Math. Program., Ser. A (2012) 135:149–193
DOI 10.1007/s10107-011-0452-4

FULL LENGTH PAPER

An augmented Lagrangian approach for sparse principal component analysis

Zhaosong Lu · Yong Zhang

Received: 12 July 2009 / Accepted: 28 February 2011 / Published online: 27 April 2011

Abstract Principal component analysis (PCA) is a widely used technique for data analysis and dimension reduction with numerous applications in science and engineering. However, the standard PCA suffers from the fact that the principal components (PCs) are usually linear combinations of all the original variables, and it is thus often difficult to interpret the PCs. To alleviate this drawback, various sparse PCA approaches were proposed in the literature (Cadima and Jolliffe in J Appl Stat 22:203–214, 1995; d'Aspremont et al. in J Mach Learn Res 9:1269–1294, 2008; d'Aspremont et al. SIAM Rev 49:434–448, 2007; Jolliffe in J Appl Stat 22:29–35, 1995; Journée et al. in J Mach Learn Res 11:517–553, 2010; Jolliffe et al. in J Comput Graph Stat 12:531–547, 2003; Moghaddam et al. in Advances in neural information processing systems 18:915–922, MIT Press, Cambridge, 2006; Shen and Huang in J Multivar Anal 99(6):1015–1034, 2008; Zou et al. in J Comput Graph Stat 15(2):265–286, 2006). Despite success in achieving sparsity, some important properties enjoyed by the standard PCA are lost in these methods such as uncorrelation of PCs and orthogonality of loading vectors. Also, the total explained variance that they attempt to maximize can be too optimistic. In this paper we propose a new formulation for sparse PCA, aiming at finding sparse and nearly uncorrelated PCs with orthogonal loading vectors while explaining as much of the total variance as possible. We also develop a novel augmented Lagrangian method for solving a class of nonsmooth constrained optimization problems, which is well suited for our formulation of sparse PCA. We show

This work was supported in part by NSERC Discovery Grant.

Z. Lu (✉) · Y. Zhang
Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
e-mail: zhaosong@sfu.ca

Y. Zhang
e-mail: yza30@sfu.ca

that it converges to a *feasible* point, and moreover under some regularity assumptions, it converges to a stationary point. Additionally, we propose two nonmonotone gradient methods for solving the augmented Lagrangian subproblems, and establish their global and local convergence. Finally, we compare our sparse PCA approach with several existing methods on synthetic (Zou et al. in J Comput Graph Stat 15(2):265–286, 2006), Pitprops (Jeffers in Appl Stat 16:225–236, 1967), and gene expression data (Chin et al in Cancer Cell 10:529C–541C, 2006), respectively. The computational results demonstrate that the sparse PCs produced by our approach substantially outperform those by other methods in terms of total explained variance, correlation of PCs, and orthogonality of loading vectors. Moreover, the experiments on random data show that our method is capable of solving large-scale problems within a reasonable amount of time.

Keywords Sparse PCA · Augmented Lagrangian method · Nonmonotone gradient methods · Nonsmooth minimization

Mathematics Subject Classification (2000) 62H20 · 62H25 · 62H30 · 90C30 · 65K05

1 Introduction

Principal component analysis (PCA) is a popular tool for data processing and dimension reduction. It has been widely used in numerous applications in science and engineering such as biology, chemistry, image processing, machine learning and so on. For example, PCA has recently been applied to human face recognition, handwritten zip code classification and gene expression data analysis (see [1,11–13]).

In essence, PCA aims at finding a few linear combinations of the original variables, called *principal components* (PCs), which point in orthogonal directions capturing as much of the variance of the variables as possible. It is well known that PCs can be found via the eigenvalue decomposition of the covariance matrix Σ. However, Σ is typically unknown in practice. Instead, the PCs can be approximately computed via the singular value decomposition (SVD) of the data matrix or the eigenvalue decomposition of the sample covariance matrix. In detail, let $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})$ be a p-dimensional random vector, and X be an $n \times p$ data matrix, which records the n observations of ξ. Without loss of generality, assume X is centered, that is, the column means of X are all 0. Then the commonly used sample covariance matrix is $\hat{\Sigma} = X^T X/(n-1)$. Suppose the eigenvalue decomposition of $\hat{\Sigma}$ is

$$\hat{\Sigma} = V D V^T.$$

Then $\eta = \xi V$ gives the PCs, and the columns of V are the corresponding loading vectors. It is worth noting that V can also be obtained by performing the SVD of X (see, for example, [31]). Clearly, the columns of V are orthonormal vectors, and moreover $V^T \hat{\Sigma} V$ is diagonal. We thus immediately see that if $\hat{\Sigma} = \Sigma$, the corresponding PCs are uncorrelated; otherwise, they can be correlated with each other (see Sect. 2

for details). We now describe several important properties of the PCs obtained by the standard PCA when Σ is well estimated by $\hat{\Sigma}$ (see also [31]):

1. The PCs sequentially capture the maximum variance of the variables approximately, thus encouraging minimal information loss as much as possible;
2. The PCs are nearly uncorrelated, so the explained variance by different PCs has small overlap;
3. The PCs point in orthogonal directions, that is, their loading vectors are orthogonal to each other.

In practice, typically the first few PCs are enough to represent the data, thus a great dimensionality reduction is achieved. In spite of the popularity and success of PCA due to these nice features, PCA has an obvious drawback, that is, PCs are usually linear combinations of all p variables and the loadings are typically nonzero. This makes it often difficult to interpret the PCs, especially when p is large. Indeed, in many applications, the original variables have concrete physical meaning. For example in biology, each variable might represent the expression level of a gene. In these cases, the interpretation of PCs would be facilitated if they were composed only from a small number of the original variables, namely, each PC involved a small number of nonzero loadings. It is thus imperative to develop sparse PCA techniques for finding the PCs with sparse loadings while enjoying the above three nice properties as much as possible.

Sparse PCA has been an active research topic for more than a decade. The first class of approaches are based on ad-hoc methods by post-processing the PCs obtained from the standard PCA mentioned above. For example, Jolliffe [17] applied various rotation techniques to the standard PCs for obtaining sparse loading vectors. Cadima and Jolliffe [7] proposed a simple thresholding approach by artificially setting to zero the standard PCs' loadings with absolute values smaller than a threshold. In recent years, optimization approaches have been proposed for finding sparse PCs. They usually formulate sparse PCA into an optimization problem, aiming at achieving the sparsity of loadings while maximizing the explained variance as much as possible. For instance, Jolliffe et al. [19] proposed an interesting algorithm, called SCoTLASS, for finding sparse orthogonal loading vectors by sequentially maximizing the approximate variance explained by each PC under the l_1-norm penalty on loading vectors. Zou et al. [31] formulated sparse PCA as a regression-type optimization problem and imposed a combination of l_1- and l_2-norm penalties on the regression coefficients. d'Aspremont et al. [10] proposed a method, called DSPCA, for finding sparse PCs by solving a sequence of semidefinite program relaxations of sparse PCA. Shen and Huang [28] recently developed an approach for computing sparse PCs by solving a sequence of rank-one matrix approximation problems under several sparsity-inducing penalties. Very recently, Journée et al. [18] formulated sparse PCA as nonconcave maximization problems with l_0- or l_1-norm sparsity-inducing penalties. They showed that these problems can be reduced into maximization of a convex function on a compact set, and they also proposed a simple but computationally efficient gradient method for finding a stationary point of the latter problems. Additionally, greedy methods were investigated for sparse PCA by Moghaddam et al. [21] and d'Aspremont et al. [9].

The PCs obtained by the above methods [7,9,10,17–19,21,28,31] are usually sparse. However, the aforementioned nice properties of the standard PCs are lost to some extent in these sparse PCs. Indeed, the likely correlation among the sparse PCs are not considered in these methods. Therefore, their sparse PCs can be quite correlated with each other. Also, the total explained variance that these methods attempt to maximize can be too optimistic as there may be some overlap among the individual variances of sparse PCs. Finally, the loading vectors of the sparse PCs given by these methods lack orthogonality except SCoTLASS [19].

In this paper we propose a new formulation for sparse PCA by taking into account the three nice properties of the standard PCA, that is, maximal total explained variance, uncorrelation of PCs, and orthogonality of loading vectors. We also explore the connection of this formulation with the standard PCA and show that it can be viewed as a certain perturbation of the standard PCA. We further propose a novel augmented Lagrangian method for solving a class of nonsmooth constrained optimization problems, which is well suited for our formulation of sparse PCA. This method differs from the classical augmented Lagrangian method in that: i) the values of the augmented Lagrangian functions at their approximate minimizers given by the method are bounded from above; and ii) the magnitude of penalty parameters outgrows that of Lagrangian multipliers (see Sect. 3.2 for details). We show that this method converges to a *feasible* point, and moreover it converges to a first-order stationary point under some regularity assumptions. (We should mention that the aforementioned two novel properties of our augmented Lagrangian method are crucial in ensuring convergence both theoretically and practically. In fact, we observed in our experiments that when one or both of these properties are dropped, the resulting method (e.g., the classical augmented Lagrangian method) almost always fails to converge to even a feasible point as applied to our formulation of sparse PCA.) We also propose two nonmonotone gradient methods for minimizing a class of nonsmooth functions over a closed convex set, which can be suitably applied to the subproblems arising in our augmented Lagrangian method. We further establish global convergence and, under a local Lipschitzian error bounds assumption [29], local linear rate of convergence for these gradient methods. Finally, we compare the sparse PCA approach proposed in this paper with several existing methods [10,18,28,31] on synthetic [31], Pitprops [16], and gene expression data [8], respectively. The computational results demonstrate that the sparse PCs obtained by our approach substantially outperform those by the other methods in terms of total explained variance, correlation of PCs, and orthogonality of loading vectors. In addition, the experiments on random data show that our method is capable of solving large-scale problems within very reasonable amount of time.

The rest of paper is organized as follows. In Sect. 2, we propose a new formulation for sparse PCA and explore the connection of this formulation with the standard PCA. In Sect. 3, we then develop a novel augmented Lagrangian method for a class of nonsmooth constrained problems, and propose two nonmonotone gradient methods for minimizing a class of nonsmooth functions over a closed convex set. In Sect. 4, we discuss the applicability and implementation details of our augmented Lagrangian method for sparse PCA. The sparse PCA approach proposed in this paper is then compared with several existing methods on synthetic [31], Pitprops [16], and gene expression data [8] in Sect. 5. Finally, we present some concluding remarks in Sect. 6.

1.1 Notation

In this paper, all vector spaces are assumed to be finite dimensional. The symbols $\Re^n$ and $\Re^n_+$ (resp., $\Re^n_-$) denote the n-dimensional Euclidean space and the nonnegative (resp., nonpositive) orthant of $\Re^n$, respectively, and $\Re_{++}$ denotes the set of positive real numbers. The space of all $m \times n$ matrices with real entries is denoted by $\Re^{m\times n}$. The space of symmetric $n \times n$ matrices is denoted by $\mathcal{S}^n$. Additionally, $\mathcal{D}^n$ denotes the space of $n \times n$ diagonal matrices. For a real matrix X, we denote by $|X|$ the absolute value of X, that is, $|X|_{ij} = |X_{ij}|$ for all ij, and by $\mathrm{sign}(X)$ the sign of X whose ijth entry equals the sign of X_{ij} for all ij. Also, the nonnegative part of X is denoted by $[X]^+$ whose ijth entry is given by $\max\{0, X_{ij}\}$ for all ij. The rank of X is denoted by $\mathrm{rank}(X)$. Further, the identity matrix and the all-ones matrix are denoted by I and E, respectively, whose dimension should be clear from the context. If $X \in \mathcal{S}^n$ is positive semidefinite, we write $X \succeq 0$. For any $X, Y \in \mathcal{S}^n$, we write $X \preceq Y$ to mean $Y - X \succeq 0$. Given matrices X and Y in $\Re^{m\times n}$, the standard inner product is defined by $X \bullet Y := \mathrm{Tr}(XY^T)$, where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix, and the component-wise product is denoted by $X \odot Y$, whose ijth entry is $X_{ij}Y_{ij}$ for all ij. $\|\cdot\|$ denotes the Euclidean norm and its associated operator norm unless it is explicitly stated otherwise. The minimal (resp., maximal) eigenvalue of an $n \times n$ symmetric matrix X are denoted by $\lambda_{\min}(X)$ (resp., $\lambda_{\max}(X)$), respectively, and $\lambda_i(X)$ denotes its ith largest eigenvalue for $i = 1, \ldots, n$. Given a vector $v \in \Re^n$, $\mathrm{Diag}(v)$ or $\mathrm{Diag}(v_1, \ldots, v_n)$ denotes a diagonal matrix whose ith diagonal element is v_i for $i = 1, \ldots, n$. Given an $n \times n$ matrix X, $\widetilde{\mathrm{Diag}}(X)$ denotes a diagonal matrix whose ith diagonal element is X_{ii} for $i = 1, \ldots, n$. Let $\mathcal{U}$ be a real vector space. Given a closed convex set $C \subseteq \mathcal{U}$, let $\mathrm{dist}(\cdot, C) : \mathcal{U} \to \Re_+$ denote the distance function to C measured in terms of $\|\cdot\|$, that is,

$$\mathrm{dist}(u, C) := \inf_{\tilde{u}\in C} \|u - \tilde{u}\| \quad \forall u \in \mathcal{U}. \tag{1}$$

2 Formulation for sparse PCA

In this section we propose a new formulation for sparse PCA by taking into account sparsity and orthogonality of loading vectors, and uncorrelation of PCs. We also address the connection of our formulation with the standard PCA.

Let $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})$ be a p-dimensional random vector with covariance matrix Σ. Suppose X is an $n \times p$ data matrix, which records the n observations of ξ. Without loss of generality, assume the column means of X are 0. Then the commonly used sample covariance matrix of ξ is $\hat{\Sigma} = X^T X/(n-1)$. For any r loading vectors represented as $V = [V_1, \ldots, V_r] \in \Re^{p\times r}$ where $1 \le r \le p$, the corresponding components are given by $\eta = (\eta^{(1)}, \ldots, \eta^{(r)}) = \xi V$, which are linear combinations of $\xi^{(1)}, \ldots, \xi^{(p)}$. Clearly, the covariance matrix of η is $V^T \Sigma V$, and thus the components $\eta^{(i)}$ and $\eta^{(j)}$ are uncorrelated if and only if the ijth entry of $V^T \Sigma V$ is zero. Also, the total explained variance by the components $\eta^{(i)}$'s equals, if they are uncorrelated, the

sum of the individual variances of $\eta^{(i)}$'s, that is,

$$\sum_{i=1}^{r} V_i^T \Sigma V_i = \mathrm{Tr}(V^T \Sigma V).$$

Recall that our aim is to find a set of sparse and orthogonal loading vectors V so that the corresponding components $\eta^{(1)}, \ldots, \eta^{(r)}$ are uncorrelated and explain as much variance of the original variables $\xi^{(1)}, \ldots, \xi^{(p)}$ as possible. It appears that our goal can be achieved by solving the following problem:

$$\begin{array}{ll} \max\limits_{V \in \Re^{n\times r}} & \mathrm{Tr}(V^T \Sigma V) - \rho \bullet |V| \\ \text{s.t.} & V^T \Sigma V \text{ is diagonal}, \\ & V^T V = I, \end{array} \tag{2}$$

where $\rho \in \Re_+^{p\times r}$ is a tunning parameter for controlling the sparsity of V. However, the covariance matrix Σ is typically unknown and can only be approximated by the sample covariance matrix $\hat{\Sigma}$. It looks plausible to modify (2) by simply replacing Σ with $\hat{\Sigma}$ at a glance. Nevertheless, such a modification would eliminate all optimal solutions V^* of (2) from consideration since $(V^*)^T \hat{\Sigma} V^*$ is generally non-diagonal. For this reason, given a sample covariance $\hat{\Sigma}$, we consider the following formulation for sparse PCA, which can be viewed as a modification of problem (2),

$$\begin{array}{ll} \max\limits_{V \in \Re^{n\times r}} & \mathrm{Tr}(V^T \hat{\Sigma} V) - \rho \bullet |V| \\ \text{s.t.} & |V_i^T \hat{\Sigma} V_j| \le \Delta_{ij} \quad \forall i \ne j, \\ & V^T V = I, \end{array} \tag{3}$$

where $\Delta_{ij} \ge 0$ $(i \ne j)$ are the parameters for controlling the correlation of the components corresponding to V. Clearly, $\Delta_{ij} = \Delta_{ji}$ for all $i \ne j$.

We next explore the connection of formulation (3) with the standard PCA. Before proceeding, we first establish a technical lemma that will be used subsequently.

Lemma 2.1 *Given any $\hat{\Sigma} \in \mathcal{S}^n$ and integer $1 \le r \le n$, consider the problem*

$$\max\{\mathrm{Tr}(V^T \hat{\Sigma} V) |\ V^T V = I,\ V \in \Re^{n\times r}\}. \tag{4}$$

The following statements hold:

(a) *The optimal value of (4) is $\sum_{i=1}^{r} \lambda_i(\hat{\Sigma})$;*
(b) *V^* is an optimal solution of (4) if and only if $V^* = SU^*Q$, where U^* is an $n \times r$ matrix whose columns consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to r largest eigenvalues of $\hat{\Sigma}$, and S and Q are arbitrary $n \times n$ and $r \times r$ orthogonal matrices with $S^T \hat{\Sigma} S = \hat{\Sigma}$.*

Proof The statement (a) holds due to equation (3.20) of Chapter 1 of [14]. We now show statement (b) also holds. Let $V^* = SU^*Q$, where S, U^* and Q are defined above.

It is straightforward to verify that $V^{*T}V^* = I$ and $\mathrm{Tr}(V^{*T}\hat{\Sigma}V^*) = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$, which together with statement (a) implies that V^* is an optimal solution of (4) and hence, the "if" part of statement (b) holds. To prove the "only if" part, suppose that V^* is an optimal solution of (4). We then observe that V^* is a critical point of the generalized Rayleigh quotient $\mathrm{Tr}(V^T\hat{\Sigma}V)$ over the Stiefel manifold $\{X \in \Re^{n\times r} : X^TX = I\}$. It follows from Theorem 3.17 of Chapter 1 of [14] that $V^* = SU^*Q$ for some S, U^* and Q that are defined above and hence the "only if" part of statement (b) holds. □

We next address the relation between the eigenvectors of $\hat{\Sigma}$ and the solutions of problem (3) when $\rho = 0$ and $\Delta_{ij} = 0$ for all $i \neq j$.

Theorem 2.2 *Suppose for problem* (3) *that* $\rho = 0$ *and* $\Delta_{ij} = 0$ *for all* $i \neq j$. *Let* f^* *be the optimal value of* (3). *Then,* $f^* = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$, *and* $V^* \in \Re^{n\times r}$ *is an optimal solution of* (3) *if and only if the columns of* V^* *consist of the orthonormal eigenvectors of* $\hat{\Sigma}$ *corresponding to* r *largest eigenvalues of* $\hat{\Sigma}$.

Proof We first show that $f^* = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$. Indeed, let U be an $n \times r$ matrix whose columns consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to r largest eigenvalues of $\hat{\Sigma}$. We then see that U is a feasible solution of (3) and $\mathrm{Tr}(U^T\hat{\Sigma}U) = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$. It follows that $f^* \geq \sum_{i=1}^r \lambda_i(\hat{\Sigma})$. On the other hand, we observe that f^* is bounded above by the optimal value of problem (4), which together with Lemma 2.1(a) yields $f^* \leq \sum_{i=1}^r \lambda_i(\hat{\Sigma})$. Thus, $f^* = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$ and U is an optimal solution of (3), which implies that the "if" part of this theorem holds. We next show that the "only if" part also holds. Suppose that V^* is an optimal solution of (3). Since problems (3) and (4) share the same optimal value, V^* is also an optimal solution of (4). It then follows from Lemma 2.1(b) that $V^* = SU^*Q$ for some $S \in \Re^{n\times n}$ and $Q \in \Re^{r\times r}$ satisfying $S^TS = I$, $S^T\hat{\Sigma}S = \hat{\Sigma}$, $Q^TQ = I$, and $U^* \in \Re^{n\times r}$ whose columns consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to r largest eigenvalues of $\hat{\Sigma}$. Clearly, $\hat{\Sigma}U^* = U^*\Lambda$ and $U^{*T}\hat{\Sigma}U^* = \Lambda$, where Λ is an $r \times r$ diagonal matrix whose diagonal consists of r largest eigenvalues of $\hat{\Sigma}$. Letting $D = V^{*T}\hat{\Sigma}V^*$ and using the above relations, we obtain that

$$D = V^{*T}\hat{\Sigma}V^* = Q^TU^{*T}(S^T\hat{\Sigma}S)U^*Q = Q^T(U^{*T}\hat{\Sigma}U^*)Q = Q^T\Lambda Q, \qquad (5)$$

which together with the relation $Q^TQ = I$, implies that D is similar to the diagonal matrix Λ. In addition, by the definition of V^*, we know that D is an $r \times r$ diagonal matrix. Thus, D and Λ share the same diagonal entries upon some permutations if necessary. It then follows that the diagonal of D consists of r largest eigenvalues of $\hat{\Sigma}$. Since $S^TS = I$ and $S^T\hat{\Sigma}S = \hat{\Sigma}$, we see that $\hat{\Sigma}S = S\hat{\Sigma}$. Using this equality along with (5) and the relations $V^* = SU^*Q$, $Q^TQ = I$, $\hat{\Sigma}U^* = U^*\Lambda$, we have

$$\hat{\Sigma}V^* = \hat{\Sigma}SU^*Q = S\hat{\Sigma}U^*Q = SU^*\Lambda Q = (SU^*Q)(Q^T\Lambda Q) = V^*D.$$

It follows that the columns of V^* consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to r largest eigenvalues of $\hat{\Sigma}$, and hence the "only if" part of this theorem holds. □

From the above theorem, we see that when $\rho = 0$ and $\Delta_{ij} = 0$ for all $i \neq j$, each solution of (3) consists of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to r largest eigenvalues of $\hat{\Sigma}$, which can be computed from the eigenvalue decomposition of $\hat{\Sigma}$. Therefore, the loading vectors obtained from (3) are the same as those given by the standard PCA when applied to $\hat{\Sigma}$. On the other hand, when ρ and Δ_{ij} for all $i \neq j$ are small, the loading vectors found by (3) can be viewed as an approximation to the ones provided by the standard PCA. We will propose suitable methods for solving (3) in Sects. 3 and 4.

3 Augmented Lagrangian method for nonsmooth constrained nonlinear programming

In this section we propose a novel augmented Lagrangian method for a class of nonsmooth constrained nonlinear programming problems, which is well suited for formulation (3) of sparse PCA. In particular, we study first-order optimality conditions in Sect. 3.1. In Sect. 3.2, we develop an augmented Lagrangian method and establish its global convergence. In Sect. 3.3, we propose two nonmonotone gradient methods for minimizing a class of nonsmooth functions over a closed convex set, which can be suitably applied to the subproblems arising in our augmented Lagrangian method. We also establish global and local convergence for these gradient methods.

3.1 First-order optimality conditions

In this subsection we introduce a class of nonsmooth constrained nonlinear programming problems and study first-order optimality conditions for them.

Consider the nonlinear programming problem

$$\begin{array}{ll} \min & f(x) + P(x) \\ \text{s.t.} & g_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \\ & x \in X. \end{array} \tag{6}$$

We assume that the functions $f : \Re^n \to \Re$, $g_i : \Re^n \to \Re$, $i = 1, \ldots, m$, and $h_i : \Re^n \to \Re$, $i = 1, \ldots, p$, are continuously differentiable, and that the function $P : \Re^n \to \Re$ is convex but not necessarily smooth, and that the set $X \subseteq \Re^n$ is closed and convex. For convenience of the subsequent presentation, we denote by Ω the feasible region of problem (6).

For the case where P is a smooth function, the first-order optimality conditions for problem (6) have been well studied in literature (see, for example, Theorem 3.25 of [27]), but there is little study when P is a nonsmooth convex function. We next aim to establish first-order optimality conditions for problem (6). Before proceeding, we describe a general constraint qualification condition for (6), that is, Robinson's condition that was proposed in [24].

Let $x \in \Re^n$ be a feasible point of problem (6). We denote the set of active inequality constraints at x as

$$\mathcal{A}(x) = \{1 \leq i \leq m : g_i(x) = 0\}.$$

In addition, x is said to satisfy *Robinson's condition* if

$$\left\{ \begin{bmatrix} g'(x)d - v \\ h'(x)d \end{bmatrix} : d \in T_X(x), v \in \Re^m, v_i \leq 0, i \in \mathcal{A}(x) \right\} = \Re^m \times \Re^p, \tag{7}$$

where $T_X(x)$ is the tangent cone to X at x, and $g'(x)$ and $h'(x)$ denote the Jacobian of the functions $g = (g_1, \ldots, g_m)$ and $h = (h_1, \ldots, h_p)$ at x, respectively. Other equivalent expressions of Robinson's condition can be found, for example, in [24,25,27].

The following result demonstrates that Robinson's condition is indeed a constraint qualification condition for problem (6), which is briefly mentioned in the proof of Theorem 3.25 of [27]. For a detailed proof of it, see [20].

Proposition 3.1 *Given a feasible point $x \in \Re^n$ of problem (6), let $T_\Omega(x)$ be the tangent cone to Ω at x and $(T_\Omega(x))^\circ$ its polar cone. If Robinson's condition (7) holds at x, then*

$$\begin{aligned} T_\Omega(x) &= \left\{ d \in T_X(x) : \begin{array}{ll} d^T \nabla g_i(x) \leq 0, & i \in \mathcal{A}(x), \\ d^T \nabla h_i(x) = 0, & i = 1, \ldots, p \end{array} \right\}, \\ (T_\Omega(x))^\circ &= \left\{ \sum_{i \in \mathcal{A}(x)} \lambda_i \nabla g_i(x) + \sum_{i=1}^{p} \mu_i \nabla h_i(x) + N_X(x) : \lambda \in \Re^m_+, \ \mu \in \Re^p \right\}, \end{aligned} \tag{8}$$

where $N_X(x)$ is the normal cone to X at x.

We are now ready to establish first-order optimality conditions for problem (6).

Theorem 3.2 *Let $x^* \in \Re^n$ be a local minimizer of problem (6) and $\partial P(x^*)$ denote the subdifferential of P at x^*. Assume that Robinson's condition (7) is satisfied at x^*. Then there exist Lagrange multipliers $\lambda \in \Re^m_+$ and $\mu \in \Re^p$ such that*

$$0 \in \nabla f(x^*) + \partial P(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{i=1}^{p} \mu_i \nabla h_i(x^*) + N_X(x^*), \tag{9}$$

and

$$\lambda_i g_i(x^*) = 0, \quad i = 1, \ldots, m. \tag{10}$$

Moreover, the set of Lagrange multipliers $(\lambda, \mu) \in \Re^m_+ \times \Re^p$ satisfying the above conditions, denoted by $\Lambda(x^*)$, is convex and compact.

Proof We first show that

$$d^T \nabla f(x^*) + P'(x^*; d) \geq 0 \quad \forall d \in T_\Omega(x^*). \tag{11}$$

Let $d \in T_\Omega(x^*)$ be arbitrarily chosen. Then, there exist sequences $\{x^k\}_{k=1}^\infty \subseteq \Omega$ and $\{t_k\}_{k=1}^\infty \subseteq \Re_{++}$ such that $t_k \downarrow 0$ and

$$d = \lim_{k\to\infty} \frac{x^k - x^*}{t_k}.$$

Thus, we have $x^k = x^* + t_k d + o(t_k)$. Using this relation along with the fact that the function f is differentiable and P is convex in $\Re^n$, we can have

$$f(x^* + t_k d) - f(x^k) = o(t_k), \quad P(x^* + t_k d) - P(x^k) = o(t_k), \tag{12}$$

where the first equality follows from the Mean Value Theorem while the second one comes from Theorem 10.4 of [26]. Clearly, $x^k \to x^*$. This together with the assumption that x^* is a local minimizer of (6), implies that

$$f(x^k) + P(x^k) \geq f(x^*) + P(x^*) \tag{13}$$

when k is sufficiently large. In view of (12) and (13), we obtain that

$$\begin{aligned} d^T \nabla f(x^*) + P'(x^*; d) &= \lim_{k\to\infty} \frac{f(x^*+t_k d) - f(x^*)}{t_k} + \lim_{k\to\infty} \frac{P(x^* + t_k d) - P(x^*)}{t_k}, \\ &= \lim_{k\to\infty} \left[\frac{f(x^k) + P(x^k) - f(x^*) - P(x^*)}{t_k} + \frac{o(t_k)}{t_k} \right], \\ &= \lim_{k\to\infty} \frac{f(x^k) + P(x^k) - f(x^*) - P(x^*)}{t_k} \geq 0, \end{aligned}$$

and hence (11) holds.

For simplicity of notations, let $T_\Omega^\circ = (T_\Omega(x^*))^\circ$ and $S = -\nabla f(x^*) - \partial P(x^*)$. We next show that $S \cap T_\Omega^\circ \neq \emptyset$. Suppose for contradiction that $S \cap T_\Omega^\circ = \emptyset$. This together with the fact that S and T_Ω° are nonempty closed convex sets and S is bounded, implies that there exists some $d \in \Re^n$ such that $d^T y \leq 0$ for any $y \in T_\Omega^\circ$, and $d^T y \geq 1$ for any $y \in S$. Clearly, we see that $d \in (T_\Omega^\circ)^\circ = T_\Omega(x^*)$, and

$$\begin{aligned} 1 \leq \inf_{y\in S} d^T y &= \inf_{z\in\partial P(x^*)} d^T(-\nabla f(x^*) - z) = -d^T \nabla f(x^*) \\ &- \sup_{z\in\partial P(x^*)} d^T z = -d^T \nabla f(x^*) - P'(x^*; d), \end{aligned}$$

which contradicts (11). Hence, we have $S \cap T_\Omega^\circ \neq \emptyset$. Using this relation, (8), the definitions of S and $\mathcal{A}(x^*)$, and letting $\lambda_i = 0$ for $i \notin \mathcal{A}(x^*)$, we easily see that (9) and (10) hold.

In view of the fact that $\partial P(x^*)$ and $N_X(x^*)$ are closed and convex, and moreover $\partial P(x^*)$ is bounded, we know that $\partial P(x^*) + N_X(x^*)$ is closed and convex. Using this result, it is straightforward to see that $\Lambda(x^*)$ is closed and convex. We next show that $\Lambda(x^*)$ is bounded. Suppose for contradiction that $\Lambda(x^*)$ is unbounded. Then, there exists a sequence $\{(\lambda^k, \mu^k)\}_{k=1}^{\infty} \subseteq \Lambda(x^*)$ such that $\|(\lambda^k, \mu^k)\| \to \infty$, and

$$0 = \nabla f(x^*) + z^k + \sum_{i=1}^{m} \lambda_i^k \nabla g_i(x^*) + \sum_{i=1}^{p} \mu_i^k \nabla h_i(x^*) + v^k \tag{14}$$

for some $\{z^k\}_{k=1}^{\infty} \subseteq \partial P(x^*)$ and $\{v^k\}_{k=1}^{\infty} \subseteq N_X(x^*)$. Let $(\bar{\lambda}^k, \bar{\mu}^k) = (\lambda^k, \mu^k)/\|(\lambda^k, \mu^k)\|$.

By passing to a subsequence if necessary, we can assume that $(\bar{\lambda}^k, \bar{\mu}^k) \to (\bar{\lambda}, \bar{\mu})$. We clearly see that $\|(\bar{\lambda}, \bar{\mu})\| = 1, \bar{\lambda} \in \Re^m_+$, and $\bar{\lambda}_i = 0$ for $i \notin \mathcal{A}(x^*)$. Note that $\partial P(x^*)$ is bounded and $N_X(x^*)$ is a closed cone. In view of this fact, and upon dividing both sides of (14) by $\|(\lambda^k, \mu^k)\|$ and taking limits on a subsequence if necessary, we obtain that

$$0 = \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(x^*) + \sum_{i=1}^{p} \bar{\mu}_i \nabla h_i(x^*) + \bar{v} \tag{15}$$

for some $\bar{v} \in N_X(x^*)$. Since Robinson's condition (7) is satisfied at x^*, there exist $d \in T_X(x^*)$ and $v \in \Re^m$ such that $v_i \leq 0$ for $i \in \mathcal{A}(x^*)$, and

$$\begin{aligned} d^T \nabla g_i(x^*) - v_i &= -\bar{\lambda}_i \quad \forall i \in \mathcal{A}(x^*), \\ d^T \nabla h_i(x^*) &= -\bar{\mu}_i, \quad i = 1, \ldots, p. \end{aligned}$$

Using these relations, (15) and the fact that $d \in T_X(x^*)$, $\bar{v} \in N_X(x^*)$, $\bar{\lambda} \in \Re^m_+$, and $\bar{\lambda}_i = 0$ for $i \notin \mathcal{A}(x^*)$, we have

$$\begin{aligned} \sum_{i=1}^{m} \bar{\lambda}_i^2 + \sum_{i=1}^{p} \bar{\mu}_i^2 &\leq -\sum_{i=1}^{m} \bar{\lambda}_i d^T \nabla g_i(x^*) - \sum_{i=1}^{p} \bar{\mu}_i d^T \nabla h_i(x^*), \\ &= -d^T \left(\sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(x^*) + \sum_{i=1}^{p} \bar{\mu}_i \nabla h_i(x^*) \right) = d^T \bar{v} \leq 0. \end{aligned}$$

It yields $(\bar{\lambda}, \bar{\mu}) = (0, 0)$, which contradicts the identity $\|(\bar{\lambda}, \bar{\mu})\| = 1$. Thus, $\Lambda(x^*)$ is bounded. □

3.2 Augmented Lagrangian method for (6)

For a convex program, it is known that under some mild assumptions, any accumulation point of the sequence generated by the classical augmented Lagrangian method is an optimal solution (e.g., see Section 6.4.3 of [27]). Nevertheless, when problem (6) is a nonconvex program, especially when the function h_i is not affine or g_i is

nonconvex, the classical augmented Lagrangian method may not even converge to a feasible point, that is, any accumulation point of the sequence generated by the method may violate some constraints of (6). We actually observed in our experiments that this ill phenomenon almost always happens when the classical augmented Lagrangian method is applied to formulation (3) of sparse PCA. To alleviate this drawback, we propose a novel augmented Lagrangian method for problem (6) and establish its global convergence in this subsection.

Throughout this subsection, we make the following assumption for problem (6).

Assumption 1 Problem (6) is feasible, and moreover at least a feasible solution, denoted by x^{feas}, is known.

It is well-known that for problem (6) the associated augmented Lagrangian function $L_\varrho(x, \lambda, \mu) : \Re^n \times \Re^m \times \Re^p \to \Re$ is given by

$$L_\varrho(x, \lambda, \mu) := w(x) + P(x), \tag{16}$$

where

$$w(x) := f(x) + \frac{1}{2\varrho}(\|[\lambda + \varrho g(x)]^+\|^2 - \|\lambda\|^2) + \mu^T h(x) + \frac{\varrho}{2}\|h(x)\|^2, \tag{17}$$

and $\varrho > 0$ is a penalty parameter (e.g., see [4,27]). Roughly speaking, an augmented Lagrangian method, when applied to problem (6), solves a sequence of subproblems in the form of

$$\min_{x \in X} L_\varrho(x, \lambda, \mu)$$

while updating the Lagrangian multipliers (λ, μ) and the penalty parameter ϱ.

Let x^{feas} be a known feasible point of (6) (see Assumption 1). We now describe the algorithm framework of a novel augmented Lagrangian method as follows.

Algorithm framework of augmented Lagrangian method:

Let $\{\epsilon_k\}$ be a positive sequence. Let $\lambda^0 \in \Re^m_+$, $\mu^0 \in \Re^p$, $\varrho_0 > 0, \tau > 0, \sigma > 1$ be given. Choose an arbitrary initial point $x^0_{\text{init}} \in X$ and constant $\Upsilon \geq \max\{f(x^{\text{feas}}), L_{\varrho_0}(x^0_{\text{init}}, \lambda^0, \mu^0)\}$. Set $k = 0$.

(1) Find an approximate solution $x^k \in X$ for the subproblem

$$\min_{x \in X} L_{\varrho_k}(x, \lambda^k, \mu^k) \tag{18}$$

such that

$$\text{dist}\left(-\nabla w(x^k), \partial P(x^k) + N_X(x^k)\right) \leq \epsilon_k, \quad L_{\varrho_k}(x^k, \lambda^k, \mu^k) \leq \Upsilon. \tag{19}$$

(2) Update Lagrange multipliers according to

$$\lambda^{k+1} := [\lambda^k + \varrho_k g(x^k)]^+, \quad \mu^{k+1} := \mu^k + \varrho_k h(x^k). \tag{20}$$

(3) Set $\varrho_{k+1} := \max\left\{\sigma\varrho_k, \|\lambda^{k+1}\|^{1+\tau}, \|\mu^{k+1}\|^{1+\tau}\right\}$.
(4) Set $k \leftarrow k+1$ and go to step 1).

end

The above augmented Lagrangian method differs from the classical augmented Lagrangian method in that: i) the values of the augmented Lagrangian functions at their approximate minimizers given by the method are uniformly bounded from above (see Step 1)); and ii) the magnitude of penalty parameters outgrows that of Lagrangian multipliers (see Step 3)). These two novel properties are crucial in ensuring the convergence of our augmented Lagrangian method both theoretically and practically. In fact, we observed in our experiments that when one or both of these steps are replaced by the counterparts of the classical augmented Lagrangian method, the resulting method almost always fails to converge to even a feasible point as applied to formulation (3) of sparse PCA.

To make the above augmented Lagrangian method complete, we need to address how to find an approximate solution $x^k \in X$ for subproblem (18) satisfying (19) as required in Step 1). We will leave this discussion to the end of this subsection. For the time being, we establish the main convergence result regarding this method for solving problem (6).

Theorem 3.3 *Assume that $\epsilon_k \to 0$. Let $\{x^k\}$ be the sequence generated by the above augmented Lagrangian method satisfying (19). Suppose that a subsequence $\{x^k\}_{k\in K}$ converges to x^*. Then, the following statements hold:*

(a) *x^* is a feasible point of problem (6);*
(b) *Further, if Robinson's condition (7) is satisfied at x^*, then the subsequence $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$ is bounded, and each accumulation point (λ^*, μ^*) of $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$ is a vector of Lagrange multipliers satisfying the first-order optimality conditions (9)–(10) at x^*.*

Proof In view of (16), (17) and the second relation in (19), we have

$$\begin{aligned} f(x^k) + P(x^k) + \frac{1}{2\varrho_k}(\|[\lambda^k + \varrho_k g(x^k)]^+\|^2 - \|\lambda^k\|^2) \\ + (\mu^k)^T h(x^k) + \frac{\varrho_k}{2}\|h(x^k)\|^2 \le \Upsilon \quad \forall k. \end{aligned}$$

It follows that

$$\begin{aligned} \|[\lambda^k/\varrho_k + g(x^k)]^+\|^2 + \|h(x^k)\|^2 \le 2[\Upsilon - f(x^k) - g(x^k) - (\mu^k)^T h(x^k)]/\varrho_k \\ + (\|\lambda_k\|/\varrho_k)^2. \end{aligned}$$

Noticing that $\varrho_0 > 0$ $\tau > 0$, and $\varrho_{k+1} = \max\left\{\sigma\varrho_k, \|\lambda^{k+1}\|^{1+\tau}, \|\mu^{k+1}\|^{1+\tau}\right\}$ for $k \ge 0$, we can observe that $\varrho_k \to \infty$ and $\|(\lambda^k, \mu^k)\|/\varrho_k \to 0$. We also know that

$\{x^k\}_{k\in K} \to x^*$, $\{g(x^k)\}_{k\in K} \to g(x^*)$ and $\{h(x^k)\}_{k\in K} \to h(x^*)$. Using these results, and upon taking limits as $k \in K \to \infty$ on both sides of the above inequality, we obtain that

$$\|[g(x^*)]^+\|^2 + \|h(x^*)\|^2 \le 0,$$

which implies that $g(x^*) \le 0$ and $h(x^*) = 0$. We also know that $x^* \in X$. It thus follows that statement (a) holds.

We next show that statement (b) also holds. Using (18), (16), (17), (20), and the first relation in (19), we have

$$\|\nabla f(x^k) + (\lambda^{k+1})^T \nabla g(x^k) + (\mu^{k+1})^T \nabla h(x^k) + z^k + v^k\| \le \epsilon_k \tag{21}$$

for some $z^k \in \partial P(x^k)$ and $v^k \in N_X(x^k)$. Suppose for contradiction that the subsequence $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$ is unbounded. By passing to a subsequence if necessary, we can assume that $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K} \to \infty$. Let $(\bar\lambda^{k+1}, \bar\mu^{k+1}) = (\lambda^{k+1}, \mu^{k+1})/\|(\lambda^{k+1}, \mu^{k+1})\|$ and $\bar v^k = v^k/\|(\lambda^{k+1}, \mu^{k+1})\|$. Recall that $\{x^k\}_{k\in K} \to x^*$. It together with Theorem 6.2.7 of [15] implies that $\cup_{k\in K}\partial P(x^k)$ is bounded, and so is $\{z^k\}_{k\in K}$. In addition, $\{g(x^k)\}_{k\in K} \to g(x^*)$ and $\{h(x^k)\}_{k\in K} \to h(x^*)$. Then, we can observe from (21) that $\{\bar v^k\}_{k\in K}$ is bounded. Without loss of generality, assume that $\{(\bar\lambda^{k+1}, \bar\mu^{k+1})\}_{k\in K} \to (\bar\lambda, \bar\mu)$ and $\{\bar v^k\}_{k\in K} \to \bar v$ (otherwise, one can consider their convergent subsequences). Clearly, $\|(\bar\lambda, \bar\mu)\| = 1$. Dividing both sides of (21) by $\|(\lambda^{k+1}, \mu^{k+1})\|$ and taking limits as $k \in k \to \infty$, we obtain that

$$\bar\lambda^T \nabla g(x^*) + \bar\mu^T \nabla h(x^*) + \bar v = 0. \tag{22}$$

Further, using the identity $\lambda^{k+1} = [\lambda^k + \varrho_k g(x^k)]^+$ and the fact that $\varrho_k \to \infty$ and $\|\lambda^k\|/\varrho_k \to 0$, we observe that $\lambda^{k+1} \in \Re^m_+$ and $\lambda_i^{k+1} = 0$ for $i \notin \mathcal{A}(x^*)$ when $k \in K$ is sufficiently large, which imply that $\bar\lambda \in \Re^m_+$ and $\bar\lambda_i = 0$ for $i \notin \mathcal{A}(x^*)$. Moreover, we have $\bar v \in N_X(x^*)$ since $N_X(x^*)$ is a closed cone. Using these results, (22), Robinson's condition (7) at x^*, and a similar argument as that in the proof of Theorem 3.2, we can obtain that $(\bar\lambda, \bar\mu) = (0, 0)$, which contradicts the identity $\|(\bar\lambda, \bar\mu)\| = 1$. Therefore, the subsequence $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$ is bounded. Using this result together with (21) and the fact $\{z^k\}_{k\in K}$ is bounded, we immediately see that $\{v^k\}_{k\in K}$ is bounded. Using semi-continuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [26] and Lemma 2.42 of [27]), and the fact $\{x^k\}_{k\in K} \to x^*$, we conclude that every accumulation point of $\{z^k\}_{k\in K}$ and $\{v^k\}_{k\in K}$ belongs to $\partial P(x^*)$ and $N_X(x^*)$, respectively. Using these results and (21), we further see that for every accumulation point (λ^*, μ^*) of $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$, there exists some $z^* \in \partial P(x^*)$ and $v^* \in N_X(x^*)$ such that

$$\nabla f(x^*) + (\lambda^*)^T \nabla g(x^*) + (\mu^*)^T \nabla h(x^*) + z^* + v^* = 0.$$

Moreover, using the identity $\lambda^{k+1} = [\lambda^k + \varrho_k g(x^k)]^+$ and the fact that $\varrho_k \to \infty$ and $\|\lambda^k\|/\varrho_k \to 0$, we easily see that $\lambda^* \in \Re^m_+$ and $\lambda_i^* = 0$ for $i \notin \mathcal{A}(x^*)$. Thus, (λ^*, μ^*) satisfies the first-order optimality conditions (9)–(10) at x^*. □

Before ending this subsection, we now briefly discuss how to find an approximate solution $x^k \in X$ for subproblem (18) satisfying (19) as required in Step 1) of the above augmented Lagrangian method. In particular, we are interested in applying the nonmonotone gradient methods proposed in Sect. 3.3 to (18). As shown in Sect. 3.3 (see Theorems 3.9 and 3.13), these methods are able to find an approximate solution $x^k \in X$ satisfying the first relation of (19). Moreover, if an initial point for these methods is properly chosen, the obtained approximate solution x^k also satisfies the second relation of (19). For example, given $k \geq 0$, let $x^k_{\rm init} \in X$ denote the initial point for solving the kth subproblem (18), and we define $x^k_{\rm init}$ for $k \geq 1$ as follows

$$x^k_{\rm init} = \begin{cases} x^{\rm feas}, & \text{if } L_{\varrho_k}(x^{k-1}, \lambda^k, \mu^k) > \Upsilon; \\ x^{k-1}, & \text{otherwise,} \end{cases}$$

where x^{k-1} is the approximate solution to the $(k-1)$th subproblem (18) satisfying (19) (with k replaced by $k-1$). Recall from Assumption 1 that $x^{\rm feas}$ is a feasible solution of (6). Thus, $g(x^{\rm feas}) \leq 0$, and $h(x^{\rm feas}) = 0$, which together with (16), (17) and the definition of Υ implies that

$$L_{\varrho_k}(x^{\rm feas}, \lambda^k, \mu^k) \;\leq\; f(x^{\rm feas}) \;\leq\; \Upsilon.$$

It follows from this inequality and the above choice of $x^k_{\rm init}$ that $L_{\varrho_k}(x^k_{\rm init}, \lambda^k, \mu^k) \leq \Upsilon$. Additionally, the nonmonotone gradient methods proposed in Sect. 3.3 possess a natural property that the objective function values at all subsequent iterates are bounded above by the one at the initial point. Therefore, we have

$$L_{\varrho_k}(x^k, \lambda^k, \mu^k) \;\leq\; L_{\varrho_k}(x^k_{\rm init}, \lambda^k, \mu^k) \;\leq\; \Upsilon,$$

and so the second relation of (19) is satisfied at x^k.

3.3 Nonmonotone gradient methods for nonsmooth minimization

In this subsection we propose two nonmonotone gradient methods for minimizing a class of nonsmooth functions over a closed convex set, which can be suitably applied to the subproblems arising in our augmented Lagrangian method detailed in Sect. 3.2. We also establish global convergence and local linear rate of convergence for these methods.

Throughout this subsection, we consider the following problem

$$\min_{x \in X} \{F(x) := f(x) + P(x)\}, \tag{23}$$

where $f : \Re^n \to \Re$ is continuously differentiable, $P : \Re^n \to \Re$ is convex but not necessarily smooth, and $X \subseteq \Re^n$ is closed and convex.

In the literature [3,23,29,30], several gradient methods were proposed for solving problem (23) or its special case. In particular, Tseng and Yun [29] studied a block

coordinate descent method for (23). Under the assumption that the gradient of f is Lipschitz continuous, Wright et al. [30] proposed a globally convergent nonmonotone gradient method for (23). In addition, for the case where f is convex and its gradient is Lipschitz continuous, Nesterov [23] and Beck and Teboulle [3] developed optimal gradient methods for (23). In this subsection, we propose two nonmonotone gradient methods for (23). These two methods are closely related to the ones proposed in [29,30], but they are not the same (see the remarks below for details). In addition, these methods can be viewed as an extension of the well-known projected gradient methods studied in [5] for smooth problems, but the methods proposed in [29,30] cannot. Before proceeding, we introduce some notations and establish some technical lemmas as follows that will be used subsequently.

We say that $x \in \Re^n$ is a *stationary point* of problem (23) if $x \in X$ and

$$0 \in \nabla f(x) + \partial P(x) + N_X(x). \tag{24}$$

Given a point $x \in \Re^n$ and $H \succ 0$, we denote by $d_H(x)$ the solution of the following problem:

$$d_H(x) := \arg\min_d \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + P(x+d) : \ x + d \in X \right\}. \tag{25}$$

The following lemma provides an alternative characterization of stationarity that will be used in our subsequent analysis.

Lemma 3.4 *For any $H \succ 0$, $x \in X$ is a stationary point of problem (23) if and only if $d_H(x) = 0$.*

Proof We first observe that (25) is a convex problem, and moreover its objective function is strictly convex. The conclusion of this lemma immediately follows from this observation and the first-order optimality condition of (25). □

The next lemma shows that $\|d_H(x)\|$ changes not too fast with H. It will be used to prove Theorems 3.10 and 3.14.

Lemma 3.5 *For any $x \in \Re^n$, $H \succ 0$, and $\tilde{H} \succ 0$, let $d = d_H(x)$ and $\tilde{d} = d_{\tilde{H}}(x)$. Then*

$$\|\tilde{d}\| \le \frac{1 + \lambda_{\max}(Q) + \sqrt{1 - 2\lambda_{\min}(Q) + \lambda_{\max}(Q)^2}}{2\lambda_{\min}(\tilde{H})} \lambda_{\max}(H) \|d\|, \tag{26}$$

where $Q = H^{-1/2} \tilde{H} H^{-1/2}$.

Proof The conclusion immediately follows from Lemma 3.2 of [29] with $J = \{1, \ldots, n\}$, $c = 1$, and $P(x) := P(x) + I_X(x)$, where I_X is the indicator function of X. □

The following lemma will be used to prove Theorems 3.10 and 3.14.

Lemma 3.6 *Given $x \in \Re^n$ and $H \succ 0$, let $g = \nabla f(x)$ and $\Delta_d = g^T d + P(x+d) - P(x)$ for all $d \in \Re^n$. Let $\sigma \in (0,1)$ be given. The following statements hold:*

(a) *If $d = d_H(x)$, then*

$$-\Delta_d \ \geq \ d^T H d \ \geq \ \lambda_{\min}(H)\|d\|^2.$$

(b) *For any $\bar{x} \in \Re^n$, $\alpha \in (0,1]$, $d = d_H(x)$, and $x' = x + \alpha d$, then*

$$(g + Hd)^T(x' - \bar{x}) + P(x') - P(\bar{x}) \leq (\alpha - 1)(d^T H d + \Delta_d).$$

(c) *If f satisfies*

$$\|\nabla f(y) - \nabla f(z)\| \leq L\|y - z\| \quad \forall y, z \in \Re^n \tag{27}$$

for some $L > 0$, then the descent condition

$$F(x + \alpha d) \leq F(x) + \sigma\alpha\Delta_d$$

is satisfied for $d = d_H(x)$, provided $0 \leq \alpha \leq \min\{1, 2(1-\sigma)\lambda_{\min}(H)/L\}$.

(d) *If f satisfies (27), then the descent condition*

$$F(x + d) \leq F(x) + \sigma\Delta_d$$

is satisfied for $d = d_{H(\theta)}(x)$, where $H(\theta) = \theta H$, provided $\theta \geq L/[2(1-\sigma)\lambda_{\min}(H)]$.

Proof The statements (a)-(c) follow from Theorem 4.1 (a) and Lemma 3.4 of [29] with $J = \{1, \ldots, n\}$, $\gamma = 0$, and $\underline{\lambda} = \lambda_{\min}(H)$. We now prove statement (d). Letting $\alpha = 1$, $d = d_{H(\theta)}(x)$ and using statement (c), we easily see that when $2(1-\sigma)\lambda_{\min}(H(\theta)) \geq 1$, $F(x+d) \leq F(x) + \sigma\Delta_d$ is satisfied, which together with the definition of $H(\theta)$ implies statement (d) holds. □

We now present the first nonmonotone gradient method for (23) as follows.

Nonmonotone gradient method I:

Choose parameters $\eta > 1, 0 < \sigma < 1, 0 < \underline{\theta} < \bar{\theta}, 0 < \underline{\lambda} \leq \bar{\lambda}$, and integer $M \geq 0$. Set $k = 0$ and choose $x^0 \in X$.

(1) Choose $\theta_k^0 \in [\underline{\theta}, \bar{\theta}]$ and $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$.
(2) For $j = 0, 1, \ldots$
  (2a) Let $\theta_k = \theta_k^0 \eta^j$. Solve (25) with $x = x^k$ and $H = \theta_k H_k$ to obtain $d^k = d_H(x)$.
  (2b) If d^k satisfies

$$F(x^k + d^k) \ \leq \ \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma\Delta_k, \tag{28}$$

go to step (3), where

$$\Delta_k := \nabla f(x^k)^T d^k + P(x^k + d^k) - P(x^k). \tag{29}$$

(3) Set $x^{k+1} = x^k + d^k$ and $k \leftarrow k + 1$.

end

Remark The above method is closely related to the one proposed in [30]. They differ from each other only in that the distinct Δ_k's are used inequality (28). In particular, the method [30] uses $\Delta_k = -\theta_k \|d^k\|^2/2$. For global convergence, the method [30], however, requires a strong assumption that the gradient of f is *Lipschitz* continuous, which is not needed for our method (see Theorem 3.9). In addition, our method can be viewed as an extension of one projected gradient method (namely, SPG1) studied in [5] for smooth problems, but their method cannot. Finally, local convergence is established for our method under the assumption that the gradient of f is Lipschitz continuous (see Theorem 3.10), but not studied for the methods in [30] and [5]. □

We next prove global convergence of the nonmonotone gradient method I. Before proceeding, we establish two technical lemmas below. The first lemma shows that if $x^k \in X$ is a nonstationary point, there exists an $\theta_k > 0$ in step 2a) so that (28) is satisfied, and hence the above method is well defined.

Lemma 3.7 *Suppose that $H_k \succ 0$ and $x^k \in X$ is a nonstationary point of problem (23). Then, there exists $\tilde{\theta} > 0$ such that $d^k = d_{H_k(\theta_k)}(x^k)$, where $H_k(\theta_k) = \theta_k H_k$, satisfies (28) whenever $\theta_k \geq \tilde{\theta}$.*

Proof For simplicity of notation, let $d(\theta) = d_{H_k(\theta)}(x^k)$, where $H_k(\theta) = \theta H_k$ for any $\theta > 0$. Since x^k is a nonstationary point of problem (23), it follows from Lemma 3.4 that $d(\theta) \neq 0$ for all $\theta > 0$. By Theorem 23.1 of [26], we know that $q(t) := \frac{1}{t}[P(x^k + td(\theta)/\|d(\theta)\|) - P(x^k)]$ is non-decreasing on $(0, \infty)$ and moreover, $\lim_{t \downarrow 0} q(t) = P'(x^k, d(\theta)/\|d(\theta)\|)$, which implies that $q(t) \geq P'(x^k, d(\theta)/\|d(\theta)\|)$ for all $t > 0$. Thus, we have

$$\frac{P(x^k + d(\theta)) - P(x^k)}{\|d(\theta)\|} = q(\|d(\theta)\|) \geq P'(x^k, d(\theta)/\|d(\theta)\|).$$

Using this inequality and (25), we further obtain that for all $\theta > 0$,

$$\begin{aligned} \theta\|d(\theta)\| &\leq -\frac{2[\nabla f(x^k)^T d(\theta) + P(x^k + d(\theta)) - P(x^k)]}{\lambda_{\min}(H_k)\|d(\theta)\|} \\ &\leq -\frac{2[\nabla f(x^k)^T d(\theta)/\|d(\theta)\| + P'(x^k, d(\theta)/\|d(\theta)\|)]}{\lambda_{\min}(H_k)} \\ &= -\frac{2F'(x^k, d(\theta)/\|d(\theta)\|)}{\lambda_{\min}(H_k)}. \end{aligned} \tag{30}$$

Thus, we easily see that the set $\tilde{S} := \{\theta\|d(\theta)\| : \theta > 0\}$ is bounded. It implies that $\|d(\theta)\| \to 0$ as $\theta \to \infty$. We claim that

$$\liminf_{\theta\to\infty} \theta\|d(\theta)\| > 0. \tag{31}$$

Suppose not. Then there exists a sequence $\{\bar{\theta}_l\} \uparrow \infty$ such that $\bar{\theta}_l\|d(\bar{\theta}_l)\| \to 0$ as $l \to \infty$. Invoking that $d(\bar{\theta}_l)$ is the optimal solution of (25) with $x = x^k$, $H = \bar{\theta}_l H_k$ and $\theta = \bar{\theta}_l$, we have

$$0 \in \nabla f(x^k) + \bar{\theta}_l H_k d(\bar{\theta}_l) + \partial P(x^k + d(\bar{\theta}_l)) + N_X(x^k + d(\bar{\theta}_l)).$$

Upon taking limits on both sides as $l \to \infty$, and using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [26] and Lemma 2.42 of [27]), and the relations $\|d(\bar{\theta}_l)\| \to 0$ and $\bar{\theta}_l\|d(\bar{\theta}_l)\| \to 0$, we see that (24) holds at x^k, which contradicts the nonstationarity of x^k. Hence, (31) holds. We observe that

$$\theta d(\theta)^T H_k d(\theta) \geq \lambda_{\min}(H_k)\theta\|d(\theta)\|^2,$$

which together with (31) and $H_k \succ 0$, implies that

$$\|d(\theta)\| = O\left(\theta d(\theta)^T H_k d(\theta)\right) \text{ as } \theta \to \infty. \tag{32}$$

This relation together with Lemma 3.6(a) implies that as $\theta \to \infty$,

$$\|d(\theta)\| = O\left(\theta d(\theta)^T H_k d(\theta)\right) = O\left(P(x^k) - \nabla f(x^k)^T d(\theta) - P(x^k + d(\theta))\right). \tag{33}$$

Using this result and the relation $\|d(\theta)\| \to 0$ as $\theta \to \infty$, we further have

$$\begin{aligned} &F(x^k + d(\theta)) - \max_{[k-M]^+ \leq i \leq k} F(x^i) \\ &\leq F(x^k + d(\theta)) - F(x^k) \\ &= f(x^k + d(\theta)) - f(x^k) + P(x^k + d(\theta)) - P(x^k) \\ &= \nabla f(x^k)^T d(\theta) + P(x^k + d(\theta)) - P(x^k) + o(\|d(\theta)\|) \\ &\leq \sigma[\nabla f(x^k)^T d(\theta) + P(x^k + d(\theta)) - P(x^k)], \end{aligned} \tag{34}$$

provided θ is sufficiently large. It implies that the conclusion holds.

The following lemma shows that the search directions $\{d^k\}$ approach zero, and the sequence of objective function values $\{F(x^k)\}$ also converges.

Lemma 3.8 *Suppose that F is bounded below in X and uniformly continuous in the the level set $\mathcal{L} = \{x \in X : F(x) \leq F(x^0)\}$. Then, the sequence $\{x^k\}$ generated by the nonmonotone gradient method I satisfies $\lim_{k\to\infty} d^k = 0$. Moreover, the sequence $\{F(x^k)\}$ converges.*

Proof We first observe that $\{x^k\} \subseteq \mathcal{L}$. Let $l(k)$ be an integer such that $[k-M]^+ \leq l(k) \leq k$ and

$$F(x^{l(k)}) = \max\{F(x^i) : [k-M]^+ \leq i \leq k\}$$

for all $k \geq 0$. We clearly observe that $F(x^{k+1}) \leq F(x^{l(k)})$ for all $k \geq 0$, which together with the definition of $l(k)$ implies that the sequence $\{F(x^{l(k)})\}$ is monotonically nonincreasing. Further, since F is bounded below in X, we have

$$\lim_{k\to\infty} F(x^{l(k)}) = F^* \tag{35}$$

for some $F^* \in \Re$. We next prove by induction that the following limits hold for all $j \geq 1$:

$$\lim_{k\to\infty} d^{l(k)-j} = 0, \quad \lim_{k\to\infty} F(x^{l(k)-j}) = F^*. \tag{36}$$

Using (28) and (29) with k replaced by $l(k)-1$, we obtain that

$$F(x^{l(k)}) \leq F(x^{l(l(k)-1)}) + \sigma\Delta_{l(k)-1}. \tag{37}$$

Replacing k and θ by $l(k)-1$ and $\theta_{l(k)-1}$ in (33), respectively, and using $H_{l(k)-1} \succeq \underline{\lambda} I$ and the definition of $\Delta_{l(k)-1}$ (see (29)), we have

$$\Delta_{l(k)-1} \leq -\underline{\lambda}\theta_{l(k)-1}\|d^{l(k)-1}\|^2.$$

The above two inequalities yield that

$$F(x^{l(k)}) \leq F(x^{l(l(k)-1)}) - \sigma\underline{\lambda}\theta_{l(k)-1}\|d^{l(k)-1}\|^2, \tag{38}$$

which together with (35) implies that $\lim_{k\to\infty}\theta_{l(k)-1}\|d^{l(k)-1}\|^2 = 0$. Further, noticing that $\theta_k \geq \underline{\theta}$ for all k, we obtain that $\lim_{k\to\infty} d^{l(k)-1} = 0$. Using this result and (35), we have

$$\lim_{k\to\infty} F(x^{l(k)-1}) = \lim_{k\to\infty} F(x^{l(k)} - d^{l(k)-1}) = \lim_{k\to\infty} F(x^{l(k)}) = F^*, \tag{39}$$

where the second equality follows from uniform continuity of F in $\mathcal{L}$. Therefore, (36) holds for $j = 1$. We now need to show that if (36) holds for j, then it also holds for $j+1$. Using a similar argument as that leading to (38), we have

$$F(x^{l(k)-j}) \leq F(x^{l(l(k)-j-1)}) - \sigma\underline{\lambda}\theta_{l(k)-j-1}\|d^{l(k)-j-1}\|^2,$$

which together with (35), the induction assumption $\lim_{k\to\infty} F(x^{l(k)-j}) = F^*$, and the fact that $\theta_{l(k)-j-1} \geq \underline{\theta}$ for all k, yields $\lim_{k\to\infty} d^{l(k)-j-1} = 0$. Using this result, the induction assumption $\lim_{k\to\infty} F(x^{l(k)-j}) = F^*$, and a similar argument as that leading to (39), we can show that $\lim_{k\to\infty} F(x^{l(k)-j-1}) = F^*$. Hence, (36) holds for $j+1$.

Finally, we will prove that $\lim_{k\to\infty} d^k = 0$ and $\lim_{k\to\infty} F(x^k) = F^*$. By the definition of $l(k)$, we see that for $k \geq M+1, k-M-1 = l(k)-j$ for some $1 \leq j \leq M+1$, which together with the first limit in (36), implies that $\lim_{k\to\infty} d^k = \lim_{k\to\infty} d^{k-M-1} = 0$. Additionally, we observe that

$$x^{l(k)} = x^{k-M-1} + \sum_{j=1}^{\bar{l}_k} d^{l(k)-j} \quad \forall k \geq M+1,$$

where $\bar{l}_k = l(k) - (k-M-1) \leq M+1$. Using the above identity, (36), and uniform continuity of F in $\mathcal{L}$, we see that $\lim_{k\to\infty} F(x^k) = \lim_{k\to\infty} F(x^{k-M-1}) = F^*$. Thus, the conclusion of this lemma holds. □

We are now ready to show that the nonmonotone gradient method I is globally convergent.

Theorem 3.9 *Suppose that F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \leq F(x^0)\}$. Then, any accumulation point of the sequence $\{x^k\}$ generated by the nonmonotone gradient method I is a stationary point of* (23).

Proof Suppose for contradiction that x^* is an accumulation point of $\{x^k\}$ that is a nonstationary point of (23). Let K be the subsequence such that $\{x^k\}_{k\in K} \to x^*$. We first claim that $\{\theta_k\}_{k\in K}$ is bounded. Suppose not. Then there exists a subsequence of $\{\theta_k\}_{k\in K}$ that goes to ∞. Without loss of generality, we assume that $\{\theta_k\}_{k\in K} \to \infty$. For simplicity of notations, let $\bar{\theta}_k = \theta_k/\eta$, $d^k(\theta) = d_{H_k(\theta)}(x^k)$ for $k \in K$ and $\theta > 0$, where $H_k(\theta) = \theta H_k$. Since $\{\theta_k\}_{k\in K} \to \infty$ and $\theta_k^0 \leq \bar{\theta}$, there exists some index $\bar{k} \geq 0$ such that $\theta_k > \theta_k^0$ for all $k \in K$ with $k \geq \bar{k}$. By the particular choice of θ_k specified in steps (2a) and (2b), we have

$$\begin{aligned} F(x^k + d^k(\bar{\theta}_k)) > & \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma[\nabla f(x^k)^T d^k(\bar{\theta}_k) \\ & + P(x^k + d^k(\bar{\theta}_k)) - P(x^k)], \end{aligned} \tag{40}$$

Using a similar argument as that leading to (30), we have

$$\bar{\theta}_k \|d^k(\bar{\theta}_k)\| \leq -\frac{2F'(x^k, d^k(\bar{\theta}_k)/\|d^k(\bar{\theta}_k)\|)}{\lambda_{\min}(H_k)} \quad \forall k \in K,$$

which along with the relations $H_k \succeq \underline{\lambda} I$ and $\{x^k\}_{k\in K} \to x^*$, implies that $\{\bar{\theta}_k \|d^k(\bar{\theta}_k)\|\}_{k\in K}$ is bounded. Since $\{\bar{\theta}_k\}_{k\in K} \to \infty$, we further have $\{\|d^k(\bar{\theta}_k)\|\}_{k\in K} \to 0$. We now claim that

$$\liminf_{k\in K, k\to\infty} \bar{\theta}_k \|d^k(\bar{\theta}_k)\| > 0. \tag{41}$$

Suppose not. By passing to a subsequence if necessary, we can assume that $\{\bar{\theta}_k \|d^k(\bar{\theta}_k)\|\}_{k\in K} \to 0$. Invoking that $d^k(\bar{\theta}_k)$ is the optimal solution of (25) with

$x = x^k$ and $H = \bar{\theta}_k H_k$, we have

$$0 \in \nabla f(x^k) + \bar{\theta}_k H_k d^k(\bar{\theta}_k) + \partial P(x^k + d^k(\bar{\theta}_k)) + N_X(x^k + d^k(\bar{\theta}_k)) \quad \forall k \in K.$$

Upon taking limits on both sides as $k \in K \to \infty$, and using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [26] and Lemma 2.42 of [27]), the relations $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, $\{\|d^k(\bar{\theta}_k)\|\}_{k\in K} \to 0$, $\{\bar{\theta}_k \|d^k(\bar{\theta}_k)\|\}_{k\in K} \to 0$ and $\{x^k\}_{k\in K} \to x^*$, we see that (24) holds at x^*, which contradicts nonstationarity of x^*. Thus, (41) holds. Now, using (41), the relation $H_k \succeq \underline{\lambda} I$, and a similar argument as for deriving (32), we obtain that $\|d^k(\bar{\theta}_k)\| = O\left(\bar{\theta}_k d^k(\bar{\theta}_k)^T H_k d^k(\bar{\theta}_k)\right)$ as $k \in K \to \infty$. Using this result and a similar argument as the one leading to (34), we have

$$\begin{aligned} F(x^k + d^k(\bar{\theta}_k)) \le & \max_{[k-M]^+ \le i \le k} F(x^i) + \sigma[\nabla f(x^k)^T d^k(\bar{\theta}_k) \\ & + P(x^k + d^k(\bar{\theta}_k)) - P(x^k)], \end{aligned}$$

provided that $k \in K$ is sufficiently large. The above inequality evidently contradicts (40). Thus, $\{\theta_k\}_{k\in K}$ is bounded.

Finally, invoking that $d^k = d^k(\theta_k)$ is the optimal solution of (25) with $x = x^k$, $H = \theta_k H_k$, we have

$$0 \in \nabla f(x^k) + \theta_k H_k d^k + \partial P(x^k + d^k) + N_X(x^k + d^k) \quad \forall k \in K. \tag{42}$$

By Lemma 3.8, we have $\{d^k\}_{k\in K} \to 0$. Upon taking limits on both sides of (42) as $k \in K \to \infty$, and using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [26] and Lemma 2.42 of [27]), and the relations $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, $\{d^k\}_{k\in K} \to 0$ and $\{x^k\}_{k\in K} \to x^*$, we see that (24) holds at x^*, which contradicts the nonstationarity of x^* that is assumed at the beginning of this proof. Therefore, the conclusion of this theorem holds. □

We next analyze the asymptotic convergence rate of the nonmonotone gradient method I under the following assumption, which is the same as the one made in [29]. In what follows, we denote by $\bar{X}$ the set of stationary points of problem (23).

Assumption 2 (a) $\bar{X} \neq \emptyset$ and, for any $\zeta \ge \min_{x\in X} F(x)$, there exists $\varpi > 0$ and $\epsilon > 0$ such that

$$\operatorname{dist}(x, \bar{X}) \le \varpi \|d_I(x)\| \quad \text{whenever} \quad F(x) \le \zeta, \ \|d_I(x)\| \le \epsilon.$$

(b) There exists $\delta > 0$ such that

$$\|x - y\| \ge \delta \quad \text{whenever} \quad x \in \bar{X}, y \in \bar{X}, F(x) \neq F(y).$$

We are ready to establish local linear rate of convergence for the nonmonotone gradient method I described above. The proof of the following theorem is inspired by the work of Tseng and Yun [29], who analyzed a similar local convergence for a coordinate gradient descent method for a class of nonsmooth minimization problems.

Theorem 3.10 *Let $l(k)$ be an integer in $[[k-M]^+, k]$ such that $F(x^{l(k)}) = \max\{F(x^i) : [k-M]^+ \le i \le k\}$ for all $k \ge 0$. Suppose that Assumption 2 holds, f satisfies (27), and F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \le F(x^0)\}$. Then, the sequence $\{x^k\}$ generated by the nonmonotone gradient method I satisfies*

$$F(x^{l(k)}) - F^* \;\le\; c(F(x^{l(l(k))-1}) - F^*),$$

provided k is sufficiently large, where $F^* = \lim_{k\to\infty} F(x^k)$ (see Lemma 3.8), and c is some constant in $(0, 1)$.

Proof Invoking $\theta_k^0 \le \bar{\theta}$ and the specific choice of θ_k, we see from Lemma 3.6(d) that $\hat{\theta} := \sup_k \theta_k < \infty$. Let $H_k(\theta) = \theta H_k$. Then, it follows from $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$ and $\theta_k \ge \underline{\theta}$ that $(\underline{\theta} \cdot \underline{\lambda}) I \preceq H_k(\theta_k) \preceq \hat{\theta}\bar{\lambda} I$. Using this relation, Lemma 3.5, $H_k \succeq \underline{\lambda} I$, and $d^k = d_{H_k(\theta_k)}(x^k)$, we obtain that

$$\|d_I(x^k)\| = O\left(\|d^k\|\right), \tag{43}$$

which together with Lemma 3.8 implies $\{d_I(x^k)\} \to 0$. Thus, for any $\epsilon > 0$, there exists some index $\bar{k}$ such that $d_I(x^{l(k)-1}) \le \epsilon$ for all $k \ge \bar{k}$. In addition, we clearly observe that $F(x^{l(k)-1}) \le F(x^0)$. Then, by Assumption 2(a) and (43), there exists some index k' such that

$$\|x^{l(k)-1} - \bar{x}^{l(k)-1}\| \le c_1 \|d^{l(k)-1}\| \quad \forall k \ge k' \tag{44}$$

for some $c_1 > 0$ and $\bar{x}^{l(k)-1} \in \bar{X}$. Note that

$$\|x^{l(k+1)-1} - x^{l(k)-1}\| \le \sum_{i=l(k)-1}^{l(k+1)-2} \|d^i\| \le \sum_{i=[k-M-1]^+}^{[k-1]^+} \|d^i\|,$$

which together with $\{d^k\} \to 0$, implies that $\|x^{l(k+1)-1} - x^{l(k)-1}\| \to 0$. Using this result, (44), and Lemma 3.8, we obtain

$$\begin{aligned}\|\bar{x}^{l(k+1)-1} - \bar{x}^{l(k)-1}\| &\le \|x^{l(k+1)-1} - \bar{x}^{l(k+1)-1}\| \\ &\quad + \|x^{l(k)-1} - \bar{x}^{l(k)-1}\| + \|x^{l(k+1)-1} - \bar{x}^{l(k)-1}\| \\ &\le c_1\|d^{l(k+1)-1}\| + c_1\|d^{l(k)-1}\| + \|x^{l(k+1)-1} - \bar{x}^{l(k)-1}\| \;\to\; 0.\end{aligned}$$

It follows from this relation and Assumption 2(b) that there exists an index $\hat{k} \ge k'$ and $v \in \Re$ such that

$$F(\bar{x}^{l(k)-1}) = v \quad \forall k \ge \hat{k}. \tag{45}$$

Then, by Lemma 5.1 of [29], we see that

$$F^* = \lim_{k\to\infty} F(x^k) = \liminf_{k\to\infty} F(x^{l(k)-1}) \;\ge\; v. \tag{46}$$

Further, using the definition of F, (27), (45), Lemma 3.6(b), and $H_k(\theta_k) \preceq \hat{\theta}\bar{\lambda}I$, we have for $k \geq \hat{k}$,

$$\begin{aligned} F(x^{l(k)}) - v &= f(x^{l(k)}) + P(x^{l(k)}) - f(\bar{x}^{l(k)-1}) - P(\bar{x}^{l(k)-1}) \\ &= \nabla f(\tilde{x}^k)^T(x^{l(k)} - \bar{x}^{l(k)-1}) + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \\ &= (\nabla f(\tilde{x}^k) - \nabla f(x^{l(k)-1})^T(x^{l(k)} - \bar{x}^{l(k)-1}) \\ &\quad -(H_{l(k)-1}(\theta_{l(k)-1})d^{l(k)-1})^T(x^{l(k)} - \bar{x}^{l(k)-1}) \\ &\quad +\Big[(\nabla f(x^{l(k)-1}) + H_{l(k)-1}(\theta_{l(k)-1})d^{l(k)-1})^T(x^{l(k)} - \bar{x}^{l(k)-1}) \\ &\qquad +P(x^{l(k)}) - P(\bar{x}^{l(k)-1})\Big] \\ &\leq L\|\tilde{x}^k - x^{l(k)-1}\|\|x^{l(k)} - \bar{x}^{l(k)-1}\| + \hat{\theta}\bar{\lambda}\|d^{l(k)-1}\|\|x^{l(k)} - \bar{x}^{l(k)-1}\|, \end{aligned} \tag{47}$$

where $\tilde{x}^k$ is some point lying on the segment joining $x^{l(k)}$ with $\bar{x}^{l(k)-1}$. It follows from (44) that, for $k \geq \hat{k}$,

$$\|\tilde{x}^k - x^{l(k)-1}\| \leq \|x^{l(k)} - x^{l(k)-1}\| + \|x^{l(k)-1} - \bar{x}^{l(k)-1}\| = (1 + c_1)\|d^{l(k)-1}\|.$$

Similarly, $\|x^{l(k)} - \bar{x}^{l(k)-1}\| \leq (1 + c_1)\|d^{l(k)-1}\|$ for $k \geq \hat{k}$. Using these inequalities, Lemma 3.6(a), $H_k(\theta_k) \succeq (\underline{\theta} \cdot \underline{\lambda})I$, and (47), we see that for $k \geq \hat{k}$,

$$F(x^{l(k)}) - v \;\leq\; -c_2\Delta_{l(k)-1}$$

for some constant $c_2 > 0$. This inequality together with (37) gives

$$F(x^{l(k)}) - v \;\leq\; c_3\left(F(x^{l(l(k)-1)}) - F(x^{l(k)})\right) \quad \forall k \geq \hat{k}, \tag{48}$$

where $c_3 = c_2/\sigma$. Using $\lim_{k\to\infty} F(x^{l(k)}) = F^*$, and upon taking limits on both sides of (48), we see that $F^* \leq v$, which together with (46) implies that $v = F^*$. Using this result and upon rearranging terms of (48), we have

$$F(x^{l(k)}) - F^* \;\leq\; c(F(x^{l(l(k))-1}) - F^*) \quad \forall k \geq \hat{k},$$

where $c = c_3/(1 + c_3)$. □

We next present the second nonmonotone gradient method for (23) as follows.

Nonmonotone gradient method II:

Choose parameters $0 < \eta < 1, 0 < \sigma < 1, 0 < \underline{\alpha} < \bar{\alpha}, 0 < \underline{\lambda} \leq \bar{\lambda}$, and integer $M \geq 0$. Set $k = 0$ and choose $x^0 \in X$.

(1) Choose $\underline{\lambda}I \preceq H_k \preceq \bar{\lambda}I$.
(2) Solve (25) with $x = x^k$ and $H = H_k$ to obtain $d^k = d_H(x)$, and compute Δ_k according to (29).

(3) Choose $\alpha_k^0 \in [\underline{\alpha}, \bar{\alpha}]$. Find the smallest integer $j \geq 0$ such that $\alpha_k = \alpha_k^0 \eta^j$ satisfies

$$x^k + \alpha_k d^k \in X, \quad F(x^k + \alpha_k d^k) \leq \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma \alpha_k \Delta_k, \tag{49}$$

where Δ_k is defined in (29).

(4) Set $x^{k+1} = x^k + \alpha_k d^k$ and $k \leftarrow k + 1$.

end

Remark The above method is closely related to the one proposed in [29]. In particular, when the entire coordinate block, that is, $J = \{1, \ldots, n\}$, is chosen for the method [29], it becomes a special case of our method with $M = 0$, which is actually a gradient descent method. Given that our method is generally a nonmonotone method when $M \geq 1$, most proofs of global and local convergence for the method [29] do not hold for our method. In addition, our method can be viewed as an extension of one projected gradient method (namely, SPG2) studied in [5] for smooth problems, but the method [29] generally cannot. □

We next prove global convergence of the nonmonotone gradient method II. Before proceeding, we establish two technical lemmas below. The first lemma shows that if $x^k \in X$ is a nonstationary point, there exists an $\alpha_k > 0$ in step (3) so that (49) is satisfied, and hence the above method is well defined.

Lemma 3.11 *Suppose that $H_k \succ 0$ and $x^k \in X$ is a nonstationary point of problem (23). Then, there exists $\tilde{\alpha} > 0$ such that $d^k = d_{H_k}(x^k)$ satisfies (49) whenever $0 < \alpha_k \leq \tilde{\alpha}$.*

Proof In view of Lemma 2.1 of [29] with $J = \{1, \ldots, n\}, c = 1, x = x^k$, and $H = H_k$, we have

$$\begin{aligned} F(x^k + \alpha d^k) &\leq F(x^k) + \alpha \Delta_k + o(\alpha) \\ &\leq \max_{[k-M]^+ \leq i \leq k} F(x^i) + \alpha \Delta_k + o(\alpha) \quad \forall \alpha \in (0, 1], \end{aligned}$$

where Δ_k is defined in (29). Using the assumption of this lemma, we see from Lemma 3.4 that $d^k \neq 0$, which together with $H_k \succ 0$ and Lemma 3.6(a) implies $\Delta_k < 0$. The conclusion of this lemma immediately follows from this relation and the above inequality. □

The following lemma shows that the scaled search directions $\{\alpha_k d^k\}$ approach zero, and the sequence of objective function values $\{F(x^k)\}$ also converges.

Lemma 3.12 *Suppose that F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \leq F(x^0)\}$. Then, the sequence $\{x^k\}$ generated by the nonmonotone gradient method II satisfies $\lim_{k\to\infty} \alpha_k d^k = 0$. Moreover, the sequence $\{F(x^k)\}$ converges.*

Proof Let $l(k)$ be defined in the proof of Lemma 3.8. We first observe that $\{x^k\} \subseteq \mathcal{L}$. Using (29), the definition of d^k, and $H_k \succeq \underline{\lambda} I$, we have

$$\Delta_k = \nabla f(x^k)^T d^k + P(x^k + d^k) - P(x^k) \leq -\frac{1}{2}(d^k)^T H_k d^k \leq -\frac{1}{2}\underline{\lambda}\|d^k\|^2, \quad (50)$$

which together with the relation $\alpha_k \leq \alpha_k^0 \leq \bar{\alpha}$, implies that

$$\alpha_k^2 \|d^k\|^2 \leq -2\bar{\alpha}\alpha_k \Delta_k / \underline{\lambda}. \quad (51)$$

By a similar argument as that leading to (35), we see that $\{x^k\}$ satisfies (35) for some F^*. We next show by induction that the following limits hold for all $j \geq 1$:

$$\lim_{k\to\infty} \alpha_{l(k)-j} d^{l(k)-j} = 0, \quad \lim_{k\to\infty} F(x^{l(k)-j}) = F^*. \quad (52)$$

Indeed, using (49) with k replaced by $l(k) - 1$, we obtain that

$$F(x^{l(k)}) \leq F(x^{l(l(k)-1)}) + \sigma \alpha_{l(k)-1} \Delta_{l(k)-1}.$$

It together with (35) immediately yields $\lim_{k\to\infty} \alpha_{l(k)-1}\Delta_{l(k)-1} = 0$. Using this result and (51), we see that the first identity of (52) holds for $j = 1$. Further, in view of this identity, (35), and uniform continuity of F in $\mathcal{L}$, we can easily see that the second identity of (52) also holds $j = 1$. We now need to show that if (52) holds for j, then it also holds for $j + 1$. First, it follows from (49) that

$$F(x^{l(k)-j}) \leq F(x^{l(l(k)-j-1)}) + \sigma \alpha_{l(k)-j-1} \Delta_{l(k)-j-1},$$

which together with (35) and the induction assumption that $\lim_{k\to\infty} F(x^{l(k)-j}) = F^*$, yields $\lim_{k\to\infty} \alpha_{l(k)-j-1}\Delta_{l(k)-j-1} = 0$. Using this result and (51), we have $\lim_{k\to\infty} \alpha_{l(k)-j-1} d^{l(k)-j-1} = 0$. In view of this identity, uniform continuity of F in $\mathcal{L}$ and the induction assumption $\lim_{k\to\infty} F(x^{l(k)-j}) = F^*$, we can easily show that $\lim_{k\to\infty} F(x^{l(k)-j-1}) = F^*$. Hence, (52) holds for $j + 1$. The conclusion of this lemma then follows from (52) and a similar argument as that in the proof of Lemma 3.8. □

We are now ready to show that the nonmonotone gradient method II is globally convergent.

Theorem 3.13 *Suppose that F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \leq F(x^0)\}$. Then, any accumulation point of the sequence $\{x^k\}$ generated by the nonmonotone gradient method II is a stationary point of (23).*

Proof Suppose for contradiction that x^* is an accumulation point of $\{x^k\}$ that is a nonstationary point of (23). Let K be the subsequence such that $\{x^k\}_{k\in K} \to x^*$. We first claim that $\liminf_{k\in K, k\to\infty} \|d^k\| > 0$. Suppose not. By passing to a subsequence

if necessary, we can assume that $\{\|d^k\|\}_{k\in K} \to 0$. Invoking that d^k is the optimal solution of (25) with $x = x^k$ and $H = H_k$, we have

$$0 \in \nabla f(x^k) + H_k d^k + \partial P(x^k + d^k) + N_X(x^k + d^k) \quad \forall k \in K.$$

Upon taking limits on both sides as $k \in K \to \infty$, and using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [26] and Lemma 2.42 of [27]) the relations $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, $\{\|d^k\|\}_{k\in K} \to 0$ and $\{x^k\}_{k\in K} \to x^*$, we see that (24) holds at x^*, which contradicts the nonstationarity of x^*. Thus, $\liminf_{k\in K, k\to\infty} \|d^k\| > 0$ holds. Further, using a similar argument as that leading to (30), we have

$$\|d^k\| \le -\frac{2F'(x^k, d^k/\|d^k\|)}{\lambda_{\min}(H_k)} \quad \forall k \in K,$$

which together with $\{x^k\}_{k\in K} \to x^*$, $H_k \succeq \underline{\lambda} I$ and $\liminf_{k\in K, k\to\infty} \|d^k\| > 0$, implies that $\{d^k\}_{k\in K}$ is bounded. Further, using (50), we see that $\limsup_{k\in K, k\to\infty} \Delta_k < 0$. Now, it follows from Lemma 3.12 and the relation $\liminf_{k\in K, k\to\infty} \|d^k\| > 0$ that $\{\alpha_k\}_{k\in K} \to 0$. Since $\alpha_k^0 \ge \underline{\alpha} > 0$, there exists some index $\bar{k} \ge 0$ such that $\alpha_k < \alpha_k^0$ and $\alpha_k < \eta$ for all $k \in K$ with $k \ge \bar{k}$. Let $\bar{\alpha}_k = \alpha_k/\eta$. Then, $\{\bar{\alpha}_k\}_{k\in K} \to 0$ and $0 < \bar{\alpha}_k \le 1$ for all $k \in K$. By the stepsize rule used in step (3), we have, for all $k \in K$ with $k \ge \bar{k}$,

$$F(x^k + \bar{\alpha}_k d^k) > \max_{[k-M]^+ \le i \le k} F(x^i) + \sigma \bar{\alpha}_k \Delta_k, \tag{53}$$

On the other hand, in view of the definition of F, (29), the boundedness of $\{d^k\}_{k\in K}$, the relation $\limsup_{k\in K, k\to\infty} \Delta_k < 0$, and the monotonicity of $(P(x^k + \alpha d^k) - P(x^k))/\alpha$, we obtain that, for sufficiently large $k \in K$,

$$\begin{aligned}
F(x^k + \bar{\alpha}_k d^k) &= f(x^k + \bar{\alpha}_k d^k) + P(x^k + \bar{\alpha}_k d^k) \\
&= f(x^k + \bar{\alpha}_k d^k) - f(x^k) + P(x^k + \bar{\alpha}_k d^k) - P(x^k) + F(x^k) \\
&= \bar{\alpha}_k \nabla f(x^k)^T d^k + o(\bar{\alpha}_k \|d^k\|) + P(x^k + \bar{\alpha}_k d^k) - P(x^k) + F(x^k) \\
&\le \bar{\alpha}_k \nabla f(x^k)^T d^k + o(\bar{\alpha}_k) + \bar{\alpha}_k [P(x^k + d^k) - P(x^k)] + \max_{[k-M]^+ \le i \le k} F(x^i) \\
&= \max_{[k-M]^+ \le i \le k} F(x^i) + \bar{\alpha}_k \Delta_k + o(\bar{\alpha}_k) \\
&< \max_{[k-M]^+ \le i \le k} F(x^i) + \sigma \bar{\alpha}_k \Delta_k,
\end{aligned}$$

which clearly contradicts (53). Therefore, the conclusion of this theorem holds. □

We next establish local linear rate of convergence for the nonmonotone gradient method II described above. The proof of the following theorem is inspired by the work of Tseng and Yun [29].

Theorem 3.14 *Let $l(k)$ be an integer in $[[k - M]^+, k]$ such that $F(x^{l(k)}) = \max\{F(x^i) : [k - M]^+ \le i \le k\}$ for all $k \ge 0$. Suppose that Assumption 2 holds,*

$\bar{\alpha} \le 1$, f satisfies (27), and F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \le F(x^0)\}$. Then, the sequence $\{x^k\}$ generated by the nonmonotone gradient method II satisfies

$$F(x^{l(k)}) - F^* \;\le\; c(F(x^{l(l(k))-1}) - F^*)$$

provided k is sufficiently large, where $F^* = \lim_{k\to\infty} F(x^k)$ (see Lemma 3.12), and c is some constant in $(0, 1)$.

Proof Since α_k is chosen by the stepsize rule used in step (3) with $\alpha_k^0 \ge \underline{\alpha} > 0$, we see from Lemma 3.6(c) that $\inf_k \alpha_k > 0$. It together with Lemma 3.12 implies that $\{d^k\} \to 0$. Further, using Lemma 3.5 and the fact that $d^k = d_{H_k}(x^k)$ and $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, we obtain that $\|d_I(x^k)\| = \Theta(\|d^k\|)$, and hence $\{d_I(x^k)\} \to 0$. Then, by a similar argument as that in the proof of Theorem 3.10, there exist $c_1 > 0$, $v \in \Re$, and $\bar{x}^{l(k)-1} \in \bar{X}$ such that

$$\|x^{l(k)-1} - \bar{x}^{l(k)-1}\| \le c_1 \|d^{l(k)-1}\|, \quad F(\bar{x}^{l(k)-1}) = v \quad \forall k \ge \hat{k},$$

where $\hat{k}$ is some index. Then, by Lemma 5.1 of [29], we see that (46) holds for $\{x^k\}$, and the above F^* and v. Further, using the definition of F, (27), Lemma 3.6(b), and $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, we have, for $k \ge \hat{k}$,

$$\begin{aligned}
F(x^{l(k)}) - v &= f(x^{l(k)}) + P(x^{l(k)}) - f(\bar{x}^{l(k)-1}) - P(\bar{x}^{l(k)-1}) \\
&= \nabla f(\tilde{x}^k)^T (x^{l(k)} - \bar{x}^{l(k)-1}) + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \\
&= (\nabla f(\tilde{x}^k) - \nabla f(x^{l(k)-1})^T (x^{l(k)} - \bar{x}^{l(k)-1}) \\
&\quad -(H_{l(k)-1} d^{l(k)-1})^T (x^{l(k)} - \bar{x}^{l(k)-1}) \\
&\quad + \Big[(\nabla f(x^{l(k)-1}) + H_{l(k)-1} d^{l(k)-1})^T (x^{l(k)} - \bar{x}^{l(k)-1}) \\
&\qquad + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \Big] \\
&\le L \|\tilde{x}^k - x^{l(k)-1}\| \|x^{l(k)} - \bar{x}^{l(k)-1}\| + \bar{\lambda} \|d^{l(k)-1}\| \|x^{l(k)} - \bar{x}^{l(k)-1}\| \\
&\quad + (\alpha_{l(k)-1} - 1) \Big[(d^{l(k)-1})^T H_{l(k)-1} d^{l(k)-1} + \Delta_{l(k)-1} \Big],
\end{aligned} \tag{54}$$

where $\tilde{x}^k$ is some point lying on the segment joining $x^{l(k)}$ with $\bar{x}^{l(k)-1}$. It follows from (44) and $\alpha_k \le 1$ that, for $k \ge \hat{k}$,

$$\|\tilde{x}^k - x^{l(k)-1}\| \le \|x^{l(k)} - x^{l(k)-1}\| + \|x^{l(k)-1} - \bar{x}^{l(k)-1}\| \le (1 + c_1)\|d^{l(k)-1}\|.$$

Similarly, $\|x^{l(k)} - \bar{x}^{l(k)-1}\| \le (1 + c_1)\|d^{l(k)-1}\|$ for $k \ge \hat{k}$. Using these inequalities, Lemma 3.6(a), $H_k \succeq \underline{\lambda} I$, $\alpha_k \le 1$, and (54), we see that, for $k \ge \hat{k}$,

$$F(x^{l(k)}) - v \;\le\; -c_2 \Delta_{l(k)-1}$$

for some constant $c_2 > 0$. The remaining proof follows similarly as that of Theorem 3.10. □

4 Augmented Lagrangian method for sparse PCA

In this section we discuss the applicability and implementation details of the augmented Lagrangian method proposed in Sect. 3 for solving sparse PCA (3).

4.1 Applicability of augmented Lagrangian method for (3)

We first observe that problem (3) can be reformulated as

$$\begin{array}{ll} \min\limits_{V\in\Re^{n\times r}} & -\mathrm{Tr}(V^T\hat{\Sigma}V)+\rho\bullet|V| \\ \text{s.t.} & V_i^T\hat{\Sigma}V_j\le\Delta_{ij}\quad\forall i\ne j, \\ & -V_i^T\hat{\Sigma}V_j\le\Delta_{ij}\quad\forall i\ne j, \\ & V^TV=I. \end{array} \tag{55}$$

Clearly, problem (55) has the same form as (6). From Sect. 3.2, we know that the sufficient conditions for convergence of our augmented Lagrangian method include: i) a feasible point is explicitly given; and ii) Robinson's condition (7) holds at an accumulation point. It is easy to observe that any $V\in\Re^{n\times r}$ consisting of r orthonormal eigenvectors of $\hat{\Sigma}$ is a feasible point of (55), and thus the first condition is trivially satisfied. Given that the accumulation points are not known beforehand, it is hard to check the second condition directly. Instead, we may check Robinson's condition at all feasible points of (55). However, due to complication of the constraints, we are only able to verify Robinson's condition at a set of feasible points below. Before proceeding, we establish a technical lemma as follows that will be used subsequently.

Lemma 4.1 *Let $V\in\Re^{n\times r}$ be a feasible solution of (55). Given any $W_1, W_2\in\mathcal{S}^r$, the system of*

$$\delta V^T\hat{\Sigma}V+V^T\hat{\Sigma}\,\delta V+\delta D=W_1, \tag{56}$$

$$\delta V^TV+V^T\delta V=W_2 \tag{57}$$

has at least one solution $(\delta V,\delta D)\in\Re^{n\times r}\times\mathcal{D}^r$ if one of the following conditions holds:

(a) *$V^T\hat{\Sigma}V$ is diagonal and $V_i^T\hat{\Sigma}V_i\ne V_j^T\hat{\Sigma}V_j$ for all $i\ne j$;*

(b) *$V^T\hat{\Sigma}(I-VV^T)\hat{\Sigma}V$ is nonsingular.*

Proof Note that the columns of V consist of r orthonormal eigenvectors. Therefore, there exist $\bar{V}\in\Re^{n\times(n-r)}$ such that $[V\ \bar{V}]\in\Re^{n\times n}$ is an orthogonal matrix. It follows that for any $\delta V\in\Re^{n\times r}$, there exists $\delta P\in\Re^{r\times r}$ and $\delta\bar{P}\in\Re^{(n-r)\times r}$ such that $\delta V=V\delta P+\bar{V}\delta\bar{P}$. Performing such a change of variable for δV, and using the fact that the matrix $[V\ \bar{V}]$ is orthogonal, we can show that the system of (56) and (57) is equivalent to

$$\delta P^TG+G\delta P+\delta\bar{P}^T\bar{G}+\bar{G}^T\delta\bar{P}+\delta D=W_1, \tag{58}$$

$$\delta P^T+\delta P=W_2, \tag{59}$$

where $G = V^T \hat{\Sigma} V$ and $\bar{G} = \bar{V}^T \hat{\Sigma} V$. The remaining proof of this lemma reduces to show that the system of (58) and (59) has at least a solution $(\delta P, \delta \bar{P}, \delta D) \in \Re^{r\times r} \times \Re^{(n-r)\times r} \times \mathcal{D}^r$ if one of conditions (a) or (b) holds.

First, we assume that condition (a) holds. Then, G is a diagonal matrix and $G_{ii} \neq G_{jj}$ for all $i \neq j$. It follows that there exists a unique $\delta P^* \in \Re^{n\times r}$ satisfying $\delta P_{ii} = (W_2)_{ii}/2$ for all i and

$$\begin{array}{rl} \delta P_{ij} G_{jj} + G_{ii} \delta P_{ij} = (W_1)_{ij} & \forall i \neq j, \\ \delta P_{ij} + \delta P_{ji} = (W_2)_{ij} & \forall i \neq j. \end{array}$$

Now, let $\delta \bar{P}^* = 0$ and $\delta D^* = \widetilde{\mathrm{Diag}}(W_1 - GW_2)$. It is easy to verify that $(\delta P^*, \delta \bar{P}^*, \delta D^*)$ is a solution of the system of (58) and (59).

We next assume that condition (b) holds. Given any $\delta \bar{P} \in \Re^{(n-r)\times r}$, there exist $\delta Y \in \Re^{(n-r)\times r}$ and $\delta Z \in \Re^{r\times r}$ such that $\bar{G}^T \delta Y = 0$ and $\delta \bar{P} = \delta Y + \bar{G} \delta Z$. Performing such a change of variable for $\delta \bar{P}$, we see that (58) can be rewritten as

$$\delta P^T G + G \delta P + \delta Z^T \bar{G}^T \bar{G} + \bar{G}^T \bar{G} \delta Z + \delta D = W_1. \tag{60}$$

Thus, it suffices to show that the system of (59) and (60) has at least a solution $(\delta P, \delta Z, \delta D) \in \Re^{r\times r} \times \Re^{r\times r} \times \mathcal{D}^r$. Using the definition of $\bar{G}$ and the fact that the matrix $[V\ \bar{V}]$ is orthogonal, we see that

$$\bar{G}^T \bar{G} = V^T \hat{\Sigma} \bar{V} \bar{V}^T \hat{\Sigma} V = V^T \hat{\Sigma} (I - VV^T) \hat{\Sigma} V,$$

which together with condition (b) implies that $\bar{G}^T \bar{G}$ is nonsingular. Now, let

$$\delta P^* = W_2/2, \quad \delta Z^* = (\bar{G}^T \bar{G})^{-1} (2W_1 - W_2 G - G W_2)/4, \quad \delta D^* = 0.$$

It is easy to verify that $(\delta P^*, \delta Z^*, \delta D^*)$ is a solution of the system of (60) and (59). Therefore, the conclusion holds. □

We are now ready to show that Robinson's condition (7) holds at a set of feasible points of (55).

Proposition 4.2 *Let $V \in \Re^{n\times r}$ be a feasible solution of (55). The Robinson's condition (7) holds at V if one of the following conditions hold:*

(a) *$\Delta_{ij} = 0$ and $V_i^T \hat{\Sigma} V_i \neq V_j^T \hat{\Sigma} V_j$ for all $i \neq j$;*
(b) *There is at least one active and one inactive inequality constraint of (55) at V, and $V^T \hat{\Sigma} (I - VV^T) \hat{\Sigma} V$ is nonsingular;*
(c) *All inequality constraints of (55) are inactive at V.*

Proof We first suppose that condition (a) holds. Then, it immediately implies that $V^T \hat{\Sigma} V$ is diagonal, and hence the condition (a) of Lemma 4.1 holds. In addition, we observe that all constraints of (55) become equality ones. Using these facts and Lemma 4.1, we see that Robinson's condition (7) holds at V. Next, we assume that condition

(b) holds. It implies that condition (b) of Lemma 4.1 holds. The conclusion then follows directly from Lemma 4.1. Finally, suppose condition (c) holds. Then, Robinson's condition (7) holds at V if and only if (57) has at least a solution $\delta V \in \Re^{n\times r}$ for any $W_2 \in \mathcal{S}^r$. Noting that $V^TV = I$, we easily see that $\delta V = VW_2/2$ is a solution of (57), and thus Robinson's condition (7) holds at V. □

From Proposition 4.2, we see that Robinson's condition (7) indeed holds at a set of feasible points of (55). Though we are not able to show that it holds at all feasible points of (55), we observe in our implementation that the accumulation points of our augmented Lagrangian method generally satisfy one of the conditions described in Proposition 4.2, and so Robinson's condition usually holds at the accumulation points. Moreover, we have never seen that our augmented Lagrangian method failed to converge for an instance in our implementation so far.

4.2 Implementation details of augmented Lagrangian method for (55)

In this section, we show how our augmented Lagrangian method proposed in Sect. 3.2 can be applied to solve problem (55) (or, equivalently, (3)). In particular, we will discuss the implementation details of outer and inner iterations of this method.

We first discuss how to efficiently evaluate the function and gradient involved in our augmented Lagrangian method for problem (55). Suppose that $\varrho > 0$ is a penalty parameter, and $\{\lambda^+_{ij}\}_{i\neq j}$ and $\{\lambda^-_{ij}\}_{i\neq j}$ are the Lagrangian multipliers for the inequality constraints of (55), respectively, and $\mu \in \mathcal{S}^r$ is the Lagrangian multipliers for the equality constraints of (55). For convenience of presentation, let $\Delta \in \mathcal{S}^r$ be the matrix whose ijth entry equals the parameter Δ_{ij} of (55) for all $i \neq j$ and diagonal entries are 0. Similarly, let λ^+ (resp., λ^-) be an $r \times r$ symmetric matrix whose ijth entry is λ^+_{ij} (resp., λ^-_{ij}) for all $i \neq j$ and diagonal entries are 0. We now define $\lambda \in \Re^{2r\times r}$ by stacking λ^+ over λ^-. Using these notations, we observe that the associated Lagrangian function for problem (55) can be rewritten as

$$L_\varrho(V, \lambda, \mu) = w(V) + \rho \bullet |V|, \tag{61}$$

where

$$w(V) = -\mathrm{Tr}(V^T\hat{\Sigma}V) + \frac{1}{2\varrho}\left(\left\|\left[\begin{pmatrix}\lambda^+\\ \lambda^-\end{pmatrix} + \varrho\begin{pmatrix}S-\Delta\\ -S-\Delta\end{pmatrix}\right]^+\right\|_F^2 - \left\|\begin{pmatrix}\lambda^+\\ \lambda^-\end{pmatrix}\right\|_F^2\right)$$
$$+\mu \bullet R + \frac{\varrho}{2}\|R\|_F^2,$$

and

$$S = V^T\hat{\Sigma}V - \widetilde{\mathrm{Diag}}(V^T\hat{\Sigma}V), \qquad R = V^TV - I. \tag{62}$$

It is not hard to verify that the gradient of $w(V)$ can be computed according to

$$\nabla w(V) = 2\left(-\hat{\Sigma}V\left(I - [\lambda^+ + \varrho S - \varrho\Delta]^+ + [\lambda^- - \varrho S - \varrho\Delta]^+\right) + V(\mu + \varrho R)\right).$$

Clearly, the main effort for the above function and gradient evaluations lies in computing $V^T\hat{\Sigma}V$ and $\hat{\Sigma}V$. When $\hat{\Sigma} \in \mathcal{S}^p$ is explicitly given, the computational complexity for evaluating these two quantities is $O(p^2r)$. In practice, we are, however, typically given the data matrix $X \in \Re^{n\times p}$. Assuming the column means of X are 0, the sample covariance matrix $\hat{\Sigma}$ can be obtained from $\hat{\Sigma} = X^TX/(n-1)$. Nevertheless, when $p \gg n$, we observe that it is not efficient to compute and store $\hat{\Sigma}$. Also, it is much cheaper to compute $V^T\hat{\Sigma}V$ and $\hat{\Sigma}V$ by using $\hat{\Sigma}$ implicitly rather than explicitly. Indeed, we can first evaluate XV, and then compute $V^T\hat{\Sigma}V$ and $\hat{\Sigma}V$ according to

$$V^T\hat{\Sigma}V = (XV)^T(XV)/(n-1), \qquad \hat{\Sigma}V = X^T(XV)/(n-1).$$

Then, the resulting overall computational complexity is $O(npr)$, which is clearly much superior to the one by using $\hat{\Sigma}$ explicitly, that is, $O(p^2r)$.

We now address initialization and termination criterion for our augmented Lagrangian method. In particular, we choose initial point V^0_{init} and feasible point V^{feas} to be the loading vectors of the r standard PCs, that is, the orthonormal eigenvectors corresponding to r largest eigenvalues of $\hat{\Sigma}$. In addition, we set initial penalty parameter and Lagrangian multipliers to be 1, and set the parameters $\tau = 0.2$ and $\sigma = 10$. We terminate our method once the constraint violation and the relative difference between the augmented Lagrangian function and the regular objective function are sufficiently small, that is,

$$\max_{i\neq j}[|V_i^T\hat{\Sigma}V_j| - \Delta_{ij}]^+ \le \epsilon_I, \quad \max_{i,j}|R_{ij}| \le \epsilon_E, \quad \frac{|L_\varrho(V,\lambda,\mu) - f(V)|}{\max(|f(V)|,1)} \le \epsilon_O, \tag{63}$$

where $f(V) = -\mathrm{Tr}(V^T\hat{\Sigma}V) + \rho \bullet |V|$, R is defined in (62), and $\epsilon_I, \epsilon_E, \epsilon_O$ are some prescribed accuracy parameters corresponding to inequality constraints, equality constraints and objective function, respectively.

We next discuss how to apply the nonmonotone gradient methods proposed in Sect. 3.3 for the augmented Lagrangian subproblems, which are in the form of

$$\min_V L_\varrho(V,\lambda,\mu), \tag{64}$$

where the function $L_\varrho(\cdot,\lambda,\mu)$ is defined in (61). Given that the implementation details of those nonmonotone gradient methods are similar, we only focus on the first one, that is, the nonmonotone gradient method I. First, the initial point for this method can be chosen according to the scheme described at the end of Sect. 3.2. In addition, given the kth iterate V^k, we choose $H_k = \beta_k^{-1}I$ according to the scheme proposed by Barzilai and Borwein [2], which was also used by Birgin et al. [5] for studying a

class of projected gradient methods. Indeed, let $0 < \beta_{\min} < \beta_{\max}$ be given. Initially, choose an arbitrary $\beta_0 \in [\beta_{\min}, \beta_{\max}]$. Then, β_k is updated as follows:

$$\beta_{k+1} = \begin{cases} \beta_{\max}, & \text{if } b_k \leq 0; \\ \max\{\beta_{\min}, \min\{\beta_{\max}, a_k/b_k\}\}, & \text{otherwise,} \end{cases}$$

where $a_k = \|V^k - V^{k-1}\|_F^2$ and $b_k = (V^k - V^{k-1}) \bullet (\nabla w(V^k) - \nabla w(V^{k-1}))$. The search direction d^k is then computed by solving subproblem (25) with $H = \theta_k H_k$ for some $\theta_k > 0$, which in the context of (18) and (61) becomes

$$d^k := \arg\min_d \left\{ \nabla w(V^k) \bullet d + \frac{1}{2\theta_k \beta_k} \|d\|_F^2 + \rho \bullet |V^k + d| \right\}. \tag{65}$$

It is not hard to verify that the optimal solution of problem (65) has a closed-form expression, which is given by

$$d^k = \text{sign}(C) \odot [|C| - \theta_k \beta_k \rho]^+ - V^k,$$

where $C = V^k - \theta_k \beta_k \nabla w(V^k)$. In addition, we see from Lemma 3.4 that the following termination criterion is suitable for this method when applied to (64):

$$\frac{\max_{ij} |d_I(V^k)|_{ij}}{\max(|L_\varrho(V^k, \lambda, \mu)|, 1)} \leq \epsilon,$$

where $d_I(V^k)$ is the solution of (65) with $\theta_k \beta_k = 1$, and ϵ is a prescribed accuracy parameter. In our numerical implementation, we set $\beta_0 = 1/\max_{ij} |d_I(V^0)|_{ij}$, $\beta_{\max} = 10^{15}$, $\beta_{\min} = 10^{-15}$ and $\epsilon = 10^{-4}$.

Finally, it shall be mentioned that for the sake of practical performance, the numerical implementation of our augmented Lagrangian method is slightly different from the one described in Sect. 3.2. In particular, we follow a similar scheme as discussed on pp. 405 of [4] to adjust penalty parameter and Lagrangian multipliers. Indeed, they are updated separately rather than simultaneously. Roughly speaking, given $\gamma \in (0, 1)$, we adjust penalty parameter only when the constraint violation is not decreased by a factor γ over the previous minimization. Similarly, we update Lagrangian multipliers only when the constraint violation is decreased by a factor γ over the previous minimization. We choose $\gamma = 0.25$ in our implementation as recommended in [4].

5 Numerical results

In this section, we conduct numerical experiments for the augmented Lagrangian method detailed in Sects. 3.2 and 4.2 for formulation (55) (or, equivalently, (3)) of sparse PCA on synthetic, random, Pitprops, and gene expression data. In particular, we compare the results of our approach with several existing sparse PCA methods in terms of total explained variance, correlation of PCs, and orthogonality of loading vectors, which include the generalized power methods (Journée et al. [18]), the DSPCA

Table 1 Sparse PCA methods used for our comparison

GPower$_{l_1}$	Single-unit sparse PCA via l_1-penalty
GPower$_{l_0}$	Single-unit sparse PCA via l_0-penalty
GPower$_{l_1,m}$	Block sparse PCA via l_1-penalty
GPower$_{l_0,m}$	Block sparse PCA via l_0-penalty
DSPCA	DSPCA algorithm
SPCA	SPCA algorithm
rSVD	sPCA-rSVD algorithm with soft thresholding
ALSPCA	Augmented Lagrangian algorithm

algorithm (d'Aspremont et al. [10]), the SPCA algorithm (Zou et al. [31]), and the sPCA-rSVD algorithm (Shen and Huang [28]). We list all the methods used in this section in Table 1. Specifically, the methods with the prefix 'GPower' are the generalized power methods studied in [18], and the method ALSPCA is the augmented Lagrangian method proposed in this paper. In addition, the codes of the GPower methods and ALSPCA are written in MATLAB. All computations in this section are performed on a Lenovo PC with an AMD Phenom(tm) IIX4 900e 2.40 GHz processor and 4 GB memory.

As discussed in Sect. 2, the PCs obtained from the standard PCA based on sample covariance matrix $\hat{\Sigma} \in \Re^{n\times p}$ are nearly uncorrelated when the sample size is sufficiently large, and the total explained variance by the first r PCs approximately equals the sum of the individual variances of PCs, that is, $\mathrm{Tr}(V^T\hat{\Sigma}V)$, where $V \in \Re^{p\times r}$ consists of the loading vectors of these PCs. However, the PCs found by sparse PCA methods may be correlated with each other, and thus the quantity $\mathrm{Tr}(V^T\hat{\Sigma}V)$ can overestimate much the total explained variance by these PCs due to the overlap among their individual variances. In response to such an overlap, two adjusted total explained variances were proposed in [31,28]. It is not hard to observe that they can be viewed as the total explained variance of a set of transformed variables from the estimated sparse PCs. Given that these transformed variables can distinguish dramatically from those sparse PCs, their total explained variances may also differ much from each other. To alleviate this drawback while taking into account the possible correlations among PCs, we naturally introduce the following *adjusted total explained variance* for sparse PCs:

$$\mathrm{AdjVar}V = \mathrm{Tr}(V^T\hat{\Sigma}V) - \sqrt{\sum_{i\neq j}(V_i^T\hat{\Sigma}V_j)^2}.$$

It is not hard to show that AdjVar ≥ 0 for any $V \in \Re^{p\times r}$ provided $\hat{\Sigma} \succeq 0$. Clearly, when the PCs are uncorrelated, it becomes the usual total explained variance, that is, $\mathrm{Tr}(V^T\hat{\Sigma}V)$. We can also define the *cumulative percentage of adjusted variance* (CPAV) for the first r sparse PCs as the quotient of the adjusted total explained variance of these PCs and the total explained variance by all standard PCs, that is, $\mathrm{AdjVar}V/\mathrm{Tr}(\hat{\Sigma})$.

Finally, we shall stress that the main purpose of this section is to compare the quality of the sparse PCs found by those methods listed in Table 1 in terms of orthogonal-

ity, uncorrelation and total explained variance. Therefore, we will not compare the speed of these methods. Nevertheless, it shall be mentioned that our method, that is, ALSPCA, is a first-order method and capable of solving large-scale problems within a reasonable amount of time as demonstrated in our experiments presented in Sect. 5.4.

5.1 Synthetic data

In this subsection we use the synthetic data introduced by Zou et al. [31] to test the effectiveness of our approach ALSPCA for finding sparse PCs.

The synthetic example [31] considers three hidden factors:

$$V_1 \sim N(0, 290),\ V_2 \sim N(0, 300),\ V_3 = -0.3V_1 + 0.925V_2 + \epsilon,\ \epsilon \sim N(0, 1),$$

where V_1, V_2 and ϵ are independent. Then the 10 observable variables are generated as follows:

$$\begin{aligned} X_i &= V_1 + \epsilon_i^1, \quad \epsilon_i^1 \sim N(0,1), \quad i = 1,2,3,4, \\ X_i &= V_2 + \epsilon_i^2, \quad \epsilon_i^2 \sim N(0,1), \quad i = 5,6,7,8, \\ X_i &= V_3 + \epsilon_i^3, \quad \epsilon_i^3 \sim N(0,1), \quad i = 9,10, \end{aligned}$$

where ϵ_i^j are independent for $j = 1, 2, 3$ and $i = 1, \ldots, 10$. We will use the actual covariance matrix of $(X_1, \ldots, X_{10})$ to find the standard and sparse PCs, respectively.

We first observe that V_1 and V_2 are independent, but V_3 is a linear combination of V_1 and V_2. Moreover, the variances of these three underlying factors V_1, V_2 and V_3 are 290, 300, and 283.8, respectively. Thus V_2 is slightly more important than V_1, and they both are more important than V_3. In addition, the first two standard PCs together explain 99.72% of the total variance (see Table 2). These observations suggest that: i) the first two sparse PCs may be sufficient to explain most of the variance; and ii) the first sparse PC recovers the most important factor V_2 using (X_5, X_6, X_7, X_8), and the second sparse PC recovers the second important factor V_1 using (X_1, X_2, X_3, X_4).

Table 2 Loadings of the first two PCs by standard PCA and ALSPCA

Variable	PCA		ALSPCA	
	PC1	PC2	PC1	PC2
X_1	0.1158	0.4785	0	0.5000
X_2	0.1158	0.4785	0	0.5000
X_3	0.1158	0.4785	0	0.5000
X_4	0.1158	0.4785	0	0.5000
X_5	−0.3955	0.1449	−0.5000	0
X_6	−0.3955	0.1449	−0.5000	0
X_7	−0.3955	0.1449	−0.5000	0
X_8	−0.3955	0.1449	−0.5000	0
X_9	−0.4005	−0.0095	0	0
X_{10}	−0.4005	−0.0095	0	0
CPAV (%)	99.72		80.46	

Synthetic data

Table 3 Loadings of the first six PCs by standard PCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	0.4038	0.2178	0.2073	0.0912	0.0826	0.1198
Length	0.4055	0.1861	0.2350	0.1027	0.1128	0.1629
Moist	0.1244	0.5406	−0.1415	−0.0784	−0.3498	−0.2759
Testsg	0.1732	0.4556	−0.3524	−0.0548	−0.3558	−0.0540
Ovensg	0.0572	−0.1701	−0.4812	−0.0491	−0.1761	0.6256
Ringtop	0.2844	−0.0142	−0.4753	0.0635	0.3158	0.0523
Ringbut	0.3998	−0.1897	−0.2531	0.0650	0.2151	0.0026
Bowmax	0.2936	−0.1892	0.2431	−0.2856	−0.1853	−0.0551
Bowdist	0.3566	0.0171	0.2076	−0.0967	0.1061	0.0342
Whorls	0.3789	−0.2485	0.1188	0.2050	−0.1564	−0.1731
Clear	−0.0111	0.2053	0.0704	−0.8036	0.3430	0.1753
Knots	−0.1151	0.3432	−0.0920	0.3008	0.6003	−0.1698
Diaknot	−0.1125	0.3085	0.3261	0.3034	−0.0799	0.6263

Pitprops data

Given that (X_5, X_6, X_7, X_8) and (X_1, X_2, X_3, X_4) are independent, these sparse PCs would be uncorrelated and orthogonal each other.

In our test, we set $r = 2$, $\Delta_{ij} = 0$ for all $i \neq j$, and $\rho = 4$ for formulation (55) of sparse PCA. In addition, we choose (63) as the termination criterion for ALSPCA with $\epsilon_I = \epsilon_O = 0.1$ and $\epsilon_E = 10^{-3}$. The results of standard PCA and ALSPCA for this example are presented in Table 2. The loadings of standard and sparse PCs are given in columns two and three, respectively, and their CPAVs are given in the last row. We clearly see that our sparse PCs are consistent with the ones predicted above. Interestingly, they are identical with the ones obtained by SPCA and DSPCA reported in [10,31]. For general data, however, these methods may perform quite differently (see Sect. 5.2).

5.2 Pitprops data

In this subsection we test the performance of our approach ALSPCA for finding sparse PCs on the Pitprops data introduced by Jeffers [16]. We also compare the results with several existing methods [10,18,28,31].

The Pitprops data [16] has 180 observations and 13 measured variables. It is a classic example that illustrates the difficulty of interpreting PCs. Recently, several sparse PCA methods [10,19,28,31] have been applied to this data set for finding *six* sparse PCs by using the actual covariance matrix. For ease of comparison, we present the standard PCs, and the sparse PCs by some of those methods in Tables 3, 4, 5, and 6, respectively. We shall mention that two groups of sparse PCs were found in [10] by DSPCA with the parameter $k_1 = 5$ or 6, and they have similar sparsity and total explained variance (see [10] for details). Thus we only present the latter one (i.e., the one with $k_1 = 6$) in Table 6. Also, we applied the GPower methods [18] to this data set for finding the PCs with the sparsity given by the largest one of those found in [10,28,31], and observed that the best result was given by GPower$_{l_0}$. Thus we only report the sparse PCs obtained by GPower$_{l_0}$ in Table 7. In addition, we present spar-

Table 4 Loadings of the first six PCs by SPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	−0.477	0	0	0	0	0
Length	−0.476	0	0	0	0	0
Moist	0	0.785	0	0	0	0
Testsg	0	0.620	0	0	0	0
Ovensg	0.177	0	0.640	0	0	0
Ringtop	0	0	0.589	0	0	0
Ringbut	−0.250	0	0.492	0	0	0
Bowmax	−0.344	−0.021	0	0	0	0
Bowdist	−0.416	0	0	0	0	0
Whorls	−0.400	0	0	0	0	0
Clear	0	0	0	−1	0	0
Knots	0	0.013	0	0	−1	0
Diaknot	0	0	−0.015	0	0	1

Pitprops data

Table 5 Loadings of the first six PCs by rSVD

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	−0.449	0	0	−0.114	0	0
Length	−0.460	0	0	−0.102	0	0
Moist	0	−0.707	0	0	0	0
Testsg	0	−0.707	0	0	0	0
Ovensg	0	0	0.550	0	0	−0.744
Ringtop	−0.199	0	0.546	−0.176	0	0
Ringbut	−0.399	0	0.366	0	0	0
Bowmax	−0.279	0	0	0.422	0	0
Bowdist	−0.380	0	0	0.283	0	0
Whorls	−0.407	0	0	0	0.231	0
Clear	0	0	0	−0.785	−0.973	0
Knots	0	0	0	−0.265	0	0.161
Diaknot	0	0	−0.515	0	0	−0.648

Pitprops data

sity, CPAV, non-orthogonality and correlation of the PCs obtained by the standard PCA and sparse PCA methods [10,18,28,31] in columns two to five of Table 11, respectively. In particular, the second and fifth columns of this table respectively give sparsity (measured by the number of zero loadings) and CPAV. The third column reports non-orthogonality, which is measured by the maximum absolute difference between 90° and the angles formed by all pairs of loading vectors. Clearly, the smaller value in this column implies the better orthogonality. The fourth column presents the maximum correlation of PCs. Though the PCs given by these sparse PCA methods all have nice sparsity, we observe from Table 11 that they are highly correlated and moreover, almost all of them are far from orthogonal except the ones given by SPCA [31]. To improve the quality of sparse PCs, we next apply our approach ALSPCA, and compare the results with these methods. For all tests below, we choose (63) as the termination criterion for ALSPCA with $\epsilon_O = 0.1$ and $\epsilon_I = \epsilon_E = 10^{-3}$.

Table 6 Loadings of the first six PCs by DSPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	−0.4907	0	0	0	0	0
Length	−0.5067	0	0	0	0	0
Moist	0	0.7071	0	0	0	0
Testsg	0	0.7071	0	0	0	0
Ovensg	0	0	0	0	−1.0000	0
Ringtop	−0.0670	0	−0.8731	0	0	0
Ringbut	−0.3566	0	−0.4841	0	0	0
Bowmax	−0.2335	0	0	0	0	0
Bowdist	−0.3861	0	0	0	0	0
Whorls	−0.4089	0	0	0	0	0
Clear	0	0	0	0	0	1.0000
Knots	0	0	0	1.0000	0	0
Diaknot	0	0	0.0569	0	0	0

Pitprops data

Table 7 Loadings of the first six PCs by GPower$_{l_0}$

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	−0.4182	0	0	0	0	0
Length	−0.4205	0	0	0	0	0
Moist	0	−0.7472	0	0	0	0
Testsg	−0.1713	−0.6646	0	0	0	0
Ovensg	0	0	0	0	−0.7877	0
Ringtop	−0.2843	0	0	0	−0.6160	0
Ringbut	−0.4039	0	0	0	0	0
Bowmax	−0.3002	0	0	0	0	0
Bowdist	−0.3677	0	0	0	0	0
Whorls	−0.3868	0	0	0	0	0
Clear	0	0	0	0	0	1.0000
Knots	0	0	0	1.0000	0	0
Diaknot	0	0	1.0000	0	0	0

Pitprops data

In the first experiment, we aim to find six nearlry uncorrelated and orthogonal sparse PCs by ALSPCA while explaining most of variance. In particular, we set $r = 6$, $\Delta_{ij} = 0.07$ for all $i \neq j$ and $\rho = 0.8$ for formulation (55) of sparse PCA. The resulting sparse PCs are presented in Table 8, and their sparsity, CPAV, non-orthogonality and correlation are reported in row seven of Table 11. We easily observe that our method ALSPCA overall outperforms the other sparse PCA methods substantially in all aspects except sparsity. Naturally, we can improve the sparsity by increasing the values of ρ, yet the total explained variance may be sacrificed as demonstrated in our next experiment.

We now attempt to find six PCs with similar correlation and orthogonality but higher sparsity than those given in the above experiment. For this purpose, we set $\Delta_{ij} = 0.07$ for all $i \neq j$ and choose $\rho = 2.1$ for problem (55) in this experiment. The resulting sparse PCs are presented in Table 9, and their CPAV, non-orthogonality and correlation

Table 8 Loadings of the first six PCs by ALSPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	0.4394	0	0	0	0	0
Length	0.4617	0	0	0	0	0
Moist	0.0419	0.4611	−0.1644	0.0688	−0.3127	0
Testsg	0.1058	0.7902	0	0	0	0
Ovensg	0.0058	0	0	0	0	0
Ringtop	0.1302	0	0.2094	0	0	0.9999
Ringbut	0.3477	0	0.0515	0	0.3240	0
Bowmax	0.2256	−0.3566	0	0	0	0
Bowdist	0.4063	0	0	0	0	0
Whorls	0.4606	0	0	0	0	−0.0125
Clear	0	0.0369	0	−0.9973	0	0
Knots	−0.1115	0.1614	−0.0762	0.0239	0.8929	0
Diaknot	−0.0487	0.0918	0.9595	0.0137	0	0

Pitprops data: Test I

Table 9 Loadings of the first six PCs by ALSPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	1.0000	0	0	0	0	0
Length	0	−0.2916	−0.1421	0	0	−0.0599
Moist	0	0.9565	−0.0433	0	0	−0.0183
Testsg	0	0	0	0.0786	−0.1330	0
Ovensg	0	0	−0.9683	0	0	0
Ringtop	0	0	0	0	0	0
Ringbut	0	0	0.1949	0	0.2369	0
Bowmax	0	0	0	0	0	0
Bowdist	0	0	0	0	0	0
Whorls	0	0	0	0	0	0
Clear	0	0	0	−0.9969	0	0
Knots	0	0	−0.0480	0.0109	0.9624	0
Diaknot	0	0	−0.0093	0	0	0.9980

Pitprops data: Test II

of these PCs are given in row eight of Table 11. Compared to the PCs found in the above experiment, the ones obtained in this experiment are much more sparse while retaining almost same correlation and orthogonality. However, their CPAV goes down dramatically. Combining the results of these two experiments, we deduce that for the Pitprops data, it seems not possible to extract six highly sparse (e.g., around 60 zero loadings), nearly orthogonal and uncorrelated PCs while explaining most of variance as they may not exist. The following experiment further sustains such a deduction.

Finally we are interested in exploring how the correlation controlling parameters $\Delta_{ij}(i \neq j)$ affect the performance of the sparse PCs. In particular, we set $\Delta_{ij} = 0.5$ for all $i \neq j$ and choose $\rho = 0.7$ for problem (55). The resulting sparse PCs are presented in Table 10, and their CPAV, non-orthogonality and correlation of these PCs are given in the last row of Table 11. We see that these PCs are highly sparse, orthogonal, and explain good amount of variance. However, they are quite correlated each other, which is actually not surprising since $\Delta_{ij}(i \neq j)$ are not small. Despite such a

Table 10 Loadings of the first six PCs by ALSPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	0.4051	0	0	0	0	0
Length	0.4248	0	0	0	0	0
Moist	0	0.7262	0	0	0	0
Testsg	0.0018	0.6875	0	0	0	0
Ovensg	0	0	−1.0000	0	0	0
Ringtop	0.1856	0	0	0	0	0
Ringbut	0.4123	0	0	0	0	0
Bowmax	0.3278	0	0	0	0	0
Bowdist	0.3830	0	0	0	0	0
Whorls	0.4437	−0.0028	0	0	0	0
Clear	0	0	0	−1.0000	0	0
Knots	0	0	0	0	1.0000	0
Diaknot	0	0	0	0	0	1.0000

Pitprops data: Test III

Table 11 Comparison of SPCA, rSVD, DSPCA, GPower$_{l_0}$ and ALSPCA

Method	Sparsity	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	87.00
SPCA	60	0.86	0.395	66.21
rSVD	53	14.76	0.459	67.04
DSPCA	63	13.63	0.573	60.97
GPower$_{l_0}$	63	10.09	0.353	64.15
ALSPCA-1	46	0.03	0.082	69.55
ALSPCA-2	60	0.03	0.084	39.42
ALSPCA-3	63	0.00	0.222	65.97

Pitprops data

drawback, these sparse PCs still overall outperform those obtained by SPCA, rSVD, DSPCA and GPower$_{l_1}$.

From the above experiments, we may conclude that for the Pitprops data, there do not exist six highly sparse, nearly orthogonal and uncorrelated PCs while explaining most of variance. Therefore, the most acceptable sparse PCs seem to be the ones given in Table 8.

5.3 Gene expression data

In this subsection we test the performance of our approach ALSPCA for finding sparse PCs on the gene expression data. We also compare the results with the GPower methods [18], which are superior to the other existing methods [10,28,31] as demonstrated in [18].

The data set used in this subsection is the publicly available gene expression data from http://www.icbp.lbl.gov/breastcancer/, and described in Chin et al. [8], consisting of 19672 gene expression measurements on 89 samples (that is, $p = 19672$, $n = 89$). We aim to extract r number of PCs with around 80% zeros by ALSPCA and GPow-

Table 12 Performance on the gene expression data for $r = 5$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	34.77
GPower$_{l_1}$	80.14	7.56	0.348	22.17
GPower$_{l_0}$	79.70	5.47	0.223	22.79
GPower$_{l_1,m}$	79.64	7.39	0.274	22.68
GPower$_{l_0,m}$	80.36	12.47	0.452	22.23
ALSPCA	80.43	0.07	0.010	20.56

Table 13 Performance on the gene expression data for $r = 10$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	46.16
GPower$_{l_1}$	80.11	4.93	0.387	31.16
GPower$_{l_0}$	79.84	4.62	0.375	31.45
GPower$_{l_1,m}$	79.95	6.31	0.332	31.80
GPower$_{l_0,m}$	80.36	6.45	0.326	31.59
ALSPCA	80.51	0.01	0.017	29.85

Table 14 Performance on the gene expression data for $r = 15$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	53.27
GPower$_{l_1}$	79.56	4.73	0.253	38.29
GPower$_{l_0}$	79.84	4.02	0.284	38.32
GPower$_{l_1,m}$	79.39	5.94	0.347	38.31
GPower$_{l_0,m}$	79.99	5.18	0.307	38.19
ALSPCA	80.16	0.01	0.014	33.92

er methods [18] for $r = 5, 10, 15, 20, 25$, respectively. For all tests below, we set $\Delta_{ij} = 0.1$ for all $i \neq j$ for problem (55) and choose (63) as the termination criterion for ALSPCA with $\epsilon_E = 0.1$ and $\epsilon_O = 0.1$.

The sparsity, CPAV, non-orthogonality and correlation of the PCs obtained by the standard PCA, ALSPCA and GPower methods are presented in columns two to five of Tables 12, 13, 14, 15 and 16 for $r = 5, 10, 15, 20, 25$, respectively. In particular, the second and fifth columns of these tables respectively give sparsity (that is, the percentage of zeros in loadings) and CPAV. The third column reports non-orthogonality, which is measured by the maximum absolute difference between 90° and the angles formed by all pairs of loading vectors. Evidently, the smaller value in this column implies the better orthogonality. The fourth column presents the maximum correlation of PCs. It is clear that the standard PCs are completely dense. We also observe that the sparse PCs given by our method are almost uncorrelated and their loading vectors are nearly orthogonal, which are consistently much superior to the GPower methods. Though the CPAV for GPower methods is better than our method, the CPAV for GPower methods may not be a close measurement of the actual total explained variance as their sparse

Table 15 Performance on the gene expression data for $r = 20$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	59.60
GPower$_{l_1}$	79.51	4.37	0.280	43.30
GPower$_{l_0}$	80.16	4.52	0.245	43.12
GPower$_{l_1,m}$	79.61	4.48	0.317	42.98
GPower$_{l_0,m}$	80.40	4.18	0.255	43.25
ALSPCA	80.66	0.11	0.037	39.59

Table 16 Performance on the gene expression data for $r = 25$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	64.67
GPower$_{l_1}$	79.48	3.60	0.237	47.74
GPower$_{l_0}$	79.94	3.05	0.296	47.76
GPower$_{l_1,m}$	79.49	5.05	0.275	47.85
GPower$_{l_0,m}$	80.39	5.00	0.237	47.45
ALSPCA	80.68	0.02	0.021	43.66

PCs are highly correlated. But for our method, the sparse PCs are almost uncorrelated and thus the CPAV can measure well their actual total explained variance.

5.4 Random data

In this subsection we conduct experiments on a set of randomly generated data to test how the size of data matrix X, the sparsity controlling parameter ρ, and the number of components r affect the computational speed of our ALSPCA method.

First, we randomly generate 100 centered data matrices X with size $n \times p$ that is specified in the tables below. For all tests, we set $\Delta_{ij} = 0.1$ for all $i \neq j$ for problem (55) and choose (63) as the termination criterion for ALSPCA with $\epsilon_E = 0.1$ and $\epsilon_O = 0.1$. In the first test, we aim to extract *five* sparse PCs by ALSPCA with $\rho = 0.001, 0.01, 0.1, 1$, respectively. In the second test, we aim to extract 5 to 25 PCs with a fixed $\rho = 0.1$ by ALSPCA. In the third test, we fix the sparsity (that is, percentage of zeros) of the PC loadings to 80% and find r number of sparse PCs by ALSPCA with $r = 5, 10, 15, 20, 25$, respectively. The average CPU times (in seconds) of ALSPCA over the above 100 instances are reported in Tables 17, 18 and 19. We observe that ALSPCA is capable of solving all problems within a reasonable amount of time. It seems that the CPU time grows linearly as the problem size, sparsity controlling parameter ρ, and number of components r increase.

6 Concluding remarks

In this paper we proposed a new formulation of sparse PCA for finding sparse and nearly uncorrelated principal components (PCs) with orthogonal loading vectors while

Table 17 Average CPU time of ALSPCA on random data for $r = 5$

$n \times p$	$\rho = 0.001$	$\rho = 0.01$	$\rho = 0.1$	$\rho = 1$
50×500	0.4	0.8	1.2	4.9
100×1000	1.2	1.5	2.4	9.5
250×2500	3.7	4.4	13.3	38.8
500×5000	8.8	13.4	15.6	65.6
750×7500	13.6	24.0	33.2	96.3

Table 18 Average CPU time of ALSPCA on random data for $\rho = 0.1$

$n \times p$	$r = 5$	$r = 10$	$r = 15$	$r = 20$	$r = 25$
50×500	1.2	12.8	24.0	37.6	48.8
100×1000	2.4	16.9	28.7	40.8	144.0
250×2500	13.4	64.2	94.8	125.1	373.6
500×5000	16.5	85.5	141.9	186.6	553.1
750×7500	38.1	96.6	217.6	328.6	798.2

Table 19 Average CPU time of ALSPCA on random data for 80% sparsity

$n \times p$	$r = 5$	$r = 10$	$r = 15$	$r = 20$	$r = 25$
50×500	11.5	26.5	33.6	43.0	49.7
100×1000	15.2	29.3	57.8	83.7	102.7
250×2500	20.7	39.5	79.7	98.0	120.0
500×5000	41.5	60.3	91.4	143.1	197.0
750×7500	55.3	90.4	141.7	208.3	255.1

explaining as much of the total variance as possible. We also developed a novel globally convergent augmented Lagrangian method for solving a class of nonsmooth constrained optimization problems, which is well suited for our formulation of sparse PCA. Additionally, we proposed two nonmonotone gradient methods for solving the augmented Lagrangian subproblems, and established their global and local convergence. Finally, we compared our sparse PCA approach with several existing methods on synthetic and real data, respectively. The computational results demonstrate that the sparse PCs produced by our approach substantially outperform those by other methods in terms of total explained variance, correlation of PCs, and orthogonality of loading vectors.

As observed in our experiments, formulation (3) is very effective in finding the desired sparse PCs. However, there remains a natural theoretical question for it. Given a set of random variables, suppose there exist sparse and uncorrelated PCs with orthogonal loading vectors while explaining most of variance of the variables. In other words, their actual covariance matrix Σ has few dominant eigenvalues and the associated orthonormal eigenvectors are sparse. Since Σ is typically unknown and only approximated by a sample covariance matrix $\hat{\Sigma}$, one natural question is whether or not there exist some suitable parameters ρ and Δ_{ij} $(i \neq j)$ so that (3) is able to recover those sparse PCs almost surely as the sample size becomes sufficiently large.

In Sect. 4 we showed that Robinson's condition (7) holds at a set of feasible points of (55). We also observed from our experiments that the accumulation points of our augmented Lagrangian method lie in this set when applied to (55), and thus it converges. However, it remains open whether or not Robinson's condition holds at all feasible points of (55).

In addition, Burer and Monteiro [6] recently applied the classical augmented Lagrangian method to a nonconvex nonlinear program (NLP) reformulation of semidefinite programs (SDP) via low-rank factorization, and they obtained some nice computational results especially for the SDP relaxations of several hard combinatorial optimization problems. However, the classical augmented Lagrangian method generally cannot guarantee the convergence to a feasible point when applied to a nonconvex NLP. Due to this and [22], their approach [6] at least theoretically may not converge to a feasible point of the primal SDP. Given that the augmented Lagrangian method proposed in this paper converges globally under some mild assumptions, it would be interesting to apply it to the NLP reformulation of SDP and compare the performance with the approach studied in [6].

Finally, the MATLAB codes of our approach for solving the sparse PCA formulation (55) (or, equivalently, (3)) are available online at www.math.sfu.ca/~zhaosong. As a future research, we will further improve their performance by conducting more extensive computational experiments and exploring more practical applications.

Acknowledgments The authors are in debt to the anonymous referees for numerous insightful comments and suggestions, which have greatly improved the paper. In addition, we would like to thank one of the referees for pointing out Theorem 3.17 of Chapter 1 of [14], which substantially shortens our original proof of Theorem 2.2.

References

1. Alter, O., Brown, P., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. **97**, 10101–10106 (2000)
2. Barzilai, J., Borwein, J.M.: Two point step size gradient methods. IMA J. Numer. Anal. **8**, 141–148 (1988)
3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
4. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific (1999)
5. Birgin, E.G., Martínez, J.M., Raydan, M.: Nonmonotone spectral projected gradient methods on convex sets. SIAM J. Optim. **10**, 1196–1211 (2000)
6. Burer, S., Monteiro, R.D.C.: A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. Math. Program. Ser. B **95**, 329–357 (2003)
7. Cadima, J., Jolliffe, I.: Loadings and correlations in the interpretation of principal components. J. Appl. Stat. **22**, 203–214 (1995)
8. Chin, K., Devries, S., Fridlyand, J., Spellman, P., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R., Qian, Z., Ryder, T.: Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell **10**, 529C–541C (2006)
9. d'Aspremont, A., Bach, F.R., El Ghaoui, L.: Optimal solutions for sparse principal component analysis. J. Mach. Learn. Res. **9**, 1269–1294 (2008)
10. d'Aspremont, A., El Ghaoui, L., Jordan, M.I., Lanckriet, G.R.G.: A direct formulation for sparse PCA using semidefinite programming. SIAM Rev. **49**, 434–448 (2007)
11. Hancock, P., Burton, A., Bruce, V.: Face processing: human perception and principal components analysis. Memory Cogn. **24**, 26–40 (1996)

12. Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L., Botstein, D.: ǵene Shavingás a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. **1**, 1–21 (2000)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, NewYork (2001)
14. Helmke, U., Moore, J.B.: Optimization and Dynamical Systems. Springer, London and New York (1994)
15. Hiriart-Urruty, J.B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms I. Comprehensive Study in Mathematics, vol. 305. Springer, New York (1993)
16. Jeffers, J.: Two case studies in the application of principal component. Appl. Stat. **16**, 225–236 (1967)
17. Jolliffe, I.: Rotation of principal components: choice of normalization constraints. J. Appl. Stat. **22**, 29–35 (1995)
18. Journée, M., Nesterov, Yu., Richtárik, P., Sepulchre, R.: Generalized power method for sparse principal component analysis. J. Mach. Learn. Res. **11**, 517–553 (2010)
19. Jolliffe, I.T., Trendafilov, N.T., Uddin, M.L.: A modified principal component technique based on the Lasso. J. Comput. Graph. Stat. **12**, 531–547 (2003)
20. Lu, Z., Zhang Y.: An Augmented Lagrangian Approach for Sparse Principal Component Analysis. Technical report, Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada, July (2009)
21. Moghaddam, B., Weiss, Y., Avidan, S. : Spectral bounds for sparse PCA: exact and greedy algorithms. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) Advances in Neural Information Processing Systems 18, pp. 915–922. MIT Press, Cambridge (2006)
22. Monteiro, R.D.C.: Private communication (2009)
23. Nesterov, Y.E.: Gradient methods for minimizing composite objective functions. CORE Discussion paper 2007/76, September 2007
24. Robinson, S.M.: Stability theory for systems of inequalities, Part 2: Differentiable nonlinear systems. SIAM J. Numer. Anal. **13**, 497–513 (1976)
25. Robinson, S.M.: Local structure of feasible sets in nonlinear programming, Part I: regularity. In: Pereira, V., Reinoza, A. (eds.) Numerical Methods Lecture Notes in Mathematics vol. 1005, Springer, Berlin (1983)
26. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
27. Ruszczyński, A.: Nonlinear Optimization. Princeton University Press, Princeton (2006)
28. Shen, H., Huang, J.Z.: Sparse principal component analysis via regularized low rank matrix approximation. J. Multivar. Anal. **99**(6), 1015–1034 (2008)
29. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. **117**, 387–423 (2009)
30. Wright, S.J., Nowak, R., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. **57**(3), 2479–2493 (2009)
31. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. J. Comput. Graph. Stat. **15**(2), 265–286 (2006)