EXACT PENALTY FUNCTION FOR $\ell_{2,1}$ NORM MINIMIZATION OVER THE STIEFEL MANIFOLD

NACHUAN XIAO*, XIN LIU†, AND YA-XIANG YUAN‡

Abstract. $\ell_{2,1}$ norm minimization with orthogonality constraints, feasible region of which is called Stiefel manifold, has wide applications in statistics and data science. The state-of-the-art approaches adopt proximal gradient technique on either Stiefel manifold or its tangent spaces. The consequent subproblem does not have closed-form solution and hence requires an iterative procedure to solve which is usually time consuming. In this paper, we discover that the Lagrangian multipliers of the orthogonality constraints in this class of problems are of closed-form expressions. By using this closed-form expression, we introduce a penalty function for this type of problems. We theoretically demonstrate the equivalence between the penalty function and the original $\ell_{2,1}$ norm minimization under mild assumptions. Based on the exact penalty function, we propose an inexact proximal gradient method in which the subproblem is of closed-form solution. The global convergence and the worst case complexity are established. Numerical experiments illustrate the numerical advantages of our method when comparing with the existing proximal-based first-order methods.

Keywords orthogonality constraint, Stiefel manifold, nonsmooth optimization, augmented Lagrangian method

Mathematics Subject Classification (2010) 15A18, 65F15, 65K05, 90C06

1. Introduction.

1.1. Problem Description.

In this paper, we focus on a class of composite optimization problems with orthogonality constraints,

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & f(X)+r(X) \\ \text{s.t.} \quad & X^\top X = I_p. \end{aligned} \tag{1.1}$$

Throughout this paper, $f:\mathbb{R}^{n\times p}\mapsto\mathbb{R}$ satisfies Assumption 1.1 listed below. The nonsmooth term

$$r(X) := ||\Gamma X||_{2,1} = \sum_{i=1}^{n} \gamma_i \left\|X_{i\cdot}\right\|_2$$

is the $\ell_{2,1}$ norm minimization, where $\Gamma\in\mathbb{R}^{n\times n}$ is a diagonal matrix with γ_i $(i=1,...,n)$ on its diagonal, and $X_{i\cdot}$ $(i=1,...,n)$ denotes the i-th row of X. The parameter $\gamma_i\geq 0$. In addition, I_p is a $p\times p$ identity matrix and the feasible region, denoted by $\mathcal{S}_{n,p}:=\{X\in\mathbb{R}^{n\times p}|X^\top X=I_p\}$, is called the Stiefel manifold.

ASSUMPTION 1.1 (blanket assumption). *$f(X)$ is differentiable and $\nabla f(X)$ is locally Lipschitz continuous.*

We are particularly interested in Problem (1.1) since it has wide applications in data science and machine learning. We list a few of them in the following.

EXAMPLE 1.2. **Sparse Variable Principle Component Analysis**

*State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, China (xnc@lsec.cc.ac.cn).

†State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, China (liuxin@lsec.cc.ac.cn).
Research is supported in part by the National Natural Science Foundation of China (No. 11971466, 11991021 and 11991020), Key Research Program of Frontier Sciences, Chinese Academy of Sciences (No. ZDBS-LY-7022), the National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences and the Youth Innovation Promotion Association, Chinese Academy of Sciences.
‡State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China (yyx@lsec.cc.ac.cn). Research supported in part by NSFC grant 11688101.

Principle component analysis (PCA) is one of the most popular multivariate data analysis techniques for dimension reduction and data mining. Its basic idea is to project the high-dimensional dataset onto a subspace spanned by a few leading eigenvectors of its covariance matrix. When the dimension is much larger than the number of samples, PCA may lead to poor estimations [8]. To enhance PCA in such situation, we may introduce the sparsity to extract lower dimensional features, i.e. nonzero rows of leading eigenvectors, see [35, 8] for instance. To this end, we adopt $\ell_{2,1}$ norm minimization term which is the most suitable regularizer to enforce the sparsity:

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & -\frac{1}{2}\mathrm{tr}\left(X^\top MX\right)+\gamma||X||_{2,1} \\ \textit{s.t.} \quad & X^\top X = I_p, \end{aligned} \tag{1.2}$$

where M is the empirical covariance matrix of the given dataset. A direct extension of Sparse Variable PCA is like the following

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & -\frac{1}{2}\mathrm{tr}\left(X^\top MX\right)+||\Gamma X||_{2,1} \\ \textit{s.t.} \quad & X^\top X = I_p, \end{aligned}$$

which is proposed by [12] and called Coordinate-independent Sparse Estimation.

EXAMPLE 1.3. **Unsupervised Feature Selection**

Feature selection aims to select a small size of features from the high dimensional dataset for a compact and accurate data representation. There are several existing works which model unsupervised feature selection as a manifold optimization problem in the form of (1.1), see [41, 34] for instance. Let $\{b_1, b_2, ..., b_N\}$ be the dataset, and we aim to find a linear classifier X, namely, the combination coefficients of different features. If many rows of X shrink to zero, a new representation $X^\top b_i$ of data point b_i ($i = 1, ..., N$) uses only a small set of selected features. Therefore, [41] suggests to compute X by solving the optimization problem with the same formulation as (1.2) but the matrix M is constructed by other means from the input data points, and the detail formula can be referred to Algorithm 1 in [41].

1.2. Existing Methods. Optimization over Stiefel manifold has drawn a lot of attention due to its wide applications in various fields, besides the examples shown in the previous subection, we refer the interested readers to [16, 36, 17, 18]. There are many existing methods for minimizing smooth objective function over Stiefel manifold, such as gradient-based methods [26, 27, 2], conjugate gradient methods [16, 1], projection-based methods [4, 14], constraint preserving updating scheme [36, 23], Newton methods [22], trust-region methods [3], multipliers correction framework [17], orthonormalization-free methods [18, 38] etc. Interested readers are referred to the book [4], a recent paper [38] and the references therein. However, the above mentioned approaches require computing the derivatives of the objective function and hence do not apply to the cases with nonsmooth objective functions.

In comparison to the case of minimizing smooth objective function, the studies on solving nonsmooth optimization problems over Stiefel manifold are relatively limited. The existing approaches can be divided into two categories. The approaches in the first category apply the nonsmooth optimization techniques to Stiefel manifold, such as Riemannian subgradient methods [20, 15, 6], inexact cyclic proximal point algorithm [5], nonsmooth trust-region method [19], gradient sampling method [21], etc. When the nonsmooth term is convex and Lipstchitz continuous, [9] proposes a manifold proximal gradient method (ManPG) in each step of which a proximal mapping is calculated in the tangent space $\mathcal{T}_{X^k}$ and then the projection of Y^{k+1} back to the Stiefel manifold is preformed to obtain the next iterate X^{k+1}. When

applied to solving (1.1), such restricted proximal mapping can be formulated as

$$Y^{k+1} = \arg\min_{D \in \mathcal{T}_{X^k}} D^\top \nabla f(X^k) + r(D) + \frac{1}{2\eta^k} \left\| D - X^k \right\|_{\mathrm{F}}^2 . \tag{1.3}$$

Obviously, Y^{k+1} in (1.3) does not have closed-form solution as the situation in Euclidean space. Therefore, [9] suggests to use semi-smooth Newton method [29, 32, 42] to calculate an approximate solution of (1.3). In each iterate of semi-smooth Newton method, it requires to solve a linear equation and the computational cost is $O(n^2p)$ in each conjugate gradient step. Hence, calculating the proximal mapping is the computational bottleneck for ManPG.

The approaches in the other category are based on splitting and alternating. The splitting method for orthogonality constrained problem (SOC) [25] introduces a block of variables to split the objective function and the orthogonality constraints apart, and then adopts alternating direction method of multipliers (ADMM) to solve the equivalent model. From the perspective of image science, the authors of [30, 31] propose similar splitting framework for fast regularization of matrix-valued images. The subproblem related to the orthogonality constraints is of closed-form solution, meanwhile the subproblem related to the objective function requires an iterative method to solve an unconstrained composite minimization. The main idea of manifold alternating direction method of multipliers (MADMM) presented by [24] consists of introducing a block of variables to split the smooth and nonsmooth terms and then apply ADMM to solve the split model. The subproblem related to the nonsmooth term is of closed-form solution, meanwhile the subproblem related to the smooth term is to minimize a smooth function over Stiefel manifold and hence can be solved by any existing solvers, such as Manopt [7] adopted in [24]. The authors in [11] propose a proximal alternating minimization based on augmented Lagrangian method (PAMAL). Different with MADMM, PAMAL introduces two block of variables to split the orthogonality constraints, smooth and nonsmooth terms, and it invokes augmented Lagrangian method (ALM) to solve the split model. As the prime subproblem is still of complicated structure, [11] suggests to use alternating minimization to solve it. As illustrated in [9], the numerical performances of all of the above mentioned splitting and alternating approaches are very sensitive with the choices of the penalty and other algorithm parameters.

1.3. Motivation and Our Contributions. For minimizing smooth objective function over the Stiefel manifold, [18] proposes proximal linearized augmented Lagrangian method (PLAM) and a parallelizable column-wise block minimization for PLAM (PCAL). Both PLAM and PCAl do not request orthonormalization in each iterate and have very few parameters to tune. In particular, PCAL is not sensitive to the penalty parameter. Recently, [38] proposes a class of exact penalty models with a compact and convex constraint for smooth optimization problem with orthogonality constraints (PenC). Both PLAM and PCAL can be viewed as algorithms for solving special PenC models. The essence of PenC is to replace the Lagrangian multipliers of the orthogonality constraints in an augmented Lagrangian penalty function by a closed-form expression with respect to the prime variables. However, this idea can hardly be extended to the nonsmooth case, as the first-order derivative of the objective function involves in the closed-form expression of the Lagrangian multipliers.

In this paper, we first prove that the Lagrangian multipliers of the orthogonality constraints in nonsmooth problem (1.1) are of closed-form expression $\Lambda(X^*)$ at an arbitrary first-order stationary point X^*, where

$$\Lambda(X) := \Phi(X^\top \nabla f(X)) + \sum_{i=1}^{n} \gamma_i S(X_{i\cdot}^\top). \tag{1.4}$$

Here, the operator $\Phi : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ is defined by $\Phi(M) = \frac{M+M^\top}{2}$. For any given $Q \in \mathbb{S}^{p\times p}$, $S_Q : \mathbb{R}^p \mapsto \mathbb{R}^{p\times p}$ is defined by

$$S_Q(x) := \begin{cases} \frac{xQx^\top}{\|x\|_2}, & \text{if } x \neq 0; \\ \mathbf{0}_{p,p}, & \text{otherwise,} \end{cases} \tag{1.5}$$

and we denote $S(x) := S_{I_p}(x)$ for brevity. Clearly, $S(x)$ is Lipschitz continuous on $\mathbb{R}^p$.

Based on this closed-form expression, we construct the following penalty model with a compact convex constraint.

$$\min_{X\in\mathcal{M}} \quad h(X) := f(X) + r(X) - \frac{1}{2}\left\langle \Lambda(X), X^\top X - I_p \right\rangle + \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2, \tag{1.6}$$

where $\mathcal{M} \subseteq \mathbb{R}^{n\times p}$ is a compact convex set that contains the feasible region $\mathcal{S}_{n,p}$. Here, we introduce an auxiliary compact convex set, because the penalty function $h(X)$ may not be bounded from below. For brevity, we call model (1.6) PenC for problem (1.1). which coincides the definition in [38] when $r(X) = 0$. We investigate the first-order stationary points set of PenC, and show that it produces additional first-order stationary point only far away from the Stiefel manifold. In addition, (1.1) and PenC share the same global minimizers under mild conditions.

We proposed a truncated proximal gradient method called PenCPG, which has two steps. The first step is to minimize a proximal linearized approximation of the penalty function $h(X)$, which is of closed-form solution. The second step is to truncate the proximal gradient step on the boundary of $\mathcal{M}$ if necessary. We establish the global convergence of PenCPG, as well as the worst case complexity. If we pursue higher precision on the feasibility rather than that of the substationarity, we can impose an orthonormalization step after the termination of PenCPG. The primarily numerical experiments show the great potential of our proposed approaches in solving sparse variable PCA problems and sparse CCA problems.

1.4. Notations. We use $\mathbb{S}^{p\times p}$ to denote the space containing all $p \times p$ real symmetric matrices. The Euclidean inner product of two matrices $X, Y \in \mathbb{R}^{n\times p}$ is defined as $\langle X, Y\rangle = \mathrm{tr}(X^\top Y)$, where $\mathrm{tr}(A)$ is the trace of a matrix $A \in \mathbb{R}^{p\times p}$. $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ represent the 2-norm and the Frobenius norm, respectively. The notations $\mathrm{diag}(A)$ and $\mathrm{Diag}(x)$ stand for the vector formed by the diagonal entries of matrix A, and the diagonal matrix with the entries of $x \in \mathbb{R}^n$ to be its diagonal, respectively. We denote the smallest eigenvalue of A by $\lambda_{\min}(A)$. The j-th column of matrix $X \in \mathbb{R}^{n\times p}$ is denoted by $X_{\cdot j}$ ($j = 1, ..., p$), and X_{ij} denotes the entry at i-th row and j-th column of X. $\mathrm{conv}\,\Omega$ is denoted as the convex hull of the set Ω. Finally, $\mathbf{0}_{n,p}$ stands for the $n \times p$ matrix with all entries being equal to zero.

1.5. Organization. The rest of this paper is organized as follows. In Section 2, we give the properties of the proposed model PenC. In Section 3, we present the proximal gradient algorithm PenCPG for solving PenC, and establish the global convergence. We discuss the extension of PenCPG for solving sparse CCA problems in Section 4. Preliminary numerical experiments are reported in Section 5. In the last section, we draw a brief conclusion.

2. Model Analyses. In this section, we explore the relationship between the original problem (1.1) and the proposed penalty model PenCPG.

2.1. Optimality Conditions. In this subsection, we aim to present the optimality conditions of the original problem (1.1) and to verify the closed-form expression of the Lagrangian multipliers of the orthogonality constraints.

We first introduce the definition of Clark subgradient for nonsmooth functions.

DEFINITION 2.1 ([13]). *For any Lipschitz continuous f on $\mathbb{R}^{n\times p}$, for any direction $D \in \mathbb{R}^{n\times p}$, the generalized directional derivative of f along D, denoted as $f^o(X, D)$, is defined by*

$$f^o(X, D) := \limsup_{Y\to X, t\to 0^+} \frac{f(Y + D) - f(Y)}{t}. \tag{2.1}$$

Based on generalized directional derivative of f, the Clark subgradient of f, denoted as $\partial f(X)$, is defined by

$$\partial f(X) := \left\{W \in \mathbb{R}^{n\times p} | \, \langle W, D\rangle \le f^o(X, D) \text{ for any } D \in \mathbb{R}^{n\times p}\right\}. \tag{2.2}$$

For brevity, we call it subgradient in the rest of this paper. As described in [40, Theorem 5.1] and [9], the optimality condition and first-order stationary point of (1.1) can be stated as the following.

DEFINITION 2.2 ([9]). *A point $X \in \mathcal{S}_{n,p}$ is called as first-order stationary point of (1.1) if and only if*

$$0 \in \mathcal{P}_{\mathcal{T}_X} \left(\nabla f(X) + \partial r(X)\right), \tag{2.3}$$

where $\mathcal{T}_X := \{Y \mid X^\top Y + Y^\top X = 0\}$ denotes the tangent space at X,

$$\mathcal{P}_{\mathcal{T}_X}(\mathcal{Y}) := \{Y - X\Phi(Y^\top X) \mid Y \in \mathcal{Y} \subseteq \mathbb{R}^{n\times p}\}$$

consists of all the projection points of $Y \in \mathcal{Y}$ onto the tangent space $\mathcal{T}_X$.

DEFINITION 2.3 ([33]).

1. *The sequential feasible direction of $\mathcal{M}$ at X is defined as*

$$\mathrm{SFD}(X) := \{d \in \mathbb{R}^{n\times p} \mid \mathcal{M} \ni X_k \to X, \quad d = \lim_{t_k\to+\infty} t_k(X_k - X)\}.$$

2. *A point $X \in \mathcal{M}$ is called as first-order stationary point of (1.6) is and only if for any $D \in \mathrm{SFD}(X)$,*

$$\liminf_{t\to 0^+} \frac{h(X + tD) - h(X)}{t} \ge 0. \tag{2.4}$$

Next, we verify the closed-form expression of the Lagrangian multipliers of the orthogonality constraints.

LEMMA 2.4. *For any $X \in \mathbb{R}^{n\times p}$ and $D \in \partial r(X)$, it holds that $X^\top D = \sum_{i=1}^{n} \gamma_i S(X_{i\cdot}^\top)$, where S is defined by (1.5).*

Proof. Firstly, for any $w \in \mathbb{R}^p$, we have

$$\partial \|w\|_2 = \begin{cases} \frac{w}{\|w\|_2}, & \|w\|_2 \neq 0; \\ \{\tilde{w} | \tilde{w} \in \mathbb{R}^p, \|\tilde{w}\|_2 \le 1\}, & \|w\|_2 = 0. \end{cases} \tag{2.5}$$

On the other hand, for any $D \in \partial r(X)$, it holds that

$$D = [\gamma^1 d_1, ..., \gamma^n d_n]^\top, \tag{2.6}$$

where $d_i \in \partial \left\|X_{i\cdot}^\top\right\|$, $i = 1, ..., n$. Combining the formula (2.5) and (2.6) together, for any $X \in \mathbb{R}^{n\times p}$, we can obtain

$$X^\top D = \sum_{i=1}^{n} X_{i\cdot}^\top D_{i\cdot} = \sum_{i=1}^{n} \gamma_i X_{i\cdot}^\top d_i^\top = \sum_{i=1, X_{i\cdot}\neq 0}^{n} \gamma_i X_{i\cdot}^\top d_i^\top + \sum_{i=1, X_{i\cdot}=0}^{n} \gamma_i X_{i\cdot}^\top d_i^\top$$

$$= \sum_{i=1, X_{i\cdot}\neq 0}^{n} \gamma_i X_{i\cdot}^\top d_i^\top = \sum_{i=1, X_{i\cdot}\neq 0}^{n} \gamma_i \frac{X_{i\cdot}^\top X_{i\cdot}}{\|X_{i\cdot}\|_2} = \sum_{i=1, X_{i\cdot}\neq 0}^{n} \gamma_i S(X_{i\cdot}^\top) = \sum_{i=1}^{n} \gamma_i S(X_{i\cdot}^\top),$$

which completes the proof. □

For brevity, we denote

$$\Lambda_f(X) := \Phi(X^\top \nabla f(X)), \qquad \Lambda_r(X) := \sum_{i=1}^{n} \gamma_i S(X_{i\cdot}^\top), \tag{2.7}$$

and $\Lambda(X) = \Lambda_f(X) + \Lambda_r(X)$. In the following, we show that $\Lambda(X)$ shares the same expression as the Lagrangian multipliers of the orthogonality constraints at any first-order stationary points.

LEMMA 2.5. *Suppose Assumption 1.1 holds, then X^* is a first-order stationary point of (1.1) if and only if it holds that*

$$\begin{cases} 0 \in \nabla f(X^*) + \partial r(X^*) - X^* \Lambda^*; \\ {X^*}^\top X^* = I_p, \end{cases} \tag{2.8}$$

where $\Lambda^* := \Lambda(X^*)$.

Proof. By Definition 2.2, suppose X^* is a first-order stationary point of (1.1). Then we have

$$0 \in \mathcal{P}_{\mathcal{T}_X} \left(\nabla f(X^*) + \partial r(X^*) \right),$$

which shows that there exists $D \in \partial r(X^*)$ such that

$$0 = \mathcal{P}_{\mathcal{T}_{X^*}} \left(\nabla f(X^*) + D \right).$$

Notice that $\mathcal{P}_{\mathcal{T}_{X^*}} \left(\nabla f(X^*) + D \right) = \nabla f(X^*) + D - X^* \Phi \left({X^*}^\top (\nabla f(X^*) + D) \right)$. As a result,

$$\begin{cases} 0 \in \nabla f(X^*) + \partial r(X^*) - X^* \Phi \left({X^*}^\top (\nabla f(X^*) + D) \right); \\ {X^*}^\top X^* = I_p. \end{cases}$$

We have $\Lambda_f(X^*) = \Phi \left({X^*}^\top \nabla f(X^*) \right)$ and $\Lambda_r(X^*) = {X^*}^\top D$ by simple calculations and Lemma 2.4, respectively, which implies that X^* satisfies relationship (2.8).

On the other hand, for any X^* satisfying (2.8), we can choose $D \in \partial r(X^*)$ such that $0 = \nabla f(X^*) + D - X^* \Lambda^*$. By using the fact ${X^*}^\top D = \Lambda_r(X^*)$ again, we have

$$\begin{aligned} 0 &= \nabla f(X^*) + D - X^* \Phi({X^*}^\top \nabla f(X^*) + {X^*}^\top D) \\ &= \mathcal{P}_{\mathcal{T}_X}(\nabla f(X) + D) \in \mathcal{P}_{\mathcal{T}_X}(\nabla f(X) + \partial r(X)), \end{aligned}$$

which implies that X^* is a first-order stationary point of (1.1). We complete the proof. □

2.2. Equivalence. In this subsection, we explore the relationship between (1.1) and the new proposed model PenC. Before starting our proof, we denote $\mathcal{L}(X, \Lambda)$ as the augmented Lagrangian function,

$$\mathcal{L}(X, \Lambda) = f(X) + r(X) - \frac{1}{2} \left\langle \Lambda, X^\top X - I_p \right\rangle + \frac{\beta}{4} \left\| X^\top X - I_p \right\|_{\mathrm{F}}^2.$$

Let $g(X) := \frac{1}{2}\langle \Lambda(X), X^\top X - I_p\rangle$, $\bar{f}(X) := f(x) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2$ and $\tilde{h}(X) = f(X) + r(X) - g(X)$. Hence, we have

$$\nabla g(X) = X\Lambda(X) + \frac{1}{2}\nabla f(X)(X^\top X - I_p) + \frac{1}{2}\nabla^2 f(X)[X(X^\top X - I_p)]$$

and $h(x) = \tilde{h}(X) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 = \bar{f}(X) + r(X) - g(X)$. For any $X \in \mathcal{S}_{n,p}$, it also holds that $\nabla g(X) = X\Lambda(X)$. Hence, we can obtain the following proposition

PROPOSITION 2.6. *Suppose Assumption 1.1 holds, if $\tilde{X}$ is a first-order stationary point of the original problem (1.1), it must be a first-order stationary point of PenC. On the other hand, if $\tilde{X}$ is a first-order stationary point of PenC satisfying $\tilde{X}^\top\tilde{X} = I_p$, it must be a first-order stationary point of (1.1).* The proof is trivial and hence omitted.

In addition, we introduce the following constants for the theoretical analyses.

- $M_0 := \sup_{X\in\mathcal{M}} \|\nabla f(X)\|_{\mathrm{F}}$;
- $M_1 := \sup_{X\in\mathcal{M}} \|\Lambda(X)\|_2$;
- $M_2 := \sup_{X\in\mathcal{M}} r(X)$;
- $C_1 := \sup_{X\in\mathcal{M}} \tilde{h}(X) - \inf_{X\in\mathcal{M}} \tilde{h}(X)$;
- $L_0 := \sup_{X,Y\in\mathcal{M}} \frac{\|\nabla f(X)-\nabla f(Y)\|_{\mathrm{F}}}{\|X-Y\|_{\mathrm{F}}}$;
- $L_1 := \sup_{X\in\mathcal{M},Y\in\mathcal{M}} \frac{\|\Lambda(X)-\Lambda(Y)\|_{\mathrm{F}}}{\|X-Y\|_{\mathrm{F}}}$;
- $L_r := \sup_{X\in\mathcal{M},D\in\partial r(X)} \|D\|_{\mathrm{F}}$;
- $\bar{\gamma} = \sum_{i=1}^n \gamma_i$.

Here, we shall mention that L_1 is well defined due to the Liptschiz continuity of S. In addition, we emphasize that all the constants defined above are independent of the penalty parameter β.

In the following, we estimate the variation of $r(X)$ after a small perturbation.

LEMMA 2.7. *For any $X \in \mathcal{M}$ and $Q \in \mathbb{S}^{p\times p}\setminus\{\mathbf{0}_{p,p}\}$ satisfying $XQ \in \mathrm{SFD}(X)$. There exist $\bar{t} > 0$ and $D \in \partial r(X)$ such that $X + tXQ \in \mathcal{M}$ and*

$$|r(X + tXQ) - r(X) - t\langle D, XQ\rangle| \le t^2 M_2 \|Q\|_2^2 \tag{2.9}$$

hold for any $t \in [0, \bar{t}]$.

Proof. Firstly, we denote $\nu_i(t) := \gamma_i||(X + tXQ)_{i\cdot}||_2$ for $i = 1, ..., n$. Hence we have $r(X + tXQ) = \sum_{i=1}^n \nu_i(t)$, $r(X) = \sum_{i=1}^n \nu_i(0)$. On the other hand, we have

$$\nu_i(t) = \gamma_i||X_{i\cdot}^\top + tQX_{i\cdot}^\top||_2 = \gamma_i\left(X_{i\cdot}X_{i\cdot}^\top + 2X_{i\cdot}QX_{i\cdot}^\top t + X_{i\cdot}Q^2X_{i\cdot}^\top t^2\right)^{\frac{1}{2}}.$$

Let $\bar{t} := \min\left\{1, \frac{1}{2||Q||_2}\right\}$. For any given $i = 1, ..., n$, if $X_{i\cdot} \neq 0$, it can be easily verify that $(X + tXQ)_{i\cdot} \neq 0$ for any $t \in [0, \bar{t}]$. It follows from the Taylor's expansion that

$$(x + \delta)^{\frac{1}{2}} = x^{\frac{1}{2}} + \frac{1}{2}x^{-\frac{1}{2}}\delta - \frac{1}{8}x^{-\frac{3}{2}}\delta^2 + \mathcal{O}(\delta^3)$$

holds for any $x \neq 0$. By setting $x := X_{i\cdot}X_{i\cdot}^\top$, $\delta := 2X_{i\cdot}QX_{i\cdot}^\top t + X_{i\cdot}Q^2X_{i\cdot}^\top t^2$, we have

$$\nu_i'(0) = \gamma_i S_Q(X_{i\cdot}^\top). \tag{2.10}$$

Similarly, by setting $x := Z_{i\cdot}Z_{i\cdot}^\top$, $\delta := -2X_{i\cdot}QX_{i\cdot}^\top t - X_{i\cdot}Q^2X_{i\cdot}^\top t^2$, we have

$$\nu_i''(t) = \gamma_i S_{Q^2}(Z_{i\cdot}^\top) - \gamma_i T_Q(Z_{i\cdot}^\top), \tag{2.11}$$

where $Z_{i\cdot} := X_{i\cdot}^\top + tQX_{i\cdot}$ and

$$T_Q(x) := \begin{cases} \frac{(xQx^\top)^2}{\|x\|_2^3}, & \text{if } x \neq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{2.12}$$

Substituting the inequalities $||QZ_{i\cdot}^\top||_2^2 \leq ||Q||_2^2||Z_{i\cdot}^\top||_2^2$ and $|(Z_{i\cdot}QZ_{i\cdot}^\top)^2| \leq ||Q||_2^2||Z_{i\cdot}^\top||_2^4$ into (2.11), we have

$$|\nu_i''(t)| \leq 2\gamma_i \|Q\|_{\mathrm{F}}^2 \left\|Z_{i\cdot}^\top\right\|_2. \tag{2.13}$$

Moreover, by the mean value theorem, there exists $\hat{t} \in [0,\, t]$ such that

$$\nu_i(t) = \nu_i(0) + t\nu_i'(0) + \frac{1}{2}t^2\nu_i''(\hat{t}). \tag{2.14}$$

Combining (2.10) and (2.13) with (2.14), we immediately obtain

$$t^2 M_2 \|Q\|_{\mathrm{F}}^2 \geq t^2 \|Q\|_2^2 \cdot r(X + \hat{t}XQ) = t^2 \sum_{i=1, X_{i\cdot}\neq 0}^{n} \gamma_i \|Q\|_{\mathrm{F}}^2 \left\|Z_{i\cdot}^\top\right\|_2 \geq \frac{1}{2}t^2 \sum_{i=1}^{n} |\nu_i''(\hat{t})|$$

$$\geq \left| \sum_{i=1, X_{i\cdot}\neq 0}^{n} (\nu_i(t) - \nu_i(0) - t\nu_i'(0)) \right| = \left| r(X + tXQ) - r(X) - t \sum_{i=1, X_{i\cdot}\neq 0}^{n} \gamma_i S(X_{i\cdot}^\top) \right|$$

$$= \left| r(X + tXQ) - r(X) - t \sum_{i=1, X_{i\cdot}\neq 0}^{n} \gamma_i \left\langle (XQ)_{i\cdot}^\top, \frac{X_{i\cdot}^\top}{\left\|X_{i\cdot}^\top\right\|_2} \right\rangle \right|$$

$$= \left| r(X + tXQ) - r(X) - t \left\langle D, XQ \right\rangle \right|,$$

where $D \in \partial r(X)$. We complete the proof. □

Next, we illustrate a property of PenC that all the first-order stationary points of PenC are located in a compact set with constant size if β is sufficiently large.

LEMMA 2.8. *Suppose Assumption 1.1 holds. Let $\tilde{X}$ be an first-order stationary point of (1.6) with $\beta \geq \max\{2(M_0 + M_1), 2pL_1\}$, then $\left\|\tilde{X}\right\|_2^2 \leq 1 + \frac{2pL_1}{\beta} \leq 2$.*

Proof. Let $\tilde{X} = U\Sigma V^\top$ be the singular value decomposition (SVD) of $\tilde{X}$, and $\sigma_1 \geq ... \geq \sigma_p$ are the diagonal entries of Σ, namely, the singular values. If $\sigma_1 \leq 1$, we have $\left\|\tilde{X}\right\|_2^2 \leq 1$ which concludes the proof. In the rest of the proof, we assume that $\sigma_1 > 1$. If $\sigma_p > 1$, we denote $X^+ = \tilde{X}$ and $X^- = \mathbf{0}_{n,p}$. Otherwise, there exists $1 \leq l < p$ satisfying $\sigma_l > 1 \geq \sigma_{l+1}$. Let $\Sigma_1 = \mathrm{Diag}(\sigma_1, ..., \sigma_l)$, $\Sigma_2 = \mathrm{Diag}(\sigma_{l+1}, ..., \sigma_p)$, and

$$X^+ := U \begin{bmatrix} \Sigma_1 & \mathbf{0}_{l,p-l} \\ \mathbf{0}_{p-l,l} & \mathbf{0}_{p-l,p-l} \end{bmatrix} V^\top \quad \text{and} \quad X^- := U \begin{bmatrix} \mathbf{0}_{l,l} & \mathbf{0}_{l,p-l} \\ \mathbf{0}_{p-l,l} & \Sigma_2 \end{bmatrix} V^\top,$$

respectively. By the definition of X^+, we can conclude that $\|X^+\|_2 = \left\|\tilde{X}\right\|_2$ and ${X^+}^\top X^- = 0$. Besides, since $\|X^-\|_2 \leq 1$, we have that $X^- \in \mathrm{conv}\,(\mathcal{S}_{n,p})$. As a result, for any $t \in [0,\, 1]$, by the convexity of $\mathcal{M}$, $\tilde{X} - tX^+ = (1-t)\tilde{X} + tX^- \in \mathcal{M}$ and hence $-X^+ \in \mathrm{SFD}(\tilde{X})$. In addition, we denote $Q := V \begin{bmatrix} I_l & \mathbf{0}_{l,p-l} \\ \mathbf{0}_{p-l,l} & \mathbf{0}_{p-l,p-l} \end{bmatrix} V^\top$, then we can conclude that $X^+ = \tilde{X}Q$. Hence, for any $D \in \partial r(\tilde{X})$, we have

$${X^+}^\top D = Q\tilde{X}^\top D = Q\Lambda_r(\tilde{X}). \tag{2.15}$$

We now assume that the inequality to be proved does not hold. Namely, $\left\|\tilde{X}\right\|_2^2 > 1 + \frac{2pL_1}{\beta}$. We set $\bar{t} = \min\left\{\frac{1}{2}, (\sigma_1^2 - 1)\sigma_1^2 \Big/ ||X^{+\top}X^+||_{\mathrm{F}}^2\right\}$ By Lemma 2.7, there exists $D \in \partial r(\tilde{X})$ such that $r(\tilde{X} - t\tilde{X}Q) - r(X) + t\left\langle D, \tilde{X}Q\right\rangle = \mathcal{O}(t^2)$ holds for any $t \in [0,\, \bar{t}]$, which implies

$$\begin{aligned}
&\mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X})) \\
&= t\left\langle X^+, \nabla f(\tilde{X}) - \tilde{X}\Lambda(\tilde{X}) + \beta\tilde{X}(\tilde{X}^\top\tilde{X} - I_p)\right\rangle - \left[r(\tilde{X} - tX^+) - r(\tilde{X})\right] \quad (2.16)\\
&= t\left\langle X^+, \nabla f(\tilde{X}) - \tilde{X}\Lambda(\tilde{X}) + \beta\tilde{X}(\tilde{X}^\top\tilde{X} - I_p) + D\right\rangle + \mathcal{O}(t^2).
\end{aligned}$$

By using the facts that $\mathrm{tr}\,(AB) = \mathrm{tr}(BA)$ holds for any two square matrices A and B with same size, $X^{+\top}\tilde{X} = X^{+\top}X^+$, $X^+ = X^+Q$, and relationship (2.15), we have the following statements hold

$$\begin{aligned}
\left\langle X^+, \beta\tilde{X}(\tilde{X}^\top\tilde{X} - I_p)\right\rangle &= \beta\mathrm{tr}\left(X^{+\top}X^+\left(X^{+\top}X^+ - Q\right)\right),\\
\left\langle X^+, \nabla f(\tilde{X}) - \tilde{X}\Lambda_f(\tilde{X})\right\rangle &= -\mathrm{tr}\left(\left(X^{+\top}X^+ - Q\right)X^{+\top}\nabla f(\tilde{X})\right),\\
\left\langle X^+, -\tilde{X}\Lambda_r(\tilde{X}) + D\right\rangle &= -\mathrm{tr}\left(\left(X^{+\top}X^+ - Q\right)\Lambda_r(\tilde{X})\right).
\end{aligned}$$

Substituting the above equalities into (2.16), we obtain

$$\begin{aligned}
&\mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X}))\\
&= t\mathrm{tr}\left(\left(X^{+\top}X^+ - Q\right)\left(\beta X^{+\top}X^+ - X^{+\top}\nabla f(\tilde{X}) - \Lambda_r(\tilde{X})\right)\right) + \mathcal{O}(t^2) \qquad (2.17)\\
&\geq t\mathrm{tr}\left(\left(X^{+\top}X^+ - Q\right)\left(\frac{\beta}{2}X^{+\top}X^+\right)\right) + \mathcal{O}(t^2) \geq \frac{t\beta}{2}\left(\left\|\tilde{X}\right\|_2^2 - 1\right)\left\|\tilde{X}\right\|_2^2 + \mathcal{O}(t^2).
\end{aligned}$$

Here the first inequality uses the fact that $\frac{\beta}{2}X^{+\top}X^+ - X^{+\top}\nabla f(\tilde{X}) - \Lambda_r(\tilde{X}) \succeq \frac{\beta}{2}X^{+\top}X^+ - \Lambda(\tilde{X}) - X^{-\top}\nabla f(\tilde{X}) \succeq 0$, which is implied by the facts that $\beta > 2(M_0 + M_1)$ and $||X^{-\top}\nabla f(\tilde{X})||_2 \leq ||X^-||_2||\nabla f(\tilde{X})||_{\mathrm{F}} \leq ||\nabla f(\tilde{X})||_{\mathrm{F}}$. The second inequality uses the definition of X^+.

Recall the definition of $\bar{t}$, for any $t \in [0,\, \bar{t}]$, it holds that

$$\begin{aligned}
&\left\|(\tilde{X} - tX^+)^\top(\tilde{X} - tX^+) - I_p\right\|_{\mathrm{F}}^2 = \left\|\tilde{X}^\top\tilde{X} - I_p + (t^2 - 2t)X^{+\top}X^+\right\|_{\mathrm{F}}^2\\
&= \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2 + 2(t^2 - 2t)\left\langle\tilde{X}^\top\tilde{X} - I_p, X^{+\top}X^+\right\rangle + (t^2 - 2t)^2||X^{+\top}X^+||_{\mathrm{F}}^2\\
&\leq \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2 - (2t - t^2)\left[2(\sigma_1^2 - 1)\sigma_1^2 - (2t - t^2)||X^{+\top}X^+||_{\mathrm{F}}^2\right]\\
&\leq \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2 - 2(2t - t^2)\left[(\sigma_1^2 - 1)\sigma_1^2 - t||X^{+\top}X^+||_{\mathrm{F}}^2\right] \leq \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Combining the above inequality with the Lipschitz continuity of $\Lambda(X)$, we have

$$\begin{aligned}
&|\mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tX^+, \Lambda(\tilde{X} - tX^+))|\\
&= \left|\frac{1}{2}\left\langle\Lambda(\tilde{X}), (\tilde{X} - tX^+)^\top(\tilde{X} - tX^+) - I_p\right\rangle\right.
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}\left|\left\langle \Lambda(\tilde{X}-tX^+), (\tilde{X}-tX^+)^\top(\tilde{X}-tX^+)-I_p\right\rangle\right| \\
&\leq \frac{1}{2}\left\|\Lambda(\tilde{X})-\Lambda(\tilde{X}-tX^+)\right\|_{\mathrm{F}}\left\|(\tilde{X}-tX^+)^\top(\tilde{X}-tX^+)-I_p\right\|_{\mathrm{F}} \qquad (2.18)\\
&\leq \frac{tL_1}{2}\left\|X^+\right\|_{\mathrm{F}}\left\|(\tilde{X}-tX^+)^\top(\tilde{X}-tX^+)-I_p\right\|_{\mathrm{F}} \\
&\leq \frac{t\sqrt{p}L_1}{2}\left\|X^+\right\|_2\left\|\tilde{X}^\top\tilde{X}-I_p\right\|_{\mathrm{F}} \leq \frac{tpL_1}{2}\max\left\{1, \left\|\tilde{X}\right\|_2^2-1\right\}\left\|\tilde{X}\right\|_2^2.
\end{aligned}
$$

Combining (2.17) with (2.18), we immediately obtain that

$$
\begin{aligned}
&h(\tilde{X})-h(\tilde{X}-tX^+) \\
&\geq \left(\mathcal{L}(\tilde{X},\Lambda(\tilde{X}))-\mathcal{L}(\tilde{X}-tX^+,\Lambda(\tilde{X}))\right) \\
&\quad -\left|\left(\mathcal{L}(\tilde{X}-tX^+,\Lambda(\tilde{X}))-\mathcal{L}(\tilde{X}-tX^+,\Lambda(\tilde{X}-tX^+))\right)\right| \\
&\geq \frac{t\beta}{2}\left(\left\|\tilde{X}\right\|_2^2-1\right)\left\|\tilde{X}\right\|_2^2-\frac{tpL_1}{2}\max\left\{1,\left\|\tilde{X}\right\|_2^2-1\right\}\left\|\tilde{X}\right\|_2^2+\mathcal{O}(t^2) \\
&\geq \frac{tpL_1}{2}\max\left\{1,\left\|\tilde{X}\right\|_2^2-1\right\}\left\|\tilde{X}\right\|_2^2+\mathcal{O}(t^2)>0
\end{aligned}
$$

holds for sufficiently small t. Here, the third inequality follows the fact that $\beta \geq 2pL_1$ and $\left\|\tilde{X}\right\|_2^2 > 1+\frac{2pL_1}{\beta}$. Therefore, we can conclude that $\liminf_{t\to 0^+}\frac{h(\tilde{X}-tX^+)-h(\tilde{X})}{t}<0$, which reveals the contradictory to the optimality condition. Hence, we complete the proof. □

We next verify a sequential feasible direction at a given first-order stationary point.

LEMMA 2.9. *Suppose Assumption 1.1 holds. Let $\tilde{X}$ be a first-order stationary point of (1.6) with $\beta \geq \max\{2(M_0+M_1), 2pL_1\}$, then it holds that $\tilde{X}-\frac{1}{2}\tilde{X}(\tilde{X}^\top\tilde{X}-I_p) \in \mathrm{conv}\,(\mathcal{S}_{n,p})$, which implies $-\tilde{X}(\tilde{X}^\top\tilde{X}-I_p)\in \mathrm{SFD}(\tilde{X})$.*

Proof. Following from Lemma 2.8, we have $||\tilde{X}||_2 \leq \sqrt{2}$. Let $\tilde{X}=U\Sigma V^\top$ be the SVD of $\tilde{X}$, namely, $U\in\mathbb{R}^{n\times p}$ and $V\in\mathbb{R}^{p\times p}$ are the orthogonal matrices and Σ is a diagonal matrix with singular values of $\tilde{X}$ on its diagonal. Let σ_i $(i=1,...,p)$ be the diagonal entries of Σ, it holds that $\sigma_i\in[0,\sqrt{2}]$. Denote $W=\tilde{X}-\frac{1}{2}\tilde{X}(\tilde{X}^\top\tilde{X}-I_p)$, and we have

$$
W=U\left(\Sigma-\frac{1}{2}\Sigma(\Sigma^2-I_p)\right)V^\top.
$$

Let $\nu(t):=t-\frac{1}{2}t(t^2-1)$, then $|\nu(\sigma_i(X))|$ $(i=1,...,p)$ are the singular values of W. By easy calculation, we can obtain all the critical points of $\nu(t)$ in $[0,\sqrt{2}]$. They are $t=0$, $t=1$ and $t=\sqrt{2}$ with function values

$$
\nu(0)=0, \quad \nu(1)=1, \quad \nu(\sqrt{2})=\frac{\sqrt{2}}{2},
$$

respectively. Hence, we can conclude that $|\nu(t)|\leq 1$ holds for any $t\in[0,\sqrt{2}]$. Namely, $\sigma_i(W)\leq 1$ holds for any $i=1,...,p$. Therefore, $\tilde{X}-\frac{1}{2}\tilde{X}(\tilde{X}^\top\tilde{X}-I_p)=W\in\mathrm{conv}\,(\mathcal{S}_{n,p})\subset\mathcal{M}$. Together with the convexity of $\mathcal{M}$, we arrive at $-\frac{1}{2}\tilde{X}(\tilde{X}^\top\tilde{X}-I_p)=W-\tilde{X}\in\mathrm{SFD}(\tilde{X})$ and complete the proof. □

Now we are ready to reveal the relationship between the original problem (1.1) and PenC.

THEOREM 2.10. *Suppose Assumption 1.1 holds. Let $\tilde{X}$ be a first-order stationary point of (1.6) with $\beta\geq\max\{2(M_0+M_1),2pL_1\}$, then either $\tilde{X}$ is a first-order stationary point of (1.1), or $\sigma_{\min}(\tilde{X}^\top\tilde{X})\leq\frac{2M_1+\sqrt{2}L_1}{2\beta}$.*

Proof. We suppose the statement is not correct. Therefore, $\tilde{X} \notin \mathcal{S}_{n,p}$ and $\tilde{X}^\top \tilde{X} \succ \frac{2M_1+\sqrt{2}L_1}{2\beta} I_p$, which implies $W := \tilde{X}(\tilde{X}^\top \tilde{X} - I_p) \neq 0$. By Lemmas 2.8 and 2.9, we have $\left\|\tilde{X}\right\|_2 \leq \sqrt{2}$ and $-W \in \mathrm{SFD}(\tilde{X})$, respectively. Let $Q = -\tilde{X}^\top \tilde{X} + I_p$, we then obtain $W = -\tilde{X}Q$.

We set $\bar{t} = \frac{1}{3}$. By Lemma 2.7 and the fact that $||Q||_2 \leq 1$, there exists $D \in \partial r(\tilde{X})$ such that $r(\tilde{X} + t\tilde{X}Q) - r(X) - t\left\langle D, \tilde{X}Q \right\rangle = \mathcal{O}(t^2)$ holds for any $t \in [0,\, \bar{t}]$, which implies

$$
\begin{aligned}
&\mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) \\
&= t\left\langle W, \nabla f(\tilde{X}) - \tilde{X}\Lambda(\tilde{X}) + \beta\tilde{X}(\tilde{X}^\top\tilde{X} - I_p)\right\rangle - \left(r(\tilde{X} - tW) - r(\tilde{X})\right) + \mathcal{O}(t^2) \\
&= t\left\langle W, \nabla f(\tilde{X}) - \tilde{X}\Lambda(\tilde{X}) + \beta\tilde{X}(\tilde{X}^\top\tilde{X} - I_p) + D\right\rangle + \mathcal{O}(t^2) \\
&= t\left\langle W, \beta\tilde{X}(\tilde{X}^\top\tilde{X} - I_p)\right\rangle + t\cdot\left\langle W, \nabla f(\tilde{X}) - \tilde{X}\Lambda_f(\tilde{X})\right\rangle \\
&\quad + t\cdot\left\langle W, -\tilde{X}\Lambda_r(\tilde{X}) + D\right\rangle + \mathcal{O}(t^2) \\
&= t\beta||W||_{\mathrm{F}}^2 - t\mathrm{tr}\left((\tilde{X}^\top\tilde{X} - I_p)^2\Lambda_f(\tilde{X})\right) - t\mathrm{tr}\left((\tilde{X}^\top\tilde{X} - I_p)^2\Lambda_r(\tilde{X})\right) \\
&= t\cdot\mathrm{tr}\left((\tilde{X}^\top\tilde{X} - I_p)^2(\beta\tilde{X}^\top\tilde{X} - \Lambda(\tilde{X}))\right) + \mathcal{O}(t^2).
\end{aligned} \tag{2.19}
$$

On the other hand, $||\tilde{X}||_2 \leq \sqrt{2}$ results in the fact that $I_p \succ I_p - 2t\tilde{X}^\top\tilde{X} + t^2\tilde{X}^\top\tilde{X}(\tilde{X}^\top\tilde{X} - I_p) \succ -I_p$ holds for any $t \in [0,\, \bar{t}]$, which further implies

$$
\begin{aligned}
&\left\|(\tilde{X} - tW)^\top(\tilde{X} - tW) - I_p\right\|_{\mathrm{F}} = \left\|\tilde{X}^\top\tilde{X} - I_p - 2t\Phi(W^\top\tilde{X}) + t^2W^\top W\right\|_{\mathrm{F}} \\
&= \left\|\left(\tilde{X}^\top\tilde{X} - I_p\right)\left(I_p - 2t\tilde{X}^\top\tilde{X} + t^2\tilde{X}^\top\tilde{X}(\tilde{X}^\top\tilde{X} - I_p)\right)\right\|_{\mathrm{F}} \leq \left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}.
\end{aligned}
$$

Together with the fact that $\|W\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\tilde{X}^\top\tilde{X}(\tilde{X}^\top\tilde{X} - I_p)^2\right) \leq \left\|\tilde{X}\right\|_2^2\left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2$ and the Lipschitz continuity of $\Lambda(X)$, we obtain

$$
\begin{aligned}
&|\mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X} - tW))| \\
&= \frac{1}{2}\left|\left\langle\Lambda(\tilde{X}) - \Lambda(\tilde{X} - tW), (\tilde{X} - tW)^\top(\tilde{X} - tW) - I_p\right\rangle\right| \\
&\leq \frac{tL_1}{2}\|W\|_{\mathrm{F}}\left\|(\tilde{X} - tW)^\top(\tilde{X} - tW) - I_p\right\|_{\mathrm{F}} \\
&\leq \frac{tL_1}{2}\left\|\tilde{X}\right\|_2\left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2 \leq \frac{\sqrt{2}tL_1}{2}\left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2.
\end{aligned} \tag{2.20}
$$

Combining (2.19) with (2.20), we conclude that

$$
\begin{aligned}
&h(\tilde{X}) - h(\tilde{X} - tW) \\
&= \mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) + \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X} - tW)) \\
&\geq \mathcal{L}(\tilde{X}, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) - |\mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X})) - \mathcal{L}(\tilde{X} - tW, \Lambda(\tilde{X} - tW))| \\
&\geq t\cdot\mathrm{tr}\left((\tilde{X}^\top\tilde{X} - I_p)^2(\beta\tilde{X}^\top\tilde{X} - \Lambda(\tilde{X}))\right) - \frac{\sqrt{2}L_1}{2}\left\|\tilde{X}^\top\tilde{X} - I_p\right\|_{\mathrm{F}}^2 + \mathcal{O}(t^2) \\
&= t\cdot\left((\tilde{X}^\top\tilde{X} - I_p)^2\left(\beta\tilde{X}^\top\tilde{X} - \Lambda(\tilde{X}) - \frac{\sqrt{2}L_1}{2}I_p\right)\right) + \mathcal{O}(t^2) > 0
\end{aligned}
$$

holds for sufficiently small t, which contradicts to the first-order stationarity of $\tilde{X}$. Here the last equality holds because the facts that $\tilde{X}^\top \tilde{X} \succ \frac{2M_1+\sqrt{2}L_1}{2\beta} I_p$, $\left\|\Lambda(\tilde{X}) + \frac{\sqrt{2}L_1}{2} I_p\right\|_2 \leq M_1 + \frac{\sqrt{2}L_1}{2}$, and $\tilde{X} \notin \mathcal{S}_{n,p}$. We complete the proof. □

Theorem 2.10 illustrates that PenC may bring about "additional" first-order stationary points which are not stationary points of (1.6). The next lemma shows that these "additional" first-order stationary points are far away from the Stiefel manifold in a certain sense.

LEMMA 2.11. *Suppose Assumption 1.1 holds. Let* $0 < \delta \leq \frac{1}{3}$ *and* $\beta \geq \max\left\{2(M_0 + M_1), 2pL_1, \left(3M_1 + \frac{3\sqrt{2}}{2}L_1\right), \frac{2C_1}{\delta^2}\right\}$. *For any* $X \in \mathcal{M}$, *it holds that*

$$\sup_{\|X^\top X - I_p\|_{\mathrm{F}} \leq \delta} h(X) < \inf_{\|X^\top X - I_p\|_{\mathrm{F}} \geq 2\delta} h(X). \tag{2.21}$$

Moreover, any global minimizer X^* *of PenC satisfies* $X^* \in \mathcal{S}_{n,p}$, *which further implies that it is a global minimizer of problem* (1.1).

Proof. For any Y, Z satisfying $\left\|Y^\top Y - I_p\right\|_{\mathrm{F}} \leq \delta \leq \frac{1}{3}$ and $\left\|Z^\top Z - I_p\right\|_{\mathrm{F}} \geq 2\delta$:

$$h(Y) \leq \sup_{X \in \mathcal{M}} \left\{ f(X) - \frac{1}{2}\mathrm{tr}\left(\Lambda(X)(X^\top X - I_p)\right) + r(X) \right\} + \frac{\delta^2\beta}{4}$$
$$h(Z) \geq \inf_{X \in \mathcal{M}} \left\{ f(X) - \frac{1}{2}\mathrm{tr}\left(\Lambda(X)(X^\top X - I_p)\right) + r(X) \right\} + \delta^2\beta.$$

As a result, we have

$$h(Y) - h(Z) \leq C_1 - \frac{3\beta\delta^2}{4} < 0, \tag{2.22}$$

illustrating that the inequality (2.21) holds.

Moreover, let X^* be a global minimizer of PenC satisfying $X^* \notin \mathcal{S}_{n,p}$, which implies $\sigma_{\min}({X^*}^\top X^*) \leq \frac{2M_1+\sqrt{2}L_1}{2\beta} \leq \frac{1}{3} \leq 1 - 2\delta$ resulting from Theorem 2.10. Hence, $\left\|{X^*}^\top X^* - I_p\right\|_{\mathrm{F}} \geq 2\delta$ which contradicts to the global optimality of X^* due to the inequality (2.21). Therefore, X^* must be feasible. Since $h(X) = f(X)$ holds for any $X \in \mathcal{S}_{n,p}$, we can conclude that X^* is also the global minimizer of (1.1), and complete the proof. □

The proof of Lemma 2.11 directly gives the following corollary, and hence its proof is omitted.

COROLLARY 2.12. *Suppose Assumption 1.1 holds. Let* $0 < \delta \leq \frac{1}{3}$ *and* $\beta \geq \max\left\{2(M_0 + M_1), 2pL_1, \left(3M_1 + \frac{3\sqrt{2}}{2}L_1\right), \frac{2C_1}{\delta^2}\right\}$. *For any given* $X^0 \in \mathcal{M}$ *satisfying* $\left\|{X^0}^\top X^0 - I_p\right\|_{\mathrm{F}} \leq \delta$, *and* $\tilde{X}$ *to be a first-order stationary point of PenC satisfying* $h(\tilde{X}) \leq h(X^0)$, *it holds that* $\tilde{X}^\top \tilde{X} = I_p$ *and* $\tilde{X}$ *is a first-order stationary point of* (1.1).

REMARK 2.13. *The equivalence illustrated in Theorem 2.10 can be extended to the cases where* Γ *is any matrix in* $\mathbb{R}^{m\times n}$ *with* $m \geq n$. *The proof can be achieved by the same techniques in proving Theorem 2.10.*

3. Algorithm. In this section, we discuss how to design a first-order method for solving PenC with $\mathcal{M}$ chosen as a ball containing the Stiefel manifold. Namely, $\mathcal{M} = \mathcal{B}_K := \{X \in \mathbb{R}^{n\times p} | \, \|X\|_{\mathrm{F}} \leq K\}$, where $K > \sqrt{p}$ is a prefixed constant.

3.1. Algorithm Framework. To utilize the composite structure of the objective function of PenC, i.e. $h(Y)$ defined by (1.6). We consider to use the proximal gradient method which is known to be efficient for problems with nonsmooth regularizer. However, in $h(Y)$,

there are not only the smooth term $f(Y)+\frac{\beta}{4}||Y^\top Y-I_p||_{\mathrm{F}}^2$ and the nonsmooth regularizer $r(Y)$, but also an inseparable and nonsmooth term

$$\frac{1}{2}\left\langle \Lambda(Y), Y^\top Y - I_p\right\rangle. \tag{3.1}$$

The term (3.1) can neither be linearized due to its nonsmoothness nor be kept as it is in the proximal mapping due to its inseparability. Therefore, we approximate (3.1) at the current iterate X^k by the following linear function

$$\left\langle Y-X^k, X^k\Lambda(X^k)\right\rangle + \frac{1}{2}\left\langle \Lambda(X^k), {X^k}^\top X^k - I_p\right\rangle. \tag{3.2}$$

The error between (3.2) and (3.1) can be estimated by the following lemma.

LEMMA 3.1. *Suppose Assumption 1.1 holds. For any $X, Y\in\mathcal{B}_K$, it holds that*

$$\begin{aligned}&\left|\left\langle \Lambda(Y), Y^\top Y - I_p\right\rangle - 2\left\langle Y-X, X\Lambda(X)\right\rangle - \left\langle \Lambda(X), X^\top X - I_p\right\rangle\right|\\ &\le L_1\left\|Y-X\right\|_{\mathrm{F}}\left\|Y^\top Y-I_p\right\|_{\mathrm{F}} + M_1\left\|Y-X\right\|_{\mathrm{F}}^2.\end{aligned} \tag{3.3}$$

Proof.

$$\begin{aligned}&\left|\left\langle \Lambda(Y), Y^\top Y - I_p\right\rangle - 2\left\langle Y-X, X\Lambda(X)\right\rangle - \left\langle \Lambda(X), X^\top X - I_p\right\rangle\right|\\ =&\left|\left\langle \Lambda(Y), Y^\top Y - I_p\right\rangle - \left\langle \Lambda(X), Y^\top Y - I_p\right\rangle + \left\langle \Lambda(X), Y^\top Y - I_p\right\rangle\right.\\ &\left. - \left\langle \Lambda(X), X^\top X - I_p\right\rangle - 2\left\langle Y-X, X\Lambda(X)\right\rangle\right|\\ \le&\left|\left\langle \Lambda(Y), Y^\top Y - I_p\right\rangle - \left\langle \Lambda(X), Y^\top Y - I_p\right\rangle\right|\\ &+\left|\left\langle \Lambda(X), Y^\top Y - I_p\right\rangle - \left\langle \Lambda(X), X^\top X - I_p\right\rangle - 2\left\langle Y-X, X\Lambda(X)\right\rangle\right|\\ \le& L_1\left\|Y-X\right\|_{\mathrm{F}}\left\|Y^\top Y-I_p\right\|_{\mathrm{F}} + \left|\left\langle \Lambda(X), (Y-X)^\top (Y-X)\right\rangle\right|\\ \le& L_1\left\|Y-X\right\|_{\mathrm{F}}\left\|Y^\top Y-I_p\right\|_{\mathrm{F}} + M_1\left\|X-Y\right\|_{\mathrm{F}}^2,\end{aligned}$$

which completes the proof. □

The above Lemma tells that (3.2) is good approximation of (3.1). Particularly, when the feasibility violation $\left\|Y^\top Y-I_p\right\|_{\mathrm{F}}$ is in the same order of $\left\|Y-X\right\|_{\mathrm{F}}$, it is a second-order approximation. Denote

$$G^k := \nabla f(X^k) - X^k\Lambda(X^k) + \beta X^k({X^k}^\top X^k - I_p) \tag{3.4}$$

and using the approximate (3.1), we can obtain propose the following "approximate" proximal mapping.

$$Y^k = \underset{Y\in\mathbb{R}^{n\times p}}{\arg\min}\quad \mathrm{tr}\left({G^k}^\top Y\right) + r(Y) + \frac{1}{2\eta^k}\left\|Y-X^k\right\|_{\mathrm{F}}^2, \tag{3.5}$$

where the global minimizer of (3.5) has the following closed-form expression.

$$Y_{i\cdot}^k = \begin{cases}\frac{\left(\left\|X_{i\cdot}^k-\eta^k G_{i\cdot}^k\right\|_2-\gamma_i\eta^k\right)}{\left\|X_{i\cdot}^k-\eta^k G_{i\cdot}^k\right\|_2}\left(X_{i\cdot}^k-\eta^k G_{i\cdot}^k\right), & \text{when } \left\|X_{i\cdot}^k-\eta^k G_{i\cdot}^k\right\|_2\ge\gamma_i\eta^k;\\ 0, & \text{otherwise.}\end{cases} \tag{3.6}$$

Then, we compute X^{k+1} by projecting Y^k onto $\mathcal{B}_K$. We are ready to bring out the framework of our Algorithm 1.

Algorithm 1: Proximal Gradient Method for PenC (PenCPG)

Input: Choose initial guess X_0, and set $k := 0$;
while *certain stopping criterion is not reached* **do**
  Calculate G^k by definition (3.4);
  Choose stepsize η^k by certain strategy;
  Compute Y^k by formulation (3.6);
  if $\left\|Y^k\right\|_{\mathrm{F}} > K$ **then**
    $X^{k+1} = \frac{K}{\|Y^k\|_{\mathrm{F}}} Y^k$;
  else
    $X^{k+1} = Y^k$;
  Set $k := k + 1$;
Output: Return X^k.

3.2. Global Convergence . In this subsection, we establish the global convergence of PenCPG. Firstly, we present some necessary lemmas.

LEMMA 3.2. *Suppose Assumption 1.1 holds. For any $X, Y \in \mathcal{B}_K$,*

$$\begin{aligned}&\bar{f}(Y) - \bar{f}(X) - \left\langle \nabla \bar{f}(X), Y - X \right\rangle \\ \leq\ & \frac{L_0 + \beta\left(\left\|X^\top X - I_p\right\|_{\mathrm{F}} + \|X\|_2^2 + \|Y\|_2^2\right)}{2} \|Y - X\|_{\mathrm{F}}^2 .\end{aligned}$$

Proof. By the definition of L_0, we have

$$f(Y) - f(X) \leq \langle Y - X, \nabla f(X) \rangle + \frac{L_0}{2} \|Y - X\|_{\mathrm{F}}^2 .$$

Besides, it follows the equality that $Y^\top Y - X^\top X = \Phi\left((X^\top + Y^\top)(Y - X)\right)$ that

$$\begin{aligned}&\left\|Y^\top Y - I_p\right\|_{\mathrm{F}}^2 - \left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 - 4\left\langle Y - X, X(X^\top X - I_p)\right\rangle \\ &= 2\left\langle Y^\top Y - X^\top X - 2\Phi((Y - X)^\top X), X^\top X - I_p\right\rangle + \left\langle Y^\top Y - X^\top X, Y^\top Y - X^\top X\right\rangle \\ &= 2\left\langle (Y - X)^\top (Y - X), X^\top X - I_p\right\rangle + \left\langle Y^\top Y - X^\top X, Y^\top Y - X^\top X\right\rangle \\ &\leq 2\left\|X^\top X - I_p\right\|_{\mathrm{F}} \|Y - X\|_{\mathrm{F}}^2 + \left\langle \Phi\left((X^\top + Y^\top)(Y - X)\right), \Phi\left((X^\top + Y^\top)(Y - X)\right)\right\rangle \\ &\leq 2\left\|X^\top X - I_p\right\|_{\mathrm{F}} \|Y - X\|_{\mathrm{F}}^2 + 2\left(\|X\|_2^2 + \|Y\|_2^2\right) \|Y - X\|_{\mathrm{F}}^2 .\end{aligned}$$

Substituting the above inequality to the relationship $\bar{f}(X) = f(X) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2$, we immediately obtain the desired statement and complete the proof. □

LEMMA 3.3. *Suppose Assumption 1.1 holds. Let $0 < \delta \leq \frac{1}{3}$, $K \geq \frac{\sqrt{6p}}{2}$ and $\beta \geq 6M_1$. Suppose $\eta^k \leq \min\left\{\frac{1}{L_0 + 4\beta + \frac{\sqrt{6}}{2}L_1 + M_1}, \frac{1}{15\left(M_0 + \frac{2\sqrt{3p}}{3}M_1 + \frac{2\sqrt{3}}{9}\beta + L_r\right)}\right\}$, $X^k \in \mathcal{B}_K$ satisfies $\left\|{X^k}^\top X^k - I_p\right\|_{\mathrm{F}} \leq \delta$, and Y^k is global minimizer of proximal mapping (3.5). Then, it holds that*

$$\left\|Y^k - X^k\right\|_{\mathrm{F}} \geq \frac{\sqrt{6}\eta^k\beta}{18}\left\|{Y^k}^\top Y^k - I_p\right\|_{\mathrm{F}} .$$

Proof. Since Y^k is the global minimizer of (3.5), through the optimality condition, we obtain that $0 \in G^k + \partial r(Y^k) + \frac{1}{\eta^k}(Y^k - X^k)$. Equivalently, there exists $D \in \partial r(Y^k)$ satisfying $0 = G^k + D + \frac{1}{\eta^k}(Y^k - X^k)$, which implies

$$\begin{aligned}
&\left\|Y^k - X^k\right\|_{\mathrm{F}} = \eta^k \left\|G^k + D\right\|_{\mathrm{F}} \leq \eta^k \left(\left\|G^k\right\|_{\mathrm{F}} + \|D\|_{\mathrm{F}}\right) \\
&\leq \eta^k \left(\left\|\nabla f(X^k)\right\|_{\mathrm{F}} + \left\|X^k\right\|_{\mathrm{F}} \left\|\Lambda(X^k)\right\|_2 + \beta \left\|X^k\right\|_2 \left\|{X^k}^\top X^k - I_p\right\|_{\mathrm{F}} + \|D\|_{\mathrm{F}}\right) \\
&\leq \eta^k \left(M_0 + \frac{2\sqrt{3p}}{3} M_1 + \frac{2\sqrt{3}}{9}\beta + L_r\right),
\end{aligned}$$

where the last inequality uses the facts that $||X^k||_2 \leq \frac{2\sqrt{3}}{3}$ and $||X^k||_{\mathrm{F}} \leq \frac{2\sqrt{3p}}{3}$ which are implied by $\left\|{X^k}^\top X^k - I_p\right\|_{\mathrm{F}} \leq \frac{1}{3}$. On the other hand, it follows from the fact $Y^k = X^k + (Y^k - X^k)$ that

$$\begin{aligned}
&\left\|{Y^k}^\top Y^k - I_p\right\|_{\mathrm{F}} \qquad (3.7) \\
&\leq \left\|{X^k}^\top X^k - I_p\right\|_{\mathrm{F}} + 2\left\|\Phi({X^k}^\top (Y^k - X^k))\right\|_{\mathrm{F}} + \left\|(Y^k - X^k)^\top (Y^k - X^k)\right\|_{\mathrm{F}} \\
&\leq \frac{1}{3} + 2||X^k||_2||Y^k - X^k||_{\mathrm{F}} + ||Y^k - X^k||_{\mathrm{F}}^2 < \frac{1}{3} + \frac{4\sqrt{3}}{3} \cdot \frac{1}{15} + \left(\frac{1}{15}\right)^2 < \frac{1}{2},
\end{aligned}$$

where the third inequality uses the fact that $\eta^k \leq 1 \Big/ 15\left(\left(M_0 + \frac{2\sqrt{3p}}{3} M_1 + \frac{2\sqrt{3}}{9}\beta + L_r\right)\right)$.

The inequality (3.7) implies that

$$\sigma_{\max}({Y^k}^\top Y^k) \leq \frac{3}{2}, \qquad \sigma_{\min}({Y^k}^\top Y^k) \geq \frac{1}{2}, \qquad \left\|Y^k\right\|_{\mathrm{F}} \leq \frac{\sqrt{6p}}{2} \leq K. \qquad (3.8)$$

The relationship (3.8) further implies

$$\begin{aligned}
&\left\|X^k\right\|_2^2 + \left\|Y^k\right\|_2^2 + \left\|{X^k}^\top X^k - I_p\right\|_{\mathrm{F}} \leq \frac{4}{3} + \frac{9}{4} + \frac{1}{3} < 4, \qquad (3.9) \\
&\beta {Y^k}^\top Y^k - \Lambda(Y^k) \succeq \frac{\beta}{2} I_p - \Lambda(Y^k) \succeq \frac{\beta}{3} I_p.
\end{aligned}$$

Denote $\hat{G} := \nabla f(Y^k) - Y^k \Lambda(Y^k) + \beta Y^k ({Y^k}^\top Y^k - I_p)$, we have

$$\begin{aligned}
&\left\|{Y^k}^\top \hat{G} + {Y^k}^\top D\right\|_{\mathrm{F}} \\
&\geq \left\|\Phi\left({Y^k}^\top \left(\nabla f(Y^k) - Y^k \Lambda(Y^k) + \beta Y^k ({Y^k}^\top Y^k - I_p) + D\right)\right)\right\|_{\mathrm{F}} \\
&\geq \left\|\Phi\left(({Y^k}^\top Y^k - I_p)\left(\beta {Y^k}^\top Y^k - \Lambda(Y^k)\right)\right)\right\|_{\mathrm{F}} \geq \frac{\beta}{3}\left\|{Y^k}^\top Y^k - I_p\right\|_{\mathrm{F}}. \qquad (3.10)
\end{aligned}$$

Substituting the relationship (3.9) into Lemma 3.2, we immediately obtain

$$\begin{aligned}
&\left\|\hat{G} - G^k\right\|_{\mathrm{F}} \leq \left\|\nabla \bar{f}(Y^k) - \nabla \bar{f}(X^k)\right\|_{\mathrm{F}} + \left\|Y^k \Lambda(Y^k) - X^k \Lambda(X^k)\right\|_{\mathrm{F}} \\
&\leq (L_0 + 4\beta)\left\|Y^k - X^k\right\|_{\mathrm{F}} + \left\|Y^k(\Lambda(Y^k) - \Lambda(X^k))\right\|_{\mathrm{F}} + \left\|(Y^k - X^k)\Lambda(X^k)\right\|_{\mathrm{F}} \\
&\leq (L_0 + 4\beta)\left\|Y^k - X^k\right\|_{\mathrm{F}} + L_1\left\|Y^k\right\|_2\left\|Y^k - X^k\right\|_{\mathrm{F}} + \left\|\Lambda(X^k)\right\|_2\left\|Y^k - X^k\right\|_{\mathrm{F}}
\end{aligned}$$

$$\leq \left(L_0 + 4\beta + \frac{\sqrt{6}}{2} L_1 + M_1 \right) \left\| Y^k - X^k \right\|_{\mathrm{F}}.$$

Together with $\eta^k \leq \frac{1}{L_0 + 4\beta + \frac{\sqrt{6}}{2} L_1 + M_1}$, we have that

$$\frac{\eta^k}{\left\| Y^k \right\|_2} \left\| {Y^k}^\top (\hat{G} - G^k) \right\|_{\mathrm{F}} \leq \frac{\eta^k \left\| Y^k \right\|_2}{\left\| Y^k \right\|_2} \left\| (\hat{G} - G^k) \right\|_{\mathrm{F}} \leq \left\| Y^k - X^k \right\|_{\mathrm{F}}. \tag{3.11}$$

Combining the inequalities (3.10) and (3.11), we have

$$\begin{aligned}
&\left\| Y^k - X^k \right\|_{\mathrm{F}} \geq \frac{1}{\left\| Y^k \right\|_2} \left\| {Y^k}^\top (Y^k - X^k) \right\|_{\mathrm{F}} = \frac{\eta^k}{\left\| Y^k \right\|_2} \left\| {Y^k}^\top G^k + {Y^k}^\top D \right\|_{\mathrm{F}} \\
&\geq \frac{\eta^k}{\left\| Y^k \right\|_2} \left\| {Y^k}^\top \hat{G} + {Y^k}^\top D \right\|_{\mathrm{F}} - \frac{\eta^k}{\left\| Y^k \right\|_2} \left\| {Y^k}^\top (\hat{G} - G^k) \right\|_{\mathrm{F}} \\
&\geq \frac{\eta^k \beta}{3 \left\| Y^k \right\|_2} \left\| {Y^k}^\top Y^k - I_p \right\|_{\mathrm{F}} - \left\| Y^k - X^k \right\|_{\mathrm{F}},
\end{aligned}$$

which implies

$$\left\| Y^k - X^k \right\|_{\mathrm{F}} \geq \frac{\eta^k \beta \left\| {Y^k}^\top Y^k - I_p \right\|_{\mathrm{F}}}{6 \left\| Y^k \right\|_2} \geq \frac{\sqrt{6} \eta^k \beta}{18} \left\| {Y^k}^\top Y^k - I_p \right\|_{\mathrm{F}},$$

and we complete the proof. □

LEMMA 3.4. *Suppose Assumption 1.1 holds. Let* $0 < \delta \leq \frac{1}{3}$, $K \geq \frac{\sqrt{6p}}{2}$ *and* $\beta \geq \max\{6M_1, 12\sqrt{6}L_1\}$. *Suppose that* $X^k \in \mathcal{B}_K$ *satisfies* $\left\| {X^k}^\top X^k - I_p \right\|_{\mathrm{F}} \leq \delta$, $\eta^k \leq \min\left\{ \frac{1}{L_0 + 4\beta + \frac{\sqrt{6}}{2} L_1 + M_1}, \frac{1}{15\left(M_0 + \frac{2\sqrt{3p}}{3} M_1 + \frac{2\sqrt{3}}{9}\beta + L_r \right)}, \frac{1}{4(L_0 + 4\beta + M_1)} \right\}$, *and* Y^k *is global minimizer of proximal mapping (3.5). Then, it holds that*

$$h(Y^k) \leq h(X^k) - \frac{1}{8\eta^k} \left\| Y^k - X^k \right\|_{\mathrm{F}}^2.$$

Proof. By using the fact that Y^k is a global minimizer of (3.5), we obtain

$$\left\langle G^k, Y^k - X^k \right\rangle + r(Y^k) + \frac{1}{2\eta^k} \left\| Y^k - X^k \right\|_{\mathrm{F}}^2 \leq r(X^k) \tag{3.12}$$

Recalling the first inequality in (3.9), we have

$$\begin{aligned}
&h(Y^k) - h(X^k) \\
&= \bar{f}(Y^k) - \bar{f}(X^k) - \frac{1}{2}\left(g(Y^k) - g(X^k) \right) + r(Y^k) - r(X^k) \\
&\leq \left\langle Y^k - X^k, \nabla \bar{f}(X^k) \right\rangle + \frac{L_0 + 4\beta}{2} \left\| Y^k - X^k \right\|_{\mathrm{F}}^2 + r(Y^k) - r(X^k) \\
&\quad - \frac{1}{2}\left(\left\langle {Y^k}^\top Y^k - I_p, \Lambda(X^k) \right\rangle - g(X^k) + g(Y^k) - \left\langle {Y^k}^\top Y^k - I_p, \Lambda(X^k) \right\rangle \right) \\
&\leq \left\langle Y^k - X^k, \nabla \bar{f}(X^k) - X^k \Lambda(X^k) \right\rangle + r(Y^k) - r(X^k) + \frac{L_0 + 4\beta}{2} \left\| Y^k - X^k \right\|_{\mathrm{F}}^2
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}\left(Y^k-X^k\right)^\top\left(Y^k-X^k\right)\Lambda(X^k)+\frac{L_1}{2}\left\|Y^k-X^k\right\|_{\mathrm{F}}\left\|{Y^k}^\top Y^k-I_p\right\|_{\mathrm{F}}\\
&\le\left\langle Y^k-X^k,G^k\right\rangle+r(Y^k)-r(X^k)+\left(\frac{L_0+4\beta+M_1}{2}+\frac{3\sqrt{6}L_1}{\eta^k\beta}\right)\left\|Y^k-X^k\right\|_{\mathrm{F}}^2.\\
&\le\left(-\frac{1}{2\eta^k}+\frac{L_0+4\beta+M_1}{2}+\frac{3\sqrt{6}L_1}{\eta^k\beta}\right)\left\|Y^k-X^k\right\|_{\mathrm{F}}^2.
\end{aligned}
$$

Here, the four inequalities are implied by Lemmas 3.2, 3.1, 3.3 and the inequality (3.12), respectively.

As a result, with $\beta\ge 12\sqrt{6}L_1$ and $\eta_k\le\frac{1}{4(L_0+4\beta+M_1)}$, we immediately obtain

$$
h(Y^k)-h(X^k)\le\left(-\frac{1}{2\eta^k}+\frac{1}{8\eta^k}+\frac{1}{4\eta^k}\right)\left\|Y^k-X^k\right\|_{\mathrm{F}}^2\le-\frac{1}{8\eta^k}\left\|Y^k-X^k\right\|_{\mathrm{F}}^2.
$$

□

LEMMA 3.5. *Suppose Assumption 1.1 holds. Let* $0<\delta\le\frac{1}{3}$, $K\ge\frac{\sqrt{6p}}{2}$ *and* $\beta\ge\max\{6M_1,\max\{2p,12\sqrt{6}\}L_1,2(M_0+M_1),\frac{2C_1}{\delta^2}\}$. *Suppose that* $\{X^k\}$ *and* $\{Y^k\}$ *are the iterate and intermediate iterate sequences generated by PenCPG, starting from the initial point* $X_0\in\mathcal{B}_K$ *satisfying* $\left\|X_0^\top X_0-I_p\right\|_{\mathrm{F}}\le\frac{\delta}{2}$, *and adopting the stepsize* $\eta^k\in[\frac{1}{2}\eta^+,\eta^+]$ *where*

$$
\eta^+=\min\left\{\frac{1}{L_0+4\beta+\frac{\sqrt{6}}{2}L_1+M_1},\frac{1}{15\left(M_0+\frac{2\sqrt{3p}}{3}M_1+\frac{2\sqrt{3}}{9}\beta+L_r\right)},\frac{1}{4\left(L_0+4\beta+M_1\right)}\right\}. \tag{3.13}
$$

Then for any $k>0$, *it holds that* $\left\|{X^k}^\top X^k-I_p\right\|_{\mathrm{F}}\le\delta$.

Proof. Suppose that $h(X^j)\le h(X^0)$ holds for any $j\le k$, which implies $h(Y^k)\le h(X^k)\le h(X^0)$ according to Lemma 3.4.

Then it holds that $Y^k\in\mathcal{B}_K$ due to (3.8) in Lemma 3.3. Then, we have $X^{k+1}=Y^k$ resulting from the Step 9 in PenCPG, which implies $h(X^{k+1})\le h(X^k)\le h(X^0)$. Hence, by mathematical induction, the inequality $h(X^k)\le h(X^0)$ holds for any $k>0$.

Recalling Lemma 2.11, we can conclude that for any $k>0$, $\left\|{X^k}^\top X^k-I_p\right\|_{\mathrm{F}}\le\delta$. □

LEMMA 3.6. *Suppose Assumption 1.1 holds. Let* $\delta\le\frac{1}{3}$, $K>\sqrt{p}$, $\beta\ge 6M_1$ *and* $\hat{\eta}\ge 0$. *Suppose that* $X\in\mathcal{B}_K$ *satisfies* $\left\|X^\top X-I_p\right\|_{\mathrm{F}}\le\delta$ *and*

$$
X=\arg\min_{Y\in\mathbb{R}^{n\times p}}\ \left\langle\nabla f(X)-X\Lambda(X)+\beta X(X^\top X-I_p),Y\right\rangle+r(Y)+\frac{1}{2\hat{\eta}}\left\|Y-X\right\|_{\mathrm{F}}^2.
$$

Then X *is a first-order stationary point of* (1.1).

Proof. By the first-order optimality condition of

$$
\min_{Y\in\mathbb{R}^{n\times p}}\ \left\langle\nabla f(X)-X\Lambda(X)+\beta X(X^\top X-I_p),Y\right\rangle+r(Y)+\frac{1}{2\hat{\eta}}\left\|Y-X\right\|_{\mathrm{F}}^2,
$$

there exists $D\in\partial r(X)$ such that $0=\nabla f(X)-X\Lambda(X)+\beta X(X^\top X-I_p)+D$. Since $\beta\ge 6M_1$, and $\left\|X^\top X-I_p\right\|_{\mathrm{F}}\le\delta$, we have $\beta X^\top X-\Lambda(X)\succeq(\frac{2\beta}{3}-M_1)I_p\succeq\frac{\beta}{2}I_p$. By simple calculation, we have

$$
\begin{aligned}
0&=\left\langle X(X^\top X-I_p),\nabla f(X)-X\Lambda(X)+\beta X(X^\top X-I_p)+D\right\rangle\\
&=\mathrm{tr}\left((X^\top X-I_p)^2(\beta X^\top X-\Lambda(X))\right)\ge\frac{\beta}{2}\left\|X^\top X-I_p\right\|_{\mathrm{F}}^2\ge 0,
\end{aligned}
$$

which leads to $X^\top X = I_p$. Therefore, we obtain

$$\begin{cases} 0 \in \nabla f(X) + \partial r(X) - X\Lambda(X); \\ X^\top X = I_p, \end{cases}$$

which implies that X is a first-order stationary point of (1.1) by Lemma 2.5. □

THEOREM 3.7. *Suppose Assumption 1.1 holds. Let* $0 < \delta \le \frac{1}{3}$, $K \ge \frac{\sqrt{6p}}{2}$ *and* $\beta \ge \max\{6M_1, \max\{2p, 12\sqrt{6}\}L_1, 2(M_0+M_1), \frac{2C_1}{\delta^2}\}$. *Suppose that* $\{X^k\}$ *is the iterate sequence generated by PenCPG, starting from the initial point* $X_0 \in \mathcal{B}_K$ *satisfying* $\left\|X_0^\top X_0 - I_p\right\|_{\mathrm{F}} \le \frac{\delta}{2}$, *and adopting the stepsize* $\eta^k \in [\frac{1}{2}\eta^+, \eta^+]$ *where* η^+ *is defined by* (3.13). *Then* $\{X^k\}$ *exists clustering point and any clustering point is a first-order stationary point of* (1.1). *More precisely, for any* $N \ge 1$, *it holds that*

$$\min_{0\le k\le N-1} \left\|X^{k+1} - X^k\right\|_{\mathrm{F}} \le \sqrt{\frac{(16C_1 + \beta\delta^2)\eta^+}{2N}}.$$

Proof. By Lemma 3.4 and Lemma 3.5, we conclude that $\left\|{X^k}^\top X^k - I_p\right\|_{\mathrm{F}} \le \delta$ and

$$h(X^{k+1}) \le h(X^k) - \frac{1}{8\eta^k}\left\|X^{k+1} - X^k\right\|_{\mathrm{F}}^2$$

hold for any $k \ge 0$. Therefore,

$$\begin{aligned} \sum_{k=0}^{N-1} \frac{\left\|X^{k+1} - X^k\right\|_{\mathrm{F}}^2}{8\eta^+} &\le \sum_{k=0}^{N-1} \frac{\left\|X^{k+1} - X^k\right\|_{\mathrm{F}}^2}{8\eta_k} \\ &\le h(X^0) - \liminf_{k\to+\infty} h(X^k) \le C_1 + \frac{\beta\delta^2}{16}, \end{aligned} \tag{3.14}$$

which implies $\lim_{k\to+\infty} \left\|X^{k+1} - X^k\right\|_{\mathrm{F}} = 0$.

Since $X^k \in \mathcal{B}_K$ and $\mathcal{B}_K$ is a compact set, we can conclude that $\{X^k\}$ exists clustering point. Let $\hat{X}$ be any clustering point of $\{X^k\}$. Due to the boundness of η_k, there exists a sequence $\{k_j\}_{j=1,2,\ldots}$ such that $X^{k_j} \to \hat{X}$, $\eta^{k_j} \to \hat{\eta}$. Recalling Lemma 3.5, $X^{k_j+1} = Y^{k_j}$, and we have

$$\begin{aligned} X^{k_j+1} = \arg\min_Y \quad & \left\langle Y, \nabla f(X^{k_j}) - X^{k_j}\Lambda(X^{k_j}) + \beta \hat{X^{k_j}}({\hat{X^{k_j}}}^\top X^{k_j} - I_p)\right\rangle \\ & + r(Y) + \frac{1}{2\eta^{k_j}}\left\|Y - X^{k_j}\right\|_{\mathrm{F}}^2. \end{aligned}$$

By the continuity of the explicit expression in (3.6), we can take limit of the both sides of the above inequality. Using the fact that $\left\|X^{k_j+1} - X^{k_j}\right\|_{\mathrm{F}} \to 0$, we obtain

$$\hat{X} = \arg\min_Y \quad \left\langle Y, \nabla f(\hat{X}) - \hat{X}\Lambda(\hat{X}) + \beta\hat{X}(\hat{X}^\top\hat{X} - I_p)\right\rangle + r(Y) + \frac{1}{2\hat{\eta}}\left\|Y - \hat{X}\right\|_{\mathrm{F}}^2.$$

Then, resulting from Lemma 3.6, $\hat{X}$ is a first-order stationary point of (1.1).

From (3.14), we have $\sum_{K=0}^{N-1}\left\|X^{k+1} - X^k\right\|_{\mathrm{F}}^2 \le \left(8C_1 + \frac{\beta\delta^2}{2}\right)\eta^+$, which immediately leads to $\min_{0\le k\le N-1}\left\|X^{k+1} - X^k\right\|_{\mathrm{F}} \le \sqrt{\frac{(16C_1+\beta\delta^2)\eta^+}{2N}}$. We complete the proof. □

REMARK 3.8. *The sublinear convergence rate of* $\left\|X^{k+1} - X^k\right\|_{\mathrm{F}}$ *illustrated in Theorem 3.7 actually tells us that PenCPG terminates after* $O(1/\epsilon^2)$ *iterations, if the stopping cri-*

terion is set as $\left\|X^{k+1}-X^k\right\|_{\mathrm{F}}<\epsilon$. *Meanwhile, we have* $\left\|{X^{k+1}}^\top X^{k+1}-I_p\right\|_{\mathrm{F}}<\frac{6\sqrt{6}}{\eta^+\beta}\epsilon$ *due to Lemma 3.3 and the fact that* $X^{k+1}=Y^k$.

3.3. Orthonormalization as Post-Process. When using the feasible approaches to solve the optimization problems with orthogonality constraints, we usually expect mild accuracy for the substationarity, but pursue high accuracy for the feasibility. Obviously, this requirement can not be reached by PenCPG. To this end, we impose an orthonormalization post-process after obtaining the last iterate X^k obtained by PenCPG. Namely,

$$X^k_{\mathrm{orth}}:=U^k(V^k)^\top, \tag{3.15}$$

where $X^k=U^k\Sigma^k(V^k)^\top$ is the economic SVD of X^k with $U^k\in\mathcal{S}_{n,p}$, $V^k\in\mathcal{S}_{p,p}$ and Σ^k is $p\times p$ diagonal matrix with the singular values of X^k on its diagonal.

The following proposition illustrates the postprocess (3.15) can further reduce the function value at the same time under mild assumptions. Consequently, X^k_{orth} is better output than X^k in any sense.

PROPOSITION 3.9. *Suppose Assumption 1.1 holds,* $\delta\le\frac{1}{3}$, $K\ge\sqrt{p+4\sqrt{p}\delta}$, $\beta\ge\max\{2(M_0+M_1),2pL_1,2\left(L_0+L_1+3M_1+2M_2\right)\}$, *and meanwhile* $X\in\mathcal{B}_K$ *satisfies* $\left\|X^\top X-I_p\right\|_{\mathrm{F}}\le\delta$. *Let* $X=U\Sigma V^\top$ *be the economic SVD of* X *and* $X_{\mathrm{orth}}:=UV^\top$, *then it holds that*

$$h(X_{\mathrm{orth}})\le h(X)-\left(\frac{\beta}{4}-\frac{1}{2}\left(L_0+L_1+3M_1+2M_2\right)\right)\left\|X^\top X-I_p\right\|_{\mathrm{F}}^2.$$

Proof. Let $T:=X-X_{\mathrm{orth}}$, we can obtain $T=X_{\mathrm{orth}}(V(\Sigma-I_p)V^\top)$ and $T-X_{\mathrm{orth}}\Phi({X_{\mathrm{orth}}}^\top T)=0$ by simple calculations. Moreover, it holds that $\|T\|_{\mathrm{F}}=\|\Sigma-I_p\|_{\mathrm{F}}\le\|(\Sigma-I_p)(\Sigma+I_p)\|_{\mathrm{F}}=\left\|X^\top X-I_p\right\|_{\mathrm{F}}$.

Recalling Lemma 2.7, there exists $\bar{t}\in\left(0,\,\frac{1}{2}\right)$ and $D\in\partial r(X_{\mathrm{orth}})$ such that the following statement holds for any $t\in(0,\bar{t}]$.

$$|r(X)-r(X_{\mathrm{orth}})-t\left\langle D,T\right\rangle|\le M_2\left\|\Sigma-I_p\right\|_{\mathrm{F}}^2\le M_2\left\|X^\top X-I_p\right\|_{\mathrm{F}}^2.$$

By Lemma 3.1, we have

$$\begin{aligned}&\left|\left\langle\Lambda(X),X^\top X-I_p\right\rangle-\left\langle\Lambda(X_{\mathrm{orth}}),{X_{\mathrm{orth}}}^\top X_{\mathrm{orth}}-I_p\right\rangle-2\left\langle T,X_{\mathrm{orth}}\Lambda(X_{\mathrm{orth}})\right\rangle\right|\\&\le L_1\|T\|_{\mathrm{F}}\left\|X^\top X-I_p\right\|_{\mathrm{F}}+M_1\|T\|_{\mathrm{F}}^2\le(L_1+M_1)\left\|X^\top X-I_p\right\|_{\mathrm{F}}^2.\end{aligned}$$

Moreover, it follows from the definition of L_0 that

$$|f(X)-f(X_{\mathrm{orth}})-\left\langle T,\nabla f(X_{\mathrm{orth}})\right\rangle|\le\frac{L_0}{2}\|T\|_{\mathrm{F}}^2\le\frac{L_0}{2}\left\|X^\top X-I_p\right\|_{\mathrm{F}}^2.$$

Combine the above three inequalities together, we obtain

$$\begin{aligned}&|[f(X)-g(X)+r(X)]-[f(X_{\mathrm{orth}})-g(X_{\mathrm{orth}})+r(X_{\mathrm{orth}})]|\\&\le|[f(X)-g(X)]-[f(X_{\mathrm{orth}})-g(X_{\mathrm{orth}})]+\left\langle D,T\right\rangle|+M_2\left\|X^\top X-I_p\right\|_{\mathrm{F}}^2\\&\le|f(X)-f(X_{\mathrm{orth}})-\left\langle T,X_{\mathrm{orth}}\Lambda(X_{\mathrm{orth}})+D\right\rangle|+\frac{1}{2}(L_1+M_1+2M_2)\left\|X^\top X-I_p\right\|_{\mathrm{F}}^2\\&\le|\left\langle T,\nabla f(X_{\mathrm{orth}})-X_{\mathrm{orth}}\Lambda(X_{\mathrm{orth}})+D\right\rangle|+\frac{1}{2}\left(L_0+L_1+M_1+2M_2\right)\left\|X^\top X-I_p\right\|_{\mathrm{F}}^2\end{aligned}$$

$$\leq \left|\mathrm{tr}\left((I_p - X^\top X)^2 \Lambda(X_{\mathrm{orth}})\right)\right| + \frac{1}{2}\left(L_0 + L_1 + M_1 + 2M_2\right)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2$$
$$\leq \frac{1}{2}\left(L_0 + L_1 + 3M_1 + 2M_2\right)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.$$

where the fourth inequality follows from the derivation of (2.17). Finally, we arrive at

$$\begin{aligned} & h(X) - h(X_{\mathrm{orth}}) \\ \geq & -\left|[f(X) - g(X) + r(X)] - [f(X_{\mathrm{orth}}) - g(X_{\mathrm{orth}}) + r(X_{\mathrm{orth}})]\right| + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 \\ \geq & \left(\frac{\beta}{4} - \frac{1}{2}\left(L_0 + L_1 + 3M_1 + 2M_2\right)\right)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2. \end{aligned}$$

□

The above proposition tells us the function value can further reduce after the post-process if the parameter β is sufficiently large and X is in a certain neighborhood of the Stiefel manifold. According to Lemma 3.5, all the iterates generated by PenCPG starting from a suitable initial guess are in such neighborhood.

4. Numerical Experiments. In this section, our main purpose is to illustrate the efficiency of PenCPG in solving the sparse variable PCA problems through comprehensive numerical experiments. Firstly, we describe the test problems, the stopping criterion and the default settings of the algorithm parameters. Then we introduce some state-of-the-art algorithms to compare with PenCPG, and describe the numerical performances of all the algorithms in comparison in solving the test problems.

All the numerical experiments in this section are run in serial in a desktop with Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz and 394GB RAM running MATLAB R2018a under Ubuntu 18.10.

4.1. Basic Setting. Firstly, we present how to generate the test problems. In this section, we focus on the spare variable PCA problem (1.2). The test problems are generated as following. We first randomly generate the data matrix $L := \mathbf{randn}(n, m)$, and set the covariance matrix as $M = LL^\top$. Besides, we choose $\gamma = b\sqrt{p + \log(n)}$ where the constant b controls the sparsity, see [19, 10] for details. The initial point X^0 is the projection of a randomly generated $n \times p$ matrix onto $\mathcal{S}_{n,p}$.

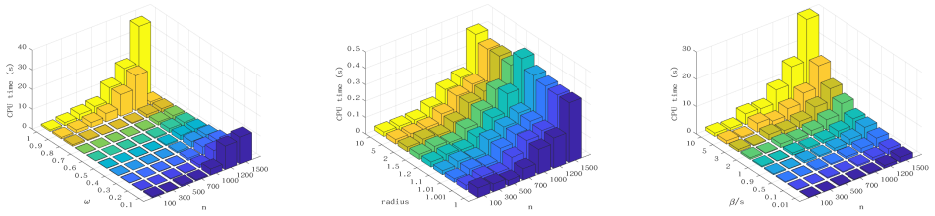
The stopping criterion for PenCPG in solving problem (1.2) is set as

$$\frac{1}{\eta^k}\left\|X^k - X^{k-1}\right\|_{\mathrm{F}} \leq 10^{-4}, \tag{4.1}$$

where the left-hand-side of (4.1) is used to measure the substationarity.

In the rest of this subsection, we determine the default choices of the penalty parameter β, the stepsize η^k, and the radius K in PenCPG through some special experiments. Theorem 3.7 suggests a sufficiently large β to guarantee the global convergence. However, large β causes slow convergence in practice. Here, we set $s := \|\nabla f(X_0)\|_{\mathrm{F}} + \bar{\gamma}$ as a basic scale of β. Theorem 3.7 also request a very restrictive stepsize η^k. In the default setting experiments, we compare two choices for the stepsizes. One is PenCPG with a fixed stepsize no greater than $\frac{1}{s}$, which is suggested in [28]. Another choice is the well-known Barzilar-Borwein (BB) stepsize. The expression of BB stepsize in our algorithm is similar to the choice in [37, 39]. However, due to the absence of the exact gradient for $f(X) - \frac{1}{2}\left\langle \Lambda(X), X^\top X - I_p\right\rangle$, the BB stepsize here is formed by the approximate gradient (3.4), namely, $\eta^k = \left\langle s^k, y^k\right\rangle / \left\langle y^k, y^k\right\rangle$, where $s^k = X^k - X^{k-1}$ and $y^k = G^k - G^{k-1}$. For convenience, we use PenCPG-BB to denote Algorithm 1 with η^k taking the BB stepsize.

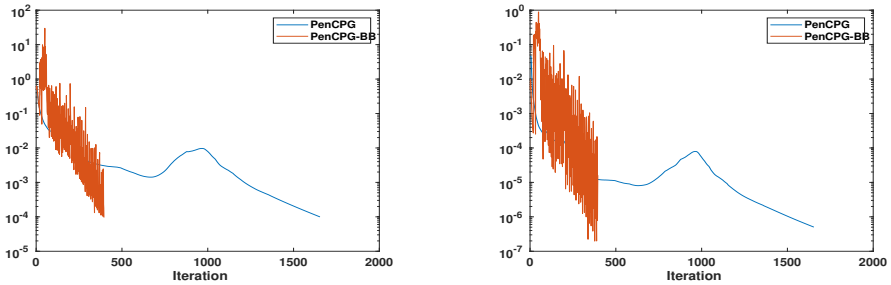
All the test problems in this subsection are generated with fixed parameters $p = 2$, $b = 0.10$ and $m = 100$. The parameter n varies among $\{100, 300, 500, 700, 1000, 1200, 1500\}$. We first fix $\beta = s$ and $K = 10\sqrt{p}$, and run PenCPG with stepsize $\eta^k = \frac{\omega}{s}$, where ω varies from 0.1 to 1. The result is shown in Figure 1(a). We can learn that $\omega = 0.5$ brings better performance than the other choices of ω. As a result, we set $\eta^k = \frac{1}{2s}$ as default setting of the fixed stepsize for PenCPG. Then we fix $\beta = s$ and $\eta^k = \frac{1}{2s}$, and run PenCPG with $K = r\sqrt{p}$, where r varies from 1 to 10. The result is illustrated in Figure 1(b). We can conclude that the performance of PenCPG is not sensitive to the parameter K. However, small K leads to frequent projections to $\mathcal{B}$ from the prime variables. Consequently, we choose $K = 10\sqrt{p}$ as default setting of parameter K. Finally, we fix $K = 10\sqrt{p}$ and $\eta^k = \frac{1}{2s}$, while β takes values from $0.1s$ to $10s$. The result is displayed in Figure 1(c). We find that the performance of PenCPG is not sensitive to β if it is not greater than s. Hence, we set $\beta = s$ as the default setting of parameter β.



(a) PenCPG on (1.2) with varying stepsize (b) PenCPG on (1.2) with varying radius (c) PenCPG on (1.2) with varying β

FIG. 1. *The CPU time of PenCPG with different combinations of parameters η^k, K and β.*

Now, we observe how the substationarity and the feasibility violations decay throughout the iterations when we call PenCPG to solve Problem 1.2. Parameters K and β are set as their default values. The numerical results are presented in Figure 2. We learn that the decay of feasibility violation has a similar tendency as the substationarity, which can partly explained by Lemma 3.3 and Lemma 3.4.

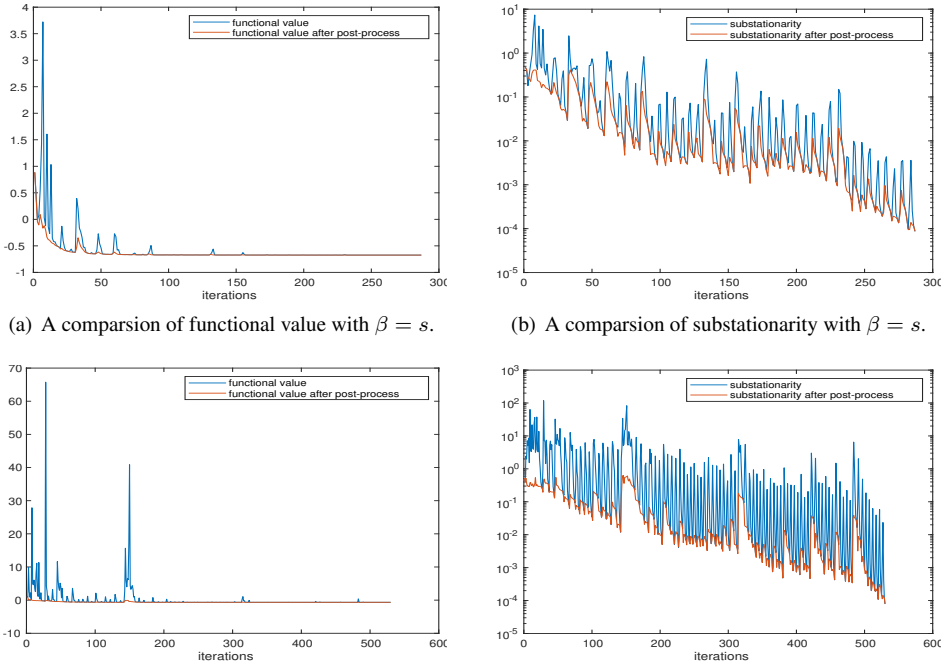


(a) A comparsion of substationarity on Problem 1.2 (b) A comparsion of feasibility on Problem 1.2

FIG. 2. *A comparison of the results of substationarity and feasibility violation for PenCPG on Problem 1.2 with $n = 500$, $p = 4$, $\gamma = 0.09$.*

Proposition 3.9 demonstrates that the post-process further reduces the function value if the parameter β is sufficiently large. In the end of this subsection, we illustrate by numerical

experiments that the post-process is of this good property when β is taken practical values. Firstly, we fix $n = 500$, $p = 4$ and $\gamma = 0.09$, and run PenCPG-BBwith $\beta = s$ and $\beta = 10s$. Then we perform the post-process at each iterate X^k and record the function value and substationarity at the point $X^k_{\rm orth}$, which is defined by (3.15). We plot the function values and substationarity of both X^k and $X^k_{\rm orth}$ together in Figure 3. We conclude that the post-process increase the accuracy in the feasibility while reduce functional value and substationarity at almost all the iterates generated by PenCPG-BB, no matter β takes small value ($\beta = s$) or large value ($\beta = 10s$).



(a) A comparsion of functional value with $\beta = s$.

(b) A comparsion of substationarity with $\beta = s$.

(c) A comparsion of functional value with $\beta = 10s$.

(d) A comparsion of substationarity with $\beta = 10s$.

FIG. 3. *A comparison of PenCPG-BB and its point-wise post-process on Problem 1.2 with* $n = 500$, $p = 4$, $\gamma = 0.09$.

4.2. Comparing with the State-of-the-Art Approaches. In this subsection, we compare PenCPG with some state-of-the-art approaches in solving the sparse variable PCA problem (1.2). The approaches in comparison includes two ADMM based algorithm, SOC [25] and PAMAL [11], and a manifold proximal gradient algorithm ManPG-Ada[9], which is an upgraded version of ManPG. All the parameters of PenCPG are set as their default values described in Section 5.1. The codes of SOC, PAMAL and ManPG-Ada are downloaded from github [1]. All the parameters of SOC, PAMAL and ManPG-Ada are set as the default values described in [9]. The stopping criterion for PenCPG , PenCPG-BB and ManPG-Ada are set as (4.1), and the stopping criterion for SOC and PAMAL adopts $\frac{\|Y^k - Z^k\|_{\rm F}}{\max\{1, \|Y^k\|_{\rm F}, \|Z^k\|_{\rm F}\}} + \frac{\|X^k - Y^k\|_{\rm F}}{\max\{1, \|X^k\|_{\rm F} + \|Y^k\|_{\rm F}\}} \le 10^{-4}$ which is comparable with (4.1) in the sense that the solution accuracy of these algorithms is in the same magnitude.

[1] https://github.com/chenshixiang/ManPG

We generate the test problems according to the way described in Section 4.1. For each parameters combination, which will be introduced later, we randomly generate 100 test problems and run the above mentioned five algorithms starting from the same randomly generated initial points. We record "Functional value", "Number of iterations", "Sparsity" and "CPU time" for each algorithm, and they stand for the average function value at the last iterate, number of iterations, percentage of zero-rows at the last iterate, and CPU time of the 100 runs, respectively.

Firstly, we fix the parameters $p = 4$, $b = 0.10$, $m = 100$, and vary n among $\{500, 1000, 1500, 2000\}$. The numerical results are plot in Figures 4(a), 4(d), 4(g) and 4(j). Secondly, we fix the parameters $n = 1000$, $b = 0.10$, $m = 100$, and vary p among $\{2, 4, 6, 8, 10, 12\}$. The numerical results are displayed in Figures 4(b), 4(e), 4(h) and 4(k). Finally, we fix the parameters $p = 4$, $n = 1000$, $m = 100$ and vary b among $\{0.08, 0.09, 0.10, 0.11, 0.12\}$. The numerical results are presented in Figures 4(c), 4(f), 4(i) and 4(l).

In all the cases, we can learn from Figure 4 that PenCPG and PenCPG-BB can obtain the same function values and sparsities, i.e. the same solution qualities, as the other three algorithms in comparison. Meanwhile, they take significantly less CPU time than the other algorithms. Besides, PenCPG-BB takes less iterations and less CPU time than PenCPG with fixed stepsize in all the cases. Therefore, we can conclude that BB stepsize is a more practical strategy in selecting the stepsize in Algorithm 1.

5. Conclusion. In this paper, we study the optimization problems with $\ell_{2,1}$-norm regularizer and orthogonality constraints, which has wide applications in data science, machine learning and image processing. The Riemannian proximal gradient based methods proposed very recently have shown great advantages in solving this type of problem comparing with the other existing approaches. However, in each iteration, the Riemannian proximal gradient methods require to solve a nonsmooth subproblem which becomes the computational bottleneck. We prove that the Lagrangian multipliers of the orthogonality constraints in this type of problem have closed-from expression and extend the exact penalty model PenC, firstly proposed in [38], with compact convex constraints to this nonsmooth case. We introduce an approximate of the gradient of the penalty function without the $\ell_{2,1}$ norm term, and consequently obtain an approximate proximal mapping with closed-form solution. Then, we propose a proximal gradient algorithm PenCPG based on the approximate proximal mapping to solve PenC. We establish the global convergence and the worst case complexity of PenCPG. Numerical experiments illustrate that PenCPG perform much better than the Riemannian proximal gradient methods and ADMM based approaches in the selected test problems. We believe that PenCPG has great potential in solving $\ell_{2,1}$ norm minimization with orthogonality constraints. In the future works, we will consider to extend PenC to general nonsmooth case, which is very challenging since the close-form expression of the Lagrangian multiplier does not exist anymore.

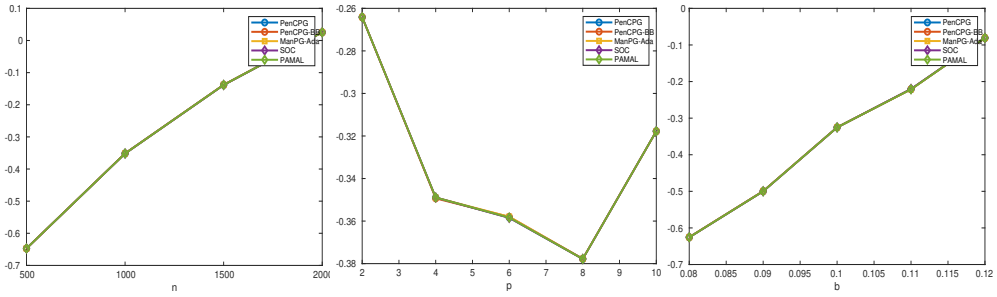
References.

[1] Traian Abrudan, Jan Eriksson, and Visa Koivunen. Conjugate gradient algorithm for optimization under unitary matrix constraint. *Signal Processing*, 89(9):1704–1714, 2009.

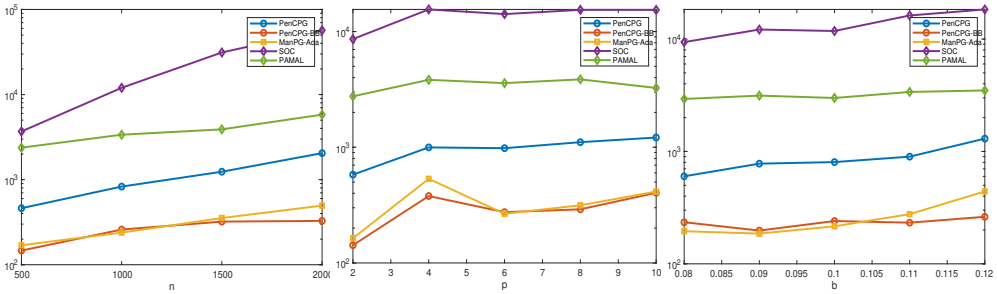
[2] Traian E Abrudan, Jan Eriksson, and Visa Koivunen. Steepest descent algorithms for optimization under unitary matrix constraint. *IEEE Transactions on Signal Processing*, 56(3):1134–1147, 2008.

[3] P-A Absil, Christopher G Baker, and Kyle A Gallivan. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007.

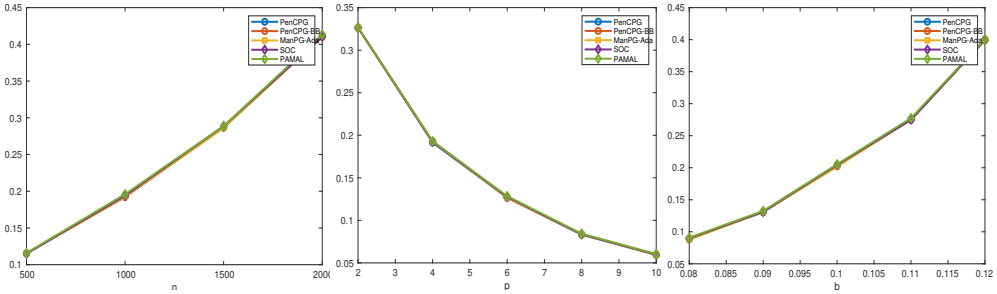
[4] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.



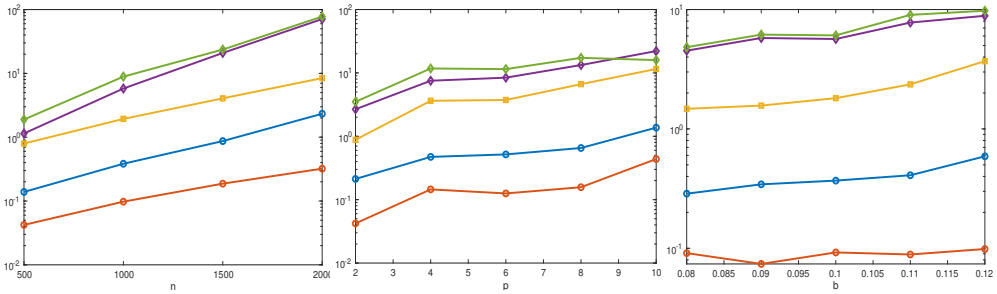
(a) Functional value (b) Functional value (c) Functional value



(d) Number of iterations (e) Number of iterations (f) Number of iterations



(g) Sparsity (h) Sparsity (i) Sparsity



(j) CPU time (k) CPU time (l) CPU time

FIG. 4. *A detailed comparison among some existing algorithms.*

[5] Miroslav Bacák, Ronny Bergmann, Gabriele Steidl, and Andreas Weinmann. A second order nonsmooth variational model for restoring manifold-valued images. *SIAM Journal on Scientific Computing*, 38(1):A567–A597, 2016.

[6] Pierre B Borckmans, S Easter Selvan, Nicolas Boumal, and P-A Absil. A riemannian subgradient algorithm for economic dispatch with valve-point effect. *Journal of Computational and Applied Mathematics*, 255:848–866, 2014.

[7] Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459, 2014.

[8] T Tony Cai, Zongming Ma, Yihong Wu, et al. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[9] Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Proximal gradient method for manifold optimization. *arXiv preprint arXiv:1811.00980*, 2018.

[10] Shixiang Chen, Shiqian Ma, Lingzhou Xue, and Hui Zou. An alternating manifold proximal gradient method for sparse pca and sparse cca. *arXiv preprint arXiv:1903.11576*, 2019.

[11] Weiqiang Chen, Hui Ji, and Yanfei You. An augmented lagrangian method for 1-regularized optimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 38(4):B570–B592, 2016.

[12] Xin Chen, Changliang Zou, R Dennis Cook, et al. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38(6):3696–3723, 2010.

[13] Frank H Clarke. *Optimization and nonsmooth analysis*, volume 5. Siam, 1990.

[14] Xiaoying Dai, Liwei Zhang, and Aihui Zhou. Adaptive step size strategy for orthogonality constrained line search methods. *arXiv preprint arXiv:1906.02883*, 2019.

[15] Gunther Dirr, Uwe Helmke, and Christian Lageman. Nonsmooth riemannian optimization with applications to sphere packing and grasping. In *Lagrangian and Hamiltonian Methods for Nonlinear Control 2006*, pages 29–45. Springer, 2007.

[16] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2): 303–353, 1998.

[17] Bin Gao, Xin Liu, Xiaojun Chen, and Ya-xiang Yuan. A new first-order algorithmic framework for optimization problems with orthogonality constraints. *SIAM Journal on Optimization*, 28(1):302–332, 2018.

[18] Bin Gao, Xin Liu, and Ya-xiang Yuan. Parallelizable algorithms for optimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 41(3): A1949–A1983, 2019.

[19] Philipp Grohs and Seyedehsomayeh Hosseini. Nonsmooth trust region algorithms for locally lipschitz functions on riemannian manifolds. *IMA Journal of Numerical Analysis*, 36(3):1167–1192, 2015.

[20] S Hosseini. Convergence of nonsmooth descent methods via kurdyka–lojasiewicz inequality on riemannian manifolds. *Hausdorff Center for Mathematics and Institute for Numerical Simulation, University of Bonn (2015,(INS Preprint No. 1523))*, 2015.

[21] Seyedehsomayeh Hosseini and André Uschmajew. A riemannian gradient sampling algorithm for nonsmooth optimization on manifolds. *SIAM Journal on Optimization*, 27(1):173–189, 2017.

[22] Jiang Hu, Andre Milzarek, Zaiwen Wen, and Yaxiang Yuan. Adaptive quadratically regularized newton method for riemannian optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1181–1207, 2018.

[23] Bo Jiang and Yu-Hong Dai. A framework of constraint preserving update schemes for optimization on stiefel manifold. *Mathematical Programming*, 153(2):535–575, 2015.

[24] Artiom Kovnatsky, Klaus Glashoff, and Michael M Bronstein. Madmm: a generic algorithm for non-smooth optimization on manifolds. In *European Conference on Computer Vision*, pages 680–696. Springer, 2016.

[25] Rongjie Lai and Stanley Osher. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58(2):431–449, 2014.

[26] Jonathan H Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650, 2002.

[27] Yasunori Nishimori and Shotaro Akaho. Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold. *Neurocomputing*, 67:106–135, 2005.

[28] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[29] Liqun Qi and Jie Sun. A nonsmooth version of newton's method. *Mathematical programming*, 58(1-3):353–367, 1993.

[30] Guy Rosman, Yu Wang, Xue Cheng Tai, Ron Kimmel, and Alfred M. Bruckstein. Fast regularization of matrix-valued images. In *European Conference on Computer Vision*, 2012.

[31] Guy Rosman, Xue Cheng Tai, Ron Kimmel, and Alfred M. Bruckstein. Augmented-lagrangian regularization of matrix-valued maps. *Methods & Applications of Analysis*, 21(1):121–138, 2014.

[32] Michael V Solodov and Benav F Svaiter. A globally convergent inexact newton method for systems of monotone equations. In *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, pages 355–369. Springer, 1998.

[33] Wenyu Sun and Ya-Xiang Yuan. *Optimization Theory and Methods*, volume 1 of *Springer Optimization and Its Applications*. Springer, New York, 2006.

[34] Jiliang Tang and Huan Liu. Unsupervised feature selection for linked social media data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 904–912. ACM, 2012.

[35] Magnus O Ulfarsson and Victor Solo. Sparse variable pca using geodesic steepest descent. *IEEE Transactions on Signal Processing*, 56(12):5823–5832, 2008.

[36] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

[37] Zaiwen Wen, Wotao Yin, Donald Goldfarb, and Yin Zhang. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. *SIAM Journal on Scientific Computing*, 32(4):1832–1857, 2010.

[38] Nachuan Xiao, Xin Liu, and Ya-xiang Yuan. A class of smooth exact penalty function methods for optimization problems with orthogonality constraints. 2020.

[39] Yunhai Xiao, Soon-Yi Wu, and Liqun Qi. Nonmonotone barzilai–borwein gradient algorithm for ℓ_1-regularized nonsmooth minimization in compressive sensing. *Journal of scientific computing*, 61(1):17–41, 2014.

[40] Wei Hong Yang, Lei-Hong Zhang, and Ruyi Song. Optimality conditions for the nonlinear programming problems on riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014.

[41] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. l2, 1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, 2011.

[42] Guanglu Zhou and Kim-Chuan Toh. Superlinear convergence of a newton-type algorithm for monotone equations. *Journal of optimization theory and applications*, 125(1): 205–221, 2005.