ORTHOGONAL TRACE-SUM MAXIMIZATION: APPLICATIONS, LOCAL ALGORITHMS, AND GLOBAL OPTIMALITY[*]

JOONG-HO WON[†], HUA ZHOU[‡], AND KENNETH LANGE[§]

Abstract. This paper studies the problem of maximizing the sum of traces of matrix quadratic forms on a product of Stiefel manifolds. This orthogonal trace-sum maximization (OTSM) problem generalizes many interesting problems such as generalized canonical correlation analysis (CCA), Procrustes analysis, and cryo-electron microscopy of the Nobel prize fame. For these applications finding global solutions is highly desirable, but it has been unclear how to find even a stationary point, let alone test its global optimality. Through a close inspection of Ky Fan's classical result [*Proc. Natl. Acad. Sci. USA*, 35 (1949), pp. 652–655] on the variational formulation of the sum of largest eigenvalues of a symmetric matrix, and a semidefinite programming (SDP) relaxation of the latter, we first provide a simple method to certify global optimality of a given stationary point of OTSM. This method only requires testing whether a symmetric matrix is positive semidefinite. A by-product of this analysis is an unexpected strong duality between Shapiro and Botha [*SIAM J. Matrix Anal. Appl.*, 9 (1988), pp. 378–383] and Zhang and Singer [*Linear Algebra Appl.*, 524 (2017), pp. 159–181]. After showing that a popular algorithm for generalized CCA and Procrustes analysis may generate oscillating iterates, we propose a simple fix that provably guarantees convergence to a stationary point. The combination of our algorithm and certificate reveals novel global optima of various instances of OTSM.

Key words. Stiefel manifold, canonical correlation analysis, Procrustes analysis, cryo-EM, semidefinite programming, MAXBET, MAXDIFF, MM algorithm, Kurdyka–Łojasiewicz inequality

AMS subject classifications. 15B10, 49K30, 49M20, 65K05, 90C26

DOI. 10.1137/20M1363388

1. Introduction.

1.1. Orthogonal trace-sum maximization. Given $S_{ij} = S_{ji}^T \in \mathbb{R}^{d_i \times d_j}$ for $i, j = 1, \ldots, m$, and $r \le \min_{i=1,\ldots,m} d_i$, we are interested in solving the following optimization problem:

$$\text{(OTSM)} \quad \text{maximize} \quad \frac{1}{2} \sum_{i,j=1}^m \text{tr}(O_i^T S_{ij} O_j) \quad \text{subject to} \quad O_i \in \mathcal{O}_{d_i,r}, \quad i = 1, \ldots, m,$$

where $\mathcal{O}_{d,r} = \{O \in \mathbb{R}^{d \times r} : O^T O = I_r\}$ is the Stiefel manifold of (partially) orthogonal matrices [8]; I_r denotes the identity matrix of order r. Since adding a positive multiple of the identity matrix to S_{ii} only increases the objective by a constant amount, without loss of generality we can assume each S_{ii} is positive semidefinite. In what follows, we call (OTSM) the *orthogonal trace-sum maximization* problem. OTSM arises in many interesting settings, as follows.

[†]Department of Statistics, Seoul National University, Seoul 08826, Korea (wonj@stats.snu.ac.kr).
[‡]Department of Biostatistics, University of California, Los Angeles, CA 90095-1766 USA (huazhou@ucla.edu).
[§]Departments of Computational Medicine, Human Genetics, and Statistics, University of California, Los Angeles, CA 90095 USA (klange@ucla.edu).

Canonical correlation analysis. Canonical correlation analysis (CCA) [18] seeks directions to maximize the correlation between two sets of n observations of variables of possibly different dimensions, $A_1 \in \mathbb{R}^{n \times d_1}$ and $A_2 \in \mathbb{R}^{n \times d_2}$:

$$\text{maximize} \quad \text{corr}(A_1 t_1, A_2 t_2) \quad \text{subject to} \quad t_i^T t_i = 1, \ i = 1, 2,$$

where $t_i \in \mathbb{R}^{d_i}$ are the optimization variables, and $\text{corr}(\cdot, \cdot)$ denotes the Pearson correlation coefficient between two sample vectors. Generalizations of CCA (i) handle more than two sets of variables $A_1, \ldots, A_m$ ($m \geq 2$), and (ii) seek partial rotation matrices (as opposed to vectors) of A_i's to achieve maximal agreement. The popular MAXDIFF and MAXBET criteria [15, 21, 31, 34] solve

$$\text{(MAXDIFF)} \quad \text{maximize} \quad \sum_{i<j} \text{tr}(O_i^T A_i^T A_j O_j) \quad \text{subject to} \quad O_i \in \mathcal{O}_{d_i, r}, \quad i = 1, \ldots, m;$$

$$\text{(MAXBET)} \quad \text{maximize} \quad \frac{1}{2} \sum_{i,j=1}^{m} \text{tr}(O_i^T A_i^T A_j O_j) \quad \text{subject to} \quad O_i \in \mathcal{O}_{d_i, r}, \quad i = 1, \ldots, m.$$

Both MAXDIFF and MAXBET are instances of (OTSM) with $S_{ij} = A_i^T A_j$ (if MAXDIFF, $S_{ii} = 0$), for $i, j = 1, \ldots, m$. It is worth noting that when $d_1 = \cdots = d_m = d = r$, i.e., the fully orthogonal case, MAXDIFF coincides with MAXBET up to an additive constant.

Procrustes analysis and little Grothendieck problem. If the variables are fully orthogonal and there exist $S_{11}, S_{12}, \ldots, S_{mm} \in \mathbb{R}^{d \times d}$ such that the symmetric $md \times md$ block matrix $\tilde{S} = (S_{ij})_{i,j=1}^m$ is positive semidefinite (denoted $S \succeq 0$), then (OTSM) reduces to the little Grothendieck problem over the orthogonal group [5], which arises in generalized Procrustes analysis [13, 14, 30]. Given a collection of n landmarks of d-dimensional images $A_i \in \mathbb{R}^{n \times d}$, $i = 1, \ldots, m$, the goal is to find orthogonal matrices that minimize the pairwise discrepancy

$$(1.1) \quad f(O_1, \ldots, O_m) = \frac{1}{2} \sum_{i,j=1}^{m} \|A_i O_i - A_j O_j\|_F^2 = -\sum_{i<j} \text{tr}(O_i^T A_i^T A_j O_j) + \text{const.}$$

subject to the constraints that $O_i \in \mathcal{O}_{d,d}$ for all i, where $\| \cdot \|_F$ is the Frobenius norm. This problem is a special case of (OTSM) with $S_{ij} = A_i^T A_j$ for $i, j = 1, \ldots, m$. Clearly, $\tilde{S} = [A_1, \ldots, A_m]^T [A_1, \ldots, A_m] \succeq 0$. When $m = 2$, problem (1.1) reduces to ordinary (partial) Procrustes analysis [10, Chapter 7].

Cryo-EM and orthogonal least squares. Another instance of (OTSM) involving fully orthogonal matrices is the least squares regression problem that minimizes the squared Frobenius norm of the difference between a given $n \times d$ matrix A_{K+1} and linear combination $\sum_{i=1}^{K} A_i O_i$ of given $n \times d$ matrices A_i with $O_i \in \mathcal{O}_{d,d}$, $i = 1, \ldots, K$. This least-squares problem has a direct application in single-particle reconstruction with cryo-electron microscopy (cryo-EM) celebrated by the 2017 Nobel Prize in Chemistry. Then we can equivalently minimize

$$(1.2) \quad \frac{1}{2} \left\| A_{K+1}(-O_{K+1}) - \sum_{i=1}^{K} A_i O_i \right\|_F^2 = \sum_{i<j} \text{tr}(O_i^T A_i^T A_j O_j) + \text{const.}$$

subject to the orthogonality constraints on $O_1, \ldots, O_{K+1}$. Any minimizer $(\tilde{O}_1, \ldots, \tilde{O}_{K+1})$ of (1.2) supplies a minimizer $(-\tilde{O}_1 \tilde{O}_{K+1}^T, \ldots, -\tilde{O}_K \tilde{O}_{K+1}^T)$ of the original problem. This is a special case of (OTSM) with $S_{ij} = -A_i^T A_j$. In cryo-EM, reconstruction

of the three-dimensional (3D) map of a particle involves estimating viewing directions of its two-dimensional (2D) projections. Retrieval of the orthogonal matrices representing the orientations is posed as the above least squares problem [6, 33, 35, 37].

1.2. Global solutions of OTSM. Each instance of (OTSM) above can be posed as a maximum likelihood estimation problem under an appropriate model. Finding its global solution is highly desirable for correct inference. While it attains a maximum because each $\mathcal{O}_{d_i,r}$ is compact and the objective function is continuous in $\mathbb{R}^{d_1 \times r} \times \cdots \times \mathbb{R}^{d_m \times r}$, (OTSM) is a nonconvex optimization problem since the constraint set $\mathcal{O}_{d_1,r} \times \cdots \times \mathcal{O}_{d_m,r}$ is nonconvex. Except for the special case of $m = 2$ in which an analytic global maximizer can be found using the singular value decomposition (SVD) [13, 34], we generally have to resort to iterative methods. The nonconvexity of the problem makes it difficult to test global optimality of a candidate (local) solution.

To add further difficulties, the global solution to (OTSM) is not unique. If $(O_1^\star, \ldots, O_m^\star)$ is a solution to (OTSM), then for any $R \in \mathcal{O}_{r,r}$, $(O_1^\star R, \ldots, O_m^\star R)$ is also a solution.

1.3. Contributions. The contributions of this paper are as follows: (i) providing a simple certificate that guarantees the *global* optimality of a *local* stationary point of problem (OTSM) (section 3) and (ii) showing that a standard algorithm for generalized CCA and Procrustes analysis may exhibit oscillation, and proposing an efficient proximal block relaxation algorithm with a convergence guarantee to a stationary point (section 4). Our certificate and duality results are developed in close analogy to the classical result by Ky Fan [11] on the variational formulation of the sum of largest eigenvalues of a symmetric matrix (section 2). (In the accompanying supplementary material file OTSM_supp.pdf [local/web 625KB], we also establish a duality between problem (OTSM) and another eigenvalue optimization problem. As a special case, a strong duality between two separately known results in the literature [28, 40] is shown.) The certificate only requires testing positive semidefiniteness of a symmetric matrix constructed from a stationary point and data. Therefore, it is simple to verify global optimality. The convergence theory for the proposed algorithm proves that the *whole* sequence $\{O^k = (O_1^k, \ldots, O_m^k)\}$ of iterates converges to a stationary point at least at a sublinear rate—this result is stronger than convergence of the objective value sequence or convergence of a subsequence of $\{O^k\}$. To the best of our knowledge, there has been no convergence result of this stronger kind for the related problems. Some numerical results of the proposed algorithm combined with the certificate are presented in section 5.

2. Preliminary: The $m = 1$ case and Ky Fan theorem. As a preparation for what follows, we review the classical results on variational formulations of the sum of r largest eigenvalues of a symmetric matrix. For a matrix S in the vector space of $d \times d$ symmetric real matrices (denoted $\mathbb{S}^d$), let $\lambda_i(S)$ be the ith largest eigenvalue of S. Then it is well known that

$$(2.1) \qquad \sum_{i=1}^{r} \lambda_i(S) = \max_{O \in \mathcal{O}_{d,r}} \text{tr}(O^T S O) = \max_{U \in \mathbb{S}^d} \left\{ \text{tr}(SU) : 0 \preceq U \preceq I_d, \ \text{tr}(U) = r \right\}.$$

The first equality is the celebrated Ky Fan theorem [11], where the involved nonconvex optimization problem over a Stiefel manifold is a special case of (OTSM) for $m = 1$. The second equality is due to [17, 25], which state that there always is a tight convex relaxation of Ky Fan's nonconvex problem. It is also well known that the dual of this

convex semidefinite programming (SDP) problem is an SDP

$$
\text{(2.2)} \qquad
\begin{aligned}
&\text{minimize} \quad rz + \operatorname{tr}(M), \\
&\text{subject to} \quad zI_d + M - L = S, \ M \succeq 0, \ L \succeq 0
\end{aligned}
$$

for variables $M, L \in \mathbb{S}^d$ and $z \in \mathbb{R}$ [3, 23, 26].

Let us examine the relation between stationary points of these optimization problems closely. For Ky Fan's nonconvex problem, the Lagrangian is

$$
\mathcal{L}(O, \Lambda) = -\operatorname{tr}(O^T S O) + \frac{1}{2} \operatorname{tr}\left[\Lambda(O^T O - I_r)\right],
$$

by rewriting the constraint $O \in \mathcal{O}_{d,r}$ as an equality constraint $O^T O = I_r$. The Lagrange multiplier matrix Λ is symmetric due to the symmetry of the corresponding constraint. Point $O \in \mathcal{O}_{d,r}$ is a stationary point if the directional derivative of $\mathcal{L}$ with respect to $W \in \mathbb{R}^{d \times r}$

$$
d_W \mathcal{L} = -\operatorname{tr}\left[(SO)^T W\right] + \operatorname{tr}\left[(O\Lambda)^T W\right]
$$

vanishes for any W, i.e., if O satisfies the necessary condition for first-order local optimality. This is equivalent to the existence of a symmetric matrix Λ satisfying

$$
\text{(2.3)} \qquad\qquad\qquad O\Lambda = SO.
$$

Further, using the constraint $O^T O = I_r$, we have a representation $\Lambda = O^T S O \in \mathbb{S}^r$. The Karush–Kuhn–Tucker (KKT) optimality conditions assert that a locally maximal point is also stationary [24, Theorem 12.1].

The second-order necessary condition for a local maximum is

$$
\text{(2.4)} \qquad\qquad d_W^2 \mathcal{L} = \operatorname{tr}(\Lambda W^T W) - \operatorname{tr}(W^T S W) \geq 0
$$

for all $W \in \mathbb{R}^{d \times r}$ such that $W^T O + O^T W = 0$ [24, Theorem 12.5]; the set of such W is the tangent space of the Stiefel manifold $\mathcal{O}_{d,r}$ at O. For the convex problem (either the primal or dual), the KKT conditions are

$$
\text{(2.5)} \qquad
\begin{aligned}
&0 \preceq U \preceq I_d, \quad \operatorname{tr}(U) = r, \quad M \succeq 0, \quad M + zI_d - L = S, \\
&\operatorname{tr}(LU) = 0, \quad \operatorname{tr}[M(I_d - U)] = 0, \quad L \succeq 0.
\end{aligned}
$$

Assume that $\bar{O}$ is a local maximizer. Let $\bar{\Lambda} = \bar{O}^T S \bar{O}$, $\bar{U} = \bar{O}\bar{O}^T$, and $\bar{M} = \bar{O}(\bar{\Lambda} - \bar{z}I_r)\bar{O}^T$ for $\bar{z} = \lambda_{\min}(\bar{\Lambda})$, the smallest eigenvalue of $\bar{\Lambda}$. Since $\bar{O}$ is a stationary point, it is easy to verify that $(\bar{U}, \bar{M}, \bar{L} = \bar{M} + \bar{z}I_d - S)$ satisfies all the KKT conditions in (2.5) but $\bar{L} \succeq 0$.[1] If $r < d$, let $\bar{O}^\perp \in \mathcal{O}_{d-r,r}$ consist of orthonormal columns that span the null space of $\bar{O}$, and choose $W = \bar{O}^\perp K$ for $K \in \mathbb{R}^{(d-r) \times r}$ so that it satisfies $W^T \bar{O} + \bar{O}^T W = 0$. Then after some algebra,

$$
\operatorname{tr}(\bar{\Lambda} W^T W) - \operatorname{tr}(W^T S W) = \operatorname{tr}(K\bar{\Lambda}K^T) - \operatorname{tr}(K^T \bar{O}^{\perp T} S \bar{O}^\perp K) \geq 0.
$$

Since K is arbitrary, choose $K = \beta v_{\min}^T$ for $\beta \in \mathbb{R}^{d-r}$ and $v_{\min} \in \mathbb{R}^r$, where the latter is the unit eigenvector of $\bar{\Lambda}$ associated with $\bar{z}$. So,

$$
\operatorname{tr}(K\bar{\Lambda}K^T) - \operatorname{tr}(K^T \bar{O}^{\perp T} S \bar{O}^\perp K) = \bar{z}\beta\beta^T - \beta^T \bar{O}^{\perp T} S \bar{O}^\perp \beta \geq 0 \quad \forall \beta \in \mathbb{R}^{d-r}.
$$

---

[1] The construction of matrices $\bar{M}$ and $\bar{L}$ is inspired by [26, Theorem 3.3], in which optimality conditions of the dual SDP (2.2) is given. In particular, if $\bar{O}$ consists of the orthonormal eigenvectors of S corresponding to the r largest eigenvalues, it follows that $\lambda_r(S) \geq \bar{z} \geq \lambda_{r+1}(S)$, $\bar{M} = \bar{O}\operatorname{diag}(\lambda_1(S) - \bar{z}, \ldots, \lambda_r(S) - \bar{z})\bar{O}^T$ and $\bar{L} = \bar{O}^\perp \operatorname{diag}(\bar{z} - \lambda_{r+1}(S), \ldots, \bar{z} - \lambda_d(S))\bar{O}^{\perp T}$.

This implies $\bar{z}I_{d-r} - \bar{O}^{\perp T} S \bar{O}^{\perp} \succeq 0$. On the other hand, as any $y \in \mathbb{R}^d$ can be written as $y = \bar{O}\alpha + \bar{O}^{\perp}\beta$ for $\alpha \in \mathbb{R}^r$ and $\beta \in \mathbb{R}^{d-r}$, it can be easily seen that

$$y^T(\bar{M} + \bar{z}I_d - S)y = \beta^T(\bar{z}I_{d-r} - \bar{O}^{\perp T}S\bar{O}^{\perp})\beta \geq 0.$$

Thus $\bar{L} = \bar{M} + \bar{z}I_d - S \succeq 0$ and $\bar{O}$ satisfies all the KKT conditions. (If $r = d$, it is immediate that $\bar{M} = S - \bar{z}I_d$, or $\bar{L} = 0$.) This implies that all the local maxima of Ky Fan's nonconvex optimization problem are global maxima.

Conversely, if a stationary point $\bar{O}$ satisfies $\bar{L} \triangleq \bar{M} + \bar{z}I_d - S \succeq 0$ for $\bar{\Lambda}$, $\bar{U}$, and $\bar{M}$ constructed as in the beginning of the previous paragraph, then it is also globally optimal, which obviously is locally maximal. In other words, $\bar{L} \succeq 0 \implies$ (global optimum) $\implies$ (local maximum) $\implies \bar{L} \succeq 0$; hence $\bar{L} \succeq 0$ is *necessary and sufficient* for a stationary point to be globally optimal.

The above analysis of Ky Fan's problem (2.1) sheds light on (OTSM) in three ways. First, there can be a tight convex relaxation to (OTSM). Second, by analyzing the KKT conditions of the convex relaxation, we may be able to certify a stationary point of (OTSM) to be globally optimal. Third, the dual of the convex relaxation may have to do with a sum of the eigenvalues of a block matrix constructed from S_{ij}'s. In what follows, we carry out an analysis of the OTSM problem inspired by the Ky Fan problem. The added complexity due to $m > 1$ reveals both similarities and differences between the two problems.

3. Certificate of global optimality.

3.1. Local optimality conditions.

3.1.1. First-order conditions. Rewriting the constraints $O_i \in \mathcal{O}_{d_i,r}$ as equality constraints $O_i^T O_i = I_r$, the Lagrangian of (OTSM) is

$$\mathcal{L}(O_1,\ldots,O_m,\Lambda_1,\ldots,\Lambda_m) = -\frac{1}{2}\sum_{i,j=1}^{m} \mathrm{tr}(O_i^T S_{ij} O_j) + \frac{1}{2}\sum_{i=1}^{m} \mathrm{tr}\left[\Lambda_i(O_i^T O_i - I_r)\right],$$

where the Lagrange multiplier matrices Λ_i are symmetric due to the symmetry of the corresponding constraints. In parallel with section 2, a point $O = (O_1,\ldots,O_m)$ is a stationary point of problem (OTSM) if the directional derivative of $\mathcal{L}$ with respect to $W = (W_1,\ldots,W_m) \in \mathbb{R}^{d_1 \times r} \times \cdots \times \mathbb{R}^{d_m \times r}$

$$(3.1) \qquad d_W\mathcal{L} = -\sum_{i=1}^{m}\sum_{j=1}^{m} \mathrm{tr}\left[(S_{ij}O_j)^T W_i\right] + \sum_{i=1}^{m} \mathrm{tr}\left[(O_i\Lambda_i)^T W_i\right]$$

vanishes for any W. A local maximum satisfies condition (3.1), which is equivalent to

$$(3.2) \qquad O_i\Lambda_i = \sum_{j=1}^{m} S_{ij}O_j, \quad i = 1,\ldots,m,$$

resembling the first-order condition (2.3) of the Ky Fan problem.

Using the constraint $O_i^T O_i = I_r$, we further have a representation for Λ_i:

$$(3.3) \qquad \Lambda_i = O_i^T\left(\sum_{j=1}^{m} S_{ij}O_j\right) = \left(\sum_{j=1}^{m} S_{ij}O_j\right)^T O_i.$$

The second equality follows from the symmetry of the Lagrange multiplier. Substituting this quantity in (3.2), we obtain

$$(3.4) \qquad O_i \left(\sum_{j=1}^m S_{ij} O_j \right)^T O_i = \sum_{j=1}^m S_{ij} O_j, \quad i = 1, \dots, m.$$

3.1.2. Second-order condition. The second-order necessary condition for local maximality of (OTSM) is

$$(3.5) \qquad d_W^2 \mathcal{L} = \sum_{i=1}^m \operatorname{tr}(\Lambda_i W_i^T W_i) - \operatorname{tr}(W^T \tilde{S} W) \geq 0$$

for all $W = [W_1^T, \dots, W_m^T]^T$, such that W_i is a tangent vector of $\mathcal{O}_{d_i,r}$ at O_i, i.e.,

$$(3.6) \qquad W_i^T O_i + O_i^T W_i = 0, \quad i = 1, \dots, m,$$

where $\tilde{S} = (S_{ij})$ is the symmetric $D \times D$ block matrix whose (i,j) block is $S_{ij} \in \mathbb{R}^{d_i \times d_j}$, $i, j = 1, \dots, m$. Again see the resemblance of condition (3.5) to the second-order condition (2.4) for Ky Fan's problem.

Unfortunately, in (OTSM) we do not have the luxury of all the local maxima being globally optimal. We revisit this issue after studying a potentially tight convex relaxation to the problem in the next subsection. As a partial result, some of the following characterization of the Lagrange multipliers associated with a local or global maximum can be deduced from [31, pp. 489–490].

PROPOSITION 3.1. *If $O = (O_1, \dots, O_m) \in \mathcal{O}_{d_1,r} \times \cdots \times \mathcal{O}_{d_m,r}$ is a local maximizer of* (OTSM), *then Λ_i as defined in (3.3) is positive semidefinite, for $i \in \{1, \dots, m\}$ such that $d_i > r$. If O is a stationary point but Λ_i is not positive semidefinite for some i, then one can find another stationary point $\tilde{O} = (\tilde{O}_1, \dots, \tilde{O}_m)$ such that $\frac{1}{2} \sum_{i,j} \operatorname{tr}(O_i S_{ij} O_j) < \frac{1}{2} \sum_{i,j} \operatorname{tr}(\tilde{O}_i S_{ij} \tilde{O}_j)$ and $\tilde{\Lambda}_i = \tilde{O}_i^T \sum_{j=1}^m S_{ij} \tilde{O}_j$ is positive semidefinite for all $i = 1, \dots, m$. Furthermore, if O is a global maximizer of* (OTSM), *then Λ_i is positive semidefinite for all i.*

A full proof is provided in Appendix A.

3.2. Semidefinite programming relaxation. By introducing an appropriate matrix variable and constraints, we can obtain an upper bound of the optimal value of (OTSM) by that of an SDP relaxation. Besides providing tight bounds, the SDP formulation paves the way toward certifying the global optimality of a local solution. If $D = \sum_{i=1}^m d_i$, then we can define a $D \times D$ matrix

$$(3.7) \qquad U \triangleq \frac{1}{m} OO^T, \quad O = [O_1^T, \dots, O_m^T]^T \in \mathbb{R}^{D \times r},$$

so that $\sum_{i<j} \operatorname{tr}(O_i^T S_{ij} O_j)$, the objective function of (OTSM), is equal to $\frac{m}{2} \operatorname{tr}(\tilde{S} U)$, where $\tilde{S} = (S_{ij})$. We can express (OTSM) in terms of the matrix U by imposing appropriate constraints. The proof of the following proposition is in Appendix A.

PROPOSITION 3.2. *Problem* (OTSM) *is equivalent to the optimization problem*
(3.8)

$$\text{maximize} \quad (m/2) \operatorname{tr}(\tilde{S} U)$$
$$\text{subject to} \quad U \succeq 0, \ \operatorname{rank}(U) = r, \ mU_{ii} \preceq I_{d_i}, \ \operatorname{tr}(mU_{ii}) = r, \quad i = 1, \dots, m,$$

where the optimization variable is a symmetric $D \times D$ matrix U, U_{ii} denotes the ith diagonal block of U whose size is $d_i \times d_i$, and $A \preceq B$ denotes the Loewner order, i.e., $B - A \succeq 0$.

By dropping the rank constraint from problem (3.8) we obtain a convex, SDP relaxation of (OTSM):

(P-SDP) $\quad\quad$ maximize $\quad (m/2) \operatorname{tr}(\tilde{S}U)$
$\quad\quad\quad\quad\quad\quad$ subject to $\quad U \succeq 0, \; mU_{ii} \preceq I_{d_i}, \; \operatorname{tr}(mU_{ii}) = r, \quad i = 1, \ldots, m.$

This relaxation is tight if the solution $U^\star$ has rank r. The solution to (OTSM) is recovered by the decomposition (3.7). The dual of (P-SDP) is easily seen to be the following SDP:

(D-SDP) $\quad\quad$ minimize $\quad \sum_{i=1}^{m}[rz_i + \operatorname{tr}(M_i)]$
$\quad\quad\quad\quad\quad\quad$ subject to $\quad Z + M - L = \tilde{S}, \quad L \succeq 0, \quad M_i \succeq 0, \quad i = 1, \ldots, m,$

where $Z = \operatorname{diag}(mz_1 I_{d_1}, \ldots, mz_m I_{d_m})$ and $M = \operatorname{diag}(mM_1, \ldots, mM_m)$. The optimization variables are $L \in \mathbb{S}^D$, $M_i \in \mathbb{S}^{d_i}$, $z_i \in \mathbb{R}$, $i = 1, \ldots, m$. Strong duality between (P-SDP) and (D-SDP) holds (e.g., Slater's condition is satisfied). A rank-r solution to the SDP relaxation (P-SDP), if it exists, yields a globally optimal solution to the original problem (OTSM). However, solving these convex programs is computationally challenging even with modern convex optimization solvers due to their lifted dimensions. Moreover, if the optimal SDP solution U has rank greater than r, the factor O in (3.7) is infeasible to the original problem (OTSM).

Thus it is natural to ask when the candidate rank-r solution (3.7) to (P-SDP) constructed from a stationary point $(O_1, \ldots, O_m)$ of (OTSM) becomes actually an optimal solution. If this is the case, then the local solution globally solves (OTSM). We explore this path in the next subsection.

3.3. Certifying global optimality of a stationary point. The KKT conditions for (P-SDP) and (D-SDP) are

(KKT-a) $\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad U \succeq 0,$

(KKT-b) $\quad\quad\quad\quad\quad\quad\quad\quad mU_{ii} \preceq I_{d_i}, \quad i = 1, \ldots, m,$

(KKT-c) $\quad\quad\quad\quad\quad\quad\quad \operatorname{tr}(mU_{ii}) = r,$

(KKT-d) $\quad\quad\quad\quad\quad\quad\quad\quad\quad M_i \succeq 0, \quad i = 1, \ldots, m,$

(KKT-e) $\quad\quad\quad\quad\quad\quad Z + M - L = \tilde{S},$

(KKT-f) $\quad\quad\quad\quad\quad\quad\quad\quad \operatorname{tr}(LU) = 0,$

(KKT-g) $\quad\quad\quad\quad \operatorname{tr}(M_i(I_{d_i} - mU_{ii})) = 0, \quad i = 1, \ldots, m,$

(KKT-h) $\quad\quad\quad\quad\quad\quad\quad\quad\quad L \succeq 0,$

where $Z = \operatorname{diag}(mz_1 I_{d_1}, \ldots, mz_m I_{d_m})$ for $z_1, \ldots, z_m \in \mathbb{R}$ and $M = \operatorname{diag}(mM_1, \ldots, mM_m)$. If any tuple (U, Z, M, L) satisfies conditions (KKT-a)–(KKT-h), then U is an optimal solution to (P-SDP) and (Z, M, L) is optimal for (D-SDP) [36].

Now suppose $\bar{O} = (\bar{O}_1, \ldots, \bar{O}_m)$ is a stationary point of (OTSM). Recalling (3.3), let the associated Lagrange multipliers be $\bar{\Lambda}_i = \sum_{j=1}^{m} \bar{O}_i^T S_{ij} \bar{O}_j$. We can find the quantities that satisfy the KKT conditions above in a similar, but not completely obvious, manner to section 2. The matrix

(3.9) $$\bar{U} \triangleq \frac{1}{m}\bar{O}\bar{O}^T$$

clearly satisfies (KKT-a), (KKT-b), and (KKT-c). Now let τ_i be the smallest eigenvalue of the symmetric matrix $\bar{\Lambda}_i$. Then

$$(3.10) \qquad \bar{M}_i \triangleq \frac{1}{m}\bar{O}_i\bar{\Lambda}_i\bar{O}_i^T - \bar{z}_i\bar{O}_i\bar{O}_i^T = \bar{O}_i\left(\frac{1}{m}\bar{\Lambda}_i - \bar{z}_iI_r\right)\bar{O}_i^T$$

satisfies (KKT-d) for any $\bar{z}_i \leq \tau_i/m$. If we define block diagonal matrices

$$(3.11) \qquad \bar{Z} = \mathrm{diag}(m\bar{z}_1I_{d_1},\ldots,m\bar{z}_mI_{d_m}), \quad \bar{M} = \mathrm{diag}(m\bar{M}_1,\ldots,m\bar{M}_m),$$

then

$$\begin{aligned}
\mathrm{tr}\left[(\bar{M}+\bar{Z})\bar{U}\right] &= \sum_{i=1}^m \mathrm{tr}\left[(\bar{M}_i+\bar{z}_iI_r)m\bar{U}_{ii}\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathrm{tr}\left[(\bar{O}_i\bar{\Lambda}_i\bar{O}_i^T - m\bar{z}_i\bar{O}_i\bar{O}_i^T + m\bar{z}_iI_r)\bar{O}_i\bar{O}_i^T\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathrm{tr}(\bar{\Lambda}_i) = \frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m \mathrm{tr}(\bar{O}_i^T S_{ij}\bar{O}_j) = \mathrm{tr}(\tilde{S}\bar{U}).
\end{aligned}$$

This satisfies (KKT-e) and (KKT-f) for $\bar{L} \triangleq \bar{M} + \bar{Z} - \tilde{S}$. Finally,

$$\mathrm{tr}(m\bar{M}_i) = \mathrm{tr}\left[\bar{O}_i(\bar{\Lambda}_i - m\bar{z}_iI_r)\bar{O}_i^T\right] = \mathrm{tr}(\bar{\Lambda}_i - m\bar{z}_iI_r),$$
$$\mathrm{tr}\left[(m\bar{M}_i)(m\bar{U}_{ii})\right] = \mathrm{tr}\left[(\bar{O}_i(\bar{\Lambda}_i - m\bar{z}_iI_r)\bar{O}_i^T)(\bar{O}_i\bar{O}_i^T)\right] = \mathrm{tr}(\bar{\Lambda}_i - m\bar{z}_iI_r);$$

thus (KKT-g) is satisfied. In short, the choices of variables (3.9), (3.10), and (3.11) satisfy all the KKT conditions except (KKT-h) for any $\bar{z}_i \leq \tau_i/m$.

To satisfy this final KKT condition, let $\bar{L}(\bar{z}_1,\ldots,\bar{z}_m) = \bar{M} + \bar{Z} - \tilde{S}$ and observe that $\bar{M} + \bar{Z} = \mathrm{diag}\left[\bar{O}_1\bar{\Lambda}_1\bar{O}_1^T + m\bar{z}_1\bar{O}_1^\perp\bar{O}_1^{\perp T},\ldots,\bar{O}_m\bar{\Lambda}_m\bar{O}_m^T + m\bar{z}_m\bar{O}_m^\perp\bar{O}_m^{\perp T}\right]$, where $O_i^\perp \in \mathcal{O}_{d_i,d_i-r}$ satisfies $O_i^\perp O_i^{\perp T} = I_{d_i} - O_iO_i^T$. If the linear matrix inequality (LMI)

$$(3.12) \qquad \bar{L}(\bar{z}_1,\ldots,\bar{z}_m) \succeq 0, \quad \bar{z}_i \leq \tau_i/m, \quad i=1,\ldots,m,$$

has a feasible point $(\bar{z}_1^\star,\ldots,\bar{z}_m^\star)$, then this is a certificate that $(\bar{O}_1,\ldots,\bar{O}_m)$ is a global maximizer of (OTSM). While, in general, LMIs are solved by interior-point methods [23], for LMI (3.12) it is unnecessary. Since $\bar{O}_i^\perp\bar{O}_i^{\perp T}$ is positive semidefinite, $\bar{L}$ is monotone (in Loewner order) in the scalars $\bar{z}_1,\ldots,\bar{z}_m$, i.e.,

$$\bar{L}(z_1,\ldots,z_m) \succeq \bar{L}(w_1,\ldots,w_m)$$

whenever $z_i \geq w_i$ for $i=1,\ldots,m$. Thus it is sufficient to check the positive semidefiniteness at values $\bar{z}_i = \tau_i/m$. If it holds, we have found a feasible point. If not, the LMI is infeasible. We state this result as the following theorem.

THEOREM 3.3. *Suppose $\bar{O} = (\bar{O}_1,\ldots,\bar{O}_m)$ is a stationary point of (OTSM). Let $\bar{\Lambda}_i = \sum_{j=1}^m \bar{O}_i^T S_{ij}\bar{O}_j$, and τ_i be the smallest eigenvalue of $\bar{\Lambda}_i$ for $i=1,\ldots,m$. Then $\bar{O}$ is a global optimum of (OTSM) if*

$$(\text{CERT}) \quad L^\star = \mathrm{diag}\left(\bar{O}_1\bar{\Lambda}_1\bar{O}_1^T + \tau_1\bar{O}_1^\perp\bar{O}_1^{\perp T},\ldots,\bar{O}_m\bar{\Lambda}_m\bar{O}_m^T + \tau_m\bar{O}_m^\perp\bar{O}_m^{\perp T}\right) - \tilde{S} \succeq 0.$$

Remark 3.4. Let $\bar{O} = m^{-1/2}[\bar{O}_1^T,\cdots,\bar{O}_m^T]^T$. It is easy to see that $\bar{O} \in \mathcal{O}_{D,r}$ and $L^\star\bar{O} = 0$ using the stationarity condition (3.2). Therefore, it suffices to test $(\bar{O}^\perp)^T L^\star\bar{O}^\perp \succeq 0$, where $\bar{O}^\perp \in \mathcal{O}_{D,D-r}$ fills out $\bar{O}$ to a fully orthogonal matrix. This matrix is $(D-r)\times(D-r)$ and may be easier to handle than the $D \times D$ matrix $L^\star$.

Example 3.5. To see the significance of Theorem 3.3, consider the example examined by Ten Berge [30, p. 270] in the context of the MAXDIFF problem ($S_{ii} = 0$, $i = 1, \ldots, m$). Let $m = 3$, $d_1 = d_2 = d_3 = d$, and set $S_{12} = -I_d$, $S_{13} = I_d$, $S_{23} = I_d$. Using Theorem 3.3, it is easy to see that any triple of (partially) orthogonal matrices $(\bar{O}_1, \bar{O}_2, \bar{O}_3) \in \mathcal{O}_{d,r} \times \mathcal{O}_{d,r} \times \mathcal{O}_{d,r}$ such that $\bar{O}_3 = \bar{O}_1 + \bar{O}_2$ satisfies (CERT) for any $r \leq d$. Indeed, for choices $\Lambda_i = I_r$ and $\tau_i = 1$ for all i,

$$L^\star = \begin{bmatrix} I_d & I_d & -I_d \\ I_d & I_d & -I_d \\ -I_d & -I_d & I_d \end{bmatrix} = \begin{bmatrix} I_d \\ I_d \\ -I_d \end{bmatrix} \begin{bmatrix} I_d & I_d & -I_d \end{bmatrix} \succeq 0.$$

Specifically, for $d = 3$ and $r = 2$, the triple

$$(3.13) \qquad \bar{O}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \bar{O}_2 = \begin{bmatrix} -1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \\ 0 & 0 \end{bmatrix}, \quad \bar{O}_3 = \begin{bmatrix} 1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & 1/2 \\ 0 & 0 \end{bmatrix}$$

is a global maximizer. The global maximum value is 3.

Unlike the Ky Fan problem of the previous section, condition (CERT) is hardly necessary for global optimality. This is a key difference of (OTSM) from the Ky Fan problem. To see this, let $\bar{O} = (\bar{O}_1, \ldots, \bar{O}_m)$ be a local maximizer, and recall $I_{d_i} = \bar{O}_i \bar{O}_i^T + \bar{O}_i^\perp \bar{O}_i^{\perp T}$. It follows from the tangency condition (3.6) that

$$\begin{aligned} \mathrm{tr}(\bar{\Lambda}_i W_i^T W_i) &= \mathrm{tr}(W_i \bar{\Lambda}_i W_i^T \bar{O}_i \bar{O}_i^T) + \mathrm{tr}(W_i \bar{\Lambda}_i W_i^T \bar{O}_i^\perp \bar{O}_i^{\perp T}) \\ &= \mathrm{tr}(\bar{O}_i^T W_i \bar{\Lambda}_i W_i^T \bar{O}_i) + \mathrm{tr}(\bar{\Lambda}_i W_i^T \bar{O}_i^\perp \bar{O}_i^{\perp T} W_i) \\ &= \mathrm{tr}(W_i^T \bar{O}_i \bar{\Lambda}_i \bar{O}_i^T W_i) + \mathrm{tr}(\bar{\Lambda}_i W_i^T \bar{O}_i^\perp \bar{O}_i^{\perp T} W_i) \\ &\geq \mathrm{tr}(W_i^T \bar{O}_i \bar{\Lambda}_i \bar{O}_i^T W_i) + \tau_i \mathrm{tr}(W_i^T \bar{O}_i^\perp \bar{O}_i^{\perp T} W_i). \end{aligned}$$

The last inequality is due to $W_i^T \bar{O}_i^\perp (\bar{O}_i^\perp)^T W_i \succeq 0$ and $\bar{\Lambda}_i \succeq \tau_i I_r$ (see, e.g., [19, pp. 482–483]). Thus the local maximality condition (3.5) is sufficiently satisfied if

$$\sum_{i=1}^m \mathrm{tr}(W_i^T \bar{O}_i \bar{\Lambda}_i \bar{O}_i^T W_i) + \sum_{i=1}^m \tau_i \mathrm{tr}\left(W_i^T \bar{O}_i^\perp \bar{O}_i^{\perp T} W_i\right) - \mathrm{tr}(W^T \tilde{S} W) \geq 0$$

for all $W = [W_1^T, \ldots, W_m^T]^T$. This will be the case if $L^\star \succeq 0$, but not only if.

Nevertheless, there are a few special cases that condition (CERT) is also necessary for global optimality.

COROLLARY 3.6. *For the MAXDIFF problem with $m = 2$, i.e., $S_{11} = 0$ and $S_{22} = 0$, if a point $(\bar{O}_1, \bar{O}_2)$ is a global maximizer, then condition* (CERT) *is satisfied. This is true for any $r \leq \min\{d_1, d_2\}$.*

COROLLARY 3.7. *For $m = 2$ and $r = 1$, if a point $(\bar{O}_1, \bar{O}_2)$ is a global maximizer of* (OTSM), *then condition* (CERT) *is satisfied.*

COROLLARY 3.8. *If S_{ij} has rank less than or equal to r with singular value decomposition $S_{ij} = V_i \Sigma_{ij} V_j^T$ for $V_i \in \mathcal{O}_{d_i,r}$, $V_j \in \mathcal{O}_{d_j,r}$, and Σ_{ij} is an $r \times r$ nonnegative diagonal matrix for $i, j = 1, \ldots, m$, then $(V_1, \ldots, V_m)$ solves* (OTSM) *globally.*

The last corollary provides a data-only sufficient condition for global optimality, which does not require computing a stationary point. Its hypothesis is satisfied, e.g., if A_i's share left singular vectors in the MAXDIFF or MAXBET problems: $A_i = R\Sigma_i V_i^T$

for some $R \in \mathcal{O}_{r,r}$. Corollary 3.7 is due to [16, Result 4], which we discuss in the next subsection. Proofs of the other corollaries are provided in Appendix A.

We also point out a special case in which a qualified local optimality implies condition (CERT). For $m = 2$ involving fully orthogonal matrices (i.e., $d_1 = d_2 = r$; note in this case (OTSM) coincides with both MAXDIFF and MAXBET), it is well-known that a locally maximal point $\bar{O} = (\bar{O}_1, \bar{O}_2)$ such that $\bar{\Lambda}_1 = \bar{O}^T S_{12} \bar{O}_2$ is positive semidefinite also implies global optimality [12, 30]. Theorem 3.3 recovers this result. To see this, observe that $\bar{\Lambda}_1 = \bar{O}_1^T S_{12} \bar{O}_2 = \bar{O}_2^T S_{21} \bar{O}_1 = \bar{\Lambda}_2$ by symmetry. Let $\bar{\Lambda} = \bar{\Lambda}_1 = \bar{\Lambda}_2$. Then,

$$L^\star = \begin{bmatrix} \bar{O}_1 \bar{\Lambda} \bar{O}_1^T & -S_{12} \\ -S_{12}^T & \bar{O}_2 \bar{\Lambda} \bar{O}_2^T \end{bmatrix} = \begin{bmatrix} \bar{O}_1 \\ -\bar{O}_2 \end{bmatrix} \bar{\Lambda} \begin{bmatrix} \bar{O}_1 \\ -\bar{O}_2 \end{bmatrix}^T \succeq 0,$$

since $S_{12} = O_1 \Lambda O_2^T = S_{21}^T$. We see the circle $L^\star \succeq 0 \implies$ (global optimum) $\implies$ (local maximum with $\bar{\Lambda} \succeq 0$) $\implies L^\star \succeq 0$ for $m = 2$ and $r = d_1 = d_2$, similar to the Ky Fan problem. In fact, a stationary point with $\bar{\Lambda} \succeq 0$ suffices, and by Proposition 3.1, any stationary point can be improved to satisfy this condition.

Table 1 summarizes the results discussed so far and their parallelism with those in section 2.

3.4. Other certificates. Sufficient conditions for global optimality of problem (OTSM) appear understudied. Here we discuss three such conditions collected from the generalized CCA and Procrustes analysis literature.

Ten Berge [30] shows that if $d_1 = \cdots = d_m = r$ (fully orthogonal) and $\bar{\Lambda}_{ij} \triangleq \bar{O}_i^T S_{ij} \bar{O}_j$ is symmetric and positive semidefinite for all $i < j$ for global optimality, then $(\bar{O}_1, \ldots, \bar{O}_m)$ is a global solution. This sufficient condition is excessively strong, and, in fact, he uses Example 3.5 to show that the condition is hardly met: no orthogonal matrices $(\bar{O}_1, \bar{O}_2, \bar{O}_3)$ exist such that $-\bar{O}_1^T \bar{O}_2$, $\bar{O}_1^T \bar{O}_3$, and $\bar{O}_2^T \bar{O}_3$ are simultaneously symmetric and positive semidefinite. Nevertheless, Ten Berge's sufficient condition is implied by Theorem 3.3. To see this, observe that $\bar{\Lambda}_i = \sum_{j=1}^m \bar{\Lambda}_{ij}$ is symmetric (and positive semidefinite), satisfying the stationarity condition (3.2). Also since $S_{ij} = \bar{O}_i \bar{\Lambda}_{ij} \bar{O}_j^T$,

$$L^\star = \mathrm{diag}(\bar{O}_1 \bar{\Lambda}_1 \bar{O}_1, \ldots, \bar{O}_m \bar{\Lambda}_m \bar{O}_m) - \tilde{S}$$

$$= \begin{bmatrix} \bar{O}_1 & & \\ & \ddots & \\ & & \bar{O}_m \end{bmatrix} \begin{bmatrix} \sum_{j \neq 1} \bar{\Lambda}_{1j} & -\bar{\Lambda}_{12} & \cdots & -\bar{\Lambda}_{1m} \\ \vdots & & \ddots & \vdots \\ -\bar{\Lambda}_{m1} & \cdots & -\bar{\Lambda}_{m,m-1} & \sum_{j \neq m} \bar{\Lambda}_{mj} \end{bmatrix} \begin{bmatrix} \bar{O}_1^T & & \\ & \ddots & \\ & & \bar{O}_m^T \end{bmatrix}.$$

It is easy to check that the middle block matrix is positive semidefinite, and so is $L^\star$.

Hanafi and Ten Berge [16] show that if unit vectors $o_i \in \mathbb{R}^{d_i}$, $i = 1, \ldots, m$, satisfy $\tilde{S}o = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)o$ with $o = (o_1^T, \ldots, o_m^T)^T$ and $\mathrm{diag}(\lambda_1 I_{d_1}, \ldots, \lambda_m I_{d_m}) - \tilde{S} \succeq 0$, then $(o_1, \ldots, o_m)$ maximizes $\frac{1}{2} \sum_{i,j=1}^m o_i^T S_{ij} o_j$ globally. Obviously, this is a special case of Theorem 3.3 for $r = 1$. Corollary 3.7 follows from this result.

If a stationary point $(\bar{O}_1, \ldots, \bar{O}_m)$ satisfies the second-order condition (3.5) for *all* $W = [W_1^T, \ldots, W_m^T]^T \in \mathbb{R}^{D \times r}$, i.e., each W_i does not necessarily observe the tangency condition (3.6) at $\bar{O}_i$, then this is obviously sufficient for the point to be globally optimal. Liu, Wang, and Wang, [21, Theorem 2.4] describe this condition in a matrix form:

$$(3.14) \qquad \mathcal{L}^\star \triangleq \mathrm{diag}(K_{d_1,r}^T (I_{d_1} \otimes \bar{\Lambda}_1) K_{d_1,r}, \ldots, K_{d_m,r}^T (I_{d_m} \otimes \bar{\Lambda}_m) K_{d_m,r}) - \mathcal{S} \succeq 0,$$

where $\mathcal{S} = (I_r \otimes S_{ij})$, and K_{mn} is the commutation matrix such that $K_{mn} \operatorname{vec} A = \operatorname{vec} A^T$ for $A \in \mathbb{R}^{m \times n}$; vec$(\cdot)$ is the usual vectorization operator, and $\otimes$ denotes the Kronecker product. Condition (3.14) can be related to Theorem 3.3 by the similarity transform of $\mathcal{L}^\star$ with $\mathcal{K} = \operatorname{diag}(K_{d_1,r}, \ldots, K_{d_m,r})$:

$$\mathcal{K}\mathcal{L}^\star\mathcal{K}^T = \operatorname{diag}(I_{d_1} \otimes \bar{\Lambda}_1, \ldots, I_{d_m} \otimes \bar{\Lambda}_m) - \mathcal{K}\mathcal{S}\mathcal{K}^T$$
$$= \operatorname{diag}(I_{d_1} \otimes \bar{\Lambda}_1, \ldots, I_{d_m} \otimes \bar{\Lambda}_m) - \tilde{\mathcal{S}} \succeq 0, \quad \tilde{\mathcal{S}} = (S_{ij} \otimes I_r).$$

The second equality uses the fact $K_{d_i,r}(I_r \otimes S_{ij})K_{d_j,r}^T = S_{ij} \otimes I_r$. Observe the resemblance of the last line to condition (CERT). When $r = 1$, these two conditions actually coincide, hence also with the Hanafi–Ten Berge condition. For $r > 1$, besides the expenses of constructing a larger matrix than $L^\star$ ($rD \times rD$ versus $D \times D$), condition (3.14) is usually stronger than Theorem 3.3. For instance, in Example 5.1 of section 5 stationary points obtained by Algorithm 4.1 in section 4 satisfy condition (CERT) for all possible values of r, but for those points $\mathcal{L}^\star \succeq 0$ only when $r = 1$.

TABLE 1
Comparison between Ky Fan's problem and the orthogonal trace-sum maximization problem (OTSM). Set Γ_2 refers to the set of locally maximal points.

	Ky Fan	OTSM
Domain	$\bar{O} \in \mathcal{O}_{d,r}$	$\bar{O} = (\bar{O}_1, \ldots, \bar{O}_m) \in \mathcal{O}_{d_1,r} \times \cdots \times \mathcal{O}_{d_m,r}$
Lagrange multiplier(s)	$\bar{\Lambda} = \bar{O}^T S \bar{O} = \bar{\Lambda}^T$	$\bar{\Lambda}_i = \sum_{j=1}^m \bar{O}_i^T S \bar{O}_j = \bar{\Lambda}_i^T$, $i = 1, \ldots, m$
Lifting matrix	$\bar{U} = \bar{O}\bar{O}^T$	$\bar{U} = \frac{1}{m}\bar{O}\bar{O}^T$
Cutoff matrix	$\bar{z} = \lambda_{\min}(\bar{\Lambda})$	$\bar{Z} = \operatorname{diag}([\tau_i I_{d_i}]_{i=1}^m)$, $\tau_i = \lambda_{\min}(\Lambda_i)$
Nonneg. part of $\bar{\Lambda}$	$\bar{M} = \bar{O}(\bar{\Lambda} - \bar{z}I_r)\bar{O}^T$	$\bar{M} = \operatorname{diag}([\bar{O}_i(\bar{\Lambda}_i - \tau_i I_r)\bar{O}_i^T]_{i=1}^m)$
Nonpos. part of $\bar{\Lambda}$	$\bar{L} = \bar{M} + \bar{z}I_d - S$	$\bar{L} = \bar{M} + \bar{Z} - \tilde{S}$
Certificate matrix	$\bar{L} \succeq 0$ ($\forall \bar{O} \in \Gamma_2$)	$\bar{L} \succeq 0$ ($\exists \bar{O} \in \Gamma_2$)

4. Proximal block relaxation algorithm. In order to apply Theorem 3.3 to verify condition (CERT), an algorithm that generates iterates converging to a stationary point is needed. Although problem (OTSM) has been studied in the generalized CCA and Procrustes analysis context for a long time, algorithms that possess this desired property appear rare. In this section, we propose such an algorithm.

4.1. Oscillation of the standard algorithm. We first point out a flaw in the block ascent algorithm widely employed in both the generalized CCA [15, 31, 32] and the Procrustes analysis contexts [10, 13, 30]. This algorithm cyclically updates each orthogonal matrix O_i with other blocks O_j, $j \neq i$, held fixed. To update the ith block in the $k + 1$st cycle, let $O^{\text{prev}} = (O_1^{k+1}, \ldots, O_{i-1}^{k+1}, O_i^k, O_{i+1}^k, \ldots, O_m^k)$, then maximize $\operatorname{tr}\left[O_i^T(\sum_{j=1}^m S_{ij}O_j^{\text{prev}})\right]$. This block update scheme is natural since the domain $\mathcal{O}_{d_1,r} \times \cdots \times \mathcal{O}_{d_m,r}$ has a product structure. Furthermore, each maximization is explicit: let us invoke the von Neumann–Fan inequality

$$\operatorname{tr}(A^T B) \leq \sum_l \sigma_l(A)\sigma_l(B),$$

which holds for any two matrices A and B of the same dimensions with the lth largest singular values $\sigma_l(A)$ and $\sigma_l(B)$, respectively; equality is attained when A and B share a simultaneous ordered SVD (see, e.g., [19]). Thus if $B = \sum_{j=1}^m S_{ij}O_j^{\text{prev}}$ has an SVD of $P_i D_i Q_i^T$, where D_i is $r \times r$ nonnegative diagonal, then the optimal choice of $A = O_i$

is $P_i Q_i^T$. The latter matrix is orthogonal. This method can be considered a linearized version of the alternating variable algorithm of [21, Algorithm 4.1].

However, convergence of this standard algorithm is not guaranteed. To be precise, let $O^k = (O_1^k, \ldots, O_m^k)$ be the kth iterate after k cycles of the algorithm. While it can be shown that the sequence of objective values $\{f(O^k)\}$ converges [31], it cannot be said that the iterates $\{O^k\}$ themselves converge. The main reason is that the map $B = \sum_{j \neq i} S_{ij} O_j \mapsto P_i Q_i^T$ is *set-valued*. If B is rank deficient, any orthonormal basis of the null space of B^T (resp., B) can be chosen as left (resp., right) singular vectors corresponding to the zero singular value; the product $P_i Q_i^T$ may not be unique [2, Proposition 7]. Each update of the ith block may place it too far from its previous location. To see the potential peril of this update scheme, let us revisit Example 3.5. Suppose the algorithm is initialized with $O^0 = (I, J, I)$, where

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \in \mathcal{O}_{3,2} \quad \text{and} \quad J = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \in \mathcal{O}_{3,2}.$$

Both $I - J$ and $I + J$ have rank 1, and $-J \in \operatorname{argmax}_{O_i \in \mathcal{O}_{3,2}} \operatorname{tr}\left[O_i^T (I - J)\right]$, $I \in \operatorname{argmax}_{O_i \in \mathcal{O}_{3,2}} \operatorname{tr}\left[O_i^T (I + J)\right]$. Taking these particular values as the outputs of an instance of the above set-valued map, we have the following sequence of O^k:

$$(I, J, I) \to (-J, I, -J) \to (-I, -J, -I) \to (J, -I, J) \to (I, J, I) \to \cdots.$$

Thus the standard algorithm *oscillates* while the objective does not change from a suboptimal value of 1 (recall the globally optimal value is 3).

4.2. Proximal regularization. We propose a simple modification of the standard algorithm that leads to a convergent algorithm. Define a bivariate function $f_i : \mathcal{O}_{d_i,r} \times (\mathcal{O}_{d_1,r} \times \cdots \times \mathcal{O}_{d_m,r}) \to \mathbb{R}$ as $f_i(O_i, \Theta) = \operatorname{tr}\left[O_i^T (\sum_{j=1}^m S_{ij} \Theta_j)\right]$ for $i = 1, \ldots, m$, where $\Theta = (\Theta_1, \ldots, \Theta_m)$. Then the objective function of (OTSM) can be denoted by $f(O) = \frac{1}{2} \sum_{i=1}^m f_i(O_i, O)$, where $O = (O_1, \ldots, O_m)$. For the update of the ith block in the $k + 1$st cycle, we consider a spherical quadratic approximation of f_i at O_i^k, i.e.,
(4.1)
$$f_i(O_i^k, O^{\text{prev}}) + \operatorname{tr}[\nabla_1 f_i(O_i^k, O^{\text{prev}})^T (O_i - O_i^k)] - \frac{1}{2\alpha} \|O_i - O_i^k\|_{\text{F}}^2 + \sum_{j \neq i} f_j(O_j^{\text{prev}}, O^{\text{prev}}),$$

where $O^{\text{prev}} = (O_1^{k+1}, \ldots, O_{i-1}^{k+1}, O_i^k, O_{i+1}^k, \ldots, O_m^k)$, and $\nabla_1 f_i(O_i, \Theta) = \sum_{j=1}^m S_{ij} \Theta_j$ is the derivative of f_i in the first variable; $\alpha > 0$ is a given constant. We then maximize this approximation with respect to O_i, with the other coordinates held fixed. This partial maximization is also explicit, as it can be easily seen that maximizing the objective (4.1) is equivalent to maximizing $\operatorname{tr}\left[O_i^T (\sum_{j=1}^m S_{ij} O_j^{\text{prev}} + \alpha^{-1} O_i^k)\right]$. Therefore, we can employ the von Neumann–Fan inequality to $A = O_i$ and $B = \sum_{j=1}^m S_{ij} O_j^{\text{prev}} + \alpha^{-1} O_i^k$. If B has an SVD of $P_i D_i Q_i^T$, then the optimal choice of A is again $P_i Q_i^T$, which is orthogonal. This fact suggests Algorithm 4.1, which includes that standard algorithm as a special case ($\alpha = +\infty$). The quadratic regularization term in objective (4.1) keeps the update O_i^{k+1} in the proximity of its previous value O_i^k, and the α moderates the degree of attraction. Algorithm 4.1 is also an instance of the minorization-maximization (MM) algorithm (see, e.g., [20]): at each update, the surrogate function defined on

$$(\mathcal{O}_{d_1,r} \times \cdots \times \mathcal{O}_{d_m,r}) \times (\mathcal{O}_{d_1,r} \times \cdots \times \mathcal{O}_{d_m,r})$$

$$g(O \mid \Theta) = f(\Theta) + \sum_{i=1}^{m} \text{tr}[\nabla_1 f_i(\Theta_i, \Theta)^T (O_i - \Theta_i)] - \frac{1}{2\alpha} \sum_{i=1}^{m} \|O_i - O_i^k\|_{\text{F}}^2$$

minorizes the objective function $f(O)$ at $\Theta = (O_1^{k+1}, \ldots, O_{i-1}^{k+1}, O_i^k, O_{i+1}^k, \ldots, O_m^k)$ for a certain range of α and is partially maximized. As a consequence of being an MM algorithm, each update monotonically improves the objective function f. Due to the compactness of each $\mathcal{O}_{d_i,r}$, actually the sequence of objective values $\{f(O^k)\}$ converges.

In the next subsection we proceed to show that the sequence of iterates $\{O^k\}$ converges to a stationary point, in contrast to the standard algorithm. In particular, in the example of the previous subsection, with any finite $\alpha > 0$ the maximizers of $\text{tr}\left[O_i^T(-J + I + \alpha^{-1}I)\right]$, $\text{tr}\left[O_i^T(-I + I + \alpha^{-1}J)\right]$, and $\text{tr}\left[O_i^T(I + J + \alpha^{-1}I)\right]$ in $\mathcal{O}_{3,2}$ are uniquely determined by I, J, and I. This yields $O^0 = (I, J, I) = O^1 = O^2 = \cdots$ in Algorithm 4.1. In fact, the point (I, J, I) is a stationary point. (Convergence to a global maximizer, e.g., (3.13), requires a good initial point. We discuss this in section 5 with another global solution.) Note, however, the map in lines 5 and 6 of Algorithm 4.1 is nevertheless set-valued, since there is no guarantee of full rank of B.

Algorithm 4.1 Proximal block relaxation algorithm for solving (OTSM).

1: Initialize $O_1, \ldots, O_m$; Set $\alpha \in (0, 1/\max_{i=1,\ldots,m} \|S_{ii}\|_2)$
2: For $k = 1, 2, \ldots$
3:     For $i = 1, \ldots, m$
4:         Set $B = \sum_{j=1}^{m} S_{ij}O_j + \alpha^{-1}O_i$
5:         Compute SVD of B as $P_i D_i Q_i^T$
6:         Set $O_i = P_i Q_i^T$
7:     End For
8:     If there is no progress, then break
9: End For
10: Return $(O_1, \ldots, O_m)$

4.3. Global convergence. Algorithm 4.1 with $\alpha > 0$ converges despite the nonuniqueness of the map in lines 5 and 6.

THEOREM 4.1. *The sequence $\{(O_1^k, \ldots, O_m^k)\}$ generated by Algorithm 4.1 converges to a stationary point of* (OTSM) *for $\alpha \in (0, 1/\max_{i=1,\ldots,m} \|S_{ii}\|_2)$; $\|\cdot\|_2$ denotes the spectral norm. Furthermore, the rate of convergence is at least sublinear.*

This result is stronger than typical global convergence results that all the limit points are stationary [19, 39], or that the gradient vanishes [1, 24]. Theorem 4.1 can be shown using Theorems 1 and 2 in Xu and Yin [38] by noting that Algorithm 4.1 falls into their "deterministic block prox-linear" class of algorithms and problem (OTSM) possesses the Kurdyka–Łojasiewicz property [4]. In the accompanying supplementary material file OTSM_supp.pdf [local/web 625KB], we provide a simpler proof utilizing the closedness [39] of the map in lines 5 and 6, and the geometry of the product of Stiefel manifolds.

Remark 4.2. In case $S_{ii} = 0$ for $i = 1, \ldots, m$, e.g., the MAXDIFF problem, the α can be chosen as an arbitrary positive constant.

5. Numerical experiments.

5.1. Setup. In this section we test Algorithm 4.1 equipped with the certificates of global optimality and suboptimality discussed in section 3 with both synthetic and real-world examples. Algorithm 4.1 was implemented in the Julia programming language [7] and run on a standard laptop computer (Macbook Pro, i5@2.4GHz, 16GB RAM). We set the proximity constant $\alpha = 1000$ and terminated the algorithm if the mean change $m^{-1} \sum_{i=1}^{m} \|O_i^k - O_i^{k-1}\|_F$ was less than 10^{-8} and the relative change of the objective function was less than 10^{-10}, or a maximum iteration of 50000 was reached. For comparison, we also tested the generic Riemanian trust-region method [1] implemented in the Manopt MATLAB toolbox [9]. The maximum number of outer iterations of this method was set to 10000. For both methods, four initialization strategies were considered:

1. ("eye") The ith block O_i^0 of the initial point takes the first r columns of I_{d_i}.
2. ("tb") Take the eigenvectors corresponding to the r largest eigenvalues of the data matrix $\tilde{S} = (S_{ij})$ to form a $D \times r$ orthogonal matrix $\tilde{V}$. Split $\tilde{V}$ into m blocks so that $\tilde{V} = [\tilde{V}_1^T, \ldots, \tilde{V}_m^T]^T$, where $\tilde{V}_i \in \mathbb{R}^{d_i \times r}$, $i = 1, \ldots, m$. Project each block $\tilde{V}_i$ to the Stiefel manifold $\mathcal{O}_{d_i, r}$ to obtain O_i^0.
3. ("sb") Replace the diagonal blocks S_{ii} of $\tilde{S}$ by $-\sum_{j=1}^{m} (S_{ij} S_{ij}^T)^{1/2}$, where $M^{1/2}$ denotes the matrix square root of the positive semidefinite matrix M. Take the eigenvectors corresponding to the r largest eigenvalues of the resulting negative semidefinite matrix to form a $D \times r$ orthogonal matrix $\tilde{V}$. Proceed as strategy "tb."
4. ("lww1") Set O_1^0 to the top r eigenvectors of S_{11}. Then set $O_k^0 = U_k Q_k$, where U_k is the top r eigenvectors of S_{kk} and Q_k is the Q factor in the QR decomposition of $U_k \sum_{j<k} S_{kj} O_j^0$, $k = 2, \ldots, m$.

The initial point of strategy "tb" coincides with that which gives the second upper bound of the orthogonal Procrustes problem [30, p. 273], and also with the starting point strategy 2 of [21, p. 1495] for the MAXBET problem. Strategy "sb" extends [28, p. 380] for the orthogonal Procrustes problem; see Lemma A.1 in Appendix A, and also the accompanying supplementary material file OTSM_supp.pdf [local/web 625KB]. Strategy "lww1" is the starting point strategy 1 by [21, p. 1494].

5.2. Small examples.

Example 5.1 (CCA of port wine data). We consider generalized CCA of the subset of the data from sensory evaluation of port wines analyzed by Hanafi and Kiers [15, Table 2]. The goal is to capture the agreement between $m = 4$ assessors in the assessment of the appearance of $n = 8$ port wines. Note that the dimensions are disparate: $d_1 = 4$, $d_2 = 3$, $d_3 = 4$, and $d_4 = 3$. The MAXDIFF criterion was tested for all possible $r = 1, 2, 3$. The results are summarized in Table 2. Algorithm 4.1 achieved global optimum for all r and for all initial point strategies. A similar phenomenon occurred with MAXBET, whose results are provided in the accompanying supplementary material file OTSM_supp.pdf [local/web 625KB], except for $r = 3$ with strategies "eye" and "lww1." On the other hand, Manopt occasionally converged to a stationary point violating the conclusion of Proposition 3.1. In addition, Algorithm 4.1 was orders of magnitudes faster than Manopt. In all cases certified to be globally optimal, the smallest eigenvalues of $L^\star$ in condition (CERT) (denoted $\lambda_{\min}(L^\star)$) were numerically zero up to the fourteenth digit after the decimal point, whereas those of $\mathcal{L}^\star$ (denoted $\lambda_{\min}(\mathcal{L}^\star)$) in condition (3.14) were often definitely negative.

TABLE 2
Port wine data, MAXDIFF. In "classification," "not loc. opt." means that the iterate at termination is stationary but violates Proposition 3.1.

r	Init	Method	Iter	Time (sec)	Obj	Classification	$\lambda\min(L^\star)$	$\lambda\min(\mathcal{L}^\star)$
1	eye	PBA	10	0.0001923	209.8	global opt.	-1.276e-14	-1.276e-14
		Manopt	9	0.07502	209.8	global opt.	-2.482e-15	-2.482e-15
	sb	PBA	10	0.0002432	209.8	global opt.	-2.101e-14	-2.101e-14
		Manopt	7	0.05024	209.8	global opt.	6.234e-15	6.234e-15
	tb	PBA	9	0.0001739	209.8	global opt.	-1.159e-14	-1.159e-14
		Manopt	5	0.03958	209.8	global opt.	-1.951e-14	-1.951e-14
	lww1	PBA	10	0.0002291	209.8	global opt.	2.599e-15	2.599e-15
		Manopt	9	0.05789	209.8	global opt.	-1.024e-15	-1.024e-15
2	eye	PBA	10	0.0002719	271.2	global opt.	-1.471e-14	-79.61
		Manopt	11	0.1084	271.2	global opt.	-2.707e-15	-79.61
	sb	PBA	9	0.0002301	271.2	global opt.	-6.812e-14	-79.61
		Manopt	7	0.08470	271.2	global opt.	-4.465e-14	-79.61
	tb	PBA	9	0.0002238	271.2	global opt.	-2.418e-14	-79.61
		Manopt	6	0.07996	271.2	global opt.	-1.992e-14	-79.61
	lww1	PBA	10	0.0002762	271.2	global opt.	-1.364e-13	-79.61
		Manopt	12	0.1024	271.2	global opt.	-1.219e-14	-79.61
3	eye	PBA	13	0.0002659	284.1	global opt.	-3.292e-14	-106.1
		Manopt	13	0.1814	280.7	not loc. opt.	–	–
	sb	PBA	9	0.0001853	284.1	global opt.	-1.370e-14	-106.1
		Manopt	6	0.1135	284.1	global opt.	-5.021e-14	-106.1
	tb	PBA	10	0.0002105	284.1	global opt.	-1.16e-13	-106.1
		Manopt	6	0.1078	284.1	global opt.	-4.389e-15	-106.1
	lww1	PBA	11	0.0002893	284.1	global opt.	4.392e-15	-106.1
		Manopt	12	0.1530	280.7	not loc. opt.	–	–

Example 5.2. We revisit Example 3.5 for $d = 3$ and $r = 1, 2, 3$. The results are summarized in Table 3. Strategies "eye" and "lww1" gave suboptimal stationary points as initial points, and both algorithms could not make progress. Strategy "sb" yielded global optima for both $r = 1, 2$. For $r = 3$, no strategy could certify global optimality. For $r = 2$, while both "sb" and "tb" were successful, Algorithm 4.1 took the full 50000 iterations to achieve the same accuracy as Manopt, which in this case took 22 outer iterations. The stationary points reached from the two initial points were all quite different from each other, and also from the analytic solution (3.13). The error $\|\bar{O}_1 + \bar{O}_2 - \bar{O}_3\|_\infty$ was between 8.877×10^{-7} and 6.146×10^{-6}. Together with the smallest eigenvalue of $L^\star$ computed being -6.674×10^{-6}, this relatively large error reflects the hardness of this problem illustrated in Example 3.5. This difficulty was also experienced with an extra run of the commercial interior-point method solver MOSEK [22] to solve the convex relaxtion (P-SDP). While the optimal objective value was 3 up to the eighth digit after the decimal point, MOSEK failed to obtain a rank-two solution. With this exception, Algorithm 4.1 terminated in a fraction of time for Manopt.

Additional examples. In the supplementary material (OTSM_supp.pdf [local/web 625KB]), Examples 5.1 and 5.2 of [21] are considered under both MAXDIFF and MAXBET criteria, and new global optima are found.

5.3. Simulation studies. Following the orthogonal Procrustes analysis model, we generated n sets of d-dimensional landmarks from the standard normal distribution independently, and randomly rotated them by m orthogonal matrices of size $d \times r$.

TABLE 3

Example 5.2. *In "classification," "stationary" means that the iterate at termination is stationary but its global optimality is not confirmed by using Theorem* 3.3.

r	Init	Method	Iter	Time (sec)	Obj	Classification	$\lambda\min(L^\star)$	$\lambda\min(\mathcal{L}^\star)$
1	eye	PBA	2	0.0005041	1.000	stationary	-1.000	-1.000
		Manopt	1	0.001597	1.000	stationary	-1.000	-1.000
	sb	PBA	12	0.0001631	1.500	global opt.	-1.608e-10	-1.608e-10
		Manopt	7	0.02612	1.500	global opt.	-1.045e-7	-1.045e-7
	tb	PBA	2	0.0001112	1.000	stationary	-1.000	-1.000
		Manopt	1	0.001039	1.000	stationary	-1.000	-1.000
	lww1	PBA	2	0.0001182	1.000	stationary	-1.000	-1.000
		Manopt	1	0.0009643	1.000	stationary	-1.000	-1.000
2	eye	PBA	2	0.0001587	2.000	stationary	-1.000	-1.000
		Manopt	1	0.001062	2.000	stationary	-1.000	-1.000
	sb	PBA	50000	0.5982	3.000	global opt.	-6.674e-6	-6.674e-6
		Manopt	22	0.2152	3.000	global opt.	-1.241e-6	-1.183e-6
	tb	PBA	50000	0.4910	3.000	global opt.	-6.674e-6	-6.674e-6
		Manopt	22	0.1925	3.000	global opt.	-1.322e-6	-1.322e-6
	lww1	PBA	2	0.0001049	2.000	stationary	-1.000	-1.000
		Manopt	1	0.0009794	2.000	stationary	-1.000	-1.000
3	eye	PBA	2	6.970e-5	3.000	stationary	-1.000	-1.000
		Manopt	1	0.001210	3.000	stationary	-1.000	-1.000
	sb	PBA	14	0.0002474	4.000	stationary	-1.000	-1.000
		Manopt	8	0.03438	4.000	stationary	-1.000	-1.000
	tb	PBA	11	0.0001370	4.000	stationary	-1.000	-1.000
		Manopt	7	0.02888	4.000	stationary	-1.000	-1.000
	lww1	PBA	2	5.438e-5	3.000	stationary	-1.000	-1.000
		Manopt	1	0.0009759	3.000	stationary	-1.000	-1.000

For this set of $n \times d$ matrices, normal error with variance σ^2 to obtain A_i was added, $i = 1, \ldots, m$. The data matrix $\tilde{S} = (S_{ij})$ was constructed with $S_{ij} = A_i^T A_j$, $i \neq j$, and $S_{ii} = 0$. Values of $m = 5$, $n = 100$, and $d \in \{10, 20, \ldots, 100\}$ were used. The noise levels considered were $\sigma \in \{0.1, 1.0, 5.0, 10.0\}$. The rank r was set to 3. Initial value strategies "sb" and "tb" were used for both Algorithm 4.1 and Manopt, as they showed good performance in the small examples. One hundred samples of random sets were generated for each combination of simulation parameters. In Figure 1, error-versus-time curves are plotted for typical instances. Algorithm 4.1 was more than an order of magnitude faster than Manopt. (For $d = 100$, Manopt did not terminate for more than three days, hence the results were omitted.) There was little difference between the two initial value strategies, hence the differences of the final objective values between Algorithm 4.1 and Manopt are plotted in Figure 1 for strategy "tb" only. The final objective values of the two algorithms agreed in most cases, while Algorithm 4.1 tended to give larger values. The proportions of certified global optima are reported in Table 4 for d that are multiples of tens. Not surprisingly, when the noise level was low both Algorithm 4.1 and Manopt almost always solved (OTSM) globally. Even if σ was as large as 10.0, the success rate was between 8% and 24%.

5.4. Real-world examples.

Example 5.3 (cryo-EM). Ab initio modeling for the single-particle reconstruction (SPR) problem in cryo-EM refers to the procedure of obtaining a preliminary 3D map of the particle in the ice from 2D images by tomographic inversion. Since each cryo-EM image is a noisy projection of the particle with unknown orientation, reliable
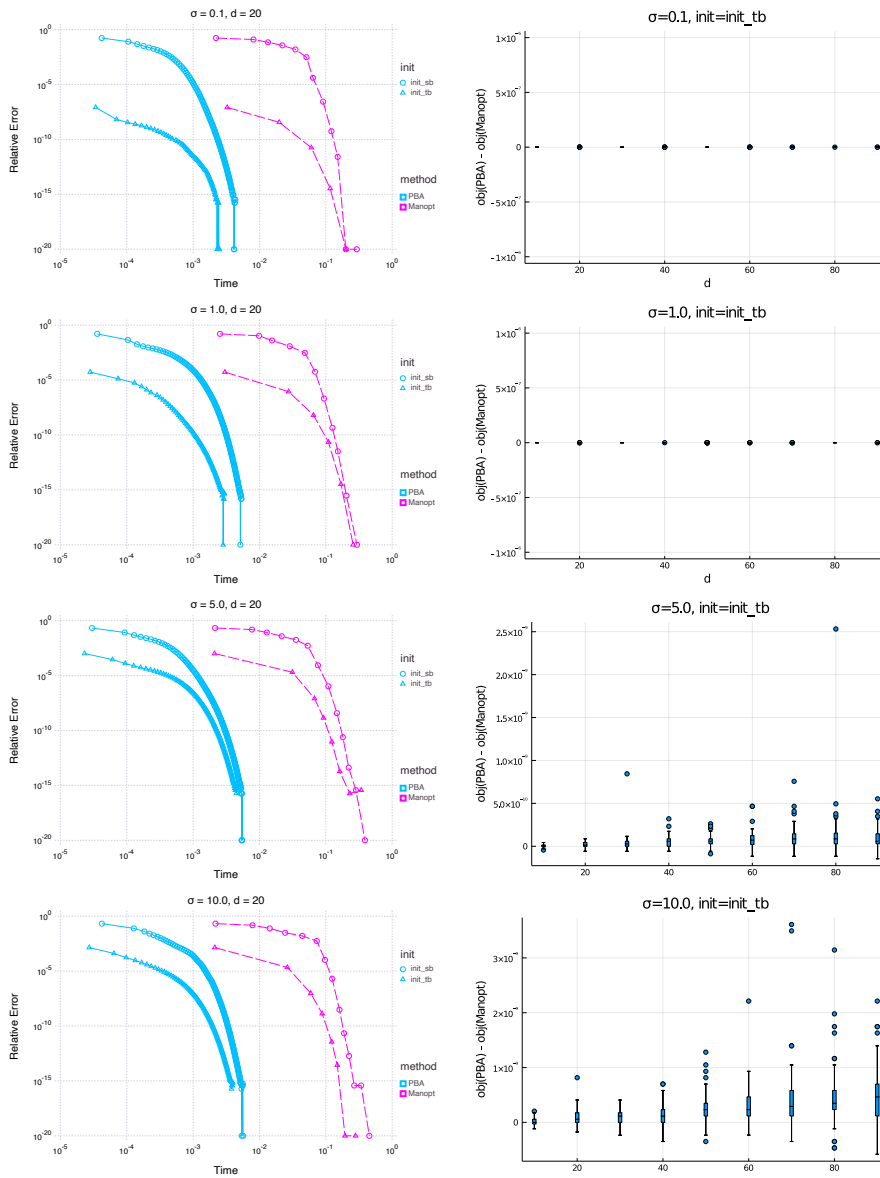
FIG. 1. *Simulation studies. Left: relative error versus wall clock time for each method and initialization strategy. Right: objective value difference between Algorithm 4.1 and Manopt at convergence.*

estimation of orientations from a collection of images is an important step in SPR. A popular approach is based on the common-lines property [6]: the Fourier slice theorem implies that any pair of projection images possesses a pair of radial lines on which their Fourier transforms coincide. Once the common lines of all the pairs among m projections are given, the orientations can be estimated via orthogonal least squares [37]. For a pair of images i and j, if the common line between images i and j appears in the direction of $c_{ij} = (\cos\theta_{ij}, \sin\theta_{ij}, 0)^T$ in image i and in $c_{ji} = (\cos\theta_{ji}, \sin\theta_{ji}, 0)^T$ in image j, then the unknown 3D rotation matrices O_i and O_j in the special orthogonal

TABLE 4
Frequency of certified global optimality, large examples.

d	Init	Method	Certification rate (σ)				d	Init	Method	Certification rate (σ)			
			0.1	1.0	5.0	10				0.1	1.0	5.0	10
10	sb	PBA	1.0	.94	.19	.17	60	sb	PBA	1.0	.78	.21	.20
		Manopt	1.0	.94	.19	.17			Manopt	1.0	.78	.21	.20
	tb	PBA	1.0	.94	.19	.17		tb	PBA	1.0	.78	.21	.20
		Manopt	1.0	.94	.19	.17			Manopt	1.0	.78	.21	.20
20	sb	PBA	1.0	.88	.14	.17	70	sb	PBA	1.0	.79	.14	.15
		Manopt	1.0	.88	.14	.17			Manopt	1.0	.79	.14	.15
	tb	PBA	1.0	.88	.14	.17		tb	PBA	1.0	.79	.14	.15
		Manopt	1.0	.88	.14	.17			Manopt	1.0	.79	.14	.15
30	sb	PBA	1.0	.91	.12	.12	80	sb	PBA	1.0	.91	.18	.14
		Manopt	1.0	.91	.12	.12			Manopt	1.0	.91	.18	.14
	tb	PBA	1.0	.91	.12	.12		tb	PBA	1.0	.91	.18	.14
		Manopt	1.0	.91	.12	.12			Manopt	1.0	.91	.18	.14
40	sb	PBA	1.0	.86	.10	.19	90	sb	PBA	1.0	.80	.21	.23
		Manopt	1.0	.86	.10	.19			Manopt	1.0	.80	.21	.23
	tb	PBA	1.0	.86	.10	.19		tb	PBA	1.0	.80	.21	.23
		Manopt	1.0	.86	.10	.19			Manopt	1.0	.80	.21	.23
50	sb	PBA	1.0	.86	.22	.21	100	sb	PBA	1.0	.76	.23	.21
		Manopt	1.0	.86	.22	.21			Manopt	-	-	-	-
	tb	PBA	1.0	.86	.22	.21		tb	PBA	1.0	.76	.23	.21
		Manopt	1.0	.86	.22	.21			Manopt	-	-	-	-

group $SO(3)$ should approximately satisfy $O_i^T c_{ij} = O_j^T c_{ji}$. Thus for estimating these matrices for all pairs among the m images, we may minimize

$$\sum_{i<j} \|O_i^T c_{ij} - O_j^T c_{ji}\|_F^2,$$

which is (OTSM) with $d_1 = \cdots = d_m = r = 3$, $S_{ij} = c_{ij}c_{ji}^T$ for $i \neq j$, and $S_{ii} = 0$, $i = 1, \ldots, m$ (i.e., MAXDIFF), but the domain is $SO(3) \times \cdots \times SO(3)$ instead of $\mathcal{O}_{3,3} \times \cdots \times \mathcal{O}_{3,3}$. Algorithm 4.1 can be trivially modified for this setting, since the projection of $B = PDQ^T \in \mathbb{R}^{r \times r}$ (full SVD) onto $SO(r)$ is $P \operatorname{diag}(1, \ldots, 1, -1)Q^T$ if the singular values of B are sorted in descending order.

We generated m noisy projections of a ribosomal subunit provided with ASPIRE software for SPR[2] that implements the orthogonal least squares method via SDP relaxation ($m = 100, 500, 1000$). The orientations of the projections were distributed uniformly over $SO(3)$. White Gaussian noise was added to the clean projections to generate noisy images of size 65 by 65 with signal-to-noise ratios (SNRs) ∞, 1, 1/2, 1/4, 1/8, and 1/16. Common-line pairs were detected with a $1°$ resolution using the functionality of ASPIRE. Due to the presence of noise, the common-line detection rate deteriorates as SNR decreases. Orientations were estimated using two methods, i.e., SDP relaxation of ASPIRE (which utilizes the SDPLR solver[3]) and Algorithm 4.1 initialized with "sb." The mean squared error of the estimated rotation matrices was computed using the formula of [37, eq. (8.2)]. Because we had difficulties in installing ASPIRE on the Macbook Pro laptop and common-line detection is computationally demanding, all the computations except for Algorithm 4.1 were conducted on a Linux

---

[2]Available at http://spr.math.princeton.edu/content/download-software.
[3]Available at http://sburer.github.io/files/SDPLR-1.03-beta.zip.

TABLE 5

Example 5.3. *"CL rate" refers to the common-line detection rate. In "classification," "station-ary" means that the iterate at termination is stationary but its global optimality is not confirmed by using Theorem* 3.3; *"not stat." means that it does not satisfy the first-order local optimality condition* (2.3).

m	SNR	CL rate	Method	MSE	Time (sec)	Iters	Classification
100	∞	0.9990	PBA	0.0000	0.0582	14	global opt.
			SDPLR	0.0000	1.3764	-	global opt.
	1	0.9442	PBA	0.0010	0.0309	13	global opt.
			SDPLR	0.0010	0.8879	-	global opt.
	1/2	0.8275	PBA	0.0148	0.0392	13	global opt.
			SDPLR	0.0148	0.8510	-	global opt.
	1/4	0.6048	PBA	0.1253	0.0333	15	global opt.
			SDPLR	0.1253	0.8089	-	global opt.
	1/8	0.3628	PBA	0.6646	0.0635	29	global opt.
			SDPLR	0.6646	0.9883	-	global opt.
	1/16	0.1834	PBA	2.1363	0.2265	110	stationary
			SDPLR	1.7991	1.3070	-	not stat.
500	∞	0.9994	PBA	0.0000	0.6996	14	global opt.
			SDPLR	0.0000	2.9501	-	global opt.
	1	0.8899	PBA	0.0038	0.6890	12	global opt.
			SDPLR	0.0038	2.3761	-	global opt.
	1/2	0.7250	PBA	0.0338	0.7755	14	global opt.
			SDPLR	0.0338	3.1371	-	global opt.
	1/4	0.4860	PBA	0.1959	0.9834	16	global opt.
			SDPLR	0.1959	5.4611	-	global opt.
	1/8	0.2678	PBA	0.7366	1.0938	20	global opt.
			SDPLR	0.7366	13.1841	-	global opt.
	1/16	0.1263	PBA	1.6252	1.5513	30	global opt.
			SDPLR	1.6252	11.2194	-	global opt.
1000	∞	0.9994	PBA	0.0000	2.5524	13	global opt.
			SDPLR	0.0000	10.1951	-	global opt.
	1	0.9177	PBA	0.0017	2.5475	13	global opt.
			SDPLR	0.0017	9.7137	-	global opt.
	1/2	0.7889	PBA	0.0188	2.4491	13	global opt.
			SDPLR	0.0188	19.7047	-	global opt.
	1/4	0.5686	PBA	0.1297	2.6885	14	global opt.
			SDPLR	0.1297	34.6575	-	global opt.
	1/8	0.3365	PBA	0.5384	3.8236	20	global opt.
			SDPLR	0.5384	67.5752	-	global opt.
	1/16	0.1687	PBA	1.3403	6.6967	35	stationary
			SDPLR	1.3403	102.6977	-	stationary

workstation with two Intel Xeon E5-2680v2@2.80GHz CPUs (256GB RAM) and eight Nvidia GTX 1080 GPUs (8GB VRAM/GPU).

The results are collected in Table 5. Except for the extremely challenging case with a low number of measurements ($m = 100$) and SNR (1/16), Algorithm 4.1 produced solutions of the same quality as ASPIRE/SDPLR in much shorter time (recall that ASPIRE was run on a much more powerful workstation), which, in turn, are certified to be globally optimal except the case $m = 1000$ and SNR $= 1/16$. Hence these solutions cannot be further improved under the least squares regime. In case the two methods disagree, the solution computed by using an SDP relaxation of the orthogonal least squares method failed to be even first-order stationary, possibly due to the rounding procedure of the SDP solution to $SO(3)$, even though the resulting MSE was lower.

Example 5.4 (generalized CCA). Our second real-world data example considers gene-level interaction analysis based on genotype data [27, 41]. Let $A_i \in \{0, 1, 2\}^{n \times d_i}$ be the genotype matrix of gene i, where n is the number of individuals and d_i is the number of single nucleotide polymorphisms (SNPs) in gene i. To test the interaction between m genes, the maximal canonical correlations among m genes, i.e., (OTSM), were computed. To demonstrate the scalability of the proximal block relaxation algorithm (Algorithm 4.1), we computed the top $r \in \{1, 2, 3\}$ generalized canonical correlations using the MAXDIFF criterion among the first $m \in \{2, \dots, 100\}$ genes on chromosome 1 of $n = 488,377$ samples from the UK Biobank [29]. (In contrast, conventional analyses [27, 41] are restricted to $m = 2$ and $r = 1$.) The numbers of SNPs d_i range from 10 to 271 with mean 34.33 in these genes. Figure 2 displays the run times, all under 15 seconds, of Algorithm 4.1 using the same convergence criteria as in section 5.1, together with the histogram of d_i's of the 100 genes. Among the 297 local solutions, 107 (36%) of them were certified to be globally optimal using Theorem 3.3.
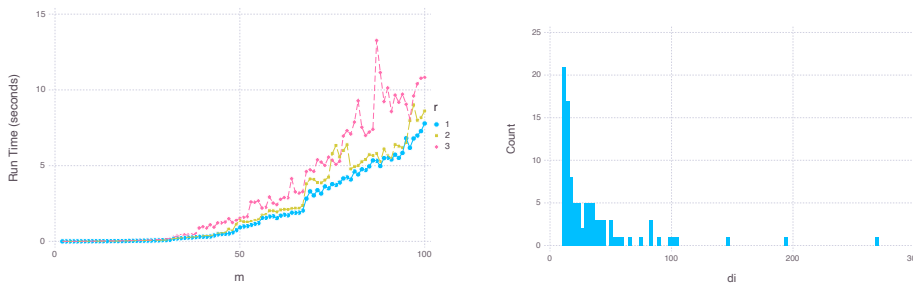

FIG. 2. *Left panel displays the run times of the proximal block relaxation algorithm (Algorithm 4.1) for finding the top r canonical correlations among the first m genes on chromosome 1 of $n = 488,377$ UK Biobank samples. Right panel shows the distribution of d_i (number of SNPs) in the first 100 genes.*

6. Conclusion. We have presented an in-depth analysis of the orthogonal trace-sum maximization (OTSM) problem, which subsumes various linear and quadratic optimization problems on a product of Stiefel manifolds. In a close analogy with classical results on eigenvalue optimization, a fairly general condition for certifying global optimality of a stationary point of the problem is derived. A practical algorithm to reach a stationary point with a global convergence guarantee is also proposed. We believe both are new to the literature. Numerical experiments show that the combination of our algorithm and certificate, with initial value strategies "sb" and "tb," can reveal global optima of various instances of OTSM. A further analysis on the probability of global optima of the algorithm, under some distributional assumption on the data, is warranted.

Appendix A. Technical proofs.

Proof of Proposition 3.1. Let $i \in \{1, \dots, m\}$ be such that $d_i > r$. Let $\tau_i = \lambda_{\min}(\Lambda_i)$, the smallest eigenvalue of the symmetric matrix Λ_i, and let $v_i \in \mathbb{R}^r$ be the associated unit eigenvector, i.e., $\Lambda_i v_i = \tau_i v_i$. If $O_i^\perp \in \mathcal{O}_{d_i, d_i - r}$ fills out O_i to a fully orthogonal matrix, then for any $x_i \in \mathbb{R}^{d_i - r}$, $W_i = O_i^\perp x_i v_i^T$ and $W_j = 0$ for $j \neq i$ satisfy the tangency condition (3.6). Then, $\operatorname{tr}(W^T \tilde{S} W) = \operatorname{tr}(W_i^T S_{ii} W_i) = $

$x_i^T O_i^{\perp T} S_{ii} O_i^\perp x_i$, and

$$\mathrm{tr}(\Lambda_i W_i^T W_i) = \mathrm{tr}(\Lambda_i v_i x_i^T O_i^{\perp T} O_i^\perp x_i v_i^T) = \mathrm{tr}(\Lambda_i v_i x_i^T x_i v_i^T) = \|x_i\|^2 v_i^T \Lambda_i v_i = \|x_i\|^2 \tau_i.$$

Further, $\mathrm{tr}(\Lambda_j W_j^T W_j) = 0$ for $j \neq i$. Thus the second-order condition (3.5) entails

$$0 \leq \sum_{i=1}^m \mathrm{tr}(\Lambda_i W_i^T W_i) - \mathrm{tr}(W^T SW) = x_i^T (\tau_i I_{d_i-r} - O_i^{\perp T} S_{ii} O_i^\perp) x_i.$$

Thus $\tau_i I_{d_i-r} \succeq O_i^{\perp T} S_{ii} O_i^\perp$. Since $O_i^{\perp T} S_{ii} O_i^\perp$ is positive semidefinite (recall $S_{ii} \succeq 0$), it follows that $\tau_i \geq 0$, proving the first claim.

Now suppose $O = (O_1, \ldots, O_m)$ is a stationary point with $\Lambda_i \not\succeq 0$ for some i. Denote the objective of (OTSM) by $f(O)$. Let the full singular value decomposition of Λ_i be $\Lambda_i = P_i D_i Q_i^T$, where $P_i, Q_i \in \mathcal{O}_{r,r}$, and $D_i \succeq 0$ is diagonal. Then it must be $P_i \neq Q_i$ since otherwise $\Lambda_i \succeq 0$. Now let $\Theta_i = O_i P_i Q_i^T \in \mathcal{O}_{d_i,r}$. Recall that $\Lambda_i = O_i^T \sum_{j=1}^m S_{ij} O_j$. It follows that

$$\mathrm{tr}\left(\Theta_i^T \sum_{j=1}^m S_{ij} O_j\right) = \mathrm{tr}\left(Q_i P_i^T O_i^T \sum_{j=1}^m S_{ij} O_j\right) = \mathrm{tr}(Q_i P_i^T \Lambda_i) = \mathrm{tr}(Q_i P_i^T P_i D_i Q_i^T)$$

$$= \mathrm{tr}(D_i)$$

$$> \mathrm{tr}(Q_i^T P_i D_i) = \mathrm{tr}(P_i D_i Q_i^T) = \mathrm{tr}(\Lambda_i) = \mathrm{tr}\left(O_i^T \sum_{j=1}^m S_{ij} O_j\right).$$

Note the update $O_i \leftarrow \Theta_i$ corresponds to Algorithm 4.1 (line 6) with $\alpha = \infty$, or the standard block ascent algorithm of section 4.1. By the ascent property of the algorithm, if we let $\Theta = (O_1, \ldots, O_{i-1}, \Theta_i, O_{i+1}, \ldots, O_m)$, then we have $f(O) < f(\Theta)$. Now if we run Algorithm 4.1 with $\alpha \in (0, 1/\max_{i=1,\ldots,m} \|S_{ii}\|_2)$, then the algorithm will converge to a stationary point $\tilde{O}$ by Theorem 4.1. Again by the ascent property of Algorithm 4.1, we see $f(O) < f(\Theta) \leq f(\tilde{O})$. If the associated Lagrange multipliers $\tilde{\Lambda}_i = \tilde{O}_i^T \sum_{j=1}^m S_{ij} \tilde{O}_j$ are not all positive semidefinite, then repeat the above procedure with $\Lambda_i \leftarrow \tilde{\Lambda}_i$ until no strict progress is possible. Eventually, we arrive at a stationary point with the desired property.

The above proof of the second claim also shows that $\Lambda_i \succeq 0$ for all i is a necessary condition for global optimality.

Proof of Proposition 3.2. It suffices to show the constraints $O_i \in \mathcal{O}_{d_i,r}$, $i = 1, \ldots, m$, are equivalent to the constraints of problem (3.8). From (3.7), clearly the former implies the latter. To show the opposite, first note that $U \succeq 0$ and $\mathrm{rank}(U) = r$ if and only if $mU = EE^T$, $E = [E_1^T, \ldots, E_m^T]^T \in \mathbb{R}^{D \times r}$, for some $E_i \in \mathbb{R}^{d_i \times r}$, $i = 1, \ldots, m$. Then $mU_{ii} = E_i E_i^T \preceq I_{d_i}$ and $\mathrm{tr}(mU_{ii}) = \mathrm{tr}(E_i^T E_i) = r$ jointly imply that all r singular values of E_i are 1. That is, $E_i \in \mathcal{O}_{d_i,r}$.

Proof of Corollary 3.6. Suppose $S_{12} = S_{21}^T \in \mathbb{R}^{d_1 \times d_2}$ has a singular value decomposition $S_{12} = U\Sigma V^T$ with $U \in \mathcal{O}_{d_1,d}$, $V \in \mathcal{O}_{d_2,d}$, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d) \in \mathbb{S}^d$, where $d = \min\{d_1, d_2\}$ and $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$. Since $O_1 O_2^T \in \mathbb{R}^{d_1 \times d_2}$ has r unit singular values and the rest are zero, the von Neumann–Fan inequality entails

$$\mathrm{tr}(O_1^T S_{12} O_2) = \mathrm{tr}[(O_1 O_2^T)^T S_{12}] \leq \sum_{i=1}^r \sigma_i,$$

with equality if and only if $O_1 = U_1 R$ (resp., $O_2 = V_1 R$), where $R \in \mathcal{O}_{r,r}$ and U_1 (resp., U_2) consists of the left (resp., right) singular vectors of S_{12} associated with $\sigma_1, \ldots, \sigma_r$ in this order. If we denote such orthogonal matrices by $\bar{O}_1$ and $\bar{O}_2$, then $\bar{O}_1^T U = [R\ 0] = \bar{O}_2^T V$ and $(\bar{O}_1, \bar{O}_2)$ is a globally optimal stationary point. The associate Lagrange multipliers are $\bar{\Lambda}_1 = \bar{O}_1^T S_{12} \bar{O}_2 = \bar{O}_2^T S_{21} \bar{O}_1 = \bar{\Lambda}_2$. Thus

$$\bar{\Lambda} = \bar{\Lambda}_1 = \bar{\Lambda}_2 = \bar{O}_1^T U \Sigma V^T \bar{O}_2 = R^T \Sigma_1 R = R^T \operatorname{diag}(\sigma_1, \ldots, \sigma_r) R,$$

and the diagonal blocks of the certificate matrix $L^\star$ are

$$L_{11}^\star = \bar{O}_1 \bar{\Lambda} \bar{O}_1^T + \sigma_r \bar{O}_1^\perp \bar{O}_1^{\perp T} = U \tilde{\Sigma} U^T, \quad L_{22}^\star = \bar{O}_2 \bar{\Lambda} \bar{O}_2^T + \sigma_r \bar{O}_2^\perp \bar{O}_2^{\perp T} = V \tilde{\Sigma} V^T,$$

where $\tilde{\Sigma} = \operatorname{diag}(\Sigma_1, I_{d-r}) \succeq \Sigma$. It follows that

$$L^\star = \begin{bmatrix} U \tilde{\Sigma} U^T & -U \Sigma V^T \\ -V \Sigma U^T & V \tilde{\Sigma} V^T \end{bmatrix} \succeq \begin{bmatrix} U \Sigma U^T & -U \Sigma V^T \\ -V \Sigma U^T & V \Sigma V^T \end{bmatrix} = \begin{bmatrix} U \\ -V \end{bmatrix} \Sigma \begin{bmatrix} U \\ -V \end{bmatrix}^T \succeq 0.$$

Proof of Corollary 3.8. The following lemma extends Theorem 2 of [28] for $S_{ii} \succeq 0$, $i = 1, \ldots, m$, and is used to prove the claim of the corollary.

LEMMA A.1. *If $S_{ij} = S_{ji}^T \in \mathbb{R}^{d_i \times d_j}$, $i, j = 1, \ldots, m$, has a singular value decomposition $S_{ij} = U_{ij} \Sigma_{ij} V_{ij}^T$ with $U_{ij} \in \mathcal{O}_{d_i,d}$, $V_{ij} \in \mathcal{O}_{d_j,d}$, and $\Sigma_{ij} = \operatorname{diag}(\sigma_1, \ldots, \sigma_d) \in \mathbb{S}^d$, where $d = \min\{d_i, d_j\}$ and $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$, then for $X = \operatorname{diag}(X_1, \ldots, X_m)$ with $X_i = \sum_{j=1}^m U_{ij} \Sigma_{ij} U_{ij}^T = \sum_{j=1}^m (S_{ij} S_{ij}^T)^{1/2}$, the symmetric matrix $X - \tilde{S}$, where $\tilde{S} = (S_{ij})$ is the data matrix, is positive semidefinite.*

Proof. Since $S_{ji} = U_{ji} \Sigma_{ji} V_{ji}^T = S_{ij}^T = V_{ij} \Sigma_{ij} U_{ij}^T$, we can set $V_{ij} = U_{ji}$ and $\Sigma_{ij} = \Sigma_{ji}$. For any $y = (y_1^T, \ldots, y_m^T)^T$ with $y_i \in \mathbb{R}^{d_i}$, let $a = \Sigma_{ij}^{1/2} U_{ij}^T y_i$ and $b = \Sigma_{ij}^{1/2} V_{ij}^T y_j = \Sigma_{ji}^{1/2} U_{ji}^T y_j$. Then from the fact $2a^T b \leq a^T a + b^T b$,

$$2y_i^T S_{ij} y_j = 2 y_i^T U_{ij} \Sigma_{ij} U_{ji}^T y_j \leq y_i^T U_{ij} \Sigma_{ij} U_{ij}^T y_i + y_j^T U_{ji} \Sigma_{ji} U_{ji}^T y_j.$$

Thus,

$$y^T (X - \tilde{S}) y = - \sum_{i,j: j \neq i} y_i^T S_{ij} y_j + \sum_{i=1}^m y_i^T (X_i - S_{ii}) y_i$$

$$\geq -\frac{1}{2} \sum_{i=1}^m y_i^T U_{ij} \Sigma_{ij} U_{ij}^T y_i - \frac{1}{2} \sum_{j=1}^m y_j^T U_{ji} \Sigma_{ji} U_{ji}^T y_j + \sum_{i=1}^m y_i^T \left(\sum_{j=1}^m (S_{ij} S_{ij}^T)^{1/2} - S_{ii} \right) y_i$$

$$= -\sum_{i=1}^m \sum_{j \neq i} y_i (S_{ij} S_{ij}^T)^{1/2} y_i + \sum_{i=1}^m \sum_{j \neq i} y_i (S_{ij} S_{ij}^T)^{1/2} y_i = 0. \qquad \square$$

Proof of Corollary 3.8. If S_{ij} has a singular value decomposition $S_{ij} = V_i \Sigma_{ij} V_j^T$, where $\Sigma_{ij} = \Sigma_{ji}$ is $r \times r$ nonnegative diagonal, then

$$X_i = \sum_{j=1}^m (S_{ij} S_{ij}^T)^{1/2} = V_i \left(\sum_{j=1}^m \Sigma_{ij} \right) V_j^T = V_i \Sigma_{i\cdot} V_i^T, \quad \Sigma_{i\cdot} = \sum_{j=1}^m \Sigma_{ij}.$$

From Lemma A.1, $X - \tilde{S} = \text{diag}(X_1, \ldots, X_m) - \tilde{S} \succeq 0$. Furthermore,

$$\sum_{j=1}^{m} S_{ij} V_j = \sum_{j=1}^{m} V_i \Sigma_{ij} V_j^T V_j = V_i \Sigma_{i\cdot}$$

and $\Sigma_{i\cdot} = \sum_{j=1}^{m} V_i^T S_{ij} V_j$. Thus we can set $\bar{O}_i = V_i$ and $\bar{\Lambda}_i = \Sigma_{i\cdot}$ in Theorem 3.3. Let τ_i be the smallest diagonal entry of $\Sigma_{i\cdot}$. Then,

$$\text{diag}\left(V_1 \Sigma_{1\cdot} V_1^T + \tau_1 V_1^\perp V_1^{\perp T}, \ldots, V_m \Sigma_{m\cdot} V_m^T + \tau_m V_m^\perp V_m^{\perp T}\right) \succeq \text{diag}(X_1, \ldots, X_m) = X,$$

hence $L^\star \succeq X - \tilde{S} \succeq 0$. $\qquad\square$

REFERENCES

[1] P.-A. ABSIL, C. G. BAKER, AND K. A. GALLIVAN, *Trust-region methods on Riemannian manifolds*, Found. Comput. Math., 7 (2007), pp. 303–330.

[2] P.-A. ABSIL AND J. MALICK, *Projection-like retractions on matrix manifolds*, SIAM J. Optim., 22 (2012), pp. 135–158, https://doi.org/10.1137/100802529.

[3] F. ALIZADEH, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51, https://doi.org/10.1137/0805002.

[4] H. ATTOUCH, J. BOLTE, P. REDONT, AND A. SOUBEYRAN, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457.

[5] A. S. BANDEIRA, C. KENNEDY, AND A. SINGER, *Approximating the little Grothendieck problem over the orthogonal and unitary groups*, Math. Program., 160 (2016), pp. 433–475.

[6] T. BENDORY, A. BARTESAGHI, AND A. SINGER, *Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities*, IEEE Signal Process. Mag., 37 (2020), pp. 58–76.

[7] J. BEZANSON, A. EDELMAN, S. KARPINSKI, AND V. B. SHAH, *Julia: A fresh approach to numerical computing*, SIAM Rev., 59 (2017), pp. 65–98, https://doi.org/10.1137/141000671.

[8] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Pure Appl. Math. 120, Academic Press, Orlando, FL, 1986.

[9] N. BOUMAL, B. MISHRA, P.-A. ABSIL, AND R. SEPULCHRE, *Manopt, a Matlab toolbox for optimization on manifolds*, J. Mach. Learn. Res., 15 (2014), pp. 1455–1459.

[10] I. L. DRYDEN AND K. V. MARDIA, *Statistical Shape Analysis: With Applications in R*, John Wiley & Sons, Hoboken, NJ, 2016.

[11] K. FAN, *On a theorem of Weyl concerning eigenvalues of linear transformations* I, Proc. Natl. Acad. Sci. USA, 35 (1949), pp. 652–655.

[12] G. H. FISCHER AND J. ROPPERT, *Ein Verfahren der Transformationsanalyse faktorenanalytischer Ergebnisse*, in Lineare Strukturen in Mathematik und Statistik, J. Roppert and G. H. Fischer, eds., Physica-Verlag, Wien, Austria, Würzburg, Germany, 1965, pp. 1–15 (in German).

[13] C. GOODALL, *Procrustes methods in the statistical analysis of shape*, J. Roy. Statist. Soc. Ser. B. Stat. Methodol., 53 (1991), pp. 285–339.

[14] J. C. GOWER, *Generalized Procrustes analysis*, Psychometrika, 40 (1975), pp. 33–51.

[15] M. HANAFI AND H. A. KIERS, *Analysis of k sets of data, with differential emphasis on agreement between and within sets*, Comput. Statist. Data Anal., 51 (2006), pp. 1491–1508.

[16] M. HANAFI AND J. M. TEN BERGE, *Global optimality of the successive Maxbet algorithm*, Psychometrika, 68 (2003), pp. 97–103.

[17] J.-B. HIRIART-URRUTY AND D. YE, *Sensitivity analysis of all eigenvalues of a symmetric matrix*, Numer. Math., 70 (1995), pp. 45–72.

[18] H. HOTELLING, *Relations between two sets of variates*, Biometrika, 28 (1936), pp. 321–377.

[19] K. LANGE, *Optimization*, 2nd ed., Springer, New York, 2013.

[20] K. LANGE, *MM Optimization Algorithms*, SIAM, Philadelphia, PA, 2016, https://doi.org/10.1137/1.9781611974409.

[21] X.-G. LIU, X.-F. WANG, AND W.-G. WANG, *Maximization of matrix trace function of product Stiefel manifolds*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1489–1506, https://doi.org/10.1137/15M100883X.

[22] MOSEK APS, *MOSEK Optimizer API for C. Version* 8.1.0 *(Revision* 34*)*, MOSEK ApS, 2017, https://docs.mosek.com/8.1/capi/index.html.

[23] Y. NESTEROV AND A. NEMIROVSKII, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, PA, 1994, https://doi.org/10.1137/1.9781611970791.

[24] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.

[25] M. L. OVERTON AND R. S. WOMERSLEY, *Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices*, Math. Program., 62 (1993), pp. 321–357.

[26] G. PATAKI, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Math. Oper. Res., 23 (1998), pp. 339–358.

[27] Q. PENG, J. ZHAO, AND F. XUE, *A gene-based method for detecting gene–gene co-association in a case–control association study*, Eur. J. Hum. Genet., 18 (2010), pp. 582–587.

[28] A. SHAPIRO AND J. BOTHA, *Dual algorithm for orthogonal Procrustes rotations*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 378–383, https://doi.org/10.1137/0609032.

[29] C. SUDLOW, J. GALLACHER, N. ALLEN, V. BERAL, P. BURTON, J. DANESH, P. DOWNEY, P. ELLIOTT, J. GREEN, M. LANDRAY, B. LIU, P. MATTHEWS, G. ONG, J. PELL, A. SILMAN, A. YOUNG, T. SPROSEN, T. PEAKMAN, AND R. COLLINS, *UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age*, PLoS Med., 12 (2015), e1001779.

[30] J. M. F. TEN BERGE, *Orthogonal Procrustes rotation for two or more matrices*, Psychometrika, 42 (1977), pp. 267–276.

[31] J. M. F. TEN BERGE, *Generalized approaches to the MAXBET problem and the MAXDIFF problem, with applications to canonical correlations*, Psychometrika, 53 (1988), pp. 487–494.

[32] J. M. F. TEN BERGE AND D. L. KNOL, *Orthogonal rotations to maximal agreement for two or more matrices of different column orders*, Psychometrika, 49 (1984), pp. 49–55.

[33] B. VAINSHTEIN AND A. GONCHAROV, *Determination of the spatial orientation of arbitrarily arranged identical particles of unknown structure from their projections*, Sov. Phys. Dokl., 31 (1986), pp. 278–280.

[34] J. P. VAN DE GEER, *Linear relations among k sets of variables*, Psychometrika, 49 (1984), pp. 79–94.

[35] M. VAN HEEL, *Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction*, Ultramicroscopy, 21 (1987), pp. 111–123.

[36] L. VANDENBERGHE AND S. BOYD, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95, https://doi.org/10.1137/1038003.

[37] L. WANG, A. SINGER, AND Z. WEN, *Orientation determination of cryo-EM images using least unsquared deviations*, SIAM J. Imaging Sci., 6 (2013), pp. 2450–2483, https://doi.org/10.1137/130916436.

[38] Y. XU AND W. YIN, *A globally convergent algorithm for nonconvex optimization based on block coordinate update*, J. Sci. Comput., 72 (2017), pp. 700–734.

[39] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[40] T. ZHANG AND A. SINGER, *Disentangling orthogonal matrices*, Linear Algebra Appl., 524 (2017), pp. 159–181.

[41] J. ZHAO, L. JIN, AND M. XIONG, *Test for interaction between two unlinked loci*, Am. J. Hum. Genet., 79 (2006), pp. 831–845.