Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 10, No. 1, pp. 21–35, January 2014

A SUCCESSIVE CONVEX APPROXIMATION APPROACH FOR SPARSE SOLUTIONS OF CONVEX PROGRAMS*

Xiaodi Bai, Xiaojin Zheng†, Xueting Cui and Xiaoling Sun

Abstract: Finding sparse solutions of optimization problems are often encountered in real-world applications when the number of nonzero entries in the decision variables has to be controlled. We study in this paper sparse solutions to general convex programs. We present a successive convex approximation method for solving the regularization formulation of sparse convex programs. This method is based on DC approximation of the ℓ_0 function and generates a sequence of approximate solutions via solving convex subproblems successively. We prove the convergence of the method to a KKT point of the approximate regularization problem. A piecewise linear DC approximation to the ℓ_0 regularization term is also derived via an ℓ_1 exact penalty function of a mixed-integer program reformulation. We report computational results of the method using different DC approximations for test problems from limited diversified mean-variance portfolio selection. Our preliminary numerical results suggest that the proposed method is promising in finding sparse solutions of convex programs.

Key words: *sparse solutions to convex program, regularization formulation, successive convex approximation, piecewise-linear DC approximation, portfolio selection*

Mathematics Subject Classification: *90C11, 90C20, 90C25*

1 Introduction

In this paper, we are concerned with finding sparse solutions to a general convex program. Mathematically, the problem can be expressed as

$$
\text{(P)} \qquad \begin{aligned} &\min\ f(x) \\ &\text{s.t. } x \in X, \\ &\qquad \|x\|_0 \le K, \end{aligned}
$$

where $f : \Re^n \to \Re$ is a differentiable convex function, $X = \{x \in \Re^n \mid g(x) \le 0\}$, $g(x) = (g_1(x), \ldots, g_m(x))^T$, $g_i : \Re^n \to \Re$, $i = 1, \ldots, m$, are differentiable convex functions, and $\|x\|_0$ denotes the ℓ_0-(quasi) norm which is defined as the number of nonzero entries in x. The sparsity constraint $\|x\|_0 \le K$ with $1 \le K < n$ is also called *cardinality* constraint which controls the number of nonzero variables of x. By introducing a 0-1 variable y_i for each x_i, the sparsity constraint can be represented by mixed-integer constraints $\sum_{i=1}^n y_i \le K$ and

*This research was supported by the NSFC grants 11101092 and 11371103, and Shanghai Pujiang Program 13PJC108.

†Corresponding author.

ISSN 1348-9151

$l_i y_i \le x_i \le u_i y_i$, $y_i \in \{0,1\}$, $i = 1,\dots,n$, where l_i and u_i are lower and upper bounds of x_i, respectively. Therefore, problem (P) can be viewed as a special mixed-integer convex program.

Finding sparse solutions of optimization problems are often encountered in real-world applications when the number of nonzero entries in the decision variables has to be limited. Applications of sparse solutions include portfolio selection [4, 8, 17, 29, 31], subset selection in multivariate regression [1, 26], signal processing and compressed sensing [7]. In the literature of sparse solutions, much attention has been devoted to sparse solutions of linear systems $Ax = b$ which can be formulated as minimizing $\|x\|_0$ subject to $Ax = b$ (see, e.g., [23] and the references therein). This problem is related to but different from problem (P) where one seeks for solutions with a given cardinality K.

An important subclass of problem (P) is cardinality-constrained quadratic programs in the following form:

$$\text{(QP)} \qquad \begin{aligned} &\min\ f(x) := x^T Q x + c^T x \\ &\text{s.t. } Ax \le b, \\ &\quad\ \ \|x\|_0 \le K, \end{aligned}$$

where Q is an $n \times n$ symmetric positive semidefinite The cardinality-constrained portfolio selection model is a special case of (QP) where the cardinality constraint limits the total number of different assets in the optimal portfolio. Branch-and-bound methods based on various relaxations and bounding techniques have been proposed for solving the mixed-integer quadratic program reformation of (QP) (see, e.g., [3, 4, 5, 24, 29, 30]). A mixed-integer quadratically constrained quadratic program reformulation is derived in [14] for a class of cardinality-constrained portfolio selection problems where the assets returns are driven by factor models. Recently, a novel geometric approach is proposed in [17] for minimizing a quadratic function with cardinality constraint. Embedded in a branch-and-bound method, this geometric approach is efficient for solving cardinality-constrained portfolio selection problems. The quadratic model (QP) also has application in indexing tracking in passive portfolio management where a small set of assets is selected to track the performance of market benchmark index. When the tracking error is measured by variance, the index tracking problem is a special case of (QP) with $f(x) = (x - x_B)^T Q (x - x_B)$, where x_B is the vector of positions of the benchmark index. In [22], an ℓ_p approximation method is presented to find an approximate optimal solution of indexing tracking problem. A continuously differentiable nonconvex piecewise quadratic approximation to ℓ_0-norm is proposed in [13] for solving index tracking problem. In [32], a successive convex approximation approach is proposed to construct convex approximations of the nonconvex feasible set of (P). Valid inequalities are also derived in [32] to strengthen the convex approximations to (P).

While cardinality constraint imposes the sparsity directly to the feasible solutions of a convex program, which is convenient for controlling the sparsity in a single model, it makes the problem difficult to solve. In fact, it has been shown in [4] that problem (P) is in general NP-hard due to the combinatorial nature of the sparsity constraint. An alternative formulation for finding sparse solutions can be obtained by integrating the ℓ_0 function into the objective function. The resulting problem is

$$(\mathrm{P}_\mu) \qquad \begin{aligned} &\min\ f(x) + \mu \|x\|_0 \\ &\text{s.t. } x \in X, \end{aligned}$$

where $\mu > 0$ is a regularization parameter. Problem (P_μ) is often called *regularization* formulation of (P). We assume that there exists feasible solution x such that $\|x\|_0 = K$.

Under this assumption, it is easy to see that the inequality sparsity constraint $\|x\|_0 \le K$ in (P) can be replaced by $\|x\|_0 = K$. It can be shown that for each $1 \le K < n$, the optimal solution of (P) can be generated via solving (P_μ) for some $\mu > 0$ (see [22]). In computation, to generate an optimal solution of (P_μ) with $\|x\|_0 = K$, we can solve (P_μ) (approximately) with different values of μ and eventually find an optimal solution of (P).

A special case of (P_μ) is ℓ_2-ℓ_0 minimization problem defined by

$$\min_{x \in \Re^n} \|Ax - b\|^2 + \mu \|x\|_0.$$

Approximating $\|x\|_0$ by the ℓ_p term $\|x\|_p^p$ with $p \in (0, 1)$ leads to the ℓ_2-ℓ_p minimization problem. The lower bound theory of nonzero entries of ℓ_2-ℓ_p minimization were discussed in [9, 10, 11]. In [18], an interior-point potential reduction algorithm was proposed to search for a local solution of ℓ_2-ℓ_p minimization.

In this paper, we focus on the regularization formulation (P_μ). We propose a successive convex approximation method for (P_μ) using a DC (difference of convex functions) approximation $\psi(x, p)$ to $\|x\|_0$. By successively linearizing the concave term in $\psi(x, p)$, we obtain a sequence of convex subproblems whose optimal solutions provide candidates of approximate solutions to (P_μ). We establish the convergence of the sequence of approximate solutions to a KKT point of the approximate regularization problem. By applying ℓ_1 exact penalty function of a mixed-integer nonlinear program reformulation of (P_μ), we also derive a piecewise linear DC approximation to the regularization term of (P_μ). We report preliminary computational results of the method for test problems of limited diversified mean-variance portfolio selection. Our numerical results suggest that the methods is promising in generating sparse solutions to convex programs.

The paper is organized as follows. In Section 2, we first describe the successive convex approximation method based on DC approximation. Some theoretical results of the method is then established. In particular, we prove the convergence of the method to a KKT point of the approximation regularization problem. In Section 3, we derive a piecewise linear DC approximation to (P_μ) via ℓ_1 exact penalty function approach to a mixed-integer nonlinear reformulation of (P_μ). Computational results are reported in Section 4. Finally, we give some concluding remarks in Section 5.

2 A Successive Convex Approximation Method

In this section, we first describe a successive convex approximation (SCA) method for problem (P_μ) using DC approximation to ℓ_0. We prove that the method converges to a KKT point of the approximation regularization problem.

Suppose that $\psi(x, p)$ is an approximation to $\|x\|_0$ satisfying the following conditions:

(C1) $\psi(x, p)$ is a nonnegative DC function of x for any $p > 0$ with DC decomposition $\psi(x, p) = c_1(x, p) - c_2(x, p)$, where $c_1(x, p)$ and $c_2(x, p)$ are convex functions of x on X;

(C2) $\lim_{p \to 0^+} \psi(x, p) = \|x\|_0$ for any $x \in X$.

Consider the following approximation of the regularization problem (P_μ):

$$(\mathrm{P}_{p,\mu}) \qquad \begin{array}{l} \min \ F(x) := f(x) + \mu \psi(x, p) \\ \text{s.t. } x \in X, \end{array}$$

where $\psi(x, p)$ satisfies (C1) and (C2). By condition (C2), the optimal solution of ($\mathrm{P}_{p,\mu}$) provides a reasonable approximation solution of the regularization problem (P_μ) when $p > 0$ is small.

Since $f(x)$ is convex and $\psi(x,p)$ is a nonconvex function of x, problem $(\mathrm{P}_{p,\mu})$ is a nonconvex optimization problem with a DC objective function and a convex feasible set. In the literature of DC optimization, most of the approaches for DC optimization problems were devoted to global optimization methods which are of branch-and-bound framework based on various partition and bounding techniques (see [20]). Recently, local methods based on sequential convex programming methods have been proposed in [19, 28] for dealing with nonlinear programming with DC constraints.

We will focus on local method based on successive convex approximation to the DC objective function $F(x)$ in $(\mathrm{P}_{p,\mu})$. The idea of the successive convex approximation method is to construct convex subproblems of $(\mathrm{P}_{p,\mu})$ via linearizing the concave term in the DC function $\psi(x,p)$. Let $\bar{x} \in X$. Since $c_2(x,p)$ is a convex function of x, we have

$$c_2(x,p) \geq c_2(\bar{x},p) + \bar{\xi}^T(x-\bar{x}), \qquad \forall x \in X,$$

where $\bar{\xi} \in \partial c_2(\bar{x},p)$. Then

$$f(x) + \mu\psi(x,p) \leq f(x) + \mu[c_1(x,p) - c_2(\bar{x},p) - \bar{\xi}^T(x-\bar{x})].$$

We obtain the following convex subproblem at $\bar{x}$:

$$(\mathrm{CP}_{p,\mu}(\bar{x})) \qquad \begin{array}{l} \min\ f(x) + \mu[c_1(x,p) - c_2(\bar{x},p) - \bar{\xi}^T(x-\bar{x})] \\ \text{s.t. } x \in X. \end{array}$$

In the sequel, we need the following assumptions:

(A1) The Slater condition holds for $X = \{x \in \Re^n \mid g(x) \leq 0\}$, i.e., the interior of X is nonempty.

(A2) $f(x)$ is a strongly convex function on X satisfying

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x-\bar{x}) + \frac{\rho}{2}\|x-\bar{x}\|^2, \qquad \forall x,\ \bar{x} \in X,$$

where $\rho > 0$ is a constant.

Lemma 2.1. *If $\bar{x}$ is an optimal solution to $(\mathrm{CP}_{p,\mu}(\bar{x}))$, then $\bar{x}$ is a KKT point of $(\mathrm{P}_{p,\mu})$.*

Proof. Let $\bar{x}$ be an optimal solution to $(\mathrm{CP}_{p,\mu}(\bar{x}))$. By assumption (A1), there exists a multiplier vector $\theta \in \Re^m_+$ such that

$$0 \in \nabla f(\bar{x}) + \mu[\partial c_1(\bar{x},p) - \bar{\xi}] + \nabla g(\bar{x})\theta, \tag{2.1}$$

$$\theta^T g(\bar{x}) = 0,\ g(\bar{x}) \leq 0, \tag{2.2}$$

where $\partial c_1(\bar{x},p)$ denotes the subgradient set of $c_1(x,p)$ at $\bar{x}$, $\bar{\xi} \in \partial c_2(\bar{x},p)$ and $\nabla g(\bar{x}) \in \Re^{n\times m}$ is the Jacobian matrix of $g(x)$ at $\bar{x}$. Since $\psi(x,p)$ is a DC function and thus is locally Lipschitz continuous on X with $\partial c_1(\bar{x},p) - \bar{\xi} \subseteq \partial\psi(\bar{x},p)$, where $\partial\psi(\bar{x},p)$ denotes the Clarke generalized gradient set of $\psi(x,p)$ at $\bar{x}$ (see [12]), we have from (2.1) that

$$0 \in \nabla f(\bar{x}) + \mu\partial\psi(\bar{x},p) + \nabla g(\bar{x})\theta.$$

This together with (2.2) implies that $\bar{x}$ is a KKT point of problem $(\mathrm{P}_{p,\mu})$. □

We are now ready to describe the successive convex approximation (SCA) method for solving $(\mathrm{P}_{p,\mu})$.

Algorithm 1 (SCA Method for $(P_{p,\mu})$).

Step 0. Choose parameter $\mu > 0$ and $p > 0$. Choose a small stopping parameter $\epsilon > 0$. Select x^0 satisfying $x^0 \in X$ and $\xi^0 \in \partial c_2(x^0, p)$. Set $k := 0$.

Step 1. Solve the following convex subproblem:

$$(\mathrm{CP}_{p,\mu}(x^k)) \qquad \begin{array}{l} \min\ f(x) + \mu[c_1(x,p) - c_2(x^k,p) - (\xi^k)^T(x - x^k)] \\ \text{s.t. } x \in X, \end{array}$$

where $\xi^k \in \partial c_2(x^k, p)$. Let x^{k+1} be an optimal solution to $(\mathrm{CP}_{p,\mu}(x^k))$ and $\theta^{k+1} \in \Re^m_+$ be the corresponding optimal multiplier vector.

Step 2. If $\|x^{k+1} - x^k\| \le \epsilon$, stop.

Step 3. Set $k := k + 1$ and go to Step 1.

Remark 2.2. The convex subproblem $(\mathrm{CP}_{p,\mu}(x^k))$ at Step 1 can be solved by any suitable convex optimization method in nonlinear programming. In particular, when $f(x)$ is a convex quadratic function and X is a polyhedron, we can solve the subproblem using efficient algorithms such as Lemke's method or interior-point method (see, e.g., [6, 27]). In our implementation of the algorithm for finding sparse solutions of quadratic programming test problems in Section 4, the convex quadratic subproblems at Step 1 are solved by the default QP solver in `CPLEX 12.3` which is an implementation of the barrier interior-point method for convex quadratic programming.

The following lemma gives an estimation of the amount of decrease in the function value of $F(x)$ at each iteration.

Lemma 2.3. *At the k-th iteration of Algorithm 1, the following inequality holds:*

$$F(x^k) - F(x^{k+1}) \ge \frac{\rho}{2}\|x^k - x^{k+1}\|^2, \tag{2.3}$$

where $F(x) = f(x) + \mu\psi(x,p)$.

Proof. Since x^{k+1} is an optimal solution of subproblem $(\mathrm{CP}_{p,\mu}(x^k))$, under assumption (A1), there exist a multiplier vector $\theta^{k+1} \in \Re^m_+$ and $\zeta^{k+1} \in \partial c_1(x^{k+1}, p)$ such that

$$\nabla f(x^{k+1}) + \mu(\zeta^{k+1} - \xi^k) + \nabla g(x^{k+1})\theta^{k+1} = 0, \tag{2.4}$$

$$(\theta^{k+1})^T g(x^{k+1}) = 0,\ g(x^{k+1}) \le 0, \tag{2.5}$$

where $\xi^k \in \partial c_2(x^k, p)$.

On the other hand, by the strong convexity of $f(x)$ and the convexity of $c_1(x)$, $c_2(x)$ and $g(x)$, we have

$$f(x^k) \ge f(x^{k+1}) + \nabla f(x^{k+1})^T(x^k - x^{k+1}) + \frac{\rho}{2}\|x^k - x^{k+1}\|^2, \tag{2.6}$$

$$c_1(x^k, p) \ge c_1(x^{k+1}, p) + (\zeta^{k+1})^T(x^k - x^{k+1}), \tag{2.7}$$

$$c_2(x^{k+1}, p) \ge c_2(x^k, p) + (\xi^k)^T(x^{k+1} - x^k), \tag{2.8}$$

$$g(x^k) \ge g(x^{k+1}) + \nabla g(x^{k+1})^T(x^k - x^{k+1}). \tag{2.9}$$

Using (2.8), we have

$$\begin{aligned}
F(x^k) - F(x^{k+1}) &= f(x^k) + \mu[c_1(x^k,p) - c_2(x^k,p)] \\
&\quad -f(x^{k+1}) - \mu[c_1(x^{k+1},p) - c_2(x^{k+1},p)] \\
&\geq f(x^k) + \mu[c_1(x^k,p) - c_2(x^k,p)] \\
&\quad -f(x^{k+1}) - \mu[c_1(x^{k+1},p) - c_2(x^k,p) - (\xi^k)^T(x^{k+1} - x^k)] \\
&\geq f(x^k) - f(x^{k+1}) + \mu[c_1(x^k,p) - c_1(x^{k+1},p) + (\xi^k)^T(x^{k+1} - x^k)] \\
&\quad +(\theta^{k+1})^T[g(x^k) - g(x^{k+1})], \qquad (2.10)
\end{aligned}$$

where the last inequality is due to (2.5) and $(\theta^{k+1})^T g(x^k) \leq 0$. Combining (2.10) with (2.6), (2.7) and (2.9), we obtain

$$\begin{aligned}
F(x^k) - F(x^{k+1}) &\geq \nabla f(x^{k+1})^T(x^k - x^{k+1}) + \frac{\rho}{2}\|x^k - x^{k+1}\|^2 \\
&\quad +\mu[(\zeta^{k+1})^T(x^k - x^{k+1}) + (\xi^k)^T(x^{k+1} - x^k)] \\
&\quad +(\theta^{k+1})^T\nabla g(x^{k+1})^T(x^k - x^{k+1}) \\
&= [\nabla f(x^{k+1}) + \mu(\zeta^{k+1} - \xi^k) + \nabla g(x^{k+1})\theta^{k+1}]^T(x^k - x^{k+1}) \\
&\quad +\frac{\rho}{2}\|x^k - x^{k+1}\|^2 \\
&= \frac{\rho}{2}\|x^k - x^{k+1}\|^2,
\end{aligned}$$

where the last equality is due to (2.4). This proves (2.3). □

Let $\epsilon = 0$. The convergence results of Algorithm 1 can be stated as follows.

Theorem 2.4. (i) *If the algorithm stops at Step 2, then x^k is a KKT point of* $(\mathrm{P}_{p,\mu})$.

(ii) *Assume that the multiplier sequence $\{\theta^k\}$ is bounded. If the algorithm generates an infinite sequence $\{x^k\}$, then any accumulation point of $\{x^k\}$ is a KKT point of* $(\mathrm{P}_{p,\mu})$.

Proof. (i) If the algorithm stops at Step 2, then $x^k = x^{k+1}$ solves subproblem $(\mathrm{CP}_{p,\mu}(x^k))$. By Lemma 2.1, x^k is a KKT point of $(\mathrm{P}_{p,\mu})$.

(ii) If the algorithm generates an infinite sequence $\{x^k\}$, by (2.3) in Lemma 2.3, $\{F(x^k)\}$ is a decreasing sequence. Also, since $f(x)$ is strongly convex over X and $\psi(x,p)$ is nonnegative, we imply that $F(x) = f(x) + \mu\psi(x,p)$ is bounded from below over X. Hence $\{F(x^k)\}$ converges and consequently $F(x^k) - F(x^{k+1}) \to 0$. It then follows from (2.3) that

$$\|x^k - x^{k+1}\| \to 0. \qquad (2.11)$$

Now, let x^* be an accumulation point of $\{x^k\}$. Then, there exists a subsequence $\{x^{k_j}\}$ such that $x^{k_j} \to x^*$ $(j \to \infty)$. By (2.11), we also have $x^{k_j+1} \to x^*$ $(j \to \infty)$. Since x^{k_j+1} is an optimal solution of the convex subproblem $(\mathrm{CP}_{p,\mu}(x^{k_j}))$, by assumption (A1), the following KKT conditions hold:

$$0 \in \nabla f(x^{k_j+1}) + \mu[\partial c_1(x^{k_j+1},p) - \xi^{k_j}] + \nabla g(x^{k_j+1})\theta^{k_j+1}, \qquad (2.12)$$

$$(\theta^{k_j+1})^T g(x^{k_j+1}) = 0,\ g(x^{k_j+1}) \leq 0 \quad i = 1,\dots,m, \qquad (2.13)$$

where $\theta^{k_j+1} \in \Re^m_+$ and $\xi^{k_j} \in \partial c_2(x^{k_j},p)$. By the assumption that $\{\theta^k\}$ is bounded and passing into a subsequence if necessary, we can assume that $\theta^{k_j+1} \to \theta^* \in \Re^m_+$. Note that the subgradient sets $\partial c_1(x,p)$ and $\partial c_2(x,p)$ are upper semicontinuous set mappings (see [2, Proposition 4.23]). Since $x^{k_j+1} \to x^*$, we have that $\limsup_{j\to\infty} \partial c_1(x^{k_j+1}, p) \subseteq \partial c_1(x^*, p)$ and $\limsup_{j\to\infty} \partial c_2(x^{k_j+1}, p) \subseteq \partial c_2(x^*, p)$. Also, since $x^{k_j} \to x^*$ and $\xi^{k_j} \in \partial c_2(x^{k_j}, p)$, by passing into a subsequence if necessary, we can assume that $\xi^{k_j} \to \xi^* \in \partial c_2(x^*, p)$. Taking limits in (2.12) and (2.13) for $j \to \infty$, we obtain that

$$\begin{aligned} &0 \in \nabla f(x^*) + \mu[\partial c_1(x^*, p) - \xi^*] + \nabla g(x^*)\theta^*, \\ &(\theta^*)^T g(x^*) = 0, \; g(x^*) \le 0. \end{aligned}$$

Thus, x^* is a KKT point of the convex subproblem ($\mathrm{CP}_{p,\mu}(x^*)$) and hence x^* is an optimal solution of ($\mathrm{CP}_{p,\mu}(x^*)$). We therefore conclude from Lemma 2.1 that x^* is a KKT point of ($\mathrm{P}_{p,\mu}$). □

Remark 2.5. When $X \subseteq \Re^n_+$, two notable examples of $\psi(x, p)$ satisfying (C1) and (C2) are ℓ_p function $\psi(x, p) = \|x\|_p^p = \sum_{i=1}^n x_i^p$ $(0 < p < 1)$ for $x \ge 0$ and the exponential function $\psi(x, p) = \sum_{i=1}^n (1 - e^{-\frac{1}{p}x_i})$ $(p > 0)$ for $x \ge 0$. In fact, we have $\lim_{p\to 0^+} \|x\|_p^p = \|x\|_0$ and $\lim_{p\to 0^+} \sum_{i=1}^n (1 - e^{-\frac{1}{p}|x_i|}) = \|x\|_0$. Also, we note that x_i^p with $0 < p < 1$ and $1 - e^{-\frac{1}{p}x_i}$ with $p > 0$ are concave functions for $x_i \ge 0$, which are special cases of DC functions. We point out that the ℓ_p function and the exponential function have been used in the literature for sparse solutions of linear systems (see [9, 10, 11, 25]).

3 A Piecewise Linear DC Approximation

In this section, we derive a new DC approximation to $\|x\|_0$ by applying ℓ_1 exact penalty function to a mixed-integer nonlinear programming reformulation of (P_μ).

Introducing $y_i \in \{0, 1\}$ to indicate x_i being zero or nonzero, for $i = 1, \ldots, n$, we can rewrite (P_μ) as a mixed-integer nonlinear program:

$$\begin{aligned} &\min \; f(x) + \mu \sum_{i=1}^n y_i \\ &\text{s.t. } x_i(1 - y_i) = 0, \; y_i \in \{0, 1\}, \; i = 1, \ldots, n, \\ &\qquad x \in X. \end{aligned} \tag{3.1}$$

In fact, from the first group of constraints of (3.1), we can see that $x_i \ne 0$ implies $y_i = 1$. Also, when $x_i = 0$, we must have $y_i = 0$ at the optimal solution of (3.1) as it gives a smaller objective value. Therefore, problem (P_μ) is equivalent to problem (3.1). The ℓ_1 exact penalty function of (3.1) with respect to constraints $x_i(1 - y_i) = 0$ $(i = 1, \ldots, n)$ is

$$P(x, y) = f(x) + \mu \left(\sum_{i=1}^n y_i + \sum_{i=1}^n \frac{1}{p_i} |x_i(1 - y_i)| \right),$$

where $p_i > 0$ $(i = 1, \ldots, n)$ are penalty parameters. Thus, we obtain the following penalty problem:

$$\begin{aligned} &\min \; P(x, y) \\ &\text{s.t. } y_i \in \{0, 1\}, \; i = 1, \ldots, n, \\ &\qquad x \in X. \end{aligned} \tag{3.2}$$

Note that

$$\begin{aligned}\min_{y\in\{0,1\}^n} P(x,y) &= f(x)+\mu\sum_{i=1}^n \min_{y_i\in\{0,1\}}(y_i+\frac{1}{p_i}|x_i(1-y_i)|)\\ &= f(x)+\mu\sum_{i=1}^n \min\{\frac{1}{p_i}|x_i|,1\}.\end{aligned}$$

Let $\psi(x,p)=\sum_{i=1}^n \phi(x_i,p_i)$, where

$$\phi(x_i,p_i)=\min\{\frac{1}{p_i}|x_i|,1\}=\begin{cases}\frac{1}{p_i}|x_i|, & |x_i|\le p_i,\\ 1, & \text{otherwise.}\end{cases} \tag{3.3}$$

Then the penalty problem (3.2) is equivalent to

$$\begin{aligned}&\min\ F(x):=f(x)+\mu\psi(x,p)\\ &\text{s.t. } x\in X.\end{aligned} \tag{3.4}$$

The following lemma summarizes the basic properties of $\psi(x,p)$.

Proposition 3.1. (i) *For any $x\in X$, $\lim_{p\to 0^+}\psi(x,p)=\|x\|_0$.*

(ii) *$\psi(x,p)$ is a nonnegative piecewise linear function of x and can be decomposed to a DC function:*

$$\psi(x,p)=c_1(x,p)-c_2(x,p), \tag{3.5}$$

where $c_1(x,p)=\sum_{i=1}^n\frac{1}{p_i}|x_i|$ and $c_2(x,p)=\sum_{i=1}^n\frac{1}{p_i}\left[\max(0,x_i-p_i)+\max(0,-x_i-p_i)\right]$ are piecewise linear convex functions of x.

Proof. (i) By (3.3), it is clear that $\lim\limits_{p_i\to 0^+}\phi(x_i,p_i)=\delta(x_i)$, where $\delta(x_i)$ is the step function defined by

$$\delta(x_i)=\begin{cases}0, & x_i=0,\\ 1 & x_i\neq 0.\end{cases}$$

Thus,

$$\lim_{p\to 0^+}\psi(x,p)=\sum_{i=1}^n\lim_{p_i\to 0^+}\phi(x_i,p_i)=\sum_{i=1}^n\delta_i(x_i)=\|x\|_0.$$

(ii) By the definition of $\phi(x_i,p_i)$ in (3.3), we have

$$\phi(x_i,p_i)=\frac{1}{p_i}|x_i|-\frac{1}{p_i}\left[\max(0,x_i-p_i)+\max(0,-x_i-p_i)\right].$$

The right-hand side of the above equation is a difference of two piecewise linear convex functions. Thus, (3.5) holds and $\psi(x,p)$ is a DC function of x. $\square$

Now, let's consider the subproblem at the kth iteration of Algorithm 1:

$$(\mathrm{CP}_{p,\mu}(x^k))\qquad \begin{aligned}&\min\ f(x)+\mu[c_1(x,p)-c_2(x^k,p)-(\xi^k)^T(x-x^k)]\\ &\text{s.t. } x\in X,\end{aligned} \tag{3.6}$$

where $c_1(x,p)$ and $c_2(x,p)$ are defined in (3.5) and $\xi^k \in \partial c_2(x^k,p)$. Since $c_1(x,p) = \sum_{i=1}^n \frac{1}{p_i}|x_i|$, the nonsmoothness of $c_1(x,p)$ in (3.6) can be eliminated by introducing $z_i = |x_i|$ for $i = 1,\dots,n$. The resulting problem is a smooth convex problem:

$$\begin{aligned} \min\ & f(x) + \mu[\sum_{i=1}^n \frac{1}{p_i} z_i - c_2(x^k,p) - (\xi^k)^T(x - x^k)] \\ \text{s.t.}\ & -z_i \le x_i \le z_i,\ i = 1,\dots,n, \\ & x \in X. \end{aligned} \tag{3.7}$$

Proposition 3.2. *Problem (3.6) is equivalent to problem (3.7).*

Proof. Note that the first constraint in (3.7) is equivalent to $z_i \ge |x_i|$ $(i = 1,\dots,n)$ and the optimal solution of (3.7) must satisfy $z_i = |x_i|$ $(i = 1,\dots,n)$. Therefore, for any optimal solution (x^*, z^*) of (3.7), x^* is an optimal solution to (3.6). Conversely, for any optimal solution x^* of (3.6), (x^*, z^*) with $z_i^* = |x_i^*|$ $(i = 1,\dots,n)$ is an optimal solution of (3.7). □

Remark 3.3. From Proposition 3.1 (ii), we see that $\psi(x,p) = \sum_{i=1}^n \phi(x_i,p_i)$ is a DC function and can be viewed as an approximation of the ℓ_0 function for small $p > 0$. This DC approximation to ℓ_0 function have been used in [32] to deal with cardinality constraint directly. The subproblem (3.7) is a smooth convex optimization problem and can be solved efficiently by convex optimization methods. In particular, when $f(x)$ is a convex quadratic function and $g_i(x)$'s are linear functions, (3.7) is a convex quadratic programming problem.

4 Computational Results

In this section, we present computational results of the successive convex approximation method (Algorithm 1) for the approximate regularization problem $(\mathrm{P}_{p,\mu})$. The method is coded in Matlab (version R2011a) and run on a PC equipped with Intel Pentium G630 CPU (2.70 GHz) and 8 GB of RAM. All the convex quadratic subproblems in Algorithm 1 are solved by the QP solver in `CPLEX 12.3` with Matlab interface (see [21]).

The test problems in our computational experiments are limited diversified mean-variance portfolio selection problems (see [4, 5]). Let μ and Q be the mean and covariance matrix of n risky assets, respectively. The limited diversified mean-variance portfolio selection model can be formulated as

$$(\mathrm{MV}) \qquad \begin{aligned} \min\ & f(x) := x^T Q x \\ \text{s.t.}\ & \|x\|_0 \le K, \\ & x \in X, \end{aligned}$$

where $\|x\|_0 \le K$ is the constraint for sparsity control and X represents the constraints of minimum return level, budget constraint and lower and upper bounds for x_i:

$$X = \{x \in \Re^n \mid \mu^T x \ge \rho,\ \sum_{i=1}^n x_i = 1,\ 0 \le x_i \le u_i,\ i = 1,\dots,n\}.$$

To find sparse solutions of problem (MV), we consider the regularization form of (MV):

$$(\mathrm{MV}_{p,\mu}) \qquad \begin{aligned} \min\ & F(x) := x^T Q x + \mu\psi(x,p) \\ \text{s.t.}\ & x \in X. \end{aligned}$$

Our test problems for ($\mathrm{MV}_{p,\mu}$) consists of 60 instances where the parameters Q, μ, ρ and u_i were created in [15, 16]. There are 30 instances for each $n = 200$ and 300. The matrix Q in the 30 instances of each n were generated with different diagonal dominance. The parameters ρ and u_i are uniformly drawn at random from intervals $[0.002, 0.01]$ and $[0.375, 0.425]$, respectively. The data files of these instances are available at: `http://www.di.unipi.it/optimize/Data/MV.html`.

In our computational experiments, we use three types of nonconvex functions $\psi(x, p)$ as approximations of $\|x\|_0$:

- Piecewise linear DC function:

$$\psi(x, p) = \sum_{i=1}^{n} \frac{1}{p_i}|x_i| - \sum_{i=1}^{n} \frac{1}{p_i}\left[\max(0, x_i - p_i) + \max(0, -x_i - p_i)\right],$$

  where $p = (p_1, \ldots, p_n)^T$ is a parameter vector.

- ℓ_p function: $\psi(x, p) = \sum_{i=1}^{n} x_i^p$, $x \in \Re_+^n$, where p is a scalar parameter with $0 < p < 1$;

- Exponential function: $\psi(x, p) = \sum_{i=1}^{n}(1 - e^{-\frac{1}{p}x_i})$, $x \in \Re_+^n$, where $p > 0$ is a scalar parameter.

In our implementation, the initial solution x^0 in Step 0 of Algorithm 1 is obtained by solving the following convex quadratic programming:

$$\begin{aligned} &\min\ x^T Q x \\ &\text{s.t. } x \in X. \end{aligned}$$

The stopping parameter ϵ in Step 2 is set as $\epsilon = 10^{-7}$. The parameter p is set as follows. For the piecewise linear DC function, $p \in \Re^n$ is set as $p_i = 2\sqrt{x_i^0 + \epsilon_1}$, $i = 1, \ldots, n$, where $\epsilon_1 = 10^{-6}$. For the ℓ_p function and the exponential function, we set $p = \frac{1}{2}$ and $p = 0.01$, respectively, as in [10, 25].

For $n = 200$ and 300, we solve problem ($\mathrm{MV}_{p,\mu}$) with different parameter μ by Algorithm 1 using the three types of nonconvex approximations to $\|x\|_0$. The numerical results are summarized in Tables 1 and 2. We explain the notations in Tables 1 and 2 as follows:

- "SCA-DC", "SCA-ℓ_p" and "SCA-exp" stand for the three versions of Algorithm 1 with $\psi(x, p)$ being the piecewise linear DC function, the ℓ_p function and the exponential function, respectively;

- "K" is the average value of cardinality (sparsity), which is the value of $\|x\|_0$, of the sparse solutions generated by Algorithm 1 for the 30 instances;

- "obj" denotes the average objective function values $f(x)$ of the sparse solutions generated by Algorithm 1 for the 30 instances;

- "$\mathrm{iter_a}$" and "$\mathrm{iter_s}$" denote the average number of iterations of Algorithm 1 and the average number of inner iterations in solving the subproblem for the 30 instances, respectively;

- "$\mathrm{time_a}$" and "$\mathrm{time_s}$" denote the average CPU time of Algorithm 1 and the average computing time in solving the subproblem at each iteration for the 30 instances, respectively.

Table 1: Average numerical results for the 30 instances with $n = 200$

SCA-DC						SCA-ℓ_p						SCA-exp					
K	obj	iter_a	time_a	iter_s	time_s	K	obj	iter_a	time_a	iter_s	time_s	K	obj	iter_a	time_a	iter_s	time_s
30.73	45.44	34	5.68	11	0.16	30.43	41.75	54	8.44	10	0.16	30.23	41.64	30	4.65	10	0.15
39.93	36.15	30	4.72	11	0.16	40.50	33.53	59	9.19	10	0.15	40.33	33.59	49	7.54	10	0.15
50.57	30.03	31	4.93	11	0.16	49.67	29.08	49	7.70	10	0.16	50.17	29.23	58	8.89	10	0.15
60.63	26.23	32	5.05	11	0.16	59.43	25.82	49	7.55	9	0.15	60.20	26.82	84	12.22	10	0.15
70.87	23.80	32	5.90	11	0.19	70.50	23.32	50	7.68	9	0.15	70.43	25.77	86	12.88	11	0.15
80.70	22.20	35	5.97	11	0.17	80.93	21.69	59	9.07	9	0.15	80.30	24.14	49	7.53	10	0.15
90.37	20.83	37	5.86	10	0.16	91.20	20.47	56	8.66	9	0.15	88.93	21.74	55	8.80	10	0.16
99.90	19.90	40	5.43	10	0.14	100.90	19.58	55	8.47	9	0.15	98.10	20.27	68	10.41	10	0.15
111.33	23.67	44	6.78	10	0.16	109.90	18.89	58	8.91	9	0.15	110.03	19.17	84	12.61	11	0.15
120.60	21.08	42	6.43	10	0.15	119.77	18.23	49	7.59	9	0.15	121.13	18.63	106	18.75	11	0.17
130.83	18.43	46	7.58	10	0.17	129.57	17.75	45	6.92	9	0.15	130.80	18.29	104	15.62	11	0.15
140.83	17.72	47	7.20	10	0.15	140.83	17.35	42	6.48	9	0.16	140.47	17.96	93	13.79	11	0.15
149.57	17.25	50	6.65	10	0.13	149.10	17.11	47	7.28	9	0.16	149.60	17.53	72	10.71	10	0.15

We see from Tables 1 and 2 that all the three versions of Algorithm 1 enjoys fast convergence with average number of iterations (subproblems solved during the algorithm) less than 100 for almost all instances. As a result, the average computation time of Algorithm 1 is no more than 20 seconds for all the instances. Among the three versions of Algorithm 1, SCA-DC uses the least CPU time and the smallest number of iterations. We also observe from Tables 1 and 2 that the average number of inner iterations in solving the subproblem is around 10 and the average computing time for each subproblem is less than half a second for all instances.

Figure 1 further illustrates the performance of Algorithm 1 using the three different functions for $\psi(x, p)$. In particular, we plot in Figure 1 the average objective value of $f(x)$ and the average computing time versus cardinality $K := \|x\|_0$ of the sparse solutions generated by Algorithm 1. It appears that the solutions generated by SCA-ℓ_p has the best trade-off between the objective value and the cardinality (sparsity), while the performance of SCA-DC is somehow between SCA-ℓ_p and SCA-exp. More precisely, we observe that for test problems with $n = 200$, SCA-exp and SCA-ℓ_p have a similar objective value-cardinality curve with SCA-ℓ_p slightly better than SCA-exp, while SCA-DC is less efficient in generating sparse solutions with smaller cardinality ($K \leq 60$). From the subfigures for $n = 300$, we also observe that SCA-DC is more efficient than SCA-exp for generating sparse solutions with smaller objective values in less computing time when the cardinality K is larger than 70.

$\boxed{5}$ Conclusions

We have presented a successive convex approximation (SCA) method for finding sparse solutions of general convex programs. Instead of tackling the cardinality constrained formulation directly, we focus on the ℓ_0 regularization form of the problem which is in some sense equivalent to the cardinality constrained problem. We have also derived a piecewise linear DC approximation formulation to the regularization problem using an ℓ_1 exact penalty function approach. We have investigated the computational performance of the proposed method us-

Table 2: Average numerical results for the 30 instances with $n = 300$

SCA-DC						SCA-ℓ_p						SCA-exp					
K	obj	iter_a	time_a	iter_s	time_s	K	obj	iter_a	time_a	iter_s	time_s	K	obj	iter_a	time_a	iter_s	time_s
30.27	65.06	33	5.88	12	0.18	30.43	59.85	56	8.80	11	0.16	30.30	59.31	35	5.23	10	0.15
40.97	49.26	35	6.11	11	0.17	40.87	46.69	60	9.54	10	0.16	40.03	47.08	48	7.18	10	0.15
50.57	41.82	30	5.31	12	0.18	49.27	40.32	56	8.92	11	0.16	50.50	39.73	64	9.53	10	0.15
60.60	36.40	31	5.53	11	0.18	60.63	34.51	56	8.89	10	0.16	60.53	35.62	86	12.82	10	0.15
70.47	32.70	31	5.59	11	0.18	69.83	31.29	57	8.88	10	0.16	70.30	33.20	93	14.18	10	0.15
79.87	29.88	31	5.44	11	0.18	80.37	28.52	60	9.35	10	0.16	79.77	31.89	92	14.10	11	0.15
89.97	27.52	32	5.59	11	0.18	89.17	26.70	59	9.28	10	0.16	90.73	31.12	97	16.89	11	0.17
99.67	25.93	32	5.69	11	0.17	99.53	25.05	58	9.01	10	0.16	100.03	30.61	96	18.08	11	0.18
110.93	24.83	33	5.74	11	0.17	109.70	23.74	55	8.54	9	0.16	110.30	30.04	86	14.68	11	0.17
120.43	23.44	35	6.07	11	0.17	119.57	22.70	56	8.62	9	0.15	120.33	28.90	63	14.38	10	0.27
130.67	23.93	37	6.37	11	0.17	130.90	21.70	64	9.85	9	0.15	130.77	25.39	54	9.85	10	0.17
140.33	22.65	38	5.95	11	0.16	140.03	21.02	61	9.49	9	0.15	140.43	22.99	69	10.31	10	0.15
150.27	21.70	40	6.20	10	0.15	149.77	20.40	62	9.62	9	0.15	149.40	21.62	87	13.00	10	0.15

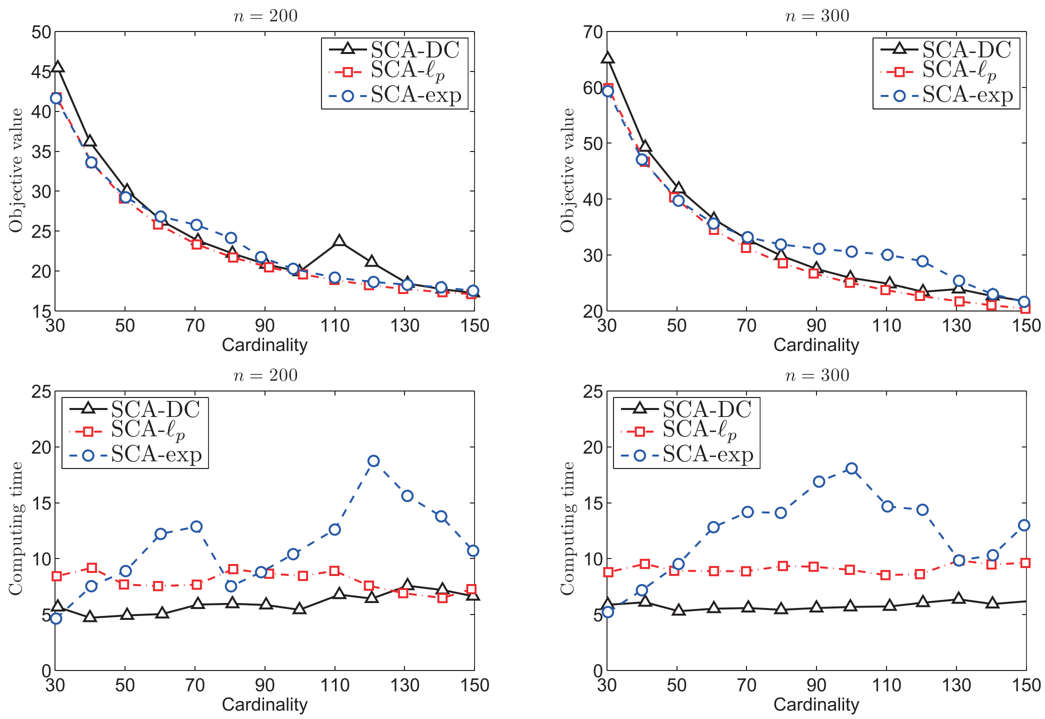

Figure 1: Average objective value and computing time (seconds) versus cardinality of sparse solutions generated by Algorithm 1 for test problems with $n = 200$ and 300

ing three different nonconvex approximations to $\|x\|_0$. Our preliminary numerical seems to suggest that the successive convex approximation method is promising in generating sparse solutions of general convex programs.

It is worth pointing out that the convex subproblem ($\mathrm{CP}_{p,\mu}(x^k)$) in Algorithm 1 can be regarded as a "weighted" ℓ_1 regularization problem when the DC function $\psi(x,p)$ reduces to a concave function. Indeed, this is the case in our computational experiment for test problem (MV), where the regularization term in ($\mathrm{CP}_{p,\mu}(x^k)$) subproblems is a linear function when using the piecewise linear DC function, the ℓ_p function and the exponential function. Therefore, the successive convex approximation method (Algorithm 1) essentially solves weighted ℓ_1 regularization subproblems iteratively with the weight parameters updated successively according to the slope of a nonconvex approximation function to $\|x\|_0$ at the iteration points.

References

[1] T.S. Arthanari and Y. Dodge, *Mathematical Programming in Statistics*, John Wiley & Sons, New York, 1993.

[2] D.P. Bertsekas, A. Nedić and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, 2003.

[3] D. Bertsimas and R. Shioda, Algorithm for cardinality-constrained quadratic optimization, *Computational Optimization and Applications* 43 (2009) 1–22.

[4] D. Bienstock, Computational study of a family of mixed-integer quadratic programming problems, *Mathematical Programming* 74 (1996) 121–140.

[5] P. Bonami and M.A. Lejeune, An exact solution approach for portfolio optimization problems under stochastic and integer constraints, *Operations Research* 57 (2009) 650–670.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004,

[7] A.M. Bruckstein, D.L. Donoho and M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, *SIAM Review* 51 (2009) 34–81.

[8] T.J. Chang, N. Meade, J.E. Beasley and Y.M. Sharaiha, Heuristics for cardinality constrained portfolio optimisation, *Computers & Operations Research* 27 (2000) 1271–1302.

[9] X. Chen, D. Ge, Z. Wang and Y. Ye, Complexity of unconstrained l_2-l_p minimization, *Mathematical Programming* 2013, DOI: 10.1007/s10107-012-0613-0.

[10] X. Chen, F. Xu and Y. Ye, Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization, *SIAM Journal on Scientific Computing* 32 (2010) 2832–2852.

[11] X. Chen and W. Zhou, Convergence of reweighted l_1 minimization algorithms and unique solution of truncated l_p minimization, Technical report, Hong Kong Polytechnic University, 2010. Available at: http://www.polyu.edu.hk/ama/staff/xjchen/IRL1-4-8.pdf.

[12] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[13] T.F. Coleman, Y.Y. Li and J. Henniger, Minimizing tracking error while restricting the number of assets, *Journal of Risk* 8 (2006) 33–55.

[14] X. T. Cui, X.J. Zheng, S.S. Zhu and X.L. Sun, Convex relaxations and MIQCQP reformulations for a class of cardinality-constrained portfolio selection problems, *Journal of Global Optimization* 56 (2013) 1409–1423.

[15] A. Frangioni and C. Gentile, Perspective cuts for a class of convex 0–1 mixed integer programs, *Mathematical Programming* 106 (2006) 225–236.

[16] A. Frangioni and C. Gentile, SDP diagonalizations and perspective cuts for a class of nonseparable MIQP, *Operations Research Letters* 35 (2007) 181–185.

[17] J.J. Gao and D. Li, Optimal cardinality constrained portfolio selection, *Operations Research* 61 (2013) 745–761.

[18] D. Ge, X. Jiang and Y. Ye, A note on complexity of ℓ_p minimization, *Mathematical Programming* 129 (2011) 285–299.

[19] L.J. Hong, Y. Yang and L.W. Zhang, Sequential convex approximations to joint chance constrained programs: A Monte Carlo approach, *Operations Research* 59 (2011) 617–630.

[20] R. Horst and Ng V. Thoai, DC programming: Overview, *Journal of Optimization Theory and Applications* 103 (1999) 1–43.

[21] IBM ILOG CPLEX, *IBM ILOG CPLEX 12. 3 User's Manual for CPLEX*, 2011.

[22] R. Jansen and R. Van Dijk, Optimal benchmark tracking with small portfolios, *The Journal of Portfolio Management* 28 (2002) 33–39.

[23] S. Jokar and M.E. Pfetsch, Exact and approximate sparse solutions of underdetermined linear equations, *SIAM Journal on Scientific Computing* 31 (2008) 23–44.

[24] D. Li, X.L. Sun and J. Wang, Optimal lot solution to cardinality constrained mean-variance formulation for portfolio selection, *Mathematical Finance* 16 (2006) 83–101.

[25] O.L. Mangasarian, Minimum-support solutions of polyhedral concave programs, *Optimization* 45 (1999) 149–162.

[26] A.J. Miller, *Subset Selection in Regression*, Second edition, Chapman and Hall, 2002.

[27] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer Verlag, New York , 1999.

[28] T.D. Quoc and M. Diehl, Sequential convex programming methods for solving nonlinear optimization problems with DC constraints, Technical report, 2011, arXiv preprint arXiv:1107.5841.

[29] D.X. Shaw, S. Liu and L. Kopman, Lagrangian relaxation procedure for cardinality-constrained portfolio optimization, *Optimization Methods & Software* 23 (2008) 411–420.

[30] J.P. Vielma, S. Ahmed and G.L. Nemhauser, A lifted linear programming branch-and-bound algorithm for mixed integer conic quadratic programs, *INFORMS Journal on Computing* 20 (2008) 438–450.

[31] M. Woodside-Oriakhi, C. Lucas and J.E. Beasley, Heuristic algorithms for the cardinality constrained efficient frontier, *European Journal of Operational Research* 213 (2011) 538–550.

[32] X.J. Zheng, X.L. Sun, D. Li and J. Sun, Successive convex approximations to cardinality-constrained convex programs: A piecewise-linear DC approach, *Computational Optimization and Applications* 2013, DOI: 10.1007/s10589-013-9582-3.

Manuscript received 14 April 2013
revised 2 September 2013
accepted for publication 13 November 2013

XIAODI BAI
Department of Management Science, School of Management
Fudan University, Shanghai 200433, P. R. China
E-mail address: xdbai@fudan.edu.cn

XIAOJIN ZHENG
School of Economics and Management, Tongji University
Shanghai 200092, P. R. China
E-mail address: xjzheng@tongji.edu.cn

XUETING CUI
Department of Applied Mathematics
Shanghai University of Finance and Economics
Shanghai 200433, P. R. China
E-mail address: cui.xueting@shufe.edu.cn

XIAOLING SUN
Department of Management Science, School of Management
Fudan University, Shanghai 200433, P. R. China
E-mail address: xls@fudan.edu.cn