Noname manuscript No.
(will be inserted by the editor)

An exact penalty approach for optimization with nonnegative orthogonality constraints

Bo Jiang · Xiang Meng · Zaiwen Wen · Xiaojun Chen

Received: date / Accepted: date

Abstract Optimization with nonnegative orthogonality constraints has wide applications in machine learning and data sciences. It is NP-hard due to some combinatorial properties of the constraints. We first propose an equivalent optimization formulation with nonnegative and multiple spherical constraints and an additional single nonlinear constraint. Various constraint qualifications, the first- and second-order optimality conditions of the equivalent formulation are discussed. By establishing a local error bound of the feasible set, we design a class of (smooth) exact penalty models via keeping the nonnegative and multiple spherical constraints. The penalty models are exact if the penalty parameter is sufficiently large but finite. A practical penalty algorithm with post-processing is then developed to approximately solve a series of subproblems with nonnegative and multiple spherical constraints. We study the asymptotic convergence and establish that any limit point is a weakly stationary point of the original problem and becomes a stationary point under some additional

The work of B. Jiang was supported by the Young Elite Scientists Sponsorship Program by CAST (2017QNRC001), the NSFC grants 11971239 and 11671036. The work of Z. Wen was supported by the NSFC grant 11831002. The work of X. Chen was supported by the Hong Kong Research Grant Council PolyU153001/18P.

Bo Jiang
School of Mathematical Sciences, Key Laboratory for NSLSCS of Jiangsu Province, Nanjing Normal University, CHINA.
E-mail: jiangbo@njnu.edu.cn

Xiang Meng
School of Mathematical Sciences, Peking University, CHINA.
E-mail: 1700010614@pku.edu.cn

Zaiwen Wen
Beijing International Center for Mathematical Research, Peking University, CHINA.
E-mail: wenzw@pku.edu.cn

Xiaojun Chen
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong.
E-mail: xiaojun.chen@polyu.edu.hk

mild conditions. Extensive numerical results on the problem of computing the orthogonal projection onto nonnegative orthogonality constraints, the orthogonal nonnegative matrix factorization problems and the K-indicators model show the effectiveness of our proposed approach.

Keywords exact penalty, nonnegative orthogonality constraint, second-order method, constraint qualification, optimality condition

Mathematics Subject Classification (2010) 65K05 · 90C30 · 90C46 · 90C90

1 Introduction

In this paper, we consider optimization with nonnegative orthogonality constraints:

$$\min_{X\in\mathbb{R}^{n\times k}} f(X) \quad \text{s.t.} \quad X^\top X = I_k,\ X \geq 0, \tag{1.1}$$

where $1 \leq k \leq n$, $f\colon \mathbb{R}^{n\times k} \to \mathbb{R}$ is continuously differentiable, I_k is the k-by-k identity matrix and $X \geq 0$ means the entrywise nonnegativity. The feasible set of (1.1) is denoted as $\mathcal{S}_+^{n,k} := \mathcal{S}^{n,k} \cap \mathbb{R}_+^{n\times k}$, where $\mathcal{S}^{n,k} := \{X \in \mathbb{R}^{n\times k} : X^\top X = I_k\}$ is the Stiefel manifold. In this paper, we adopt the usual Euclidean metric as the Riemannian metric on $\mathcal{S}^{n,k}$. The set $\mathcal{S}_+^{n,k}$ is geodesically convex in $\mathcal{S}^{n,k}$ if $k = 1$ and is geodesically non-convex if $k \geq 2$; see Examples 5.4.1 and 5.4.2 in [1] and Definition 11.2 in [12]. The non-negativity in $\mathcal{S}_+^{n,k}$ destroys the smoothness of $\mathcal{S}^{n,k}$ and introduces some combinatorial features. Specifically, a matrix $X \in \mathcal{S}_+^{n,k}$ means that each row of X has at most one positive element and each column of X takes the unit norm. Problem (1.1) has captured a wide variety of applications and interests, see [7,33,41,59,65] and the references therein.

Due to the combinatorial features, solving (1.1) is generally NP-hard. Actually, problem (1.1) covers some classical NP-hard problems, such as the problem of checking copositivity of a symmetric matrix [30], the quadratic assignment problem and the more general optimization over permutation matrices [33] as special cases. Besides, the constraint $X \in \mathcal{S}_+^{n,k}$ also appears in the k-means clustering [14,18], the min-cut problem [51], etc. Several typical instances of problem (1.1) are briefly reviewed as follows.

1.1 Applications

We mainly introduce three classes of problem (1.1). The first one is the so-called *trace minimization with nonnegative orthogonality constraints*, formulated as

$$\min_{X\in\mathcal{S}_+^{n,k}} \operatorname{tr}(X^\top M X), \tag{1.2}$$

where $M \in \mathbb{R}^{n\times n}$ is symmetric. If $M = -AA^\top$ with $A \in \mathbb{R}^{n\times r}$ being some data matrix, (1.2) is known as *nonnegative principal component analysis* [63].

If $M = D - W$ with W being a similarity matrix corresponding to n objects and D is a diagonal matrix having the same main diagonal as $W\mathbf{e}$, where $\mathbf{e}$ is the all-one vector, (1.2) is known as the *nonnegative Laplacian embedding* [42]. If $M = D - W + \mu R$ with some particularly chosen matrix R and nonnegative regularization parameter μ, (1.2) is known as the *discriminative nonnegative spectral clustering* [60].

The second one is the *orthogonal nonnegative matrix factorization (ONMF)* [26]. Given the data matrix $A \in \mathbb{R}_+^{n\times r}$, ONMF solves

$$\min_{X\in\mathcal{S}_+^{n,k}, Y\in\mathbb{R}_+^{r\times k}} \|A - XY^\top\|_{\mathrm{F}}^2. \tag{1.3}$$

Based on the idea of approximating the data matrix A by its nonnegative subspace projection, Yang and Oja [61] proposed the *orthonormal projective nonnegative matrix factorization (OPNMF)* model as follows:

$$\min_{X\in\mathcal{S}_+^{n,k}} \|A - XX^\top A\|_{\mathrm{F}}^2. \tag{1.4}$$

Models (1.3) and (1.4) are equivalent since the optimal solutions $\bar{X}$ and $\bar{Y}$ of (1.3) satisfy the relation $\bar{Y} = A^\top \bar{X}$. Yang and Oja [61] also proposed a special OPNMF model by replacing the Frobenius norm in (1.4) by the Kullback-Leibler divergence of A and $XX^\top A$. The orthogonal symmetric non-negative matrix factorization models were considered in [37].

The third one is an efficient K-indicators model for data clustering proposed by Chen et al. [20]. Let $U \in \mathcal{S}^{n,k}$ be the features matrix extracted from the data matrix $A \in \mathbb{R}^{n\times r}$, the K-indicators model in [20] reads

$$\min_{X\in\mathcal{S}_+^{n,k}, Y\in\mathcal{S}^{k,k}} \|UY - X\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \|X_{i,:}\|_0 = 1, i \in [n], \tag{1.5}$$

where $\|X_{i,:}\|_0$ is the number of nonzero elements in the i-th row of X, namely, $X_{i,:}$.

1.2 Related works

Optimization on the Stiefel manifold [1,56] has already been well explored. However, a systematic study on problem (1.1) is lacking in the literature albeit it captures many applications. The existing works rarely considered the general problem (1.1), and most of them focused on some special formulations of (1.1). We briefly review some main existing methods. For solving the ONMF model (1.3), motivated by the multiplicative update methods for nonnegative matrix factorization, Ding et al. [26] and Yoo and Choi [62] gave two different multiplicative update schemes. By establishing the equivalence of ONMF with a weighted variant of spherical k-means, Pompoli et al. [50] proposed an EM-like algorithm. Pompoli et al. [50] also designed an augmented Lagrangian method via penalizing the nonnegative constraints but keeping the orthogonality constraints. Li et al. [39] and Wang et al. [54,55] considered the nonconvex

penalty approach by keeping the nonnegative constraints. Some theoretical properties of the nonconvex penalty model were investigated in [55] but the results may not be applied directly to the general problem (1.1). Zhang et al. [66] proposed a greedy orthogonal pivoting algorithm which can promote exact orthogonality. For solving OPNMF model (1.4), Yang and Oja [61] designed a specific multiplicative update method. Pan and Ng [49] introduced a convex relaxation model, wherein the relaxed model is solved by the alternating direction method of multipliers. We remark that the multiplicative update scheme for solving problem (1.3) or (1.4) highly depends on the specific formulation of the objective function, so it is not easy to extend this class of methods to solve the general problem (1.1). In addition, Chen et al. [20] proposed a semi-convex relaxation model and construct a double-layered alternating projection scheme to solve the K-indicators model (1.5). For the case where $k = n$, Wen and Yin [56] designed an augmented Lagrangian method by penalizing the nonnegative constraints but keeping the orthogonality constraints for solving the quadratic assignment problem, and Jiang et al. [33] developed an efficient ℓ_p regularization methods for optimization over permutation matrices.

1.3 Our contribution

Let $[k] := \{1, \ldots, k\}$, r be an arbitrary positive integer and $V \in \mathbb{R}^{k\times r}$ be a constant matrix satisfying

$$\|V\|_{\mathtt{F}} = 1 \quad \text{and} \quad \omega := \min_{i,j\in[k]} [VV^\top]_{ij} > 0. \tag{1.6}$$

By well exploring the structure of $\mathcal{S}_+^{n,k}$, we give a new characterization of $\mathcal{S}_+^{n,k}$ as

$$\mathcal{S}_+^{n,k} = \mathcal{X}_V := \mathcal{OB}_+^{n,k} \cap \{X \in \mathbb{R}^{n\times k} : \|XV\|_{\mathtt{F}} = 1\}, \tag{1.7}$$

where $\mathcal{OB}_+^{n,k} = \{X \in \mathbb{R}^{n\times k} : \|\mathbf{x}_j\| = 1, \mathbf{x}_j \geq 0, j \in [k]\}$, in which $\mathbf{x}_j$ denotes the j-th column of X. Based on this equivalent characterization, a reformulation of problem (1.1) is given as

$$\min_{X\in\mathcal{OB}_+^{n,k}} f(X) \quad \text{s.t.} \quad \|XV\|_{\mathtt{F}} = 1. \tag{1.8}$$

We show that the classical constraint qualifications (CQs) including cone-continuity property (CCP) and Abadie CQ (ACQ) only hold when $\|X\|_0 = n$ while the weakest Guignard CQ (GCQ) always holds. The first- and second-order optimality conditions are also given for problem (1.8). We then explore the relationship between problems (1.1) and (1.8) and show that the two formulations not only share the same minimizers but also the same optimality conditions.

To motivate the exact penalty approach, we prove that a local error bound with exponent 1/2 holds for $\mathcal{S}_+^{n,k}$. Therefore, via keeping the simple constraints $\mathcal{OB}_+^{n,k}$ and penalizing the constraint $\|XV\|_{\mathtt{F}} = 1$, we propose a class of exact penalty models:

$$\min_{X\in\mathcal{OB}_+^{n,k}} \{P_\theta(X) := f(X) + \sigma\left(\zeta_q(X)+\epsilon\right)^p\}, \tag{1.9}$$

where $\zeta_q(X) := \|XV\|_{\mathrm{F}}^q - 1$, $\sigma > 0$ is the penalty parameter and $p, q > 0$ and $\epsilon \geq 0$ are the model parameters. For simplicity of notation, throughout this paper, we use $\theta := \{\sigma, p, q, \epsilon\}$. An important feature of (1.9) is that it allows smooth penalty by choosing appropriate model parameters, such as choosing $p \geq 1$ and $\epsilon = 0$. We show that if the penalty parameter σ is chosen to be larger than a positive constant, the optimal solution of the exact penalty problem (possibly a postprocessing will be invoked) is also optimal for the original problem. A more general exact penalty model (3.9) is also discussed. Then we develop a practical exact penalty algorithm which approximately solves a series of penalty subproblems of the form (1.9) and performs a postprocessing procedure to further improve the solution quality. We study the asymptotic convergence of the penalty algorithm and show that any limit point of the sequence generated by the algorithm is a weakly stationary point of (1.8). We also provide some mild conditions under which the limit point is a stationary point of (1.8). To solve the subproblem (1.9) efficiently, we develop a second-order algorithm for solving optimization over $\mathcal{OB}_+^{n,k}$, which is of independent interest.

The reason of using (1.8) rather than (1.1) is that it can better motivate us to design the exact penalty approach. Simply speaking, the constraints $\mathcal{OB}_+^{n,k}$ in (1.9) is separable with respect to the columns of X and only one simple constraint $\|XV\|_{\mathrm{F}} = 1$ has to be penalized in our approach. The penalty term in (1.9) is general and it reduces to be quadratic if we set $p = 1$, $\epsilon = 0$ and $q = 2$. Traditionally, penalizing the nonnegative constraints but keeping the orthogonality constraints makes the constraints of the subproblem coupled, while penalizing the orthogonality constraints but keeping the nonnegative constraints will introduce quartic or nonsmooth penalty terms. Therefore, our framework is quite different from traditional exact penalty approaches applied to (1.1) directly.

We also discuss how to use the proposed penalty algorithmic framework to solve a two block model

$$\min_{X\in\mathcal{S}_+^{n,k}, Y\in\mathcal{Y}} f(X, Y), \tag{1.10}$$

where $\mathcal{Y}$ is some simple closed set of a finite-dimensional Euclidean space such that the orthogonal projection onto the set $\mathcal{Y}$ is easy to compute. Finally, numerical results on the problem of computing the orthogonal projection onto $\mathcal{S}_+^{n,k}$ and the ONMF model on synthetic data, text clustering, hyperspectral unmixing and the K-indicators model demonstrate the efficiency of our approach.

1.4 Organization

The rest of this paper is organized as follows. A new characterization of $\mathcal{S}_+^{n,k}$ and the equivalent reformulation of problem (1.1) are given in section 2. We propose the exact penalty model in section 3. A practical penalty algorithm together with its convergence results and a second-order method for solving the penalty subproblem is presented in section 4. A variety of numerical results are presented in section 5. Finally, we make some concluding remarks in section 6.

1.5 Notations

For a positive integer n, we denote $[n] := \{1, \ldots, n\}$. The j-th column (resp. i-th row) of a matrix Z with appropriate dimension is denoted by $Z_{:,j}$ (resp. $Z_{i,:}$). For simplicity, we also denote $\mathbf{z}_j := Z_{:,j}$. The number of nonzero elements of Z is $\|Z\|_0$. The Frobenius norm of Z is $\|Z\|_{\mathtt{F}}$ while the 2-norm of a vector z is $\|z\|$. For $z \in \mathbb{R}^n$, $\mathsf{Diag}(z) \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the main diagonal being z. For $Z \in \mathbb{R}^{n\times n}$, $\mathsf{diag}(Z) \in \mathbb{R}^n$ is the main diagonal of Z. For simplicity, we use $\mathsf{Diag}(Z)$ to denote $\mathsf{Diag}(\mathsf{diag}(Z))$. Let $\mathsf{Off}(Z) = Z - \mathsf{Diag}(Z)$. The inner product between two matrices A and B with the same sizes is $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. The notation $0 \le A \perp B \ge 0$ means that $A \ge 0$ and $B \ge 0$ component-wisely and $A \circ B = 0$, where $\circ$ means the Hadamard product operation. Similarly, $\min(A, B)$ takes the minimum of matrices A and B component-wisely. Let $\mathcal{S}$ be a nonempty, closed and possibly nonconvex set of a finite-dimensional Euclidean space $\mathcal{E}$. The orthogonal projection operator is the set-valued mapping $\mathit{\Pi}_{\mathcal{S}} : \mathcal{E} \rightrightarrows \mathcal{E}$ with $\mathit{\Pi}_{\mathcal{S}}(x) = \{u \in \mathcal{S} : \|u - x\|_{\mathcal{E}} = \mathrm{dist}(x, \mathcal{S}) := \min_{y\in\mathcal{S}} \|y - x\|_{\mathcal{E}}\}$, where $\|\cdot\|_{\mathcal{E}}$ is the endowed norm on $\mathcal{E}$. If $\mathcal{S}$ is closed and convex, $\mathit{\Pi}_{\mathcal{S}} : \mathcal{E} \to \mathcal{E}$ becomes a single-valued mapping and we identify $\mathit{\Pi}_{\mathcal{S}}(x) = \mathrm{argmin}_{y\in\mathcal{S}} \|y - x\|_{\mathcal{E}}$.

2 Reformulation of problem (1.1)

Let f^* and $\mathcal{X}^*$ be the optimal value and the optimal solution set of problem (1.1) or (1.8), respectively. We define

$$\mathsf{sgn}(\mathcal{S}_+^{n,k}) := \{H \in \{0,1\}^{n,k} : H = \mathsf{sgn}(X) \text{ with } X \in \mathcal{S}_+^{n,k}\},$$

where $\mathsf{sgn}(X)_{ij} = 1$ if $X_{ij} > 0$ and $\mathsf{sgn}(X)_{ij} = 0$ otherwise. The set $\mathsf{sgn}(\mathcal{X}^*)$ is defined accordingly. For ease of reference, we state a blanket assumption on problem (1.1) or (1.8).

Assumption 1 We assume that $\emptyset \ne \mathsf{sgn}(\mathcal{S}_+^{n,k}) \setminus \mathsf{sgn}(\mathcal{X}^*) := \{H \in \{0,1\}^{n,k} : H \in \mathsf{sgn}(\mathcal{S}_+^{n,k})$ but $H \notin \mathsf{sgn}(\mathcal{X}^*)\}$, namely, the constant $\chi_f := \tilde{f}^* - f^* > 0$ with

$$\tilde{f}^* = \min_{X\in\mathbb{R}^{n\times k}} f(X) \quad \text{s.t.} \quad X \in \mathcal{S}_+^{n,k}, \quad \mathsf{sgn}(X) \in \mathsf{sgn}(\mathcal{S}_+^{n,k}) \setminus \mathsf{sgn}(\mathcal{X}^*).$$

If Assumption 1 does not hold, then $\mathsf{sgn}(\mathcal{S}_+^{n,k}) \setminus \mathsf{sgn}(\mathcal{X}^*) = \emptyset$, which with $\mathsf{sgn}(\mathcal{X}^*) \subseteq \mathsf{sgn}(\mathcal{S}_+^{n,k})$ tells $\mathsf{sgn}(\mathcal{S}_+^{n,k}) = \mathsf{sgn}(\mathcal{X}^*)$. In this case, problem (1.1) or (1.8) is trivial in the sense that any $X \in \mathcal{S}_+^{n,k}$ with $\|X\|_0 = k$ is a global minimizer. However, we can verify that Assumption 1 holds for the test problems in section 5 by randomly choosing some matrices with k nonzero elements in $\mathcal{S}_+^{n,k}$ with different sign matrices and comparing their function values.

For $X \in \mathcal{S}_+^{n,k}$, we define $\mathsf{supp}(X) := \{(i,j) \in [n] \times [k] : X_{ij} \neq 0\}$ and $\Omega_0(X) = \{(i,j) \in [n] \times [k] : X_{ij} = 0\}$. The set $\Omega_0(X)$ is split into two disjoint sets as $\Omega_0'(X) = \{(i,j) \in \Omega_0(X) : \|X_{i,:}\| > 0\}$ and $\Omega_0''(X) = \{(i,j) \in \Omega_0(X) : \|X_{i,:}\| = 0\}$.

We first give an equivalent algebraic characterization of $\mathcal{S}_+^{n,k}$.

Lemma 2.1 *For any $X \in \mathcal{OB}_+^{n,k}$ and $V \in \mathbb{R}^{k\times r}$ satisfying (1.6), there holds that $\|XV\|_{\mathrm{F}} \geq 1$, where the equality holds if and only if $X \in \mathcal{S}_+^{n,k}$. Furthermore, the characterization (1.7) holds.*

Proof With $\|V\|_{\mathrm{F}} = 1$ and $X \in \mathcal{OB}_+^{n,k}$, we have

$$\|XV\|_{\mathrm{F}}^2 - 1 = \langle VV^\top, X^\top X - I_k\rangle = \sum_{i,j\in[k], i\neq j} [VV^\top]_{ij}(\mathbf{x}_i^\top \mathbf{x}_j),$$

which with $VV^\top > 0$ implies that $\|XV\|_{\mathrm{F}}^2 - 1 \geq 0$. The equality holds if and only if $\mathbf{x}_i^\top \mathbf{x}_j = 0$ for $i,j \in [k]$ and $i \neq j$, which with $X \in \mathcal{OB}_+^{n,k}$ means that $X \in \mathcal{S}_+^{n,k}$. Hence (1.7) follows directly. The proof is completed. □

With the equivalent characterization (1.7) of $\mathcal{S}_+^{n,k}$ and Lemma 2.3, we reformulate problem (1.1) as problem (1.8). Throughout this paper, we mainly focus on the formulation (1.8) since it gives us more insight to design our exact penalty approach. We are now going to discuss the CQs and first- and second-order optimality conditions (1.8) and investigate the relationship between the two formulations (1.1) and (1.8).

2.1 Constraint qualifications of problem (1.8)

In this subsection, we investigate several CQs of problem (1.8) which are important to establish the optimality conditions. We mainly consider, GCQ, which is the weakest CQ, ACQ and CCP, which is the weakest strict CQ [3]. Note that the following implications hold: CCP $\Longrightarrow$ ACQ $\Longrightarrow$ GCQ.

We first give the expression of the tangent cone $\mathcal{T}_{\mathcal{X}_V}(X)$ and linearized cone $\mathcal{L}_{\mathcal{X}_V}(X)$ at $X \in \mathcal{X}_V$. Following the definition of linearized cone, we have

$$\mathcal{L}_{\mathcal{X}_V}(X) = \left\{ D \in \mathbb{R}^{n\times k} : \begin{array}{l} \mathbf{x}_j^\top \mathbf{d}_j = 0 \ \forall j \in [k], \\ D_{ij} \geq 0 \ \forall (i,j) \in \Omega_0(X), \langle D, XVV^\top\rangle = 0 \end{array} \right\}. \quad (2.1)$$

With the choice of V and (1.7), $\langle D, XVV^\top\rangle = 0$ tells that $\mathbf{d}_i^\top \sum_{j\in[k]}(VV^\top)_{ji}\mathbf{x}_j = 0$ which further implies that $D_{ij} = 0$ if $(i,j) \in \Omega_0'(X)$. This together with the definitions of $\Omega_0'(X)$ and $\Omega_0''(X)$ and (2.1) yields

$$\mathcal{L}_{\mathcal{X}_V}(X) = \left\{ D \in \mathbb{R}^{n\times k} : \begin{array}{l} \mathbf{x}_j^\top \mathbf{d}_j = 0 \ \forall j \in [k], D_{ij} = 0 \ \forall (i,j) \in \Omega_0'(X), \\ D_{ij} \geq 0 \ \forall (i,j) \in \Omega_0''(X) \end{array} \right\}. \tag{2.2}$$

The tangent cone at X is given as $\mathcal{T}_{\mathcal{X}_V}(X) = \{D \in \mathbb{R}^{n\times k} : \exists \alpha^l > 0, \alpha^l \to 0, D^l \to D$ such that $X^l := X + \alpha^l D^l \in \mathcal{X}_V\}$. Clearly we have $\mathcal{T}_{\mathcal{X}_V}(X) \subseteq \mathcal{L}_{\mathcal{X}_V}(X)$. For each l and fixed $i \in [n]$, there is at most one element of $\{(X^l - X)_{ij} : (i,j) \in \Omega_0''(X)\}$ being nonzero. Hence, any $D \in \mathcal{T}_{\mathcal{X}_V}(X)$ must satisfy $\|D_{i,:}\|_0 \leq 1$ if $X_{i,:} = 0$. On the other hand, for any $D \in \mathcal{L}_{\mathcal{X}_V}(X)$ with $\|D_{i,:}\|_0 \leq 1$ if $X_{i,:} = 0$, choosing $D^l \equiv D$, $\alpha^l = 1/l$ and X^l as $\mathbf{x}_j^l = (\mathbf{x}_j + \alpha^l \mathbf{d}_j)/\|\mathbf{x}_j + \alpha^l \mathbf{d}_j\|$, it is clear that $X^l \in \mathcal{X}_V$. This means that $D \in \mathcal{T}_{\mathcal{X}_V}(X)$. In summary, we have

$$\mathcal{T}_{\mathcal{X}_V}(X) = \mathcal{L}_{\mathcal{X}_V}(X) \cap \left\{ D \in \mathbb{R}^{n\times k} : \|D_{i,:}\|_0 \leq 1 \text{ if } X_{i,:} = 0 \ \forall i \in [n] \right\}. \tag{2.3}$$

We now discuss the CQs in the following lemma.

Lemma 2.2 *Consider a feasible $\bar{X} \in \mathcal{X}_V$ of (1.8). If $k = 1$, then the linear independence constraint qualification (LICQ) holds at $\bar{X}$; if $2 \leq k \leq n$ and $\|\bar{X}\|_0 = n$, then CCP holds; if $2 \leq k \leq n$ and $\|\bar{X}\|_0 < n$, then GCQ holds but ACQ fails to hold.*

Proof Case I. $k = 1$. It is straightforward to check that LICQ holds at $\bar{X}$.

Case II. $2 \leq k \leq n$ and $\|\bar{X}\|_0 = n$, namely, each row of $\bar{X}$ has exactly one positive element. In this case $\Omega_0'(\bar{X}) = \Omega_0(\bar{X})$ and $\Omega_0''(\bar{X}) = \emptyset$. For a sequence $\{X^l\} \subset \mathcal{X}_V$ and $X^l \to \bar{X}$, we consider the closed convex cone (see equation (2.11) in [3] for its definition), which is related to CCP, as

$$\mathcal{K}_{\mathcal{X}_V}(X^l) = \left\{ X^l \mathsf{Diag}(\Lambda) + \lambda X^l VV^\top - \sum_{(i,j)\in\Omega_0(\bar{X})} Z_{ij}\mathbf{E}_{ij} : \Lambda \in \mathbb{R}^k, \lambda \in \mathbb{R}, Z_{ij} \in \mathbb{R}_+ \right\},$$

where $\mathbf{E}_{ij} \in \mathbb{R}^{n\times k}$ with (i,j) element being one while the remaining elements being zeros. Since $\mathcal{X}_V \ni X^l \to \bar{X}$ and $\Omega_0''(\bar{X}) = \emptyset$, we have $\Omega_0(X^l) = \Omega_0(\bar{X})$ and $\mathsf{supp}(X^l) = \mathsf{supp}(\bar{X})$ for sufficiently large l. Thus $X^l VV^\top = X^l \mathsf{Diag}(VV^\top) + \sum_{(i,j)\in\Omega_0(\bar{X})} Z_{ij}\mathbf{E}_{ij}$ for some $Z_{ij} \in \mathbb{R}_+$. By some easy calculations, one has

$$\mathcal{K}_{\mathcal{X}_V}(X^l) = \left\{ X^l \mathsf{Diag}(\Lambda) + \sum_{(i,j)\in\Omega_0(\bar{X})} Z_{ij}\mathbf{E}_{ij} : \Lambda \in \mathbb{R}^k, Z_{ij} \in \mathbb{R} \right\} \tag{2.4}$$

for sufficiently large l, which with $X^l \to \bar{X}$ implies that $\limsup_{X^l\to\bar{X}} \mathcal{K}_{\mathcal{X}_V}(X^l) \subset \mathcal{K}_{\mathcal{X}_V}(\bar{X})$. This means that CCP holds in this case. See Theorem 3.2 in [3].

Case III. $2 \le k \le n$ and $\|\bar{X}\|_0 < n$. In this case, $\Omega_0''(\bar{X}) \neq \emptyset$. By definition, it is easy to verify that the polar cones of $\mathcal{T}(X)$ and $\mathcal{L}(X)$ coincide, namely,

$$\begin{aligned}&\mathcal{T}_{\mathcal{X}_V}(X)^\circ = \mathcal{L}_{\mathcal{X}_V}(X)^\circ\\ &= \left\{ D \in \mathbb{R}^{n\times k} : \begin{array}{l} D_{ij} = \lambda_j X_{ij}, \lambda_j \in \mathbb{R}\ \forall (i,j) \in \mathsf{supp}(X),\\ D_{ij} \le 0\ \forall (i,j) \in \Omega_0''(X) \end{array} \right\}.\end{aligned}$$

This means GCQ holds. Recalling $\Omega_0''(\bar{X}) \neq \emptyset$, we know from (2.2) and (2.3) that $\mathcal{T}_{\mathcal{X}_V}(X) \subsetneq \mathcal{L}_{\mathcal{X}_V}(X)$, which tells that ACQ does not hold. The proof is completed. □

2.2 Optimality conditions of problem (1.8)

Since the oblique manifold $\mathcal{OB}^{n,k} := \{X \in \mathbb{R}^{n\times k} : \|\mathbf{x}_j\| = 1, j \in [k]\}$ is embedded in the Euclidean space $\mathbb{R}^{n\times k}$, we adopt the Riemannian metric on the tangent space as the usual Euclidean metric. The Riemannian gradient and Riemannian Hessian [1] are given as

$$\operatorname{grad} f(X) = \nabla f(X) - X\mathsf{Diag}\big(X^\top \nabla f(X)\big) \tag{2.5}$$

and

$$\langle D_1, \operatorname{Hess} f(X)[D_2]\rangle = \big\langle D_1, \nabla^2 f(X)[D_2]\big\rangle - \big\langle D_1, D_2 \mathsf{Diag}\big(X^\top \nabla f(X)\big)\big\rangle, \tag{2.6}$$

where $\langle\cdot,\cdot\rangle$ is the usual Euclidean inner product, D_1 and D_2 are in the tangent space $\mathcal{T}_{\mathcal{OB}^{n,k}}(X) := \{D \in \mathbb{R}^{n\times k} : \mathbf{x}_j^\top \mathbf{d}_j = 0,\ j \in [k]\}$.

Theorem 2.1 (First-order necessary conditions) *Suppose that $\bar{X} \in \mathcal{X}_V$ is a local minimizer of (1.8). Then $\bar{X}$ is a stationary point of (1.8), namely, $-\nabla f(\bar{X}) \in \mathcal{L}_{\mathcal{X}_V}(\bar{X})^\circ$, which can be further represented as*

$$[\operatorname{grad} f(\bar{X})]_{ij} = 0 \quad \forall (i,j) \in \mathsf{supp}(\bar{X}), \tag{2.7a}$$

$$[\nabla f(\bar{X})]_{ij} \ge 0 \quad \forall (i,j) \in \Omega_0''(\bar{X}). \tag{2.7b}$$

Proof Lemma 2.2 tells that GCQ holds at $\bar{X}$. Thus, $\bar{X}$ must be a stationary point and $-\nabla f(\bar{X}) \in \mathcal{L}_{\mathcal{X}_V}(\bar{X})^\circ$ due to [9, Proposition 3.3.14]. Hence, there exists Lagrange multiplier vector $\bar{\Lambda} \in \mathbb{R}^k$ corresponding to $\|\mathbf{x}_j\| = 1, j \in [k]$, Lagrange multiplier $\bar{\lambda} \in \mathbb{R}$ corresponding to $\|XV\|_{\mathtt{F}} = 1$ and Lagrange multiplier matrix $\bar{Z} \in \mathbb{R}^{n\times k}_+$ corresponding to $X \ge 0$ such that $0 \le \bar{X} \perp \bar{Z} \ge 0$ and $\nabla_X L(\bar{X}, \bar{\Lambda}, \bar{Z}, \bar{\lambda}) = 0$, namely,

$$\nabla f(\bar{X}) - \bar{X}\mathsf{Diag}\big(\bar{\Lambda} - \bar{\lambda}\mathsf{diag}(VV^\top)\big) - \bar{Z} + \bar{\lambda}\bar{X}\mathsf{Off}(VV^\top) = 0. \tag{2.8}$$

Here, the Lagrangian function is given as

$$L(X, \bar{\Lambda}, \bar{Z}, \bar{\lambda}) = f(X) - \sum_{j\in[k]} \bar{\Lambda}_j(\|\mathbf{x}_j\| - 1) - \big\langle \bar{Z}, X\big\rangle + \bar{\lambda}(\|XV\|_{\mathtt{F}} - 1). \tag{2.9}$$

Multiplying $\bar{X}^\top$ on both sides of (2.8) and then performing the $\mathsf{diag}(\cdot)$ operator, with $\bar{X}^\top\bar{X} = I_k$, we have $\bar{\Lambda} - \bar{\lambda}\mathsf{diag}(VV^\top) = \mathsf{diag}(\bar{X}^\top\nabla f(\bar{X}))$, which again with (2.8) and (2.5) implies that $\bar{Z} = \mathrm{grad}\, f(\bar{X}) + \bar{\lambda}\bar{X}\mathsf{Off}(VV^\top)$. Recalling $\bar{X} \in \mathcal{X}_V$, it is easy to verify that $[\bar{X}\mathsf{Off}(VV^\top)]_{ij} = 0$, $\forall (i,j) \in \mathsf{supp}(\bar{X}) \cup \Omega_0''(\bar{X})$, $[\bar{X}\mathsf{Off}(VV^\top)]_{ij} > 0$ $\forall (i,j) \in \Omega_0'(\bar{X})$ and $[\mathrm{grad}\, f(\bar{X})]_{ij} = [\nabla f(\bar{X})]_{ij}$, $\forall (i,j) \in \Omega_0'(\bar{X}) \cup \Omega_0''(\bar{X})$. Hence, we have

$$\bar{Z}_{ij} = \begin{cases} [\mathrm{grad}\, f(\bar{X})]_{ij} & (i,j) \in \mathsf{supp}(\bar{X}), \\ [\nabla f(\bar{X})]_{ij} + \bar{\lambda}[\bar{X}\mathsf{Off}(VV^\top)]_{ij} & (i,j) \in \Omega_0'(\bar{X}), \\ [\nabla f(\bar{X})]_{ij} & (i,j) \in \Omega_0''(\bar{X}). \end{cases} \tag{2.10}$$

Besides, we can always choose $\bar{\lambda} \geq \lambda(\bar{X}) := \max_{(i.j)\in\Omega_0'(\bar{X})} \frac{-[\nabla f(\bar{X})]_{ij}}{[\bar{X}\mathsf{Off}(VV^\top)]_{ij}}$ such that $\bar{Z}_{ij} \geq 0$, $\forall (i,j) \in \Omega_0'(\bar{X})$. Thus we arrive at the equivalent formulation (2.7). □

We borrow the idea from mathematical programs with complementarity constraints, see [2, Definition 2.2] for instance, to define a weakly stationary point $\bar{X}$ of problem (1.8).

Definition 1 We call $\bar{X} \in \mathcal{X}_V$ a weakly stationary point of problem (1.8) if (2.7a) holds at $\bar{X}$.

Note that a weakly stationary point $\bar{X}$ has no requirements on the sign of the Lagrange multiplier corresponding to the constraint $X_{ij} \geq 0$ with $(i,j) \in \Omega_0''(\bar{X})$. Such multiplier is equal to $[\nabla f(\bar{X})]_{ij}$; see the third case in (2.10). Actually, similar to the proof of Theorem 2.1, one can verify that a weakly stationary point $\bar{X}$ is a stationary point of the following relaxed problem

$$\begin{array}{ll} \min\limits_{X\in\mathbb{R}^{n\times k}} & f(X) \\ \text{s.t.} & \|XV\|_{\mathtt{F}} = 1, \|\mathbf{x}_j\| = 1, j \in [k], \\ & X_{ij} \geq 0 \ \forall (i,j) \in \mathsf{supp}(\bar{X}) \cup \Omega_0'(\bar{X}), \\ & X_{ij} = 0 \ \forall (i,j) \in \Omega_0''(\bar{X}). \end{array}$$

Due to page limit, we omit the details here. In the case when $\Omega_0''(\bar{X}) = \emptyset$, namely, $\|\bar{X}\|_0 = n$ or $\nabla f(\bar{X}) \geq 0$ always holds, then the weakly stationary point $\bar{X}$ becomes a stationary point of problem (1.8).

We now assume that f in problem (1.8) is twice continuously differentiable. The set of all sequential null constraint directions at a stationary point $\bar{X}$ (see Definition 8.3.1 in [53]) of problem (1.8) is given as

$$\mathcal{N}_{\mathcal{X}_V}(\bar{X}, \bar{Z}) = \left\{ D \in \mathbb{R}^{n\times k} : \begin{array}{l} X^l := \bar{X} + \alpha^l D^l \in \mathcal{X}_V, \alpha^l > 0, \alpha^l \to 0, D^l \to D, \\ X^l_{ij} = 0 \text{ if } \bar{Z}_{ij} > 0, X^l_{ij} \geq 0 \text{ if } \bar{Z}_{ij} = 0 \end{array} \right\}.$$

Notice that $\mathcal{N}_{\mathcal{X}_V}(\bar{X}, \bar{Z}) \subseteq \mathcal{T}_{\mathcal{X}_V}(\bar{X})$, with (2.3), (2.7) and (2.10), and we have

$$\mathcal{N}_{\mathcal{X}_V}(\bar{X}, \bar{Z}) = \mathcal{T}_{\mathcal{X}_V}(\bar{X}) \cap \mathcal{D}(\bar{X}), \tag{2.11}$$

where $\mathcal{D}(\bar{X}) = \left\{D \in \mathbb{R}^{n\times k} : D_{ij} = 0 \text{ if } [\nabla f(\bar{X})]_{ij} > 0 \ \forall (i,j) \in \Omega_0''(\bar{X})\right\}$. Since $\mathcal{N}_{\mathcal{X}_V}$ is independent of Z, we write $\mathcal{N}_{\mathcal{X}_V}(\bar{X}, \bar{Z})$ as $\mathcal{N}_{\mathcal{X}_V}(\bar{X})$ for short. Similarly, we have the set of all linearized null constraint directions at $\bar{X}$, also known as the critical cone, $\mathcal{C}_{\mathcal{X}_V}(\bar{X}) = \mathcal{L}_{\mathcal{X}_V}(\bar{X}) \cap \{D \in \mathbb{R}^{n\times k} : D_{ij} = 0 \text{ if } \bar{Z}_{ij} > 0, (i,j) \in \Omega_0(\bar{X})\}$. Using (2.2), (2.7) and (2.10), we further have

$$\mathcal{C}_{\mathcal{X}_V}(\bar{X}) = \mathcal{L}_{\mathcal{X}_V}(\bar{X}) \cap \mathcal{D}(\bar{X}). \tag{2.12}$$

We are now ready to establish the second-order optimality conditions as follows.

Theorem 2.2 (Second-order necessary conditions) *If $\bar{X} \in \mathcal{X}_V$ is a local minimizer of problem (1.8), then*

$$\left\langle D, \operatorname{Hess} f(\bar{X})[D]\right\rangle \geq 0, \quad \text{for all } D \in \mathcal{N}_{\mathcal{X}_V}(\bar{X}). \tag{2.13}$$

Proof The proof of Theorem 2.1 tells $\bar{\Lambda} - \bar{\lambda}\mathsf{diag}(VV^\top) = \mathsf{diag}(\bar{X}^\top \nabla f(\bar{X}))$. By [53, Theorem 8.3.3] and the fact that $\bar{X}$ is a local minimizer of problem (1.8), we have from (2.9) and (2.6) that

$$\left\langle D, \nabla^2_{XX} L(\bar{X}, \bar{\Lambda}, \bar{Z}, \bar{\lambda})[D]\right\rangle = \left\langle D, \operatorname{Hess} f(\bar{X})[D] + \bar{\lambda} D\mathsf{Off}(VV^\top)\right\rangle \geq 0, \tag{2.14}$$

for all $D \in \mathcal{N}_{\mathcal{X}_V}(\bar{X})$. For $D \in \mathcal{N}_{\mathcal{X}_V}(\bar{X})$, we know from (2.3) and (2.11) that $D^\top D$ must be diagonal. Thus,

$$\left\langle D, D\mathsf{Off}(VV^\top)\right\rangle = \operatorname{tr}(D^\top D\mathsf{Off}(VV^\top)) = 0, \tag{2.15}$$

which with (2.14) implies (2.13). The proof is completed. □

Theorem 2.3 (Second-order sufficient conditions) *Suppose that $\bar{X} \in \mathcal{X}_V$ is a stationary point of problem (1.8) and that there exists a Lagrange multiplier $\bar{\lambda}$ associated to $\|XV\|_{\mathbb{F}} = 1$ with $\bar{\lambda} \geq \lambda(\bar{X})$ such that*

$$\left\langle D, \operatorname{Hess} f(\bar{X})[D] + \bar{\lambda} D\mathsf{Off}(VV^\top)\right\rangle > 0, \quad \text{for all } D \in \mathcal{C}_{\mathcal{X}_V}(\bar{X}) \setminus \{0\}. \tag{2.16}$$

Then $\bar{X}$ is a strict local minimizer of (1.8).

Proof It follows directly from, for instance [53, Theorems 8.3.4]. □

Remark 2.1 Consider the case when $\Omega_0''(\bar{X}) = \emptyset$, namely, $\|\bar{X}\|_0 = n$. Following from (2.11), (2.12) and $\mathcal{T}_{\mathcal{X}_V}(\bar{X}) = \mathcal{L}_{\mathcal{X}_V}(\bar{X})$, we have $\mathcal{N}_{\mathcal{X}_V}(\bar{X}) = \mathcal{C}_{\mathcal{X}_V}(\bar{X}) = \mathcal{L}_{\mathcal{X}_V}(\bar{X})$. Recalling (2.15), we thus know that (2.13) and (2.16) become $\left\langle D, \operatorname{Hess} f(\bar{X})[D]\right\rangle \geq 0 \ \forall D \in \mathcal{L}_{\mathcal{X}_V}(\bar{X})$ and $\left\langle D, \operatorname{Hess} f(\bar{X})[D]\right\rangle > 0$ $\forall D \in \mathcal{L}_{\mathcal{X}_V}(\bar{X})/\{0\}$, respectively.

2.3 Relationship between problems (1.1) and (1.8)

It is clear that formulations (1.1) and (1.8) share the same minimizers. Moreover, the two problems share the same stationary points.

Lemma 2.3 *(i) The statements in Lemma 2.2 hold for problem (1.1); (ii) Problems (1.1) and (1.8) share the same minimizers and optimality conditions.*

Proof We first claim that that problems (1.1) and (1.8) have the same tangent and linearized cones. Obviously, we know $\mathcal{T}_{\mathcal{S}_+^{n,k}}(X) = \mathcal{T}_{\mathcal{X}_V}(X)$. For the linearized cone, we have

$$\mathcal{L}_{\mathcal{S}_+^{n,k}}(X) = \{D \in \mathbb{R}^{n\times k} : X^\top D + D^\top X = 0, D_{ij} \geq 0 \ \forall (i,j) \in \Omega_0(X)\}.$$

The linear equation above tells that $\mathbf{x}_l^\top \mathbf{d}_j + \mathbf{d}_l^\top \mathbf{x}_j = 0 \ \forall l, j \in [k]$. With $X \in \mathcal{S}_+^{n,k}$ and $D_{ij} \geq 0 \ \forall (i,j) \in \Omega_0(X)$, we further know that $\mathbf{x}_l^\top \mathbf{d}_j \geq 0$. Therefore, we have $\mathbf{x}_l^\top \mathbf{d}_j = 0 \ \forall l, j \in [k]$ and thus $\mathbf{x}_j^\top \mathbf{d}_j = 0 \ \forall j \in [k]$ and $D_{ij} = 0 \ \forall (i,j) \in \Omega_0'(X)$. This means that $\mathcal{L}_{\mathcal{S}_+^{n,k}}(X) \subseteq \mathcal{L}_{\mathcal{X}_V}(X)$. On the other hand, it is easy to see that $D \in \mathcal{L}_{\mathcal{X}_V}(X)$ must imply that $D \in \mathcal{L}_{\mathcal{S}_+^{n,k}}(X)$. Hence, we have $\mathcal{L}_{\mathcal{S}_+^{n,k}}(X) = \mathcal{L}_{\mathcal{X}_V}(X)$. Besides, by some easy calculations, the cones $\mathcal{K}_{\mathcal{S}_+^{n,k}}(X)$ and $\mathcal{K}_{\mathcal{X}_V}(X)$ coincide, see (2.4) for the definition. This completes the proof of (i).

The proof of (ii) can be verified since $\mathcal{T}_{\mathcal{X}_V}(\bar{X}) = \mathcal{T}_{\mathcal{S}_+^{n,k}}(\bar{X})$ and $\mathcal{N}_{\mathcal{S}_+^{n,k}}(\bar{X}) = \mathcal{N}_{\mathcal{X}_V}(\bar{X})$ and $\mathcal{C}_{\mathcal{S}_+^{n,k}}(\bar{X}) = \mathcal{C}_{\mathcal{X}_V}(\bar{X})$. The details are omitted to save space. □

Based on the above lemma, problem (1.1) can be equivalently written as problem (1.8).

3 An exact penalty approach

As pointed by [16], the exact penalty methods are efficient for solving difficulty nonlinear programs especially when the standard CQs are not satisfied, see the references therein for some successful examples.We first present the exact penalty properties. Let X_θ be a global minimizer of (1.9) and denote X_θ^{R} as the matrix returned by Procedure 1 in section 3.1 with an input X_θ. The solution quality can be further improved by solving an auxiliary problem constructed from X_θ^{R} as

$$X_\theta^{\diamondsuit} = \arg \min_{X \in \mathcal{OB}_+^{n,k}} f(X) \quad \text{s.t.} \quad X_{ij} = 0 \text{ if } (i,j) \notin \mathsf{supp}(X_\theta^{\mathsf{R}}). \tag{3.1}$$

Let $L_f \geq 0$ be the Lipschitz constant of f, namely,

$$|f(X_1) - f(X_2)| \leq L_f \|X_1 - X_2\|_{\mathtt{F}}, \quad \forall X_1, X_2 \in \mathcal{OB}_+^{n,k}. \tag{3.2}$$

Such L_f exists since the convex hull of $\mathcal{OB}_+^{n,k}$ is compact.

The next theorem shows that if σ is chosen sufficiently large but finite, the optimal sign matrix can be obtained from X_θ^{R}, thus $X_\theta^{\diamondsuit}$ is also a solution of (1.8).

Theorem 3.1 *Under Assumption 1, we choose*

$$\sigma > \varrho_q^{2p} L_f \nu, \tag{3.3}$$

where $\nu = (\sqrt{2k})^{1-2p}$ *if* $0 < p \leq 1/2$ *and* $\epsilon = 0$, $\nu = (\kappa_f)^{1-2p}$ *if* $p > 1/2$ *and* $\epsilon = 0$, *and* $\nu = \frac{\sqrt{2k}}{(\kappa_f)^{2p} - (\varrho_q \sqrt{\epsilon})^{2p}}$ *if* $p > 0$ *and* $0 < \epsilon < \kappa_f^2/\varrho_q^2$. *Here* $\kappa_f = \chi_f / L_f$ *and the constant* ϱ_q *is defined later in Lemma 3.1. Then it holds that* (i) $\mathsf{sgn}(X_\theta^{\mathsf{R}}) \in \mathsf{sgn}(\mathcal{X}^*)$; (ii) $X_\theta^{\diamondsuit}$ *is a global minimizer of the problem* (1.8).

Remark 3.1 We further explain Theorem 3.1 on the smooth penalty function in (1.9). There are two main ingredients for establishing the exact penalty property: the error bound estimation (3.4) and the "combinatorial nature" of problem (1.8). The latter one is the key and it means that each $X \in \mathcal{S}_+^{n,k}$ corresponds to a unique 0-1 matrix $\mathsf{sgn}(X)$, and the cardinality of $\mathsf{sgn}(\mathcal{S}_+^{n,k})$ is finite. Thus, in nontrivial case, the optimal $\mathsf{sgn}(\mathcal{X}^*)$ and the non-optimal $\mathsf{sgn}(\mathcal{S}_+^{n,k}) \setminus \mathsf{sgn}(\mathcal{X}^*)$ are entirely different in the sense that the minimal f over them have positive gap, namely, $\tilde{f}^* > f^*$, see Assumption 1. Therefore, once we find a matrix $X \in \mathcal{S}_+^{n,k}$ with a function value smaller than $\tilde{f}^*$ but probably still larger than f^*, the job is almost done since one element $\mathsf{sgn}(X) \in \mathsf{sgn}(\mathcal{X}^*)$ is already in hand and a solution of the auxiliary problem (3.1) is enough to recover the optimal solution of (1.8).

Remark 3.2 Consider the case when X_θ is only a local minimizer of the subproblem (1.9). If $X_\theta^\diamond$ is a local minimizer of the subproblem (1.9) or $\|X_\theta^\diamond\|_0 = n$, we can say that $X_\theta^\diamond$ is a local minimizer of the problem (1.8). Otherwise, it is still not clear whether $X_\theta^\diamond$ is a local minimizer of the problem (1.8). However, our practical algorithm, namely, Algorithm 2, can return a local minimizer of the problem (1.8) in finite steps provided that the solution of each subproblem (1.9) is a local minimizer. See Remark 4.3 and Corollary 4.2 for details.

We next investigate the error bound for $\mathcal{S}_+^{n,k}$ in section 3.1, then give the proof of Theorem 3.1 via establishing a class of general exact penalty models in section 3.2.

3.1 Error bound for $\mathcal{S}_+^{n,k}$

It is well known that the error bound plays a key role in establishing the exact penalty results, see [43] for more discussion. By [45, Theorem 16.7], we know that there exist positive scalars ρ and γ such that $\mathrm{dist}(X, \mathcal{S}_+^{n,k}) \leq \rho(\zeta_2(X))^\gamma, \forall X \in \mathcal{OB}_+^{n,k}$. However, the exponent γ is not immediately clear for our case. We next show that the exponent is $\gamma = 1/2$. Our key step is based

on rounding Procedure 1. The basic idea for rounding is simply keeping one largest element in each row and setting the remaining elements to be zeros, and then doing normalization such that each column takes the unit norm. Here, we use the convention that $0/0 = 0$.

Procedure 1: A procedure for rounding $X \in \mathcal{OB}_+^{n,k}$ to be $X^{\mathsf{R}} \in \mathcal{S}_+^{n,k}$.

1 Initialization: Set $H \in \mathbb{R}^{n\times k}$ as a zero matrix.
2 For $i \in [n]$, set $H_{ij^*} = 1$ with j^* is the smallest index in the set $\operatorname{argmax}_{j\in[k]} X_{ij}$.
3 Set the j-th column of X^{R} as $\mathbf{x}_j^{\mathsf{R}} = \frac{\mathbf{x}_j \circ \mathbf{h}_j}{\|\mathbf{x}_j \circ \mathbf{h}_j\|}$, $\quad j \in [k]$.
4 Reset $X^{\mathsf{R}} = I_{n,k}$ if $X^{\mathsf{R}} \notin \mathcal{S}_+^{n,k}$.

Lemma 3.1 *For any $X \in \mathcal{OB}_+^{n,k}$, we have $X^{\mathsf{R}} \in \mathcal{S}_+^{n,k}$ and*

$$\operatorname{dist}(X, \mathcal{S}_+^{n,k}) \le \|X^{\mathsf{R}} - X\|_{\mathtt{F}} \le \varrho_q \sqrt{\zeta_q(X)}, \tag{3.4}$$

where $\varrho_q = (2k\tilde{\varrho}_q/\omega)^{\frac{1}{2}}$ with ω defined in (1.6), and $\tilde{\varrho}_q$ is 1 if $q \ge 2$, and is $\frac{\sqrt{k}+1}{q}$ if $1 \le q < 2$, and is $\frac{2\sqrt{k}(\sqrt{k}+1)}{q(q+1)}$ if $0 < q \le 1$.

Proof We first focus on $q = 2$. Recalling $\|V\|_{\mathtt{F}} = 1$ and $\omega > 0$, we have

$$\zeta_2(X) = \sum_{j\in[k]} \mathbf{x}_j^\top \left(\sum_{l\in[k]\setminus\{j\}} (VV^\top)_{jl} \mathbf{x}_l \right) \ge \omega \sum_{j\in[k]} \left(\mathbf{x}_j^\top \sum_{l\in[k]\setminus\{j\}} \mathbf{x}_l \right). \tag{3.5}$$

The proof of (3.4) for $q = 2$ is split into two cases.

Case I. $\zeta_2(X) \ge \omega$. Since $X^{\mathsf{R}} \in \mathcal{S}_+^{n,k}$ and $X \in \mathcal{OB}_+^{n,k}$, we obtain $\|X^{\mathsf{R}}\|_{\mathtt{F}}^2 = \|X\|_{\mathtt{F}}^2 = k$ and thus $\|X - X^{\mathsf{R}}\|_{\mathtt{F}}^2 \le 2k$. Hence, there holds $\|X - X^{\mathsf{R}}\|_{\mathtt{F}} \le \sqrt{2k} \le \varrho\sqrt{\zeta_2(X)}$.

Case II. $\zeta_2(X) < \omega$. First, we prove that X^{R} generated by Line 3 lies in $\mathcal{S}_+^{n,k}$. Clearly, it follows from Line 2 that each row of H has at most one element being 1. We now claim that each column of H has at least one element being 1. Otherwise, without loss of generality, we assume $\mathbf{h}_1 = 0$. This together with Line 2 implies that $X_{i1} \le \max_{l\in[k]\setminus\{1\}} X_{il}$, $\forall i \in [n]$, which with (3.5) tells that $\zeta_2(X) \ge \omega \sum_{i\in[n]} X_{i1} \max_{l\in[k]\setminus\{j\}} X_{il} \ge \omega \sum_{i\in[n]} X_{i1}^2 = \omega\|\mathbf{x}_1\|^2 = \omega$. This gives a contradiction to $\zeta_2(X) < \omega$. The similar arguments tell that $\mathbf{x}_j \circ \mathbf{h}_j \ne 0$ for $j \in [k]$. In summary, we know that $\|\mathbf{h}_j\|_0 \ge 1, \forall j \in [k]$ and $\mathbf{h}_i^\top \mathbf{h}_j = 0, \forall i, j \in [k]$ and $i \ne j$ and

$$\mathbf{x}_j \circ \mathbf{h}_j \ne 0, \quad (\mathbf{x}_j \circ \mathbf{h}_j)^\top (\mathbf{x}_j \circ (\mathbf{e} - \mathbf{h}_j)) = 0, \quad \forall j \in [k]. \tag{3.6}$$

Therefore, using the construction of X^{R} in Line 3, we must have $X^{\mathsf{R}} \in \mathcal{S}_+^{n,k}$. Using Line 3, (3.6), and the decomposition $\mathbf{x}_j = \mathbf{x}_j \circ \mathbf{h}_j + \mathbf{x}_j \circ (\mathbf{e} - \mathbf{h}_j)$, we

obtain $\|\mathbf{x}_j - \mathbf{x}_j^{\mathsf{R}}\|^2 \le 2\|\mathbf{x}_j \circ (\mathbf{e} - \mathbf{h}_j)\|^2$. With Line 2, we have

$$\|\mathbf{x}_j \circ (\mathbf{e} - \mathbf{h}_j)\|^2 = \sum_{i\in[n], H_{ij}=0} X_{ij}^2 \le \sum_{i\in[n]} X_{ij} \max_{l\in[k]\setminus\{j\}} X_{il} \le \mathbf{x}_j^\top \sum_{l\in[k]\setminus\{j\}} \mathbf{x}_l,$$

which with (3.5) implies $\|X - X^{\mathsf{R}}\|_{\mathsf{F}}^2 = \sum_{j\in[k]} \|\mathbf{x}_j - \mathbf{x}_j^{\mathsf{R}}\|^2 \le 2\sum_{j\in[k]} \|\mathbf{x}_j \circ (\mathbf{e} - \mathbf{h}_j)\|^2 \le 2k\zeta_2(X)/\omega \le \varrho^2\zeta_2(X)$. Combining the above two cases gives (3.4) for $q = 2$.

It is ready to prove (3.4) for general q. For $X \in \mathcal{OB}_+^{n,k}$, there holds that $1 \le \|XV\|_{\mathsf{F}} \le \|X\|_2\|V\|_{\mathsf{F}} \le \sqrt{k}$. We consider three cases. Case I. $q \in [2, +\infty)$. It is easy to have $\zeta_q(X) \ge \zeta_2(X)$. Case II. $q \in [1, 2)$. We first have $\zeta_1(X) = \frac{\zeta_2(X)}{\|XV\|_{\mathsf{F}}+1} \ge \frac{\zeta_2(X)}{\sqrt{k}+1}$. Then we have $\zeta_q(X) = (1 + \zeta_1(X))^q - 1 \ge q\zeta_1(X) \ge \frac{q}{\sqrt{k}+1}\zeta_2(X)$, where the first inequality uses the fact that $(1 + a)^q - 1 > qa$ for $a \in (0, +\infty)$ and $q \in [1, 2)$. Case III. $q \in (0, 1)$. Since $\|XV\|_{\mathsf{F}} = 1 + \zeta_1(X) \ge 1 + \frac{\zeta_1(X)}{\sqrt{k}}$, we have $\zeta_q(X) \ge \left(1 + \frac{\zeta_1(X)}{\sqrt{k}}\right)^q - 1 \ge \frac{q(q+1)}{2\sqrt{k}}\zeta_1(X) \ge \frac{q(q+1)}{2\sqrt{k}(\sqrt{k}+1)}\zeta_2(X)$, where the second inequality uses the fact that $(1 + a)^q - 1 \ge \frac{q(q+1)}{2}a$ for $a \in (0, 1)$, $q \in (0, 1)$. Combining the above three cases, we have $\zeta_2(X) \le \tilde{\varrho}_q\zeta_q(X)$, which with (3.4) for $q = 2$ implies that (3.4) holds for general q. □

We remark that the order 1/2 in the local error bound (3.4) is the best.

Example 3.1 Take $q = 2$ and $V = 1/\sqrt{2}\,[1\ 1]^\top$. Let $0 < \epsilon \ll 1$. Consider $X \in \mathbb{R}^{3\times 2}$ with $X_{11} = X_{22} = \sqrt{1 - \epsilon^2 - \epsilon}$, $X_{12} = X_{21} = \epsilon$, $X_{31} = X_{32} = \sqrt{\epsilon}$ and $\widehat{X} \in \mathbb{R}^{3\times 2}$ with $\widehat{X}_{11} = \sqrt{(1 - \epsilon^2 - \epsilon)/(1 - \epsilon^2)}$, $\widehat{X}_{22} = 1$, $\widehat{X}_{31} = \sqrt{\epsilon/(1 - \epsilon^2)}$, $\widehat{X}_{12} = \widehat{X}_{21} = \widehat{X}_{32} = 0$. It is easy to see that $\widehat{X} \in \Pi_{\mathcal{S}_+^{n,k}}(X)$ and $\mathrm{dist}(X, \mathcal{S}_+^{n,k}) = \|\widehat{X} - X\|_{\mathsf{F}} \approx \sqrt{\epsilon}$ while $\zeta_2(X) = x_1^\top x_2 \approx 3\epsilon$.

3.2 A general exact penalty model

Let $0 \le Q_0 < \kappa_f = \chi_f/L_f$ be a constant and $\Psi : [Q_0, +\infty) \to \mathbb{R}_+$ be strictly increasing. Choose $Q : \mathcal{OB}_+^{n,k} \to \mathbb{R}_+$ such that

$$Q(X) \ge \varrho_q\sqrt{\zeta_q(X)} \quad \forall X \in \mathcal{OB}_+^{n,k}, \tag{3.7a}$$

$$Q(X) \equiv Q_0 \quad \forall X \in \mathcal{S}_+^{n,k}, \quad Q(X) \ge Q_0 \quad \forall X \in \mathcal{OB}_+^{n,k}. \tag{3.7b}$$

Note that (3.7a) and (3.4) imply that

$$Q(X) \ge \|X^{\mathsf{R}} - X\|_{\mathsf{F}} \ge \mathrm{dist}(X, \mathcal{S}_+^{n,k}), \quad \forall X \in \mathcal{OB}_+^{n,k}. \tag{3.8}$$

Our general penalty model, including (1.9) as a special case, is given as

$$\min_{X\in\mathcal{OB}_+^{n,k}} f(X) + \sigma\Psi(Q(X)). \tag{3.9}$$

Let $X_{\sigma,\Psi}$ be a global minimizer of (3.9), and $X_{\sigma,\Psi}^{\mathsf{R}}$ be the matrix returned by Procedure 1 with an input $X_{\sigma,\Psi}$.

Lemma 3.2 *For the penalty model (3.9), it holds*

$$f(X^*) \le f(X^{\mathsf{R}}_{\sigma,\Psi}) \le f(X^*) + L_f \Upsilon_{\sigma,Q_0,\Psi}, \tag{3.10}$$

where X^* is a global minimizer of problem (1.8) and

$$\Upsilon_{\sigma,Q_0,\Psi} := \max_{z\in\mathbb{R}} \ z \text{ s.t. } \Psi(z) \le \Psi(Q_0) + \frac{L_f}{\sigma} z, 0 \le z \le \Psi^{-1}\big(\Psi(Q_0) + \sqrt{2k} L_f/\sigma\big). \tag{3.11}$$

Proof Using the Lipschitz continuity of f in (3.2), we have

$$f(X^{\mathsf{R}}_{\sigma,\Psi}) \le f(X_{\sigma,\Psi}) + L_f \|X^{\mathsf{R}}_{\sigma,\Psi} - X_{\sigma,\Psi}\|_{\mathsf{F}} \le f(X_{\sigma,\Psi}) + L_f Q(X_{\sigma,\Psi}), \tag{3.12}$$

where the second inequality is due to (3.8). By the optimality of $X_{\sigma,\Psi}$, we obtain

$$f(X_{\sigma,\Psi}) + \sigma\Psi\left(Q(X_{\sigma,\Psi})\right) \le f(X) + \sigma\Psi\left(Q(X)\right) = f(X) + \sigma\Psi(Q_0) \quad \forall X \in \mathcal{X}_V. \tag{3.13}$$

Taking $X = X^*$ in (3.13) and using the strictly increasing property of Ψ and (3.7b), we have $f(X_{\sigma,\Psi}) \le f(X^*)$. Hence, we know from (3.12) that

$$f(X^*) \le f(X^{\mathsf{R}}_{\sigma,\Psi}) \le f(X^*) + L_f Q(X_{\sigma,\Psi}). \tag{3.14}$$

The remaining is to estimate $Q(X_{\sigma,\Psi})$. Taking X to be $X^{\mathsf{R}}_{\sigma,\Psi}$ in (3.13), we get

$$\Psi\left(Q(X_{\sigma,\Psi})\right) \le \Psi(Q_0) + \frac{f(X^{\mathsf{R}}_{\sigma,\Psi}) - f(X_{\sigma,\Psi})}{\sigma} \le \Psi(Q_0) + \frac{L_f \|X^{\mathsf{R}}_{\sigma,\Psi} - X_{\sigma,\Psi}\|_{\mathsf{F}}}{\sigma}, \tag{3.15}$$

where the second inequality is due to (3.2). Since $X_{\sigma,\Psi} \in \mathcal{OB}^{n,k}_{+}$, it is easy to see that $\|X^{\mathsf{R}}_{\sigma,\Psi} - X_{\sigma,\Psi}\|_{\mathsf{F}} \le \sqrt{2k}$. Thus, we have from (3.15) that $\Psi\left(Q(X_{\sigma,\Psi})\right) \le \Psi(Q_0) + \sqrt{2k} L_f/\sigma$. Since Ψ is strictly increasing, we obtain

$$Q(X_{\sigma,\Psi}) \le \Psi^{-1}\left(\Psi(Q_0) + \sqrt{2k} L_f/\sigma\right). \tag{3.16}$$

On the other hand, recalling (3.8), we have from (3.15) that $\Psi\left(Q(X_{\sigma,\Psi})\right) \le \Psi(Q_0) + \frac{L_f}{\sigma} Q(X_{\sigma,\Psi})$, which together with (3.16) and (3.14) establishes (3.10). The proof is completed. □

Let $X^{\diamondsuit}_{\sigma,\Psi}$ be a global minimizer of the problem (3.1) with $X^{\mathsf{R}}_{\theta} = X^{\mathsf{R}}_{\sigma,\Psi}$. We now have the following exact penalty property.

Theorem 3.2 *Suppose Assumption 1 holds and $\sigma > 0$ is chosen such that*

$$\Upsilon_{\sigma,Q_0,\Psi} < \kappa_f. \tag{3.17}$$

Then it holds that (i) $\mathsf{sgn}(X^{\mathsf{R}}_{\sigma,\Psi}) \in \mathsf{sgn}(\mathcal{X}^*)$; (ii) $X^{\diamondsuit}_{\sigma,\Psi}$ *is a global minimizer of problem (1.8), namely, $f(X^{\diamondsuit}_{\sigma,\Psi}) = f(X^*)$.*

Proof We first claim that $\mathsf{sgn}(X^{\mathsf{R}}_{\sigma,\Psi}) \in \mathsf{sgn}(\mathcal{X}^*)$. Otherwise, it follows from Assumption 1 that $f(X^{\mathsf{R}}_{\sigma,\Psi}) \geq f(X^*) + \chi_f$. By using (3.10) and $\kappa_f = \chi_f/L_f$, we thus have $\Upsilon_{\sigma,Q_0,\Psi} \geq \kappa_f$, which makes a contradiction to (3.17). Using $\mathsf{sgn}(X^{\mathsf{R}}_{\sigma,\Psi}) \in \mathsf{sgn}(\mathcal{X}^*)$ and the definition of $X^{\diamondsuit}_{\sigma,\Psi}$, see problem (3.1) with X^{R}_{θ} being $X^{\mathsf{R}}_{\sigma,\Psi}$, we know that $X^{\diamondsuit}_{\sigma,\Psi}$ is a global minimizer of problem (1.8). The proof is completed. □

It follows from (3.11) that $\Upsilon_{\sigma,Q_0,\Psi} \leq \Psi^{-1}\big(\Psi(Q_0) + \sqrt{2k}L_f/\sigma\big)$. To make (3.17) hold, we can choose $0 \leq Q_0 < \kappa_f$ and $\sigma > \sqrt{2k}L_f\big(\Psi(\kappa_f) - \Psi(Q_0)\big)^{-1}$. For some particular $\Psi(\cdot)$, we next show that this lower bound can be improved.

Proof of Theorem 3.1 Let us choose $Q(X) = \varrho_q\sqrt{\zeta_q(X) + \epsilon}$ and $\Psi(z) = (z/\varrho_q)^{2p}$ with $0 \leq \epsilon < \kappa_f^2/\varrho_q^2$ and $Q_0 = \varrho_q\sqrt{\epsilon}$. By Theorem 3.2, we only need to prove $\Upsilon_{\sigma,Q_0,\Psi} < \kappa_f$ if $\sigma > \varrho_q^{2p}L_f\nu$. We consider three cases.

Case I. $\epsilon = 0$ and $0 < p \leq 1/2$. Since $\sigma > \varrho_q^{2p}L_f\nu = (\sqrt{2k})^{1-2p}\varrho_q^{2p}L_f$, we have from $\Psi(z) \leq \Psi(Q_0) + \frac{L_f}{\sigma}z$ that $z = 0$ or $z > \sqrt{2k}$ and have from $0 \leq z \leq \Psi^{-1}(\Psi(Q_0) + \sqrt{2k}L_f/\sigma)$ that $0 \leq z < \sqrt{2k}$. By definition (3.11), we have $\Upsilon_{\sigma,Q_0,\Psi} = 0 < \kappa_f$.

Case II. $\epsilon = 0$ and $p > 1/2$. Using the definition of χ_f in Assumption 1, (3.2) and $\|X - Y\|_{\mathrm{F}} \leq \sqrt{2k}$ for $X, Y \in \mathcal{OB}^{n,k}_{+}$, we have $\chi_f \leq \sqrt{2k}L_f$, i.e., $\kappa_f \leq \sqrt{2k}$. By $\sigma > \varrho_q^{2p}L_f\nu = (\kappa_f)^{1-2p}\varrho_q^{2p}L_f$, it is easy to obtain from (3.11) that $\Upsilon_{\sigma,Q_0,\Psi} < \kappa_f$.

Case III. $0 < \epsilon < \kappa_f^2/\varrho_q^2$, $p > 0$. We thus have $\varrho_q^{2p}L_f\nu = \sqrt{2k}L_f\big(\Psi(\kappa_f) - \Psi(Q_0)\big)^{-1}$. Thus, we have $\Upsilon_{\sigma,Q_0,\Psi} < \kappa_f$ by (3.11). The proof is completed. □

Remark 3.3 The threshold $\varrho_q^{2p}L_f\nu$ of the exact penalty parameter depends on the parameter $\kappa_f = \chi_f/L_f$, except that for model (1.9) with $p \in (0, 1/2]$ and $\epsilon = 0$, $\varrho_q^{2p}L_f\nu$ is independent of κ_f but the corresponding penalty term is nonsmooth. Usually, estimating L_f is possible for the instances such as the orthogonal nonnegative matrix factorization models (1.3) and (1.4). However, computing χ_f is hard since it needs to know f^* in advance and solves an optimization problem. In fact, calculating a threshold of the exact penalty parameter is not always easy, see [8,44,28,21] for some convex and nonconvex examples. In practice, we simply solve approximately a series of problems (1.9) with an increasing σ; see section 4 for a detailed description.

A few more remarks on the exact penalty model (1.9) are listed in order. (i) To make the objective function in problem (1.9) smooth, we need to choose $\epsilon \in (0, \kappa_f^2/\varrho_q^2)$ for $p \in (0, 1)$. As for $p \in [1, +\infty)$, we can simply choose $\epsilon = 0$. (ii) By directly using the results in [21, Lemma 3.1], we can show that a global minimizer of (1.9) with $p = 1/2$ and $\epsilon = 0$ is also a global minimizer of (1.8) under the condition that $\sigma > \varrho_q L_f$. However, the results therein do not apply to the general $\Psi(\cdot)$ and Q_0. By contrast, our results in Theorem 3.2 or Theorem 3.1 allow more flexible choices of $\Psi(\cdot)$ and Q_0 or p and ϵ. (iii) The multiple spherical constraints in model (1.9) are not only important

to establish the exact penalty property but also make model (1.9) working over a compact set. It should be mentioned that for a variant of ONMF problem (1.3), [55] proposed an exact penalty model without keeping the multiple spherical constraints. However, their results only hold on this special formulation (1.3) rather than the general problem (1.1). Besides, Gao et al. [29] used a customized augmented Lagrangian type method to solve optimization with orthogonality constraints but without the nonnegative constraints. The multiple spherical constraints are also kept therein to make their method more robust. However, their method can not be directly used to solve problem (1.1) or (1.8) due to the nonnegative constraints, which make the problem totally different.

4 A practical exact penalty algorithm

We now focus on the exact penalty model (1.9). A practical exact penalty method, named as EP4Orth+, is presented in Algorithm 2. We adopt the way in [21] to choose a feasible initial point X^{feas}. In each iteration, the penalty parameter σ is dynamically increased and we find an approximate stationary point X^t satisfying the approximate first-order optimality condition (4.1) and the sufficient descent condition (4.2). Such X^t can be found in a finite number of iterations; see section 4.1 for more discussion. To obtain an exact orthogonal nonnegative matrix and improve the solution quality, we perform a postprocessing procedure at the end of the algorithm.

Algorithm 2: EP4Orth+: A practical exact penalty method for solving (1.8)

Initialization: Choose an initial point $X^0 \in \mathcal{OB}_+^{n,k}$, $X^{\text{feas}} \in \mathcal{S}_+^{n,k}$ and $p, q > 0$. Choose a positive integer $t_{\max}$ and $\text{tol}^{\text{feas}}, \varepsilon_0^{\text{grad}}, \varepsilon_{\min}^{\text{grad}} \in [0, 1)$, $\sigma_0 > 0$ and $\gamma_2 > 1$. Choose $\epsilon_0 > 0$, $\gamma_1 = \gamma_2^{-1/p}$ if $p \in (0, 1)$ and set $\epsilon_0 = 0$, $\gamma_1 = 1$ if $p \geq 1$. Choose $\eta \in (0, \gamma_2^{-1/p})$ and set $X^{0,0} = X^0$.

for $t = 0, 1, 2, \ldots, t_{\max}$ **do**

  If $P_{\sigma_t,p,q,\epsilon_t}(X^{t,0}) > P_{\sigma_t,p,q,\epsilon_t}(X^{\text{feas}})$, set $X^{t,0} = X^{\text{feas}}$.

  Starting from $X^{t,0}$, we find an approximate stationary point X^t of (1.9) with $\theta = \theta_t := \{\sigma_t, p, q, \epsilon_t\}$ such that

$$\|\min(X^t, \operatorname{grad} P_{\theta_t}(X^t))\|_{\text{F}} \leq \varepsilon_t^{\text{grad}}, \tag{4.1}$$

$$P_{\theta_t}(X^t) \leq P_{\theta_t}(X^{t,0}). \tag{4.2}$$

  if $\|X^t V\|_{\text{F}}^2 - 1 \leq \text{tol}^{\text{feas}}$ **then**

    break

  end

  Set $\epsilon_{t+1} = \gamma_1 \epsilon_t$, $\sigma_{t+1} = \gamma_2 \sigma_t$, $\varepsilon_{t+1}^{\text{grad}} = \max\{\eta \varepsilon_t^{\text{grad}}, \varepsilon_{\min}^{\text{grad}}\}$ and $X^{t+1,0} = X^t$.

end

Set $X^{\text{R}} = (X^t)^{\text{R}}$ using Procedure 1 and solve (3.1) approximately with $X_\theta^{\text{R}} = X^{\text{R}}$ to get $X^\diamondsuit$ such that $f(X^\diamondsuit) \leq f(X^{\text{R}})$. `// Postprocessing`

It should be mentioned that the postprocessing procedure is always easy to perform. Consider a separable function $f(X) = \sum_{j=1}^{k} f_j(\mathbf{x}_j)$, where $f_j : \mathbb{R}^n \to \mathbb{R}$. Let $\mathbf{x}_j^{\mathsf{R}}$ be the j-th column of X^{R}. The corresponding problem (3.1) is split to the form of

$$\min_{\mathbf{x}_j \in \mathbb{R}^n} \ f_j(\mathbf{x}_j) \quad \text{s.t.} \quad \mathbf{x}_j^\top \mathbf{x}_j = 1, \mathbf{x}_j \geq 0, (\mathbf{x}_j)_i = 0 \text{ if } i \notin \mathsf{supp}(\mathbf{x}_j^{\mathsf{R}}).$$

(i) If $f(X)$ is $-\langle C, X\rangle$ as in the K-indicators model (1.5) with fixed Y, the j-th column of the global minimizer $X^\diamondsuit$ is $\mathbf{x}_j^\diamondsuit = \Pi_{\mathcal{OB}_+^{n,k}}(\mathbf{c}_j \circ \mathbf{h}_j^{\mathsf{R}})$, where $\mathbf{h}_j^{\mathsf{R}}$ is the j-th column of $\mathsf{sgn}(X^{\mathsf{R}})$. (ii) Consider $f(X) = -\mathrm{tr}(X^\top M X)$ with $M = M^\top \geq 0$ in the ONMF models (1.3) and (1.4). Then $(\mathbf{x}_j^\diamondsuit)_i = 0$ if $i \notin \mathsf{supp}(\mathbf{x}_j^{\mathsf{R}})$ and $(\mathbf{x}_j^\diamondsuit)_{\mathsf{supp}(\mathbf{x}_j^{\mathsf{R}})}$ is the dominant eigenvector of M_j, which is a principal submatrix of M whose rows and columns indices are both $\mathsf{supp}(\mathbf{x}_j^{\mathsf{R}})$. Since $M_j \geq 0$, there always exists a nonnegative dominant eigenvector due to Perron-Frobenius theorem, see [19, Theorem 1.1]. When f is a general smooth function, we can use the nonconvex gradient projection method to get an approximate stationary point $X^\diamondsuit$ with $f(X^\diamondsuit) \leq f(X^{\mathsf{R}})$.

One can also solve (1.8) by the augmented Lagrangian method, and the objective function in the subproblem (1.9) becomes $L_{\sigma,\lambda}(X) := f(X) + \lambda(\|XV\|_{\mathtt{F}} - 1) + \sigma(\|XV\|_{\mathtt{F}} - 1)^2$ and the Lagrange multiplier is updated by $\lambda^{t+1} = \lambda^t + \sigma(\|X^t V\|_{\mathtt{F}} - 1)$. Based on our preliminary numerical tests, we find that it has a similar performance to our Algorithm 2 and sometimes the augmented Lagrangian method can return a feasible solution quickly but with worse function values. Therefore, we focus on the exact penalty approach in this paper. More interestingly, the augumented Lagrangian function $L_{\sigma,\lambda}(X)$ has some close connections with our penalty approach. Note that we can always choose a nonnegative Lagrange multiplier because $\|XV\|_{\mathtt{F}} = 1$ is equivalent to $\|XV\|_{\mathtt{F}} \leq 1$ for $X \in \mathcal{OB}_+^{n,k}$; see Lemma 2.1. Choosing $\lambda = 2\sigma > 0$, then $L_{\sigma,\lambda}(X)$ becomes $P_\theta(X)$ in (1.9) with $q = 2$ and $p = 1$; choosing a fixed $\lambda \geq 0$, then $L_{\sigma,\lambda}(X)$ falls into a special instance of (3.9) with $Q(X) = \varrho_q\sqrt{\|XV\|_{\mathtt{F}} - 1}$ and $\Psi(t) = \lambda(t/\varrho_q)^2 + \sigma(t/\varrho_q)^4$, and the threshold penalty value in Theorem 3.2 can be estimated accordingly.

We next discuss how to solve the subproblem (1.9) in section 4.1 and then give the convergence analysis of Algorithm 2 in section 4.2.

4.1 Optimization over $\mathcal{OB}_+^{n,k}$

The penalty subproblem (1.9) with suitable choices of parameters is a special instance of optimization over $\mathcal{OB}_+^{n,k}$, namely,

$$\min_{X \in \mathcal{OB}_+^{n,k}} \ h(X), \tag{4.3}$$

where $h\colon \mathbb{R}^{n\times k} \to \mathbb{R}$ is continuously differentiable. Note that LICQ holds at any $X \in \mathcal{OB}_+^{n,k}$, namely, X satisfying the constraints $\|\mathbf{x}_j\| = 1, \mathbf{x}_j \geq 0, j \in [k]$.

Similar to the discussion in section 2.2, we establish the optimality conditions for (4.3). To save the space, we omit some details here.

Theorem 4.1 (First-order necessary conditions) *If $\bar{X} \in \mathcal{OB}_+^{n,k}$ is a local minimizer of problem (4.3), then $\bar{X}$ is a stationary point, namely, $0 \leq \bar{X} \perp \operatorname{grad} h(\bar{X}) \geq 0$, which is equivalent to $\min(\bar{X}, \operatorname{grad} h(\bar{X})) = 0$.*

Remark 4.1 The two first-order necessary conditions in Theorem 4.1 are further equivalent to

$$\Pi_{\mathcal{L}_{\mathcal{OB}_+^{n,k}}(\bar{X})}(-\operatorname{grad} h(\bar{X})) = \Pi_{\mathbf{T}(\bar{X})}(-\operatorname{grad} h(\bar{X})) = 0,$$

where $\mathcal{L}_{\mathcal{OB}_+^{n,k}}(\bar{X}) = \{D \in \mathbb{R}^{n\times k} : \bar{\mathbf{x}}_j^\top \mathbf{d}_j = 0,\ D_{ij} \geq 0 \text{ if } \bar{X}_{ij} = 0\}$ and $\mathbf{T}(\bar{X}) = \{D \in \mathbb{R}^{n\times k} : \bar{\mathbf{x}}_j^\top \mathbf{d}_j = 0,\ \bar{\mathbf{x}}_j + \mathbf{d}_j \geq 0\}$.

Theorem 4.2 (Second-order necessary and sufficient conditions)

i) If $\bar{X}$ is a local minimizer of problem (4.3) then $\langle D, \operatorname{Hess} h(\bar{X})[D]\rangle \geq 0,\ \forall D \in \mathcal{C}_{\mathcal{OB}_+^{n,k}}(\bar{X})$, where $\operatorname{Hess} h(\bar{X})[D]$ is obtained by specializing (2.6) to $h(\bar{X})$ and the critical cone $\mathcal{C}_{\mathcal{OB}_+^{n,k}}(\bar{X}) = \mathcal{L}_{\mathcal{OB}_+^{n,k}}(\bar{X}) \cap \{D \in \mathbb{R}^{n\times k} : D_{ij} = 0 \text{ if } [\nabla f(\bar{X})]_{ij} > 0 \text{ and } \bar{X}_{ij} = 0\}$.

ii) If $\bar{X}$ is a stationary point of (4.3) and $\langle D, \operatorname{Hess} h(\bar{X})[D]\rangle > 0,\ \forall D \in \mathcal{C}_{\mathcal{OB}_+^{n,k}}(\bar{X})/\{0\}$, then $\bar{X}$ is a strict local minimizer of (4.3).

We first introduce a first-order nonconvex gradient projection method:

$$X^{l+1} \in \Pi_{\mathcal{OB}_+^{n,k}}(X^l - \alpha^l \nabla h(X^l)), \quad \alpha^l > 0, \tag{4.4}$$

where the stepsize α^l can be determined by either monotone or non-monotone linesearch, see [64] and the references therein. Note that $\Pi_{\mathcal{OB}_+^{n,k}}(\cdot)$ is explicitly available and can be computed in $O(nk)$ flops, one can refer [65] for instance. Since $\mathcal{OB}_+^{n,k}$ is compact, by Theorem 3.1 in [58], any limit point of the sequence $\{X^l\}$ generated by (4.4) is a stationary point of (4.3). If further h is a KL function, by Theorem 5.3 in [5], the sequence $\{X^l\}$ converges to a stationary point of (4.3) for $\alpha^l \in (\underline{\alpha}, \frac{1}{L_h} - \underline{\alpha})$ where $\underline{\alpha} \in (0, \frac{1}{2L_h})$ and L_h is the Lipschitz constant of ∇h on $\mathbb{R}^{n\times k}$.

We next adopt the adaptive quadratically regularized Newton method [31, 32] to solve (4.3) in order to accelerate the convergence of the first-order nonconvex gradient projection method. At the l-th iteration, we construct a quadratically regularized subproblem as

$$\min_{X \in \mathcal{OB}_+^{n,k}} m_l(X), \tag{4.5}$$

where $m_l(X) := \langle \nabla h(X^l), X - X^l\rangle + \frac{1}{2}\langle X - X^l, \nabla^2 h(X^l)[X - X^l]\rangle + \frac{\tau_l}{2}\|X - X^l\|_{\mathbf{F}}^2$. It holds $\operatorname{grad} m_l(X^l) = \operatorname{grad} h(X^l)$ and $\operatorname{Hess} m_l(X^l)[D] = \operatorname{Hess} h(X^l)[D]$

$+\tau_l D$, $\forall D \in \mathcal{T}_{\mathcal{OB}^{n,k}}(X^l)$. Since $\mathcal{OB}_+^{n,k}$ is compact, there exist positive constants κ_g and κ_H such that $\|\nabla h(X)\|_{\mathrm{F}} \le \kappa_g$ and $\|\nabla^2 h(X)\| \le \kappa_H$, $\forall X \in \mathcal{OB}_+^{n,k}$. Instead of solving the subproblem (4.5) exactly, motivated by condition (3.1) in [31], we only compute an approximate solution $Y^l \in \mathcal{OB}_+^{n,k}$ of (4.5) satisfying

$$m_l(Y^l) \le -\frac{a}{\kappa_H + \kappa_g + \tau_l} \|\Pi_{\mathbf{T}(X^l)}(-\operatorname{grad} h(X^l))\|_{\mathrm{F}}^2, \tag{4.6}$$

where a is a positive constant. Then, we calculate the ratio ρ_l between the predicted reduction and the actual reduction to determine whether the trial point Y^l is accepted or not. The complete algorithm is presented in Algorithm 3. By

Algorithm 3: An adaptive regularized Newton method for (4.3)

Initialization: Choose $X^0 \in \mathcal{OB}_+^{n,k}$, a tolerance $\epsilon > 0$ and an initial regularization parameter $\tau^0 > 0$. Choose $0 < \eta_1 \le \eta_2 < 1$, $0 < \beta_0 < 1 < \beta_1 < \beta_2$. Set $l := 0$.
while $\|\min(X^l, \operatorname{grad} h(X^l))\|_{\mathrm{F}} > \epsilon$ **do**
  Solve (4.5) (by, e.g., Algorithm 4) to obtain a trial point Y^l satisfying (4.6).
  Calculate $\rho_l = \big(h(Y^l) - h(X^l)\big)/m_l(Y^l)$.
  Set $X^{l+1} := Y^l$ if $\rho_l \ge \eta_1$ and set $X^{l+1} := X^l$ otherwise.
  Choose $\tau_{l+1} \in (0, \beta_0\tau_l]$ if $\rho_l \ge \eta_2$; choose $\tau_{l+1} \in [\beta_0\tau_l, \beta_1\tau_l]$ if $\eta_1 \le \rho_l \le \eta_2$; choose $\tau_{l+1} \in [\beta_1\tau_l, \beta_2\tau_l]$ otherwise.
  Set $l := l + 1$.
end

following the proof of [31, Theorem 4], we can establish $\|\Pi_{\mathbf{T}(X^l)}(-\operatorname{grad} h(X^l))\|_{\mathrm{F}} = 0$ for some $l > 0$ or $\lim_{l\to\infty} \|\Pi_{\mathbf{T}(X^l)}(-\operatorname{grad} h(X^l))\|_{\mathrm{F}} = 0$.

We now show that the inexact condition (4.6) is well defined. Let c_1 be a positive constant. For a direction $D^l \in \mathbf{T}(X^l)$ satisfying

$$\langle \operatorname{grad} h(X^l), D^l \rangle \le -c_1 \|\Pi_{\mathbf{T}(X^l)}(-\operatorname{grad} h(X^l))\|_{\mathrm{F}} \|D^l\|_{\mathrm{F}}, \tag{4.7}$$

we compute

$$Y^l = \Pi_{\mathcal{OB}_+^{n,k}}\left(X^l + \alpha_l D^l\right) \quad \text{with} \quad \alpha_l = \frac{2c_1(1-c_2)\|\Pi_{\mathbf{T}(X^l)}(-\operatorname{grad} h(X^l))\|_{\mathrm{F}}}{(\kappa_g + \kappa_H + \tau_l)\|D\|_{\mathrm{F}}}, \tag{4.8}$$

where $c_2 \in (0,1)$ is a constant. For any $D \in \mathbf{T}(X)$, it is easy to verify that $\langle \nabla h(X), D\rangle = \langle \operatorname{grad} h(X), D\rangle$ and

$$\|\Pi_{\mathcal{OB}_+^{n,k}}(X+D) - X\|_{\mathrm{F}} \le \|D\|_{\mathrm{F}}, \quad \|\Pi_{\mathcal{OB}_+^{n,k}}(X+D) - X - D\|_{\mathrm{F}} \le \frac{1}{2}\|D\|_{\mathrm{F}}^2. \tag{4.9}$$

It follows from the property (4.9) and the arguments in [13, Lemma 2.10] that such Y^l satisfies (4.6) with $a = 2c_1^2c_2(1-c_2)$. For sake of saving space, we omit the tedious details here.

We give two particular choices of D^l satisfying (4.7). The first one is a single projected gradient step $D^l = \Pi_{\mathbf{T}(X^l)}(-\operatorname{grad} h(X^l))$. Then (4.7) holds with $c_1 = 1$. The second one is from the Newton subproblem motived by [32]:

$$\min_{D\in\mathbf{T}(X^l)} \left\langle \operatorname{grad} m_l(X^l), D\right\rangle + \frac{1}{2}\left\langle D, \operatorname{Hess} m_l(X^l)[D]\right\rangle. \tag{4.10}$$

Setting $D = Z - X^l$ reformulates problem (4.10) as

$$\min_{Z\in\Delta(X^l)} \left\langle \operatorname{grad} m_l(X^l), Z - X^l\right\rangle + \frac{1}{2}\left\langle Z - X^l, \operatorname{Hess} m_l(X^l)[Z - X^l]\right\rangle, \tag{4.11}$$

where $\Delta(X^l) := \{Z \in \mathbb{R}^{n\times k} : (\mathbf{x}^l)_j^\top \mathbf{z}_j = 1, \mathbf{z}_j \geq 0, j \in [k]\}$. The first-order optimality condition is the following nonsmooth equation:

$$\mathcal{F}(Z) := Z - \Pi_{\Delta(X^l)}\big(Z - \alpha\big(\operatorname{grad} m_l(X^l) + \operatorname{Hess} m_l(X^l)[Z - X^l]\big)\big) = 0,$$

where $\alpha > 0$ is a constant. Denote $C := Z - \alpha\big(\operatorname{grad} m_l(X^l) + \operatorname{Hess} m_l(X^l)[Z - X^l]\big)$ for simplicity. Thanks to [40], we can efficiently compute the the HS generalized Jacobian $\mathcal{P}_C(\cdot)$ of $\Pi_{\Delta(X)}(\cdot)$ efficiently. Define a linear operator $\Xi : \mathbb{R}^{n\times k} \to \mathbb{R}^{n\times k}$ by $[\Xi(H)]_{ij} = 0$ if $[\Pi_{\Delta(X^l)}(C)]_{ij} = 0$ and $[\Xi(H)]_{ij} = H_{ij}$ otherwise for any $H \in \mathbb{R}^{n\times k}$. We simply denote $\Xi[\mathbf{h}_j] = (\Xi(H))_{j,:}$, $\forall j \in [k]$. Following Proposition 3 in [40] yields the HS-Jacobian of $\Pi_{\Delta(X)}(\cdot)$ at C as $\mathcal{P}_C(H) = \Xi(H) - \Xi(X)M$, where M is diagonal with $M_{jj} = \mathbf{x}_j^\top \Xi[\mathbf{h}_j]/\mathbf{x}_j^\top \Xi[\mathbf{x}_j]$, $\forall j \in [k]$. Hence, we have the HS-Jacobian of $\mathcal{F}$ at Z as

$$\partial\mathcal{F}(Z)[H] = H - \mathcal{P}_C\left(H - \alpha \operatorname{Hess} m_l(X^l)[H]\right), \quad \forall H \in \mathbb{R}^{n\times k}. \tag{4.12}$$

We then apply the adaptive semi-smooth Newton (ASSN) method in [57,46] to generate a sequence $\{Z^j\}$ to solve (4.11). It satisfies $\lim_{j\to\infty} \|\mathcal{F}(Z^j)\| = 0$ by virtue of Theorem 3.10 in [46] under some reasonable assumptions. If $\tau_l > \kappa_H + \kappa_g$ and $\alpha \in (0, \frac{2}{\kappa_H+\kappa_g+\tau_l})$, by [57, Theorem 3.4] and the follow-up comments, we can show that the the limit point of $\{D^j := Z^j - X^l\}$ satisfies (4.7) with $c_1 = \frac{\tau_l-\kappa_H-\kappa_g}{\tau_l+\kappa_H+\kappa_g+1}$. However, since $\operatorname{Hess} m_l(X^l)$ may not be positive definite in other cases, it is still not clear whether (4.7) holds or not although the numerical performance is well. For completeness, a brief description of the second approach (4.10) for solving the subproblem (4.5) is outlined in Algorithm 4.

4.2 Convergence Analysis

We now study the asymptotic convergence of Algorithm 2 without postprocessing.

Theorem 4.3 *Let $\{X^t\}$ be the sequence generated by Algorithm 2 with $t_{\max} = \infty$, $\mathrm{tol}^{\mathrm{feas}} = -1$ and $\varepsilon_{\min}^{\mathrm{grad}} = 0$. If X^∞ is a limit point of $\{X^t\}$, then X^∞ is a weakly stationary point of problem (1.8).*

Algorithm 4: An ASSN based method for inexactly solving (4.5)

Initialization: Set $\delta_1 \in (0,1)$, $Z^0 = X_l$, $\tau_l > 0$ and $j = 0$.
while *"not converged"*, **do**
  Compute the HS-Jacobian $\partial\mathcal{F}(Z^j)$ according to (4.12).
  Solve (inexactly) the linear system $\partial\mathcal{F}(Z^j)[H^j] + \mu_j H^j = -F(Z^j)$.
  Compute $U^{j+1} = Z^j + H^j$. Set $Z^{j+1} = U^{j+1}$ if the residual $\|F(U^{j+1})\|_{\mathrm{F}}$ is reduced sufficiently. Otherwise, set $Z^{j+1} = Z^j$. Update μ_{j+1}.
  Set $j = j + 1$.
end
Set $D^l = Z^{j+1} - X^l$. If (4.7) holds, find the smallest nonnegative integer m such that $Y^l := \Pi_{\mathcal{OB}_+^{n,k}}(X^l + \delta_1^m D^l)$ satisfies (4.6). Otherwise, set $Y^l := X^l$.

Proof Note that $t_{\max} = \infty$ and $\mathrm{tol}^{\mathrm{feas}} = -1$, the algorithm does not stop within a finite number of iterations. Since $\{X^t\}$ is bounded, without loss of generality, throughout the proof we assume that $\{X^t\}$ converges to X^∞. By (4.2) and $P_{\theta_t}(X^{t,0}) \le P_{\theta_t}(X^{\mathrm{feas}})$, we obtain

$$f(X^t) + \sigma_t(\zeta_q(X^t) + \epsilon_t)^p \le f(X^{\mathrm{feas}}) + \sigma_t \epsilon_t^p. \tag{4.13}$$

If $p \ge 1$, it holds $\epsilon_t = 0$ and thus $\zeta_p(X^t) \to 0$. If $p \in (0,1)$, using $(a+b)^p - a^p \ge (1-2^{-p})b^p$ for $b > a > 0$ with $a = \epsilon_t$ and $b = \zeta_q(X)$, we also have $\zeta_p(X^t) \to 0$ from (4.13). It follows from the proof of Lemma 3.1 that $\zeta_2(X^t) \le \tilde{\varrho}_q \zeta_q(X^t)$ and thus $\zeta_2(X^t) \to 0$.

Denote $c_t := pq(\zeta_q(X^t) + \epsilon_t)^{p-1}\|XV\|_{\mathrm{F}}^{q-2}$. Some easy calculations yield

$$\mathrm{grad}\, P_{\theta_t}(X^t) = \mathrm{grad}\, f(X^t) + \sigma_t \mathrm{grad}\,(\zeta_q(X^t) + \epsilon_t)^p \tag{4.14}$$

with

$$\mathrm{grad}\,(\zeta_q(X^t) + \epsilon_t)^p = c_t X^t \left(\mathsf{Off}(VV^\top) - \mathsf{Diag}\left(\left((X^t)^\top X^t - I_k\right) VV^\top\right)\right). \tag{4.15}$$

Denote $\bar{\omega} := \max_{i,j\in[k]}[VV^\top]_{ij}$. For each X^t, we have

$$\omega \max_{l\in[k]\setminus\{j\}} X_{il}^t \le [X^t \mathsf{Off}(VV^\top)]_{ij} \le (k-1)\bar{\omega} \max_{l\in[k]\setminus\{j\}} X_{il}^t \quad \forall (i,j) \in [n]\times[k], \tag{4.16}$$

$$0 \le [X^t \mathsf{Diag}\left(\left((X^t)^\top X^t - I_k\right) VV^\top\right)]_{ij} \le X_{ij}^t \zeta_2(X^t) \quad \forall (i,j) \in [n]\times[k]. \tag{4.17}$$

We consider the following two cases.

Case I. The sequence $\{\sigma_t c_t\}$ is unbounded. Since $\{X^t\}$ converges to X^∞, there exists sufficiently large integer T_1 such that for every $t > T_1$, there holds that

$$\zeta_2(X^t) \le \frac{1}{4}\omega X_{\min}^\infty \quad \text{and} \quad X_{ij}^t \ge \frac{1}{2} X_{\min}^\infty \ \forall (i,j) \in \mathsf{supp}(X^\infty), \tag{4.18}$$

where $X_{\min}^\infty := \min_{(i,j)\in\mathsf{supp}(X^\infty)} X^\infty$. For any $(i,j) \in \Omega_0'(X^\infty)$, noting that $X^\infty \in \mathcal{S}_+^{n,k}$, with (4.18), for $t > T_1$, we have that $\max_{l\in[k]\setminus\{j\}} X_{il}^t \ge X_{\min}^\infty/2$.

With the first assertion in (4.16), for $t > T_1$ and $(i,j) \in \Omega_0'(X^\infty)$, there holds that

$$\left[X^t \mathsf{Off}(VV^\top)\right]_{ij} \geq \frac{1}{2}\omega X^\infty_{\min}. \tag{4.19}$$

With (4.17) and the first assertion in (4.18), and noting $X^t_{ij} \leq 1$, we derive for $t > T_1$ and $(i,j) \in \Omega_0'(X^\infty)$ that

$$[X^t \mathsf{Diag}\left(\left((X^t)^\top X^t - I_k\right) VV^\top\right)]_{ij} \leq \frac{1}{4}\omega X^\infty_{\min}. \tag{4.20}$$

Combining (4.19), (4.20) and (4.15), for $t > T$, we have

$$\sigma_t[\mathrm{grad}\,(\zeta_q(X^t) + \epsilon_t)^p]_{ij} \geq \frac{1}{4}\omega X^\infty_{\min}\sigma_t c_t \quad \forall (i,j) \in \Omega_0'(X^\infty), \tag{4.21}$$

which with the unboundedness of $\{\sigma_t c_{\epsilon_t}(X^t)\}$ yields $\limsup_{t\to\infty} \sigma_t[\mathrm{grad}\,(\zeta_q(X^t)+\epsilon_t)^p]_{ij} = \infty$, $\forall (i,j) \in \Omega_0'(X^\infty)$. Since $\mathrm{grad}\, f(X^t)$ is bounded and due to (4.14), we finally have $\limsup_{t\to\infty}\ [\mathrm{grad}\, P_{\theta_t}(X^t)]_{ij} = \infty\ \forall (i,j) \in \Omega_0'(X^\infty)$, which with (4.1) implies that there exists sufficiently large integer T_2 and subindices $\{t_l\}$ with $t_l > T_2$ such that $X^{t_l}_{ij} \leq \varepsilon^{\mathrm{grad}}_{t_l}\ \forall (i,j) \in \Omega_0'(X^\infty)$. Using (4.15), (4.16) and (4.17), we obtain that for any $(i,j) \in \mathsf{supp}(X^\infty)$,

$$-\sigma_{t_l} c_{t_l}\zeta_2(X^{t_l}) \leq [\sigma_{t_l}\mathrm{grad}\,(\zeta_q(X^{t_l}) + \epsilon_{t_l})^p]_{ij} \leq (k-1)\bar{\omega}\sigma_{t_l} c_{t_l}\varepsilon^{\mathrm{grad}}_{t_l}. \tag{4.22}$$

With the choice of η, γ_1 and (4.13), it is easy to verify that $\sigma_{t_l} c_{t_l}\varepsilon^{\mathrm{grad}}_{t_l} \to 0$. Since

$$\begin{aligned}\zeta_2(X^{t_l}) &= \mathrm{tr}(((X^t)^\top(X^t) - I_k)VV^\top)\\ &\leq k^2 n\bar{\omega}(\varepsilon^{\mathrm{grad}}_{t_l} + \max_{\substack{(i,j_1),(i,j_2)\in\Omega_0''(X^\infty)\\ j_1\neq j_2}} X^{t_l}_{ij_1}X^{t_l}_{ij_2}), \quad t_l > T_2\end{aligned}$$

we can show the leftmost term of (4.22) tends to 0 by proving that, for any $i \in [n]$, $j_1, j_2 \in [k]$ with $j_1 \neq j_2$ such that $\|X^\infty_{i,:}\| = 0$, there holds that $\lim_{l\to\infty} \sigma_{t_l} c_{t_l} X^{t_l}_{ij_1} X^{t_l}_{ij_2} = 0$. Obviously, we can focus on the case in which $\min\{X^{t_l}_{ij_1}, X^{t_l}_{ij_2}\} > \varepsilon^{\mathrm{grad}}_{t_l}$. The approximate optimality conditions (4.1) together with (4.14) and (4.15) gives

$$[\sigma_{t_l} c_{t_l} X^{t_l}\mathsf{Off}(VV^\top)]_{ij_1} \leq \varepsilon^{\mathrm{grad}}_{t_l} + \left|[\mathrm{grad}\, f(X^{t_l})]_{ij_1}\right| + \sigma_{t_l} c_{t_l}\zeta_2(X^{t_l})X^{t_l}_{ij_1}.$$

We have from (4.13) and the choice of γ_1 that $\sigma_{t_l} pq(\zeta_q(X^{t_l})+\epsilon_{t_l})^p \leq \sqrt{2k}L_f + \sigma_t\epsilon_t^p$ is bounded. It follows from the proof of Lemma 3.1 that $\sigma_{t_l} c_{t_l}\zeta_2(X^{t_l}) \leq \sigma_{t_l} pq(\zeta_q(X^{t_l}) + \epsilon_{t_l})^p\|X^{t_l}V\|_{\mathrm{F}}^{q-2}\tilde{\varrho}_q$ is also bounded. Thus

$$\lim_{l\to\infty} \sigma_{t_l} c_{t_l} X^{t_l}_{ij_1} X^{t_l}_{ij_2} \leq \lim_{l\to\infty} X^{t_l}_{ij_1}[\sigma_{t_l} c_{t_l} X^{t_l}\mathsf{Off}(VV^\top)]_{ij_1} = 0.$$

Together with (4.22), the previous equation implies

$$\lim_{l\to\infty}\ [\sigma_{t_l}\mathrm{grad}\,(\zeta_q(X^{t_l}) + \epsilon_{t_l})^p]_{ij} = 0 \quad \forall (i,j) \in \mathsf{supp}(X^\infty). \tag{4.23}$$

On the other hand, it follows from (4.1) that

$$\lim_{l\to\infty}[\mathrm{grad}\, P_{\theta_t}(X^{t_l})]_{ij}=0 \quad \forall (i,j)\in \mathsf{supp}(X^\infty). \tag{4.24}$$

Combining (4.23) and (4.24), we have from (4.14) that $\lim_{l\to\infty}[\mathrm{grad}\, f(X^{t_l})]_{ij}=0\,\forall (i,j)\in\mathsf{supp}(X^\infty)$. Considering that $X^{t_l}\to X^\infty$, we arrive at the conclusion in this case.

Case II. The sequence $\{\sigma_t c_t\}$ is bounded by a constant, say, $\bar{N}$. Similar to Case I, for any $\mu\in(0,1)$, there exists sufficiently large integer T_3 such that for every $t>T_3$ and $(i,j)\in\mathsf{supp}(X^\infty)$, there holds that

$$\zeta_2(X^t)\le \mu\min\{\omega X^\infty_{\min},1\} \text{ and } X^t_{ij}\ge \frac{1}{2}X^\infty_{\min}\ge \varepsilon_t^{\mathrm{grad}} \quad \forall (i,j)\in\mathsf{supp}(X^\infty). \tag{4.25}$$

The second assertion above together with (4.1) tells that for $t>T_3$ there holds that

$$|[\mathrm{grad}\, P_{\theta_t}(X^t)]_{ij}|\le \varepsilon_t^{\mathrm{grad}} \quad \forall (i,j)\in\mathsf{supp}(X^\infty). \tag{4.26}$$

Similar to Case I, for $(i,j)\in\Omega_0'(X^\infty)$ and $t>T_3$, we have $\max_{l\in[k]\setminus\{j\}}X^t_{il}\ge X^\infty_{\min}/2$ and thus

$$X^t_{ij}\le \frac{2}{X^\infty_{\min}}\max_{l\in[k]\setminus\{j\}}X^t_{il}X^t_{ij}\le \frac{2}{X^\infty_{\min}\omega}[VV^\top]_{j'j}(\mathbf{x}^t_{j'})^\top\mathbf{x}^t_j\le \frac{2\zeta_2(X^t)}{X^\infty_{\min}\omega}, \tag{4.27}$$

where $j'=\mathrm{argmax}_{l\in[k]\setminus\{j\}}X^t_{il}$ and the last inequality uses the fact that $\zeta_2(X^t)=\sum_{i,j\in[k],i\neq j}[VV^\top]_{ij}((\mathbf{x}^t_i)^\top\mathbf{x}^t_j)$ which appears in the proof of Lemma 2.1. By the first assertion in (4.25) and (4.27), we have for $t>T_3$ that $X^t_{ij}\le 2\mu\,\forall (i,j)\in\Omega_0'(X^\infty)$. Noting that $X^\infty\in\mathcal{S}_+^{n,k}$, this together with (4.16) implies that

$$\left[X^t\mathsf{Off}(VV^\top)\right]_{ij}\le 2(k-1)\bar{\omega}\mu \quad \forall (i,j)\in\mathsf{supp}(X^\infty). \tag{4.28}$$

Again using (4.17) and noting that $X^t_{ij}\le 1$ and (4.25), we have

$$[X^t\mathsf{Diag}\left(\left((X^t)^\top X^t-I_k\right)VV^\top\right)]_{ij}\le\mu \quad \forall (i,j)\in\mathsf{supp}(X^\infty). \tag{4.29}$$

Combining (4.28) and (4.29), we have from (4.15) that for $t>T_3$ there holds that

$$\left|[\sigma_t\mathrm{grad}\,(\zeta_q(X^t)+\epsilon_t)^p]_{ij}\right|\le \bar{N}\left(2(k-1)\bar{\omega}+1\right)\mu \quad \forall (i,j)\in\mathsf{supp}(X^\infty).$$

Consequently, by (4.14) and (4.26), for $t>T_3$, we have for each $(i,j)\in\mathsf{supp}(X^\infty)$ that $|[\mathrm{grad}\, f(X^t)]_{ij}|\le \bar{N}\left(2(k-1)\bar{\omega}+1\right)\mu+\varepsilon_t^{\mathrm{grad}}$. Due to the arbitrariness of μ, we conclude from $X^t\to X^\infty$ that $|[\mathrm{grad}\, f(X^\infty)]_{ij}|\le \lim_{t\to\infty}\varepsilon_t^{\mathrm{grad}}=0\,\forall (i,j)\in\mathsf{supp}(X^\infty)$. The proof is completed. □

Example 4.1 Consider problem (1.8) with $f(X) = \langle C, X\rangle$ and $C \in \mathbb{R}^{3\times 2}$ with $C_{11} = C_{12} = -1$ and $C_{21} = C_{22} = C_{31} = C_{32} = 0$. For Algorithm 2, we set $V = 1/\sqrt{2}\,[1\ 1]^\top$, $p = 1$, $q = 2$, $\epsilon_0 = 0$, $\mathrm{tol}^{\mathrm{feas}} = 0$ and $\mathrm{tol}_t^{\mathrm{sub}} = 0$ for all $t \geq 0$. By some easy calculations, we see that X^t with $X^t_{11} = X^t_{12} = 1/\sigma_t$, $X_{22} = X_{31} = 0$ and $X_{21} = X_{32} = \sqrt{1 - 1/\sigma_t^2}$ is a stationary point of (1.9) with $\sigma = \sigma_t$ and it also satisfies (4.2). Unfortunately, the limit point X^∞ of $\{X^t\}$ is not a stationary point of problem (1.8) under the above settings. This counter-example tells that X^∞ is a stationary point only if some additional conditions are satisfied. On the other hand, we observe numerically that the point $\bar{X}$ found by our algorithm satisfies $\|\bar{X}\|_0 = n$, namely, $\Omega_0''(\bar{X}) = \emptyset$. This implies that all the weakly stationary points encountered in numerical experiments are stationary points.

Theorem 4.4 *Under conditions of Theorem 4.3, if additionally there holds that*

$$\lim_{t\to\infty} \sigma_t(\zeta_q(X^t) + \epsilon_t)^{p-1} X^t_{ij} = 0 \quad \forall (i,j) \in \Omega_0''(X^\infty), \tag{4.30}$$

then X^∞ is a stationary point of problem (1.8).

Proof By Theorem 4.3 and Theorem 2.1, it remains to verify the correctness of the equation $[\mathrm{grad}\, f(X^\infty)]_{ij} \geq 0\ \forall (i,j) \in \Omega''(X^\infty)$. For any $(i,j) \in \Omega_0''(X^\infty)$, it is easy to see that $(i,l) \in \Omega_0''(X^\infty)$ for each $l \in [k]$. Together with (4.15) and (4.16), we have

$$\begin{aligned} &\sigma_t[\mathrm{grad}\,(\zeta_q(X^t) + \epsilon_t)^p]_{ij} \\ &\leq pq(k-1)\bar{\omega}\|X^tV\|_{\mathrm{F}}^{q-2}\left(\sigma_t(\zeta_q(X^t) + \epsilon_t)^{p-1} \max_{(i,j)\in\Omega_0''(X^\infty)} X^t_{ij}\right), \end{aligned}$$

which with (4.30) and $\lim_{t\to\infty}\|X^tV\|_{\mathrm{F}} = 1$ gives $\lim_{t\to\infty}\sigma_t[\mathrm{grad}\,(\zeta_q(X^t) + \epsilon_t)^p]_{ij} \leq 0\ \forall (i,j) \in \Omega_0''(X^\infty)$. By (4.1), we have $[\mathrm{grad}\, P_{\theta_t}(X^t)]_{ij} \geq -\varepsilon_t^{\mathrm{grad}}$ for any $(i,j) \in [n]\times[k]$. Consequently, it follows from (4.14) that for any $(i,j) \in \Omega_0''(X^\infty)$ there holds

$$[\mathrm{grad}\, f(X^\infty)]_{ij} = \lim_{t\to\infty}[\mathrm{grad}\, f(X^t)]_{ij} \geq \lim_{t\to\infty}[\mathrm{grad}\, P_{\theta_t}(X^t)]_{ij} \geq 0.$$

The proof is completed. □

Furthermore, Theorem 4.3 gives the following corollary.

Corollary 4.1 *Consider the same conditions as in Theorem 4.3. If additionally $\|X^\infty\|_0 = n$, then X^∞ is a stationary point of problem (1.8).*

Remark 4.2 Here, we present a different understanding of the above corollary. Lemma 2.2 tells that CCP holds at X^∞ if $\|X^\infty\|_0 = n$. By the approximate optimality conditions (4.1) and the choice of the penalty term, X^∞ is an approximate KKT point (see [3] for its definition). Recalling that CCP is a strict CQ, a CQ which guarantees an approximate KKT point as a KKT point (see [3] for details), we thus know that X^∞ must be a stationary point.

Theorem 4.4 implies that the limit point X^∞ is stationary as long as X^t_{ij} decays to 0 sufficiently fast on $(i,j) \in \Omega_0''(X^\infty)$, while the following theorem improves the convergence result by requiring a better solution of the subproblem.

Definition 2 We say $\bar{X} \in \mathcal{OB}_+^{n,k}$ satisfies the weak second-order optimality conditions (WSOC) of (1.9) if $\min\left(\bar{X}, \operatorname{grad} h(\bar{X})\right) = 0$ and $\langle D, \operatorname{Hess} h(\bar{X})[D]\rangle \geq 0$ for any $D \in \tilde{\mathcal{C}}_{\mathcal{OB}_+^{n,k}}(\bar{X}) := \{D \in \mathbb{R}^{n\times k} : \bar{\mathbf{x}}_j^\top \mathbf{d}_j = 0, j \in [k], D_{ij} = 0 \text{ if } \bar{X}_{ij} = 0\}$.

Theorem 4.5 *Let $\{X^t\}$ be the sequence generated by Algorithm 2 with $t_{\max} = \infty$, $\mathrm{tol}^{\mathrm{feas}} = 0$, $p \leq 1$ and $\varepsilon_0^{\mathrm{grad}} = \varepsilon_{\min}^{\mathrm{grad}} = 0$. If X^t satisfies the WSOC conditions of the subproblem (1.9), then the algorithm stops at some $\tilde{t}$ iteration and $X^{\tilde{t}}$ is a stationary point of problem (1.8).*

Proof We prove it by contradiction. Suppose that $X^t \notin \mathcal{S}_+^{n,k}$ for every t. Without loss of generality, we assume $X^t \to X^\infty$. Since $p \leq 1$, the sequence $\{\sigma_t c_t\}$ tends to infinity. Following Case I in the proof of Theorem 4.3, we have $X^t_{ij} = 0, \forall (i,j) \in \Omega_0'(X^\infty)$ for large enough t since X^t satisfies conditions (4.1) with $\varepsilon_t^{\mathrm{grad}} = 0$.

Since $X^t \notin \mathcal{S}_+^{n,k}$ for any t, there must exist $i_1 \in [n]$, $j_1, j_2 \in [k]$ and a subsequence $\{t_l\}$ such that $j_1 \neq j_2$, $\|X^\infty_{i_1,:}\| = 0$, $X^{t_l}_{i_1,j_1}, X^{t_l}_{i_1,j_2} \neq 0$ and $\mathsf{supp}(X^\infty) \subset \mathsf{supp}(X^{t_l})$ for all l. Since X^{t_l} satisfies the WSOC conditions of (1.9), denote $\theta_{t_l} = \{\sigma_{t_l}, p, q, \epsilon_{t_l}\}$, we have

$$\langle D, \operatorname{Hess} P_{\theta_{t_l}}(X^{t_l})[D]\rangle \geq 0, \ \forall D \in \tilde{\mathcal{C}}_{\mathcal{OB}_+^{n,k}}(X^{t_l}). \tag{4.31}$$

Choosing D^{t_l} with $D^{t_l}_{i_1 j_1} = 1, D^{t_l}_{i_1 j_2} = -1, D^{t_l}_{k_1 j_1} = -X^{t_l}_{i_1 j_1}/X^{t_l}_{k_1 j_1}$ and $D^{t_l}_{k_2 j_2} = X^{t_l}_{i_1 j_2}/X^{t_l}_{k_2 j_2}$ and the remaining elements of D^{t_l} being zeros. Here, the indices k_1, k_2 are chosen such that $(k_1, j_1), (k_2, j_2) \in \mathsf{supp}(X^\infty)$. One can check via direct calculation that $D^{t_l} \in \tilde{\mathcal{C}}_{\mathcal{OB}_+^{n,k}}(X^{t_l})$ and

$$\left\langle D^{t_l}, D^{t_l}\mathsf{Off}(VV^\top)\right\rangle \leq -\omega, \ \left|\left\langle D^{t_l}, X^{t_l}VV^\top\right\rangle\right| \leq 2k\bar{\omega} \max_{(i,j)\in\Omega_0''(X^\infty)} X^{t_l}_{ij}. \tag{4.32}$$

Substituting D^{t_l} into the WSOC condition (4.31) yields

$$\left\langle D^{t_l}, \operatorname{Hess} f(X^{t_l})[D^{t_l}]\right\rangle + \sigma_{t_l}\left\langle D^{t_l}, \operatorname{Hess}\left(\zeta_q(X^{t_l}) + \epsilon_{t_l}\right)^p)[D^{t_l}]\right\rangle \geq 0, \tag{4.33}$$

where

$$\begin{aligned}
&\left\langle D^{t_l}, \operatorname{Hess}\left(\zeta_q(X^{t_l}) + \epsilon_{t_l}\right)^p)[D^{t_l}]\right\rangle \\
&= c_{t_l}\Bigg(a|\langle D^{t_l}, X^{t_l}VV^\top\rangle|^2 \\
&\quad + \left\langle D^{t_l}, D^{t_l}\left(\mathsf{Off}(VV^\top) - \mathsf{Diag}\left(((X^{t_l})^\top X^{t_l} - I_k)VV^\top\right)\right)\right\rangle\Bigg)
\end{aligned}$$

with $a = \frac{q-2}{\|X^{t_l}V\|_{\mathrm{F}}^2} + \frac{(p-1)\|X^{t_l}V\|_{\mathrm{F}}^{q-2}}{\zeta_q(X^{t_l})+\epsilon_{t_l}}$. Since $p \leq 1$, we can drop the term with respect to $p-1$ when deriving a upper bound for the right-hand side of the above assertion. A closer check then reveals that the term $\langle D^{t_l}, D^{t_l}\mathsf{Off}(VV^\top)\rangle$ dominates the equation in the brackets since others tend to 0, according to $X^{t_l} \to X^\infty \in \mathcal{S}_+^{n,k}$ and (4.32). Hence, we obtain

$$\lim_{l\to\infty} \sigma_{t_l} \left\langle D^{t_l}, \mathrm{Hess}\,(\zeta_q(X^{t_l}) + \epsilon_{t_l})^p)[D^{t_l}]\right\rangle \leq \sigma_{t_l} c_{t_l} \langle D^{t_l}, D^{t_l}\mathsf{Off}(VV^\top)\rangle = -\infty.$$

Together with (4.33) and the fact that $\langle D^{t_l}, \mathrm{Hess}\, f(X^{t_l})[D^{t_l}]\rangle$ is bounded, we reach a contradiction. Therefore, the algorithm stops at some $\tilde{t}$ iteration. Since (4.1) holds at $X^{\tilde{t}}$ with $\varepsilon_{\tilde{t}}^{\mathrm{grad}} = 0$ we know that $X^{\tilde{t}}$ must be a stationary point of probelm (1.8). The proof is completed. □

Remark 4.3 Notice that $\mathcal{N}_{\mathcal{X}_V}(X) \subset \mathcal{C}_{\mathcal{OB}_+^{n,k}}(X)$. For $X \in \mathcal{S}_+^{n,k}$, there always holds that $\left\langle D, XVV^\top\right\rangle = \left\langle D, D\mathsf{Off}(VV^\top)\right\rangle = 0$ for any $D \in \mathcal{N}_{\mathcal{X}_V}(X)$ and thus $\langle D, \mathrm{Hess}\, f(X)[D]\rangle = \langle D, \mathrm{Hess}\, P_\theta(X)[D]\rangle \ \ \forall D \in \mathcal{N}_{\mathcal{X}_V}(X)$. Thus, when $p \leq 1$, $X^{\tilde{t}}$ in Theorem 4.5 satisfies the second order necessary conditions (2.13) of problem (1.8) as long as WSOC conditions therein replaced by the second order necessary conditions of the subproblem (1.9). Moreover, we immediately arrive at the following corollary which characterizes the relationship between the local minimizers of the subproblem and the original problem (1.8). The key issue is the feasibility of X^t after finite iterations. Guaranteeing that the local minimizer of the penalty subproblem with sufficiently large penalty value is feasible to the original problem is not an easy task, we usually need some strong CQs, such as LICQ or MFCQ. With these CQs, there exists some universal threshold penalty value (always not explicitly computable), see [23, Theorem 2], [24, Theorem 4], [22, Proposition 7], [25, Theorem 5.3], [27, Theorem 6], to name a few. However, these CQs do not hold for our problem.

Corollary 4.2 *Let $\{X^t\}$ be the sequence generated by Algorithm 2 with $t_{\max} = \infty$, $\mathrm{tol}^{\mathrm{feas}} = 0$, $p \leq 1$ and $\varepsilon_0^{\mathrm{grad}} = \varepsilon_{\min}^{\mathrm{grad}} = 0$. If X^t is a local minimizer of the subproblem (1.9), then the algorithm stops at some $\tilde{t}$ iteration and $X^{\tilde{t}}$ is a local minimizer of problem (1.8).*

Moreover, our exact penalty approach can be applied to the general problem (1.10). The subproblem (1.9) becomes

$$\min_{X\in\mathcal{OB}_+^{n,k}, Y\in\mathcal{Y}} \left\{P_\theta(X,Y) := f(X,Y) + \sigma\left(\|XV\|_{\mathrm{F}}^q - 1 + \epsilon\right)^p\right\}. \tag{4.34}$$

The corresponding algorithm is almost the same as Algorithm 2 except that condition (4.1) is replaced by

$$\begin{aligned}&\left\|\min(X^t, \mathrm{grad}_X P_{\theta_t}(X^t, Y^t))\right\|_{\mathrm{F}} \leq \varepsilon_t^{\mathrm{grad}},\\ &\mathrm{dist}\big(Y^t, \Pi_{\mathcal{Y}}(Y^t - \nabla_Y P_{\theta_t}(X^t, Y^t))\big) \leq \varepsilon_t^{\mathrm{grad}}\end{aligned}$$

and condition (4.2) becomes $P_{\theta_t}(X^t, Y^t) \le P_{\theta_t}(X^{t,0}, Y^{t,0})$. To obtain a point satisfying the above conditions, we can employ the proximal alternating linearized minimization (PALM) method in [11]. One can also use the proximal alternating minimization scheme [4], wherein the X-subproblem can be approximately solved by the second-order method Algorithm 3. In this case, we can extend the convergence results to this general model by following almost the same proof.

5 Numerical experiments

In this section, we present a variety of numerical results to evaluate the performance of our proposed method. All experiments are performed in Windows 10 on an Intel Core 4 Quad CPU at 2.30 GHZ with 8 GB of RAM. All codes are written in MATLAB R2018b. The matrix V is simply taken as $V = \mathbf{e}/\sqrt{k}$, and the choice of parameters for Algorithm 2 are set as follows: $p = 1$, $q = 2$, $\varepsilon_0^{\rm grad} = 10^{-1}$, $\varepsilon_{\rm min}^{\rm grad} = 10^{-7}$, $t_{\max} = 300$; the choices of γ_2, σ_0, η, $\mathrm{tol}^{\rm feas}$ and X^0 are given in each subsection. In our implementation, instead of using (4.1), we use the stopping condition when the distance between two consecutive iterations is small, namely, $\|X^{l+1} - X^l\|_{\rm F} \le \varepsilon_t^{\rm grad}$.

5.1 Computing projection onto $\mathcal{S}_+^{n,k}$

Given $C \in \mathbb{R}^{n\times k}$, we consider to compute its projection onto $\mathcal{S}_+^{n,k}$, which is formulated as

$$\min_{X\in\mathcal{S}_+^{n,k}} \|X - C\|_{\rm F}^2. \tag{5.1}$$

The associated exact penalty model (1.9) with $p = 1$, $q = 2$, and $\epsilon = 0$ becomes

$$\min_{X\in\mathcal{OB}_+^{n,k}} -\frac{1}{\sigma_t}\langle C, X\rangle + \frac{1}{2}\|XV\|_{\rm F}^2. \tag{5.2}$$

The Lipschitz constant of the gradient of the objective function in (5.2) is 1 since to $VV^\top \preceq I_k$. Thus we can simply use the nonconvex gradient projection scheme $X^{l+1} \in \Pi_{\mathcal{OB}_+^{n,k}}\big(X^l - \alpha(X^lVV^\top - C/\sigma_t)\big)$ with $\alpha \equiv 0.99$ to solve the subproblem (5.2). It is always difficult to seek the projection globally for a general matrix C. Due to Proposition A.1, we can construct a family of matrices with unique and known projection. For a given $B \in \mathcal{S}_+^{n,k}$, the MATLAB codes for generating C is given as

```
X = (B>0).*(1+rand(n,k)); Xstar = X./sqrt(sum(X.*X));
d = 0.5+3*rand(k,1); L = xi*((d*d').^0.5).*rand(k,k);
L(sub2ind([k,k],1:k,1:k))=d; C=Xtar*L;
```

The parameter $\xi \in [0, 1]$ controls the magnitude of noise level. Larger ξ makes it harder to find the ground truth X^*. Let $X^\diamondsuit$ be the solution generated by EP4Orth+, namely, Algorithm 2. Note that the postprocessing problem (3.1) has closed form solution. Define gap $= \|X^\diamondsuit - C\|_{\mathrm{F}}/\|X^* - C\|_{\mathrm{F}} - 1$ as a measure of the solution quality. For each ξ, n and k, we run 50 times of EP4Orth+, and the initial point is generated by rounding C via Procedure 1. We choose $\gamma_2 = 5$, $\mathrm{tol}^{\mathrm{feas}} = 10^{-8}$, $\sigma_0 = 10^{-2}$, $\eta = 0.8$. It should be mentioned that, by setting $\mathrm{tol}^{\mathrm{feas}} = 10^{-8}$, we can always find a nearly feasible solution even without using the postprocessing. The averaged results are reported in Table 1, wherein the "suc" means the total number of instances for which the gap is zero. From this table, we can see that for small ξ, EP4Orth+ can solve all 50 instances to a zero gap, while for large ξ it can only solve some instances to a zero gap. However, for all cases, EP4Orth+ always returns an orthogonal nonnegative matrix with a small gap.

Table 1 Numerical results on computing projection onto $\mathcal{S}_+^{n,k}$, "t" means the time in seconds, "nproj" means the number of gradient projection steps.

	$n = 2000, k = 10$				$n = 2000, k = 50$				$n = 2000, k = 100$			
ξ	suc	gap	t	nproj	suc	gap	t	nproj	suc	gap	t	nproj
0.80	50	0.0e0	0.01	24.5	50	0.0e0	0.06	62.7	50	0.0e0	0.52	95.7
0.90	50	0.0e0	0.01	28.7	50	0.0e0	0.07	82.1	50	0.0e0	0.66	134.6
0.95	49	7.2e-5	0.01	31.9	46	2.1e-4	0.09	112.2	49	6.6e-7	0.87	184.8
0.98	43	8.9e-4	0.01	33.8	22	5.0e-4	0.11	156.3	19	8.0e-4	1.23	268.2
1.00	37	1.2e-3	0.01	38.1	0	2.6e-3	0.12	170.3	0	2.6e-3	1.43	317.5

	$n = 2000, k = 200$				$n = 2000, k = 300$				$n = 2000, k = 400$			
ξ	suc	gap	time	nproj	suc	gap	time	nproj	suc	gap	time	nproj
0.80	50	0.0e0	1.38	144.7	50	0.0e0	2.62	186.7	50	0.0e0	3.79	211.7
0.90	50	0.0e0	1.6	207.6	50	0.0e0	3.43	276.0	50	0.0e0	5.39	328.7
0.95	50	0.0e0	2.42	295.1	50	0.0e0	5.13	424.9	50	0.0e0	7.74	483.0
0.98	23	4.5e-4	3.93	489.2	20	2.5e-4	8.60	718.6	24	1.7e-4	15.42	962.2
1.00	0	1.9e-3	5.07	636.9	0	1.8e-3	11.31	951.3	0	1.6e-3	20.86	1324.0

5.2 Orthogonal nonnegative matrix factorization

We compare our proposed method with uni-orthogonal NMF (U-onmf) [26], orthonormal projective nonnegative matrix factorization (OPNMF) [61], orthogonal nonnegatively penalized matrix factorization (ONP-MF) [50] and EM-like algorithm for ONMF (EM-onmf) [50]. In addition to the above methods, we also compare our method with K-means, which is considered as a benchmark in clustering problems. We implement U-onmf by ourselves since the original code is not available. We adopt the implementation of OPNMF from `https://github.com/asotiras/brainparts`. The codes of ONP-MF and EM-onmf can be downloaded from `https://github.com/filippo-p/onmf`. As to K-means, we call the MATLAB function `kmeans` directly. Note

that our proposed method and OPNMF solve the equivalent formulation (1.4) while the remaining methods solve directly (1.3). Considering that the objective function in (1.4) is quartic, to make the subproblem (1.9) easier to solve, one can consider the Gauss-Newton technique as

$$\|A-XX^\top A\|_{\mathrm{F}}^2 = \|A-X\tilde{X}^\top A-\tilde{X}S^\top A-SS^\top A\|_{\mathrm{F}}^2 \approx \|A-X\tilde{X}^\top A-\tilde{X}S^\top A\|_{\mathrm{F}}^2,$$

where $S = X - \tilde{X}$. By neglecting the term $\tilde{X}S^\top A$, we obtain a partial Gauss-Newton approximation, namely, $\|A - XX^\top A\|_{\mathrm{F}}^2 \approx \|A - X\tilde{X}^\top A\|_{\mathrm{F}}^2$. Moreover, if $X \in \mathcal{S}_+^{n,k}$, we know that $\|A - XX^\top A\|_{\mathrm{F}}^2 = \|A - X(X^\top X)^{-1}X^\top A\|_{\mathrm{F}}^2$. Hence, to make the approximation robust, we consider $\|A - XX^\top A\|_{\mathrm{F}}^2 \approx \|A - X(\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top A\|_{\mathrm{F}}^2$. The objective function in subproblem (1.9) at t-th iteration with $p = 1$, $q = 2$, and $\epsilon = 0$ becomes $\|A - X(Y^t)^\top\|_{\mathrm{F}}^2 + \sigma_t\|XV\|_{\mathrm{F}}^2$ with $Y^t = \Pi_{\mathbb{R}_+^{r\times k}}\big(A^\top\tilde{X}^t((\tilde{X}^t)^\top\tilde{X}^t)^{-1}\big)$.

In some datasets, the matrix A maybe degenerated, namely, there exists a row (column) of A with all zero entries. This causes a division by zero error when running the U-onmf method. Thus we will first remove such degenerate rows and columns of A. For K-means and EN-onmf, the initial points are chosen randomly. The other methods adopt the SVD-based initializations [15]. The time cost of generating initial points is relatively low compared to that of the rest parts. In practice, we utilize the nonconvex projection gradient method (4.4) if $\zeta(X^{t,0}) = \|X^{t,0}V\|_{\mathrm{F}}^2 - 1 > \bar{\zeta}$, otherwise switch to Algorithm 3. We adopt the Barzilai-Borwein stepsize [6] and use the nonmonotone line search [64] in the gradient projection iteration (4.4). We set $\sigma_0 = 10^{-3}$, $\eta = 0.98$, and choose $\mathrm{tol}^{\mathrm{feas}} = 10^{-8}$, $\bar{\zeta} = 5$ in section 5.2.1 and $\mathrm{tol}^{\mathrm{feas}} = 0.3$, $\bar{\zeta} = 0.6$ in section 5.2.2. In section 5.2.1, we set $\gamma_2 = 1.05$ if $\|X^tV\|_{\mathrm{F}}^2 > 2$ and $\gamma_2 = 1.03$, otherwise. In section 5.2.2, we set $\gamma_2 = 1.1 \times 1.05$ if $\|X^tV\|_{\mathrm{F}}^2 > 2$ and $\gamma_2 = 1.1 \times 1.03$, otherwise. The main parameters of Algorithm 3 are chosen as $\eta_1 = 0.01$, $\eta_2 = 0.9$, $\beta_0 = 0.98$, $\beta_1 = 1$, and $\beta_2 = 1.3$.

5.2.1 Text and image clustering

We evaluate algorithms on text and image datasets adopted from [17], they are available at `http://www.cad.zju.edu.cn/home/dengcai/Data/data.html`. Since the original text dataset is too huge and disproportionate, we extract some subsets from original data to make it suitable for testing clustering algorithms. The details of modification are provided in section 6.2.1 in [34] due to space limits. For text datasets, every article is assigned with a vector, which reflects the frequency of each word in the article. While for image datasets, a vector represents the gray level of each pixel in a picture. The data matrix A is comprised of these vectors. Any solution $X^* \in \mathcal{S}_+^{n,k}$ of ONMF indicates a partition (clustering result) of the dataset. The scale of each dataset is given in Table 2, in which "data" denotes the number of rows of data matrix A and "features" stands for the number of columns.

We consider three criteria to compare the performance of clustering results: purity, entropy and NMI. The definitions of the three criteria are skipped for space consideration, see also section 6.2.1 in [34] and the references therein.

Table 2 Description of each dataset.

Name	Reuters-t10	Reuters-t20	TDT2-t10	TDT2-t20	NewsG-t5	MNIST	Yale
data	1897	2402	1477	1721	2344	4000	165
features	12444	13568	22181	23674	14475	784	1024
clusters	10	20	10	20	5	10	5

Generally, a better clustering result has smaller entropy, larger purity and NMI. Note that we will not calculate "feasi" for K-means and EN-onmf, as they only generate the clustering results instead of solutions of ONMF problem. Since we aim to show that our algorithm can generate a solution with high quality and small feasibility violation, we remove the postprocessing in EP4Orth+ for fair comparison. For random algorithms, their results are averaged over 10 runs. In Table 3, we report text and images clustering results. We can observe from this table that our proposed method performs very well. Specifically, the clustering results given by our proposed method has the highest purity and NMI in most of cases (being close for the rest dataset). As to the speed, our method is faster than U-onmf and ONP-MF for most of cases, and it is especially efficient on text dataset. Besides, the feasibility violation of the solution returned by our method is very small, while those returned by the other methods are always very large. On the other hand, K-means is the fastest among all algorithms and performs well on image datasets MNIST and Yale, but it results poorly when applying to text dataset; EM-onmf and OPNMF are efficient but their performance is slightly worse than ours.

Table 3 Text clustering results on real datasets. In the table, "c1", "c2" and "c3" stand for "purity" (%), "NMI" (%) and "entropy" (%), respectively; "t" means the time in seconds. The term "feasi" means the feasibility violation with feasi $:= ||\hat{X}^\top \hat{X} - I_k||_{\mathrm{F}} + ||\min(\hat{X}, 0)||_{\mathrm{F}}$, where $\hat{X}$ is the solution generated by the corresponding algorithm. Results marked in bold mean better performance in the corresponding index.

	EP4Orth+					U-onmf					K-means			
datasets	c1	c2	c3	feasi	t	c1	c2	c3	feasi	t	c1	c2	c3	t
Ret-t10	**73.1**	**60.7**	**37.9**	3e-14	9	72.7	59.2	39.4	0.6	55	36.9	22.2	75.1	4
Ret-t20	**65.5**	56.3	38.4	1e-15	25	60.6	52.7	41.7	0.9	149	33.9	17.4	79.8	4
TDT2-t10	**85.7**	70.0	20.8	7e-12	9	80.9	65.7	22.8	0.5	115	41.1	17.8	70.5	4
TDT2-t20	**82.3**	**69.6**	**18.1**	3e-15	18	79.3	64.3	21.2	0.7	299	39.1	18.6	65.8	7
NewsG-t5	41.5	**22.8**	**77.1**	1e-15	7	39.3	14.9	85.0	0.2	18	21.1	0.4	99.5	2
MNIST	**60.1**	**48.9**	**51.0**	1e-15	26	50.0	41.9	58.0	1.0	39	55.4	45.2	54.7	0.9
Yale	**44.8**	**47.9**	**52.1**	9e-12	2	43.7	45.9	54.0	1.2	2	40.8	44.1	55.9	0.1

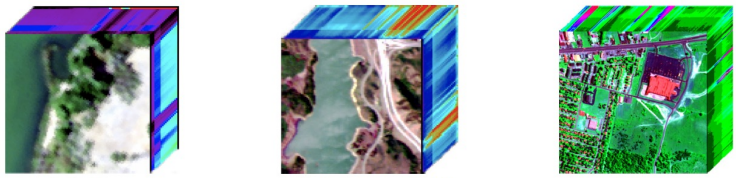
	OPNMF					ONP-MF					EM-onmf			
datasets	c1	c2	c3	feasi	t	c1	c2	c3	feasi	t	c1	c2	c3	t
Ret-t10	72.0	58.7	39.9	1.1	15	66.9	52.8	45.6	3e-3	82	71.3	58.6	39.9	17
Ret-t20	62.9	54.6	40.0	1.8	24	62.0	53.5	41.6	4e-3	386	64.1	**57.4**	**37.8**	30
TDT2-t10	82.2	64.3	24.4	0.9	10	82.9	65.3	23.8	3e-3	133	85.0	**71.3**	**20.1**	21
TDT2-t20	79.1	62.5	21.4	1.1	14	81.1	65.0	20.4	4e-3	542	80.8	67.2	19.3	25
NewsG-t5	37.1	13.1	86.7	0.4	11	**42.9**	22.6	77.2	2e-3	44	35.7	15.4	84.5	14
NMIST	55.1	44.1	55.9	1.3	218	57.4	46.1	53.8	5e-2	61	56.3	47.8	52.2	4
Yale	43.7	45.4	54.6	1.4	4	40.0	43.6	56.6	1e-2	10	38.1	41.7	58.3	0.1

5.2.2 Hyperspectral unmixing

A set of images taken on the same object at different wavelengths is called a hyperspectral image. At a given wavelength, images are generated by surveying reflectance on each single pixel. Hyperspectral unmixing plays an essential role in hyperspectral image analysis [10]. Let n be the total number of pixels. It assumes that each pixel spectrum $\mathbf{a}_i \in \mathbb{R}^r_+$ with $i \in [n]$ is a composite of k spectral bases $\{\mathbf{y}_j\}_{j=1}^k \in \mathbb{R}^r_+$, where r is the number of wavelengths. Each spectral base is denoted as an endmember, which represents the pure spectrum. For example, a spectral base could be the spectrum of "rock", "tree", etc. Linear mixture model [35,68] approximates the pixel spectrum $\mathbf{a}_i$ by a linear combination of endmembers as $\mathbf{a}_i = Y\tilde{\mathbf{x}}_i + \mathbf{r}_i$, where $\tilde{\mathbf{x}}_i \in \mathbb{R}^k_+$ is called the abundance vector corresponding to pixel $\mathbf{a}_i$, $\mathbf{r}_i \in \mathbb{R}^r$ is a residual term and $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_k] \in \mathbb{R}^{r\times k}_+$ is the endmember matrix. When ONMF is applied to hyperspectral unmixing, we assume that both endmember and abundances remain unknown. In addition, each pixel only corresponds to one material. That is to say, $\tilde{\mathbf{x}}_i$ only has one non-zero element. For all the pixels combined together, the ONMF formulation of hyperspectral image unmixing becomes (1.3). Specifically, we adopt the formulation in [50,68], where the matrix $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\top \in \mathbb{R}^{n\times r}_+$ represents n pixels observed at r different wavelengths and each column of A corresponds to an image observed at a given wavelength. This approach can be seen as unfolding a 3-order tensor for representing a set of images observed at r wavelengths to a matrix by vectorizing the 2-D image. The matix $X \in \mathcal{S}^{n,k}_+$ is the abundance matrix having $\tilde{\mathbf{x}}_i^\top$ as its i-th row.

We test algorithms on three widely used hyperspectral image datasets, Samson, Jasper Ridge and Urban [68]. They are widely used datasets in the hyperspectral unmixing study and can be downloaded at `http://www.escience.cn/people/feiyunZHU/Dataset_GT.html`. Since the sizes of the first two images are huge, we choose a region in each image. This process is common in the context of hyperspectral unmixing. For Samson, a region which contains 95×95 pixels is chosen, starting from the $(252, 332)$-th pixel in original image. We choose a subimage of Jasper Ridge with 100×100 pixels, whose first pixel corresponds to the $(105, 269)$-th pixel in the original image. The refined Samson has 156 wavelengths and it contains three endmembers: water, tree and rock. The refined Jasper Ridge has 198 wavelengths, and its endmembers include water, tree, dirt and road. Urban is the largest hyperspectral data with 307×307 pixels observed at 162 wavelengths, and there are four endmembers: asphal, grass, tree and roof.

Since the groundtruth of abundance matrix X does not satisfy the orthogonality constraints, the criteria utilized in the preceding subsection are not appropriate to measure the quality of hyperspectral unmixing. Here we consider spectral angle distance (SAD) (see for instance [68]) to evaluate the performance of algorithms. SAD uses an angle distance between ground truth and estimated endmembers to measure the accuracy of endmember estimation. It is defined as $\text{SAD} := \frac{1}{k}\sum_{i=1}^k \arccos\left(\frac{y_i^\top \hat{y}_i}{\|y_i\|\|\hat{y}_i\|}\right)$, where $\hat{y}_i$ and y_i are



(a) Samson (b) Jasper Ridge (c) Urban

Fig. 1 Three real hyperspectral images

estimation of i-th endmember and its corresponding ground truth. Smaller SAD corresponds to better performance. Since other algorithms cannot generate a solution of problem (1.3) with small feasibility violation, to keep a fair comparison, we perform the rounding procedure and postprocessing on the solution generated by each method. We report in Table 4 the SAD and time cost for the three hyperspectral image datasets. From this table, we know that the efficiency of the proposed method is competitive to other algorithms. Particularly, our method achieves satisfying SAD among all algorithms. Besides, although EM-onmf is faster than our method on these datasets, the unmixing quality given by EM-onmf is unstable. The unmixing results of Samson,

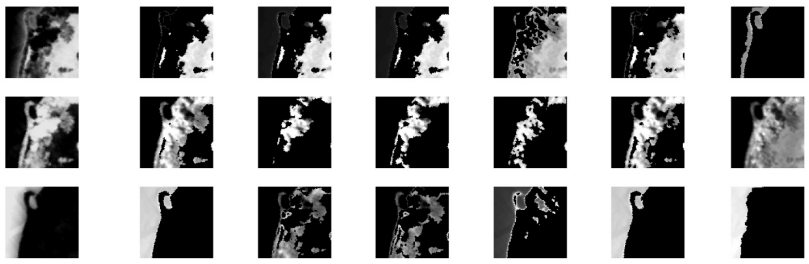
Table 4 Results on the hyperspectral image datasets. For "Samson", $r = 156, n = 95 \times 95, k = 3$, for "Jasper Ridge", $r = 198, n = 100 \times 100, k = 4$, for "Urban", $r = 162, n = 307 \times 307, k = 4$.

	Samson		Jasper Ridge		Urban	
method	SAD	time(s)	SAD	time(s)	SAD	time(s)
EP4Orth+	**0.081**	1.0	**0.150**	1.3	0.114	22
U-onmf	0.365	10	0.306	19	0.128	99
OPNMF	0.348	44	0.336	85	0.132	545
K-means	0.296	0.2	0.174	0.4	0.266	4
ONP-MF	0.085	16	0.276	34	0.112	339
EM-onmf	0.196	0.4	0.192	0.8	**0.091**	17

Jasper Ridge and Urban in illustrated in Figs. 2, 3 and 4, respectively. For Samson image, our method and ONP-MF are able to separate three endmembers, while the rest methods mix them together. For Jasper Ridge image, none of the methods can identify the road endmember, while our method and K-means can split water from other endmembers completely. All of algorithms perform relatively well on Urban dataset except for K-means, being able to separate four endmembers.

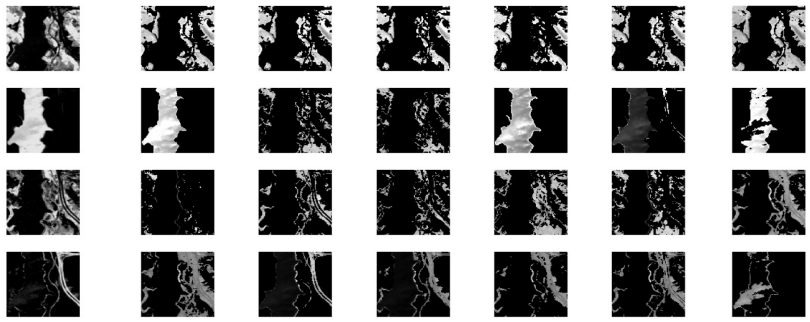
5.3 K-indicators model

We first remove the zero norm constraints from (1.5). The exact penalty model (4.34) with $p = 1$, $q = 2$, and $\epsilon = 0$ for solving the K-indicator model (1.5)



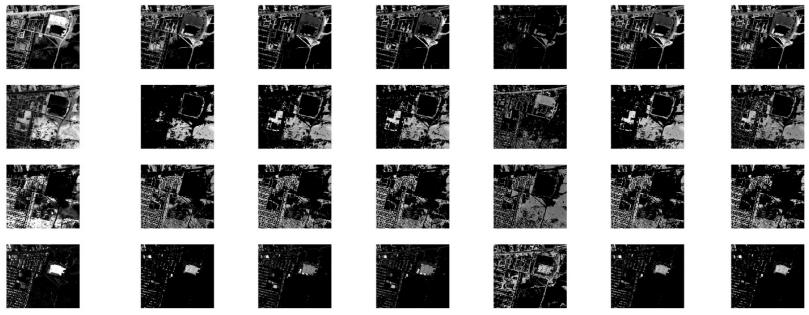
(a) truth (b) EP4Orth+ (c) U-onmf (d) OPNMF (e) K-means (f) ONP-MF (g) EM-onmf

Fig. 2 Unmixing results of Samson, from top to bottom: rock, tree, water.



(a) truth (b) EP4Orth+ (c) U-onmf (d) OPNMF (e) K-means (f) ONP-MF (g) EM-onmf

Fig. 3 Unmixing results of Jasper Ridge, from top to bottom: tree, water, dirt, road



(a) truth (b) EP4Orth+ (c) U-onmf (d) OPNMF (e) K-means (f) ONP-MF (g) EM-onmf

Fig. 4 Unmixing results of Urban, from top to bottom: asphal, grass, tree, roof

becomes

$$\min_{X\in\mathcal{OB}_{+}^{n,k},Y\in\mathcal{S}^{k,k}} \left\{\widehat{P}_\sigma(X,Y) := \|UY - X\|_{\mathrm{F}}^2 + \sigma\|XV\|_{\mathrm{F}}^2\right\}, \tag{5.3}$$

which is further equivalent to

$$\min_{X\in\mathcal{OB}_+^{n,k},Y\in\mathcal{S}^{k,k}} \left\{ P_\sigma(X,Y) := -\frac{1}{\sigma}\langle UY, X\rangle + \frac{1}{2}\|XV\|_{\mathrm{F}}^2 \right\}. \tag{5.4}$$

With a fixed Y, (5.4) is exactly (5.2) with $C = UY$. Similar to the discussion therein, we obtain the main PALM iterations [11] for solving (5.4) in Algorithm 2 as

$$Y^{l+1} = \Pi_{\mathcal{S}^{k,k}}\left(\beta^{-1}Y^l + U^\top X^l\right), \quad \beta > 0, \tag{5.5a}$$

$$X^{l+1} \in \Pi_{\mathcal{OB}_+^{n,k}}\left(X^l - \alpha\left(X^l VV^\top - UY^{l+1}/\sigma\right)\right), \quad 0 < \alpha < 1. \tag{5.5b}$$

Theorem 1 in [11] tells that the sequence $\{(X^l, Y^l)\}$ generated by (5.5a) and (5.5b) converges to a stationary point of (5.4). However, we find the convergence is slow if we fix the constant stepsizes α and β. Note that the closed form solution of (5.4) with respect to Y for a fixed $X = X^l$ is $\Pi_{\mathcal{S}^{k,k}}\left(U^\top X^l\right)$, which corresponds to setting $\beta = +\infty$ in (5.5a). For the tested problem, by some easy calculations, we can see $\alpha_{\mathtt{LBB}}^l \geq 1$. The practical PALM iterations for solving (5.4) is thus given as

$$Y^{l+1} = \Pi_{\mathcal{S}^{k,k}}\left(U^\top X^l\right). \tag{5.6a}$$

$$X^{l+1} \in \Pi_{\mathcal{OB}_+^{n,k}}\left(X^l - \alpha^l\left(X^l VV^\top - UY^{l+1}/\sigma\right)\right), \tag{5.6b}$$

where $\alpha^l = \min\{\alpha_{\mathtt{LBB}}^l, 10k\}$. The flops for (5.6a) and (5.6b) are $2nk^2 + O(k^3)$ and $2nk^2 + O(nk)$, respectively.

Chen et al. [20] proposed a semi-convex relaxation model to solve (1.5). Their intermediate model corresponds to (5.3) with $\sigma = 0$ and $\mathcal{OB}_+^{n,k}$ replaced by $\{X \in \mathbb{R}^{n\times k} : 0 \leq X \leq 1\}$. A double-layered alternating projection framework was investigated in [20] to solve the relaxation model. The method was named KindAP. To evaluate the efficiency of our method, we compare it with KindAP (downloaded from `https://github.com/yangyuchen0340/Kind`) on data clustering problems. We adopt eight datasets in Table 5 and perform post-processing on all datasets according to the KindAP algorithm [20]. We set $\gamma_2 = 10$, $\sigma_0 = 10$, $\eta = 0.5$ and $\mathrm{tol}^{\mathrm{feas}} = 0.1$ and $X^0 = \Pi_{\mathcal{OB}_+^{n,k}}(U)$ in our method. The initial points of KindAP is set as $\Pi_{\mathbb{R}_+^{n,k}}(U)$. Similar as in section 5.2.1, purity, entropy and NMI are adopted to judge the performance of proposed algorithms. The results are presented in Table 5.

It shows that the clustering results given by our methods are comparable to that provided by KindAP, which means both methods are able to solve (1.5) with a relatively high quality. On the other hand, our algorithm is generally 2-5 times faster than KindAP on most of the datasets. Our algorithm is especially efficient on datasets birch, worm, omniglot and UKbench, in which the numbers of samples and clusters are relatively large. Although we relax the zero norm constraints from problem (1.5), the obtained matrix X is always feasible to (1.5). By contrast, the matrix X returned by KindAP may not be

Table 5 Comparison of KindAP and our methods on data clustering problems. In the table, "a" and "b" stand for KindAP and EP4Orth+, respectively. Results marked in bold mean better performance in the corresponding index.

datasets	n	k	purity(%)		NMI(%)		entropy(%)		time(s)	
			a	b	a	b	a	b	a	b
CIFAR100-test [36]	10000	100	69.42	**69.44**	71.34	**71.36**	28.66	**28.64**	0.63	**0.41**
CIFAR100-train [36]	50000	100	99.63	**99.63**	99.57	**99.57**	0.43	**0.43**	3.13	**1.66**
COIL100 [47]	7200	100	**91.93**	**91.93**	97.30	**97.41**	2.70	**2.59**	1.47	**0.44**
birch [67]	100000	100	83.23	**83.44**	93.78	**93.84**	6.22	**6.16**	16.47	**5.22**
worm1 [52]	186432	67	47.87	**48.81**	**62.63**	62.39	**37.37**	37.61	48.78	**5.76**
worm2 [52]	165720	62	49.34	**50.31**	**63.50**	63.47	**36.50**	36.53	25.56	**6.46**
omniglot [38]	17853	1623	21.95	**21.97**	70.86	**70.94**	29.14	**29.06**	1176	**432**
UKBench [48]	10200	2550	90.64	**91.04**	97.64	**97.76**	2.36	**2.24**	3215	**1268**

an orthogonal nonnegative matrices although it always satisfies the zero norm constraints.

To end this section, we make some remarks on the numerical performance. Overall, our proposed algorithm can make some effective improvement over the best baselines in terms of the solution quality and speed. In addition, we provide an exact penalty algorithmic framework with convergence guarantee for solving optimization with nonnegative orthogonality while the existing methods focused on some specific models of (1.1) or (1.10).

6 Concluding remarks

In this paper, we consider optimization with nonnegative and orthogonality constraints. We focus on an equivalent formulation of the concerned problem, and show that the two formulations share the same minimizers and first- and second-order optimality conditions. By estimating a local error bound of $\mathcal{S}_+^{n,k}$, we provide a general class of exact and possibly smooth penalty models as well as a practical penalty algorithm with postprocessing. We investigate the asymptotic convergence of the penalty method and show that any limit point is a weakly stationary point of the concerned problem and becomes a stationary point under some more mild conditions. A second-order method for solving the penalty subproblem, namely, optimization with nonnegative and multiple spherical constraints, is also given. Our numerical results show that the proposed penalty method performs well for the problem of computing the orthogonal projection onto nonnegative orthogonality constraints, ONMF and the K-indicators model and it can always return high quality orthogonal nonnegative matrices.

Acknowledgements

The authors are grateful to the Co-Editor Dr. Adrian Lewis, the Associate Editor and the anonymous reviewers for their valuable comments and suggestions that helped to improve the quality of our manuscript.

A Construction of problem (5.1) with unique solution

Proposition A.1 *Choose $X^* \in \mathcal{S}_+^{n,k}$ and $L \in \mathbb{R}^{k\times k}$ with positive diagonal elements satisfying $L_{ii}L_{jj} > \max\{L_{ij}, L_{ji}, 0\}^2$ $\forall i, j \in [k], i \neq j$. Then the optimal solution of (5.1) with $C = X^*L^\top$ is unique and equals to X^*.*

Proof For simplicity of notation, we use $\sum_i$ to denote $\sum_{i\in[k]}$ in the proof. Since problem (5.1) is equivalent to $\max_{X\in\mathcal{S}_+^{n,k}} \langle C, X\rangle$, we only need to show that $\langle C, Y\rangle < \langle C, X^*\rangle = \sum_i L_{ii}$, $\forall\, \mathcal{S}_+^{n,k} \ni Y \neq X^*$. Let $Z = \mathsf{sgn}(Y)$ and $P = \Pi_{\mathbb{R}_+^n}(L)$. We have

$$\langle C, Y\rangle = \mathrm{tr}(L(X^*)^\top Y) = \sum\nolimits_i \sum\nolimits_j L_{ji}\mathbf{y}_i^\top \mathbf{x}_j^* \leq \sum\nolimits_i \sum\nolimits_j P_{ji}\mathbf{y}_i^T(\mathbf{x}_j^* \circ \mathbf{z}_i). \tag{A.1}$$

Define $w_{ji} = \|\mathbf{x}_j^* \circ \mathbf{z}_i\|^2$. With $X^* \in \mathcal{S}_+^{n,k}$, we have $\|\sum_j P_{ji}(\mathbf{x}_j^* \circ \mathbf{z}_i)\| = (\sum_i P_{ji}^2 w_{ji})^{1/2}$. Using the Cauchy-Schwarz inequality, $\|\mathbf{y}_i\| = 1$ and the requirements on L, we have

$$\sum\nolimits_j P_{ji}\mathbf{y}_i^T(\mathbf{x}_j^* \circ \mathbf{z}_i) \leq \Big(\sum\nolimits_j P_{ji}^2 w_{ji}\Big)^{\frac{1}{2}} \leq P_{ii}\Big(\sum\nolimits_j \frac{P_{jj}}{P_{ii}} w_{ji}\Big)^{\frac{1}{2}}. \tag{A.2}$$

With (A.1) and $\langle C, X^*\rangle = \sum_i L_{ii} = \sum_i P_{ii}$, we further have

$$\langle C, Y\rangle \leq \sum\nolimits_i P_{ii}\Big(\sum\nolimits_j \frac{P_{jj}}{P_{ii}} w_{ji}\Big)^{\frac{1}{2}} \leq \Big(\sum\nolimits_i P_{ii}\Big)^{\frac{1}{2}}\Big(\sum\nolimits_i \sum\nolimits_j P_{jj} w_{ji}\Big)^{\frac{1}{2}} \leq \langle C, X^*\rangle, \tag{A.3}$$

where the second inequality uses the fact that $\sum_i a_i x_i^{1/2} \leq (\sum_i a_i)^{1/2}(\sum_i a_i x_i)^{1/2}$ for $a_i > 0$ and $x_i \geq 0$, and the third inequality uses $\sum_i w_{ji} \leq 1$. Obviously, the equalities in (A.2) and (A.3) hold if and only if $Y = X^*$. The proof is completed. □

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization algorithms on matrix manifolds. Princeton University Press (2009)
2. Andreani, R., Haeser, G., Secchin, L.D., Silva, P.J.: New sequential optimality conditions for mathematical programs with complementarity constraints and algorithmic consequences. SIAM J. Optim. **29**(4), 3201–3230 (2019)
3. Andreani, R., Martínez, J.M., Ramos, A., Silva, P.J.: A cone-continuity constraint qualification and algorithmic consequences. SIAM J. Optim. **26**(1), 96–110 (2016)
4. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
5. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. Math. Program. **137**(1), 91–129 (2013)
6. Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA J. Numer. Anal. **8**(1), 141–148 (1988)
7. Bergmann, R., Herzog, R.: Intrinsic formulation of KKT conditions and constraint qualifications on smooth manifolds. SIAM J. Optim. **29**(4), 2423–2444 (2019)
8. Bertsekas, D.P.: Constrained optimization and Lagrange multiplier methods. Academic press (1996)
9. Bertsekas, D.P.: Nonlinear programming. Athena Scientific (1999)
10. Bioucasdias, J.M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., Chanussot, J.: Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(2), 354–379 (2012)

11. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1-2), 459–494 (2014)
12. Boumal, N.: An introduction to optimization on smooth manifolds. Available online, Aug (2020)
13. Boumal, N., Absil, P.A., Cartis, C.: Global rates of convergence for nonconvex optimization on manifolds. IMA J. Numer. Anal. **39**(1), 1–33 (2019)
14. Boutsidis, C., Drineas, P., Mahoney, M.W.: Unsupervised feature selection for the k-means clustering problem. In: NeurIPS, pp. 153–161 (2009)
15. Boutsidis, C., Gallopoulos, E.: SVD based initialization: a head start for nonnegative matrix factorization. Pattern Recogn. **41**(4), 1350–1362 (2008)
16. Byrd, R.H., Lopez-Calva, G., Nocedal, J.: A line search exact penalty method using steering rules. Math. Program. **133**(1-2), 39–73 (2012)
17. Cai, D., Mei, Q., Han, J., Zhai, C.: Modeling hidden topics on document manifold. In: Proceedings of the 17th ACM CIKM, pp. 911–920. ACM (2008)
18. Carson, T., Mixon, D.G., Villar, S.: Manifold optimization for k-means clustering. In: SampTA, pp. 73–77. IEEE (2017)
19. Chang, K.C., Pearson, K., Zhang, T.: Perron-Frobenius theorem for nonnegative tensors. Commun. Math. Sci. **6**(2), 507–520 (2008)
20. Chen, F., Yang, Y., Xu, L., Zhang, T., Zhang, Y.: Big-data clustering: K-means or k-indicators? arXiv:1906.00938 (2019)
21. Chen, X., Lu, Z., Pong, T.K.: Penalty methods for a class of non-Lipschitz optimization problems. SIAM J. Optim. **26**(3), 1465–1492 (2016)
22. Di Pillo, G.: Exact penalty methods. In: Algorithms for Continuous Optimization, pp. 209–253. Springer (1994)
23. Di Pillo, G., Grippo, L.: A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints. SIAM J. Control Optim. **23**(1), 72–84 (1985)
24. Di Pillo, G., Grippo, L.: An exact penalty function method with global convergence properties for nonlinear programming problems. Math. Program. **36**(1), 1–18 (1986)
25. Di Pillo, G., Lucidi, S.: An augmented Lagrangian function with improved exactness properties. SIAM J. Optim. **12**(2), 376–406 (2002)
26. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering. In: Proceedings of the 12th ACM SIGKDD, pp. 126–135. ACM (2006)
27. Estrin, R., Friedlander, M.P., Orban, D., Saunders, M.A.: Implementing a smooth exact penalty function for general constrained nonlinear optimization. SIAM J. Sci. Comput. **42**(3), A1836–A1859 (2020)
28. Friedlander, M.P., Tseng, P.: Exact regularization of convex programs. SIAM J. Optim. **18**(4), 1326–1350 (2008)
29. Gao, B., Liu, X., Yuan, Y.: Parallelizable algorithms for optimization problems with orthogonality constraints. SIAM J. Sci. Comput. **41**(3), A1949–A1983 (2019)
30. Hiriart-Urruty, J.B., Seeger, A.: A variational approach to copositive matrices. SIAM Review **52**(4), 593–629 (2010)
31. Hu, J., Jiang, B., Lin, L., Wen, Z., Yuan, Y.: Structured quasi-Newton methods for optimization with orthogonality constraints. SIAM J. Sci. Comput. **41**(4), A2239–A2269 (2019)
32. Hu, J., Milzarek, A., Wen, Z., Yuan, Y.: Adaptive quadratically regularized Newton method for Riemannian optimization. SIAM J. Matrix Anal. Appl. **39**(3), 1181–1207 (2018)
33. Jiang, B., Liu, Y.F., Wen, Z.: l_p-norm regularization algorithms for optimization over permutation matrices. SIAM J. Optim. **26**(4), 2284–2313 (2016)
34. Jiang, B., Meng, X., Wen, Z., Chen, X.: An exact penalty approach for optimization with nonnegative orthogonality constraints. arXiv 1907.12424v2 (2020)
35. Keshava, N., Mustard, J.F.: Spectral unmixing. IEEE Signal Process. Mag. **19**(1), 44–57 (2002)
36. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
37. Kuang, D., Ding, C., Park, H.: Symmetric nonnegative matrix factorization for graph clustering. In: Proceedings of the 2012 SDM, pp. 106–117. SIAM (2012)
38. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)

39. Li, B., Zhou, G., Cichocki, A.: Two efficient algorithms for approximately orthogonal nonnegative matrix factorization. IEEE Signal Process. Lett. **22**(7), 843–846 (2015)
40. Li, X., Sun, D., Toh, K.C.: On the efficient computation of a generalized Jacobian of the projector over the Birkhoff polytope. Math. Program. **179**, 419–446 (2020)
41. Liu, C., Boumal, N.: Simple algorithms for optimization on Riemannian manifolds with constraints. Appl. Math. Opt. **82**, 949–981 (2020)
42. Luo, D., Ding, C., Huang, H., Li, T.: Non-negative Laplacian embedding. In: 2009 Ninth ICDM, pp. 337–346. IEEE (2009)
43. Luo, Z.Q., Pang, J.S., Ralph, D.: Mathematical programs with equilibrium constraints. Cambridge University Press (1996)
44. Luo, Z.Q., Pang, J.S., Ralph, D., Wu, S.Q.: Exact penalization and stationarity conditions of mathematical programs with equilibrium constraints. Math. Program. **75**(1), 19–76 (1996)
45. Luo, Z.Q., Sturm, J.F.: Error bounds for quadratic systems. In: High performance optimization, pp. 383–404. Springer (2000)
46. Milzarek, A., Xiao, X., Cen, S., Wen, Z., Ulbrich, M.: A stochastic semismooth Newton method for nonsmooth nonconvex optimization. SIAM J. Optim. **29**(4), 2916–2948 (2019)
47. Nene, S.A., Nayar, S.K., Murase, H.: Columbia object image library (coil-100) (1996)
48. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR'06, vol. 2, pp. 2161–2168. IEEE (2006)
49. Pan, J., Ng, M.K.: Orthogonal nonnegative matrix factorization by sparsity and nuclear norm optimization. SIAM J. Matrix Anal. Appl. **39**(2), 856–875 (2018)
50. Pompili, F., Gillis, N., Absil, P.A., Glineur, F.: Two algorithms for orthogonal nonnegative matrix factorization with application to clustering. Neurocomputing **141**, 15–25 (2014)
51. Povh, J., Rendl, F.: A copositive programming approach to graph partitioning. SIAM J. Optim. **18**(1), 223–241 (2007)
52. Sieranoja, S., Fränti, P.: Fast and general density peaks clustering. Pattern Recogn. Lett. **128**, 551–558 (2019)
53. Sun, W., Yuan, Y.: Optimization theory and methods: nonlinear programming, vol. 1. Springer Science & Business Media (2006)
54. Wang, S., Chang, T.H., Cui, Y., Pang, J.S.: Clustering by orthogonal non-negative matrix factorization: a sequential non-convex penalty approach. In: ICASSP, pp. 5576–5580 (2019)
55. Wang, S., Chang, T.H., Cui, Y., Pang, J.S.: Clustering by orthogonal NMF model and non-convex penalty optimization. IEEE Trans. Signal Process. (2021)
56. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Math. Program. **142**(1), 397–434 (2013)
57. Xiao, X., Li, Y., Wen, Z., Zhang, L.: A regularized semi-smooth Newton method with projection steps for composite convex programs. J. Sci. Comput. **76**, 364–389 (2016)
58. Yang, L.: Proximal gradient method with extrapolation and line search for a class of nonconvex and nonsmooth problems. arXiv:1711.06831 (2017)
59. Yang, W.H., Zhang, L.H., Song, R.: Optimality conditions for the nonlinear programming problems on Riemannian manifolds. Pac. J. Optim. **10**(2), 415–434 (2014)
60. Yang, Y., Yang, Y., Shen, H.T., Zhang, Y., Du, X., Zhou, X.: Discriminative nonnegative spectral clustering with out-of-sample extension. IEEE Trans. Knowl. Data Eng. **25**(8), 1760–1771 (2012)
61. Yang, Z., Oja, E.: Linear and nonlinear projective nonnegative matrix factorization. IEEE Trans. Neural Netw. **21**(5), 734–749 (2010)
62. Yoo, J., Choi, S.: Orthogonal nonnegative matrix factorization: multiplicative updates on Stiefel manifolds. In: IDEAL, pp. 140–147. Springer (2008)
63. Zass, R., Shashua, A.: Nonnegative sparse PCA. In: NeurIPS, pp. 1561–1568 (2007)
64. Zhang, H., Hager, W.W.: A nonmonotone line search technique and its application to unconstrained optimization. SIAM J. Optim. **14**(4), 1043–1056 (2004)
65. Zhang, J., Liu, H., Wen, Z., Zhang, S.: A sparse completely positive relaxation of the modularity maximization for community detection. SIAM J. Sci. Comput. **40**(5), A3091–A3120 (2018)

66. Zhang, K., Zhang, S., Liu, J., Wang, J., Zhang, J.: Greedy orthogonal pivoting algorithm for non-negative matrix factorization. In: ICML, pp. 7493–7501. PMLR (2019)
67. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: A new data clustering algorithm and its applications. Data Min. Knowl. Disc. **1**(2), 141–182 (1997)
68. Zhu, F., Wang, Y., Fan, B., Xiang, S., Meng, G., Pan, C.: Spectral unmixing via data-guided sparsity. IEEE Trans. Image Process. **23**(12), 5412–5427 (2014)