Noname manuscript No.
(will be inserted by the editor)

Optimal Portfolio Selections via $\ell_{1,2}$-norm Regularization

Hongxin Zhao · Lingchen Kong · Hou-Duo Qi

Received: date / Accepted: date

Abstract There has been much research about regularizing optimal portfolio selections through ℓ_1 norm and/or ℓ_2-norm squared. The common consensuses are (i) ℓ_1 leads to sparse portfolios and there exists a theoretical bound that limits extreme shorting of assets; (ii) ℓ_2 (norm-squared) stabilizes the computation by improving the condition number of the problem resulting in strong out-of-sample performance; and (iii) there exist efficient numerical algorithms for those regularized portfolios with closed-form solutions each step. When combined such as in the well-known elastic net regularization, theoretical bounds are difficult to derive so as to limit extreme shorting of assets. In this paper, we propose a minimum variance portfolio with the regularization of ℓ_1 and ℓ_2 norm combined (namely $\ell_{1,2}$-norm). The new regularization enjoys the best of the two regularizations of ℓ_1 norm and ℓ_2-norm squared. In particular, we derive a theoretical bound that limits short-sells and develop a closed-form formula for the proximal term of the $\ell_{1,2}$ norm. A fast proximal augmented Lagrange method is applied to solve the $\ell_{1,2}$-norm regularized problem. Extensive numerical experiments confirm that the new model often results in high Sharpe ratio, low turnover and small amount of short sells when compared with several existing models on six datasets.

The paper was supported in part by 111 Project of China (B16002), IEC/NSFC/191543 and NSFC(12071022).

H. Zhao
Department of Applied Mathematics, Beijing Jiaotong University, Beijing, 100044, China.
E-mail: hxzhao@bjtu.edu.cn

L. Kong
Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, China.
E-mail: lchkong@bjtu.edu.cn

H.-D Qi
School of Mathematical Sciences, University of Southampton, SO17 1BJ, UK. E-mail: hdqi@soton.ac.uk

Keywords Portfolio optimization · minimum variance portfolio · $\ell_{1,2}$-norm regularization · proximal augmented Lagrange method · out-of-sample performance

Mathematics Subject Classification (2020) 90C20 · 90C25 · 90C90

1 Introduction

It has long been known that Markowitz's mean variance model [25] is sensitive to the estimation errors in the expected returns and the covariance matrix, leading to its poor out-of-sample performance, see DeMiguel et al. [10], Green and Hollifield [16]. Various approaches have been proposed to enhance its performance, see, e.g., Ben et al. [3], Lhabitant [22], Maillet [24] and Olivier and Wolf [28]. One outstanding approach is to construct the minimum variance portfolio (MVP) by regularizing its portfolio weight vector $\mathbf{w} \in \mathbb{R}^N$ (we assume there are N risky assets in the portfolio). The rationale of this approach is well justified by Jaganathan and Ma [18] for two reasons. The first reason is that there exists strong empirical evidence that the estimates of the expected returns and covariance matrix can be contaminated by large errors, especially in large portfolios, see Olivier and Wolf [29,30]. However, the estimation errors in the covariance matrix may be significantly less than those for the expected returns, see Merton [27]. The second reason is that constraining on the weights vector (e.g., enforcing lower and upper bounds on each individual weight w_i) can actually lead to a portfolio with reduced risk compared to the unconstrained MVP [18, Prop. 1]. Those two strong arguments have initiated extensive research on what is now known as the regularization methods for constructing robust and stable portfolios. For example, DeMiguel et al. [9] proposed a general framework for norm-constrained portfolio selection problems. Behr et al. [2] imposed the upper and lower portfolio weight constraints. Shen et al. [32] considered a doubly regularized technique which is the combination of the ℓ_1 norm and the squared ℓ_2 norm. Fastrich et al. [12] analyzed the regularization technique in portfolio optimization and Kremer et al. [20] studied the sorted ℓ_1-norm regularization. A large number of relevant references can be found in those papers.

This paper contributes to the literature a new regularization method that is related to the ℓ_1 norm and ℓ_2 norm of the weight vector ($\|\mathbf{w}\|_1 := |w_1| + \cdots + |w_N|$ and $\|\mathbf{w}\|_2 := \sqrt{w_1^2 + \cdots + w_N^2}$). The use of ℓ_1-norm in portfolio selection has been well studied. It is well established in statistics and optimization that the ℓ_1-norm penalty promotes sparsity among the decision variables [6] and this was confirmed by Brodie et al. [5] for constructing sparse portfolios, often called Lasso portfolios in literature due to its deep root in the Lasso method in statistics [35]. Moreover, ℓ_1-norm yields certain theoretical bounds that regularize the amount of short-sells (Brodie et al. [5]) and control the risks originated from the estimation errors in the expected returns and covariance matrix (Fan et al. [11]). In other words, the estimation errors cannot be propagated excessively over any algorithmic process. Limiting extreme positions also reduces the overall transaction cost. DeMiguel et al. [9], [10] and Fan et al. [11] have conducted detailed study on treating ℓ_1 norm as a constraint, i.e., $\|\mathbf{w}\|_1 \le c$, where c is a given constant (called the exposure constraint in [11]). In particular, when $c = 1$, the ℓ_1 norm constraint, together with the budget constraint $w_1 + \cdots + w_N = 1$, enforces the nonnegativity constraints $w_i \ge 0$, $i = 1, \ldots, N$, recovering the portfolio of Jagannathan and Ma [18]. However, when the non-negativity constraints are present, the ℓ_1-norm penalty does not work because in this case $\|\mathbf{w}\|_1$ is constant ($= 1$). Furthermore, the Lasso portfolios are often highly sensitive to multicollinearity, see Giuzio and Paterlini [14] for detailed comments.

DeMiguel et al. [9] also studied the ℓ_2 norm constraint ($\|\mathbf{w}\|_2 \le \delta$). When penalized to the objective, the squared norm $\|\mathbf{w}\|_2^2$ will evidently improve the condition number of the quadratic risk term $\mathbf{w}^T(V + I_N)\mathbf{w}$, where V is an estimator of covariance matrix and I_N is the identity matrix of size N. This leads to stable portfolios that avoid extreme positions. When $\delta = 1$, the ℓ_2 norm constraint reduces to the equally weighted portfolio $w_i = 1/N$ for all i [10]. Since ℓ_2 norm promotes dense solutions, it improves diversification of the resulting portfolio, see Takeda et al. [33]. Because of this density effect, it may result in many small weights (often too small to be practically implemented). Another negative effect is that it may incur higher transaction cost. Given ℓ_1 norm tends to reduce the transaction cost (Brodie [5]), a natural approach is to combine ℓ_1 norm and ℓ_2 norm to construct both stable and sparse portfolios.

This idea of combining ℓ_1 and ℓ_2 norm together has also been extensively investigated. Yen and Yen [37] considered the case of ℓ_1 norm with the squared ℓ_2 norm ($\lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2$, with $\lambda_1, \lambda_2 \ge 0$ being penalty parameters), which is known as the elastic net in statistics [39]. Furthermore, a coordinate-wise descent algorithm was proposed to solve this type of problems. Yen [36] further conducted a detailed analysis of this approach. It is interesting to note that the weighted elastic net has a nice interpretation in robust optimization [17], where a split-Bregman algorithm was proposed. However, the budge constraint $\mathbf{1}_N^T\mathbf{w} = 1$ was not included in the model of [17]. Using uncorrelated data, Li [23] explained that the squared ℓ_2 norm is equivalent to enforcing a scaling factor to the ℓ_1 norm and hence will not harm the sparsity induced by the ℓ_1 norm in this case. In terms of algorithmic design, the most useful fact in using ℓ_1-norm and ℓ_2-norm squared is the separability in its variables. They can be put in a uniform form $\sum_{i=1}^N \rho_i(w_i)$ with each $\rho_i(\cdot)$ being a nonnegative function. For the ℓ_1-norm and ℓ_2-norm squared, $\rho_i(w_i) = \lambda_1|w_i| + \lambda_2 w_i^2$. This separability property has been exploited by Fastrich et al. [12]) to develop more general models and is also the mathematical reason behind the coordinate-wise descent algorithm proposed in Yen and Yen [37]. However, the benefit of employing ℓ_2-norm squared comes at a heavy cost in that theoretical bounds such as [5, Eq. (7)] for the ℓ_1-norm penalty are hard to establish. Its effect on the sparsity of the resulting portfolio has also not been well understood. The deep reason for those shortcomings is that the magnitude of the ℓ_2-norm squared is of the squared order of the ℓ_1-norm, i.e., $\|\mathbf{w}\|_2^2 = O(\|\mathbf{w}\|_1^2)$.

In this paper, we propose a new combination called $\ell_{1,2}$-norm: $(\lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2)$ and establish its theoretical and algorithmic justification. The main criticism is probably on the fact that ℓ_2-norm is not differentiable while its squared form $\|\mathbf{w}\|_2^2$ is continuously differentiable and strongly convex. As we will illustrate that we will not lose too much. For example, $\|\mathbf{w}\|_2$ is continuously differentiable everywhere except when $\mathbf{w} = 0$, which is not a feasible portfolio anyway. Moreover, we will gain a lot. The main contributions of our investigation are summarized as follows.

(i) We will explain that the ℓ_2-norm $\|\mathbf{w}\|_2$ is strongly convex over the subspace of $(N-1)$ dimensions that is perpendicular to the one-dimensional space Span$\{\mathbf{w}\}$ spanned by the vector $\mathbf{w} \neq 0$. The implication of this result is that the ℓ_2-norm still improves the condition number of various objectives encountered in portfolio selections as long as we have not reached an optimal solution yet.

(ii) Since the ℓ_2-norm $\|\mathbf{w}\|_2$ has the same magnitude as the ℓ_1-norm $\|\mathbf{w}\|_1$, i.e., $\|\mathbf{w}\|_2 = O(\|\mathbf{w}\|_1)$, we are able to establish a theoretical bound similar to that of Brodie et al. [5, Eq. (7)]. This bound shows that the amount of short-selling can be controlled by appropriately adjusting the penalty parameters λ_1 and λ_2 and hence prevents extreme positions from happening.

(iii) It is widely known that the reason why the ℓ_1-norm promotes sparsity is that its proximal operator (now known as the soft-thresholding operator) shrinks small quantities to zero. It is interesting to note that the $\ell_{1,2}$-norm also admits a proximal operator that is a product of two shrinkage operators, of which one is the soft-thresholding operator. This means that the $\ell_{1,2}$-norm is likely to promote as sparse solutions as the ℓ_1-norm alone would. We will also compare our proximal formula of $\ell_{1,2}$-norm with that of [38] (see Remark 3).

(iv) The availability of the proximal operator for $\ell_{1,2}$-norm opens a large venue for applying existing algorithms in convex optimization [1] to the $\ell_{1,2}$-norm regularized portfolio selections. In particular, we develop a proximal Augmented Lagrange Method (pALM) and demonstrate its efficiency using six real world data sets including DJIA, NASDAQ, SP500, Russell2000, Russell3000 and FF100. Extensive numerical experiments confirm that the new model often results in high Sharpe ratio, low turnover and small amount of short sells when compared with several existing models on those data sets.

The paper is organized as follows. In the next section, we review some norm-penalized portfolio models with comments on their advantages and disadvantages. Section 3 aims to justify the use of the new $\ell_{1,2}$-norm regularized portfolio. In particular, we show that it inherits from the Lasso portfolio a result that bounds the short positions, see (3.3). We also prove that the $\ell_{1,2}$ norm admits a closed-form solution for its proximal operator, which is a composition of the double soft-thresholding shrinkage operators, respectively from the ℓ_1 and ℓ_2 norm, see Prop. 1. In Section 4, we develop a proximal augmented

Lagrange method for solving our $\ell_{1,2}$-norm regularized problem. Numerical experiments are reported in Section 5. We conclude the paper in Section 6.

Notation: We use boldfaced small letters to denote vectors, e.g., $\mathbf{w} \in \mathbb{R}^N$ is a column vector of N elements of w_i, $i = 1, \ldots, N$. The transpose of $\mathbf{w}$ is denoted as $\mathbf{w}^T$, which is a row vector. In particular, $\mathbf{1}_n$ is the vector of all ones of size n. For a vector $\mathbf{a} \in \mathbb{R}^N$, we define its absolute value vector by $|\mathbf{a}|^T := (|a_1|, \cdots, |a_N|)$. For two vectors $\mathbf{a}$ and $\mathbf{b}$ of the same length, $\langle \mathbf{a}, \mathbf{b} \rangle$ denotes the standard inner product and $\mathbf{a} \circ \mathbf{b}$ denotes a new vector of the componentwise product of $\mathbf{a}$ and $\mathbf{b}$, i.e., $(\mathbf{a} \circ \mathbf{b})_i = a_i b_i$. We further define the weighted ℓ_1-norm with weight $\mathbf{c} > 0$ by

$$\|\mathbf{w}\|_{1,\mathbf{c}} := \|\mathbf{c} \circ \mathbf{w}\|_1 = \sum_{i=1}^{N} c_i |w_i|.$$

For a scalar $x \in \mathbb{R}$, we define $x_+ := \max(0, x)$ and the sign function $\mathrm{sgn}(x)$ takes the value 1 if $x > 0$, 0 if $x = 0$ and -1 if $x < 0$. For given $\lambda > 0$, we denote the soft thresholding operator by

$$\mathcal{S}_\lambda(x) := \Big(|x| - \lambda\Big)_+ \mathrm{sgn}(x) \qquad \forall\, x \in \mathbb{R} \tag{1.1}$$

2 Norm-penalized Portfolio Selections

In this section, we describe several norm-penalized portfolio selection models that have motivated our research with a view of possible extension and numerical comparison later on. Suppose there are N risky assets with returns r_i, $i = 1, \ldots, N$, sampled over T periods. Let $\mathbf{r}_t^T = (r_{1,t}, \ldots, r_{N,t})$ be the sample return vector during the period t. Let $\boldsymbol{\mu}$ and V be the theoretical return vector and covariance matrix of the return vector $\mathbf{r}$. The risk associated with a portfolio $\mathbf{w}$ among those N assets is $\mathbf{w}^T V \mathbf{w}$. Usually, both $\boldsymbol{\mu}$ and V are estimated from the sampled data $\{\mathbf{r}_t\}_1^N$. As pointed out in the introduction, the Markowitz mean-variance model is more sensitive to the estimation error in $\boldsymbol{\mu}$ than that in V.

A significant model that departs from the Markowitz model is the minimum variance portfolio (MVP) with nonnegative constraints on the weights by Jagannathan and Ma [18]:

$$\min\ \mathbf{w}^T V \mathbf{w} \quad \text{s.t.} \quad \mathbf{1}_N^T \mathbf{w} = 1, \quad \mathbf{w} \geq 0. \tag{2.1}$$

DeMiguel et al. [9] considered more general constraints on the weight vector $\mathbf{w}$:

$$\min\ \mathbf{w}^T V \mathbf{w} \quad \text{s.t.} \quad \mathbf{1}_N^T \mathbf{w} = 1, \quad \|\mathbf{w}\| \leq \delta, \tag{2.2}$$

where the norm $\|\cdot\|$ can be ℓ_1, ℓ_2 norm or even A-norm $\|\cdot\|_A$ with $\|\mathbf{w}\|_A^2 = \mathbf{w}^T A \mathbf{w}$ and A is positive definite, and $\delta > 0$ is a pre-set constant. The ℓ_1-norm constraint is also treated as an exposure constraint by Fan et al. [11]. Empirical evidences show that the constrained MVP models (2.1) and (2.2) yield much better out-of-sample performance than that of the Markowitz model. Extended research on norm-constrained portfolios has been done, see the review part of Behr et al. [2].

Another important development is the Lasso-style portfolio proposed by Brodie et al. [5]:

$$\mathbf{w}^{[\lambda]} = \arg\min \ \|\overline{\mu}\mathbf{1}_T - R\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \quad \text{s.t. } \mathbf{1}_N^T\mathbf{w} = 1, \ \ \boldsymbol{\mu}^T\mathbf{1}_N = \overline{\mu}, \tag{2.3}$$

where $\overline{\mu}$ is an expected return set by the investors, $\lambda > 0$ is a penalty parameter and R is the $T \times N$ return data matrix whose tth row is $\mathbf{r}_t^T$, $t = 1, \ldots, T$. It has long been known that ℓ_1 norm promotes sparsity among decision variables and it can be explained through the KKT condition of (2.3), see Dai and Wen [8]. One important property of the Lasso portfolio is

$$(\lambda_1 - \lambda_2)(\|\mathbf{w}^{[\lambda_2]}\|_1 - \|\mathbf{w}^{[\lambda_1]}\|_1) \geq 0. \tag{2.4}$$

This bound was used to argue that when λ is above a threshold λ_0, the portfolio $\mathbf{w}^{[\lambda]}$ will have no short selling, for more comments, see the part below Eq. (7) of Brodie et al. [5]. Another important implication is as follows. We note that

$$\|\mathbf{w}\|_1 = 1 + 2\sum_{i:\ w_i<0} |w_i| = 1 + 2p\%, \quad \forall\ \mathbf{w} \in \mathbb{R}^n, \tag{2.5}$$

where $p\%$ denotes the total amount of short sells in $\mathbf{w}$. Consider the case $\lambda_1 > \lambda_2$. Then the bound (2.4) translates into

$$p_{\lambda_1}\% \leq p_{\lambda_2}\%,$$

which means that as the penalty parameter increases in (2.3), the total amount of short sells will not increase. Hence, the Lasso portfolio is unlikely to lead to extreme positions as the penalty parameter increases.

The elastic-net regularized portfolios were considered in Yen and Yen [37], Yen [36], Ho et al. [17], and Li [23]. Combined with MVP, the elastic-net regularization leads to the following optimization problem:

$$\min \ \frac{1}{2}\mathbf{w}^T V\mathbf{w} + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2, \quad \text{s.t.} \quad \mathbf{1}_N^T\mathbf{w} = 1. \tag{2.6}$$

Excellent out-of-sample performance by the elastic-net portfolio has been reported. It is worth pointing out that the elastic-net portfolio does not guarantee a bound similar to (2.4) from the Lasso portfolio. All of the models reviewed so far can be cast as Quadratic Programming (QP), which makes them computationally very attractive. Advances in machine learning algorithms have opened a possibility to make non-QP models practically implementable and a great effort has been made in Ben et al. [3], Perrin and Roncalli [31]. In this paper, we propose a new non-QP model that is closely related to the elastic-net portfolio (2.6) with the squared ℓ_2 norm being replaced by the ℓ_2 norm itself:

$$\min \ f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T V\mathbf{w} + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2, \quad \text{s.t.} \quad \mathbf{1}_N^T\mathbf{w} = 1. \tag{2.7}$$

The nonlinearity comes from the ℓ_2 norm. The model (2.7) is viable based on the principles discussed in Perrin and Roncalli [31]. It is convex. It can be solved by fast algorithms and it enjoys some favourable and unique properties, which will be investigated in the next section.

3 $\ell_{1,2}$-norm Penalized MVP

The main purpose of this section is to justify the use of $\ell_{1,2}$ norm. Our main motivation is the elastic-net regularization, which improves the condition number of the resulting optimization problem and has the scaled soft-thresholding operator as its proximal operator. The latter property leads to fast algorithms, see [37] and [17]. We show that the $\ell_{1,2}$ norm enjoys similar properties, but in a more involved manner. We also prove that it can bound the amount of short positions that the elastic-net regularization fails to guarantee.

3.1 Improving the condition number

For simplicity, let

$$g(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T V\mathbf{w} + \lambda_2\|\mathbf{w}\|_2.$$

It is easy to see that $g(\mathbf{w})$ is differentiable as long as $\mathbf{w} \neq 0$. The Hessian matrix takes the following form:

$$\nabla^2 g(\mathbf{w}) = V + \frac{\lambda_2}{\|\mathbf{w}\|_2^3}\Big(\|\mathbf{w}\|_2^2 I_N - \mathbf{w}\mathbf{w}^T\Big).$$

Let $\widetilde{\mathbf{w}} := \mathbf{w}/\|\mathbf{w}\|_2$ (normalized $\mathbf{w}$) and U be a $N\times(N-1)$ matrix satisfying

$$U^T\mathbf{w} = 0 \qquad \text{and} \qquad U^TU = I_{N-1}.$$

In other words, U is a (column) orthonormal matrix and each column of U is orthogonal to $\mathbf{w}$. Then we have the following eigenvalue-eigenvector decomposition:

$$\nabla^2 g(\mathbf{w}) = V + \frac{\lambda_2}{\|\mathbf{w}\|_2}[U,\ \widetilde{\mathbf{w}}]\begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix}\begin{bmatrix} U^T \\ \widetilde{\mathbf{w}}^T \end{bmatrix}.$$

The Hessian matrix of $g(\cdot)$ is V plus a positive semidefinite matrix. Let $\mathbf{x}$ be an orthonormal eigenvector of $\nabla^2 g(\mathbf{w})$ corresponding to its smallest eigenvalue $\lambda_{\min}(\nabla^2 g(\mathbf{w}))$. We have

$$\begin{aligned}\lambda_{\min}(\nabla^2 g(\mathbf{w})) &= \mathbf{x}^T\nabla^2 g(\mathbf{w})\mathbf{x} \\ &= \mathbf{x}^T V\mathbf{x} + \frac{\lambda_2}{\|\mathbf{w}\|_2}\Big(1 - \langle \mathbf{w}/\|\mathbf{w}\|,\ \mathbf{x}\rangle^2\Big) \\ &\geq \lambda_{\min}(V) + \frac{\lambda_2}{\|\mathbf{w}\|_2}(1-\cos^2(\theta)),\end{aligned}$$

where $\lambda_{\min}(V)$ is the smallest eigenvalue of V and θ is the angle between $\mathbf{x}$ and $\mathbf{w}$. It is obvious to see that if $\mathbf{w}$ is not perfectly correlated with $\mathbf{x}$ (i.e., $cos(\theta) = 1$ or -1), then $\lambda_{\min}(\nabla^2 g(\mathbf{w}))$ is strictly bigger than $\lambda_{\min}(V)$. Consequently, the condition number of the Hessian matrix $\nabla^2 g(\mathbf{w})$ is improved under the condition:

$$\cos^2(\theta) \leq 1 - \frac{1}{\kappa(V)}, \text{ where } \kappa(V) := \frac{\lambda_{\max}(V)}{\lambda_{\min}(V)} \text{ is the condition number of } V. \tag{3.1}$$

We explain it below. We note that

$$\lambda_{\max}(\nabla^2 g(\mathbf{w})) \leq \lambda_{\max}(V) + \frac{\lambda_2}{\|\mathbf{w}\|_2}.$$

Then under the condition (3.1) we obtain

$$\kappa(\nabla^2 g(\mathbf{w})) = \frac{\lambda_{\max}(\nabla^2 g(\mathbf{w}))}{\lambda_{\min}(\nabla^2 g(\mathbf{w}))} \leq \frac{\lambda_{\max}(V) + \frac{\lambda_2}{\|\mathbf{w}\|_2}}{\lambda_{\min}(V) + \frac{\lambda_2}{\|\mathbf{w}\|_2}(1 - \cos^2(\theta))} \leq \frac{\lambda_{\max}(V)}{\lambda_{\min}(V)} = \kappa(V).$$

If the condition number $\kappa(V)$ is large, then the condition (3.1) is likely to hold and hence the condition number of the Hessian matrix $\nabla^2 g(\mathbf{w})$ is likely improved. This is the reason why $\ell_{1,2}$ regularized portfolio is computationally stable.

3.2 Bounding short positions

This part studies the effect of $\ell_{1,2}$ norm on short positions of a portfolio. Let us consider the following model:

$$\mathbf{w}^{[\beta]} \in \arg\min \left\{ \mathcal{E}(\mathbf{w}) + \beta\Big(\alpha_1 \|\mathbf{w}\|_1 + \alpha_2 \|\mathbf{w}\|_2\Big) \;\Big|\; \mathbf{1}_N^T \mathbf{w} = 1 \right\},$$

where $\mathcal{E}(\cdot)$ is a risk/utility function of a portfolio and $\beta > 0$ is a penalty factor common to both ℓ_1 and ℓ_2 norms. Suppose we have two optimal portfolios $\mathbf{w}^{[\beta_1]}$ and $\mathbf{w}^{[\beta_2]}$, $\beta_1 > \beta_2$. We have

$$\begin{aligned}
&\mathcal{E}(\mathbf{w}^{[\beta_1]}) + \beta_1(\alpha_1\|\mathbf{w}^{[\beta_1]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_1]}\|_2) \\
&\leq \mathcal{E}(\mathbf{w}^{[\beta_2]}) + \beta_1(\alpha_1\|\mathbf{w}^{[\beta_2]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_2]}\|_2) \\
&= \mathcal{E}(\mathbf{w}^{[\beta_2]}) + \beta_2(\alpha_1\|\mathbf{w}^{[\beta_2]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_2]}\|_2) + (\beta_1 - \beta_2)(\alpha_1\|\mathbf{w}^{[\beta_2]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_2]}\|_2) \\
&\leq \mathcal{E}(\mathbf{w}^{[\beta_1]}) + \beta_2(\alpha_1\|\mathbf{w}^{[\beta_1]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_1]}\|_2) + (\beta_1 - \beta_2)(\alpha_1\|\mathbf{w}^{[\beta_2]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_2]}\|_2) \\
&= \mathcal{E}(\mathbf{w}^{[\beta_1]}) + \beta_1(\alpha_1\|\mathbf{w}^{[\beta_1]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_1]}\|_2) \\
&\quad + (\beta_1 - \beta_2)\Big(\alpha_1\|\mathbf{w}^{[\beta_2]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_2]}\|_2 - \alpha_1\|\mathbf{w}^{[\beta_1]}\|_1 - \alpha_2\|\mathbf{w}^{[\beta_1]}\|_2\Big).
\end{aligned}$$

Therefore,

$$(\beta_1 - \beta_2)\Big(\alpha_1\|\mathbf{w}^{[\beta_2]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_2]}\|_2 - \alpha_1\|\mathbf{w}^{[\beta_1]}\|_1 - \alpha_2\|\mathbf{w}^{[\beta_1]}\|_2\Big) \geq 0.$$

Since $\beta_1 > \beta_2$, we must have

$$\alpha_1\|\mathbf{w}^{[\beta_2]}\|_1 + \alpha_2\|\mathbf{w}^{[\beta_2]}\|_2 - \alpha_1\|\mathbf{w}^{[\beta_1]}\|_1 - \alpha_2\|\mathbf{w}^{[\beta_1]}\|_2 \geq 0. \tag{3.2}$$

Using the bounds $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{N}\|\mathbf{x}\|_2$ for $\mathbf{x} \in \mathbb{R}^N$, the inequality (3.2) implies

$$(\alpha_1 + \alpha_2)\|\mathbf{w}^{[\beta_2]}\|_1 - \left(\alpha_1 + \frac{\alpha_2}{\sqrt{N}}\right)\|\mathbf{w}^{[\beta_1]}\|_1 \geq 0,$$

which further yields

$$\|\mathbf{w}^{[\beta_1]}\|_1 \leq \frac{\alpha_1 + \alpha_2}{\alpha_1 + \alpha_2/\sqrt{N}}\|\mathbf{w}^{[\beta_2]}\|_1,$$

or equivalently

$$2\sum_{i:\, w_i^{[\beta_1]}<0} \left|w_i^{[\beta_1]}\right| \leq \frac{\alpha_1 + \alpha_2}{\alpha_1 + \alpha_2/\sqrt{N}}\|\mathbf{w}^{[\beta_2]}\|_1 - 1. \tag{3.3}$$

We note that (3.3) generalizes the bound of Brodie et al. [5] on short positions. In particular, when $\alpha_2 = 0$, (3.3) becomes [5, Eq. (7)]. One important observation of Brodie et al. [5] is that when $\mathbf{w}^{[\beta_2]} \geq 0$ (no short positions), then $\mathbf{w}^{[\beta_1]}$ must not have any short positions for any $\beta_1 > \beta_2$. This implies that when the penalty parameter α_1 is above a certain threshold, the ℓ_1 norm penalty all produces the same optimal portfolio. This property does not theoretically hold for the $\ell_{1,2}$-norm portfolio, but the shorting positions can be quantified. We illustrate this point below.

Let $p\%$ denote the percentage of short positions within the portfolio $\mathbf{w}$:

$$\|\mathbf{w}\|_1 = 1 + 2\sum_{i:\, w_i<0} |w_i| = 1 + 2p\%.$$

In particular, $p_1\%$ and $p_2\%$ are the respective percentages of short positions corresponding to the portfolios $\mathbf{w}^{[\beta_1]}$ and $\mathbf{w}^{[\beta_2]}$. Suppose $\alpha_1 = m\alpha_2$. It follows from (3.3) that

$$2\sum_{i:\, w_i^{[\beta_1]}<0} \left|w_i^{[\beta_1]}\right| \leq \frac{(m+1)\sqrt{N}}{m\sqrt{N}+1}\|\mathbf{w}^{[\beta_2]}\|_1 - 1.$$

Simplifying the above inequality leads to

$$\begin{aligned} p_1\% = \sum_{i:\, w_i^{[\beta_1]}<0} \left|w_i^{[\beta_1]}\right| &\leq \frac{1}{2}\frac{\sqrt{N}-1}{m\sqrt{N}+1} + \left(1 + \frac{\sqrt{N}-1}{m\sqrt{N}+1}\right) \times p_2\% \\ &< \frac{1}{2m} + \left(1 + \frac{1}{m}\right) p_2\%. \end{aligned}$$

If $m = 5$, the total amount of short positions for the portfolio $\mathbf{w}^{[\beta_1]}$ is roughly no more 10% than that of $\mathbf{w}^{[\beta_2]}$. If $m = 50$, the extra amount can be bounded by 1%. In practice, this bounding inequality even becomes tighter. For SP500 data (see the numerical part), Figure 3.2 demonstrates the bounding relationship between $p_1\%$ and $p_2\%$ corresponding to two cases where $m = 5$ and 50 with β_2 varying from 10^{-2} to 10^2. It is consistently observed that $p_1\% \leq p_2\%$.

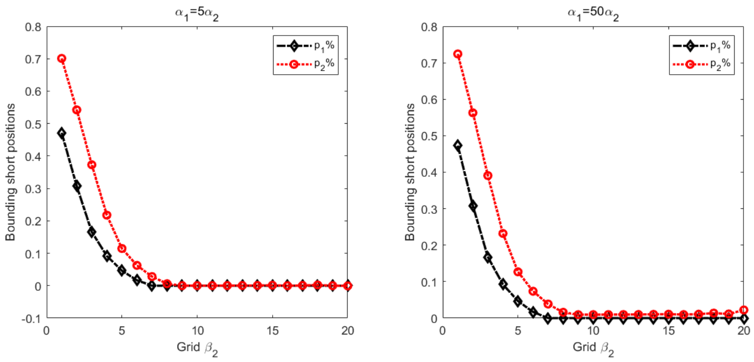

Fig. 1 The boundary of short positions for SP500 (data from July 27, 2015 to Dec. 18, 2017).

3.3 Proximal operator of $\ell_{1,2}$ norm

We first define the vectorized soft-thresholding operator (1.1) as follows. For a given positive vector $\mathbf{c} \in \mathbb{R}^N$ and a vector $\mathbf{b} \in \mathbb{R}^N$, we defined $\mathcal{S}_{\mathbf{c}}(\mathbf{b}) \in \mathbb{R}^N$ by

$$\Big(\mathcal{S}_{\mathbf{c}}(\mathbf{b})\Big)_i := \mathcal{S}_{c_i}(|b_i|), \quad i = 1, \dots, N.$$

It is easy to verify that for a given $\lambda > 0$,

$$\mathcal{S}_{\lambda\mathbf{c}}(\mathbf{b}) = \lambda \mathcal{S}_{\mathbf{c}}(\mathbf{b}/\lambda).$$

Proposition 1 *Consider the problem:*

$$\min_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{x}-\mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_{1,\mathbf{c}} + \gamma\|\mathbf{x}\|_2,$$

where $\mathbf{b} \in \mathbb{R}^N$, the positive weight vector $\mathbf{c} \in \mathbb{R}^N$, and $\alpha, \gamma > 0$ are given. The optimal solution is given by

$$\mathbf{x}^* = \mathcal{S}_{(\alpha,\gamma,\mathbf{c})}(\mathbf{b}) := \begin{cases} \left(1 - \frac{\gamma}{\|\mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b})\|_2}\right)_+ \mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b}), & \text{if } \mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b}) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

In the particular case when $\mathbf{c} = \mathbf{1}_N$ (uniformly weighted ℓ_1 norm), the solution becomes

$$\mathbf{x}^* = \mathcal{S}_{(\alpha,\gamma)}(\mathbf{b}) := \begin{cases} \left(1 - \frac{\gamma}{\|\mathcal{S}_{\alpha}(\mathbf{b})\|_2}\right)_+ \mathcal{S}_{\alpha}(\mathbf{b}), & \text{if } \mathcal{S}_{\alpha}(\mathbf{b}) \neq 0 \\ 0, & \text{otherwise.} \end{cases}, \tag{3.5}$$

where

$$\mathcal{S}_\alpha(\mathbf{b}) := (\mathcal{S}_\alpha(b_1), \ldots, \mathcal{S}_\alpha(b_N)).$$

Proof. Since the problem is strongly convex, there is a unique solution, denoted as $\mathbf{x}^*$. A necessary and sufficient condition for $\mathbf{x}^*$ is that there exist $\mathbf{u} \in \partial\|\mathbf{x}^*\|_1$ and $\mathbf{v} \in \partial\|\mathbf{x}^*\|_2$ such that

$$\mathbf{x}^* - \mathbf{b} + \alpha(\mathbf{c} \circ \mathbf{u}) + \gamma\mathbf{v} = 0. \tag{3.6}$$

The subgradients of ℓ_1 and ℓ_2-norm have simple representations as follows:

$$\partial\|\mathbf{x}^*\|_2 = \begin{cases} \{\mathbf{x}^*/\|\mathbf{x}^*\|_2\} & \text{if } \mathbf{x}^* \neq 0 \\ \left\{\mathbf{y} \in \mathbb{R}^N \,|\|\mathbf{y}\|_2 \leq 1\right\} & \text{otherwise} \end{cases}$$

and

$$\partial\|\mathbf{x}^*\|_1 = \left\{ \mathbf{y} \in \mathbb{R}^N \;\middle|\; \begin{matrix} y_i = \operatorname{sgn}(x_i^*) \text{ if } x_i^* \neq 0 \\ y_i \in [-1,1] \quad \text{otherwise.} \end{matrix} \right\}$$

In order to compute $\mathbf{x}^*$, we consider the following two cases.

Case 1. $\mathbf{x}^* \neq 0$. The condition (3.6) becomes

$$\left(1 + \frac{\gamma}{\|\mathbf{x}^*\|_2}\right)\mathbf{x}^* = \mathbf{b} - \alpha(\mathbf{c} \circ \mathbf{u}). \tag{3.7}$$

Writing (3.7) componentwise, we have

$$\left(1 + \frac{\gamma}{\|\mathbf{x}^*\|_2}\right)x_i^* = b_i - \alpha c_i u_i, \qquad i = 1, \ldots, N. \tag{3.8}$$

Consider two sub-cases.

Subcase 1.1. We consider $|b_j| \leq \alpha c_j$. The choice $x_j^* = 0$ and $u_j = b_j/(\alpha c_j)$ satisfies (3.8) for $i = j$. Moreover, $|u_j| \leq 1$ and $u_j \in \partial|x_j^*|$. In this case, we have

$$\left(1 + \frac{\gamma}{\|\mathbf{x}^*\|_2}\right)x_j^* = 0 = \underbrace{\Big(|b_j| - \alpha c_j\Big)_+}_{=0} \operatorname{sgn}(b_j) = 0. \tag{3.9}$$

Subcase 1.2. We consider the remaining case $|b_j| > \alpha c_j$. We choose $u_j = \operatorname{sgn}(b_j)$. Based on (3.8), we must have

$$\left(1 + \frac{\gamma}{\|\mathbf{x}^*\|_2}\right)x_j^* = \underbrace{\Big(|b_j| - \alpha c_j\Big)}_{>0}\operatorname{sgn}(b_j) = \Big(|b_j| - \alpha c_j\Big)_+\operatorname{sgn}(b_j). \tag{3.10}$$

Furthermore, $x_j^* \neq 0$ and it has the same sign as b_j. Hence $u_j \in \partial|x_j^*|$.

We need to show those choices of u_j and $x_j^* = 0$ for the subcase 1.1 are consistent with the KKT condition (3.8). Putting (3.9) and (3.10) in the vector form, we have

$$\left(1 + \frac{\gamma}{\|\mathbf{x}^*\|_2}\right)\mathbf{x}^* = \Big(|\mathbf{b}| - \alpha\mathbf{c}\Big)_+\operatorname{sgn}(\mathbf{b}) = \mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b}). \tag{3.11}$$

Computing the ℓ_2 norm on both sides yields

$$\|\mathbf{x}^*\|_2 = \|\mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b})\|_2 - \gamma.$$

Substituting it back to (3.11) we obtain

$$\mathbf{x}^* = \left(1 - \frac{\gamma}{\|\mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b})\|_2}\right) \mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b}). \tag{3.12}$$

Case 2. $\mathbf{x}^* = 0$. In this case, the KKT condition (3.6) becomes

$$\gamma\mathbf{v} = \mathbf{b} - \alpha(\mathbf{c} \circ \mathbf{u}),$$

which implies

$$1 \ge \|\mathbf{v}\|_2 = \frac{\|\mathbf{b} - \alpha(\mathbf{c} \circ \mathbf{u})\|_2}{\gamma} \ge \min_{\|\mathbf{y}\| \le 1} \frac{\|\mathbf{b} - \alpha(\mathbf{c} \circ \mathbf{y})\|_2}{\gamma} = \frac{\|\mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b})\|_2}{\gamma}$$

Hence, when $\mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b}) \neq 0$, a necessary condition for $\mathbf{x}^* = 0$ is

$$1 - \frac{\gamma}{\|\mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b})\|_2} \le 0$$

The solution takes the following form

$$0 = \mathbf{x}^* = \begin{cases} \left(1 - \frac{\gamma}{\|\mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b})\|_2}\right)_+ \mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b}) & \text{if } \mathcal{S}_{\alpha\mathbf{c}}(\mathbf{b}) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

The representation in (3.12) and (3.13) gives a unifying representation (3.4) and (3.5) is just a simple consequence of (3.4). □

Remark 1 **(On weighted ℓ_2 norm)** We only considered the weighted ℓ_1-norm $\|\mathbf{w}\|_{1,\mathbf{c}}$, and deliberately left out the weighted form for the ℓ_2 norm. The reason is that the closed-form solution proved in Prop. 1 is not valid any more. This is because ℓ_2 norm does not have the separability property that allows it be reformulated as the sum of functions of individual weights w_i.

Remark 2 **(On the double shrinkage operators)** It is interesting to note that the proximal operator for the $\ell_{1,2}$ norm has double shrinkage operators, one is from the ℓ_1 norm and the other is from the ℓ_2 norm. The product of them in fact may enlarge the shrinking region than that from the ℓ_1 norm (the soft-thresholding operator). Let us illustrate this point via a small example. Consider the following problem in two dimensions $\mathbf{x} = (x_1, x_2)$ with $\alpha = \gamma = 1$:

$$\min f(x_1, x_2) = \frac{1}{2}\|\mathbf{x} - \mathbf{b}\|_2^2 + \|\mathbf{x}\|_1 + \|\mathbf{x}\|_2, \qquad \mathbf{b} = (b_1, 1 + \sqrt{3}/2). \tag{3.14}$$

We want to study the behaviour of the solution when b_1 varies. It is easy to calculate that

$$\mathcal{S}_1(\mathbf{b}) = \begin{pmatrix} \mathcal{S}_1(b_1) = (|b_1| - 1)_+ \mathrm{sgn}(b_1) \\ \mathcal{S}_1(b_2) = (|b_2| - 1)_+ \mathrm{sgn}(b_2) = \sqrt{3}/2 \end{pmatrix}.$$

Hence, $\|\mathcal{S}_1(\mathbf{b})\|_2 = \sqrt{(|b_1|-1)_+^2 + 3/4} > 0$. For simplicity, we let $(y_1, y_2) = \mathcal{S}_{(1,1)}(\mathbf{b})$. According to the solution formula (3.5), we have

$$y_1 = \left(1 - \frac{1}{\sqrt{(|b_1|-1)_+^2 + 3/4}}\right)_+ \Big(|b_1| - 1)_+ \mathrm{sgn}(b_1)$$

$$y_2 = \frac{\sqrt{3}}{2}\left(1 - \frac{1}{\sqrt{(|b_1|-1)_+^2 + 3/4}}\right)_+ .$$

Without the ℓ_2 norm, the above problem becomes the ℓ_1-norm regularized problem (i.e., the Lasso problem). The corresponding solution is

$$y_1 = (|b_1|-1)_+\mathrm{sgn}(b_1), \qquad y_2 = (|b_2|-1)_+\mathrm{sgn}(b_2) = \sqrt{3}/2.$$

When the ℓ_2 norm is replaced by its squared $\|x\|_2^2$ (i.e., the elastic-net problem), the corresponding solution is just a scaling of the Lasso solution:

$$y_1 = \frac{1}{3}(|b_1|-1)_+\mathrm{sgn}(b_1), \qquad y_2 = \frac{1}{3}(|b_2|-1)_+\mathrm{sgn}(b_2) = \sqrt{3}/6.$$

Those solutions each as a function of b_1 are plotted in Fig. 2. The effect of $\ell_{1,2}$ norm on the sparsity of the resulting solutions can be clearly appreciated. It renders a larger region where both y_1 and y_2 may become zero. For both the Lasso and elastic-net regularization, the region for y_1 to be zero is $-1 \le b_1 \le 1$, while it is $-3/2 \le b_1 \le 3/2$ for the $\ell_{1,2}$-norm regularization. For the y_2 part, it remains positive for both the Lasso and elastic-net regularization no matter how b_1 changes, while for the $\ell_{1,2}$-norm regularization, y_2 becomes zero when $-3/2 \le b_1 \le -1/2$ or $-1/2 \le b_1 \le 3/2$.

Remark 3 On Zhang-Jiang-Luo formula [38]. Zhang et al. also studied the proximal operator for the $\ell_{1,2}$-norm and proposed a computational formula, which is similar to ours. But there is an importance difference between the two. We explain it below. Let us consider the simple case where $\mathbf{c} = \mathbf{1}_N$. We apply the Zhang-Jiang-Luo formula [38, Eq. (17)] to the proximal problem in Prop. 1 (i.e., in [38], let $f_2(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{b}\|_2^2$, $\mathbf{w}_J = \gamma$ and $\alpha = 1$), we get

$$\mathbf{x}^* = \begin{cases} 0, & \text{if } \|\beta(\mathbf{b})\|_2 \le \gamma \\ \left(1 - \frac{\gamma}{\|\beta(\mathbf{b})\|_2}\right)\beta(\mathbf{b}), & \text{otherwise,} \end{cases} \tag{3.15}$$

where the vector $\beta(\mathbf{b})$ is defined as in [38, Eq. 12] (their λ is our α):

$$\beta_j(\mathbf{b}) := \begin{cases} 0, & \text{if } \mathbf{b} \in \gamma\mathcal{B} + \alpha\mathcal{B}_\infty \quad \text{Case 1} \\ 0, & \text{if } |b_j| \le \gamma \quad \text{Case 2} \\ \Big(b_j - \gamma\mathrm{sgn}(b_j)\Big), & \text{otherwise,} \quad \text{Case 3} \end{cases}$$

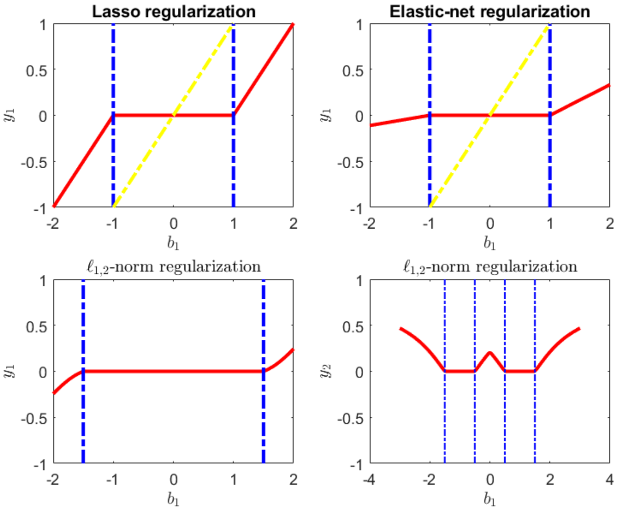

Fig. 2 Behaviour of three shrinkage operators for Problem (3.14): Lasso, Elastic-net and $\ell_{1,2}$. The yellow dotted lines are for the diagonal $x = y$, which shows that the elastic shrinkage operator is just a scaling of the Lasso shrinkage operator (top panel). The $\ell_{1,2}$ shrinkage operator has a larger region where y_1 takes the zero value than that by Lasso and Elastic-net. The second component y_2 also has a chance to become zero, while in the case of Lasso and Elastic-net, the y_2 part remains positive.

and

$$\mathcal{B} := \{\mathbf{s} \in \mathbb{R}^n \mid \|\mathbf{s}\|_2 \le 1\} \quad \text{and} \quad \mathcal{B}_\infty := \{\mathbf{s} \in \mathbb{R}^n \mid \|\mathbf{s}\|_\infty \le 1\}\,.$$

As noted in the paragraph below [38, Eq. 12], for Case 2 and Case 3, $\beta(\mathbf{b})$ is the soft-thresholding operator. For those two cases, the two formulae (3.5) and (3.15) are same. However, there is Case 1 in Zhang-Jiang-Luo formula and the proof in [38] makes use of this particular case. Moreover, checking Case 1 needs solving a quadratic programming problem (see [38, Eq. 15]). This is the key difference. Our formula only needs the soft-thresholding operator and its proof is based on the classical KKT conditions.

4 Optimization methods

In this section, we derive a proximal augmented Lagrange method to efficiently solve $\ell_{1,2}$ norm regularized MVP. We consider the equivalent reformulation of (2.7):

$$\min f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T V\mathbf{w} + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2, \qquad \text{s.t.} \quad \frac{1}{\sqrt{N}}\mathbf{1}_N^T\mathbf{w} = \frac{1}{\sqrt{N}}. \quad (4.1)$$

Note that we scaled the constraint by a factor of $1/\sqrt{N}$. This is because the matrix $\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ has 1 as its largest eigenvalue and this fact will be used in our algorithm. The augmented Lagrange function for the problem (4.1) is

$$\mathcal{L}_c(\mathbf{w},\eta) := f(\mathbf{w}) - \frac{\eta}{\sqrt{N}}(\mathbf{1}_N^T\mathbf{w}-1) + \frac{c}{2N}\|\mathbf{1}_N^T\mathbf{w}-1\|_2^2,$$

where $\eta \in \mathbb{R}$ is the Lagrange multiplier and $c > 0$ is a penalty parameter. We let $\lambda_{\max}$ be the largest eigenvalue of V. We define the positive semidefinite matrix $\mathcal{Q}$ by

$$\mathcal{Q} := \Big(1+\lambda_{\max}/c\Big)I_N - \Big(\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T + \frac{1}{c}V\Big).$$

The vector norm defined by $\mathcal{Q}$ is $\|\mathbf{w}\|_{\mathcal{Q}}^2 := \mathbf{w}^T\mathcal{Q}\mathbf{w}$ for $\mathbf{w} \in \mathbb{R}^N$. We apply the proximal augmented Lagrange method (pALM) studied in Fazel et al. [13] to our problem.

Algorithm 1 pALM: Proximal Augmented Lagrange Method

```
Set the problem parameter: $\lambda_1 > 0$ and $\lambda_2 > 0$. Initialize $\mathbf{w}^0$, $\eta_0 > 0$ and choose the
    penalty parameter $c > 0$ and a constant $\nu \in (0,2)$. set the iteration index $k := 0$.
Compute
    $$\mathbf{w}^{k+1} = \arg\min\Big\{\mathcal{L}_c(\mathbf{w},\eta_k) + \frac{c}{2}\|\mathbf{w}-\mathbf{w}^k\|_{\mathcal{Q}}^2\Big\}. \tag{4.2}$$
Update $\eta_k$ by
    $$\eta_{k+1} = \eta_k - \nu c\Big(\mathbf{1}_N^T\mathbf{w}^{k+1}-1\Big).$$
Continue the update until a stopping criterion is met.
```

Obviously, the main computational task of Alg. 1 is in (4.2), which we show has a closed-form solution. By making use of the definition of $\mathcal{Q}$ matrix, we arrive at (some computational details were omitted):

$$\begin{aligned}
\mathbf{w}^{k+1} &= \arg\min \frac{1}{2}\mathbf{w}^T\Big[V+\frac{c}{N}\mathbf{1}_N\mathbf{1}_N^T\Big]\mathbf{w} - \frac{1}{\sqrt{N}}\eta_k(\mathbf{1}_N^T\mathbf{w}-1) - \frac{c}{N}\mathbf{1}_N^T\mathbf{w} \qquad (4.3)\\
&\quad + \frac{c}{2}\mathbf{w}^TQ\mathbf{w} - c(Q\mathbf{w}^k)^T\mathbf{w} + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2\\
&= \arg\min \frac{1}{2}(\lambda_{\max}+c)\|\mathbf{w}\|_2^2 - (\lambda_{\max}+c)\langle\mathbf{w},\ \mathbf{w}^k\rangle - (\eta_k/\sqrt{N}+c/N)\langle\mathbf{1}_N,\ \mathbf{w}\rangle\\
&\quad + \langle\mathbf{w},\ (V+(c/N)(\mathbf{1}_N\mathbf{1}_N^T))\mathbf{w}^k\rangle + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2\\
&= \arg\min \frac{1}{2}\|\mathbf{w}\|_2^2 - \langle\mathbf{w},\ \mathbf{b}^k\rangle + \frac{\lambda_1}{\lambda_{\max}+c}\|\mathbf{w}\|_1 + \frac{\lambda_2}{\lambda_{\max}+c}\|\mathbf{w}\|_2\\
&= \mathcal{S}_{(\alpha,\gamma)}(\mathbf{b}^k),
\end{aligned} \tag{4.4}$$

where we defined

$$\mathbf{b}^k := \mathbf{w}^k - \frac{c}{\lambda_{\max}+c}\left[\left(\frac{V}{c}+\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)\mathbf{w}^k - \left(\frac{1}{N}+\frac{\eta_k}{c\sqrt{N}}\right)\mathbf{1}_N\right]$$

and

$$\alpha := \frac{\lambda_1}{\lambda_{\max}+c}, \qquad \gamma := \frac{\lambda_2}{\lambda_{\max}+c}.$$

The last equation (4.4) used the formula (3.5) in Prop. 1.

By following the convergence analysis in Fazel et al. [13, Appendix B], the sequence $\{(\mathbf{w}^k, \eta_k)\}$ generated by pALM converges with a limit $\mathbf{w}^*$ of $\{\mathbf{w}^k\}$ being the optimal solution of (4.1). Hence, a natural termination criterion for pALM is when the relative change among the iterates falls under a small threshold:

$$\mathrm{Tol}_k := \frac{\|\mathbf{w}^k - \mathbf{w}^{k-1}\|_2 + |\eta_k - \eta_{k-1}|}{\|\mathbf{w}^{k-1}\|_2 + |\eta_{k-1}|} \le \epsilon,$$

where $\epsilon > 0$ is an error allowed (e.g., 10^{-3}).

Remark 4 (On the choice of the steplength ν.) According to the general convergence result [13, Thm. B.1(d)], Alg. 1 converges as long as $\nu < 2$. It is well received in optimization that longer stepsize seems to make convergence faster. We wonder whether a smaller stepsize may lead to significant benefit for portfolio problems. We tested two choices of ν: $\nu = 1.618 \approx (\sqrt{5}+1)/2$ and $\nu = 1.999$ on the six data sets listed in the next section. Our general observation is that there is not much difference between the two cases in terms of the convergence speed, solution quality and out-of-sample performances. For example, Table 1 lists the proportions of zero positions obtained for the six data sets. We observe that the difference is insignificant though the choice of $\nu = 1.618$ led to a slightly higher proportions in all cases. The out-of-sample performance indicators (their definitions are in the next sections) only differ in the third decimal points as reported in Table 2 and hence did not show any meaningful difference. Therefore, in our numerical test in the next section, we used $\nu = 1.999$.

Table 1 Proportion of zero positions (%) of optimal L12 portfolio for two scenarios: $\nu = 1.999$ and $\nu = 1.618$. The test are on six data sets (weekly data from Jan. 5, 2015 to Oct. 26, 2020, $\tau = 120$; monthly data from Nov., 1978 to June, 2020; $\tau = 240$.)

	DJIA	NASDAQ	SP500	Russell2000	Russell3000	FF100
ν	N=29	N=95	N=336	N=1340	N=2166	N=100
1.999	14.06	46.11	42.93	79.20	81.18	71.82
1.618	14.15	46.22	43.15	79.86	81.60	72.66

Table 2 Out-of-sample performance indicators: variance ($\widehat{\sigma}^2$) ($(\%)^2$), Sharp ratio ($\widehat{SR}$), turnover ($TURN$) and the average short positions (ASP) of portfolio strategies by setting different value of ν (weekly data from June 22, 2015 to Nov. 18, 2019, $\tau = 60$; monthly data from Oct., 2009 to Oct., 2019; $\tau = 72$.).

		L12			
Data	ν	$\widehat{\sigma}^2$	$\widehat{SR}$	$TURN$	ASP
DJIA	1.999	1.9838	0.2360	0.0565	0
	1.618	1.9824	0.2360	0.0566	0
NASDAQ	1.999	2.2587	0.1982	0.1061	0
	1.618	2.2585	0.1977	0.1055	0
SP500	1.999	1.5219	0.1555	0.1239	0.0001
	1.618	1.5221	0.1547	0.1192	0.0001
Russell2000	1.999	1.8567	0.3085	0.1560	0.0006
	1.618	1.8345	0.3121	0.1563	0.0005
Russell3000	1.999	1.4684	0.3115	0.1489	0.0007
	1.618	1.4604	0.3134	0.1450	0.0006
FF100	1.999	10.1471	0.2977	0.1226	0.0210
	1.618	10.1133	0.2968	0.1268	0.0212

5 Numerical Results

This section reports extensive numerical experiments that show the advantages of the $\ell_{1,2}$ regularization over some existing models. In Section 5.1, we collect six real data sets, explain some existing models to be compared with, and describe the performance measures to be used. In Section 5.2, we demonstrate that the $\ell_{1,2}$ regularization leads to sparse and stable portfolios. In Section 5.3, we compare $\ell_{1,2}$ portfolio against 8 popular portfolios in terms of out-of-sample performance measures. Finally, in Section 5.4, we demonstrate the speed of the proposed pALM algorithm. A general message is that the proposed $\ell_{1,2}$ regularized portfolio model can yield stable and sparse portfolio with strong out-of-sample performance and can be efficiently solved by a purposely developed algorithm. All of our computations were conducted in Matlab R2019a environment, on a PC with Intel(R) Core(TM) i5-7200U CPU (2.50GHz, 4 CPUs) and 4G RAM processors.

5.1 Models of comparison, data and performance measures

(a) 8 **MVP models compared.** We compare the out-of-sample performance of 8 minimum-variance portfolio models across six real data sets of weekly and monthly returns. Those models are well studied and we divided them into three groups, which are summarized in Table 3. The first group includes norm-regularized portfolio, namely, the ℓ_2-regularized MVP corresponding to $\ell_{1,2}$-regularization MVP when $\lambda_1 = 0$ in (2.7), the ℓ_1-regularized MVP of Brodie et al. [5] and the elastic-net regularized MVP of Yen and Yen[37]. The second group includes the MVP with/without the short-sale constraints considered by Jagannathan and Ma [18], and the equally-weighted portfolio of DeMiguel et al. [10]. The last group consists of two portfolios that used the shrinkage technique to estimate the covariance matrix. The first is from Olivier and Wolf [29] where the covariance matrix is a combination of the sample covariance matrix and the identity matrix. The second is from Olivier and Wolf [30], where the covariance matrix is a mixture of the sample covariance matrix and the single-factor model.

Table 3 List of Portfolio strategies Considered.

Group	Model	*Abbr.*	*Refer.*
(1)	Norm-regular Portfolio		
	MVP with the $\ell_{1,2}$ regularization	L12	this paper
	MVP with ℓ_2-regularization	L2	this paper
	MVP with ℓ_1-regularization	L1	Brodie et al. [5]
	MVP with the Elastic Net regularization	EN	Yen and Yen[37]
(2)	Benchmarks Portfolio		
	MVP with shortsale-constrained	SC	Jagannathan and Ma [18]
	MVP with shortsale-unconstrained	SU	Jagannathan and Ma [18]
	Equally-weighted (1/N) portfolio	EW	DeMiguel et al. [10]
(3)	Shrinkage of Covariance		
	Mixture of sample covariance and identity matrix	SCID	Olivier and Wolf [29]
	Mixture of sample covariance and 1-factor matrix	SC1F	Olivier and Wolf [30]

(b) 6 **data sets tested**. Table 4 lists the six data sets: DJIA30 [21], NASDAQ100 [7], SP500 [8], Russell2000 [11], Russell3000 [34] and FF100 [37]. All data were obtained from Yahoo finance[1] and Ken French's website[2]. In all cases, we removed those assets that have missing values.

Table 4 Information of the six real data sets.

#	Dataset	Stocks	Time period	Source	Frequency
1	DJIA30	29	01/04/2015-30/10/2020	Yahoo finance	Weekly
2	NASDAQ100	95	01/04/2015-30/10/2020	Yahoo finance	Weekly
3	SP500	336	01/04/2015-30/10/2020	Yahoo finance	Weekly
4	Russell2000	1340	01/04/2015-30/10/2020	Yahoo finance	Weekly
5	Russell3000	2166	01/04/2015-30/10/2020	Yahoo finance	Weekly
6	FF100	100	11/1978-06/2020	K.French	Monthly

(c) Measuring the out-of-sample performance and its setup. We largely follow the procedures in [5], [9] and [10] to conduct our comparison.

Let T be the length of a data set and τ be the window length (e.g., $\tau = 120$) used to construct the optimal portfolio by a model. In each period $(t + 1)$,

[1] https://finance.yahoo.com/

[2] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

$t = \tau, ..., T-1$, we compute different portfolio over the previous τ periods. We then compute the *out-of-sample* return in the $(t+1)th$ period based on the obtained portfolio. We repeat this procedure until we reached the end of the data set. In this way, we will get a series of $(T-\tau)$ portfolio vectors for each model listed in Table 3. To make it precise, let $\mathbf{w}_t^s$ be the optimal portfolio obtained by the portfolio strategy s over the date from $t-\tau+1, \ldots, t$. The *out-of-sample* return in the $t+1$ period is computed as $\mathbf{r}_{t+1}^s = \mathbf{w}_t^{s\intercal}\mathbf{r}_{t+1}$, where $\mathbf{r}_{t+1}$ is the return in the $(t+1)$th period. Thus, we obtain a time series of $(T-\tau-1)$ periods out-of-sample returns for all strategies.

We evaluate the out-of-sample performance of each portfolio strategy by using four quantities: (i) the out-of-sample portfolio variance ($\widehat{\sigma}^2$), (ii) the out-of-sample portfolio Sharpe ratio ($\widehat{SR}$), (iii) portfolio turnover ($TURN$) and (iv) the average short positions (ASP).

$$(\widehat{\sigma}^s)^2 = \frac{1}{T-\tau-1}\sum_{t=\tau}^{T-1}(\mathbf{w}_t^{s\intercal}\mathbf{r}_{t+1} - \widehat{\mu}^s)^2, \tag{5.1}$$

$$\text{where} \quad \widehat{\mu}^s = \frac{1}{T-\tau}\sum_{t=\tau}^{T-1}\mathbf{w}_t^{s\intercal}\mathbf{r}_{t+1}, \quad \widehat{SR}^s = \frac{\widehat{\mu}}{\widehat{\sigma}^s}. \tag{5.2}$$

We let $w_{i,t}^s$ be the ith component of $\mathbf{w}_t^s$ (the holding on the asset i). Define

$$TURN = \frac{1}{T-\tau-1}\sum_{t=\tau}^{T-1}\sum_{i=1}^{n}\left(|w_{i,t+1}^s - w_{i,t^+}^s|\right), \tag{5.3}$$

where

$$w_{i,t^+}^s = \frac{w_{i,t}^s(1+r_{i,t+1})}{\sum_{j=1}^N(1+r_{i,t+1})}.$$

The average short positions over the $(T-\tau)$ periods is defined as

$$ASP = \frac{1}{T-\tau-1}\sum_{t=\tau}^{T-1}\frac{\|\mathbf{w}_t^s\|_1 - 1}{2}. \tag{5.4}$$

We further consider some quantities studied in [37] on the profiles of portfolio weights: PAP represents the proportion of active positions and PSP is the proportion of shortsale positions respectively defined as

$$PAP_t = \frac{|S_t^+ \bigcup S_t^-|}{N}, \quad PSP_t = \frac{|S_t^-|}{N}, \tag{5.5}$$

where $S_t^+ = \{i : w_{i,t} > 0\}$ and $S_t^- = \{i : w_{i,t} < 0\}$.

5.2 Sparse portfolio by the $\ell_{1,2}$ regularization

This section is to understand the behaviour of $\ell_{1,2}$ MVP in terms of sparse portfolio generated. Our general conclusion is that not only it generates sparse portfolios, but also the sparsity can be controlled by adjusting its parameters. Let us use the data DJIA to demonstrate this feature.

Figure 3 shows the sparsity of portfolio weights, proportion of active positions (PAP) and proportion of shortsale positions (PSP). The penalty parameter were set as follows. $\lambda_1 = \lambda\alpha$, $\lambda_2 = \lambda\beta$ where $\alpha + \beta = 1$ and λ varies from 10^{-2} to 10^{1}, $c = 1$ and $tol = 10^{-4}$. When $\alpha = 1$, it reduces to the ℓ_1-regularized portfolio. When $\alpha = 0$, the value of proportion of active positions (PAP) is approximately 1. When $\alpha > 0$, PAP declines as λ increases. For example, when $\alpha = 0.5$ and λ is large, the number of active constituents increases and few portfolio has negative weights.

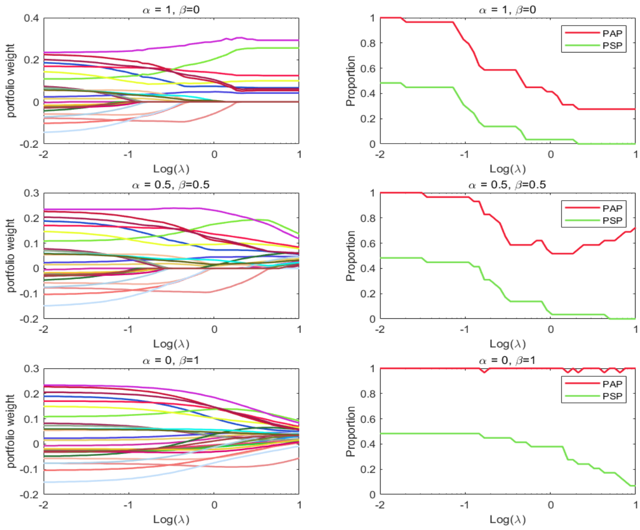

Fig. 3 L12 portfolio weights, PAP and PSP (data from Jul. 02, 2018 to Oct. 19, 2020).

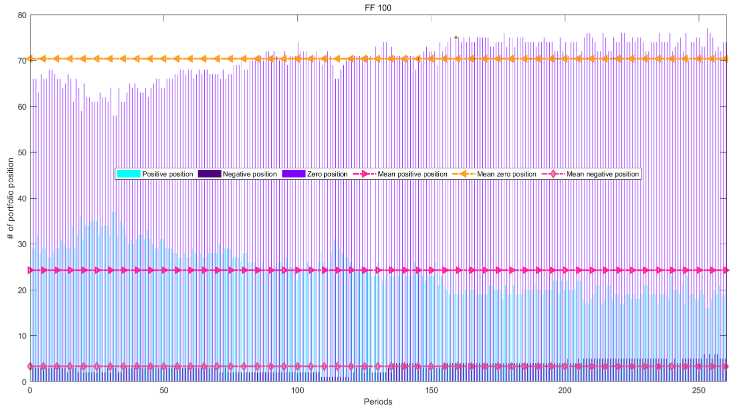

Fig. 4 Positive, negative and zero positions in optimal L12 portfolios (data from Nov., 1978 to Jun., 2020, $\tau = 240$).

Fig. 4 and Fig. 5 report the number of positive, negative and zero positions in optimal portfolios obtained from L12 model by using FF100 and Russell3000 data sets. We can see that in each data set the number of zero positions is at least three times as many the number of positive positions in L12 portfolio. Specifically, there are about 70.34% and 81.18% zero positions in FF100 and Russell3000, respectively. The setting of parameters are $\lambda_1 = 3$, $\lambda_2 = 3$, $c = 10$ and $tol = 10^{-4}$.

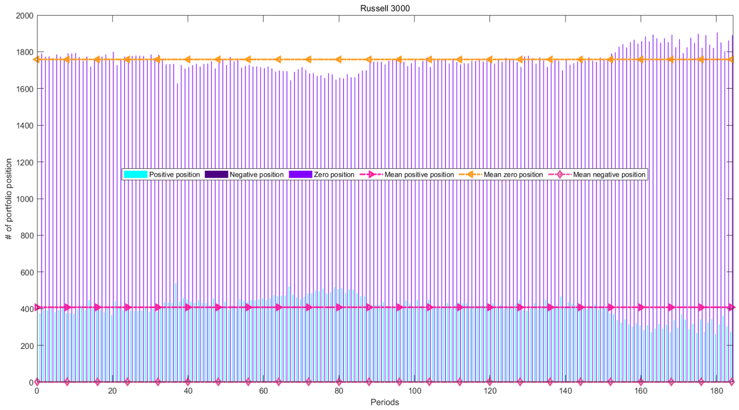

Fig. 5 Positive, negative and zero positions in optimal L12 portfolios (data from Jan. 05, 2015 to Oct. 26, 2020, $\tau = 120$).

We further conducted similar experiments on randomly selected data from all 6 data sets listed in Table 4. The number of positive, negative and zero positions, as well as the proportion of active position (PAP) and the proportion of shortsale positions (PSP) in the optimal L12 portfolio are reported in Table 5. It can be clearly observed that the L12 portfolio selection model can yield sparse portfolios for a wide range of parameter values. For instance, for the data of FF100, we chose $\lambda_1 = 3$, $\lambda_2 = 3$, $c = 10$ and $tol = 10^{-4}$. For DJIA, NASDAQ, SP and Russell, we chose $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $c = 1$ and $tol = 10^{-4}$. They all resulted in reasonably sparse portfolios.

Table 5 Sparse portfolio (randomly choose DJIA data from May 01, 2017 to Aug. 19, 2019; NASDAQ data from Apr. 13, 2015 to Nov. 18, 2019; SP500 data from Nov. 28, 2016 to Dec. 23, 2019; Russell data from Nov. 28, 2016 to Dec. 03, 2018; FF100 data from June, 2009 to June, 2019.).

	Posi.	*Nega.*	*Zer.*	*PAP*(%)	*PSP*(%)
DJIA	15	3	11	62.07	10.34
NASDAQ	31	18	46	51.58	18.95
SP500	88	27	221	34.23	8.04
Russell2000	237	49	1054	21.34	3.66
Russell3000	333	49	1784	17.64	2.26
FF100	19	1	78	20.41	1.02

5.3 Out-of-sample performance comparison

The out-of-sample Sharpe ratio considers return and risk at the same time, so it is a comprehensive measure of portfolio performance. Figure 6 shows the Sharpe ratio of L12, L1, EN and three benchmark portfolios which are widely used in practice: EW, SC and SU. We use DJIA and NASDAQ data sets. The penalty parameters are $\lambda_1 = \lambda_2 = \lambda$, and varying the value of λ from $10^{-1.5}$ to 10, $c = 10$ and $tol = 10^{-4}$. We can see that, in some periods, the Sharpe ratio of L12 is almost always higher than that by L1 and slightly higher than that of EN.

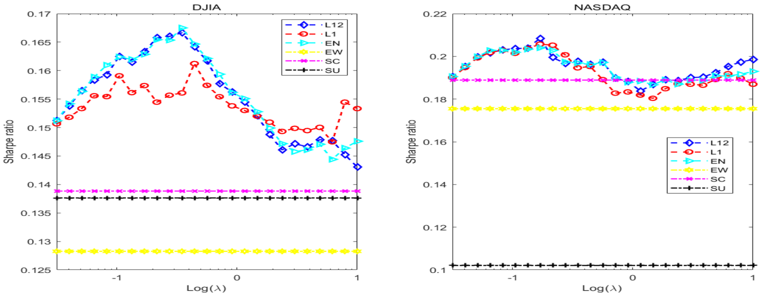

Fig. 6 The Sharpe ratio of L12, L1, EN and three benchmark portfolio (DJIA data from Dec. 18, 2017 to Apr. 06, 2020, $\tau = 60$; NASDAQ data from Jan. 05, 2015 to Oct. 29, 2018, $\tau = 120$). Since λ plays no role in EN, EW, SC, and SU, their Sharpe ratio is a straight line against λ.

Figure 7 shows the Sharpe ratio of three benchmark portfolios and the L12 portfolio with different $\alpha = 1$, 0.8, 0.5, 0.3 and 0. We use FF100 and SP500 data sets. The penalty parameters are $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)$. The value of λ varying from 10^{-2} to 10^2, $c = 10$ and $tol = 10^{-4}$. For each α, as λ increases, the Sharpe ratio of L12 increases first and then declines. However, the Sharpe ratio of L12 is always above that of the SU portfolio and it can be higher than that of all the three benchmark portfolios when a suitable λ is chosen.

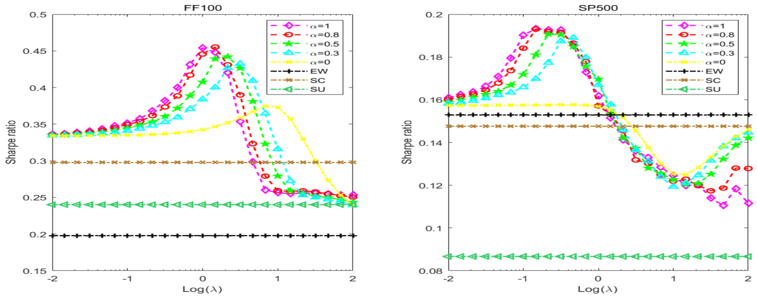

Fig. 7 The Sharpe ratio of L12 and three benchmark portfolio (FF100 data from Nov., 1995 to Dec., 2007; SP500 data from Apr. 13, 2015 to Oct. 09, 2017; $\tau = 60$).

Table 6 reports the out-of-sample performance by using four quantities defined in Section 5.1 and we set $\lambda_1 = 3$, $\lambda_2 = 3$, $c = 10$ and $tol = 10^{-4}$. We can observe that EW and SU portfolios have the largest variance on average with $6.03(\%)^2$ and $5.96(\%)^2$ respectively. The variance of EN and L12 portfolio are similar across all data sets. L2 portfolio achieves smaller variances across all portfolio strategies. In addition, we observe that the Sharpe ratios of the various portfolios on average are 25.41% (SC), 25.36% (EN), 25.12% (L12),

Table 6 Portfolio out-of-sample variance ($\widehat{\sigma}^2$) ((%)^2), Sharp ratio ($\widehat{SR}$), turnover ($TURN$) and the average short positions (ASP) of different portfolio selection strategies (weekly data from June 22, 2015 to Nov. 18, 2019, $\tau = 60$; monthly data from Oct., 2009 to Oct., 2019; $\tau = 72$.).

		DJIA	NASDAQ	SP500	Russell2000	Russell300	FF100
		N=29	N=95	N=336	N=1340	N=2166	N=100
L12	$\widehat{\sigma}^2$	1.98381	2.25869	1.52188	1.85671	1.46840	10.1471
	$\widehat{SR}$	0.23595	0.19817	0.15550	0.30849	0.31146	0.29769
	$TURN$	0.05646	0.10606	0.12390	0.15597	0.14891	0.12259
	ASP	8.35E-06	8.14E-06	1.09E-04	6.01E-04	6.67E-04	0.02096
L2	$\widehat{\sigma}^2$	2.01456	2.11008	1.38388	1.73384	1.25711	7.00065
	$\widehat{SR}$	0.23559	0.18545	0.19713	0.23118	0.28039	0.31461
	$TURN$	0.07234	0.20522	0.24692	0.21357	0.20233	0.60469
	ASP	0.06460	0.38253	0.49899	0.43966	0.44889	1.82499
L1	$\widehat{\sigma}^2$	2.01439	2.13231	1.47325	1.83275	1.44224	9.25632
	$\widehat{SR}$	0.20761	0.15524	0.15753	0.32592	0.32511	0.31517
	$TURN$	0.13299	0.15676	0.16584	0.17724	0.17568	0.18922
	ASP	9.93E+00	-3.78E-08	1.06E-04	7.39E-04	7.73E-04	0.02976
EN	$\widehat{\sigma}^2$	1.94038	2.17864	1.49136	1.83431	1.44526	9.90031
	$\widehat{SR}$	0.23317	0.18662	0.15461	0.32114	0.32428	0.30181
	$TURN$	0.07566	0.12855	0.15532	0.17211	0.16984	0.12261
	ASP	5.29E-06	8.37E-07	1.04E-04	7.42E-04	7.49E-04	0.02447
SC	$\widehat{\sigma}^2$	2.07429	2.37054	1.63826	1.85500	1.45764	8.30524
	$\widehat{SR}$	0.18748	0.10532	0.12759	0.38125	0.35991	0.36345
	$TURN$	0.15456	0.18364	0.30801	0.43402	0.47122	0.14687
	ASP	1.04E-10	-1.38E-12	1.00E-13	1.29E-13	6.03E-13	7.18E-13
SU	$\widehat{\sigma}^2$	3.49337	4.78759	1.89031	3.18212	2.36929	20.0196
	$\widehat{SR}$	0.18248	0.09557	0.17626	0.11328	0.12944	0.19851
	$TURN$	0.71041	2.34017	0.67355	0.30893	0.28456	6.51852
	ASP	1.21327	2.13683	0.93632	0.39211	0.38312	6.73961
EW	$\widehat{\sigma}^2$	2.42987	3.57957	2.54673	4.02917	3.27845	20.3124
	$\widehat{SR}$	0.20654	0.20759	0.12039	0.14270	0.15859	0.19955
	$TURN$	0.01425	0.02100	0.02034	0.03062	0.02709	0.02101
	ASP	1.12E-16	1.12E-16	-1.12E-16	4.47E-16	-3.35E-16	0
SCID	$\widehat{\sigma}^2$	2.35213	3.01854	1.42864	1.68951	1.21899	8.33069
	$\widehat{SR}$	0.24806	0.11569	0.20955	0.23836	0.29046	0.27559
	$TURN$	0.28054	0.71968	0.50116	0.23481	0.22195	1.27282
	ASP	0.53436	1.35843	1.05423	0.49395	0.50159	3.40904
SC1F	$\widehat{\sigma}^2$	2.15511	2.66150	1.41161	1.69618	1.22063	7.04642
	$\widehat{SR}$	0.25468	0.13452	0.20870	0.23706	0.28955	0.28684
	$TURN$	0.20129	0.87781	0.58113	0.25516	0.23405	1.32178
	ASP	0.37529	1.05698	0.99627	0.49364	0.50109	2.72615

24.78% (L1), 24.07% (L2), 23.52% (SC1F), 22.96% (SCID), 17.26% (EW) and 14.92% (SU). We see that the L12 portfolio does not result in a significantly different out-of-sample Sharpe ratio when compared with EN and SC, however, it is higher than the rest portfolio strategies.

As for the portfolio turnover, unsurprisingly, the long only EW portfolio strategy exhibits the lowest turnover of all portfolio strategies, amounting to 2.24%. However, the L12 portfolio strategy is also desirable since it achieves the lowest turnover except EW portfolio, ranging between 5.65%

(DJIA) and 15.60% (Russell2000). The highest average turnover is generated by SU amounting on average to 180.60%, meaning that it is very costly to use SU portfolio. The turnover of norm constrained portfolio are comparatively similar, amounting on average to 11.90% (L12), 25.75% (L2), 16.63% (L1) and 13.73% (EN), respectively. The turnover of the remaining portfolio strategies ranges between 28.31% (SC) and 57.85% (SC1F) on average.

The high turnover of SU is reflected in the enormous average short positions of over 196.69% on average across the six data sets. The second two highest average short positions are by SCID and SC1F, respectively amounting to 122.53% and 102.49%. The average short positions of SC and EW portfolios are on average approximate to 0% across the six data sets. And the average short positions of the other portfolio strategies are in the range between 0.37% (L12) and 60.99% (L2) across the six data sets. Therefore, given the moderate turnover and the average short positions, the proposed L12 strategy represents a practically implementable method that outperforms the portfolio strategies listed in Table 3 consistently and significantly.

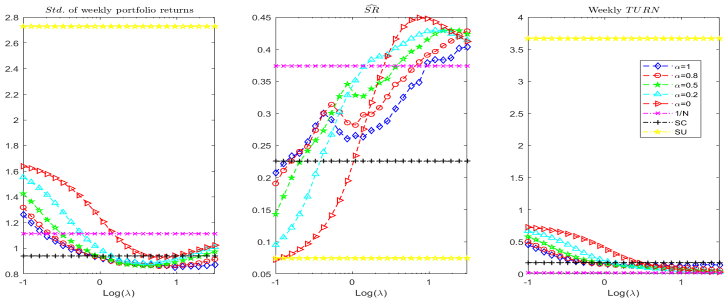

Fig. 8 Standard deviation ($Std.$) of weekly portfolio returns, Sharpe ratio ($\widehat{SR}$), weekly turnover ($TURN$). (NASDAQ data from Oct. 05, 2015 to Jan. 22, 2018, $\tau = 72.$).

At the end of this subsection, we compare the standard deviation ($Std.$), Sharpe ratio ($\widehat{SR}$) and turnover ($TURN$) of the L12 with three benchmark portfolios: EW, SC, and (SU). We use NASDAQ and FF100 data sets. The penalty parameters are $\lambda_1 = \lambda\alpha$, $\lambda_2 = \lambda(1 - \alpha)$. We set $\alpha = 1, 0.8, 0.5, 0.2, 0$ and vary λ from 10^{-1} to $10^{1.5}$, $c = 10$, $tol = 10^{-4}$. Figure 8 and 9 show that, as λ increases, $Std.$ of L12 portfolio declines to a minimum and then converges to a level around where the SC portfolio holds. The trend of $\widehat{SR}$ of L12 is on increase first and then it declines, however, L12 can yield a higher $\widehat{SR}$ than the benchmark portfolios. The trend of $TURN$ of L12 is in decline consistently and converges to a level where the EW portfolio holds. Therefore, we can conclude that for those tested datasets L12 portfolio has lower variance, higher Sharpe ratio and lower turnover than benchmark portfolios, especially when the value of λ is in the range between 1 and 10.

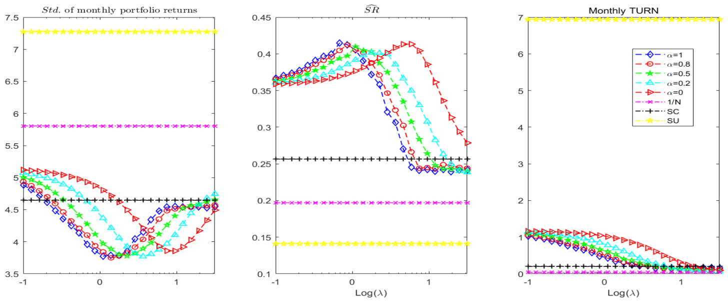

Fig. 9 Standard deviation (*Std.*) of monthly portfolio returns, Sharpe ratio ($\widehat{SR}$), monthly turnover ($TURN$). (FF100 data from Oct., 1993 to Oct., 2003, $\tau = 72$.).

5.4 Speed comparison with existing solvers

The L12 model being convex can be solved by existing optimization solvers. The purpose of this subsection is to demonstrate that the proposed algorithm pALM can produce high quality of solutions and outperform several well-known convex optimization solvers in terms of the time taken. The existing algorithms used are the alternating direction method of multipliers (ADMM), the fast iterative shrinkage-thresholding algorithm (FISTA), and the projected subgradient method (PSM). We refer the reader to Beck [1] for more details about these algorithms.

In order to measure the solution quality, we take the solution produced by CVX package [15] as our benchmark solution and compute the difference between it and the one obtained by each method considered. The smaller the difference is, the better quality the solution is. Let $\mathbf{w}_{cvx,t}$ be the weight vector at period t produced by CVX, and let $\mathbf{w}_{alg,t}$ be the weight vector for the same period produced by one of the algorithms among pALM, ADMM, FISTA and PSM. We define the cumulative difference between the weight vector obtained and $\mathbf{w}_{cvx}$ (see [37] for more explanation of the measure):

$$\sum_{t=\tau+1}^{T} \|\mathbf{w}_{cvx,t} - \mathbf{w}_{alg,t}\|_1, \tag{5.6}$$

We used data NASDAQ and Russell 3000 to conduct the comparison. We set the penalty parameters $\lambda_1 = \lambda_2 = \lambda$, and vary λ from $10^{0.5}$ to $10^{2.5}$, $c = 100$, $tol = 10^{-3}$. Figure 10 illustrates the results. We can see the value of the cumulative difference are in decline with the growth of λ. It can also be seen that the curve formed by cvx-pALM is smooth and strictly decreasing as λ increases. In particular, it is below all other curves when $\lambda \geq 10$. This means that solution quality of pALM is the best among those obtained by the tested algorithms.

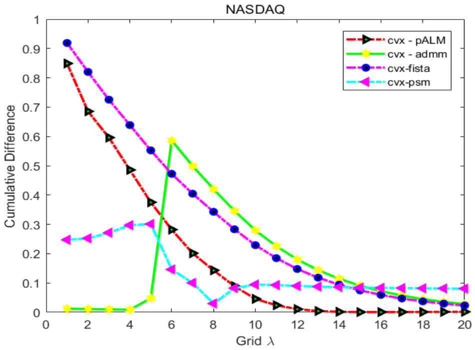

Fig. 10 Cumulative difference between solution from pALM, ADMM, FISTA, PSM and CVX (NASDAQ data from May 01, 2015 to Nov. 23, 2020.).

Table 7 reports the average running time for each method over 100 runs against the 6 data sets. It can be clearly seen that pALM shows growing advantage over the others as the problem size grows. For example, for the case of Russell3000 where we have 2166 assets, the time (less than 1 second) used by pALM is a fraction of that by others (at least 10 seconds). The parameters for these experiments are set as $\lambda_1 = 10$, $\lambda_2 = 10$, $c = 100$ and $tol = 10^{-3}$.

Table 7 Average CPUtime (seconds) with L12 portfolio in six real-world market. Each simulation is run 100 times (weekly data from May 01, 2015 to Nov. 23, 2020 and monthly data from Nov., 1978 to June, 2020.).

	DJIA	NASDAQ	SP500	Russell2000	Russell300	FF100
	N=29	N=95	N=336	N=1340	N=2166	N=100
pALM	0.00746	0.00833	0.02521	0.13972	0.19167	0.00619
ADMM	0.05108	0.45986	4.74670	105.935	795.817	0.53085
FISTA	0.00800	0.04339	0.73759	25.0486	141.848	0.05827
PSM	0.09555	0.13224	0.53499	3.84220	9.63697	0.14679
CVX	0.57184	0.64192	2.33361	16.7944	29.5056	0.67345

6 Conclusion

In this paper, we proposed a new portfolio selection model which is the minimum variance portfolio with $\ell_{1,2}$ regularization. We derived a theoretical bound that limits short sells and developed a closed form formula for the proximal term of the $\ell_{1,2}$ norm. Numerical results show that $\ell_{1,2}$ portfolio can lead to small turnover and small amount of short positions. Moreover, $\ell_{1,2}$ can achieve higher Sharpe ratio when compared with several benchmark portfolios. We also designed a proximal augmented Lagrange method (pALM) to solve $\ell_{1,2}$ MVP and showed that the algorithm is efficient and fast when compared with other algorithms. In particular, for the case there are thousands of assets, pALM took a fraction of the time used by other popular algorithms to solve the model.

When there are no short-sales allowed (i.e., $\mathbf{w} \geq 0$), the ℓ_1 norm $\|\mathbf{w}\|_1 = 1$ is constant. In this case, the ℓ_1 norm being used in the objective would not contribute to the sparse solution. An important strategy for this case is to penalize the ℓ_0 norm $\|\mathbf{w}\|_0$ directly to the objective. An interesting research topic is to combine this penalty approach with the ℓ_2-norm regularization and study the resulting portfolios.

Acknowledgements We thank the editor and two referees for their detailed comments that have improved the quality of the paper. In particular, one referee points us to the much relevant paper [38], which leads to the comment Remark 3 in the paper.

References

1. A. Beck, *First-Order Methods in Optimization*. SIAM and Mathematical Optimization Society, 2017.
2. P. Behr, A. Guettler, and F. Miebs, *On portfolio optimization: imposing the right constraints*. Journal of Banking and Finance, 37 (2013), 1232–1242.
3. G.-Y. Ben, E.K. Noureddine and A.EB. LIM, *Machine learning and portfolio optimization*. Management Science, 64.3 (2018) 1136–1154.
4. A. Borodin, R. El-Yaniv and V. Gogan, *Can we learn to beat the best stock*. Journal of Artificial Intelligence Research, 21 (2004), 579-594.
5. J. Brodie, I. Daubechies, C. De Mol, D. Giannone and I. Loris, *Sparse and stable Markowitz portfolios*. Proceedings of the National Academy of Sciences of the United States of America, 106 (2009), 12267–12272.
6. E. J. Candés and T. Tao, *Decoding by linear programming*. IEEE Transactions on Information Theory, 51 (2005), 4203–4215.
7. R.K. Chou and H. Chung, *Decimalization, trading costs, and information transmission between ETFs and index futures*. Journal of Futures Markets: Futures, Options, and Other Derivative Products, 26(2) (2006), 131-151.
8. Z. Dai and F. Wen, *Some improved sparse and stable portfolio optimization problems*. Finance Research Letters 27 (2018), 46–52.
9. V. DeMiguel, L. Garlappi, F.J. Nogales, and R. Uppal, *A generalized approach to portoflio optimization: improving performance by constraining portfolio norms*. Management Science, 55 (2009), 798–812.
10. 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), 1915–1953.
11. J. Fan, J. Zhang, and K. Yu, *Vast portfolio selection with gross exposure constraints*. Journal of the American Statistical Association, 107 (2012), 592–606.
12. B. Fastrich, S. Paterlini and P. Winker, *Constructing optimal sparse portfolios using regularization methods*. Comput. Management Sci., 12 (2015), 417–434.
13. M. Fazel, T.K. Pong, D. Sun and P. Tseng, *Hankel matrix rank mininization with applications to system identification and realization*. SIAM Journal on Matrix Analysis and Applications, 34 (2013), 946–977.
14. M. Giuzio and S. Paterlini, *Un-diversifying during crises: Is it a good idea?*. Computational Management Science, 16 (2019), 401–432.
15. M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming, version 1.21*. http://cvxr.com/cvx. (2010).
16. R.C. Green and B. Hollifield, *When will mean-variance efficient portfolios be well diversified?*. The Journal of Finance, 47.5 (1992), 1785–1809.

17. M. Ho, Z. Sun and J. Xin, *Weighted elastic net penalized mean-variance portfolio design and computation.* SIAM Journal on Financial Mathematics, 6 (2015), 1220–1244.
18. R. Jagannathan and T. Ma, *Risk reduction in large portfolios: why imposing the wrong constraints helps.* The Journal of Finance, 58 (2003), 57–72.
19. M.J. Kim, Y. Lee, J.H. Kim, and W.C. Kim, *Sparse tangent portfolio selection via semi-definite relaxation.* Operations Research Letters, 44.4 (2016), 540–543.
20. P.J. Kremer, S. Lee, M. Bogdan and S. Paterlini, *Sparse portfolio selection via the sorted ℓ_1-Norm.* Journal of Banking and Finance, 110 (2020), 105687.
21. Z.R. Lai, P.Y. Yang, L. Fang and X. Wu, *Short-term sparse portfolio optimization based on alternating direction method of multipliers.* The Journal of Machine Learning Research, 19(1) (2018), 2547-2574.
22. F.-S. Lhabitant, *Portfolio Diversification.* ISTE Press Ltd and Elsevier Ltd., 2017.
23. J. Li, *Sparse and stable portfolio selection with parameter uncertainty.* Journal of Business and Economic Statistics, 33 (2015), 381–392.
24. B. Maillet, S. Tokpavi, and B. Vaucher, *Global minimum variance portfolio optimisation under some model risk: A robust regression-based approach.* European Journal of Operational Research, 244.1 (2015), 289–299.
25. H. Markowitz, *Portfolio selection.* The Journal of Finance, 7 (1952), 77–91.
26. H. Markowitz, *Portfolio Selection: Efficient Diversification of Investment.* New York: John Wiley & Sons, 1959.
27. R. Merton, *On estimating the expected return on the market: an exploratory investigation.* Journal of Financial Economics, 8 (1980), 323–361.
28. L. Olivier and M. Wolf, *Analytical nonlinear shrinkage of large-dimensional covariance matrices.* Annals of Statistics, 48.5 (2020), 3043–3065.
29. L. Olivier, and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices.* Journal of Multivariate Analysis, 88.2 (2004), 365–411.
30. L. Olivier, and M. Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection.* Journal of Empirical Finance, 10.5 (2003), 603–621.
31. S. Perrin and T. Roncalli, *Machine learning optimization: algorithms and portfolio allocation.* In: Machine Learning for Asset Management: New Developments and Financial Applications (2020): 261-328.
32. W. Shen, J. Wang and S. Ma., *Doubly Regularized Portfolio with Risk Minimization.* In: AAAI, (2014), 1286-1292.
33. A. Takeda, M. Niranjan, J.Y. Gotoh, and Y. Kawahara, *Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios.* Computational Management Science 10 (2013), 21–49.
34. Y. Teng, L. Yang, B. Yu and X. Song, *A penalty PALM method for sparse portfolio selection problems.* Optimization Methods and Software, 32(1) (2017), 126-147.
35. R. Tibshirani, *Regression shrinkage and selection via the Lasso.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58 (1996), 267-–288.
36. Y.-M. Yen, *Sparse weighted-norm minimum variance portfolios.* Review of Finance, 2015, 1–29.
37. Y.-M. Yen and T.J. Yen, *Solving norm constrained portfolio optimization via coordinate-wise descent algorithms.* Computational Statistics and Data Analysis, 76 (2014), 737–759.
38. H. Zhang, J. Jiang and Z.-Q. Luo, *On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems.* Journal of the Operations Research Society of China 1 (2013), 163–186.
39. H. Zou and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2005), 301–320.