CrossMark

Constructing optimal sparse portfolios using regularization methods

B. Fastrich · S. Paterlini · P. Winker

Abstract Mean-variance portfolios have been criticized because of unsatisfying out-of-sample performance and the presence of extreme and unstable asset weights, especially when the number of securities is large. The bad performance is caused by estimation errors in inputs parameters, that is the covariance matrix and the expected return vector. Recent studies show that imposing a penalty on the 1-norm of the asset weights vector (i.e. ℓ_1-regularization) not only regularizes the problem, thereby improving the out-of-sample performance, but also allows to automatically select a subset of assets to invest in. However, ℓ_1-regularization might lead to the construction of biased solutions. We propose a new, simple type of penalty that explicitly considers financial information and then we consider several alternative penalties, that allow to improve on the ℓ_1-regularization approach. By using U.S.-stock market data, we show empirically that the proposed penalties can lead to the construction of portfolios with

B. Fastrich · P. Winker
Department of Economics, University of Giessen, Licher Strasse 64, 35394 Giessen, Germany
e-mail: Bjoern.Fastrich@wirtschaft.uni-giessen.de

P. Winker
e-mail: Peter.Winker@wirtschaft.uni-giessen.de

S. Paterlini (✉)
Department of Finance and Accounting, European Business School,
Gustav-Stresemann-Ring 3, 65189 Wiesbaden, Germany
e-mail: sandra.paterlini@ebs.edu

P. Winker
Center for European Economic Research (ZEW), L 7, 1, 68161 Mannheim, Germany

Springer

an out-of-sample performance superior to several state-of-art benchmarks, especially in high dimensional problems.

Keywords Minimum variance portfolio · Statistical regularization · Lasso · Non-convex penalties

1 Introduction

The Markowitz mean-variance portfolio model (Markowitz 1952) indisputably constitutes a milestone in modern finance. Given a set of assets with expected return vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, Markowitz' model aims to find the optimal asset weight vector that minimizes the portfolio variance, subject to the constraint that the portfolio exhibits a desired portfolio return. Since $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, some estimates $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ must be obtained from a finite sample of data to compute the optimal asset allocation vector.

As financial literature has largely shown, using sample estimates can hardly provide reliable out-of-sample asset allocations in practical implementations (e.g., Jobson and Korkie 1980; Best and Grauer 1991; Broadie 1993; Britten-Jones 1999; DeMiguel et al. 2009a). Frankfurter et al. (1971), Dickinson (1974), Jobson and Korkie (1980), and Frost and Savarino (1988) already provided strong empirical evidence that the estimates of the expected portfolio return and variance are very unreliable. By now, it is commonly accepted that estimation errors in the expected return estimates are much larger than those in the covariance matrix estimates (e.g. Merton 1980). Hence, in this investigation, we choose to join the large group of researchers who have more recently focused on the minimum-variance portfolio (MVP), which relies solely on the covariance structure and neglects the estimation of expected returns altogether (see, e.g., Chan et al. 1999; Jagannathan and Ma 2003; Ledoit and Wolf 2003; DeMiguel and Nogales 2009; Fan et al. 2012; Fernandes et al. 2012; Behr et al. 2012). Despite evidence on MVPs shows that they perform better out-of-sample than portfolios that consider asset means (Jorion 1986; Jagannathan and Ma 2003; DeMiguel et al. 2009a), they still suffer considerably from estimation errors (Chan et al. 1999; Jagannathan and Ma 2003; Ledoit and Wolf 2003).

In principle, the opportunity for allowing portfolio managers to select their constituents from an extremely large universe of assets is desirable, because it corresponds with a large potential to diversify risk. Unfortunately, exploiting such potential is very difficult, as the presence of estimation error increases and the out-of-sample performance worsens when the dimensionality K increases relative to the number of observations T. Thus, if, as it often happens, K is large compared to T, only an insufficient amount of data is available to precisely estimate the $K(K + 1)/2$ parameters in $\widehat{\boldsymbol{\Sigma}}$. For example, Ledoit and Wolf (2003) observe that small sample problems in covariance matrix estimation, such as severe estimation errors and numerical instability, may occur whenever T is not at least ten times larger than K.[1] As a consequence, a portfolio manager might not be able to (optimally) benefit from a high diversification

---

[1] Note that sample covariance matrix estimates are singular when $T < K$.

potential while (somehow) selecting constituents to obtain portfolios of a manageable size. Nevertheless, as a manageable portfolio size is essential, the ability to select good assets is an important requisite.

One research stream has recently focused on shrinking asset allocation weights by using regularization methods. Among the first contributors, Brodie et al. (2009) and DeMiguel et al. (2009b) use ℓ_1-regularization to obtain stable and sparse (i.e. with few active weights) portfolios, which is an adaptation of the Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996).[2] Both studies provide some insights into the earlier work of Jagannathan and Ma (2003) on no-shortselling portfolios, because ℓ_1-regularization allows a limit on the total shorting of a portfolio. Moreover, the empirical results show that the newly proposed portfolio construction methods outperform many strategies in terms of the out-of-sample portfolio risk and Sharpe ratio. Fan et al. (2012) provide both theoretical and empirical evidence supporting the use of ℓ_1-regularization to identify sparse and stable portfolios by limiting the gross exposure, showing that this causes no accumulation of estimation errors, the result of which is an outperformance compared to standard Markowitz portfolios. Recent examples of regularization techniques applied in the Markowitz framework are Yen (2010), Carrasco and Noumon (2011), Yen and Yen (2011), and Fernandes et al. (2012), while in index tracking are Giomouridis and Paterlini (2010) and Fastrich et al. (2014).

The large appeal of using ℓ_1-regularization in portfolio optimization is the possibility to estimate (numerically stable) asset weights and select the portfolio constituents in a single step by solving a convex optimization problem. An ℓ_1-penalty term in the objective function prevents an overfitting of the portfolio to erroneous estimates, thereby preventing the estimation errors from entering the portfolio without restriction and allowing a better out-of-sample performance. However, Fan and Li (2001) show that the ℓ_1-penalty, as a linear function of absolute coefficients, tends to produce biased estimates for large (absolute) coefficients. As a remedy, they propose using penalties that are singular at the origin, just like the ℓ_1-penalty, in order to promote sparsity, but non-convex, in order to countervail bias.[3] An alternative to non-convex approaches, which can still retain the oracle property, has been suggested by Zou (2006). His approach is now known as the adaptive LASSO and has proven to be able to prevent bias while preserving convexity of the optimization problem, and thus clearly alleviates the optimization challenge as compared to the non-convex approaches.

This work contributes to the literature on portfolio regularization by proposing a new, simple type of convex penalty, which is inspired by the adaptive LASSO and explicitly considers financial information to optimally determine the portfolio composition. Moreover, we are, to our knowledge, the first to apply non-convex penalties in the Markowitz framework to identify sparse and stable portfolios. We particularly focus on four types of non-convex penalty functions and provide a financial interpre-

---

[2] The LASSO relies on imposing a constraint on the ℓ_1-norm the regression coefficients. In this paper, ℓ_1-regularization is used synonymously.

[3] They claimed that a good penalty function should result in an estimator with three properties: unbiasedness, sparsity, and continuity. They also coined the term oracle property, which, in a nutshell, means that a strategy performs as well as when the true underlying model is known a priori.

tation of their effect on asset weights. Extensive empirical investigations with various U.S.-stock market data sets and covariance matrix estimators support the validity of our proposals.

The paper is organized as follows: Sect. 2 introduces the asset selection framework as well as the (new) penalty functions and provides their economic motivations. Section 3 discusses the experimental set-up, while Sect. 4 comments on the main empirical findings on small and large data sets. Finally, Sect. 5 concludes this work.

2 Regularization approaches for minimum variance portfolios

Given a set of K assets and a penalty function $\rho(\cdot)$, the regularized minimum-variance problem can be stated as:

$$\boldsymbol{w}^* = \underset{\boldsymbol{w} \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ \boldsymbol{w}' \boldsymbol{\Sigma} \boldsymbol{w} + \lambda \sum_{i=1}^{K} \rho(w_i) \right\} \tag{1}$$

$$\text{subject to} \qquad \mathbf{1}'_K \boldsymbol{w} = 1, \tag{2}$$

where $\boldsymbol{w}^*$ is the optimal (and potentially sparse) $(K \times 1)$-vector of asset weights, $\mathbf{1}_K$ is a $(K \times 1)$-vector of ones and λ is the regularization parameter that controls the intensity of the penalty and thereby the sparsity of the optimal portfolio.

Assuming we estimate $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\Sigma}}$ and we set $\lambda = 0$, the solution to problem (1)–(2) is simply the MVP, where the optimized portfolio weights vector $\boldsymbol{w}^*$ is (over)fitted to the correlation structure in $\widehat{\boldsymbol{\Sigma}}$. Such a fit implies unlimited trust in the precision of the estimate $\widehat{\boldsymbol{\Sigma}}$, which is obviously very naive. In fact, especially when K is large relative to sample size T, $\widehat{\boldsymbol{\Sigma}}$ contains a considerable amount of estimation errors that suggests the presence of a diversification potential that is actually non-existent and typically leads to large absolute components in $\boldsymbol{w}^*$. The latter will cause the penalty term $\sum_{i=1}^{K} \rho(w_i)$ to be large, which then prevents the unreal diversification potential in $\widehat{\boldsymbol{\Sigma}}$ from being excessively exploited whenever $\lambda > 0$. The larger λ, the smaller is the number of active weights and the total amount of shorting. The optimal solution $\boldsymbol{w}^*$ is thus determined by a trade-off between the estimated portfolio risk and the corresponding penalty term, whose magnitude is controlled by λ.

In this work, we focus on penalty functions $\rho(\cdot)$ that are singular at the origin and thus allow a shrinkage of the components in $\boldsymbol{w}$ to exactly zero. Hence, the corresponding approaches not only stabilize the problem to improve the out-of-sample performance, but simultaneously also conduct the asset selection step.

2.1 LASSO and weighted LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is the only one of the proposed approaches that has already received considerable attention in the portfolio optimization context (Brodie et al. 2009; DeMiguel et al. 2009b; Yen 2010; Yen and Yen 2011; Carrasco and Noumon 2011; Fan et al. 2012; Fernandes et al. 2012). Hence, we choose it as a benchmark to test the validity of the newly proposed approaches.

(a) LASSO-penalty **(b)** $w8Las$-penalty
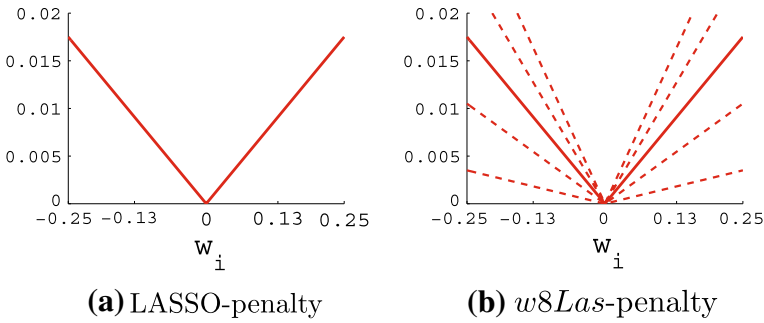
Fig. 1 The Lasso **(a)** and w8Las penalty **(b)** functions

Problem formulation (1)–(2) constructs ℓ_1-regularized (LASSO) portfolios if the weighted penalty term is proportional to the 1-norm of the asset weight vector (i.e. $||\boldsymbol{w}||_1$), such that:

$$\lambda \sum_{i=1}^{K} \rho(w_i) = \lambda(|w_1| + |w_2| + \cdots + |w_K|) = \lambda||\boldsymbol{w}||_1 \tag{3}$$

Due to the budget constraint (2), the minimum value that $||\boldsymbol{w}||_1$ can be shrunk to is one. This is only possible when the portfolio weights are shrunk towards zero until they are all non-negative, identifying the so-called no-shortsale portfolio (Jagannathan and Ma 2003). Hence, only if the portfolio exhibits many and/or large negative weights, $||\boldsymbol{w}||_1$ takes on large values. In fact, we can re-write (3) as:

$$\lambda||\boldsymbol{w}||_1 = \lambda \left(1 - 2 \sum_{j=\{i:i=1,\ldots,K \ | \ w_i<0\}} w_j \right). \tag{4}$$

Equation (4) shows that the LASSO-penalty can only increase when the shorting increases and that its minimum value is λ. Increasing values of λ cause the construction of portfolios with less shorting, or more precisely, with a shrunken ℓ_1-norm of the portfolio weight vector. This prevents the estimation errors contained in $\widehat{\boldsymbol{\Sigma}}$ from entering unhindered in the portfolio weight vector. Hence, the economic intuition of the LASSO-regularized portfolios is to limit the shorting of the whole portfolio instead of the shorting of individual assets via additional constraints.[4] Note that while the intensity of shrinkage is controlled by the value of λ, the decision which assets to shrink and to which relative extent is determined by the estimated correlation structure. Finally, from the optimization viewpoint, as Fig. 1a shows, introducing the lasso penalty still results in a convex optimization problem, which can be easily solved for large dimensional problems.

Zou (2006) proposed a new version of the convex LASSO, displayed in Fig. 1b, to countervail the difficulties of the LASSO that are related to potentially biased estimates

---

[4] This is a very useful approach for practitioners, among whom the so-called 130/30, 120/20, and 110/10 portfolios are very popular. These portfolios consist of, e.g., 130 percent long and 30 percent short positions.

B. Fastrich et al.

of large true coefficients (Fan and Li 2001). The idea is to replace the equal penalty that is applied to all coefficients (here portfolio weights) with a penalization-scheme that can vary among the K portfolio weights. This can be easily achieved by introducing a weight ω_i so that, in equation (3), each absolute portfolio weight $|w_i|$ is assigned an individual regularization parameter $\lambda_i = \lambda \omega_i$. This explains the abbreviation $w8Las$ used henceforth.[5] If $\forall\, i\ \omega_i = 1$, then the $w8Las$-penalty is identical to the LASSO-penalty. In contrast, whenever there is reason to assume that asset i can contribute to the desired portfolio performance by more (less) than is implied by the LASSO-regularized portfolio weights, the factor ω_i should be reduced (increased) in order to favor (discriminate against) this asset by a lower (higher) slope of the penalty function.

In general, the economic intuition is to over- or underweight some assets in comparison to the LASSO in order to improve performance. Specifically, this intuition depends on the method used to determine the ω_i, for which no "blueprint" exists in a portfolio optimization context. In the original work of Zou (2006) the weights are computed by $\omega_i = 1/|\widehat{\beta_i}|^\gamma$, where $\widehat{\beta_i}$ is an OLS-estimate and $\gamma > 0$. This is now known as the adaptive LASSO. In a financial context, Brodie et al. (2009) suggest some ideas of how to use the additional flexibility to consider transaction costs (e.g. by determining ω_i according to the bid-ask spread of asset i). But, to our knowledge, no implementation and empirical results have been provided so far in a financial context. We interpret the $w8Las$-penalty as a possibility not only to regularize the MVP, but also to incorporate investment styles of any kind, while still dealing with a convex optimization problem. In particular, we suggest determining the (individual) regularization weights λ_i by considering specific financial time series properties that are ignored when, e.g. $T = 250$, historical observations are used to estimate one (constant) covariance matrix. The components of the latter can be thought of as average (co)variances of the sample with T observations, while trends and clusters in the volatilities (heteroscedasticity) remain unexploited when the portfolio is constructed. Figure 2a shows one year of daily returns for the guns-industry portfolio, which is one of the 48 industry portfolios provided by Kenneth French (thin line, left axis). The remaining lines correspond to the right axis. The thick dashed line represents the volatility of the sample covariance matrix computed with the 250 observations shown, while the thick solid line corresponds to the historical return volatility of the most recent 4 months of daily data. Clearly, a trend of a decreasing volatility can be observed and is supported by a moving average over the most recent 42 observations of the 4-month-volatility (thick dashed-dotted line). This scenario indicates that the volatility in the near future will also be below the value of the covariance matrix estimate. This should materialize in both a smaller penalty λ_i and, consequently, a larger portfolio weight in comparison to the LASSO.

We operationalize this idea by extracting three signals from the data that determine the individual factors ω_i, $i = 1, \ldots, K$. One component for the signals is the historical return volatility over the last $T_1 = 84$ observations, at time t denoted by $\widehat{\sigma}(T_1)_{i,t}$—the thick solid line in Fig. 2a. Another component is a moving average over the last $T_2 = 42$ observations of $\widehat{\sigma}(T_1)_{i,t}$, at time t denoted by $MA\big[\widehat{\sigma}(T_1)_i, T_2\big]_t$—the thick dashed-dotted line in Fig. 2a. The last component for the three signals are the main-

---

[5] Obviously, $w8$ abbreviates weighted and Las LASSO.

(a) Guns-industry portfolio

(b) $\theta_4 = 1.5 > 1$, strong strategy impact.
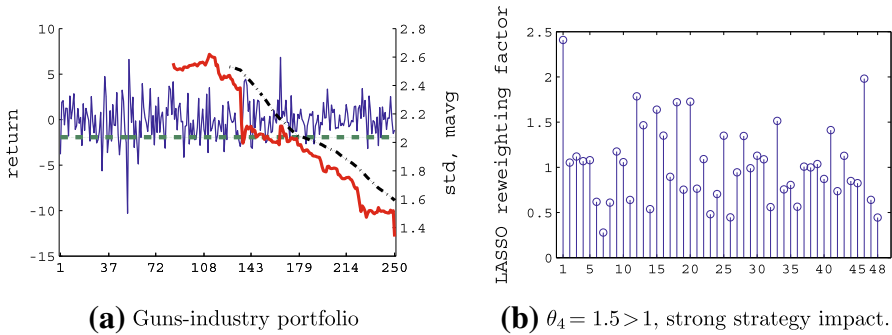
Fig. 2 **a** One year of daily returns from 23.08.2002 to 23.08.2003 (*thin line*), full-period standard deviation obtained from the covariance matrix estimator (*thick dashed line*), standard deviation based on the last 4 months of daily return observations (*thick solid line*), 2 months moving average of the *thick solid line* (*thick dashed-dotted line*). All values are reported in percent. **b** Re-weighting factors ω_i of the $K = 48$ individual constituents

diagonal components of the covariance matrix estimate $\widehat{\boldsymbol{\Sigma}}_{i,i}$—the thick dashed line in Fig. 2a.

We define the first signal Υ_1 as:

$$\Upsilon_{1,i,T} = MA\big[\widehat{\sigma}(T_1)_i, T_2\big]_T - \sqrt{\widehat{\boldsymbol{\Sigma}}_{i,i}}, \tag{5}$$

where T represents the current end of the sample. The smaller the value of $\Upsilon_{1,i,T}$ is, the smaller the regularization weight and thus the penalty for asset i should ceteris paribus be, because its out-of-sample volatility is expected to be smaller than implied by the covariance matrix estimate.

The second and the third signals attempt to capture recent trends in the historical volatility $\widehat{\sigma}(T_1)_{i,t}$. For that, we define Υ_2 as:

$$\Upsilon_{2,i,T} = \widehat{\sigma}(T_1)_{i,T} - MA\big[\widehat{\sigma}(T_1)_i, T_2\big]_T. \tag{6}$$

Whenever $\Upsilon_{2,i,T} < 0$, the recent return volatility is smaller than its average, thereby indicating a negative short-term tendency and vice versa. A negative (positive) *short-term* tendency should ceteris paribus decrease (increase) λ_i and thus the penalty. Analogously, we try to capture *longer-term* tendencies by the third signal Υ_3:

$$\Upsilon_{3,i,T_3,T} = MA\big[\widehat{\sigma}(T_1)_i, T_2\big]_T - MA\big[\widehat{\sigma}(T_1)_i, T_2\big]_{T-T_3}, \tag{7}$$

i.e. by the difference between the moving average at time T and at time $T-T_3$, where $T_3 = 42$.

The three signals (5)–(7) shown in Fig. 2a are now combined to obtain the individual factors ω_i to reweight the LASSO-regularization parameter λ according to $\lambda_i = \lambda \omega_i$. First, the K values per signal are shifted by the absolute value of the most negative realization in order to avoid negativity (denoted by a plus-sign in the superscript). Then, the shifted signals are weighted and scaled by their mean to obtain the individual factors ω_i:

$$\omega_i = \left(\frac{\theta_1 \Upsilon_{1,i,T}^+ + \theta_2 \Upsilon_{2,i,T}^+ + \theta_3 \Upsilon_{3,i,T_3,T}^+}{\frac{1}{K} \sum_{i=1}^{K} \theta_1 \Upsilon_{1,i,T}^+ + \theta_2 \Upsilon_{2,i,T}^+ + \theta_3 \Upsilon_{3,i,T_3,T}^+} \right)^{\theta_4}. \tag{8}$$

Parameters $\theta_1, \theta_2, \theta_3 > 0$ determine the relative importance of the (shifted) signals while parameter $\theta_4 > 0$ controls the general impact of the strategy on the portfolio weight vector. Figure 2b shows the factor ω_i for all assets $i = 1, \ldots, 48$, when $\theta_1 = 0.3$, $\theta_2 = 0.5, \theta_3 = 0.2$ and $\theta_4 = 1.5$. Asset $i = 26$ corresponds to Fig. 2a (the guns portfolio) and exhibits a factor of about $\omega_{26} = 0.5$. In other words, the individual regularization parameter $\lambda_{26} = \lambda \omega_{26}$ is half the size as it is in the LASSO-approach. Consequently, the favorable expectations with respect to the (near-future) volatility of the guns industry will encourage investment in this sector. In case the portfolio allocation ought to be more similar to the LASSO-portfolio, parameter θ_4 can be reduced to a value of, e.g., $\theta_4 = 0.5$. Then, the individual factors ω_i are shrunk towards a value of one.

Note that beyond the simple strategy we suggest, the penalty construction allows for a simple processing of all kinds of signals, may they be gained from (time series) econometrics, fundamental or technical analysis, or expert knowledge.

2.2 Non-convex penalties

Statistical research has shown that the class of non-convex penalties provides an alternative approach to countervail the difficulties of the LASSO related to potentially biased estimates of large absolute coefficients. In particular, Gasso et al. (2009) have already shown the good model selection performance of non-convex penalties when K and the level of correlation are high. The common idea behind this class is to penalize gains in small (absolute) portfolio weights more heavily than gains in large (absolute) portfolio weights. This causes the portfolio weights to take on more extreme long and short positions than in the LASSO-penalty, because it allows a reduction of the estimated portfolio risk without increasing the value of the penalty term (by much). The economic intuition behind the non-convex penalties is based on the fact that if the true correlation of assets is high, shorting can reduce the risk, since it accounts for true similarities of the assets instead of being the result of overfitting. Analogously, large portfolio weights tend to be appropriate if the true correlations are small. Now, if a correlation structure is "strong enough" to grow absolute portfolio weights—against the counteracting penalty—large enough, it is considered reliable and should therefore enter the portfolio to a greater extend. The ℓ_q- and the Log-penalty provide a particularly strong incentive to avoid small and presumably dispensable positions in favor of selecting a small subset of presumably indispensable assets. This tendency to construct very sparse and less diversified portfolios coincides with the suggestion of Fernholz et al. (1998) to use the ℓ_q-norm as a diversity measure for portfolios.

Table 1 reports the mathematical descriptions of the non-convex penalties and their corresponding domains, while Fig. 3 shows all the non-convex penalties considered in this work for a single asset weight. Figure 3a presents the non-convex SCAD-penalty (Fan and Li 2001), which is linear for small, and constant for large $|w_i|$. In between the linear and the constant domain, there is a "softly clipped" domain. A discontinuous and piecewise linear approximation of the SCAD-penalty is sometimes referred to as

Table 1 Regularization penalties

Penalty $\lambda\rho(w_i)$	Domain
LASSO $= \lambda|w_i|$	All
$w8Las = \lambda\omega_i|w_i|$	All
SCAD $= \begin{cases} \lambda|w_i| \\ \frac{-|w_i|^2+2a\lambda|w_i|-\lambda^2}{2(a-1)} \\ \frac{(a+1)\lambda^2}{2} \end{cases}$	$\begin{aligned} |w_i| &\leq \lambda \\ \lambda &< |w_i| \leq a\lambda \\ a\lambda &< |w_i| \end{aligned}$
Zhang $= \begin{cases} \lambda|w_i| \\ \lambda\eta \end{cases}$	$\begin{aligned} |w_i| &< \eta \\ \eta &\leq |w_i| \end{aligned}$
$L_q = \lambda|w_i|^q,\ 0<q<1$	All
Log $= \begin{aligned} \lambda\ln(|w_i|+\phi) \\ -\lambda\ln(\phi) \end{aligned}$	All

the Zhang-penalty (see, e.g., Gasso et al. 2009) and shown by Fig. 3b. It introduces further discontinuities at locations η and $-\eta$ that separate the linear and the constant penalty domain. The fact that η is independent of the value λ can be beneficial for parameterization.[6] Figure 3c shows the ℓ_q-penalty (Frank and Friedman 1993; Fu 1998; Knight and Fu 2000), which nests classical subset selection when $q = 0$, the LASSO when $q = 1$, and the Ridge Regression (Hoerl and Kennard 1970) when $q = 2$. However, a simultaneous asset selection and portfolio weight estimation is only possible when $0 < q \leq 1$, which is the domain considered in this paper.[7] Then, compared to the LASSO, the penalty imposed is relatively large for an increase in small absolute weights, which is a characteristic that is exhibited neither by the SCAD nor by the Zhang-penalty, and imposes a higher incentive on the portfolio-construction process to avoid small positions. Figure 3d shows the Log-penalty (see, e.g. Weston et al. 2003), which can be considered an approximation of the ℓ_0-penalty and showed a good sparsity recovering capability (Candes et al. 2008).

3 Data and experimental setup

3.1 Data sets

We consider daily observations of five different data sets, summarized in Table 2, to evaluate the proposed portfolio approaches. Data sets, that represent the U.S. stock market at different levels of aggregation, are characterized by a different number of constituents: from now on, the 48 industry portfolios and the 98 firm portfolios provided by Kenneth French are referred to as *small* data sets,[8] while the remaining

---

[6] It is relatively easy to choose a value for η because it can be directly interpreted as the threshold that determines how large an absolute portfolio weight must (at least) be to be constantly penalized.

[7] The parameter ϕ is simply a very small increment that prevents division by zero.

[8] In the latter case firms were sorted according to the market (ME) and the book-to-market (BE/ME) value to obtain the firm portfolios. The data sets may be downloaded from the homepage of Kenneth French (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html), where information on the portfolio construction methods is also provided.
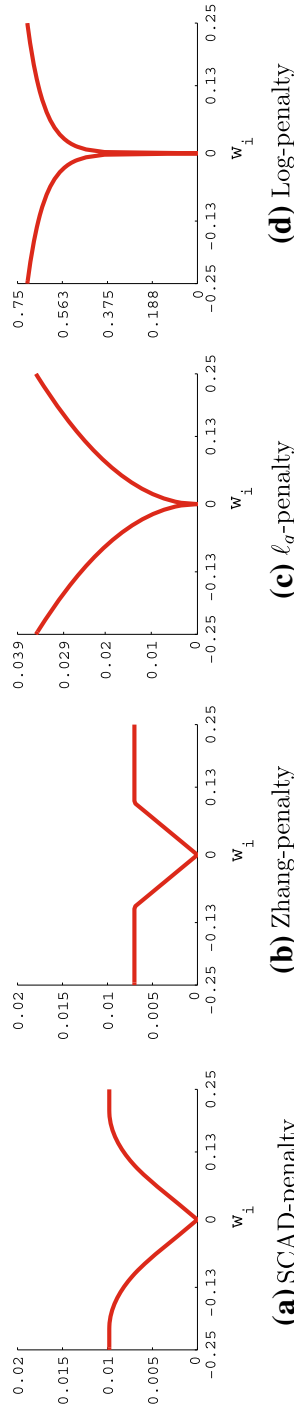
Fig. 3 The four non-convex penalty functions

(a) SCAD-penalty

(b) Zhang-penalty

(c) ℓ_q-penalty

(d) Log-penalty

Table 2 U.S. stock market data sets for the period 23.08.02 to 27.03.08

Data set	Source	T	K	$\bar{r}$	$\hat{\sigma}$	$\hat{S}$	$\widehat{Ku}$
F&F 48: industry portfolios	K. French	1,401	48	11.83	14.95	−0.0935	4.27
F&F 100: firm portfolios (ME & BE/ME)	K. French	1,401	98	8.42	15.03	−0.0037	5.12
S&P 200: largest firms (w.r.t. ME)	Datastream	1,401	200	6.57	14.79	0.0487	5.32
S&P 500: largest firms (w.r.t. ME)	Datastream	1,401	500	6.57	14.77	0.0410	5.13
S&P 1,036: largest firms (w.r.t. ME)	Datastream	1,401	1036	6.39	14.88	0.0380	4.99

Table reports the data sets under consideration, the source of the data, the number of assets (K), and the number of observations (T) in each data set. For the S&P data sets, value weighted indices are computed whose return distributions are characterized by the mean p.a. ($\bar{r}$), the standard deviation p.a. ($\hat{\sigma}$), the skewness ($\hat{S}$), and the kurtosis ($\widehat{Ku}$) given in the last four columns. For the F&F48 data set, an equally weighted index is computed due to the non-availability of market values or an appropriate analogue, while the returns of the portfolios in the F&F100 data set are weighted by the average firm size that a portfolio is composed of. The S&P indices are market value weighted. The weighting schemes are updated daily and applied the following day

ones, which include the 200, 500, and 1,036 largest individual firms (with respect to the market value on March 27, 2008) of the S&P 1500 constituents list, are referred to as *large* data sets.[9]

3.2 Backtesting procedure, performance measures, and benchmarks

We backtest the out-of-sample performance of the proposed methods with a moving time window procedure, where $\tau = 250$ in-sample observations (corresponding to one year of market data) are used to form a portfolio. The optimized portfolio allocations are then kept unchanged for the subsequent 21 trading days (corresponding to 1 month of market data) and the out-of-sample returns are recorded. After holding the portfolios unchanged for 1 month, the time window is moved forward, so that the formerly out-of-sample days become part of the in-sample window and the oldest observations drop out. The updated in-sample window is then used to form a new portfolio, according to which the funds are reallocated. The $T = 1,401$ observations allow for the construction of $\Gamma = 54$ portfolios with the corresponding out-of-sample returns, which are then used to compute the out of sample portfolio variance (i.e. $s^2 = \frac{1}{T-\tau-1} \sum_{t=\tau+1}^{T} (r_t - \bar{r})^2$), the Sharpe Ratio (i.e. $SR = \frac{\bar{r}}{\sqrt{s^2}}$) and the Value-at-Risk (i.e. VaR_{1-p}) as the $F_r^{-1}(p)$, that is value of the inverse cumulated empirical distribution function of the daily out-of-sample returns at probability level p. Moreover, to compare the portfolio composition, we also compute the number of active positions (i.e. $No.act. = \frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} |\{i \mid w_{i,\gamma} \neq 0 \ \forall \ i\}|$), the total amount of shorting (i.e. $Short = \frac{1}{\Gamma} \sum_{j=\{i \mid w_{i,\gamma} < 0 \ \forall \ i\}} -w_{j,\gamma}$) and the turnover (i.e. $TO = \frac{1}{\Gamma-1} \sum_{\gamma=2}^{\Gamma} \sum_{i=1}^{K} |w_{i,\gamma} - w_{i,\gamma-1}|$).

---

[9] The largest data set consists of 1,036 assets, because in the course of data preparation, penny stocks and series with missing data were excluded.

For comparative evaluations, we consider the following standard benchmarks: (1) the shortsale-unconstrained MVP, denoted MVPssu, the shortsale-constrained MVP, denoted MVPssc, the market value weighted portfolio, denoted mvw, and the equally weighted portfolio, denoted 1oK. The above performance measures are reported for the benchmarks, the only exception being the mvw in the case of the F&F 48 data set, where neither the market values of the firms that compose the industry portfolios nor an appropriate analogue is provided by Kenneth French.[10]

3.3 Covariance matrix estimators

For determining the minimum variance portfolio, we focus on two frequently used covariance matrix estimators: (1) the sample estimator and (2) a three-factor model estimator (Chan et al. 1999). Note that we keep the in-sample window size $\tau = 250$ fixed throughout all experiments, which makes the sample estimator exhibit a poor quality quickly when K/T grows. We do not choose to simply increase τ to improve the quality of the estimator, because, in reality, portfolio selection problems contain thousands of assets to choose from, so that tens of thousands of observations are required. Time series of this length are hardly available or, if they are, contain in a sense "outdated" observations. Instead, the three-factor model estimator imposes a factor structure on the returns and is thus clearly more stable, allowing for a better quality even for large ratios K/T. The cost of the higher (matrix) stability of this estimators is its lower precision.

3.4 Optimization algorithm

The optimization task is very challenging when the penalty function in problem (1)–(2) is specified by one of the non-convex approaches. We modify the DC-programming approach proposed by Gasso et al. (2009) to tackle the non-convex optimization problems. The optimization relies on using an iterative primal dual-approach for the non-convex penalty functions considered, where in each iteration a convex primal problem is solved with a gradient projection algorithm proposed by Figueirdo et al. (2007). More details on the optimization procedure can be found in the supplementary materials.

4 Empirical analysis

Prior to optimizing problem formulation (1)–(2) for any of the six penalization approaches, a value of the regularization parameter must be chosen. Section 4.1 provides some insights on the relationship between the choice of λ and the portfolio risk and shorting profiles, while Sect. 4.2 reports results in a realistic investment context, when cross-validation is used to choose the optimal value of λ^*. For each approach, we set a grid of 30 ascending values starting from zero. The largest element in each set is chosen such that the resulting portfolios exhibit only few active positions and a high out-of-sample portfolio variance. In this way, it is most likely that the intervals spanned by zero and the largest regularization parameters cover λ^*.

---

[10] This is different for the 98 firm portfolios (see the caption of Table 2).

4.1 Portfolio risk and shorting profiles

As a first set of experiments, we fix a priori a grid of 30 possible values for λ and determine in each rolling window the optimal portfolio in correspondance of a given value of λ. Figure 4 shows for the S&P 1,036 data set the portfolio variance (labeled as risk) and the mean shorting of all approaches in comparison to the LASSO, which serves as a benchmark and is shown in all graphs. For each approach the entire spectrum of 30 portfolios of different sizes is shown—corresponding to 30 regularization parameters fixed a priori.[11] Note that the leftmost LASSO-portfolio corresponds to the MVPssu and the rightmost to the MVPssc. The profile of the LASSO shows that the risk (s^2 and VaR) can be minimized by investing neither in the MVPssu nor the MVPssc, but instead in a regularized portfolio. The reason for this is that the imposed penalties prevent estimation errors in $\widehat{\Sigma}$ from fully and freely entering the portfolio weights and deteriorating the performance. Comparing the risk profiles among the various penalties reveals that not only our $w8Las$ but also all non-convex penalties are capable of constructing portfolios with a lower minimum risk than the LASSO. This is a first indication for the success of the portfolio regularization strategies proposed in this paper. While the minimum risk portfolio of the Zhang and the $w8Las$ approach exhibit about as many active positions as the LASSO, the (other) non-convex approaches construct clearly smaller and thus more practical portfolios, especially when dealing with a large pool of candidate assets. That the minimum risk portfolios of the latter group exhibit about the same amount of shorting but fewer constituents than the LASSO, is a consequence of the more uncompromising fit of the (left) active positions to the estimated correlation structure. This characteristic is especially beneficial in large data sets where the complexity of the underlying asset selection problem is sufficiently high. Then, it is advantageous to maintain presumably relevant assets in the portfolio with relatively large absolute weights, while an increasing regularization parameter can shrink the absolute weights only of presumably less relevant assets. In contrast, the LASSO always shrinks all weights, causing a "biased estimate" of the portfolio vector. The asset-specific penalties in the $w8Las$ approach circumvent this problem. However, the $w8Las$ performs especially well in the small data sets as it can be seen in Sect. 4.2 and in the supplementary materials.[12]

4.2 Choosing the regularization parameter by cross validation

As a second set of experiments, we consider a real-world context where the portfolio manager has to determine and update his portfolio allocation every month. Then, we invest in only *one* of the 30 possible optimal portfolio by applying tenfold cross-validation to choose a suited value of λ prior to the investment decision in each period. The cross-validation procedure is as follows: 21 observations are randomly picked from the in-sample data, portfolios are optimized on the remaining 229 observations

---

[11] The abscissa denotes the number of active positions instead of the values of λ, since the scaling of the latter differs among the penalties.

[12] Note that the latter contain detailed results for five data sets and three covariance matrix estimators. These results do not only support the above findings but also provide various additional insights.

(a) *w8Las* vs. LASSO

(b) SCAD vs. LASSO

(c) Zhang vs. LASSO

(d) ℓ_q $(q=0.5)$ vs. LASSO

(e) Log vs. LASSO
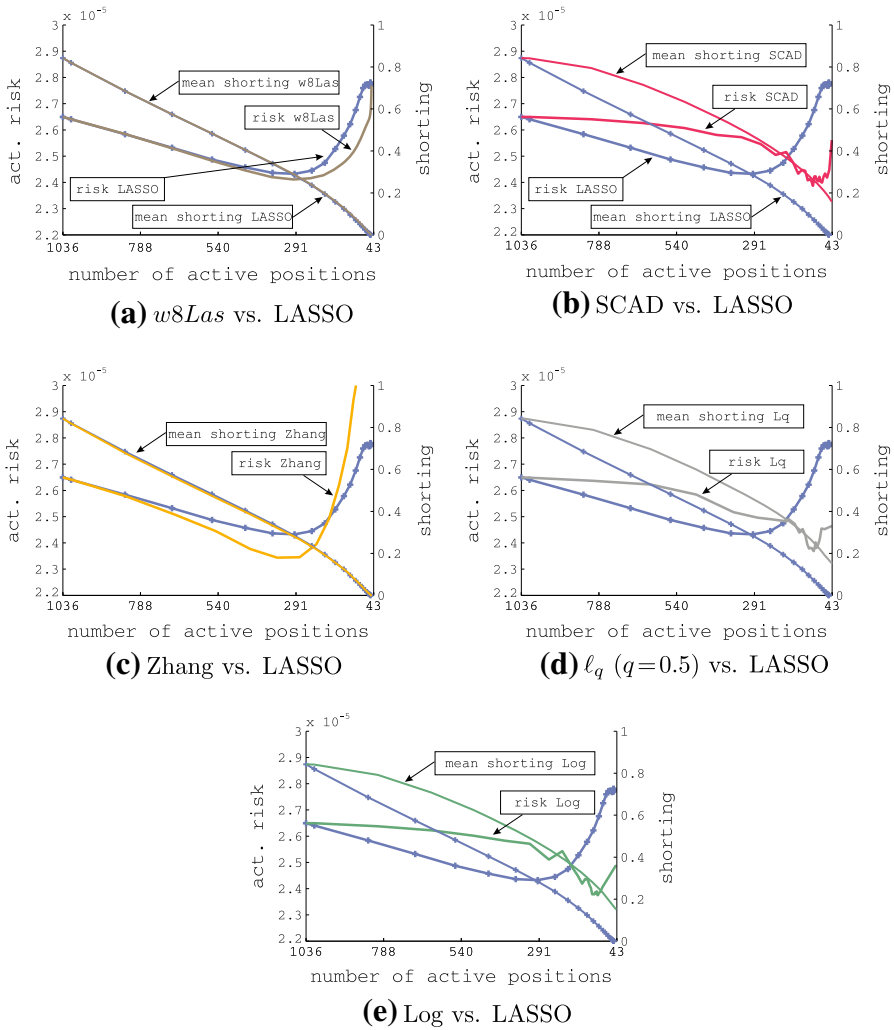
Fig. 4 Risk and shorting profiles of portfolios constructed by the LASSO and the competing regularization strategies based on the three-factor model covariance matrix computed with the S&P 1,036 data set. The risk and shorting profile of the LASSO is shown in all graphs with markers [see the tags in graphs (**a**) and (**b**)]

for all 30 regularization parameters, and the portfolio variance is computed using the 21 picked observations. This is done ten times and the chosen λ is the one that corresponds to the smallest average portfolio variance.

4.2.1 Small data sets

Table 3 shows the results for the sample covariance matrix estimator in a realistic setting when ten-fold cross-validation is applied to the current in-sample data to determine the optimal λ. The results found in the previous section can generally be confirmed,

Table 3 Sample covariance matrix - Cross-validation choice of λ

	MVPssu	MVPssc	mvw	1oK	Lasso	w8Las	Log	ℓ_q	Zhang	SCAD
Panel A: F&F 48 industry portfolios										
$s^2 \cdot 10^5$	2.890	3.471	–	7.061	2.669	2.624	2.734	2.732	2.875	2.803
$VaR \cdot 10^2$	0.879	1.017	–	1.446	0.834	0.820	0.816	0.806	0.850	0.818
SR	0.058	0.053	–	0.041	0.062	0.060	0.067	0.070	0.060	0.064
$No.\,act.$	48.0	9.1	–	48.0	25.1	27.0	15.5	15.2	21.7	15.4
$Short$	1.17	0.00	–	0.00	0.42	0.48	0.42	0.42	0.47	0.42
TO	1.06	0.23	–	0.00	0.56	0.64	0.74	0.77	0.85	0.74
Panel B: F&F 100 ME-BE-portfolios										
$s^2 \cdot 10^5$	1.984	4.633	6.389	8.414	1.667	1.654	1.928	1.959	1.903	1.928
$VaR \cdot 10^2$	0.659	1.163	1.352	1.555	0.576	0.585	0.616	0.615	0.608	0.616
SR	0.102	0.059	0.027	0.028	0.096	0.101	0.101	0.103	0.092	0.100
$No.\,act.$	98.0	5.6	98.0	98.0	46.2	48.4	27.5	25.0	28.6	27.6
$Short$	3.19	0.00	0.00	0.00	1.19	1.26	1.28	1.19	1.28	1.27
TO	3.04	0.31	0.03	0.00	1.39	1.55	2.20	2.10	2.13	2.20

Table shows in the upper part of both panels the performance measures based on the out-of-sample returns, while the lower parts summarize the portfolio compositions

i.e. in Table 3 all regularization approaches outperform all benchmarks with respect to the risk (s^2 and VaR). Note that the stabilizing effect of regularization increases with K/T. Compared to the best benchmark (MVPssu), s^2 can be reduced by about 20 % for F&F 100 in Panel B compared to about 10 % for F&F 48 in Panel A. The Sharpe ratios of the regularized strategies are higher as well, with the only exception being the MVPssu in Panel B. Our proposed $w8Las$-strategy shows the lowest variance, which is a result we obtained in all experiments using the small data sets (see supplementary materials). The non-convex strategies consist of fewer active positions with about the same amount of shorting (or slightly more). This shows the capability of non-convex approaches of identifying sparse solution by avoiding (over)fitting to the estimated correlation structure. Moreover, the non-convex regularized portfolios tend to exhibit higher Sharpe ratios while their risks are overall slightly higher than those of the convex approaches. Thus, considering the fact that the need for solutions with very few active position might not be as important in small data sets as in large dataset, empirical results for F&F48 and F&F100 points out the benefit of using the LASSO and the newly proposed w8Las strategies.

4.2.2 Large datasets

Table 4 shows that the cross-validation approach works well for the large data sets. In Panel B for S&P 500 and C for S&P 1,036, the portfolio variances and Value-at-Risks of all regularization approaches reach lower levels of portfolio variance than MVPssu, MVPssc, mvw and 1oK, while having a much smaller number of active positions. This shows that the possibility of having a stronger shrinkage in some periods but not in others can be beneficial. The same is not true for the S&P 200 data set in Panel

Table 4 Three-factor model covariance matrix—Cross-validation choice of λ

	MVPssu	MVPssc	mvw	1oK	Lasso	w8Las	Log	ℓ_q	Zhang	SCAD
Panel A: S&P 200 individual firms										
$s^2 \cdot 10^5$	3.007	3.162	6.023	6.524	2.843	2.808	3.017	3.009	2.777	2.942
$VaR \cdot 10^2$	0.885	0.898	1.312	1.348	0.828	0.824	0.893	0.916	0.843	0.881
SR	0.054	0.062	0.018	0.050	0.049	0.050	0.054	0.048	0.049	0.054
$No.\ act.$	200.0	54.9	200.0	200.0	82.6	91.1	66.1	65.6	93.9	64.8
$Short$	0.75	0.00	0.00	0.00	0.26	0.29	0.38	0.38	0.32	0.39
TO	0.57	0.52	0.04	0.00	0.59	0.68	0.96	0.98	0.73	0.90
Panel B: S&P 500 individual firms										
$s^2 \cdot 10^5$	2.883	3.796	6.081	6.799	2.529	2.495	2.617	2.601	2.538	2.643
$VaR \cdot 10^2$	0.923	1.071	1.335	1.385	0.834	0.835	0.794	0.814	0.847	0.842
SR	0.031	0.042	0.018	0.045	0.043	0.043	0.043	0.049	0.042	0.036
$No.\ act.$	500.0	278.6	500.0	500.0	131.9	147.6	102.8	108.1	151.6	101.0
$Short$	0.83	0.00	0.00	0.00	0.20	0.24	0.33	0.35	0.24	0.33
TO	0.61	0.22	0.04	0.00	0.69	0.75	1.11	1.04	0.80	1.09
Panel C: S&P 1036 individual firms										
$s^2 \cdot 10^5$	2.649	4.593	6.254	9.001	2.382	2.379	2.343	2.356	2.485	2.369
$VaR \cdot 10^2$	0.833	1.166	1.352	1.566	0.802	0.792	0.775	0.789	0.819	0.754
SR	0.031	0.031	0.016	0.028	0.054	0.050	0.041	0.045	0.050	0.044
$No.\ act.$	1,036.0	572.4	1,036.0	1,036.0	276.7	308.3	179.6	153.8	298.7	161.3
$Short$	0.84	0.00	0.00	0.00	0.26	0.30	0.33	0.31	0.28	0.31
TO	0.65	0.22	0.04	0.00	0.84	0.89	1.30	1.13	0.87	1.26

Table shows results of the four benchmarks and the six regularization approaches for the three large data sets and the three-factor model covariance matrix

A, where the Log- and the ℓ_q-regularized portfolios exhibit slightly higher risks than the MVPssu. However, this fits the picture that the non-convex approaches can be mostly beneficial when dealing with large data sets. In fact, the number of non-convex regularized portfolios that outperform the convex ones is only one in Panel A, while it is three in Panel C. However, this relationship cannot be shown as clearly as it can for the a priori choice of λ in Fig. 4. Further research on how to optimally choose λ is needed.

5 Conclusion

In this work, we show that the performance of the minimum-variance portfolio can be substantially improved by using regularization methods that aim to control (over)fitting of the asset weight vector to the estimation errors in the covariance matrix. Our main contributions and findings can be smmarized as follows.

First, we propose a new type of a (convex) penalty whose construction allows for easy processing of all kinds of signals to optimized portfolios, may they be gained from

(time series) econometrics, fundamental or technical analysis, or expert knowledge. We have shown that over- or underweighting assets based on a simple strategy works well, especially when dealing with small data sets. It almost always leads to the construction of portfolios with a smaller variance and often a higher Sharpe ratio than all benchmarks and even the LASSO-portfolios. The individual penalization of asset weights was also shown to partly compensate for the disadvantages of the LASSO in larger data sets (even though the non-convex approaches dominated).

Second, we contribute to the existing literature on portfolio-regularization by providing empirical results for several non-convex penalty functions that have not yet been examined in a portfolio optimization context. It turned out that these approaches perform very well when dealing with very large data sets, where they not only outperformed standard benchmarks but also the (convex) "state-of-the-art" LASSO approach. The success of these approaches stems from their ability to maintain relevant assets in the portfolio with large absolute weights, while only the weights of the remaining assets are shrunk. This can be thought of as a financial interpretation of the oracle property-notion used in the statistical regularization literature and allows for a better exploitation of the higher potential to diversify portfolio risk in larger data sets.

Third, we show that the regularization parameter can be determined successfully via ten-fold cross-validation. However, several cross-validation criteria, considering different objective functions, can be specified and might allow for a further improvement of the portfolio performances. So far, our results refer to the minimum variance portfolios. Extending the portfolio set-up by considering alternative risk-measures, such as tail-related or one-sided ones, is currently high on our agenda for future research.

References

Behr P, Guettler A, Truebenbach F (2012) Using industry momentum to improve portfolio performance. J Bank Financ 36(5):1414–1423

Best M, Grauer J (1991) On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results. Rev Financ Stud 4(2):315–342

Britten-Jones M (1999) The sampling error in estimates of mean-variance efficient portfolio weights. Ann Oper Res 54(2):655–671

Broadie M (1993) Computing efficient frontiers using estimated parameters. Ann Oper Res 45(1):2158

Brodie J, Daubechies I, DeMol C, Giannone D, Loris D (2009) Sparse and stable Markowitz portfolios. Proc Natl Acad Sci USA 106(30):12267–12272

Candes E, Waking M, Boyed S (2008) Enhancing sparsity by reweighted l_1 minimization. J Fourier Anal Appl 14(5):877–905

Carrasco M, Noumon N (2011) Optimal portfolio selection using regularization. Working Paper University of Montreal. Available from http://www.unc.edu/maguilar/metrics/carrasco.pdf

Chan L, Karceski J, Lakonishok J (1999) On portfolio optimization: forecasting covariances and choosing the risk model. Rev Financ Stud 12(5):937–974

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

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

DeMiguel V, Nogales F (2009) Portfolio selection with robust estimation. Oper Res 57(3):560–577

Dickinson J (1974) The reliability of estimation procedures in portfolio analysis. J Financ Quant Anal 9(3):447–462

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360

Fan J, Zhang J, Yu K (2012) Vast portfolio selection with gross exposure constraints. J Am Stat Assoc 107(498):592–606

Fastrich B, Paterlini S, Winker P (2014) Cardinality versus q-norm constraints for index tracking. Quantitative Finance 14(11):2019–2032

Fernandes M, Rocha G, Souza T (2012) Regularized minimum-variance portfolios using asset group information, pp 1–28. Available from http://webspace.qmul.ac.uk/tsouza/index_arquivos/Page497.htm

Fernholz R, Garvy R, Hannon J (1998) Diversity weighted indexing. J Portf Manag 24(2):74–82

Figueirdo M, Nowak R, Wright S (2007) Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J Sel Top Signal Process 1(4):586–596

Frank I, Friedman J (1993) A statistical view of some chemometrics regression tools. Technometrics 35(2):109–135

Frankfurter G, Phillips H, Seagle J (1971) Portfolio selection: the effects of uncertain means, variances, and covariances. J Financ Quant Anal 6(5):1251–1262

Frost P, Savarino J (1988) For better performance: constrain portfolio weights. J Portf Manag 15(1):29–34

Fu J (1998) Penalized regression: the bridge versus the lasso. J Comput Graph Stat 7(3):397–416

Gasso G, Rakotomamonjy A, Canu S (2009) Recovering sparse signals with a certain family of nonconvex penalties and DC programming. IEEE Trans Signal Process 57(12):4686–4698

Giomouridis D, Paterlini S (2010) Regular(ized) hedge funds. J Financ Res 33(3):223–247

Hoerl A, Kennard R (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1):55–67

Jagannathan R, Ma T (2003) Risk reduction in large portfolios: Why imposing the wrong constraints helps. J Financ 58(4):1651–1683

Jobson J, Korkie R (1980) Estimation for Markowitz efficient portfolios. J Am Stat Assoc 75(371):544–554

Jorion P (1986) Bayes-Stein estimation for portfolio analysis. J Financ Quant Anal 21(3):279–292

Knight K, Fu W (2000) Asymptotics for lasso-type estimators. Ann Stat 28(5):1356–1378

Ledoit O, Wolf M (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J Empir Financ 10(5):603–621

Markowitz H (1952) Portfolio selection. J Financ 7(1):77–91

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

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. R Stat Soc 58(1):267–288

Weston J, Elisseeff A, Schölkopf B (2003) Use of the zero-norm with linear models and kernel methods. J Mach Learn Res 3:1439–1461

Yen Y-M (2010) A note on sparse minimum variance portfolios and coordinate-wise descent algorithms, pp 1–27. Available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1604093

Yen Y-M, Yen T-J (2011) Solving norm constrained portfolio optimizations via coordinate-wise descent algorithms, pp 1–41. Available from http://personal.lse.ac.uk/yen/sp_090111.pdf

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429