Int. J. Appl. Math. Comput. Sci., *2025*, *Vol. 35*, *No. 1*, *157–177*
DOI: 10.61822/amcs-2025-0012

A NOVEL NONCONVEX PENALTY METHOD FOR A RANK CONSTRAINED MATRIX OPTIMIZATION PROBLEM AND ITS APPLICATIONS

WENJUAN ZHANG [a], JIAYI YAO [a], FENG XIAO [b,*], YUPING WANG [a], YULIAN WU [c]

[a]School of Science
Xi'an Technological University
Xi'an, Shaanxi, 710021, China
e-mail: zhangwenjuan@xatu.edu.cn

[b]School of Computer Science
Xi'an Technological University
Xi'an, Shaanxi, 710021, China
e-mail:19456013@qq.com

[c]Department of Health Management
Xi'an Medical University
Xi'an, Shaanxi, 710021, China
e-mail:317903600@qq.com

The rank constrained nonconvex nonsmooth matrix optimization problem is an important and challenging issue. To solve it, we first design a penalty model in which the penalty term can be expressed as a sum of specific functions defined on smallest singular values of the matrix in question. We prove that the global minimizers of this penalty model are the same as those of the original problem. Second, we propose a flexible factorization format for the penalty function, such that the model enjoys the merit of fast computation in a SVD-free manner. We further prove that the factorization format problem is equivalent to the penalty one. A Bregman proximal gradient (BPG) method is developed for optimizing the factorization model. Third, we use two application problems as examples to illustrate that the problem considered has a wide application. Finally, some numerical experiments are conducted, and their results indicates the effectiveness of the proposed method.

Keywords: low rank, penalty method, nonconvex optimization, nonsmooth optimization, Bregman proximal gradient.

1. Introduction

The low-rank constrained matrix optimization problem has a wide application in the fields of financial engineering, image and video processing, machine learning and data mining, and so on, because low rank constraints can be used to find the intrinsic compact structure embedded in the large-scale data (Ji *et al.*, 2013; Li and White, 2001; Nguyen, 2017; Wang *et al.*, 2022). In financial engineering, it is used to optimize portfolios and manage risk, especially in the estimation of covariance matrices in asset pricing and risk assessment. In the field of image and video processing, low-rank matrix estimation helps to extract signals from noise and is used for data compression. In machine learning and data mining, low-rank matrix factorization technology is widely used in recommendation systems and data completion, which improves the accuracy and efficiency of data analysis. In this paper, we consider the following rank constrained optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} f(\mathbf{X}) \tag{1a}$$

subject to

$$\operatorname{rank}(\mathbf{X}) \le r, \quad \mathbf{X}^{\mathrm{T}}\mathbf{X} \preceq \mathbf{I}_n, \tag{1b}$$

where $f : \mathbb{R}^{m\times n} \to \mathbb{R}^{+}$ is proper, lower semicontinuous and continuously differentiable. Here $r \le \min\{m, n\}$

[*]Corresponding author

is a given positive integer. The rank of matrix $\mathbf{X}$ is the number of its nonzero singular values, i.e.,

$$\operatorname{rank}(\mathbf{X}) = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{X})^0,$$

where $\sigma_i(\mathbf{X})$ is the i-th largest singular value and

$$x^0 = \begin{cases} 1 & \text{if } x \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbf{I}_n$ is the $n \times n$ identity matrix. $\mathbf{X}^{\mathrm{T}}\mathbf{X} \preceq \mathbf{I}_n$ means that $\mathbf{I}_n - \mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a positive semidefinite symmetric matrix, which can be easily achieved by multiplying the matrix $\mathbf{X}$ by a normalized constant. The feasible region of problem (1) will be denoted by

$$\Omega = \left\{ \mathbf{X} \in \mathbb{R}^{m\times n} \middle| \operatorname{rank}(\mathbf{X}) \leq r,\ \mathbf{X}^{\mathrm{T}}\mathbf{X} \preceq \mathbf{I}_n \right\}.$$

The rank function is discontinuous and nonconvex, generally making problem (1) difficult to solve. In the past few decades, various approximations of the rank function have been extensively studied. The nuclear norm is a well-known convex approximation to the rank function, which is the l_1-norm of the singular value vector (Candes *et al.*, 2008; Gao and Sun, 2010; Recht *et al.*, 2011; Recht *et al.*, 2010). It has been shown that under certain incoherence assumptions on the singular vectors of the matrix, the nuclear norm regularization problem produces a near-optimal low-rank approximate solution (Bolte *et al.*, 2014). On the other hand, Fan and Li (2001) show that l_1-norm over-penalizes the large elements of a vector. In addition, they also proposed three criteria for determining a good penalty function: unbiasedness, sparsity and continuity at the origin. The l_1-norm satisfies both sparsity and continuity requirements, but it is biased. Similarly, the nuclear norm excessively penalizes large singular values and is biased.

In recent years, nonconvex penalties have attracted a lot of attention in sparse and low-rank learning problems (Ülkü and Kizgut, 2018; Zhong *et al.*, 2022), because researchers believe that a possible solution to a nonconvex problem can make up for the deficiency of a unique solution to a convex problem. Therefore, the l_1-norm can be replaced by the l_p-norm with $0 < p < 1$ (Chartrand, 2007; Ji *et al.*, 2013; Liang *et al.*, 2022; Lu *et al.* , 2015a; Xu *et al.*, 2012), and the Schatten p-norm

$$\|\mathbf{X}\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{X})^p \right)^{1/p} \qquad (0 < p < 1)$$

is proposed for approximating the rank function. There are other nonconvex norms, such as the sum of the logarithms of singular values

$$\|\mathbf{X}\|_{\log} = \sum_{i=1}^{\min\{m,n\}} \log(\sigma_i(\mathbf{X})),$$

cf. (Candes *et al.*, 2008; Fazel *et al.*, 2003).

Jia *et al.* (2020) proposed a generalized unitarily invariant gauge (GUIG) regularization

$$G_g(\mathbf{X}) = \inf \left\{ \sum_{i=1}^{d} g(|\lambda_i|) : \mathbf{X} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i {\mathbf{v}_i}^{\mathrm{T}},\ \|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1 \right\} \tag{2}$$

for fast low-rank matrix recovery, where d is a parameter satisfying $\operatorname{rank}(\mathbf{X}) \leq d \leq \min\{m,n\}$; $g(\cdot) : \mathbb{R}^{+} = [0\ {+\infty}) \to \mathbb{R}$ is a bounded function. The rank-one matrix decomposition $\mathbf{X} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i {\mathbf{v}_i}^{\mathrm{T}}$ is not unique. SVD leads to a decomposition, which imposes orthogonal constraint on the factors $\mathbf{u}_i$ and $\mathbf{v}_i$. Note that in (2), the orthogonality of the factors $\mathbf{u}_i$ and $\mathbf{v}_i$ is not enforced. The authors presented some conditions for function g under which $G_g(\mathbf{X})$ can be expressed as $G_g(\mathbf{X}) = \sum_{i=1}^{d} g[\sigma_i(\mathbf{X})]$ ($g(0) = 0$ is assumed). This regularization term is more general and covers the cases such as the rank function ($g(x) = x^0$), the nuclear norm ($g(x) = x$), the Schatten p-norm ($g(x) = x^p\ (0 < p < 1)$), and the log sum of singular values ($g(x) = \log x$).

All of the above works consider the low-rank constraint from the viewpoint of enforcing the sparsity of singular value vector by minimizing the sum of specific functions with the first few largest singular values. However, simply taking these functions as penalty terms and adding to the objective function $f(\mathbf{X})$ does not guarantee that a solution satisfies the constraint $\operatorname{rank}(\mathbf{X}) \leq r$ since they aim to find an approximated solution with the lowest possible rank rather than meet the low-rank constraints exactly. In some problems that require the rank of the approximation matrix to be as small as possible, we can loosely set r to a sufficiently small value according to the practical problems. However, in some problems with strict requirements on the rank, for example, in the sensor localization problem, the rank of the approximation matrix is required to be strictly less than or equal to 3 since the matrix is composed of the position coordinates of the sensors. In these problems, the constraint $\operatorname{rank}(\mathbf{X}) \leq r$ needs to be exactly satisfied.

Inspired by the relation

$$\operatorname{rank}(\mathbf{X}) \leq r \Leftrightarrow \sum_{i=r+1}^{\min\{m,n\}} \sigma_i(\mathbf{X}) = 0, \tag{3}$$

some authors propose to realize the low-rank constraint by minimizing the sum of specific functions with the last $\min\{m,n\} - r$ smallest singular values. A penalty term $\sum_{i=r+1}^{\min\{m,n\}} \sigma_i(\mathbf{X})$ has been used by Gao and Sun (2010) for solving a nearest low-rank correlation matrix problem. In addition, the penalty term $\sum_{i=r+1}^{\min\{m,n\}} \sigma_i^p(\mathbf{X})\,(0<p<1)$ has been discussed by Liu *et al.* (2020) for solving a semidefinite-box constrained low-rank matrix optimization problems. Unfortunately, the optimization process for solving these models includes singular value decomposition (SVD), which always involves great computational cost. Additionally, the computation procedure for solving these model, especially in the fractional order norm case, is also very complicated, and, therefore, unsuitable for large-scale problems.

We propose the following penalty model for the problem (1):

$$\min_{\mathbf{X}\in\bar{\mathcal{C}}} F_\mu(\mathbf{X}) := f(\mathbf{X}) + \mu R_g(\mathbf{X}), \tag{4}$$

where the penalty function

$$R_g(\mathbf{X}) = \inf\Bigg\{\sum_{r+1}^{d} g(|\lambda_i|) : \mathbf{X} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i {\mathbf{v}_i}^T,\ \|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1\Bigg\}.$$

The feasible region $\bar{\mathcal{C}}$ denotes the closure of $\mathcal{C} = \{\mathbf{X}\in\mathbb{R}^{m\times n} |\ \mathbf{X}^{\mathrm{T}}\mathbf{X} \prec \mathbf{I}\}$ which is a nonempty, convex and open set in $\mathbb{R}^{m\times n}$. We prove that the penalty term $R_g(\mathbf{X})$ can be expressed as $R_g(\mathbf{X}) = \sum_{i=r+1}^{d} g[\sigma_i(\mathbf{X})]$ under some conditions on function g. Therefore, it generalizes the penalty functions used by Gao and Sun (2010) and Liu *et al.* (2020), just like the way the penalty function (2) generalizes the rank function, the nuclear norm and Schatten-p norm, etc. Then the problem (4) can be reformulated as

$$\min_{\mathbf{X}\in\bar{\mathcal{C}}} F_\mu^1(\mathbf{X}) := f(\mathbf{X}) + \mu \sum_{i=r+1}^{d} g[\sigma_i(\mathbf{X})]. \tag{5}$$

We further prove that problem (5) is an exact penalty reformulation for problem (1) in terms of global solutions. In addition, a flexible bilinear factorization formation for the proposed penalty problem (4) is constructed with fast computation in an SVD-free manner as follows:

$$\begin{aligned}\min_{\mathbf{U}\mathbf{V}^{\mathrm{T}}\in\bar{\mathcal{C}}} & F_\mu^2(\mathbf{U},\mathbf{V}) \\ &= f\left(\mathbf{U}\mathbf{V}^{\mathrm{T}}\right) \\ &\quad + \mu \sum_{i=r+1}^{d} \Big[g_1\left(\|\mathbf{U}_{:,i}\|_2\right) + g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\Big],\end{aligned} \tag{6}$$

where matrix $\mathbf{X}$ is decomposed as $\mathbf{X} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ with $\mathbf{U}$ and $\mathbf{V}$ as two factors. $\mathbf{U}_{:,i}$ and $\mathbf{V}_{:,i}$ are the i-th columns of the matrices $\mathbf{U}$ and $\mathbf{V}$, respectively; g_1 and g_2 are two functions satisfying some conditions associated with g.

In recent years, first-order algorithms have become an important tool for solving large-scale optimization problems, especially when low to medium accuracy is sufficient (Sulaiman *et al.*, 2024). However, most first-order algorithms generally assume that the objective function has a global Lipschitz continuous gradient, which is a very strict condition hindering its application in areas where the assumptions do not hold or are not reasonable by practical considerations (Yang *et al.*, 2023). There are many application problems whose objective functions do not have this property, such as quadratic inverse problems, D-optimal experimental design (Atwood, 1969), and Poisson inverse problems (Bertero *et al.*, 2009). Bauschke *et al.* (2016) deal with nonglobal Lipschitz continuous gradients by replacing the usual quadratic upper bound functions of the gradient Lipschitz functions with the more general Bregman measure. The corresponding BPG (Bregman proximal gradient) method was proposed by Bauschke *et al.* (2016) with guaranteed complexity and global convergence properties for convex composite optimization problem. In the work of Lu *et al.* (2015b), a similar idea was independently proposed for the convergence of the BPG algorithm for solving convex combination problems in Banach spaces. Bolte *et al.* (2018) extend the BPG to the nonconvex case, that is, minimizes the sum of an extended valued function and a C^1 function.

Instead of the commonly used alternating minimization method for optimizing the factorization format problem, a direct minimization method, namely the BPG method developed by Bolte *et al.* (2018), is used in this paper for the minimization problem (6). The direct method is far more efficient since it only uses a single update for $\mathbf{U}$ and $\mathbf{V}$, rather than several updates involved by alternating minimization within each main iteration (Alain, 2013). Here, neither global gradient Lipschitz nor convexity is needed to be satisfied by the objective function f.

The main contributions of this paper are summarized as follows:

(i) We propose a novel penalty function $R_g(\mathbf{X})$ for the rank constraint of problem (1), and prove that the penalty term can be expressed as $R_g(\mathbf{X}) = \sum_{i=r+1}^{d} g[\sigma_i(\mathbf{X})]$ under some conditions for function g (stated by Theorem 1). We further prove that the global optimal solution set of the penalty problem (5) is the same as that of the original problem (1) (stated by Theorem 2). The proposed penalty function generalizes several existing exact penalty functions induced by relation (3).

(ii) We construct a flexible bilinear factorization format for the proposed penalty function (presented in Theorem 3), which enables the penalty problem to be fast solved in a simple and SVD-free manner. We also prove that the solution set of the factorization formation (6) is the same as that of the proposed penalty problem (4) (stated by Theorem 4).

(iii) A BPG method is developed for directly optimizing the problem (6). This method does not require global gradient Lipschitz and convexity of the objective function. Therefore, it can be flexibly used under very mild conditions.

The rest of this paper is organized as follows. In Section 2, the equivalence between (4) and (5) is declared under some conditions for function g. We further prove the equivalence between the optimal solution sets of (5) and (1), (4) and (6). In Section 3, we present a convergent BPG algorithm for solving problem (6). In Section 4, we take two applications as examples to illustrate that the proposed model can cover a variety of applications. Numerical experiments are presented in Section 5 to further testify the effectiveness of the novel method. Section 6 concludes the whole work. To make the paper self-contained, we provide appendices which include the material related to the convergence of the BPG algorithm and a general list of notations.

2. Proposed penalty model

In this section, we first design a novel penalty term $R_g(\mathbf{X})$, and prove that it can be expressed as $R_g(\mathbf{X}) = \sum_{i=r+1}^{d} g[\sigma_i(\mathbf{X})]$, thus converting the original problem into an exact penalty minimization problem. Furthermore, we transform the penalty function into a bilinear factorization format. Finally, we prove that these problems are equivalent in terms of the global minimizers.

2.1. Proposed penalty function and equivalence between (4) and (5).

Remark 1. Assume that there exists a decomposition $\mathbf{X} = \sum_{i=1}^{d} \lambda_i^* \mathbf{u}_i^* (\mathbf{v}_i^*)^{\mathrm{T}}$ with $\|\mathbf{u}_i^*\|_2 = 1$ and $\|\mathbf{v}_i^*\|_2 = 1$ satisfying $\sum_{i=1}^{d} g(|\lambda_i^*|) = G_g(\mathbf{X})$ as defined by (2). Then we can verify that $\sum_{i=r+1}^{d} g(|\lambda_i^*|)$ achieves the infimum of $\sum_{i=r+1}^{d} g(|\lambda_i|)$ among all the decompositions of $\mathbf{X}$, i.e.,

$$\sum_{i=r+1}^{d} g(|\lambda_i^*|) = R_g(\mathbf{X}) = \inf\Big\{ \sum_{i=r+1}^{d} g(|\lambda_i|) : \mathbf{X} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}, \|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1 \Big\}.$$

Indeed, if there exists another decomposition $\mathbf{X} = \sum_{i=1}^{d} \lambda_i^{**} \mathbf{u}_i^{**} (\mathbf{v}_i^{**})^{\mathrm{T}}$ with $\|\mathbf{u}_i^{**}\|_2 = 1$ and $\|\mathbf{v}_i^{**}\|_2 = 1$, such that

$$\sum_{i=r+1}^{d} g(|\lambda_i^{**}|) = \inf\Big\{ \sum_{i=r+1}^{d} g(|\lambda_i|) : \mathbf{X} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}, \|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1 \Big\},$$

then

$$\sum_{i=r+1}^{d} g(|\lambda_i^{**}|) \le \sum_{i=r+1}^{d} g(|\lambda_i^*|).$$

By taking the decomposition

$$\mathbf{X} = \sum_{i=1}^{r} \lambda_i^* \mathbf{u}_i^* (\mathbf{v}_i^*)^{\mathrm{T}} + \sum_{i=r+1}^{d} \lambda_i^{**} \mathbf{u}_i^{**} (\mathbf{v}_i^{**})^{\mathrm{T}},$$

we immediately have

$$\begin{aligned} \sum_{i=1}^{r} g(|\lambda_i^*|) + \sum_{i=r+1}^{d} g(|\lambda_i^{**}|) &\le \sum_{i=1}^{r} g(|\lambda_i^*|) + \sum_{i=r+1}^{d} g(|\lambda_i^*|) \\ &= \sum_{i=1}^{d} g(|\lambda_i^*|). \end{aligned} \tag{7}$$

On the other hand,

$$\sum_{i=1}^{r} g(|\lambda_i^*|) + \sum_{i=r+1}^{d} g(|\lambda_i^{**}|) \ge \sum_{i=1}^{d} g(|\lambda_i^*|)$$

since $\sum_{i=1}^{d} g(|\lambda_i^*|)$ achieves the infimum of $\sum_{i=1}^{d} g(|\lambda_i|)$ among all the decompositions of $\mathbf{X}$. Therefore, we have

$$\sum_{i=1}^{r} g(|\lambda_i^*|) + \sum_{i=r+1}^{d} g(|\lambda_i^{**}|) = \sum_{i=1}^{d} g(|\lambda_i^*|),$$

which implies

$$\sum_{i=r+1}^{d} g(|\lambda_i^{**}|) = \sum_{i=r+1}^{d} g(|\lambda_i^*|),$$

that is to say, $\sum_{i=r+1}^{d} g(|\lambda_i^*|)$ achieves the infimum of $\sum_{i=r+1}^{d} g(|\lambda_i|)$ among all the decompositions of $\mathbf{X}$.

Similarly, we can conclude that $\sum_{i=1}^{r} g(|\lambda_i^*|)$ achieves the infimum of $\sum_{i=1}^{r} g(|\lambda_i|)$ among all the decompositions of $\mathbf{X}$.

In the work of Jia *et al.* (2020), it is proposed that under some conditions on the function g, $G_g(\mathbf{X})$ can be expressed as

$$G_g(\mathbf{X}) = \sum_{i=1}^{d} g(|\lambda_i^*|) = \sum_{i=1}^{d} g[\sigma_i(\mathbf{X})] \tag{8}$$

if $(\lambda_i^*)_{i=1}^d$ achieves the minimum of $\sum_{i=1}^d g(|\lambda_i|)$ among all the decompositions of $\mathbf{X}$. We shall prove

$$\sum_{i=1}^{r} g(|\lambda_i^*|) = \sum_{i=1}^{r} g[\sigma_i(\mathbf{X})]. \tag{9}$$

Theorem 1 states that under some conditions, this conclusion is true. Before giving Theorem 1, we list the relevant definition and all the lemmas used to prove the theorem.

Definition 1. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. We say that $\mathbf{x}$ is *majorized* by $\mathbf{y}$, denoted by $\mathbf{x} \prec \mathbf{y}$, if for $1 \le k < n$ the following holds:

$$\sum_{j=1}^{k} \mathbf{x}_j^{\downarrow} \le \sum_{j=1}^{k} \mathbf{y}_j^{\downarrow}, \tag{10a}$$

$$\sum_{j=1}^{n} \mathbf{x}_j^{\downarrow} = \sum_{j=1}^{n} \mathbf{y}_j^{\downarrow}, \tag{10b}$$

where $\mathbf{x}^{\downarrow}$ is the vector obtained by rearranging the coordinates of $\mathbf{x}$ in descending order.

Lemma 1. (Schur's theorem (cf. Bhatia, 2011)) *Let $\mathbf{A}$ be an $n \times n$ Hermitian matrix. Let $\operatorname{diag}(\mathbf{A})$ denote the vector whose coordinates are the diagonal entries of $\mathbf{A}$, and let $\lambda(\mathbf{A})$ denote the vector whose coordinates are the eigenvalues of $\mathbf{A}$ specified in any order. Then*

$$\operatorname{diag}(\mathbf{A}) \prec \lambda(\mathbf{A}). \tag{11}$$

Lemma 2. (Horn and Johnson, 1990, Thm. II.3.1) *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. The following two conditions are equivalent:*

(i) $\mathbf{x} \prec \mathbf{y}$,

(ii) $\sum_{i=1}^{n} \phi(x_i) \le \sum_{i=1}^{n} \phi(y_i)$ *for all convex functions ϕ from $\mathbb{R}$ to $\mathbb{R}$.*

Lemma 3. (Horn and Johnson, 1990, Thm.3.3.14(c)) *Let $\mathbf{A} \in \mathbb{R}^{m\times l}$ and $\mathbf{B} \in \mathbb{R}^{n\times l}$ be given. Then, for any real-valued function f such that $\phi(t) = f(e^t)$ is increasing and convex, we have*

$$\sum_{i=1}^{k} f\left(\sigma_i\left(\mathbf{A}\mathbf{B}^{\mathrm{T}}\right)\right) \le \sum_{i=1}^{k} f\left(\sigma_i(\mathbf{A})\,\sigma_i\left(\mathbf{B}^{\mathrm{T}}\right)\right), \tag{12}$$

where $1 \le k \le q$ and $q = \min\{m, n, l\}$.

Theorem 1. *Given a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, if there exists a decomposition $\mathbf{X} = \sum_{i=1}^d \lambda_i^* \mathbf{u}_i^* (\mathbf{v}_i^*)^T$ with $\|\mathbf{u}_i^*\|_2 = 1$ and $\|\mathbf{v}_i^*\|_2 = 1$ satisfying $\sum_{i=1}^d g(|\lambda_i^*|) = G_g(\mathbf{X})$, then (9) holds if the bounded function g satisfies the following conditions:*

(i) g is concave and monotonically ascending in $(0, +\infty)$, and $g(0) = 0$,

(ii) the function $\vartheta(t) \equiv g(e^t)$ is convex.

What is more, by subtracting (9) from (8), we get

$$\sum_{i=r+1}^{d} g(|\lambda_i^*|) = \sum_{i=r+1}^{d} g[\sigma_i(\mathbf{X})]. \tag{13}$$

Proof. Given a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, consider its decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} = \hat{\mathbf{U}}\,|\mathbf{\Sigma}|\,\hat{\mathbf{V}}^{\mathrm{T}}$, where $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with elements $\lambda_1, \lambda_2, \cdots, \lambda_d$ such that $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_d|$. The matrix $\mathbf{U}, \hat{\mathbf{U}} \in \mathbb{R}^{m\times d}$ and $\mathbf{V}, \hat{\mathbf{V}} \in R^{n\times d}$ are of unit l_2 column norm.

Let $\mathbf{\Lambda} \in \mathbb{R}^{d\times d}$ be a diagonal matrix with elements $\beta_1, \beta_2, \ldots, \beta_d$, satisfying $\beta_i = |\lambda_i|$ as $1 \le i \le r$, and $\beta_i = 0$ as $r+1 \le i \le d$. Let $\mathbf{A} = \hat{\mathbf{U}}\mathbf{\Lambda}^{(1/2\,)}$, and $\mathbf{B} = \mathbf{\Lambda}^{(1/2\,)}\hat{\mathbf{V}}^{\mathrm{T}}$. Write $\bar{\mathbf{A}} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\bar{\mathbf{B}} = \mathbf{B}\mathbf{B}^{\mathrm{T}}$. From Lemma 1, we have

$$\begin{aligned} \operatorname{diag}(\mathbf{\Lambda}) &= \operatorname{diag}\left(\bar{\mathbf{A}}\right) \prec \sigma\left(\bar{\mathbf{A}}\right) = \sigma^2(\mathbf{A}), \\ \operatorname{diag}(\mathbf{\Lambda}) &= \operatorname{diag}\left(\bar{\mathbf{B}}\right) \prec \sigma\left(\bar{\mathbf{B}}\right) = \sigma^2(\mathbf{B}). \end{aligned}$$

Let $\sigma(\cdot)$ denote the singular value vector in descending order. Since $\beta_1 \ge \beta_2 \ge \cdots \ge \beta_d$, we have $\operatorname{diag}(\mathbf{\Lambda}) \prec \left(\sigma^2(\mathbf{B}) + \sigma^2(\mathbf{A})\right)/2$. For a concave function g, we have

$$\begin{aligned} \sum_{i=1}^{r} g(|\lambda_i|) &= \sum_{i=1}^{d} g(\beta_i) \\ &\ge \sum_{i=1}^{d} g\left(\frac{\sigma_i^2(\mathbf{A}) + \sigma_i^2(\mathbf{B})}{2}\right) \\ &\ge \sum_{i=1}^{d} g(\sigma_i(\mathbf{A})\,\sigma_i(\mathbf{B})) \\ &\ge \sum_{i=1}^{d} g(\sigma_i(\mathbf{A}\mathbf{B})) \\ &= \sum_{i=1}^{d} g\left(\sigma_i\left(\hat{\mathbf{U}}\mathbf{\Lambda}\hat{\mathbf{V}}^{\mathrm{T}}\right)\right) \\ &= \sum_{i=1}^{r} g(\sigma_i(\mathbf{X})). \end{aligned}$$

The first inequality results from Lemma 2 since $-g$ is a convex function. The second inequality holds because of the fact that the function g is increasing, and $a^2 + b^2 \geq 2ab$. The third inequality stems from Lemma 3. All the inequalities become equalities if and only if $\mathbf{X} = \mathbf{U\Sigma V}^{\mathrm{T}}$ is the SVD. Since $\sum_{i=1}^{r} g\left(|\lambda_i^*|\right)$ achieves the minimum of $\sum_{i=1}^{r} g\left(|\lambda_i|\right)$ among all the decompositions of $\mathbf{X}$ due to Remark 1, we conclude that

$$\sum_{i=1}^{r} g\left(|\lambda_i^*|\right) = \sum_{i=1}^{r} g\left[\sigma_i\left(\mathbf{X}\right)\right].$$

The proof is completed. ■

Combining Eqn. (13) with Remark 1, we conclude that the proposed penalty function can be expressed as a sum of specific functions with the $d-r$ smallest singular values of matrix $\mathbf{X}$, i.e.,

$$R_g\left(\mathbf{X}\right) = \sum_{i=r+1}^{d} g\left[\sigma_i\left(\mathbf{X}\right)\right]. \tag{14}$$

Hence we deduce that the penalty term $R_g\left(\mathbf{X}\right)$ proposed here is a generalization of those presented by Gao and Sun (2010) and Liu *et al.* (2020) when g is taken as $g\left(x\right) = x$ and $g\left(x\right) = x^p\left(0 < p < 1\right)$, respectively. The equivalence between the penalty models (4) and (5) is verified by (14).

It should be noted that the conditions of Theorem 1 can be easily satisfied by many functions. For example, the functions $g\left(x\right) = x$, $g\left(x\right) = x^0$ and the widely used nonconvex functions $g\left(x\right) = x^p\left(0 < p < 1\right)$ all satisfy these conditions. The nonconvex function $g\left(x\right) = \log x$ meets these conditions if we revise its value at zero as $g\left(0\right) = 0$.

2.2. Equivalence between (5) and (1).

In this section, we build a relationship between the penalty model (5) and the original problem (1). The following theorem shows that model (5) is an exact penalty reformulation of problem (1) in terms of global minimizers.

Lemma 4. *Let $\mathbf{X}_\Omega$ be a projection of $\mathbf{X} \in \bar{\mathcal{C}}$ ($\bar{\mathcal{C}} = \left\{\mathbf{X} \in \mathbb{R}^{m\times n} | \mathbf{X}^{\mathrm{T}}\mathbf{X} \preceq \mathbf{I}\right\}$) onto Ω, and let g satisfy $g\left(x\right) \geq x$ for any $x \in [0,1]$. Then*

$$\|\mathbf{X} - \mathbf{X}_\Omega\|_{\mathrm{F}} \leq \sum_{i=r+1}^{d} g\left(\sigma_i\left(\mathbf{X}\right)\right). \tag{15}$$

Proof. Using Proposition 2.6 of Lu *et al.* (2017), it is not hard to prove that

$$\|\mathbf{X} - \mathbf{X}_\Omega\|_{\mathrm{F}} = \sqrt{\sum_{i=r+1}^{d} \sigma_i^2\left(\mathbf{X}\right)}, \quad \forall\, \mathbf{X} \in \bar{\mathcal{C}}.$$

Notice from $\mathbf{X} \in \bar{\mathcal{C}}$ that $0 \leq \sigma_i\left(\mathbf{X}\right) \leq 1$ holds for all $1 \leq i \leq d$. In view of the fact $g\left(x\right) \geq x$ for any $x \in [0,1]$, one can observe that

$$\begin{aligned}\sqrt{\sum_{i=r+1}^{d} \sigma_i^2\left(\mathbf{X}\right)} &\leq \sum_{i=r+1}^{d} \sigma_i\left(\mathbf{X}\right)\\ &\leq \sum_{i=r+1}^{d} g\left(\sigma_i\left(\mathbf{X}\right)\right), \quad \forall\, \mathbf{X} \in \bar{\mathcal{C}}.\end{aligned}$$

It then follows that (15) holds as desired. This completes the proof. ■

Theorem 2. *If g satisfies $g\left(x\right) \geq x$ for any $x \in [0,1]$, then for any $\mu > L_f$, where L_f is the Lipschitz constant of f, problems (1) and (5) have the same global minimizers.*

Proof. Recall that f is assumed to be continuously differentiable in $\bar{\mathcal{C}}$. It follows that f is Lipschitz continuous in $\bar{\mathcal{C}}$, that is, there exists some constant $L_f > 0$ such that

$$|f\left(\mathbf{X}\right) - f\left(\mathbf{Y}\right)| \leq L_f \|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}.$$

For the first part, let $\hat{\mathbf{X}}$ be a global minimizer of problem (1) and $\mathbf{X}$ be an arbitrary matrix in $\bar{\mathcal{C}}$. We let $\mathbf{X}_\Omega$ denote a projection of $\mathbf{X}$ onto Ω. Thus, we know from the global optimality of $\hat{\mathbf{X}}$ that $f\left(\mathbf{X}_\Omega\right) \geq f\left(\hat{\mathbf{X}}\right)$. Using this relation and the Lipschitz continuity of f, we have

$$\begin{aligned} f(\hat{\mathbf{X}}) - f(\mathbf{X}) &= f(\hat{\mathbf{X}}) - f\left(\mathbf{X}_\Omega\right)\\ &\quad + f\left(\mathbf{X}_\Omega\right) - f\left(\mathbf{X}\right)\\ &\leq f\left(\mathbf{X}_\Omega\right) - f\left(\mathbf{X}\right)\\ &\leq L_f\|\mathbf{X} - \mathbf{X}_\Omega\|_{\mathrm{F}}.\end{aligned}$$

This, together with Lemma 4, $\mu \geq L_f$, and rank$(\hat{\mathbf{X}}) \leq r$, implies that

$$\begin{aligned} F_\mu^1\left(\mathbf{X}\right) &= f\left(\mathbf{X}\right) + \mu \sum_{i=r+1}^{d} g\left(\sigma_i\left(\mathbf{X}\right)\right)\\ &\geq f\left(\mathbf{X}\right) + L_f\|\mathbf{X} - \mathbf{X}_\Omega\|_{\mathrm{F}}\\ &\geq f(\hat{\mathbf{X}})\\ &= f(\hat{\mathbf{X}}) + \mu \sum_{i=r+1}^{d} g(\sigma_i(\hat{\mathbf{X}}))\\ &= F_\mu^1(\hat{\mathbf{X}}),\end{aligned}$$

which together with the arbitrariness of $\mathbf{X} \in \bar{\mathcal{C}}$ and $\hat{\mathbf{X}} \in \bar{\mathcal{C}}$ implies that $\hat{\mathbf{X}}$ is a global minimizer of problem (5).

For the second part, assume that $\mu > L_f$. Let $\hat{\mathbf{X}}$ be a global minimizer of problem (5) and $\hat{\mathbf{X}}_\Omega$ be a projection of $\hat{\mathbf{X}}$ onto Ω. It is easy to observe that if $\hat{\mathbf{X}} \in \Omega$, then it is a global minimizer of problem (1). Thus, it suffices to prove that $\hat{\mathbf{X}} \in \Omega$. Suppose for contradiction that $\hat{\mathbf{X}} \notin \Omega$. Then we have $\left\|\hat{\mathbf{X}} - \hat{\mathbf{X}}_{\mathbf{\Omega}}\right\| > 0$, and hence

$$\begin{aligned} f(\hat{\mathbf{X}}_{\mathbf{\Omega}}) &\le f(\hat{\mathbf{X}}) + L_f \left\|\hat{\mathbf{X}} - \hat{\mathbf{X}}_{\mathbf{\Omega}}\right\|_{\mathrm{F}} \\ &< f(\hat{\mathbf{X}}) + \mu \left\|\hat{\mathbf{X}} - \hat{\mathbf{X}}_{\mathbf{\Omega}}\right\|_{\mathrm{F}} \\ &\le f(\hat{\mathbf{X}}) + \mu \sum_{i=r+1}^{d} g(\sigma_i(\hat{\mathbf{X}})) \\ &\le f(\hat{\mathbf{X}}_{\mathbf{\Omega}}) + \mu \sum_{i=r+1}^{d} g(\sigma_i(\hat{\mathbf{X}}_{\mathbf{\Omega}})) \\ &= f(\hat{\mathbf{X}}_{\mathbf{\Omega}}). \end{aligned}$$

The first inequality follows from the Lipschitz continuity of f. The second inequality is due to $\mu > L_f$. The third inequality is due to Lemma 4. Then the last inequality follows from the global optimality of $\hat{\mathbf{X}}$ in problem (5). These inequalities immediately lead to a contradiction $f(\hat{\mathbf{X}}_{\mathbf{\Omega}}) < f(\hat{\mathbf{X}}_{\mathbf{\Omega}})$. This completes the proof. ■

The condition of Theorem 2, that is, $g(x) \ge x$ for any $x \in [0,1]$, can be satisfied by functions $g(x) = x^0$ and $g(x) = x^p$ $(0 < p \le 1)$. The function $g(x) = \log x$ does not meet this condition.

2.3. Equivalence between (4) and (6).

We further prove that the penalty function $R_g(\mathbf{X})$ can be expressed in a bilinear factorization form, which enables the penalty problem (4) being efficiently solved without using SVD.

Theorem 3. *If there exist functions g_1 and g_2 such that*

$$g(x) = \min_{\substack{x=ab\\ a,b\ge 0}} g_1(a) + g_2(b), \quad x \in [0, +\infty), \tag{16}$$

then $R_g(\mathbf{X})$ can be represented as

$$\begin{aligned} &R_g(\mathbf{X}) \\ &= \min_{\mathbf{X}=\mathbf{U}\mathbf{V}^{\mathrm{T}}} \sum_{i=r+1}^{d} \left[g_1\left(\|\mathbf{U}_{:,i}\|_2\right) + g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\right]. \end{aligned} \tag{17}$$

Proof. Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i {\mathbf{v}_i}^{\mathrm{T}}$ be a decomposition of $\mathbf{X}$, satisfying $\|\mathbf{u}_i\|_2 = 1$, $\|\mathbf{v}_i\|_2 = 1$. According to (16), for $i = r+1, \dots, d$, we have

$$\mathrm{g}(|\lambda_i|) = \min_{\substack{|\lambda_i|=\lambda_i^1\lambda_i^2\\ \lambda_i^1,\lambda_i^2\ge 0}} g_1(\lambda_i^1) + g_2\left(\lambda_i^2\right).$$

For any decomposition $\lambda_i = \lambda_i^1\lambda_i^2$ of each λ_i, $\mathbf{X}$ can be written as

$$\begin{aligned} \mathbf{X} &= \sum_{i=1}^{d} \lambda_i \mathbf{u}_i {\mathbf{v}_i}^{\mathrm{T}} \\ &= \sum_{i=1}^{d} \left(\lambda_i^1 \mathbf{u}_i\right)\left(\lambda_i^2 \mathbf{v}_i\right)^{\mathrm{T}} \\ &= \sum_{i=1}^{d} \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i^{\mathrm{T}} = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^{\mathrm{T}}, \end{aligned}$$

where $\|\tilde{\mathbf{u}}_i\|_2 = \left|\lambda_i^1\right|$, $\|\tilde{\mathbf{v}}_i\|_2 = \left|\lambda_i^2\right|$, and $\tilde{\mathbf{U}}$, $\tilde{\mathbf{V}}$ are two matrices composed of columns $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$, respectively. Therefore,

$$\mathrm{g}(|\lambda_i|) = \min_{|\lambda_i|=\|\tilde{\mathbf{u}}_i\|_2\|\tilde{\mathbf{v}}_i\|_2} g_1(\|\tilde{\mathbf{u}}_i\|_2) + g_2\left(\|\tilde{\mathbf{v}}_i\|_2\right),$$

$i = r+1, \dots, d$. Adding these from $i = r+1$ to $i = d$, we immediately get

$$\begin{aligned} &\sum_{i=r+1}^{d} g\left(|\lambda_i|\right) \\ &= \sum_{i=r+1}^{d} \min_{|\lambda_i|=\|\tilde{\mathbf{u}}_i\|_2\|\tilde{\mathbf{v}}_i\|_2} [g_1(\|\tilde{\mathbf{u}}_i\|_2) + g_2\left(\|\tilde{\mathbf{v}}_i\|_2\right)]. \end{aligned}$$

Taking minima for all the decompositions of $\mathbf{X}$, we then have

$$\begin{aligned} &\inf_{\lambda=(\lambda_1,\dots,\lambda_d)} \sum_{i=r+1}^{d} g\left(|\lambda_i|\right) \\ &= \min_{\mathbf{X}=\mathbf{U}\mathbf{V}^{\mathrm{T}}} \sum_{i=r+1}^{d} \left[g_1(\left\|\mathbf{U}_{:,i}\right\|_2) + g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\right]. \end{aligned}$$

The proof is completed. ■

Many widely used functions, such as $g(x) = x^0$, $g(x) = x^p$ $(0 < p \le 1)$ and $g(x) = \log x$, all have the factorization expression shown in (16), for example,

$$\begin{aligned} x^0 &= \min_{\substack{x=ab\\ a,b\ge 0}} a^0 + b^0, \\ x^p &= \min_{\substack{x=ab\\ a,b\ge 0}} (p/p_1)\, a^{p_1} + (p/p_2)\, b^{p_2}, \\ \log x &= \min_{\substack{x=ab\\ a,b>0}} \log a + \log b. \end{aligned}$$

Under the conditions of Theorem 3, the penalty model (4) becomes

$$\begin{aligned} &\min_{\mathbf{X}\in\bar{\mathcal{C}}} F_\mu(\mathbf{X}) \\ &\quad = f(\mathbf{X}) + \mu \min_{\mathbf{X}=\mathbf{U}\mathbf{V}^{\mathrm{T}}} \sum_{i=r+1}^{d} \left[g_1\left(\|\mathbf{U}_{:,i}\|_2\right)\right. \\ &\qquad \left. + g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\right]. \end{aligned} \tag{18}$$

We further relax it to a bilinear optimization problem as follows:

$$\begin{aligned}\min_{\mathbf{U}\mathbf{V}^{\mathrm{T}}\in\bar{\mathcal{C}}} & F_\mu^2\left(\mathbf{U},\mathbf{V}\right)\\ &= f\left(\mathbf{U}\mathbf{V}^{\mathrm{T}}\right)+\mu\sum_{i=r+1}^{d}\left[g_1\left(\|\mathbf{U}_{:,i}\|_2\right)\right.\\ &\quad\left.+g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\right],\end{aligned}$$

which is exactly the problem (6). The following theorem shows that the solution sets of problems (4) and (6) are equivalent and the optimal objective function values are the same.

Theorem 4. *Suppose that $\hat{\mathbf{X}}$ is a solution to problem (4), in which function g satisfies the conditions of Theorem 3. Then there exists a decomposition $\hat{\mathbf{X}}=\tilde{\mathbf{U}}\tilde{\mathbf{V}}^T$ such that $(\tilde{\mathbf{U}},\ \tilde{\mathbf{V}})$ is the solution of (6). Let $(\hat{\mathbf{U}},\ \hat{\mathbf{V}})$ be a solution to problem (6). Then $\bar{\mathbf{X}}=\hat{\mathbf{U}}(\hat{\mathbf{V}})^T$ is also a solution to (4). Furthermore, we have that $F_\mu(\hat{\mathbf{X}})=F_\mu^2(\hat{\mathbf{U}},\hat{\mathbf{V}})$.*

Proof. On account of Theorem 3, assume that there exists a decomposition $\hat{\mathbf{X}}=\tilde{\mathbf{U}}\tilde{\mathbf{V}}^{\mathrm{T}}$ such that

$$\begin{aligned}R_g&(\hat{\mathbf{X}})\\ &=\sum_{i=r+1}^{d}\left[g_1(\|\tilde{\mathbf{U}}_{:,i}\|_2)+g_2(\|\tilde{\mathbf{V}}_{:,i}\|_2)\right]\\ &=\min_{\hat{\mathbf{X}}=\mathbf{U}\mathbf{V}^{\mathrm{T}}}\sum_{i=r+1}^{d}\left[g_1\left(\|\mathbf{U}_{:,i}\|_2\right)+g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\right].\end{aligned}$$

Then

$$\begin{aligned}F_\mu(\hat{\mathbf{X}})&=f(\hat{\mathbf{X}})+\mu R_g(\hat{\mathbf{X}})\\ &=f(\tilde{\mathbf{U}}\tilde{\mathbf{V}}^{\mathrm{T}})\\ &\quad+\mu\sum_{i=r+1}^{d}\left[g_1(\|\tilde{\mathbf{U}}_{:,i}\|_2)+g_2(\|\tilde{\mathbf{V}}_{:,i}\|_2)\right]\\ &=F_\mu^2(\tilde{\mathbf{U}},\tilde{\mathbf{V}})\geq F_\mu^2(\hat{\mathbf{U}},\hat{\mathbf{V}}).\end{aligned}$$

Write $\bar{\mathbf{X}}=\hat{\mathbf{U}}\left(\hat{\mathbf{V}}\right)^{\mathrm{T}}$. According to Theorems 3 again, we have

$$\begin{aligned}F_\mu^2&(\hat{\mathbf{U}},\hat{\mathbf{V}})\\ &=f(\hat{\mathbf{U}}(\hat{\mathbf{V}})^{\mathrm{T}})\\ &\quad+\mu\sum_{i=r+1}^{d}\left[g_1(\|\hat{\mathbf{U}}_{:,i}\|_2)+g_2(\|\hat{\mathbf{V}}_{:,i}\|_2)\right]\\ &\geq f(\hat{\mathbf{U}}(\hat{\mathbf{V}})^{\mathrm{T}})\\ &\quad+\mu\min_{\bar{\mathbf{X}}=\mathbf{U}\mathbf{V}^{\mathrm{T}}}\sum_{i=r+1}^{d}\left[g_1(\|\mathbf{U}_{:,i}\|_2)+g_2(\|\mathbf{V}_{:,i}\|_2)\right]\\ &=f(\bar{\mathbf{X}})+\mu R_g(\bar{\mathbf{X}})\\ &=F_\mu(\bar{\mathbf{X}})\geq F_\mu(\hat{\mathbf{X}}).\end{aligned}$$

Then we have $F_\mu(\hat{\mathbf{X}})=F_\mu^2(\hat{\mathbf{U}},\hat{\mathbf{V}})$. Consequently, $F_\mu^2(\hat{\mathbf{U}},\hat{\mathbf{V}})=F_\mu^2(\tilde{\mathbf{U}},\tilde{\mathbf{V}})$ and $F_\mu(\hat{\mathbf{X}})=F_\mu(\bar{\mathbf{X}})$ according to the above inequalities. Therefore $(\tilde{\mathbf{U}},\tilde{\mathbf{V}})$ is a solution to problem (6), and $\bar{\mathbf{X}}$ is a solution of problem (4). The proof is completed. ■

Figure 1 visually shows the relationship between the problems mentioned above.

Remark 2. As for the selection of the function $g\left(x\right)\left(x\in[0,\ +\infty)\right)$, the widely used functions, such as $g\left(x\right)=x^0$ and $g\left(x\right)=x^p\left(0<p\leq 1\right)$, all satisfy the conditions in this section. Then the equivalence among models (1), (4), (5) and (6) can be built. For function $g\left(x\right)=\log x$, only the equivalence among models (4), (5), (6) can be built since it does not satisfy $g\left(x\right)\geq x$ for $x\in[0,\ 1]$. We usually choose the functions $g\left(x\right)=x^p\ (0<p\leq 1)$ to solve problem (1) because $g\left(x\right)=x^0$ and the corresponding g_1 and g_2 are discontinuous at zero.

3. Optimization algorithm

To solve problem (6), $\mathbf{U}$ and $\mathbf{V}$ are directly optimized by the BPG method in this section. Let $\mathbf{Z}=\left(\mathbf{U}^{\mathrm{T}},\mathbf{V}^{\mathrm{T}}\right)^{\mathrm{T}}$, $p\left(\mathbf{Z}\right)=\mu\sum_{i=r+1}^{d}\left[g_1\left(\|\mathbf{U}_{:,i}\|_2\right)+g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\right]$ and $f\left(\mathbf{Z}\right)=f\left(\mathbf{U}\mathbf{V}^{\mathrm{T}}\right)$, we consider the following composite optimization problem:

$$\min_{\mathbf{Z}\in\bar{\mathcal{C}}'}F_\mu^2(\mathbf{Z})=f\left(\mathbf{Z}\right)+p\left(\mathbf{Z}\right),\tag{19}$$

where $\bar{\mathcal{C}}'=\left\{\mathbf{Z}=\left(\mathbf{U}^{\mathrm{T}},\mathbf{V}^{\mathrm{T}}\right)^{\mathrm{T}}\middle|\mathbf{U}\mathbf{V}^{\mathrm{T}}\in\bar{\mathcal{C}}\right\}$ and $\bar{\mathcal{C}}=\left\{\mathbf{X}\in\mathbb{R}^{m\times n}\middle|\mathbf{X}^{\mathrm{T}}\mathbf{X}\preceq\mathbf{I}\right\}$. By using the BPG method, we can get a sequence of approximate solutions, that is, for $k=0,1,\ldots,$

$$\begin{aligned}\mathbf{Z}^{k+1}&\in T_\tau\left(\mathbf{Z}^k\right)\\ &=\arg\min_{\mathbf{Y}\in\bar{\mathcal{C}}'}\left\{\left\langle\nabla f\left(\mathbf{Z}^k\right),\mathbf{Y}-\mathbf{Z}^k\right\rangle\right.\\ &\qquad\left.+\tau D_h\left(\mathbf{Y},\mathbf{Z}^k\right)+p\left(\mathbf{Y}\right)\right\},\end{aligned}\tag{20}$$

where $T_\tau\left(\mathbf{Z}\right)$ is the BPG map

$$\begin{aligned}T_\tau\left(\mathbf{Z}\right):=\arg\min_{\mathbf{Y}\in\bar{\mathcal{C}}'}&\left\{\left\langle\nabla f\left(\mathbf{Z}\right),\ \mathbf{Y}-\mathbf{Z}\right\rangle\right.\\ &\left.+\tau D_h\left(\mathbf{Y},\mathbf{Z}\right)+p\left(\mathbf{Y}\right)\right\}.\end{aligned}$$

The proximity measure $D_h:\mathbb{R}^{m\times n}\times\mathbb{R}^{m\times n}\to\mathbb{R}^+$ is the so-called Bregman distance and is defined as

$$D_h\left(\mathbf{Y},\mathbf{Z}\right)=h\left(\mathbf{Y}\right)-h\left(\mathbf{Z}\right)-\left\langle\nabla h\left(\mathbf{Z}\right),\ \mathbf{Y}-\mathbf{Z}\right\rangle,$$

where $h:\mathbb{R}^{m\times n}\to(-\infty,+\infty]$ is a kernel generating distance. For early pivotal results on Bregman distances, associated proximal-based algorithms and a lot of instances of kernels h, we refer the reader to the works

$$\begin{tikzcd}[column sep=huge, row sep=huge]
\begin{array}{c} \min\limits_{\mathrm{X}} f(\mathrm{X}) \\ \text{s.t. } \operatorname{rank}(\mathrm{X}) \leq r, \\ \mathrm{X}^{\mathrm{T}}\mathrm{X} \preceq \mathrm{I} \end{array} \arrow[r, leftrightarrow, line width=2pt, "\begin{array}{c} g(x) \geq x, \\ x \in [0,1] \\ \mu > L_f \end{array}"] & \min\limits_{\mathrm{X}\in\bar{\mathcal{C}}} F_\mu^1(\mathrm{X}) := f(\mathrm{X}) + \mu\sum_{i=r+1}^{d} g\left[\sigma_i(\mathrm{X})\right] \arrow[d, Leftrightarrow, "\begin{array}{l} \text{(1) } g \text{ is concave and monotonically} \\ \text{ascending in } (0,+\infty), \text{ and } g(0)=0. \\ \text{(2) } g(e^t) \text{ is convex.} \end{array}"] \\
\begin{array}{l} \min\limits_{\mathrm{UV}^{\mathrm{T}}\in\bar{\mathcal{C}}} F_\mu^2(\mathrm{U},\mathrm{V}) = f\left(\mathrm{UV}^{\mathrm{T}}\right) \\ +\mu\sum\limits_{i=r+1}^{d}\left[g_1\left(\left\|\mathrm{U}_{:,i}\right\|_2\right) + g_2\left(\left\|\mathrm{V}_{:,i}\right\|_2\right)\right] \end{array} \arrow[r, leftrightarrow, line width=2pt, "g(x) = \min\limits_{\substack{x=ab \\ a,b\geq 0}} g_1(a) + g_2(b)"] & \min\limits_{\mathrm{X}\in\bar{\mathcal{C}}} F_\mu(\mathrm{X}) := f(\mathrm{X}) + \mu R_g(\mathrm{X})
\end{tikzcd}$$

Fig. 1. Relationship between different models. The hollow arrow indicates that the corresponding two problems are equivalent. The black solid arrow indicates that the global minimal solution sets of the two problems are the same. The texts next to the arrows are the required conditions for the equivalence relationship.

of Bauschke *et al.* (2016), Bauschke and Borwein (1997), Censor and Zenios (1992), Chen and Teboulle (1993), Eckstein (1993) and Teboulle (1992). The gradient of f is

$$\nabla f(\mathbf{Z}) = \left((\nabla f_{\mathbf{U}})^{\mathrm{T}},\ (\nabla f_{\mathbf{V}})^{\mathrm{T}}\right)^{\mathrm{T}}.$$

By deleting the irrelated terms with the minimization of $\mathbf{Y}$, and letting $P_\tau\left(\mathbf{Z}^k\right) = \nabla f\left(\mathbf{Z}^k\right) - \tau\nabla h\left(\mathbf{Z}^k\right)$, the iteration (20) can be simplified as

$$\mathbf{Z}^{k+1} \in \arg\min_{\mathbf{Y}\in\bar{\mathcal{C}}'} \left\{\left\langle P_\tau\left(\mathbf{Z}^k\right), \mathbf{Y}\right\rangle + \tau h(\mathbf{Y}) + p(\mathbf{Y})\right\}. \tag{21}$$

Under some wild assumptions, which can be found in Appendix A, and the L-smooth adaptable condition, there exists $L > 0$ such that $Lh - f$ is convex on $\mathcal{C}'$, the convergence of iteration (21) falls into the scope considered by Bolte *et al.* (2018). It is worth mentioning that the conditions associated with h are easily fulfilled. Although a lot of classical f are not global gradient Lipschitz functions, we can always find a h such that these conditions hold (please refer to Bolte *et al.* (2018)). Especially, when ∇f is locally Lipschitz continuous on any bounded subset of $\mathbb{R}^{m\times n}$, the global convergence, meaning that any bounded sequence $\left\{\mathbf{Z}^k\right\}_{k=0}^{\infty}$ generated by the BPG converges to a critical point of $F_\mu^2(\mathbf{Z})$, can be obtained using the KL property (Bolte *et al.*, 2014) of $F_\mu^2(\mathbf{Z})$. When f does not even meet the local gradient Lipschitz condition, the descent of the function values can still be guaranteed. The convergence results and relative materials presented by Bolte *et al.* (2018) can also be found in Appendix A.

4. Applications

This section illustrates the potential applications of our approach. To this end, we consider two classes of application problems. One is the nearest low-rank correlation matrix problem, and the other is the quadratic inverse problem. We shall show that our method is effective for both.

4.1. Nearest low-rank correlation matrix problem.

The nearest low-rank correlation matrix estimation is often used in the fields of financial engineering, data compression, data mining etc. For example, it is used for the estimation of covariance matrices in asset pricing and risk assessment in financial engineering. In the field of data compression, it is also used to estimate covariance matrices in the widely used principal component analysis for dimensionality reduction. In data mining, it is used to estimate the correlation matrix between items or users to eliminate redundancy.

The nearest low-rank correlation problem (Horn and Johnson, 1990; Liu *et al.*, 2020) can be formulated as follows:

$$\min_{\mathbf{X}\in\mathcal{S}^n} \frac{1}{2}\left\|\mathbf{H}\circ(\mathbf{X}-\mathbf{C})\right\|_{\mathrm{F}}^2 \tag{22a}$$

subject to

$$\mathbf{0} \preceq \mathbf{X} \preceq n\mathbf{I}_n, \tag{22b}$$

$$\operatorname{diag}(\mathbf{X}) = \mathbf{e}, \tag{22c}$$

$$\operatorname{rank}(\mathbf{X}) \leq r, \tag{22d}$$

where $\mathbf{H}$ is a given weight matrix belonging to the set $\mathcal{S}^n$ of $n \times n$ symmetric positive semidefinite matrices. $\mathbf{C} \in \mathcal{S}^n$ is a given correlation matrix, and $\mathbf{e}$ is the all ones vector. Upon changing the variable $\mathbf{Y} = \mathbf{X}/n$, we solve problem (22) by the bilinear relaxation model of the proposed penalty method in the form of

$$\min_{\mathbf{UV}^{\mathrm{T}}\in\bar{\mathcal{C}}} F_\mu^2(\mathbf{U},\mathbf{V})$$

$$
\begin{aligned}
&= \frac{1}{2}\left\|\mathbf{H}\circ\left(\mathbf{U}\mathbf{V}^{\mathrm{T}}-\frac{1}{n}\mathbf{C}\right)\right\|_{\mathrm{F}}^{2}\\
&+\frac{\mu_{1,k}}{2}\left\|\operatorname{diag}\left(\mathbf{U}\mathbf{V}^{\mathrm{T}}\right)-\frac{1}{n}\mathbf{e}\right\|_{2}^{2}\\
&+\mu_{2,k}\sum_{i=r+1}^{d}\left[g_1\left(\|\mathbf{U}_{:,i}\|_2\right)+g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\right],
\end{aligned}
\tag{23}
$$

where $\mu_{1,k},\mu_{2,k} > 0$ are penalty parameters for $k = 0,1,\dots$, and $\mathcal{C} = \left\{\mathbf{Y}\in\mathcal{S}^n \middle| \mathbf{Y}^{\mathrm{T}}\mathbf{Y} \prec \mathbf{I}_n\right\}$. This problem coincides with the form of problem (19) with

$$
\begin{aligned}
f(\mathbf{Z}) = {}& \frac{1}{2}\left\|\mathbf{H}\circ\left(\mathbf{U}\mathbf{V}^{\mathrm{T}}-\frac{1}{n}\mathbf{C}\ \right)\right\|_{\mathrm{F}}^{2}\\
&+\frac{\mu_{1,k}}{2}\left\|\operatorname{diag}\left(\mathbf{U}\mathbf{V}^{\mathrm{T}}\right)-\frac{1}{n}\mathbf{e}\ \right\|_{2}^{2}
\end{aligned}
$$

and

$$
p(\mathbf{Z}) = \mu_{2,k}\sum_{i=r+1}^{d}\left[g_1\left(\|\mathbf{U}_{:,i}\|_2\right)+g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\right].
$$

Since $f(\mathbf{Z})$ has a local Lipschitz gradient on the bounded region $\mathcal{C}' = \left\{\mathbf{Z} = \left(\mathbf{U}^{\mathrm{T}},\mathbf{V}^{\mathrm{T}}\right)^{\mathrm{T}} \middle| \mathbf{U}\mathbf{V}^{\mathrm{T}}\in\mathcal{C}\right\}$ (Liu *et al*., 2020), the BPG algorithm is obviously applicable by taking $h(\mathbf{Z}) = \|\mathbf{Z}\|_{\mathrm{F}}^2/2$.

By the first-order stability condition of (21), we obtain

$$
\nabla p\left(\mathbf{Z}^{k+1}\right)+\tau\nabla h\left(\mathbf{Z}^{k+1}\right) = -P_\tau\left(\mathbf{Z}^{k}\right). \tag{24}
$$

We calculate

$$
\begin{aligned}
&P_\tau(\mathbf{Z})\\
&= \nabla f(\mathbf{Z})-\tau\nabla h(\mathbf{Z})\\
&= \begin{pmatrix}
\mathbf{H}\circ\left(\mathbf{U}(\mathbf{V})^{\mathrm{T}}-\dfrac{1}{n}\mathbf{C}\right)\mathbf{V}\\
+\mu_{1,k}\left(\operatorname{Diag}\left(\mathbf{U}(\mathbf{V})^{\mathrm{T}}\right)-\dfrac{1}{n}\mathbf{I}_n\right)\mathbf{V}-\tau\mathbf{U}\\
\mathbf{H}\circ\left(\mathbf{U}(\mathbf{V})^{\mathrm{T}}-\dfrac{1}{n}\mathbf{C}\right)\mathbf{U}\\
+\mu_{1,k}\left(\operatorname{Diag}\left(\mathbf{U}(\mathbf{V})^{\mathrm{T}}\right)-\dfrac{1}{n}\mathbf{I}_n\right)\mathbf{U}-\tau\mathbf{V}
\end{pmatrix}.
\end{aligned}
\tag{25}
$$

Applying (24) and the separability on columns, we have

$$
\begin{cases}
\mathbf{U}_{:,i}^{k+1} = -\dfrac{1}{\tau}\left(P_\tau^{1,k}\right)_{:,i},\\
\mathbf{V}_{:,i}^{k+1} = -\dfrac{1}{\tau}\left(P_\tau^{2,k}\right)_{:,i}
\end{cases}
\tag{26}
$$

for columns $i = 1,2,\dots,r$ and

$$
\begin{cases}
\left[\mu_{2,k}\dfrac{g_1'\left(\left\|\mathbf{U}_{:,i}^{k+1}\right\|_2\right)}{\left\|\mathbf{U}_{:,i}^{k+1}\right\|_2}+\tau\right]\mathbf{U}_{:,i}^{k+1}\\
= -\left(P_\tau^{1,k}\right)_{:,i}\\
\left[\mu_{2,k}\dfrac{g_2'\left(\left\|\mathbf{V}_{:,i}^{k+1}\right\|_2\right)}{\left\|\mathbf{V}_{:,i}^{k+1}\right\|_2}+\tau\right]\mathbf{V}_{:,i}^{k+1}\\
= -\left(P_\tau^{2,k}\right)_{:,i}
\end{cases}
\tag{27}
$$

for columns $i = r+1,\dots,d$. Here $P_\tau^{1,k}$ is the upper block of $P_\tau\left(\mathbf{Z}^k\right)$, and $P_\tau^{2,k}$ is the lower block of $P_\tau\left(\mathbf{Z}^k\right)$. For any column of $i = r+1,\dots,d$, due to the linear correlation between $\left(P_\tau^{1,k}\right)_{:,i}$ and $\mathbf{U}_{:,i}^{k+1}$, $\left(P_\tau^{2,k}\right)_{:,i}$ and $\mathbf{V}_{:,i}^{k+1}$ shown by (27), consider the following two cases:

(i) If $\left(P_\tau^{1,k}\right)_{:,i}=\mathbf{0}$, then $\mathbf{U}_{:,i}^{k+1} = \mathbf{0}$; if $\left(P_\tau^{2,k}\right)_{:,i}=\mathbf{0}$, then $\mathbf{V}_{:,i}^{k+1}=\mathbf{0}$.

(ii) If $\left(P_\tau^{1,k}\right)_{:,i} \neq \mathbf{0}$, then $\mathbf{U}_{:,i}^{k+1} = -{t_i}^*\left(P_\tau^{1,k}\right)_{:,i}$; if $\left(P_\tau^{2,k}\right)_{:,i} \neq \mathbf{0}$, then $\mathbf{V}_{:,i}^{k+1} = -s_i^*\left(P_\tau^{2,k}\right)_{:,i}$, where s_i^*, ${t_i}^*$ is the positive root of the following system of equations:

$$
\begin{cases}
\dfrac{\mu_{2,k}g_1'\left(t_i\left\|\left(P_\tau^{1,k}\right)_{:,i}\right\|_2\right)}{\left\|\left(P_\tau^{1,k}\right)_{:,i}\right\|_2}+\tau t_i-1 = 0\\
\dfrac{\mu_{2,k}g_2'\left(s_i\left\|\left(P_\tau^{2,k}\right)_{:,i}\right\|_2\right)}{\left\|\left(P_\tau^{2,k}\right)_{:,i}\right\|_2}+\tau s_i-1 = 0
\end{cases}
\tag{28}
$$

Let $\hat{\mathbf{Y}} = \hat{\mathbf{U}}(\hat{\mathbf{V}})^{\mathrm{T}}$ with $(\hat{\mathbf{U}},\ \hat{\mathbf{V}})$ being an approximated solution of problem (23) obtained by the above method. We use the same post-processing strategy as Liu *et al*. (2020) to further obtain an approximated solution $\hat{\mathbf{X}}$ of problem (22): let $\mathbf{D}\in\mathcal{S}^n$ be a diagonal matrix with $D_{ii} = 1/\sqrt{n\hat{Y}_{ii}}\ (i = 1,\dots,n)$ and $\hat{\mathbf{X}} = n(\mathbf{D}\hat{\mathbf{Y}}\mathbf{D})$. One can observe that $\hat{\mathbf{X}}$ preserves the rank of $\hat{\mathbf{Y}}$ while having all ones on its diagonal.

To sum up, the nearest low-rank correlation matrix problem can be effectively solved using Algorithm 1. Its convergence falls into the scope of Proposition A1 in Appendix A. Since $f(\mathbf{Z})$ has a local Lipschitz gradient on the bounded region $\mathcal{C}'$, the L-smooth adaptable property holds. It is also easy to see that Assumption A1 holds. We need to choose functions g_1 and g_2 such that the penalty term $p(\mathbf{Z})$ satisfying $h(\mathbf{Z})+(1/\tau\ )\,p(\mathbf{Z})$ is supercoercive for all $\tau > 0$, and this is true for the widely used functions $g(x) = x^0$ $\left(g_1(a) = a^0,\ g_2(b) = b^0\ \right)$ and $g(x) = x^p\ (0 < p \leq 1)\ (g_1(a) = (p/p_1\ )\,a^{p_1},\ g_2(b) = (p/p_2\ )\,b^{p_2},$

$1/p = 1/p_1 + 1/p_2, p_1, p_2 \in \mathbf{Z}^+$), since either $p(\mathbf{Z})/\|\mathbf{Z}\|_{\mathrm{F}}$ is lower bounded or its limit is ∞, and $\lim_{\|\mathbf{Z}\|_{\mathrm{F}}\to\infty} h(\mathbf{Z})/\|\mathbf{Z}\|_{\mathrm{F}} = +\infty$. Furthermore, the assumption $T_\tau(\mathbf{Z}) \subset \mathcal{C}'$ has to be fulfilled. Please refer to Bolte *et al*. (2018) for how to guarantee this condition. However, to keep our presentation simple and transparent, these technical issues will not be pursued here.

Remark 3. Note that $\tau > L$ is needed in the convergence results of Proposition A1 and Theorem A1 in Appendix A.

Algorithm 1. Nearest low-rank correlation matrix.

```
Require: $\mathbf{H}$, $\mathbf{C}$, $\rho = 1.1$.
Initialize:$\mathbf{U}^{0,0} = \mathbf{V}^{0,0} = \mathbf{P}\sqrt{\mathbf{D}}$ where $\mathbf{Y}^{0,0} = \mathbf{C}/n = \mathbf{PDP}^{\mathrm{T}}$, $\mu_{1,0}$, $\mu_{2,0}$, $\varepsilon_0$, $\tau_{0,0} = 1$, $k = 0$.
while $\frac{\|\mathrm{diag}(\mathbf{Y}^{k,0})-\mathbf{e}/n\|}{\max\{\|\mathbf{Y}^{k,0}\|_{\mathrm{F}},1\}} > 10^{-4}$ and $\sum_{i=r+1}^{d} g(\sigma_i(\mathbf{Y}^{k,0})) > 10^{-4}$ do
    Let $j = 0$.
    while $\frac{\|\mathbf{Y}^{k,j+1}-\mathbf{Y}^{k,j}\|_{\mathrm{F}}}{\max\{\|\mathbf{Y}^{k,j+1}\|_{\mathrm{F}},1\}} > \varepsilon_k$ do
        Set $\mathbf{Z}^{k,j} = ((\mathbf{U}^{k,j})^{\mathrm{T}}, (\mathbf{V}^{k,j})^{\mathrm{T}})^{\mathrm{T}}$, compute $P_{\tau_{k,j}}(\mathbf{Z}^{k,j})$ using (25).
        For $i = 1, 2, \ldots, r$, obtain $\mathbf{U}_{:,i}^{k,j+1}$ by (26);
        For $i = r+1, \ldots, d$ ,
        if $(P_{\tau_{k,j}}^{1})_{:,i}$=0 then
            $\mathbf{U}_{:,i}^{k,j+1}$=0,
        else
            $\mathbf{U}_{:,i}^{k,j+1} = -t_i{}^*(P_{\tau_{k,j}}^{1})_{:,i}$.
        end if
        if $(P_{\tau_{k,j}}^{2})_{:,i}$=0 then
            $\mathbf{V}_{:,i}^{k,j+1}$=0,
        else
            $\mathbf{V}_{:,i}^{k,j+1} = -s_i^*(P_{\tau_{k,j}}^{2})_{:,i}$.
        end if
        where $s_i^*$, $t_i^*$ is the positive root of (28).
        Update $\mathbf{Y}^{k,j+1} = \mathbf{U}^{k,j+1}(\mathbf{V}^{k,j+1})^{\mathrm{T}}$, $\tau_{k,j+1} = \min\{\rho\tau_{k,j}, 10^{-6}\}$, $j = j+1$.
    end while
    Set $\mathbf{U}^{k+1,0} = \mathbf{U}^{k,J}$, $\mathbf{V}^{k+1,0} = \mathbf{V}^{k,J}$, $\mathbf{Y}^{k+1,0} = \mathbf{Y}^{k,J}$, $\tau_{k+1,0} = \tau_{k,J}$ ($J$ is the total number of iterations in the current inner loop), $\varepsilon_{k+1} = \max\{0.2\varepsilon_k,\ 10^{-4}\}$.
    if $\frac{\|\mathrm{diag}(\mathbf{Y}^{k+1,0})-\mathbf{e}/n\|}{\max\{\|\mathbf{Y}^{k+1,0}\|_{\mathrm{F}},1\}} > 10^{-4}$ then
        $\mu_{1,k+1} = 5\mu_{1,k}$.
    end if
    if $\sum_{i=r+1}^{d} g(\sigma_i(\mathbf{Y}^{k+1,0})) > 10^{-4}$ then
        $\mu_{2,k+1} = 5\mu_{2,k}$.
    end if
    $k = k+1$.
end while
```

Here L is the constant included in the L-smooth adaptable condition, which means that τ must be chosen sufficiently large. Since only the existence of L is described in Liu *et al.* (2020) for the nearest low-rank correlation matrix problem without knowing its value, we make the parameter τ incrementally larger in Algorithm 1.

4.2. Quadratic inverse problem. Quadratic inverse problems are widely used in medical imaging, geophysical exploration, material science, and control engineering. In medical imaging, the quadratic inverse problem technique is used in the reconstruction of computer tomography and magnetic resonance imaging, which significantly improves the accuracy of diagnosis. In geophysical exploration, scientists are able to invert underground structures and discover mineral, oil and gas resources by processing the data of seismic waves, gravity and magnetism. In materials science, a quadratic inverse problem technique is used to reconstruct defects and structures inside materials to ensure the quality of materials and components through nondestructive testing methods. In control engineering, system identification techniques use quadratic inverse problems to invert the dynamic model of the system by observing the input and output data, so as to optimize the control strategy.

The quadratic inverse problem can be generally described as follows. Given a symmetric matrix $\mathbf{A}_i \in \mathbb{R}^{m\times m}$ and possibly noisy measurements $\mathbf{B}_i \in \mathbb{R}^{n\times n}$ for $i = 1, 2, \ldots, l$, our goal is to find a solution $\mathbf{X} \in \mathbb{R}^{m\times n}$ that satisfies the following equation:

$$\mathbf{X}^{\mathrm{T}}\mathbf{A}_i\mathbf{X} = \mathbf{B}_i, \quad i = 1, 2, \ldots, l. \tag{29}$$

The famous phase retrieval problem is a particular case of this problem, which has been extensively studied in the literature (Beck and Eldar, 2012; Luke, 2017). Applying the least-squares method to quantify the error and taking a low-rank constraint on matrix $\mathbf{X}$, the problem can then be rewritten as the following nonconvex problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} f(\mathbf{X}) = \frac{1}{4}\sum_{i=1}^{l} \left\|\mathbf{X}^{\mathrm{T}}\mathbf{A}_i\mathbf{X} - \mathbf{B}_i\right\|_{\mathrm{F}}^2 \tag{30a}$$

subject to

$$\mathrm{rank}(\mathbf{X}) \le r. \tag{30b}$$

By applying the bilinear relaxation model for the proposed penalty method, the above problem can be solved by

$$\begin{aligned}\min_{\mathbf{UV}^{\mathrm{T}}\in\bar{\mathcal{C}}} & F_\mu^2(\mathbf{U},\mathbf{V}) \\ &= \frac{1}{4}\sum_{i=1}^{l}\left\|\mathbf{VU}^{\mathrm{T}}\mathbf{A}_i\mathbf{UV}^{\mathrm{T}} - \mathbf{B}_i\right\|_{\mathrm{F}}^2 \\ &\quad + \mu\sum_{i=r+1}^{d}\left[g_1(\|\mathbf{U}_{:,i}\|_2) + g_2(\|\mathbf{V}_{:,i}\|_2)\right]\end{aligned} \tag{31}$$

with $\bar{\mathcal{C}} = \mathbb{R}^{m\times n}$.

This problem coincides with the form of problem (19) with

$$f(\mathbf{Z}) = \frac{1}{4}\sum_{i=1}^{l} \left\|\mathbf{V}\mathbf{U}^{\mathrm{T}}\mathbf{A}_i\mathbf{U}\mathbf{V}^{\mathrm{T}} - \mathbf{B}_i\right\|_{\mathrm{F}}^2$$

and

$$p(\mathbf{Z}) = \mu \sum_{i=r+1}^{d} \left[g_1\left(\|\mathbf{U}_{:,i}\|_2\right) + g_2\left(\|\mathbf{V}_{:,i}\|_2\right)\right].$$

By taking

$$h(\mathbf{Z}) = \frac{1}{4}\|\mathbf{Z}\|_{\mathrm{F}}^4 + \frac{1}{2}\|\mathbf{Z}\|_{\mathrm{F}}^2, \tag{32}$$

we can prove that $Lh - f$ is convex on $\mathcal{C}$ through a similar procedure as in Bolte *et al.* (2018) if

$$L \geq \sum_{i=1}^{l} 3\,\|\mathbf{A}_i\|_{\mathrm{F}}^2 + \|\mathbf{A}_i\|_{\mathrm{F}}\|\mathbf{B}_i\|_{\mathrm{F}}.$$

By the expressions of f and h, we calculate the following terms:

$$\begin{aligned}\nabla h(\mathbf{Z}) &= \left(\|\mathbf{Z}\|_{\mathrm{F}}^2 + 1\right)\mathbf{Z} \\ &= \begin{pmatrix}\left(\|\mathbf{U}\|_{\mathrm{F}}^2 + \|\mathbf{V}\|_{\mathrm{F}}^2 + 1\right)\mathbf{U} \\ \left(\|\mathbf{U}\|_{\mathrm{F}}^2 + \|\mathbf{V}\|_{\mathrm{F}}^2 + 1\right)\mathbf{V}\end{pmatrix},\end{aligned} \tag{33}$$

$$\begin{aligned}P_\tau(\mathbf{Z}) &= \nabla f(\mathbf{Z}) - \tau\nabla h(\mathbf{Z}) \\ &= \begin{pmatrix}\displaystyle\sum_{i=1}^{l}\mathbf{A}_i\mathbf{U}\mathbf{V}^{\mathrm{T}}\left(\mathbf{V}\mathbf{U}^{\mathrm{T}}\mathbf{A}_i\mathbf{U}\mathbf{V}^{\mathrm{T}}\text{-}\mathbf{B}_i\right)\mathbf{V} \\ -\tau\left(\|\mathbf{U}\|_{\mathrm{F}}^2 + \|\mathbf{V}\|_{\mathrm{F}}^2 + 1\right)\mathbf{U} \\ \displaystyle\sum_{i=1}^{l}\left(\mathbf{V}\mathbf{U}^{\mathrm{T}}\mathbf{A}_i\mathbf{U}\mathbf{V}^{\mathrm{T}} - \mathbf{B}_i\right)^{\mathrm{T}}\mathbf{V}\mathbf{U}^{\mathrm{T}}\mathbf{A}_i^{\mathrm{T}}\mathbf{U} \\ -\tau\left(\|\mathbf{U}\|_{\mathrm{F}}^2 + \|\mathbf{V}\|_{\mathrm{F}}^2 + 1\right)\mathbf{V}\end{pmatrix}.\end{aligned} \tag{34}$$

Substituting (33) and (34) into (24), due to the separability of columns, we can obtain, for each column with index $i = 1, 2, \ldots, r$,

$$\begin{cases}\tau\left(\left\|\mathbf{U}^{k+1}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{V}^{k+1}\right\|_{\mathrm{F}}^2 + 1\right)\mathbf{U}_{:,i}^{k+1} = -\left(P_\tau^{1,k}\right)_{:,i}, \\ \tau\left(\left\|\mathbf{U}^{k+1}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{V}^{k+1}\right\|_{\mathrm{F}}^2 + 1\right)\mathbf{V}_{:,i}^{k+1} = -\left(P_\tau^{2,k}\right)_{:,i}\end{cases} \tag{35}$$

and for each column with index $i = r+1, \ldots, d$,

$$\begin{cases}\Big[\mu g_1'\left(\left\|\mathbf{U}_{:,i}^{k+1}\right\|_2\right)/\left\|\mathbf{U}_{:,i}^{k+1}\right\|_2 \\ +\tau\left(\left\|\mathbf{U}^{k+1}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{V}^{k+1}\right\|_{\mathrm{F}}^2 + 1\right)\Big]\mathbf{U}_{:,i}^{k+1} \\ = -\left(P_\tau^{1,k}\right)_{:,i}, \\ \Big[\mu g_2'\left(\left\|\mathbf{V}_{:,i}^{k+1}\right\|_2\right)/\left\|\mathbf{V}_{:,i}^{k+1}\right\|_2 \\ +\tau\left(\left\|\mathbf{U}^{k+1}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{V}^{k+1}\right\|_{\mathrm{F}}^2 + 1\right)\Big]\mathbf{V}_{:,i}^{k+1} \\ = -\left(P_\tau^{2,k}\right)_{:,i}.\end{cases} \tag{36}$$

Now, for all columns with index $i = 1, \ldots, d$, according to the linear correlation described by (35) and (36), we consider the following two cases:

(i) If $\left(P_\tau^{1,k}\right)_{:,i}$=$\mathbf{0}$, then $\mathbf{U}_{:,i}^{k+1} = \mathbf{0}$; if $\left(P_\tau^{2,k}\right)_{:,i}$=$\mathbf{0}$, then $\mathbf{V}_{:,i}^{k+1}$=$\mathbf{0}$.

(ii) If $\left(P_\tau^{1,k}\right)_{:,i} \neq \mathbf{0}$, then $\mathbf{U}_{:,i}^{k+1} = -t_i^*\left(P_\tau^{1,k}\right)_{:,i}$; if $\left(P_\tau^{2,k}\right)_{:,i} \neq \mathbf{0}$, then $\mathbf{V}_{:,i}^{k+1} = -s_i^*\left(P_\tau^{2,k}\right)_{:,i}$, where s_i^*, ${t_i}^*$ are the positive roots of the following equations:

$$\begin{cases}\tau t_i\Big(\sum_{j=1}^{d} t_j^2\left\|\left(P_\tau^{1,k}\right)_{:,j}\right\|_2^2 \\ \quad+\sum_{j=1}^{d} s_j^2\left\|\left(P_\tau^{2,k}\right)_{:,j}\right\|_2^2\text{+}1\Big) - 1 = 0, \\ \tau s_i\Big(\sum_{j=1}^{d} t_j^2\left\|\left(P_\tau^{1,k}\right)_{:,j}\right\|_2^2 \\ \quad+\sum_{j=1}^{d} s_j^2\left\|\left(P_\tau^{2,k}\right)_{:,j}\right\|_2^2\text{+}1\Big) - 1 = 0,\end{cases} \tag{37}$$

$$\begin{cases}\mu g_1'\left(t_i\left\|\left(P_\tau^{1,k}\right)_{:,i}\right\|_2\right)/\left\|\left(P_\tau^{1,k}\right)_{:,i}\right\|_2 \\ \quad+\tau t_i\Big(\sum_{j=1}^{d} t_j^2\left\|\left(P_\tau^{1,k}\right)_{:,j}\right\|_2^2 \\ \quad+\sum_{j=1}^{d} s_j^2\left\|\left(P_\tau^{2,k}\right)_{:,j}\right\|_2^2\text{+}1\Big) - 1 = 0, \\ \mu g_2'\left(s_i\left\|\left(P_\tau^{2,k}\right)_{:,i}\right\|_2\right)/\left\|\left(P_\tau^{2,k}\right)_{:,i}\right\|_2 \\ \quad+\tau s_i\Big(\sum_{j=1}^{d} s_j^2\left\|\left(P_\tau^{2,k}\right)_{:,j}\right\|_2^2 \\ \quad+\sum_{j=1}^{d} t_j^2\left\|\left(P_\tau^{1,k}\right)_{:,j}\right\|_2^2\text{+}1\Big) - 1 = 0.\end{cases} \tag{38}$$

To sum up, the quadratic inverse problem can be effectively solved using Algorithm 2. In order to apply the convergence results referred to in Appendix A, we observe that h given above is 1-strongly convex on $\mathbb{R}^{m\times n}$ and it easy to see that the L-smooth adaptable property and Assumptions A1 and A4 hold (Bolte *et al.*, 2018). The supercoercive property of $h(\mathbf{Z}) + (1/\tau\,)\,p(\mathbf{Z})$ for all $\tau > 0$ holds for the common functions: (i) $g_1(a) = (p/p_1)\, a^{p_1}$, $g_2(b) = (p/p_2)\, b^{p_2}$, $(0 < p \le 1,\ 1/p = 1/p_1 + 1/p_2\ (p_1, p_2 \in \mathbb{Z}^+))$; (ii) $g_1(a) = a^0$, $g_2(b) = b^0$, since we obviously have $\lim_{\|\mathbf{Z}\|_\mathrm{F} \to \infty} h(\mathbf{Z})/\|\mathbf{Z}\|_\mathrm{F} = \infty$. Furthermore, the function f is a real polynomial, hence semi-algebraic (Bolte *et al.*, 2014). The functions g_1, g_2 mentioned above are all semi-algebraic (Bolte *et al.*, 2014), and $\|\cdot\|_2$ is a semi-algebraic function, therefore, since the addition and composition of semi-algebraic functions result in a semi-algebraic function; it follows that for model (31), the objective F_μ^2 is a KL function, and BPG can be applied on the problem to produce a globally convergent sequence which converges to a critical point of F_μ^2.

Algorithm 2. Quadratic inverse problem.

```
Require: $\mathbf{A}_i \in \mathbb{R}^{m\times m}$ and $\mathbf{B}_i \in \mathbb{R}^{n\times n}$ for $i = 1, 2, \ldots, l$, $L \ge \sum_{i=1}^{l} \Big(3\,\|\mathbf{A}_i\|_\mathrm{F}^2 + \|\mathbf{A}_i\|_\mathrm{F}\|\mathbf{B}_i\|_\mathrm{F}\Big)$, $\tau > L$ .
  Initialize:$\mathbf{U}^{0,0} \in \mathbb{R}^{m\times d}$, $\mathbf{V}^{0,0} \in \mathbb{R}^{n\times d}$, $\mathbf{X}^{0,0} = \mathbf{U}^{0,0}\big(\mathbf{V}^{0,0}\big)^\mathrm{T}$, $\mu_0, \varepsilon_0, k = 0$.
  while $\sum_{i=r+1}^{d} g\left(\sigma_i\left(\mathbf{X}^{k,0}\right)\right) > 0.1$ do
    Let $j = 0$.
    while $\frac{\|\mathbf{X}^{k,j+1} - \mathbf{X}^{k,j}\|_\mathrm{F}}{\max\{\|\mathbf{X}^{k,j+1}\|_\mathrm{F}, 1\}} > \varepsilon_k$ do
      Set $\mathbf{Z}^{k,j} = \Big(\big(\mathbf{U}^{k,j}\big)^\mathrm{T}, \big(\mathbf{V}^{k,j}\big)^\mathrm{T}\Big)^\mathrm{T}$, compute $P_\tau\left(\mathbf{Z}^{k,j}\right)$ using (34).
      For $i = 1, \ldots, d$,
      if $\left(P_\tau^1\right)_{:,i}$=0 then
        $\mathbf{U}_{:,i}^{k,j+1} = \mathbf{0}$,
      else
        $\mathbf{U}_{:,i}^{k,j+1} = -t_i{}^*\left(P_\tau^1\right)_{:,i}$.
      end if
      if $\left(P_\tau^2\right)_{:,i}$=0 then
        $\mathbf{V}_{:,i}^{k,j+1}$=0,
      else
        $\mathbf{V}_{:,i}^{k,j+1} = -s_i^*\left(P_\tau^2\right)_{:,i}$.
      end if
      where $s_i^*$, $t_i^*$ are the positive roots of (37) and (38).
      $j = j + 1$.
    end while
    Set $\mathbf{U}^{k+1,0} = \mathbf{U}^{k,J}$, $\mathbf{V}^{k+1,0} = \mathbf{V}^{k,J}$, $\mathbf{X}^{k+1,0} = \mathbf{X}^{k,J}$ ($J$ is the total number of iterations in the current inner loop), $\varepsilon_{k+1} = \max\left\{0.5\varepsilon_k,\ 10^{-4}\right\}$.

    if $\sum_{i=r+1}^{d} g\left(\sigma_i\left(\mathbf{X}^{k+1,0}\right)\right) > 0.1$ then
      $\mu_{k+1} = 10\mu_k$.
    end if
    $k = k + 1$.
  end while
```

5. Numerical experiments

In this section, experiments are conducted on the above two applications to test the effectiveness and efficiency of the proposed method. For the nearest low-rank correlation matrix problem, we conduct numerical experiments and comparisons in Section 5.1. Meanwhile, an extensive experiment is performed for the quadratic inverse problem in Section 5.2. All the experiments are performed on a PC with Windows 10 LTSC, Intel(R) Core(TM) i7-5650U CPU (2.20 GHz), and 8G RAM. The code is written in Matlab 2018a.

For the following two groups of experiments, we adopt the following common settings. The function

$$g(x) = \min_{\substack{x=ab \\ a,b\ge 0}} g_1(a) + g_2(b)$$

is chosen as

$$g(x) = x^p \quad (0 < p \le 1, \quad x \in [0,\ \infty))$$

with

$$g_1(a) = \frac{p}{p_1} a^{p_1}$$

and

$$g_2(b) = \frac{p}{p_2} b^{p_2}$$

satisfying $\frac{1}{p} = \frac{1}{p_1} + \frac{1}{p_2}$; for example, in case $p = 1$, we simply choose $g_1(a) = a^2/2$ and $g_2(b) = b^2/2$; in case $p = 0.5$, we choose $g_1(a) = a/2$ and $g_2(b) = b/2$; and in case $p = 2/3$, $g_1(a) = 2a/3$ and $g_2(b) = b^2/3$ are selected.

5.1. Experiments on the nearest low-rank correlation matrix problem.

We now verify the performance of the method presented in Section 4.1. For fair comparison, we use exactly the same experimental setup as Liu *et al.* (2020) as follows. Choose $\mathbf{H}$ as the all-ones matrix, and $\mathbf{C}$ is with $C_{ij} = 0.5 + 0.5e^{-0.05|i-j|}$ for $i, j = 1, \ldots, n$. The initial values of parameters are $\mu_{1,0} = 0.5$, $\mu_{2,0} = 0.5$, $\varepsilon_0 = 10^{-3}$. To evaluate the performance of the competing methods, we adopt the same quantity $residue = \|\mathbf{H} \circ (\hat{\mathbf{X}} - \mathbf{C})\|_\mathrm{F}$. Since $\operatorname{rank}(\mathbf{X}) \le d \le n$, and we do not know the rank of matrix $\mathbf{X}$, we choose $d = n$ to reduce the selection of parameters.

Figure 2 shows the singular values of the ground-truth correlation matrix $\mathbf{C}$ and the nearest low-rank correlation matrix $\hat{\mathbf{X}}$. We present experimental results for different matrix sizes, and different values p and r. From Fig. 2(a) it can be observed that the last $n - r$ singular values of $\hat{\mathbf{X}}$ are all zeros, which strictly coincides with the constraint $\operatorname{rank}(\mathbf{X}) \le r$ in all cases of the matrix size. Figure 2(b) shows that the suggested approach satisfies the rank constraint condition well in the three cases of the p value. Figures 2(a) and

(b) also show the effect of our method on different r values. Figure 2(c) shows the result of the singular values after adopting the penalty function $\sum_{i=1}^{d} \sigma_i{}^p(\mathbf{X})$ with different p values. Except for the penalty function, other experimental settings are the same as shown in Fig. 2(b). The rank constraint cannot be satisfied as $p = 1$ and $p = 2/3$. As $p = 1/2$, although the rank constraint condition is satisfied, in contrast, it is satisfied more compactly by using our penalty function. The penalty function $\sum_{i=1}^{d} \sigma_i^p(\mathbf{X})$ yields an approximation of $\mathbf{C}$ with the lowest possible rank while the penalty proposed here meets the low-rank constraint exactly.

Figure 3 shows variation in the objective function values with the number of iterations within the first outer loop, corresponding to all cases in Fig. 2. The objective function value decreases rapidly in a few iterations and finally almost stabilizes at the smallest value in all cases.

In Table 1, we compare CPU times and residues of our method with the methods proposed by Gao and Sun (2010) and Liu *et al.* (2020) denoted by PenCorr and $\mathrm{PM}_{0.5}$, respectively. Note that $\mathrm{PM}_{0.5}$ achieves better results with $p = 0.5$ than PM_1 with $p = 1$; thus, we only select $\mathrm{PM}_{0.5}$ as the compared method. The best residue and the least time cost in each row are shown in boldface. One can see from Table 1 that the proposed method outperforms $\mathrm{PM}_{0.5}$ and PenCorr a little bit in terms of residues in most cases, and costs far less CPU time than the other two methods.

5.2. Experiments on the quadratic inverse problem.

In this section, experiments are performed on synthetic data to verify the effectiveness and computational efficiency of the proposed method for quadratic inverse problems. The ground-truth low rank data is generated by $\mathbf{X} = \mathbf{U}_{\text{true}}\mathbf{V}_{\text{true}}^{\mathrm{T}}$, where $\mathbf{U}_{\text{true}} \in \mathbb{R}^{m\times r}$ and $\mathbf{V}_{\text{true}} \in \mathbb{R}^{n\times r}$ are i.i.d. data sampled from the normal distribution $\mathcal{N}(0,\ 1)$ and normalized on columns. To ensure $\mathbf{X}$ satisfying the constraint $(\mathbf{X})^{\mathrm{T}}\mathbf{X} \preceq \mathbf{I}_n$, we use the eigenvalue decomposition of $\mathbf{X}(\mathbf{X})^{\mathrm{T}}$ and $(\mathbf{X})^{\mathrm{T}}\mathbf{X}$, namely, $\mathbf{X}(\mathbf{X})^{\mathrm{T}} = \mathbf{P}_1\mathbf{D}\mathbf{P}_1^{\mathrm{T}}$, $(\mathbf{X})^{\mathrm{T}}\mathbf{X} = \mathbf{P}_2\mathbf{D}\mathbf{P}_2^{\mathrm{T}}$, and the truth $\mathbf{X}_{\text{true}}$ is set as $\mathbf{X}_{\text{true}} = \mathbf{P}_1\sqrt{\mathbf{D}/\|\mathbf{D}\|_{\mathrm{F}}}\,\mathbf{P}_2^{\mathrm{T}}$. The intrinsic rank of $\mathbf{X}_{\text{true}}$ satisfies $\operatorname{rank}(\mathbf{X}) \le r$. Set $\mathbf{A}_i{=}\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}} \in \mathbb{R}^{m\times m}$ $(i = 1, 2, \dots, 10)$, in which the entries of $\mathbf{a}_i$ are i.i.d. data sampled from the normal distribution $\mathcal{N}(0,\ 1)$ for ten times. The matrices $\mathbf{B}_i \in \mathbb{R}^{n\times n}$ are generated from $\mathbf{X}_{\text{true}}^{\mathrm{T}}\mathbf{A}_i\mathbf{X}_{\text{true}} = \mathbf{B}_i$ $(i = 1, \dots, 10)$.

In this experimental scenario, we initialize $\mathbf{U}^{0,0} \in \mathbb{R}^{m\times d}$ and $\mathbf{V}^{0,0} \in \mathbb{R}^{n\times d}$ through i.i.d. randomly sampling from the normal distribution $\mathcal{N}(0,\ 1)$, and $\mathbf{U}^{0,0}$ and $\mathbf{V}^{0,0}$ are normalized on columns. To ensure the initial $\mathbf{X}^{0,0} = \mathbf{U}^{0,0}\left(\mathbf{V}^{0,0}\right)^{\mathrm{T}}$ satisfying the constraint $\left(\mathbf{X}^{0,0}\right)^{\mathrm{T}}\mathbf{X}^{0,0} \preceq \mathbf{I}_n$, we use the same operation as on $\mathbf{X}$. We set

$$L = \sum_{i=1}^{l} 3\,\|\mathbf{A}_i\|_{\mathrm{F}}^2 + \|\mathbf{A}_i\|_{\mathrm{F}}\|\mathbf{B}_i\|_{\mathrm{F}},$$

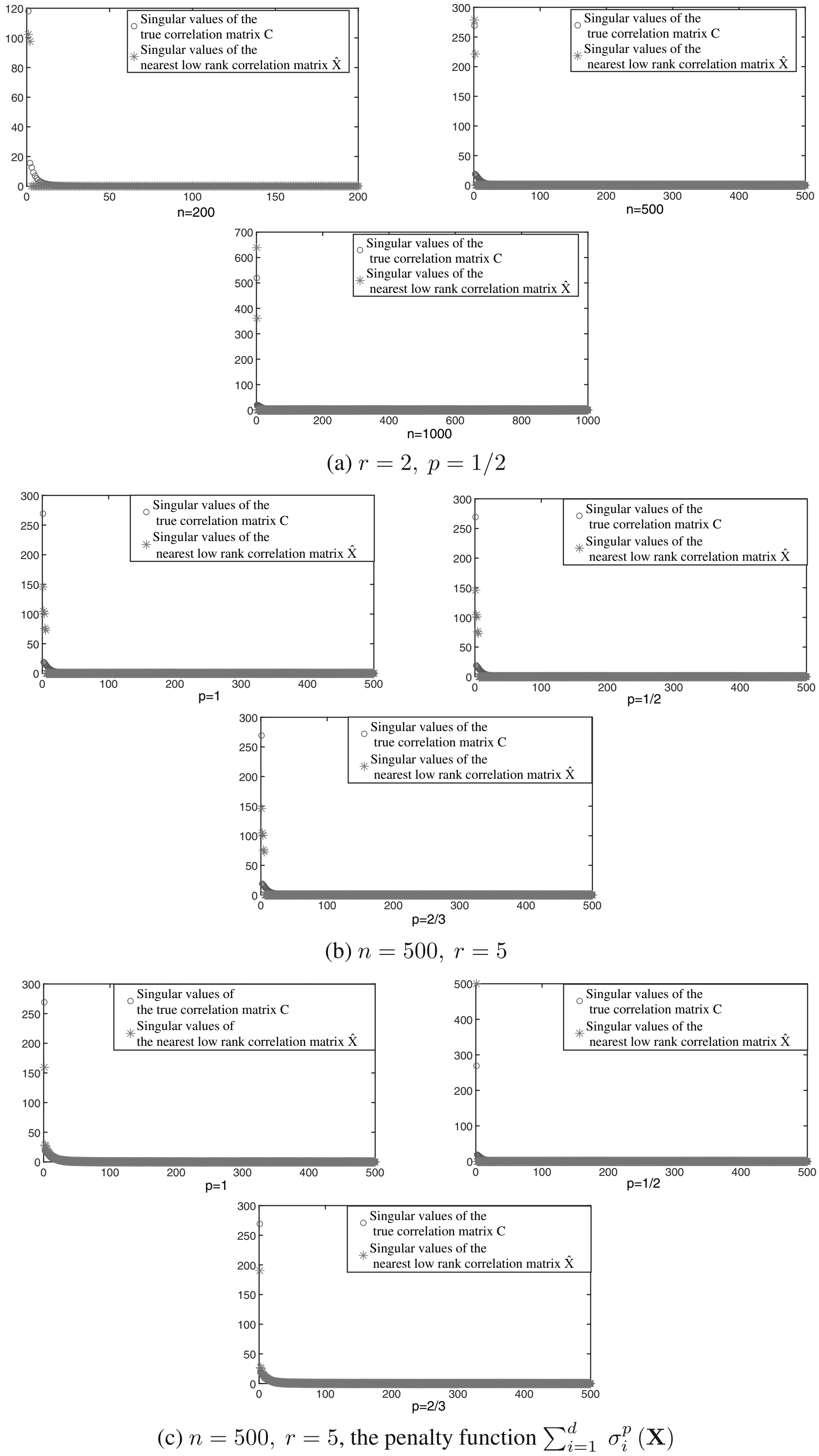
and $\tau{=}1.1L$. The parameters are initialized as $\mu_0 = 10^{-4}$, $\varepsilon_0 = 0.1$. We evaluate the accuracy of the estimated $\hat{\mathbf{X}}$ to ground-truth $\mathbf{X}_{\text{true}}$ by $residue = \|\hat{\mathbf{X}} - \mathbf{X}_{\text{true}}\|_{\mathrm{F}}/\|\mathbf{X}_{\text{true}}\|_{\mathrm{F}}$, and a lower value indicates a more accurate result. Since the objective function of this problem is highly nonconvex and nonsmooth, convergence is slow when dealing with a large-size matrix, so we choose a smaller d to speed up calculations. We set $d = [1.25r]$ (round up to an integer) in the same way as done by Jia *et al.* (2020).

The singular values of the ground-truth matrix $\mathbf{X}_{\text{true}}$ and the estimated matrix $\hat{\mathbf{X}}$ are shown in Fig. 4. Figure 4(a) shows the presented method meets the rank constraint well in all matrix sizes. The choice of parameter p is illustrated in Fig. 4(b). In all the three cases, the result satisfying the rank constraint is good. Figures 4(a) and (b) also indicate the good effect of our method on different r values. Figure 4(c) shows the result of adopting the penalty function $\sum_{i=1}^{d} \sigma_i^p(\mathbf{X})$ with different p values. Except for the penalty function, all other experiment settings are the same as in Fig. 4(b). In all the three cases of p, the rank constraint condition can not be satisfied. The result of the penalty function $\sum_{i=1}^{d} \sigma_i^p(\mathbf{X})$ is affected by the value of d, while in the same settings the proposed penalty function is not affected by this value.

Figure 5 shows variation in the objective function values with the number of iterations within the first outer loop. The objective function values gradually decrease and converge in all the listed three cases. Similar results can be obtained with different r and p values. In Table 2, we report CPU times and residues for different p values. The best residue and the least time cost in each row are expressed in boldface. One can see that as $p = 2/3$, the proposed method yields the best residues in most cases. The least amount of CPU time is corresponds to $p = 0.5$ in most cases. As $p = 1$, the residue results are worse than other two cases.

6. Conclusion

We present a general exact penalty method for a class of nonconvex nonsmooth matrix optimization problems with a rank constraint. We prove that the penalty function can be expressed as a sum of specific functions with the smallest singular values, and thus the penalty problem is equivalent to the original problem. Then, the proposed penalty problem is transformed into an equivalent bilinear factorization form to avoid SVD computation. Furthermore, a BPG algorithm is designed

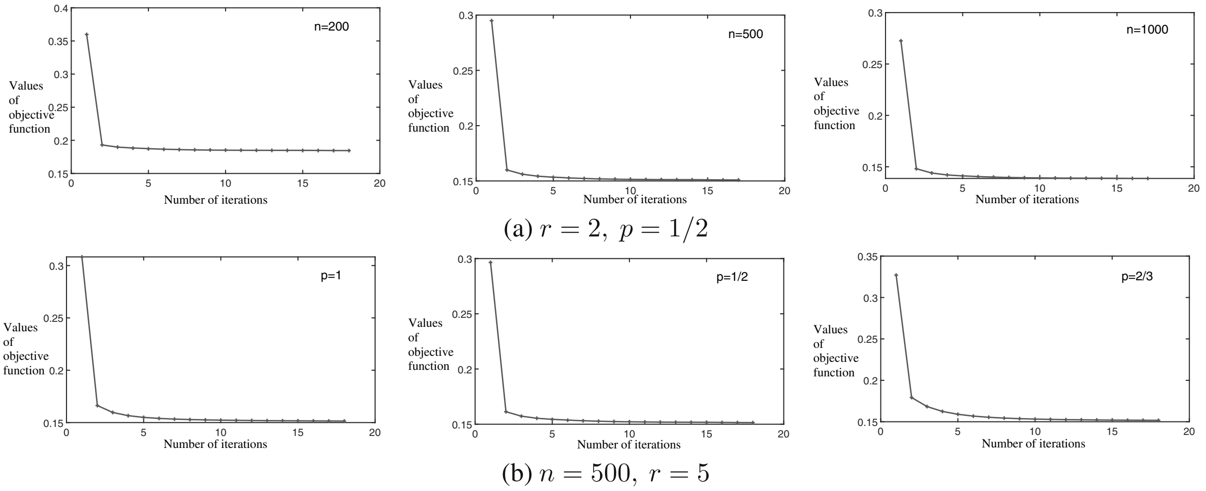


(a) $r = 2,\ p = 1/2$

(b) $n = 500,\ r = 5$

(c) $n = 500,\ r = 5$, the penalty function $\sum_{i=1}^{d}\ \sigma_i^p\,(\mathbf{X})$

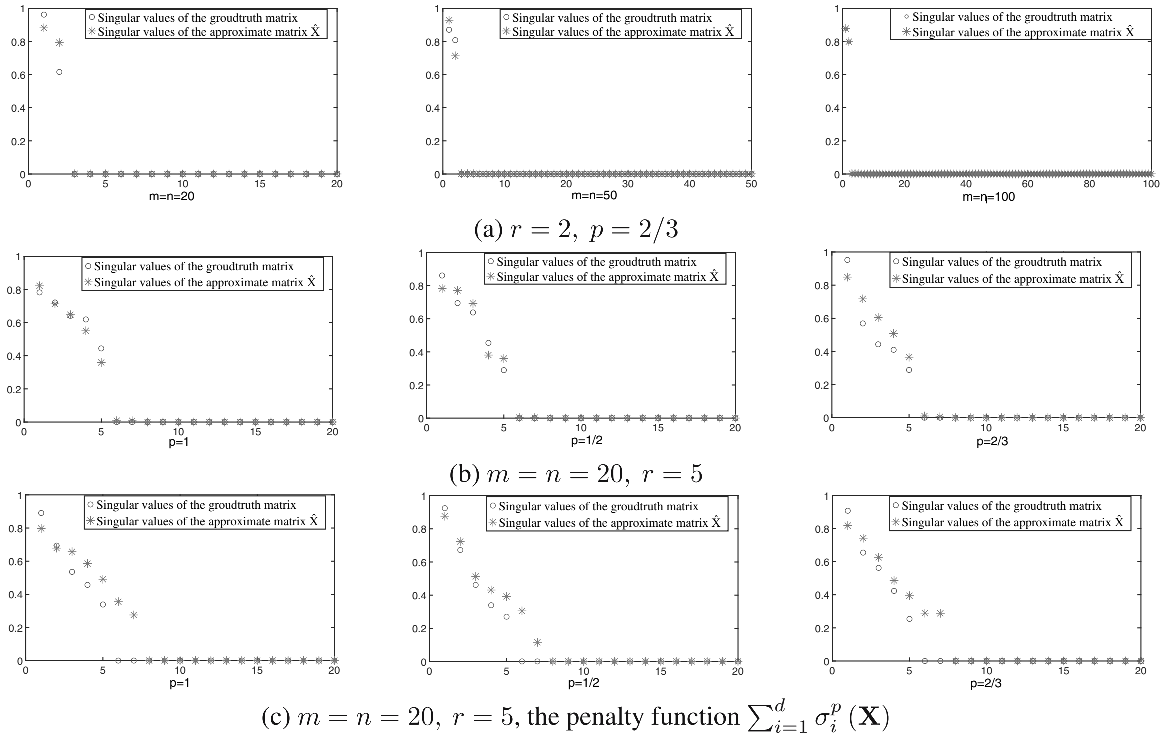
Fig. 2. Comparison for keeping the rank constraint.



(a) $r = 2,\ p = 1/2$

(b) $n = 500,\ r = 5$

Fig. 3. Variation in the objective function values.



(a) $r = 2,\ p = 2/3$

(b) $m = n = 20,\ r = 5$

(c) $m = n = 20,\ r = 5$, the penalty function $\sum_{i=1}^{d} \sigma_i^p\left(\mathbf{X}\right)$

Fig. 4. Comparison for keeping the rank constraint.

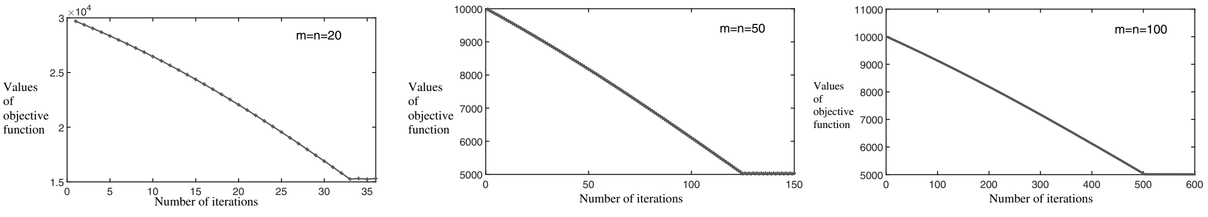


Fig. 5. Variation in the objective function values with the number of iterations in the case of $r = 2, p = 1/2$.

Table 1. Comparison of the residue and CPU time.

n	r	PM$_{0.5}$		Pencorr		Our method ($p = 0.5$)	
		Time(s)	Residue	Time(s)	Residue	Time(s)	Residue
500	2	389.323	156.4053	437.293	156.4172	**179.416**	**155.0347**
	5	303.377	78.8307	407.338	78.8342	**116.829**	**78.5631**
	10	249.849	38.6845	358.938	38.6852	**24.3952**	**38.6033**
	15	187.383	23.2497	273.484	**23.2463**	**5.0740**	23.7481
	20	102.384	15.7106	183.484	15.7080	**5.2267**	**15.5352**
1000	2	694.838	332.7649	795.483	332.8054	**447.2027**	**332.1289**
	5	539.940	189.3868	628.485	189.3978	**318.1396**	**189.0017**
	10	429.330	110.7867	510.384	110.7868	**228.4321**	**110.6066**
	15	354.393	**74.7463**	463.844	74.7494	**154.7981**	74.7688
	20	239.209	54.1675	345.243	54.1680	**88.8603**	**53.8598**
1500	2	3081.834	509.4009	4183.374	509.4665	**1193.7**	**508.8373**
	5	2846.374	301.1784	3629.385	301.1892	**1057.1**	**300.8431**
	10	2459.539	188.5594	2937.596	188.5554	**780.2371**	**188.2148**
	15	2084.380	135.3811	2849.382	135.3820	**584.9598**	**134.8534**
	20	1639.373	103.1023	2084.283	103.1043	**448.7425**	**103.0713**
2000	2	4373.293	686.1070	5930.293	686.1815	**2088.967**	**686.0691**
	5	4028.373	413.0689	4837.382	413.0763	**1877.233**	**412.8227**
	10	3547.383	**267.3751**	4293.162	267.3920	**1688.929**	267.4098
	15	3048.733	198.6823	3703.293	198.6795	**1460.605**	**198.1249**
	20	2694.383	156.1624	3493.056	156.1522	**1058.554**	**156.0143**

Table 2. Comparison of the residue and CPU time.

$m = n$	r	$p = 1$		$p = 0.5$		$p = 2/3$	
		Time(s)	Residue	Time(s)	Residue	Time(s)	Residue
20	2	41.782	1.4991	**9.343**	1.4424	19.418	**1.4084**
	5	**32.560**	1.6726	36.582	1.4084	59.168	**1.3969**
	10	52.089	1.4946	**31.831**	1.2083	76.261	**1.1727**
	15	98.453	1.5919	**64.886**	1.4627	141.322	**1.4022**
50	2	185.838	1.4516	**40.468**	1.4215	97.300	**1.3558**
	5	209.674	1.5977	**190.735**	1.2848	360.791	**1.2840**
	10	276.910	1.4413	**186.278**	1.2572	456.382	**1.2246**
	15	571.255	1.5529	**409.519**	1.4763	874.754	**1.4602**
100	2	472.116	1.4212	**167.855**	1.4039	378.653	**1.3970**
	5	791.636	1.5920	**736.154**	**1.2546**	1533.720	1.3140
	10	999.878	1.4517	**887.41**	1.2157	1825.149	**1.1776**
	15	2320.564	1.6003	**1852.654**	1.4349	3578.863	**1.4314**

for fast solving the factorization problem. The numerical experiments on two application problems are conducted and the results indicate the effectiveness and efficiency of the proposed method. There are still some issues to be addressed in the future. The first is that, when the objective is just continuously differentiable, only a descent of the objective function values is proved in the situation of nonconvex optimization, and the global convergence needs to be guaranteed. The second is how to design a parameter learning method to avoid parameter tuning.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (no. 62171361) and a Natural Science Basic Research Program of Shaanxi Province, China (2021JM-440).

References

Alain, R. (2013). Direct optimization of the dictionary learning problem, *IEEE Transactions on Signal Processing* **61**(22): 5495–5506.

Atwood, C.L. (1969). Optimal and efficient designs of experiments, *The Annals of Mathematical Statistics* **40**(5): 1570–1602.

Bauschke, H.H., Bolte, J. and Teboulle, M. (2016). A descent lemma beyond Lipschitz gradient continuity: First order methods revisited and applications, *Mathematics of Operations Research* **42**(2): 330–348.

Bauschke, H.H. and Borwein, J.M. (1997). Legendre functions and the method of Bregman projections, *Journal of Convex Analysis* **4**(1): 27–67.

Beck, A. and Eldar, Y. (2012). Sparsity constrained nonlinear optimization: Optimality conditions and algorithms, *SIAM Journal on Optimization* **23**: 1480–1509.

Bertero, M., Boccacci, P., Desidera, G. and Vicidomini, G. (2009). Image deblurring with poisson data: From cells to galaxies, *Inverse Problems* **25**(12): 123006.

Bhatia, R. (2011). *Matrix Analysis*, World Books Publishing Corporation, Beijing.

Bolte, J., Sabach, S. and M.Teboulle (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems, *Mathematical Programming* **146**(1): 459–494.

Bolte, J., Sabach, S. and M.Teboulle (2018). First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems, *SIAM Journal on Optimization* **28**(3): 2131–2151.

Candes, E.J., Wakin, M.B. and Boyd., S.P. (2008). Enhancing sparsity by reweighted ℓ_1 minimization, *Journal of Fourier Analysis and Applications* **14**(5): 877–905.

Censor, Y. and Zenios, S.A. (1992). Proximal minimization algorithm with d-functions, *Journal of Optimization Theory and Applications* **73**(3): 451–464.

Chartrand, R. (2007). Exact reconstruction of sparse signals via nonconvex minimization, *IEEE Signal Processing Letters* **14**(10): 707–710.

Chen, G. and Teboulle, M. (1993). Convergence analysis of a proximal-like minimization algorithm using Bregman functions, *SIAM Journal on Optimization* **3**(3): 538–543.

Eckstein, J. (1993). Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming, *Mathematics of Operations Research* **18**(1): 202–226.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456): 1348–1361.

Fazel, M., Hindi, H. and Boyd, S.P. (2003). Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices, *Proceedings of the 2003 American Control Conference, Denver, USA*, pp. 2156–2162.

Gao, Y. and Sun, D. (2010). A majorized penalty approach for calibrating rank constrained correlation matrix problems, *2010 IEEE International Conference on Computing, Guiyang, Guizhou, China*.

Horn, R.A. and Johnson, C.R. (1990). *Matrix Analysis*, Cambridge University Press, Cambridge.

Ji, S., Sze, K.-F., Zhou, Z., So, A. M.-C. and Ye, Y. (2013). Beyond convex relaxation: A polynomial-time non-convex optimization approach to network localization, *2013 Proceedings IEEE INFOCOM, Turin, Italy*, pp. 2499–2507, DOI: 10.1109/INFCOM.2013.6567056.

Jia, X., Feng, X. and Wang, W. (2020). Generalized unitarily invariant gauge regularization for fast low-rank matrix recovery, *IEEE Transactions on Neural Networks and Learning Systems* **32**(4): 1627–1641.

Li, J.-R. and White, J. (2001). Reduction of large circuit models via low rank approximate gramians, *International Journal of Applied Mathematics and Computer Science* **11**(5): 1151–1171.

Liang, Z., Zeng, D. and Guo, S. (2022). A fusion representation for face learning by low-rank constrain and high-frequency texture components, *Pattern Recognition Letters* **155**: 48–53, DOI:10.1016/j.patrec.2022.01.022.

Liu, T., Lu, Z. and Chen, X. (2020). An exact penalty method for semidefinite-box-constrained low-rank matrix optimization problems, *IMA Journal of Numerical Analysis* **40**(1): 563–586.

Lu, Y., Zhang, L. and Wu, J. (2015a). A smoothing majorization method for matrix minimization, *Optimization Methods and Software* **30**(1): 682–705.

Lu, Z., Zhang, Y. and Li, X. (2015b). Penalty decomposition methods for rank minimization, *Optimization Methods and Software* **30**(3): 531–558.

Lu, Z., Zhang, Y. and Lu, J. (2017). ℓ_p Regularized low-rank approximation via iterative reweighted singular value minimization, *Computational Optimization and Applications* **68**(3): 619–642.

Luke, D.R. (2017). Phase retrieval. What's new?, *SIAG/OPT Views and News* **25**(1): 1–5.

Nguyen, Q.V. (2017). Forward-backward splitting with Bregman distances, *Vietnam Journal of Mathematics* **45**(3): 519–539.

Recht, B., Fazel, M. and Parrilo, P.A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, *SIAM Review* **52**(3): 471–501.

Recht, B., Xu, W. and Hassibi, B. (2011). Null space conditions and thresholds for rank minimization, *Mathematical Programming* **127**: 175–202.

Sulaiman, I.M., Kaelo, P., Khalid, R. and Nawawi, M.K.M. (2024). A descent generalized RMIL spectral gradient algorithm for optimization problems, *International Journal of Applied Mathematics and Computer Science* **34**(2): 225–233, DOI: 10.61822/amcs-2024-0016.

Teboulle, M. (1992). Entropic proximal mappings with application to nonlinear programming, *Mathematics of Operations Research* **17**(3): 670–690.

Ülkü, I. and Kizgut, E. (2018). Large-scale hyperspectral image compression via sparse representations based on online learning, *International Journal of Applied Mathematics and Computer Science* **28**(1): 197–207, DOI: 10.2478/amcs-2018-0015.

Wang, S., Xiao, S. and Zhu, W. (2022). Multi-view fuzzy clustering of deep random walk and sparse low-rank embedding, *Information Sciences* **586**: 224–238.

Xu, Z., Chang, X., Xu, F. and Zhang, H. (2012). $L_{1/2}$ regularization: A thresholding representation theory and a fast solver, *IEEE Transactions on Neural Networks and Learning Systems* **586**(7): 1013–1027.

Yang, S., Tan, Y., Dong, R. and Tan, Q. (2023). Nonsmooth optimization control based on a sandwich model with hysteresis for piezo-positioning systems, *International Journal of Applied Mathematics and Computer Science* **33**(3): 449–461, DOI: 10.34768/amcs-2023-0033.

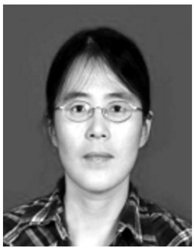
Zhong, Y., Li, C., Li, Z. and Duan, X. (2022). A proximal based algorithm for piecewise sparse approximation with application to scattered data fitting, *International Journal of Applied Mathematics and Computer Science* **32**(4): 671–682, DOI: 10.34768/amcs-2022-0046.

Wenjuan Zhang received his MS and PhD degrees from Xidian University, Xi'an, China, in 2005 and 2013, respectively. She was then a visiting scholar with the Department of Mathematics, University of Florida. She is currently an associate professor with the School of Science, Xi'an Technological University. Her research interests include modeling and numerical optimization methods in machine learning, computer vision, and image processing.

Jiayi Yao is currently pursuing her MS degree in the School of Science, Xi'an Technological University. Her research interests include low-rank optimization methods in subspace clustering.

Feng Xiao received his PhD degree from Northwestern University, Xi'an, China, in 2012. He is currently a professor at the Institute of Weapons Science and Technology, Xi'an Technological University. His research interests include deep learning in computer vision.

Yuping Wang is a professor at the School of Computer Science and Technology, Xidian University, China. He received his PhD degree from the Department of Mathematics, Xi'an Jiaotong University, China, in 1993. Currently, his research interests include the theory and applications of optimization methods, evolutionary computation, data mining and machine learning.



Yulian Wu received her BS degree in the School of Science at the Shandong University of Technology in 2003, and her MS and PhD degrees in the School of Science at Xidian University in 2006 and 2014, respectively. She is currently an associate professor in Xi'an Medical University. Her research interests include image processing and deep learning.

Appendix A

Convergence of the BPG algorithm (Bolte *et al.*, 2018)

Consider the following nonconvex composite minimization problem:

$$\inf_{\mathbf{X}} \left\{ F\left(\mathbf{X}\right) = f\left(\mathbf{X}\right) + p\left(\mathbf{X}\right) : \mathbf{X} \in \overline{\mathcal{C}} \right\}. \tag{A1}$$

It consists in minimizing the sum of two nonconvex functions: an extended real valued function p and a continuously differentiable function f. $\overline{\mathcal{C}}$ denotes the closure of $\mathcal{C}$ which is a nonempty, convex and open set in $\mathbb{R}^{m\times n}$.

Definition A1. (*L-smooth adaptable*) A pair (f, h) is called L-smooth adaptable on $\mathcal{C}$ if there exists $L > 0$ such that $Lh - f$ are convex on $\mathcal{C}$.

Definition A2. (*Kernel generating distance*) Let $\mathcal{C}$ be a nonempty, convex and open subset of $\mathbb{R}^{m\times n}$. Associated with $\mathcal{C}$, a function $h : \mathbb{R}^{m\times n} \to (-\infty, +\infty]$ is called a kernel generating distance if it satisfies the following:

(i) h is proper, lower semicontinuous and convex, with $\operatorname{dom} h \subset \overline{\mathcal{C}}$ and $\operatorname{dom} \partial h = \mathcal{C}$.

(ii) h is C^1 on $\operatorname{int}\operatorname{dom} h \equiv \mathcal{C}$.

We denote by $\mathcal{G}\left(C\right)$ the class of kernel generating distances.

Assumption A1.

(i) $h \in \mathcal{G}(\mathcal{C})$ with $\bar{\mathcal{C}} = \overline{\operatorname{dom} h}$.

(ii) $p : \mathbb{R}^{m\times n} \to (-\infty, +\infty]$ is a proper and lower semicontinuous function with $\operatorname{dom} p \cap \mathcal{C} \neq \varnothing$.

(iii) $f : \mathbb{R}^{m\times n} \to (-\infty, +\infty]$ is a proper and lower semicontinuous function with $\operatorname{dom} h \subset \operatorname{dom} f$, which is C^1 on $\mathcal{C}$.

(iv) $\inf \left\{ f(\mathbf{X}) + p(\mathbf{X}) : \mathbf{X} \in \overline{\mathcal{C}} \right\} > -\infty$.

Assumption A2. The function $h + \left(\frac{1}{\tau}\right) p$ is supercoercive for all $\tau > 0$, that is,

$$\lim_{\|\mathbf{Z}\|_{\mathrm{F}} \to \infty} \frac{h(\mathbf{Z}) + \frac{1}{\tau} p(\mathbf{Z})}{\|\mathbf{Z}\|_{\mathrm{F}}} = \infty. \tag{A2}$$

Assumption A2 is quite a standard coercivity condition for guaranteeing the well-posedness of T_τ, which is stated in the following result.

Lemma A1. (Well-posedness of T_τ) *Suppose that $h + (1/\tau) p$ is supercoercive. Then the set $T_\tau(\mathbf{Z})$ is a nonempty and compact subset of $\mathbb{R}^{m\times n}$.*

Assumption A3. For all $\mathbf{Z} \in \mathcal{C}$, we have $T_\tau(\mathbf{Z}) \subset \mathcal{C}$.

Proposition A1. *Assume that the following assumptions hold:*

(i) (f, h) is L-smooth adaptable on $\mathcal{C}$;

(ii) Assumptions A1, A2 and A3 hold.

Let $\left\{\mathbf{X}^k\right\}_{k\in N}$ be a sequence generated by BPG with $\tau > L$ for solving problem (A1). Then the following assertions hold:

(i) The sequence $F\left(\mathbf{X}^k\right)$ is nonincreasing.

(ii) $\sum_{k=1}^{\infty} D_h\left(\mathbf{X}^k, \mathbf{X}^{k-1}\right) < \infty$ and hence the sequence $\left\{D_h\left(\mathbf{X}^k, \mathbf{X}^{k-1}\right)\right\}_{k\in\mathsf{N}}$ converges to zero.

Consider problem (A1) defined on $\mathcal{C} = \mathbb{R}^{m\times n}$, namely,

$$\inf_{\mathbf{X}} \left\{ f(\mathbf{X}) + p(\mathbf{X}) : \mathbf{X} \in \mathbb{R}^{m\times n} \right\}. \tag{A3}$$

Assumption A4.

(i) $\operatorname{dom} h = \mathbb{R}^{m\times n}$ and h is σ-strongly convex on $\mathbb{R}^{m\times n}$;

(ii) ∇h and ∇f are Lipschitz continuous on any bounded subset of $\mathbb{R}^{m\times n}$.

Let $\eta \in (0, +\infty]$. We denote by Φ_η be the class of all concave and continuous functions $\varphi : [0, \eta) \to \mathbb{R}^+$ which satisfy the following conditions:

(i) $\varphi(0) = 0$;

(ii) φ is C^1 on $(0, \eta)$ and continuous at 0;

(iii) for all $s \in (0, \eta)$:$\varphi'(s) > 0$.

Definition A3. (*Kurdyka–Lojasiewicz property (Bolte* et al.*, 2014)*) Let $F : \mathbb{R}^{m\times n} \to (-\infty, +\infty]$ be proper and lower semicontinuous

(i) $\bar{\mathbf{X}} \in \operatorname{dom} \partial F := \left\{ \mathbf{X} \in \mathbb{R}^{m\times n} : \partial F\left(\bar{\mathbf{X}}\right) \neq 0 \right\}$ is defined as a KL point of function F if there exists a neighborhood $\mathbf{U}$ of $\bar{\mathbf{X}}$, $\eta > 0$ and $\varphi \in \Phi_\eta$ such that for all $\mathbf{X}$ in the following intersection

$$\mathbf{U} \cap \left\{ \mathbf{X} | \, F\left(\bar{\mathbf{X}}\right) < F(\mathbf{X}) < F\left(\bar{\mathbf{X}}\right) + \eta \right\},$$

one has

$$\varphi'\left(F(\mathbf{X}) - F\left(\bar{\mathbf{X}}\right)\right) \operatorname{dist}(0, \partial F(\mathbf{X})) \geq 1. \tag{A4}$$

(ii) If F satisfies the KL property at each point of $\operatorname{dom} \partial F$ then F is called a KL function.

It is easy to establish that the KL property holds in the neighborhood of noncritical points. Thus, the truly relevant aspect of this property is when $\bar{\mathbf{X}}$ is critical, i.e., when $0 \in \partial F\left(\bar{\mathbf{X}}\right)$. In that case it warrants that F is sharpened up to a reparameterization of its values. A remarkable aspect of KL functions is that they are ubiquitous in applications, for example, semi-algebraic, subanalytic and log-exp are KL functions (Bolte *et al.*, 2014).

Theorem A1. (Convergence theorem of BPG) *Assume that the following assumptions hold: (i) (f, h) is L-smooth adaptable on $\mathcal{C}$, (ii) Assumptions A1, A2 and A4 are met. Let $\left\{\mathbf{X}^k\right\}_{k\in N}$ be a sequence generated by BPG which is assumed to be bounded and let $\tau > L$. The following assertions hold:*

*(i) **Subsequential convergence:** Any limit point of the sequence $\left\{\mathbf{X}^k\right\}_{k\in N}$ is a critical point of the objective function.*

*(ii) **Global convergence:** Suppose that the objective function satisfies the KL property on its definition domain. Then, the sequence $\left\{\mathbf{X}^k\right\}_{k\in N}$ has finite length and converges to a critical point of the objective function.*

Assumption A3 is automatically fulfilled since $\mathcal{C} = \mathbb{R}^{m\times n}$.

Appendix B
Adopted notation

Table 1. List of special symbols.

$\operatorname{rank}(\mathbf{X})$	Rank of matrix $\mathbf{X}$
$\sigma(\mathbf{X})$	Singular value vector of matrix $\mathbf{X}$ in descending order
$\\|\mathbf{x}\\|_2$	Vector l_2 norm
$\\|\cdot\\|_{\mathrm{F}}$	Frobenius norm of matrix $\mathbf{X}$
$\mathbf{A} \preceq \mathbf{B}$	$\mathbf{B} - \mathbf{A}$ is a positive semidefinite symmetric matrix
$\\|\mathbf{X}\\|_p$	Schatten p-norm of matrix $\mathbf{X}$
$\\|\mathbf{X}\\|_{\log}$	Sum of logarithms of the singular values of matrix $\mathbf{X}$
$\mathbf{X}_{:,i}$	Column i-th of matrix $\mathbf{X}$
$\operatorname{diag}(\mathbf{X})$	Vector whose coordinates are the diagonal entries of matrix $\mathbf{X}$
$\mathbf{x}^{\downarrow}$	Vector by rearranging the coordinates of $\mathbf{x}$ in descending order
$\mathbf{x} \prec \mathbf{y}$	Vector $\mathbf{x}$ is majorized by vector $\mathbf{y}$
$\lambda(\mathbf{X})$	Vector whose coordinates are the eigenvalues of matrix $\mathbf{X}$ specified in any order
$T_\tau(\mathbf{Z})$	BPG map
$D_h(\mathbf{Y}, \mathbf{Z})$	Bregman distance
$\nabla f(\mathbf{Z})$	Gradient of functional f on $\mathbf{Z}$
$\nabla f_{\mathbf{U}}$	Partial derivative of functional f on $\mathbf{U}$
$\mathcal{S}^n$	Set of $n \times n$ symmetric positive semidefinite matrices
$\mathbf{e}$	All-ones vector
$\mathbf{A} \circ \mathbf{B}$	Element-wise multiplication of matrices
$\operatorname{Diag}(\mathbf{X})$	Diagonal matrix whose diagonal entries are the diagonal entries of matrix $\mathbf{X}$
$\mathbf{I}_n$	Identity matrix $n \times n$

Received: 25 March 2024
Revised: 20 June 2024
Re-revised: 4 July 2024
Accepted: 10 October 2024