( r \) for the time period of interest. We shall assume that \( \Gamma \) is symmetric and positive semi-definite. With this assumption the mean-variance optimization problem is a convex quadratic program whose solution, \( w^* \), satisfies

\[
    \Gamma w^* = \mu
\]

2.2 Weighted Elastic Net Penalized Portfolio

In this section we augment the mean-variance criterion with the sum of a weighted \( \ell_1 \) and the square of a weighted \( \ell_2 \) penalty and insert estimates of \( \mu \) and \( \Gamma \) in place of the unknown true values. The penalty terms in the new portfolio selection criterion will be referred to as a weighted elastic net. 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

\[
    ||w||_{\beta,\ell_1} + ||w||_{\alpha,\ell_2}^2
\]
where
\[ \|w\|_{\beta, \ell_1} = \sum_{k=1}^{N} \beta_k |w_k| \]  
(4)
and
\[ \|w\|_{\alpha, \ell_2}^2 = \sum_{k=1}^{N} \alpha_k |w_k|^2. \]  
(5)
Thus the weighted elastic net penalized mean-variance criterion may be written as
\[ \min_w w^T \hat{\Gamma} w - w^T \hat{\mu} + \|w\|_{\beta, \ell_1} + \|w\|_{\alpha, \ell_2}^2 \]  
(6)
where \( \hat{\Gamma} \) and \( \hat{\mu} \) are estimates of \( \Gamma \) and \( \mu \) respectively.

### 2.3 Motivation

Several justifications for using \( \ell_1 \) and squared \( \ell_2 \) 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 \( \ell_1 \) 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 [12] the authors show that estimation risk due to errors in the mean return estimation is bounded above by
\[ \|\mu - \hat{\mu}\|_{\infty} \|w\|_{\ell_1} \]  
(7)
and use that upper bound as a rationale for promoting small \( \|w\|_{\ell_1} \). In [22] it is mentioned that a benefit of using a uniformly weighted \( \ell_2 \) norm penalty is to stabilize the inverse covariance matrix which is often ill-conditioned in financial applications.

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 uncertainties in the parameter estimates in the optimization criterion.

One way to account for risk due to parameter estimation errors is to assume the true covariance and mean belong to the following uncertainty sets
\[ A = \{ R : R_{i,j} = \hat{\Gamma}_{i,j} + e_{i,j}; |e_{i,j}| \leq \Delta_{i,j}; R \succeq 0 \} \]
\[ B = \{ v : v_i = \hat{\mu}_i + c_i; |c_i| \leq \beta_i \} \]
where the matrix \( \Delta \) is symmetric and diagonally dominant with \( \Delta_{i,j} \geq 0 \) for all \( i, j \). This condition on \( \Delta \) ensures that a matrix, \( R \), of the form
\[ R_{i,j} = \begin{cases} 
\hat{\Gamma}_{i,i} + \Delta_{i,i} & \text{if } i = j \\
\hat{\Gamma}_{i,j} \pm \Delta_{i,j} & \text{if } i \neq j
\end{cases} 
\]
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 \[15\]

\[
\min_{w} \max_{R \in A, v \in B} w^T R w - v^T w.
\] (8)

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

\[
\max_{R \in A, v \in B} w^T R w - v^T w
\] (9)

is convex in \(w\). Performing the inner maximization with respect to \(\mu\) reduces the problem to

\[
\min_{w} \max_{R \in A} w^T R w + \sum_{i=1}^{N} (\hat{\mu}_i + \beta_i \text{sgn}(w_i)) w_i
\] (10)

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

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

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

\[
\min_{w} w^T \hat{\Gamma} w - w^T \hat{\mu} + \|w\|_{\beta, \ell_1} + \|w\|_{\alpha, \ell_2}
\] (11)

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

\[
|w|_i = |w_i|.
\] (12)

Thus we see that problem (5) is equivalent to augmenting the mean-variance criterion with a weighted pairwise elastic net penalty \[24\].

When \(\Delta\) equals the diagonal matrix \(D_\alpha\) where

\[
D_\alpha = \begin{pmatrix}
\alpha_1 & 0 & \ldots & 0 \\
0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \ldots & 0 & \alpha_N
\end{pmatrix}
\] (13)

the criterion simplifies to the weighted elastic net penalized problem defined in problem (6)

\[
\min_{w} w^T \hat{\Gamma} w - w^T \hat{\mu} + \|w\|_{\beta, \ell_1} + \|w\|_{\alpha, \ell_2}^2
\] (14)

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

**Theorem 1** The weighted elastic net penalized problem in (6) is equivalent to the robust optimization problem in (8), when \(\Delta = D_\alpha\).
2.4 Data Driven Calibration of Weighting Parameters

We now address the problem of selecting the weighting parameters $\alpha$ and $\beta$. Recall that Theorem 1 states that problems (6) and (8) are equivalent. This implies that $\alpha$ and $\beta$ represent the level of uncertainty in the mean and variance of each asset. Thus we propose setting $\alpha$ and $\beta$ 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 [10]. Bootstrapping is a non-parametric approach that has been applied to portfolio optimization [27] and calibration of robust portfolio optimization problems [31]. 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 $T_{\text{train}}$ time samples of past training data to estimate $\mu_i$ and $\Gamma_{i,i}$, using estimators $f_{\mu_i}$ and $f_{\Gamma_{i,i}}$ respectively. Common choices for $f_{\mu_i}$ 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 $\mu_i$ 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 $T_{\text{train}}$. Here $k \in \{1, \ldots, K\}$ and $m \in \{1, \ldots, T_{\text{train}}\}$. Under the condition that the estimators $f_{\mu_i}$ 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

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

and

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

respectively. Here $r_i(t)$ is the return of the $i$th asset in the $t$th training sample. The percentiles of the empirical distributions of $\{\Gamma_{i,\text{err}}(k)|k = 1 \ldots K\}$ and $\{\mu_{i,\text{err}}(k)|k = 1 \ldots K\}$ can then be referenced to derive $\alpha_i$ and $\beta_i$. For example, suppose an investor chooses an estimation risk protection factor of $0 \leq p \leq 1$. Then a corresponding bootstrap percentile of $p$ can be selected and the values for $\alpha_i$ and $\beta_i$ can be defined as

$$\alpha_i = \min \{x : |\{n : \Gamma_{i,\text{err}}(n) \leq x\}| \leq pK\}$$

and

$$\beta_i = \min \{x : |\{n : \mu_{i,\text{err}}(n) \leq x\}| \leq pK\}$$

where $K$ is the number of bootstrap estimates. Here $p = 0$ corresponds to an unpenalized mean-variance criterion whereas $p = 1$ corresponds to a heavily weighted elastic net penalty which assumes a worse case uncertainty in $\Gamma$ and $\mu$.

3 Numerical methods

In this section we review some numerical algorithms for determining solutions of (6). First we review an application of the Split-Bregman algorithm [16] for solving (6). Then we propose a novel Adaptive Support Split-Bregman approach which solves (6) 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 (6). These conditions are then be used to design a numerical algorithm for determining the solution of (6).

Let \( \Psi(w) \) denote the objective function for the weighted elastic net portfolio problem (6)

\[
\Psi(w) = w^T \hat{\Gamma} w - w^T \hat{\mu} + \|w\|^2_{\beta,1} + \|w\|^2_{\alpha,2}
\]

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

\[
0 \in \partial \Psi(w^*)
\]

where \( \partial \Psi(w) \) is the sub-gradient of \( \Psi \) evaluated at \( w \). Note that since \( R \) is positive definite, \( \Psi \) is strictly convex and thus there is a unique solution to (6).

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 (6). 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

\[
\alpha_o = \min\{\alpha_i : 0 \leq i \leq N\}.
\]

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

**Theorem 2** Let \( w^* \) be the solution of (6). Suppose that \( \tilde{w} \) satisfies

\[
\sum_{i \in \text{supp}(\tilde{w})} \left( \frac{\partial}{\partial w_i} (w^T R w - w^T \hat{\mu} + \|w\|_{\beta,1}) \right)_{w=\tilde{w}}^2 \leq 2\epsilon \alpha_o.
\]

and

\[
-\beta_i \leq \frac{\partial}{\partial w_i} (w^T R w - w^T \hat{\mu}) \bigg|_{w=\tilde{w}} \leq \beta_i
\]

for all \( i \notin \text{supp}(\tilde{w}) \). Then

\[
\Psi(\tilde{w}) \leq \Psi(w^*) + \epsilon
\]

**Proof** 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

\[
\text{supp}_\epsilon(w) = \{i \in \text{supp}(w) : |w_i| < \epsilon\}.
\]

With this definition we have the following corollary which suggests as a more practical stopping rule than Theorem 2
Theorem 3 Let $M \geq 2\|R\|_{\ell_2}$ and let $\epsilon > 0$ be given. Choose $\eta < \frac{\epsilon}{\sqrt{NM}}$. Let $w^*$ be the solution the of (6). Suppose that $\tilde{w}$ satisfies

$$
\sum_{i \in \text{supp}(\tilde{w}) \setminus \text{supp}_\eta(\tilde{w})} \left( \frac{\partial}{\partial w_i} \left( w^T R w - w^T \hat{\mu} \right) \bigg|_{w=\tilde{w}} \right)^2 \leq 2\epsilon \alpha.
$$

and

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

for $i \in \text{supp}_\eta(\tilde{w})$ and

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

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

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

where

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

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 (6) is the Split-Bregman algorithm. The Split-Bregman algorithm was introduced in [16] for problems involving $\ell_1$ regularization such as (6). When using the Split-Bregman method to solve (6) we solve an equivalent problem

$$
\begin{aligned}
\min_{w,d} & \quad w^T R w - w^T \hat{\mu} + \|d\|_{\ell_1} \\
\text{s.t.} & \quad d = \psi(w)
\end{aligned}
$$

where $R = \rho \hat{\Gamma} + D_\alpha$ and where $\psi(w) = (\beta_1 w_1, \ldots, \beta_N w_N)$. The Split-Bregman algorithm applied to (27) is

Algorithm 1 Split Bregman Algorithm for solving (27)

Initialize: $k = 1$, $b^k = 0$, $w^k = 0$, $d^k = 0$

while $\|w^k - w^{k-1}\|_{\ell_2} > \text{tol}$ do

$w^{k+1} = \min_w w^T R w - w^T \hat{\mu} + \frac{\epsilon}{\alpha} \|d^k - \psi(w) - b^k\|_{\ell_2}^2$

$d^{k+1} = \min_d \frac{\epsilon}{\alpha} \|d - \psi(w^{k+1}) - b^k\|_{\ell_2} + \|d\|_{\ell_1}$

$b_i^{k+1} = b_i^k + \beta_i w_i^{k+1} - d_i^{k+1}$

$k = k + 1$

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

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

where

\[ \text{shrink}(x, \gamma) = \frac{x}{|x|} \cdot \max(|x| - \gamma, 0) \]

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 (27)

**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

\[
\begin{align*}
  w^{k+1} &= \min_w w^T R w - w^T \hat{\mu} + \frac{\lambda}{2} ||d^k - \psi(w) - b^k||_{\ell_2}^2 \\
  d^{k+1} &= \min_d \frac{\lambda}{2} ||d - \psi(w^{k+1}) + b^k||_{\ell_2}^2 + ||d||_{\ell_1} \\
  b_i^{k+1} &= b_i^k + \beta_i w_i^{k+1} - d_i^{k+1} \\
  k &= k + 1
\end{align*}
\]

**end while**

**Output** \( \zeta \) and \( d^k \) where \( \zeta \) 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 computationally 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 \( \ell_1 \) regularization term can result in sparse portfolios i.e. the solution of (6) 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 6. 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 (6) and \( I = \text{supp}(w^*) \) is known a priori (before computing the solution). Then the problem (6) can be relaxed to the equivalent problem

\[
\min_{w} w^T R_I w - w^T \mu_I + ||w||_{\beta, \ell_1}
\]

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 which attempts to solve (6) on smaller subspaces, \( I \), where \( \text{supp}(w^*) \subset I \) 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 (6) if and only if \(|(2Rw^*)_i - \hat{\mu}_i| \leq \beta_i \) for all \( i \notin \text{supp}(w^*) \) and \((2Rw^*)_i - \hat{\mu}_i + \beta_i \text{sgn}(w^*_i) = 0 \) for all \( i \in \text{supp}(w^*) \).

**Proof 3** Suppose \( w^* \) solves (6) and let \( i \in \text{supp}(w^*) \). 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

\[
0 = \frac{\partial}{\partial w_i} \Psi(w)|_{w=w^*} = 2(Rw^*)_i - \hat{\mu}_i + \beta_i \text{sgn}(w^*_i)
\]

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

$$\Psi(w) - \Psi(w^*) \geq \sum_{i \in \text{supp}(w^*)} ((2Rw^*)_i - \hat{\mu}_i + \beta_i \text{sgn}(w^*_i)) (w_i - w^*_i) +$$

$$+ \sum_{i \notin \text{supp}(w^*)} ((2Rw^*)_i - \hat{\mu}_i) w_i + \beta_i |w_i| \geq 0$$

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 $i \notin \text{supp}(x)$, and $(2Rx)_i - \hat{\mu}_i > \beta_i$. Then the objective function in (6) can be reduced by adding $i$ into $\text{supp}(x)$. Thus $x$ is not optimal and we should incorporate $i$ into $\text{supp}(x)$.

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 (27) obtained with Algorithm 7. Then

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

**(Proof 4)** By Algorithm 7 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

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

This suggests that the prolongation of $b$ can be defined from equation (28). For example suppose $(\bar{w}, \bar{d}, \bar{b})$ solves (27) on a restricted domain $I \subset \{1, 2, \ldots, N\}$ and let $w$ and $d$ represent the prolongation of $\bar{w}$ and $\bar{d}$ to a set $J \supset I$ i.e.

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

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

Then taking a cue from equation (28) the prolongation of $\bar{b}$ may be defined as

$$b_i = (-2R_{|J|}w + \hat{\mu}_{|J|})_i/\beta_i \lambda.$$
The Adaptive Support Split Bregman Algorithm for solving (27) is given below.

**Algorithm 3** Adaptive Support Split Bregman Algorithm for solving (27)

Initialize: \( k = 0, w^0 = 0, d^0 = 0, b^0 = 0, \epsilon > 0, M > 0 \)

Define \( D^0 = 2Rw^0 - \hat{\mu} \)

while \( |D^k_i| > \beta_i \) for any \( i \notin \text{supp}(w^k) \) AND \( k < N \) do

Define the set \( J^k = \{ D^k_i : i \notin \text{supp}(w^k) \} \)

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 \( w^k_{|I^k}, b^k_{|I^k}, d^k_{|I^k} \) and tolerance \( \epsilon \)

Set \( (w^{k+1}, d^{k+1}) \) to the prolongation of output of previous step

Set \( b^{k+1} = -2(Rw^{k+1} - \hat{\mu})i/(\beta_i \lambda) \),

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 (6) and let \( w' \) be a solution produced by Algorithm 3 for \( \epsilon = \text{tol} \). Then

\[
\Psi(w') \leq \Psi(w^*) + \text{tol}. \tag{32}
\]

**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 2, \( w' \) satisfies the conditions of Theorem 2 with \( \epsilon = \text{tol} \). Thus by Theorem 2 \( \Psi(w') < \Psi(w^*) + \text{tol} \). 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 \( \Psi(w') < \Psi(w^*) + \text{tol} \).

Now we compare the execution speed of Adaptive Support Split-Bregman algorithm, the Modified Split-Bregman algorithm (Algorithm 2), FISTA [2] and a multi-level algorithm proposed in [32]. For the multi-level algorithm proposed in [32] 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 (6) 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 [32] |
|-----------|---------------|--------------------------------|---------------|-------|-----------------------|
| 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                        | 11.6 sec      | 3.6 sec | 1.5 sec               |
| 2000      | 853           | 1.8 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      | 35.7 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 [32] |
|-----------|---------------|--------------------------------|---------------|-------|-----------------------|
| 500       | 53            | 0.05 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 (6) on daily return data from 1600 U.S. stocks collected between January 1, 2001 and July 1, 2014. The results are then compared with an unpenalized mean-variance approach and a uniformly weighted elastic net penalized portfolio.

In our experiments we compute new portfolios every 21 trading days using daily returns from the prior 84 trading days as training data for parameter estimation and calibration of the elastic net weights. The Sharpe ratios of the resulting daily portfolio returns over the subsequent 21 trading days are then computed after each rebalancing. This results in a total of 157 Sharpe ratio data points.

4.1 Parameter Estimation

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 [9]. We estimate the covariance matrix using the technique described in [20]. In that paper the following shrinkage estimator for $\Gamma$ is proposed

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

(33)
where $\Gamma_S$ is the sample average covariance obtained from the training data and where $\rho_1, \rho_2 > 0$. In our experiments we use the optimal values of $\rho_1 > 0$ and $\rho_2 > 0$ which are derived in [20]. Note that this choice of shrinkage target guarantees that $\hat{\Gamma}$ will be positive definite.

For estimation of the mean we use a James-Stein estimator [8, 17] which was proposed for portfolio optimization in [19]. Using the James-Stein approach we compute the estimate of $\mu$ using the equation

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

(34)

Here $\mu_S$ is the sample mean vector and $\eta$ is the average of the sample means i.e.

$$\eta = \frac{1}{N} \sum_{i=1}^{N} \mu_{S,i}$$

(35)

The value of $\rho$ is set according to [19] as

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

(36)

The weights for the weighted elastic net penalty are calibrated using the bootstrap technique described in section 2.4. For cases when we use a uniformly weighted elastic net penalty ($\alpha_i = \alpha_j$ and $\beta_i = \beta_j$ all $i$ and $j$) we average the weights obtained using the bootstrap technique in section 2.4.

4.2 Sharpe Ratio performance

We now present results from our experiment. We tested the following 5 techniques: 1) unpenalized 2) weighted elastic net penalized, 3) weighted $\ell_1$ penalized, 4) uniformly weighted $\ell_1$ penalized, 5) uniformly weighted elastic net. For each technique we employed the shrinkage estimators described above to estimate the mean and covariance parameters.

In Figure 2 we compare the distribution of Sharpe ratios obtained using a weighted elastic net penalty with uniformly weighted penalties. The figure demonstrates that the weighted elastic net penalized criterion improves Sharpe ratio performance. This supports our intuition that modeling distinct uncertainty levels for each individual asset provides more than benefit than assuming a constant uncertainty across all assets.

In Figures 3 and 4 we present the average and median daily Sharpe ratios obtained in each test period for various bootstrap percentiles $P$. As in the previous figure we see that a weighted elastic net penalty results in higher Sharpe ratios compared with the other criteria. For a bootstrap percentile of 70% the weighted elastic net penalized criterion improves the mean Sharpe ratio versus the unpenalized criterion with a $p$-value of $< 4\%$. The figure also shows that the weighted penalties performs better when additional uncertainty is modeled (i.e. higher bootstrap percentile), whereas in the uniformly penalized criteria we see that performance decreases as more uncertainty is modeled. Finally we observe that the weighted elastic net criterion outperforms the weighted $\ell_1$ criterion for the majority of bootstrap percentages which is consistent with the intuition that modeling the additional uncertainty in variance will improve performance.
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 a data-driven criterion for calibration of the elastic net weights based on bootstrapping. 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 can result in higher out-of-sample Sharpe ratio than the uniformly weighted elastic net penalized criterion and the unpenalized criterion. 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 general pairwise elastic net penalty in (11) 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 $\Gamma$ 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 (11) have not been extensively reported on in the literature. We plan to investigate these questions in future work.
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 (6) as a quadratic program.

A.1 Quadratic Program Reformulation

Problem (6) can be reformulated as a quadratic program with linear inequality constraints by introducing an auxiliary variable $d$.

$$\min_{w,d} \Phi(w, d)$$

s.t. $-d_i \leq w_i$

$$-d_i \leq -w_i$$

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

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

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 (37). Our first task is to derive a lower bound on the Lagrangian for a fixed $\lambda$ and when $d = |w|$. First note that $R$ is symmetric positive definite whose smallest eigenvalue is greater than $\alpha_o$ 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

$$\Phi(w, d) \geq L(w, d, \lambda) = L(\tilde{w}, \tilde{d}, \lambda) + \nabla_w L(\tilde{w}, \tilde{d}, \lambda)^T (w - \tilde{w}) + \nabla_d L(\tilde{w}, \tilde{d}, \lambda)^T (d - \tilde{d}) + \nabla_w L(\tilde{w}, \tilde{d}, \lambda)^T (w - \tilde{w}) + \nabla_d L(\tilde{w}, \tilde{d}, \lambda)^T (d - \tilde{d})$$

$$\geq L(\tilde{w}, \tilde{d}, \lambda) + \nabla_w L(\tilde{w}, \tilde{d}, \lambda)^T (w - \tilde{w}) + \nabla_d L(\tilde{w}, \tilde{d}, \lambda)^T (d - \tilde{d}) + \alpha_o \|w - \tilde{w}\|_2^2 + \frac{1}{2} \alpha_o \|d - \tilde{d}\|_2^2$$

$$\geq L(\tilde{w}, \tilde{d}, \lambda) + \nabla_w L(\tilde{w}, \tilde{d}, \lambda)^T (w - \tilde{w}) + \nabla_d L(\tilde{w}, \tilde{d}, \lambda)^T (d - \tilde{d}) + \frac{1}{2} \alpha_o \|w - \tilde{w}\|_2^2 + \frac{1}{2} \alpha_o \|d - \tilde{d}\|_2^2$$

(39)

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 $\|\nabla_{w,d} L(\tilde{w}, \tilde{d}, \lambda)\|_{\ell^2} \leq 2\epsilon\alpha_o$. Then $L(\tilde{w}, \tilde{d}, \lambda) \leq \Phi(w^*, d^*) + \epsilon$ where $w^*$ solves (6) and $d_i^* = |w_i^*|$ for all $i$.

**Proof 6** By equation (39) we have

\[
\Phi(w^*, d^*) \geq L(w^*, d^*, \lambda) \geq L(\tilde{w}, \tilde{d}, \lambda) + \nabla_w L(\tilde{w}, \tilde{d})^T (w^* - \tilde{w}) + \nabla_d L(\tilde{w}, \tilde{d})^T (d^* - \tilde{d}) + \frac{1}{2\alpha_o} \|w^* - \tilde{w}\|^2_{\ell^2} + \frac{1}{2\alpha_o} \|d^* - \tilde{d}\|^2_{\ell^2}
\]

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

\[
\begin{align*}
x_1 + x_2 &= b \\
-x_1 + x_2 &= a
\end{align*}
\]

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 (6).

**Proof 7** of Theorem 2

Choose $d^*$ and $\tilde{d}$ such that $d_i^* = |w_i^*|$ and $\tilde{d}_i = |\tilde{w}_i|$. For $i \in \text{supp}(\tilde{w})$ define $\lambda$ such that

\[
\lambda_i = \begin{cases} 
0 & \text{if } w_i > 0, i \in \text{supp}(\tilde{w}) \\
\beta & \text{if } w_i < 0, i \in \text{supp}(\tilde{w})
\end{cases}
\]

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

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

\[
\lambda_i + \lambda_{i+N} = \beta_i
\]

and

\[
-\lambda_i + \lambda_{i+N} = -\frac{\partial}{\partial w_i} \left( w^T R w - w^T \hat{\mu} \right) \bigg|_{w = \tilde{w}}
\]

By Lemma 8 equation (21) implies that such a $\lambda_i, \lambda_{i+N}$ exists.

Let us form the Lagrangian $L(w, d, \lambda)$ as in equation (38). Then for $i \in \text{supp}(\tilde{w})$

\[
\frac{\partial}{\partial w_i} L(w, d, \lambda)_{|_{\tilde{w}, \tilde{d}}} = \frac{\partial}{\partial w_i} \left( w^T R w - w^T \hat{\mu} + \|w\|_{\beta, \ell^1} \right) \bigg|_{w = \tilde{w}}
\]
and

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

For $i \notin \text{supp}(\tilde{w})$ we have by equation (41)

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

and by equation (40)

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

It then follows from equation (20) that

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

and so by Lemma 7 and our choice of $\lambda$ we have that

$$\Phi(\tilde{w}, \tilde{d}) = L(\tilde{w}, \tilde{d}, \lambda) \leq \Phi(w^*, d^*) + \epsilon$$

This clearly implies that

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

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 $\|\zeta - \tilde{w}\|_{\ell_\infty} \leq \|\zeta - \tilde{w}\|_{\ell_2} \leq \frac{\epsilon\sqrt{\epsilon\alpha_o}}{M}$. It follows that

$$\sum_{i \in \text{supp}(\zeta)} \left( \frac{\partial}{\partial w_1} (w^T R w - w^T \hat{\mu} + \|w\|_{\beta, \ell_1})_{w=\zeta} \right)^2 \leq \left( \sqrt{2} + 1 \right)^2 \alpha_o \epsilon.$$

and

$$-\beta_i \leq \frac{\partial}{\partial w_i} (w^T R w - w^T \hat{\mu})_{w=\zeta} \leq \beta_i$$

for all $i \notin \text{supp}(\zeta)$. So by Theorem 2 we have that $\zeta$ satisfies (26).

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 Sciences, 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] 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.

[7] 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.

[8] 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.

[9] B. Efron and C. Morris, *Stein’s Paradox in Statistics*, Scientific American, 237 (1977), pp. 119–127.

[10] 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.

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

[12] 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.

[13] B. Fastrich, S. Paterlini, and P. Winker, *Constructing optimal sparse portfolios using regularization methods*, Computational Management Science, (Submitted to).

[14] 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.

[15] D. Goldfarb and G. Iyengar, *Robust Portfolio Selection Problems*, Mathematics of Operations Research, 28 (2003), pp. 1–38.

[16] T. Goldstein and S. Osher, *The Split Bregman Method for $\ell_1$ Regularized problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 323–343.

[17] 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.

[18] J. Jobson and B. Korkie, *Estimation for Markowitz Efficient Portfolios*, Journal of the American Statistical Association, 75 (1980), pp. 544–554.

[19] P. Jorion, *Bayes-Stein Estimation for Portfolio Analysis*, The Journal of Financial and Quantitative Analysis, 21 (1986), pp. 279–292.
[20] O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, Journal of Multivariate Analysis, 88 (2004), p. 365411.

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

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

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

[24] 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.

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

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

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

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

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

[30] 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.

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

[32] E. Treister and I. Yavneh, *A Multilevel Iterated-Shrinkage Approach to $\ell_1$ Penalized Least-Squares Minimization*, IEEE Trans. Signal Proc., 60 (2012), pp. 6319–6329.

[33] Y. Yu-Min, *Three essays in financial econometrics*, PhD thesis, University of London, 2012.