An Augmented Lagrangian Approach for Sparse Principal Component Analysis

Zhaosong Lu[*]   Yong Zhang [†]

July 12, 2009

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 literature [15, 6, 17, 28, 8, 25, 18, 7, 16]. 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 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, random, 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.

Key words: sparse PCA, augmented Lagrangian method, nonmonotone gradient methods, nonsmooth minimization

AMS 2000 subject classification: 62H20, 62H25, 62H30, 90C30, 65K05

---

[*]Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. (email: zhaosong@sfu.ca). This author was supported in part by NSERC Discovery Grant.

[†]Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. (email: yza30@sfu.ca).

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 [10, 12, 1, 11]).

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

$$\hat{\Sigma} = VDV^T.$$

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, [28]). 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 Section 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 [28]):

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 [15] applied various rotation techniques to the standard PCs for obtaining sparse loading vectors. Cadima and Jolliffe [6] 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. [17] 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. [28] 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. [8] proposed a method, called DSPCA, for finding sparse PCs by solving a sequence of semidefinite program relaxations of sparse PCA. Shen and Huang [25] 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. [16] 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. [18] and d'Aspremont et al. [7].

The PCs obtained by the above methods [15, 6, 17, 28, 8, 25, 18, 7, 16] 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 [17].

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, un-correlation 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 per-turbation 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 Section 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 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 [26], local linear rate of convergence for these gradient methods. Finally, we compare the sparse PCA approach proposed in this paper with several existing methods [28, 8, 25, 16] on synthetic, random, and real data, 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.

The rest of paper is organized as follows. In Section 2, we propose a new formulation for sparse PCA and explore the connection of this formulation with the standard PCA. In Section 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 Section 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, random, and real data in Section 5. Finally, we present some concluding remarks in Section 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 $\text{sign}(X)$ 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 $\text{rank}(X)$. 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 $X \bullet Y := \text{Tr}(XY^T)$, where $\text{Tr}(\cdot)$ 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 $v \in \Re^n$, $\text{Diag}(v)$ or $\text{Diag}(v_1, \ldots, v_n)$ denotes a diagonal matrix whose ith diagonal element is v_i for $i = 1, \ldots, n$. Given an $n \times n$ matrix X, $\widetilde{\text{Diag}}(X)$ 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 $\text{dist}(\cdot, C) : \mathcal{U} \to \Re_+$ denote the distance function to C measured in terms of $\|\cdot\|$, that is,

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

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,

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

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:

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

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),

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

5

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

We next explore the connection of formulation (3) with the standard PCA. Before proceeding, we state a technical lemma as follows that will be used subsequently. Its proof can be found in [20].

Lemma 2.1 *Given any $\hat{\Sigma} \in \mathcal{S}^n$ and integer $1 \leq r \leq n$, define*

$$\underline{i}_r = \max\{1 \leq i \leq n : \lambda_i(\hat{\Sigma}) > \lambda_r(\hat{\Sigma})\}, \qquad \bar{i}_r = \max\{1 \leq i \leq n : \lambda_i(\hat{\Sigma}) = \lambda_r(\hat{\Sigma})\}, \quad (4)$$

and let f^* be the optimal value of

$$\max\{\operatorname{Tr}(\hat{\Sigma}Y) : \ 0 \preceq Y \preceq I, \ \operatorname{Tr}(Y) = r\}. \quad (5)$$

Then, $f^* = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$, and Y^* is an optimal solution of (5) if and only if $Y^* = U_1^* U_1^{*T} + U_2^* P^* U_2^{*T}$, where $P^* \in \mathcal{S}^{\bar{i}_r - \underline{i}_r}$ satisfies $0 \preceq P^* \preceq I$ and $\operatorname{Tr}(P^*) = r - \underline{i}_r$, and $U_1^* \in \Re^{n \times \underline{i}_r}$ and $U_2^* \in \Re^{n \times (\bar{i}_r - \underline{i}_r)}$ are the matrices whose columns consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to the eigenvalues $(\lambda_1(\hat{\Sigma}), \ldots, \lambda_{\underline{i}_r}(\hat{\Sigma}))$ and $(\lambda_{\underline{i}_r+1}(\hat{\Sigma}), \ldots, \lambda_{\bar{i}_r}(\hat{\Sigma}))$, respectively.

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 $i \neq j$.

Proposition 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 $\operatorname{Tr}(U^T \hat{\Sigma} U) = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$. 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

$$\max\{\operatorname{Tr}(V^T \hat{\Sigma} V) : \ V^T V = I, \ V \in \Re^{n \times r}\}.$$

We know from [9] that its optimal value equals $\sum_{i=1}^r \lambda_i(\hat{\Sigma})$. Therefore, $f^* = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$ holds and U is an optimal solution of (3). It also implies that the "if" part of this proposition holds. We next show that the "only if" part also holds. Let $V^* \in \Re^{n \times r}$ be an optimal solution of (3), and define $Y^* = V^* V^{*T}$. Then, we have $V^{*T} V^* = I$, which yields $0 \preceq Y^* \preceq I$ and $\operatorname{Tr}(Y^*) = r$. Hence, Y^* is a feasible solution of (5). Using the fact that $f^* = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$, we then have

$$\operatorname{Tr}(\hat{\Sigma} Y^*) = \operatorname{Tr}(V^{*T} \hat{\Sigma} V^*) = \sum_{i=1}^r \lambda_i(\hat{\Sigma}),$$

which together with Lemma 2.1 implies that Y^* is an optimal solution of (5). Let $\underline{i}_r$ and $\bar{i}_r$ be defined in (4). Then, it follows from Lemma 2.1 that $Y^* = U_1^* U_1^{*T} + U_2^* P^* U_2^{*T}$, where

6

$P^* \in \mathcal{S}^{\bar{\imath}_r - \underline{\imath}_r}$ satisfies $0 \preceq P^* \preceq I$ and $\mathrm{Tr}(P^*) = r - \underline{\imath}_r$, and $U_1^* \in \Re^{n \times \underline{\imath}_r}$ and $U_2^* \in \Re^{n \times (\bar{\imath}_r - \underline{\imath}_r)}$ are the matrices whose columns consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to the eigenvalues $(\lambda_1(\hat{\Sigma}), \ldots, \lambda_{\underline{\imath}_r}(\hat{\Sigma}))$ and $(\lambda_{\underline{\imath}_r + 1}(\hat{\Sigma}), \ldots, \lambda_{\bar{\imath}_r}(\hat{\Sigma}))$, respectively. Thus, we have

$$\hat{\Sigma} U_1^* = U_1^* \Lambda, \quad \hat{\Sigma} U_2^* = \lambda_r(\hat{\Sigma}) U_2^*, \tag{6}$$

where $\Lambda = \mathrm{Diag}(\lambda_1(\hat{\Sigma}), \ldots, \lambda_{\underline{\imath}_r}(\hat{\Sigma}))$. In addition, it is easy to show that $\mathrm{rank}(Y^*) = \underline{\imath}_r + \mathrm{rank}(P^*)$. Since $Y^* = V^* V^{*T}$ and $V^{*T} V^* = I$, we can observe that $\mathrm{rank}(Y^*) = r$. Hence, $\mathrm{rank}(P^*) = r - \underline{\imath}_r$, which implies that P^* has only $r - \underline{\imath}_r$ nonzero eigenvalues. Using this fact and the relations $0 \preceq P^* \preceq I$ and $\mathrm{Tr}(P^*) = r - \underline{\imath}_r$, we can further conclude that $r - \underline{\imath}_r$ eigenvalues of P^* are 1 and the rest are 0. Therefore, there exists $W \in \Re^{(\bar{\imath}_r - \underline{\imath}_r) \times (r - \underline{\imath}_r)}$ such that

$$W^T W = I, \quad P^* = W W^T. \tag{7}$$

It together with $Y^* = U_1^* U_1^{*T} + U_2^* P^* U_2^{*T}$ implies that $Y^* = U^* U^{*T}$, where $U^* = [U_1^* \ U_2^* W]$. In view of (7) and the identities $U_1^{*T} U_1^* = I$, $U_2^{*T} U_2^* = I$ and $U_1^{*T} U_2^* = 0$, we see that $U^{*T} U^* = I$. Using this result, and the relations $V^{*T} V^* = I$ and $Y^* = U^* U^{*T} = V^* V^{*T}$, it is not hard to see that the columns of U^* and V^* form an orthonormal basis for the range space of Y^*, respectively. Thus, $V^* = U^* Q$ for some $Q \in \Re^{r \times r}$ satisfying $Q^T Q = I$. Now, let $D = V^{*T} \hat{\Sigma} V^*$. By the definition of V^*, we know that D is an $r \times r$ diagonal matrix. Moreover, in view of (6), (7), the definition of U^*, and the relations $V^* = U^* Q$, $U_1^{*T} U_1^* = I$, $U_2^{*T} U_2^* = I$ and $U_1^{*T} U_2^* = 0$, we have

$$
\begin{aligned}
D &= V^{*T} \hat{\Sigma} V^* = Q^T U^{*T} \hat{\Sigma} U^* Q = Q^T \begin{bmatrix} U_1^{*T} \\ W^T U_2^{*T} \end{bmatrix} \hat{\Sigma} [U_1^* \ U_2^* W] Q \\
&= Q^T \begin{bmatrix} \Lambda & 0 \\ 0 & \lambda_r(\hat{\Sigma}) I \end{bmatrix} Q,
\end{aligned} \tag{8}
$$

which together with $Q^T Q = I$ implies that D is similar to the diagonal matrix appearing on the right-hand side of (8). Hence, the diagonal elements of D consist of r largest eigenvalues of $\hat{\Sigma}$. In addition, let $Q_1 \in \Re^{\underline{\imath}_r \times r}$ and $Q_2 \in \Re^{(r - \underline{\imath}_r) \times r}$ be the submatrices corresponding to the first $\underline{\imath}_r$ and the last $r - \underline{\imath}_r$ rows of Q, respectively. Then, in view of the definition of U^* and $V^* = U^* Q$, we have

$$[U_1^* \ U_2^* W] = U^* = V^* Q^T = [V^* Q_1^T \ V^* Q_2^T].$$

Thus, we obtain that $U_1^* = V^* Q_1^T$ and $U_2^* W = V^* Q_2^T$. Using these identities, (6), (8), and the relation $V^* = U^* Q$, we have

$$
\begin{aligned}
\hat{\Sigma} V^* &= \hat{\Sigma} U^* Q = \hat{\Sigma} [U_1^* \ U_2^* W] Q = [U_1^* \Lambda \ \lambda_r(\hat{\Sigma}) U_2^* W] Q \\
&= [V^* Q_1^T \Lambda \ \lambda_r(\hat{\Sigma}) V^* Q_2^T] Q = V^* Q^T \begin{bmatrix} \Lambda & 0 \\ 0 & \lambda_r(\hat{\Sigma}) I \end{bmatrix} Q = V^* D.
\end{aligned}
$$

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

From the above proposition, 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 Sections 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 Subsection 3.1. In Subsection 3.2, we develop an augmented Lagrangian method and establish its global convergence. In Subsection 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

$$
\begin{aligned}
\min \quad & f(x) + P(x) \\
\text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m, \\
& h_i(x) = 0, \quad i = 1, \ldots, p, \\
& x \in X.
\end{aligned}
\tag{9}
$$

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 (9).

Before establishing first-order optimality conditions for problem (9), we describe a general constraint qualification condition for (9), that is, Robinson's condition that was proposed in [21].

Let $x \in \Re^n$ be a feasible point of problem (9). 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

$$\left\{ \begin{bmatrix} g'(x)d - v \\ h'(x)d \end{bmatrix} : d \in T_X(x), v \in \Re^m, v_i \leq 0, i \in \mathcal{A}(x) \right\} = \Re^m \times \Re^p, \qquad (10)$$

where $g'(x)$ 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 [21, 22, 24].

The following proposition demonstrates that Robinson's condition is indeed a constraint qualification condition for problem (9). For the sake of completeness, we include a brief proof for it.

Proposition 3.1 *Given a feasible point $x \in \Re^n$ of problem (9), let $T_\Omega(x)$ be the tangent cone to Ω at x, and $(T_\Omega(x))^\circ$ be its polar cone. If Robinson's condition (10) holds at x, then*

$$T_\Omega(x) = \left\{ d \in T_X(x) : \begin{array}{ll} d^T \nabla g_i(x) \leq 0, & i \in \mathcal{A}(x), \\ d^T \nabla h_i(x) = 0, & i = 1, \ldots, p \end{array} \right\},$$

$$(T_\Omega(x))^\circ = \left\{ \sum_{i \in \mathcal{A}(x)} \lambda_i \nabla g_i(x) + \sum_{i=1}^{p} \mu_i \nabla h_i(x) + N_X(x) : \lambda \in \Re^m_+, \ \mu \in \Re^p \right\}, \qquad (11)$$

where $T_X(x)$ and $N_X(x)$ are the tangent and normal cones to X at x, respectively.

Proof. By Theorem A.10 of [24], we see that Robinson's condition (10) implies that the assumption of Theorem 3.15 of [24] is satisfied with

$$x_0 = x, \quad X_0 = X, \quad Y_0 = \Re^m_- \times \Re^p, \quad g(\cdot) = (g_1(\cdot); \ldots; g_m(\cdot); h_1(\cdot); \ldots; h_p(\cdot)).$$

The first statement then follows from Theorem 3.15 of [24] with the above x_0, X_0, Y_0 and $g(\cdot)$. Further, let $A(x)$ denote the matrix whose rows are the gradients of all active constraints at x in the same order as they appear in (9). Then, Robinson's condition (10) implies that the assumptions of Theorem 2.36 of [24] are satisfied with

$$A = A(x), \quad K_1 = T_X(x), \quad K_2 = \Re^{|\mathcal{A}(x)|}_- \times \Re^p.$$

Let $K = \{d \in K_1 : Ad \in K_2\}$. Then, it follows from Theorem 2.36 of [24] that

$$(T_\Omega(x))^\circ = K^\circ = K_1^\circ + \{A^T \xi : \xi \in K_2^\circ\},$$

which together with the identity $(T_X(x))^\circ = N_X(x)$ and the definitions of A, K_1 and K_2, implies that the second statement holds. ∎

We are now ready to establish first-order optimality conditions for problem (9).

Theorem 3.2 *Let $x^* \in \Re^n$ be a local minimizer of problem (9). Assume that Robinson's condition (10) is satisfied at x^*. Then there exist Lagrange multipliers $\lambda \in \Re_+^m$ and $\mu \in \Re^p$ such that*

$$0 \in \nabla f(x^*) + \partial P(x^*) + \sum_{i=1}^m \lambda_i \nabla g_i(x^*) + \sum_{i=1}^p \mu_i \nabla h_i(x^*) + N_X(x^*), \tag{12}$$

and

$$\lambda_i g_i(x^*) = 0, \quad i = 1, \ldots, m. \tag{13}$$

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

$$d^T \nabla f(x^*) + P'(x^*; d) \geq 0 \quad \forall d \in T_\Omega(x^*). \tag{14}$$

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

$$f(x^* + t_k d) - f(x^k) = o(t_k), \quad P(x^* + t_k d) - P(x^k) = o(t_k), \tag{15}$$

where the first equality follows from the Mean Value Theorem while the second one comes from Theorem 10.4 of [23]. Clearly, $x^k \to x^*$. This together with the assumption that x^* is a local minimizer of (9), implies that

$$f(x^k) + P(x^k) \geq f(x^*) + P(x^*) \tag{16}$$

when k is sufficiently large. In view of (15) and (16), 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 (14) 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 $d^T y \le 0$ for any $y \in T_\Omega^\circ$, and $d^T y \ge 1$ for any $y \in S$. Clearly, we see that $d \in (T_\Omega^\circ)^\circ = T_\Omega(x^*)$, and

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

which contradicts (14). Hence, we have $S \cap T_\Omega^\circ \ne \emptyset$. Using this relation, (11), the definitions of S and $\mathcal{A}(x^*)$, and letting $\lambda_i = 0$ for $i \notin \mathcal{A}(x^*)$, we easily see that (12) and (13) 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

$$0 = \nabla f(x^*) + z^k + \sum_{i=1}^m \lambda_i^k \nabla g_i(x^*) + \sum_{i=1}^p \mu_i^k \nabla h_i(x^*) + v^k \tag{17}$$

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 (17) by $\|(\lambda^k, \mu^k)\|$ and taking limits on a subsequence if necessary, we obtain that

$$0 = \sum_{i=1}^m \bar\lambda_i \nabla g_i(x^*) + \sum_{i=1}^p \bar\mu_i \nabla h_i(x^*) + \bar v \tag{18}$$

for some $\bar v \in N_X(x^*)$. Since Robinson's condition (10) is satisfied at x^*, there exist $d \in T_X(x^*)$ and $v \in \Re^m$ such that $v_i \le 0$ for $i \in \mathcal{A}(x^*)$, and

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

Using these relations, (18) 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

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

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 (9)

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 []). Nevertheless, when problem (9) 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. To alleviate this drawback, we propose a novel augmented Lagrangian method for problem (9) and establish its global convergence in this subsection.

Throughout this subsection, we make the following assumption for problem (9).

Assumption 1 *Problem (9) is feasible, and moreover at least a feasible solution, denoted by x^{feas}, is known.*

It is well-known that for problem (9) the associated augmented Lagrangian function $L_\varrho(x, \lambda, \mu) : \Re^n \times \Re^m \times \Re^p \to \Re$ is given by

$$L_\varrho(x, \lambda, \mu) := w(x) + P(x), \tag{19}$$

where

$$w(x) := f(x) + \frac{1}{2\varrho}(\|[\lambda + \varrho g(x)]^+\|^2 - \|\lambda\|^2) + \mu^T h(x) + \frac{\varrho}{2}\|h(x)\|^2, \tag{20}$$

and $\varrho > 0$ is a penalty parameter (e.g., see [3, 24]). Roughly speaking, an augmented Lagrangian method, when applied to problem (9), 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 (9) (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 decreasing 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

$$\min_{x \in X} L_{\varrho_k}(x, \lambda^k, \mu^k) \tag{21}$$

such that

$$\text{dist}\left(-\nabla w(x^k), \partial P(x^k) + N_X(x^k)\right) \leq \epsilon_k, \quad L_{\varrho_k}(x^k, \lambda^k, \mu^k) \leq \Upsilon. \tag{22}$$

2) Update Lagrange multipliers according to

$$\lambda^{k+1} := [\lambda^k + \varrho_k g(x^k)]^+, \quad \mu^{k+1} := \mu^k + \varrho_k h(x^k). \tag{23}$$

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 bounded from above (see Step 1)); and ii) the magnitude of penalty parameters outgrows that of Lagrangian multipliers (see Step 2)).

In addition, to make the above augmented Lagrangian method complete, we need address how to find an approximate solution $x^k \in X$ for subproblem (21) satisfying (22) 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 (9).

Theorem 3.3 *Assume that $\epsilon_k \to 0$. Let $\{x^k\}$ be the sequence generated by the above augmented Lagrangian method satisfying (22). 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 (9);*

(b) *Further, if Robinson's condition (10) 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 the vector of Lagrange multipliers satisfying the first-order optimality conditions (12)-(13) at x^*.*

Proof. In view of (19), (20) and the second relation in (22), we have

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

It follows that

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

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.

13

We next show that statement (b) also holds. Using (21), (19), (20), (23), and the first relation in (22), we have

$$\|\nabla f(x^k) + (\lambda^{k+1})^T \nabla g(x^k) + (\mu^{k+1})^T \nabla h(x^k) + z^k + v^k\| \leq \epsilon_k \tag{24}$$

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 [13] 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 (24) 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 (24) by $\|(\lambda^{k+1}, \mu^{k+1})\|$ and taking limits as $k \in k \to \infty$, we obtain that

$$\bar{\lambda}^T \nabla g(x^*) + \bar{\mu}^T \nabla h(x^*) + \bar{v} = 0. \tag{25}$$

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, (25), Robinson's condition (10) 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 (24) and the fact $\{z^k\}_{k\in K}$ is bounded, we immediately see that $\{v^k\}_{k\in K}$ is bounded. Using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [23] and Lemma 2.42 of [24]), 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 (24), 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 (12)-(13) at x^*. ∎

Before ending this subsection, we now briefly discuss how to find an approximate solution $x^k \in X$ for subproblem (21) satisfying (22) as required in Step 1) of the above augmented Lagrangian method. In particular, we are interested in applying the nonmonotone gradient methods proposed in Subsection 3.3 to (21). As shown in Subsection 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 (22). Moreover, if an initial point for these methods is properly chosen, the obtained approximate solution x^k also satisfies the second relation of (22). For example, given $k \geq 0$, let $x_{\text{init}}^k \in X$ denote the initial point for solving the kth subproblem (21), and we define

14

x^k_{init} for $k \geq 1$ as follows

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

where x^{k-1} is the approximate solution to the $(k-1)$th subproblem (21) satisfying (22) (with k replaced by $k-1$). Recall from Assumption 1 that x^{feas} is a feasible solution of (9). Thus, $g(x^{\text{feas}}) \leq 0$, and $h(x^{\text{feas}}) = 0$, which together with (19), (20) and the definition of Υ implies that

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

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

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

and so the second relation of (22) 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 Subsection 3.2. We also establish global convergence and local linear rate of convergence for these methods. It shall be mentioned that these two methods are closely related to the ones proposed in [27] and [26], 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 [4] for smooth problems, but the methods proposed in [27] and [26] cannot.

Throughout this subsection, we consider the following problem

$$\min_{x \in X}\{F(x) := f(x) + P(x)\}, \tag{26}$$

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.

We say that $x \in \Re^n$ is a *stationary point* of problem (26) if $x \in X$ and

$$0 \in \nabla f(x) + \partial P(x) + N_X(x). \tag{27}$$

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

$$d_H(x) := \arg\min_d \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + P(x + d) : \ x + d \in X \right\}. \tag{28}$$

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 (26) if and only if $d_H(x) = 0$.*

Proof. We first observe that (28) 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 (28). ∎

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*

$$\|\tilde{d}\| \leq \frac{1 + \lambda_{\max}(Q) + \sqrt{1 - 2\lambda_{\min}(Q) + \lambda_{\max}(Q)^2}}{2\lambda_{\min}(\tilde{H})} \lambda_{\max}(H)\|d\|, \qquad (29)$$

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

Proof. The conclusion immediately follows from Lemma 3.2 of [26] 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
$$\|\nabla f(y) - \nabla f(z)\| \leq L\|y - z\| \quad \forall y, z \in \Re^n \qquad (30)$$
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 (30), then the descent condition
$$F(x + d) \leq F(x) + \sigma\Delta_d$$
is satisfied for $d = d_{H(\alpha)}(x)$, where $H(\alpha) = \alpha H$, provided $\alpha \geq L/[2(1 - \sigma)\lambda_{\min}(H)]$.

16

Proof. The statements (a)-(c) follow from Theorem 4.1 (a) and Lemma 3.4 of [26] with $J = \{1, \ldots, n\}$, $\gamma = 0$, and $\underline{\lambda} = \lambda_{\min}(H)$. We now prove statement (d). Let $d = d_{H(\alpha)}(x)$, where $H(\alpha) = \alpha H$ for $\alpha > 0$. It follows from statement (a) that $\|d\|^2 \leq -\Delta_d/(\alpha\lambda_{\min}(H))$. Using this relation, (30), and the definitions of F and Δ_d, we have

$$
\begin{aligned}
F(x + d) - F(x) &= f(x + d) - f(x) + P(x + d) - P(x) \\
&= \nabla f(x)^T d + P(x + d) - P(x) + \int_0^1 d^T(\nabla f(x + td) - \nabla f(x))dt \\
&\leq \Delta_d + \int_0^1 \|\nabla f(x + td) - \nabla f(x)\|\|d\|dt \\
&\leq \Delta_d + L\|d\|^2/2 \leq [1 - L/(2\alpha\lambda_{\min}(H))]\Delta_d,
\end{aligned}
$$

which together with $\alpha \geq L/[2(1 - \sigma)\lambda_{\min}(H)]$, immediately implies that statement (d) holds. ∎

We now present the first nonmonotone gradient method for (26) as follows.

Nonmonotone gradient method I:

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

1) Choose $\alpha_k^0 \in [\underline{\alpha}, \bar{\alpha}]$ and $\underline{\lambda}I \preceq H_k \preceq \bar{\lambda}I$.

2) For $j = 0, 1, \ldots$

    2a) Let $\alpha_k = \alpha_k^0 \eta^j$. Solve (28) with $x = x^k$ and $H = \alpha_k H_k$ to obtain $d^k = d_H(x)$.

    2b) If d^k satisfies
$$
F(x^k + d^k) \leq \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma\Delta_k, \tag{31}
$$

    go to step 3), where

$$
\Delta_k := \nabla f(x^k)^T d^k + P(x^k + d^k) - P(x^k). \tag{32}
$$

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 [27]. They differ from each other only in that the distinct Δ_k's are used in (31). Indeed, the latter method uses $\Delta_k = -\alpha_k\|d^k\|^2/2$. Nevertheless, for global convergence, the method [27] needs a strong assumption that the function f is *Lipschitz* continuously differentiable, which is not required for our method (see Theorem 3.9). In addition, our method can be viewed as an extension of the first projected gradient method (namely, SPG1) studied in [4] for smooth problems, but their method cannot. Finally, local convergence is established for our method (see Theorem 3.10), but not studied for their method. ∎

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 $\alpha_k > 0$ in step 2a) so that (31) 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 (26). Then, there exists $\tilde{\alpha} > 0$ such that $d^k = d_{H_k(\alpha_k)}(x^k)$, where $H_k(\alpha_k) = \alpha_k H_k$, satisfies (31) whenever $\alpha_k \geq \tilde{\alpha}$.*

Proof. For simplicity of notation, let $d(\alpha) = d_{H_k(\alpha)}(x^k)$, where $H_k(\alpha) = \alpha H_k$ for any $\alpha > 0$. Then, it follows from (28) that for all $\alpha > 0$,

$$\alpha \|d(\alpha)\| \leq -\frac{2[\nabla f(x^k)^T d(\alpha) + P(x^k + d(\alpha)) - P(x^k)]}{\lambda_{\min}(H_k)\|d(\alpha)\|} \leq -\frac{2F'(x^k, d(\alpha)/\|d(\alpha)\|)}{\lambda_{\min}(H_k)}. \tag{33}$$

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

$$\liminf_{\alpha \to \infty} \alpha\|d(\alpha)\| > 0. \tag{34}$$

Suppose not. Then there exists a sequence $\{\bar{\alpha}_l\} \uparrow \infty$ such that $\bar{\alpha}_l\|d(\bar{\alpha}_l)\| \to 0$ as $l \to \infty$. Invoking that $d(\bar{\alpha}_l)$ is the optimal solution of (28) with $x = x^k$, $H = \bar{\alpha}_l H_k$ and $\alpha = \bar{\alpha}_l$, we have

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

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 [23] and Lemma 2.42 of [24]), and the relations $\|d(\bar{\alpha}_l)\| \to 0$ and $\bar{\alpha}_l\|d(\bar{\alpha}_l)\| \to 0$, we see that (27) holds at x^k, which contradicts the nonstationarity of x^k. Hence, (34) holds. We observe that

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

which together with (34) and $H_k \succ 0$, implies that $\|d(\alpha)\| = O\left(\alpha d(\alpha)^T H_k d(\alpha)\right)$ as $\alpha \to \infty$. In addition, using the boundedness of $\tilde{S}$ and $H_k \succ 0$, we have $\alpha d(\alpha)^T H_k d(\alpha) = O\left(\|d(\alpha)\|\right)$ as $\alpha \to \infty$. Thus, we obtain that

$$\alpha d(\alpha)^T H_k d(\alpha) = \Theta(\|d(\alpha)\|) \quad \text{as } \alpha \to \infty. \tag{35}$$

This relation together with Lemma 3.6(a) implies that

$$P(x^k) - \nabla f(x^k)^T d(\alpha) - P(x^k + d(\alpha)) \geq \alpha d(\alpha)^T H_k d(\alpha) = \Theta(\|d(\alpha)\|). \tag{36}$$

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

$$
\begin{aligned}
F(x^k + d(\alpha)) - \max_{[k-M]^+ \leq i \leq k} F(x^i) &\leq F(x^k + d(\alpha)) - F(x^k) \\
&= f(x^k + d(\alpha)) - f(x^k) + P(x^k + d(\alpha)) - P(x^k) \\
&= \nabla f(x^k)^T d(\alpha) + P(x^k + d(\alpha)) - P(x^k) + o(\|d(\alpha)\|) \\
&\leq \sigma[\nabla f(x^k)^T d(\alpha) + P(x^k + d(\alpha)) - P(x^k)], \quad (37)
\end{aligned}
$$

18

provided α 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

$$\lim_{k\to\infty} F(x^{l(k)}) = F^* \tag{38}$$

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

$$\lim_{k\to\infty} d^{l(k)-j} = 0, \qquad \lim_{k\to\infty} F(x^{l(k)-j}) = F^*. \tag{39}$$

Using (31) and (32) with k replaced by $l(k) - 1$, we obtain that

$$F(x^{l(k)}) \leq F(x^{l(l(k)-1)}) + \sigma\Delta_{l(k)-1}. \tag{40}$$

Replacing k and α by $l(k) - 1$ and $\alpha_{l(k)-1}$ in (36), respectively, and using $H_{l(k)-1} \succeq \underline{\lambda}I$ and the definition of $\Delta_{l(k)-1}$ (see (32)), we have

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

The above two inequalities yield that

$$F(x^{l(k)}) \leq F(x^{l(l(k)-1)}) - \sigma\underline{\lambda}\alpha_{l(k)-1}\|d^{l(k)-1}\|^2, \tag{41}$$

which together with (38) implies that $\lim_{k\to\infty} \alpha_{l(k)-1}\|d^{l(k)-1}\|^2 = 0$. Further, noticing that $\alpha_k \geq \underline{\alpha}$ for all k, we obtain that $\lim_{k\to\infty} d^{l(k)-1} = 0$. Using this result and (38), we have

$$\lim_{k\to\infty} F(x^{l(k)-1}) = \lim_{k\to\infty} F(x^{l(k)} - d^{l(k)-1}) = \lim_{k\to\infty} F(x^{l(k)}) = F^*, \tag{42}$$

where the second equality follows from uniform continuity of F in $\mathcal{L}$. Therefore, (39) holds for $j = 1$. We now need to show that if (39) holds for j, then it also holds for $j + 1$. Using a similar argument as that leading to (41), we have

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

19

which together with (38), the induction assumption $\lim_{k\to\infty} F(x^{l(k)-j}) = F^*$, and the fact that $\alpha_{l(k)-j-1} \geq \underline{\alpha}$ 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 (42), we can show that $\lim_{k\to\infty} F(x^{l(k)-j-1}) = F^*$. Hence, (39) 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 (39), 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, (39), 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 (26).*

Proof. Suppose for contradiction that x^* is an accumulation point of $\{x^k\}$ that is a non-stationary point of (26). Let K be the subsequence such that $\{x^k\}_{k \in K} \to x^*$. We first claim that $\{\alpha_k\}_{k \in K}$ is bounded. Suppose not. Then there exists a subsequence of $\{\alpha_k\}_{k \in K}$ that goes to ∞. Without loss of generality, we assume that $\{\alpha_k\}_{k \in K} \to \infty$. For simplicity of notations, let $\bar{\alpha}_k = \alpha_k/\eta$, $d^k(\alpha) = d_{H_k(\alpha)}(x^k)$ for $k \in K$ and $\alpha > 0$, where $H_k(\alpha) = \alpha H_k$. Since $\{\alpha_k\}_{k \in K} \to \infty$ and $\alpha_k^0 \leq \bar{\alpha}$, there exists some index $\bar{k} \geq 0$ such that $\alpha_k > \alpha_k^0$ for all $k \in K$ with $k \geq \bar{k}$. By the particular choice of α_k specified in steps (2a) and (2b), we have

$$F(x^k + d^k(\bar{\alpha}_k)) > \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma[\nabla f(x^k)^T d^k(\bar{\alpha}_k) + P(x^k + d^k(\bar{\alpha}_k)) - P(x^k)], \quad (43)$$

Using a similar argument as that leading to (33), we have

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

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

$$\liminf_{k \in K, k \to \infty} \bar{\alpha}_k \|d^k(\bar{\alpha}_k)\| > 0. \quad (44)$$

Suppose not. By passing to a subsequence if necessary, we can assume that $\{\bar{\alpha}_k \|d^k(\bar{\alpha}_k)\|\}_{k \in K} \to 0$. Invoking that $d^k(\bar{\alpha}_k)$ is the optimal solution of (28) with $x = x^k$ and $H = \bar{\alpha}_k H_k$, we have

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

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 [23] and Lemma 2.42 of [24]), the relations $\underline{\lambda}I \preceq H_k \preceq \bar{\lambda}I$, $\{\|d^k(\bar{\alpha}_k)\|\}_{k \in K} \to 0$, $\{\bar{\alpha}_k\|d^k(\bar{\alpha}_k)\|\}_{k \in K} \to 0$ and $\{x^k\}_{k \in K} \to x^*$, we see that (27) holds at x^*, which contradicts nonstationarity of x^*. Thus, (44) holds. Now, using (44), the relation $H_k \succeq \underline{\lambda}I$, boundedness of $\{\bar{\alpha}_k\|d^k(\bar{\alpha}_k)\|\}_{k \in K}$, and a similar argument as that leading to (35), we observe that $\bar{\alpha}_k d^k(\bar{\alpha}_k)^T H_k d^k(\bar{\alpha}_k) = \Theta(\|d^k(\bar{\alpha}_k)\|)$ as $k \in K \to \infty$. Using this result and a similar argument as that leading to (37), we have

$$F(x^k + d^k(\bar{\alpha}_k)) \leq \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma[\nabla f(x^k)^T d^k(\bar{\alpha}_k) + P(x^k + d^k(\bar{\alpha}_k)) - P(x^k)],$$

provided that $k \in K$ is sufficiently large. The above inequality evidently contradicts (43). Thus, $\{\alpha_k\}_{k \in K}$ is bounded.

Finally, invoking that $d^k = d^k(\alpha_k)$ is the optimal solution of (28) with $x = x^k$, $H = \alpha_k H_k$, we have

$$0 \in \nabla f(x^k) + \alpha_k H_k d^k + \partial P(x^k + d^k) + N_X(x^k + d^k) \quad \forall k \in K. \tag{45}$$

By Lemma 3.8, we have $\{d^k\}_{k \in K} \to 0$. Upon taking limits on both sides of (45) as $k \in K \to \infty$, and using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [23] and Lemma 2.42 of [24]), 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 (27) 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 that made in [26]. In what follows, we denote by $\bar{X}$ the set of stationary points of problem (26).

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

$$\mathrm{dist}(x, \bar{X}) \leq \tau\|d_I(x)\| \quad \textit{whenever} \quad F(x) \leq \zeta, \; \|d_I(x)\| \leq \epsilon.$$

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

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

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 [26], who analyzed a similar local convergence for a coordinate gradient descent method for a class of nonsmooth minimization problems.

Theorem 3.10 *Suppose that f satisfies (30), and 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 I satisfies*

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

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

$$\|d_I(x^k)\| = \Theta(\|d^k\|), \tag{46}$$

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 (46), there exists some index k' such that

$$\|x^{l(k)-1} - \bar{x}^{l(k)-1}\| \le c_1 \|d^{l(k)-1}\| \quad \forall k \ge k' \tag{47}$$

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, (47), and Lemma 3.8, we obtain

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

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

$$F(\bar{x}^{l(k)-1}) = v \quad \forall k \ge \hat{k}. \tag{48}$$

Then, by Lemma 5.1 of [26], we see that

$$F^* = \lim_{k \to \infty} F(x^k) = \liminf_{k \to \infty} F(x^{l(k)-1}) \ge v. \tag{49}$$

Further, using the definition of F, (30), (48), Lemma 3.6(b), and $H_k(\alpha_k) \preceq \hat{\alpha} \bar{\lambda} I$, we have for $k \ge \hat{k}$,

$$
\begin{aligned}
F(x^{l(k)}) - v &= f(x^{l(k)}) + P(x^{l(k)}) - f(\bar{x}^{l(k)-1}) - P(\bar{x}^{l(k)-1}) \\
&= \nabla f(\tilde{x}^k)^T (x^{l(k)} - \bar{x}^{l(k)-1}) + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \\
&= (\nabla f(\tilde{x}^k) - \nabla f(x^{l(k)-1})^T (x^{l(k)} - \bar{x}^{l(k)-1}) - (H_{l(k)-1}(\alpha_{l(k)-1}) d^{l(k)-1})^T (x^{l(k)} - \bar{x}^{l(k)-1}) \\
&\quad + \left[(\nabla f(x^{l(k)-1}) + H_{l(k)-1}(\alpha_{l(k)-1}) d^{l(k)-1})^T (x^{l(k)} - \bar{x}^{l(k)-1}) + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \right] \\
&\le L \|\tilde{x}^k - x^{l(k)-1}\| \|x^{l(k)} - \bar{x}^{l(k)-1}\| + \hat{\alpha} \bar{\lambda} \|d^{l(k)-1}\| \|x^{l(k)} - \bar{x}^{l(k)-1}\|,
\end{aligned}
\tag{50}
$$

where $\tilde{x}^k$ is some point lying on the segment joining $x^{l(k)}$ with $\bar{x}^{l(k)-1}$. It follows from (47) that, for $k \ge \hat{k}$,

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

22

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), $H_k(\alpha_k) \succeq (\underline{\alpha} \cdot \underline{\lambda})I$, and (50), 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 (40) gives

$$F(x^{l(k)}) - v \leq c_3 \left(F(x^{l(l(k)-1)}) - F(x^{l(k)})\right) \quad \forall k \geq \hat{k}, \tag{51}$$

where $c_3 = c_2/\sigma$. Using $\lim_{k\to\infty} F(x^{l(k)}) = F^*$, and upon taking limits on both sides of (51), we see that $F^* \leq v$, which together with (49) implies that $v = F^*$. Using this result and upon rearranging terms of (51), 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 (26) 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 (28) with $x = x^k$ and $H = H_k$ to obtain $d^k = d_H(x)$, and compute Δ_k according to (32).

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

$$F(x^k + \alpha_k d^k) \leq \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma \alpha_k \Delta_k, \tag{52}$$

where Δ_k is defined in (32).

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 [26]. When the entire coordinate block, that is, $J = \{1, \ldots, n\}$ is chosen for the latter method, it becomes a special case of our method with $M = 0$, which is a gradient descent method. Given that our method is generally a nonmonotone method especially when $M \geq 1$, most proofs of global and local convergence for the method [26] do not hold for our method directly. In addition, our method can be viewed as an extension of the second projected gradient method (namely, SPG2) studied in [4] for smooth problems, but the method [26] 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 (52) 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 (26). Then, there exists $\tilde{\alpha} > 0$ such that $d^k = d_{H_k}(x^k)$ satisfies (52) whenever $0 < \alpha_k \leq \tilde{\alpha}$.*

Proof. In view of Lemma 2.1 of [26] 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 (32). 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 (32), the definition of d^k, and $H_k \succeq \underline{\lambda} I$, we have

$$
\Delta_k = \nabla f(x^k)^T d^k + P(x^k + d^k) - P(x^k) \leq -\frac{1}{2}(d^k)^T H_k d^k \leq -\frac{1}{2}\underline{\lambda}\|d^k\|^2, \tag{53}
$$

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

$$
\alpha_k^2 \|d^k\|^2 \leq -2\bar{\alpha}\alpha_k \Delta_k / \underline{\lambda}. \tag{54}
$$

By a similar argument as that leading to (38), we see that $\{x^k\}$ satisfies (38) for some F^*. We next show by induction that the following limits hold for all $j \geq 1$:

$$
\lim_{k \to \infty} \alpha_{l(k)-j} d^{l(k)-j} = 0, \quad \lim_{k \to \infty} F(x^{l(k)-j}) = F^*. \tag{55}
$$

Indeed, using (52) 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 (38) immediately yields $\lim_{k \to \infty} \alpha_{l(k)-1} \Delta_{l(k)-1} = 0$. Using this result and (54), we see that the first identity of (55) holds for $j = 1$. Further, in view of this identity, (38),

and uniform continuity of F in $\mathcal{L}$, we can easily see that the second identity of (55) also holds $j = 1$. We now need to show that if (55) holds for j, then it also holds for $j + 1$. First, it follows from (52) 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 (38) 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 (54), 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, (55) holds for $j + 1$. The conclusion of this lemma then follows from (55) and a similar argument as that in the proof of Lemma 3.12. ∎

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 (26).*

Proof. Suppose for contradiction that x^* is an accumulation point of $\{x^k\}$ that is a non-stationary point of (26). 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 (28) 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 [23] and Lemma 2.42 of [24]) 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 (27) 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 (33), we have

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

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 $\|d^k\| = \Theta(1)$ for $k \in K$. Further, using (53), 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 $\alpha_k^0 \geq \underline{\alpha} > 0$, there exists some index $\bar{k} \geq 0$ such that $\alpha_k < \alpha_k^0$ and $\alpha_k < \eta$ for all $k \in K$ with $k \geq \bar{k}$. Let $\bar{\alpha}_k = \alpha_k/\eta$. Then, $\{\bar{\alpha}_k\}_{k\in K} \to 0$ and $0 < \bar{\alpha}_k \leq 1$ for all $k \in K$. By the stepsize rule used in step (3), we have, for all $k \in K$ with $k \geq \bar{k}$,

$$F(x^k + \bar{\alpha}_k d^k) > \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma\bar{\alpha}_k\Delta_k, \tag{56}$$

On the other hand, in view of the definition of F, (32), the relations $\|d^k\| = \Theta(1)$ and $\limsup_{k \in K, k \to \infty} \Delta_k < 0$, and the monotonicity of $(P(x^k + \alpha d^k) - P(x^k))/\alpha$, we obtain that, for sufficiently large $k \in K$,

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

which clearly contradicts (56). 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 [26].

Theorem 3.14 *Suppose that $\bar{\alpha} \leq 1$, f satisfies (30), and 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*

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

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 Armijo rule with $\alpha_k^0 \geq \underline{\alpha} > 0$, 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

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

where $\hat{k}$ is some index. Then, by Lemma 5.1 of [26], we see that (49) holds for $\{x^k\}$, and the above F^* and v. Further, using the definition of F, (30), Lemma 3.6(b), and $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$,

we have, for $k \geq \hat{k}$,

$$
\begin{aligned}
F(x^{l(k)}) - v &= f(x^{l(k)}) + P(x^{l(k)}) - f(\bar{x}^{l(k)-1}) - P(\bar{x}^{l(k)-1}) \\
&= \nabla f(\tilde{x}^k)^T(x^{l(k)} - \bar{x}^{l(k)-1}) + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \\
&= (\nabla f(\tilde{x}^k) - \nabla f(x^{l(k)-1})^T(x^{l(k)} - \bar{x}^{l(k)-1}) - (H_{l(k)-1}d^{l(k)-1})^T(x^{l(k)} - \bar{x}^{l(k)-1}) \\
&\quad + \left[(\nabla f(x^{l(k)-1}) + H_{l(k)-1}d^{l(k)-1})^T(x^{l(k)} - \bar{x}^{l(k)-1}) + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \right] \\
&\leq L\|\tilde{x}^k - x^{l(k)-1}\| \|x^{l(k)} - \bar{x}^{l(k)-1}\| + \bar{\lambda}\|d^{l(k)-1}\| \|x^{l(k)} - \bar{x}^{l(k)-1}\| \\
&\quad + (\alpha_{l(k)-1} - 1) \left[(d^{l(k)-1})^T H_{l(k)-1}d^{l(k)-1} + \Delta_{l(k)-1} \right],
\end{aligned}
\tag{57}
$$

where $\tilde{x}^k$ is some point lying on the segment joining $x^{l(k)}$ with $\bar{x}^{l(k)-1}$. It follows from (47) and $\alpha_k \leq 1$ that, for $k \geq \hat{k}$,

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

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), $H_k \succeq \underline{\lambda}I$, $\alpha_k \leq 1$, and (57), we see that, for $k \geq \hat{k}$,

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

for some constant $c_2 > 0$. 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 Section 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

$$
\begin{aligned}
\min_{V \in \Re^{n \times r}} \quad & -\text{Tr}(V^T\hat{\Sigma}V) + \rho \bullet |V| \\
\text{s.t.} \quad & V_i^T\hat{\Sigma}V_j \leq \Delta_{ij} \quad \forall i \neq j, \\
& -V_i^T\hat{\Sigma}V_j \leq \Delta_{ij} \quad \forall i \neq j, \\
& V^TV = I.
\end{aligned}
\tag{58}
$$

Clearly, problem (58) has the same form as (9). From Subsection 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 (10) 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 (58), 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 (58). 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 (58). Given any W_1, $W_2 \in \mathcal{S}^r$, the system of*

$$\delta V^T \hat{\Sigma} V + V^T \hat{\Sigma} \, \delta V + \delta D = W_1, \tag{59}$$

$$\delta V^T V + V^T \delta V = W_2 \tag{60}$$

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 $V_i^T \hat{\Sigma} V_i \neq V_j^T \hat{\Sigma} V_j$ for all $i \neq j$;

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 (59) and (60) is equivalent to

$$\delta P^T G + G \delta P + \delta \bar{P}^T \bar{G} + \bar{G}^T \delta \bar{P} + \delta D = W_1, \tag{61}$$

$$\delta P^T + \delta P = W_2, \tag{62}$$

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 (61) and (62) 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

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

Now, let $\delta \bar{P}^* = 0$ and $\delta D^* = \widetilde{\text{Diag}}(W_1 - GW_2)$. It is easy to verify that $(\delta P^*, \delta \bar{P}^*, \delta D^*)$ is a solution of the system of (61) and (62).

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 $\delta \bar{P} = \delta Y + \bar{G} \delta Z$. Peforming such a change of variable for $\delta \bar{P}$, we see that (61) can be rewritten as

$$\delta P^T G + G \delta P + \delta Z^T \bar{G}^T \bar{G} + \bar{G}^T \bar{G} \delta Z + \delta D = W_1. \tag{63}$$

Thus, it suffices to show that the system of (62) and (63) 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

$$\delta P^* = W_2/2, \quad \delta Z^* = (\bar{G}^T\bar{G})^{-1}(2W_1 - W_2G - GW_2)/4, \quad \delta D^* = 0.$$

It is easy to verify that $(\delta P^*, \delta Z^*, \delta D^*)$ is a solution of the system of (63) and (62). Therefore, the conclusion holds. ∎

We are now ready to show that Robinson's condition (10) holds at a set of feasible points of (58).

Proposition 4.2 *Let $V \in \Re^{n \times r}$ be a feasible solution of (58). The Robinson's condition (10) 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 (58) at V, and $V^T \hat{\Sigma}(I - VV^T)\hat{\Sigma}V$ is nonsingular;*

c) *All inequality constraints of (58) 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 (58) become equality ones. Using these facts and Lemma 4.1, we see that Robinson's condition (10) 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 (10) holds at V if and only if (60) 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 (60), and thus Robinson's condition (10) holds at V. ∎

From Proposition 4.2, we see that Robinson's condition (10) indeed holds at a set of feasible points of (58). Though we are not able to show that it holds at all feasible points of (58), 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 (58)

In this section, we show how our augmented Lagrangian method proposed in Subsection 3.2 can be applied to solve problem (58) (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 (58). 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 (58), respectively, and $\mu \in \mathcal{S}^r$ is the Lagrangian multipliers for the equality constraints of (58). For convenience of presentation, let $\Delta \in \mathcal{S}^r$ be the matrix whose ijth entry equals the parameter Δ_{ij} of (58) 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 (58) can be rewritten as

$$L_\varrho(V, \lambda, \mu) = w(V) + \rho \bullet |V|, \tag{64}$$

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

$$S = V^T \hat{\Sigma} V - \widetilde{\mathrm{Diag}}(V^T \hat{\Sigma} V), \qquad R = V^T V - I. \tag{65}$$

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 $O(p^2 r)$. 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 $\hat{\Sigma} = X^T X / (n-1)$. 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 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, $O(p^2 r)$.

We now address initialization and termination criterion for our augmented Lagrangian method. In particular, we choose initial point V_{init}^0 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,

$$\max_{i \neq j}[|V_i^T \hat{\Sigma} V_j| - \Delta_{ij}]^+ \leq \epsilon_I, \quad \max_{i,j}|R_{ij}| \leq \epsilon_E, \quad \frac{|L_\varrho(V, \lambda, \mu) - f(V)|}{\max(|f(V)|, 1)} \leq \epsilon_O, \qquad (66)$$

where $f(V) = -\mathrm{Tr}(V^T \hat{\Sigma} V) + \rho \bullet |V|$, R is defined in (65), and ϵ_I, ϵ_E, ϵ_O are some pre-scribed accuracy parameters corresponding to inequality constraints, equality constraints and objective function, respectively.

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

$$\min_V L_\varrho(V, \lambda, \mu), \qquad (67)$$

where the function $L_\varrho(\cdot, \lambda, \mu)$ is defined in (64). Given that the implementation details of those nonmonotone gradient methods are similar, we only focus on the second one, that is, the nonmonotone gradient method II. First, the initial point for this method can be chosen according to the scheme described at the end of Subsection 3.2. In addition, given an iterate V^k, the search direction d^k is computed by solving subproblem (28) with $H = \alpha_k^{-1} I$, which becomes, in the context of (21) and (64),

$$d^k := \arg\min_d \left\{ \nabla w(V^k) \bullet d + \frac{1}{2\alpha_k}\|d\|_F^2 + \rho \bullet |V^k + d| \right\}. \qquad (68)$$

Here, $\alpha_k > 0$ is chosen according to the scheme proposed by Barzilai and Borwein [2], which is also used by Birgin et al. [4] for studying a class of projected gradient methods. Let $0 < \alpha_{\min} < \alpha_{\max}$ be given. Initially, choose an arbitrary $\alpha_0 \in [\alpha_{\min}, \alpha_{\max}]$. Then, α_k is updated as follows:

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

where $a_k = \|V^k - V^{k-1}\|_F^2$ and $b_k = (V^k - V^{k-1}) \bullet (\nabla w(V^k) - \nabla w(V^{k-1}))$. It is not hard to verify that the optimal solution of problem (68) has a closed-form expression, which is given by

$$d^k = \mathrm{sign}(C) \odot \left[|C| - \alpha^k \rho\right]^+ - V^k,$$

where $C = V^k - \alpha^k \nabla w(V^k)$. In addition, we see from Lemma 3.4 that the following termination criterion is suitable for this method when applied to (67):

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

where $d_I(V^k)$ is the solution of (68) with $\alpha_k = 1$, and ϵ is a prescribed accuracy parameter. In our numerical implementation, we set $\alpha^0 = 1/\max_{ij}|d_I(V^0)|_{ij}$, $\alpha_{\max} = 1$, $\alpha_{\min} = 10^{-15}$ and $\epsilon = 10^{-4}$.

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

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 Subsection 3.2. In particular, we follow a similar scheme as discussed on pp. 405 of [3] 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 [3].

5   Numerical results

In this section, we conduct numerical experiments for the augmented Lagrangian method detailed in Subsections 3.2 and 4.2 for formulation (58) (or, equivalently, (3)) of sparse PCA on synthetic, random, and real 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. [16]), the DSPCA algorithm (d'Aspremont et al. [8]), the SPCA algorithm (Zou et al. [28]), and the sPCA-rSVD algorithm (Shen and Huang [25]). We now 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 [16], and the method ALSPCA is the augmented Lagrangian method proposed in this paper.

As discussed in Section 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, $\text{Tr}(V^T \hat{\Sigma} V)$, 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 $\text{Tr}(V^T \hat{\Sigma} V)$ can overestimate much the total explained variance by the PCs due to the overlap among their individual variances. In attempt to deal with such an overlap, two adjusted total explained variances were proposed in [28, 25]. 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 be dramatically different from those sparse PCs, their total explained variances may

differ much from each other as well. To alleviate this drawback while taking into account the possible correlations among PCs, we introduce the following *adjusted total explained variance* for sparse PCs:

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

Clearly, when the PCs are uncorrelated, it becomes the usual total explained variance, that is, $\text{Tr}(V^T \hat{\Sigma} V)$. We 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, $\text{AdjVar}V / \text{Tr}(\hat{\Sigma})$.

Finally, we shall stress that the sole purpose of this section is to compare the performance of those methods listed in Table 1 for finding the sparse PCs that nearly enjoy the three important properties possessed by the standard PCA (see Section 1). 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 observed in our experiments.

5.1   Synthetic data

In this subsection we use the synthetic data introduced by Zou et al. [28] to test the effectiveness of our approach ALSPCA for finding sparse PCs.

The synthetic example [28] considers three hidden factors:

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

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

$$X_i = V_1 + \epsilon_i^1, \qquad \epsilon_i^1 \sim N(0, 1), \qquad i = 1, 2, 3, 4,$$

$$X_i = V_2 + \epsilon_i^2, \qquad \epsilon_i^2 \sim N(0, 1), \qquad i = 5, 6, 7, 8,$$

$$X_i = V_3 + \epsilon_i^3, \qquad \epsilon_i^3 \sim N(0, 1), \qquad i = 9, 10,$$

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 see 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 the three underlying factors V_1, V_2 and V_3 are 290, 300, and 283.8, respectively. Therefore V_2 is slightly more important than V_1, and they both are much more important than V_3. In addition, the first two standard PCs together explain 99.72% of the total variance (see Table 2). These facts suggest that the first two sparse PCs be sufficient to explain most of the variance. Ideally, the first sparse PC recovers the factor V_2 only using (X_5, X_6, X_7, X_8), and the second sparse PC recovers the factor V_1 only using (X_1, X_2, X_3, X_4). Moreover, 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.

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

In our test, we set $r = 2$, $\Delta_{ij} = 0$ for all $i \neq j$, and $\rho = 4$ for formulation (58) of sparse PCA. In addition, we choose (66) 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 [28, 8]. For general data, however, these methods may perform quite differently (see Subsection 5.3).

5.2 Random data

In this subsection, we compare the performance of the GPower methods [16] and our ALSPCA method for finding sparse PCs on a set of randomly generated data. First, we randomly generate 100 centered data matrices X with the size of 20×20. Throughout this subsection, we set $\Delta_{ij} = 0.5$ for all $i \neq j$ for formulation (58) of sparse PCA, and choose (66) as the termination criterion for ALSPCA with $\epsilon_I = 0.1$, $\epsilon_E = 0.1$ and $\epsilon_O = 0.1$.

In the first test, we aim to find the first *three* sparse PCs with the average number of zero loadings around 30 (50% sparsity). To meet this purpose, the tunning parameter ρ for problem (58) and the parameters for the GPower methods are properly chosen. The results of the GPower methods and ALSPCA for the above randomly generated instances are presented in Table 3. The name of each method is given in column one. The sparsity measured by the number of zero loadings averaged over all instances is given in column two. The column three gives the average amount of non-orthogonality of the loading vectors, 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. In addition, the maximum correlation and CPAV of sparse PCs averaged over all instances are presented in columns four and five, respectively. From Table 3, we see that the average number of zero loadings of the first three sparse PCs for all methods are almost same, which are around 30. We also observe that the sparse PCs given by our method ALSPCA are almost uncorrelated and their loading vectors are nearly orthogonal, which are much superior to the GPower methods.

34

Table 3: Comparison of GPower and ALSPCA

Method	Sparsity	Non-orthogonality	Correlation	CPAV (%)
GPower$_{l_1}$	30.73	3.480	0.104	39.06
GPower$_{l_0}$	30.37	3.540	0.111	38.74
GPower$_{l_1,m}$	30.61	5.220	0.161	38.12
GPower$_{l_0,m}$	30.51	5.216	0.153	37.99
ALSPCA	30.91	1.007	0.038	40.79

Random data: Test I

Table 4: Comparison of GPower and ALSPCA

Method	Sparsity	Non-orthogonality	Correlation	CPAV (%)
GPower$_{l_1}$	60.58	5.766	0.168	63.46
GPower$_{l_0}$	60.53	6.031	0.169	63.27
GPower$_{l_1,m}$	60.02	8.665	0.265	62.34
GPower$_{l_0,m}$	60.14	9.146	0.272	62.00
ALSPCA	60.74	1.177	0.066	64.62

Random data: Test II

Moreover, our sparse PCs have better CPAV than the others.

Our aim of the second test is to find the first *six* sparse PCs with the average number of zero loadings around 60 (50% sparsity). To reach this goal, the tunning parameter ρ for problem (58) and the parameters for the GPower methods are appropriately chosen. The results of the GPower methods and ALSPCA for the above randomly generated instances are presented in Table 4. Each column of Table 4 has the same meaning as Table 3. First, we clearly see that our method substantially outperforms the GPower methods in terms of uncorrelation of PCs, orthogonality of loading vectors and CPAV. Further, in comparison with Table 3, we observe that as the number of PCs increases, the CPAV grows accordingly for all methods, and moreover, the sparse PCs given by the GPower methods become much more correlated and non-orthogonal each other. But the performance of our sparse PCs almost remains same. This phenomenon is actually not surprising, given that uncorrelation and orthogonality are not well taken into account in the GPower methods.

5.3 Pitprops data

In this subsection we test the performance of our approach ALSPCA for finding sparse PCs on the real data, that is, Pitprops data. We also compare the results with several existing methods [28, 8, 25, 16].

The Pitprops data introduced by Jeffers [14] has 180 observations and 13 measured variables. It is a classic example that illustrates the difficulty of interpreting PCs. Recently, several sparse PCA methods [17, 28, 25, 8] 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 some of those sparse PCs in Tables 5-8, respectively. It shall be mentioned that two groups of sparse PCs were found by DSPCA with the parameter $k_1 = 5$ or 6, and they have similar sparsity and total explained variance (see [8] for details). Thus, we only present the latter one (i.e., the one with $k_1 = 6$) in Table 8. Also, we applied the GPower methods [16] to this data for finding the PCs with the largest sparsity of the ones given in [28, 25, 8],

35

Table 5: 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	0.1128	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

Table 6: 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

and found the best result was given by GPower$_{l_1}$. Thus, we only report the six sparse PCs obtained by GPower$_{l_1}$ in Table 9. In addition, we present sparsity, CPAV, non-orthogonality and correlation of the PCs obtained by the standard PCA and sparse PCA methods [28, 25, 8, 16] in columns two to five of Table 13, respectively. In more details, the second and fifth columns 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. The fourth column presents the maximum correlation of PCs. Though the PCs obtained by these sparse PCA methods have nice sparsity, we observe from Tables 13 that they are much correlated, and moreover almost all of them are far from orthogonal each other except the ones given by SPCA [28]. 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 (66) as the termination criterion for ALSPCA with $\epsilon_O = 0.1$ and $\epsilon_I = \epsilon_E = 10^{-3}$.

In the first experiment, we aim to find six 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 (58) of sparse PCA. The resulting sparse PCs are presented in Table 10, and their sparsity, CPAV, non-orthogonality and correlation are reported in row

Table 7: 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

Table 8: 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

seven of Table 13. 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 (58) in this experiment. The resulting sparse PCs are presented in Table 11. The CPAV, non-orthogonality and correlation of these PCs are given in row eight of Table 13. In comparison with the ones given in the above experiment, the PCs 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

Table 9: Loadings of the first six PCs by GPower$_{l_1}$

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

Table 10: 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

choose $\rho = 0.7$ for problem (58). The obtained sparse PCs are presented in Table 12. The CPAV, non-orthogonality and correlation of these PCs are given in the last row of Table 13. We see that these PCs are highly sparse, orthogonal, and explain good amount of variance. However, they are quite correlated with each other, which is actually not surprising, given that $\Delta_{ij}(i \neq j)$ are not small. Despite such a drawback, we observe that these sparse PCs still overally 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 10.

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 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

Table 11: 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

Table 12: 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

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, random, 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 about 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 in practice 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 Section 4 we showed that Robinson's condition (10) holds at a set of feasible points of

Table 13: Comparison of SPCA, rSVD, DSPCA, GPower$_{l_1}$ 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$_{l_1}$	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

(58). Also, we observed from our experiments that the accumulation points of our augmented Lagrangian method lie in this set when applied to (58), and thus it converges. However, it remains open whether or not Robinson's condition holds at all feasible points of (58).

In addition, Burer and Monteiro [5] recently applied the classical augmented Lagrangian method to a nonconvex nonlinear program (NLP) reformulation of semidefinite programs (SDP), 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 converging to a feasible point when applied to a nonconvex NLP. Due this and [19], at least theoretically, their approach [5] 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 [5].

Finally, the codes of our approach for solving the sparse PCA formulation (58) (or, equivalently, (3)) are written in Matlab, which 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.

Acknowledgement

References

[1] O. Alter, P. Brown, and D. Botstein. Singular value decomposition for Genome-Wide expression data Processing and Modeling. *Proceedings of the National Academy of Sciences*, 97:10101–10106, 2000.

[2] J. Barzilai and J. M. Borwein. Two point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148, 1988.

[3] D. P. Bertsekas *Nonlinear Programming*. Athena Scientific, 1999.

[4] E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 4:1196–1211, 2000.

[5] S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming, Series B*, 95:329–357, 2003.

[6] J. Cadima and I. Jolliffe. Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22:203–214, 1995.

[7] A. d'Aspremont, F. R. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.

[8] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49:434–448, 2007.

[9] K. Fan. On a theorem of Weyl concerning the eigenvalues of linear transformations. *Proceedings of the National Academy of the Sciences of U.S.A.*, 35: 652–655 (1949).

[10] P. Hancock, A. Burton, and V. Bruce. Face processing: Human perception and principal components analysis. *Memory and Cognition*, 24:26–40, 1996.

[11] T. Hastie, R. Tibshirani, M. Eisen, P. Brown, D. Ross, U. Scherf, J. Weinstein, A. Alizadeh, L. Staudt, and D. Botstein. ǵene Shavingás a method for identifying distinct sets of genes with similar Expression Patterns. *Genome Biology*, 1:1–21, 2000.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data mining, Inference and Prediction*. New York: Springer Verlag.

[13] J. B. Hiriart–Urruty and C. Lemaréchal. *Convex Analysis and Minimization algorithms I*. Comprehensive Study in Mathematics, volume 305, Springer-Verlag, New York, 1993.

[14] J. Jeffers. Two Case Studies in the Application of Principal Component. *Applied Statistics*, 16:225–236, 1967.

[15] I. Jolliffe. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22:29–35, 1995.

[16] M. Journée, Yu. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. CORE Discussion Paper 2008/70. Submitted to *Journal of Machine Learning Research*.

[17] I. T. Jolliffe, N. T. Trendafilov, and M. L. Uddin. A modified principal component technique based on the Lasso. *Journal of Computational and Graphical Statistics*, 12:531–547.

[18] B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 915–922. MIT Press, Cambridge, MA, 2006.

[19] R. D. C. Monteiro. *private communication*, 2009.

[20] M. L. Overton and R. S. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62:321–357, 1993.

[21] S. M. Robinson. Stability theory for systems of inequalities, Part 2: Differentiable nonlinear systems. *SIAM Journal on Numerical Analysis*, 13:497–513, 1976.

[22] S. M. Robinson. Local structure of feasible sets in nonlinear programming, Part I: Regularity. In V. Pereira and A. Reinoza, editors, *Numerical Methods Lecture Notes in Mathematics*, vol. 1005, Springer-Verlag, Berlin, 1983.

[23] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[24] A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006.

[25] H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.

[26] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423, 2009.

[27] S J. Wright, R. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. To appear in *IEEE Transactions on Signal Processing*.

[28] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.