SIAM J. OPTIM.
Vol. 30, No. 1, pp. 439–463

PRINCIPAL COMPONENT ANALYSIS BY OPTIMIZATION OF SYMMETRIC FUNCTIONS HAS NO SPURIOUS LOCAL OPTIMA*

ARMIN EFTEKHARI† AND RAPHAEL A. HAUSER‡

Abstract. Principal component analysis (PCA) finds the best linear representation of data and is an indispensable tool in many learning and inference tasks. Classically, principal components of a dataset are interpreted as the directions that preserve most of its "energy," an interpretation that is theoretically underpinned by the celebrated Eckart–Young–Mirsky theorem. This paper introduces many other ways of performing PCA, with various geometric interpretations, and proves that the corresponding family of nonconvex programs has no spurious local optima, while possessing only strict saddle points. These programs therefore loosely behave like convex problems and can be efficiently solved to global optimality, for example, with certain variants of the stochastic gradient descent. Beyond providing new geometric interpretations and enhancing our theoretical understanding of PCA, our findings might pave the way for entirely new approaches to structured dimensionality reduction, such as sparse PCA and nonnegative matrix factorization. More specifically, we study an unconstrained formulation of PCA using determinant optimization that might provide an elegant alternative to the deflating scheme commonly used in sparse PCA.

Key words. principal component analysis, nonconvex optimization, saddle point property

AMS subject classifications. 90C26, 15A23

DOI. 10.1137/18M1188495

1. Introduction.

Let $A \in \mathbb{R}^{m\times n}$ be a data matrix, with rows corresponding to m different data vectors and columns corresponding to n different features. Successful dimensionality reduction is at the heart of classification, regression, and other learning tasks that often suffer from the "curse of dimensionality," where having a small number of training samples in relation to the data dimension (namely, $m \ll n$) typically leads to overfitting [19].

To reduce the dimension of data from n to $p \leq n$, consider a matrix X with orthonormal columns. Then the rows of $AX \in \mathbb{R}^{m\times p}$ correspond to the data vectors (namely, the rows of A) projected onto the column span of X, which we denote by range(X). In particular, the new data matrix AX has reduced dimension p, while the number m of projected data vectors is unchanged; see Figure 1. Principal component analysis (PCA) is one of the oldest dimensionality reduction techniques that can be traced back to the work of Pearson [35] and Hotelling [23], motivated by the observation that often data lives near a lower-dimensional subspace of $\mathbb{R}^n$; see Figure 2. PCA identifies this subspace by finding a suitable matrix X that retains in AX as much as possible of the energy of A, and the optimal X is called the *loading matrix*. The columns of the loading matrix also reveal the hidden correlations between different features by identifying groups of variables that occur with jointly positive or jointly negative weights, for example, in gene expression data [2].

*Received by the editors May 21, 2018; accepted for publication (in revised form) November 26, 2019; published electronically February 6, 2020.

https://doi.org/10.1137/18M1188495

Funding: The work of the authors was supported by EPSRC grant EP/N510129/1. The work of the first author was also supported by the Alan Turing Institute through Turing Seed Funding grant SF019.

†EPFL, Lausanne, NW1 2DB, Switzerland (armin.eftekhari@gmail.com).

‡Mathematical Institute, Oxford University, Oxford, Oxfordshire, OX1 3LB, UK (hauser@maths.ox.ac.uk).

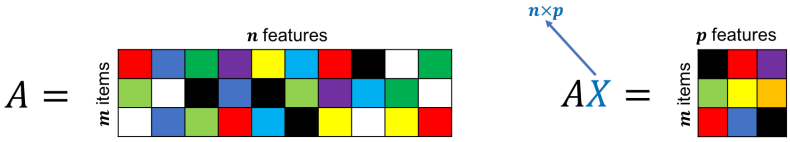

FIG. 1. *This figure illustrates the simple and powerful concept of linear dimensionality reduction. The left panel shows a data matrix A, with rows corresponding to m different data vectors and columns corresponding to n different features. For a matrix $X \in \mathbb{R}^{p\times r}$, the right panel shows the projected data matrix AX, containing again m data vectors (rows) but with only $p \leq n$ features (columns).*

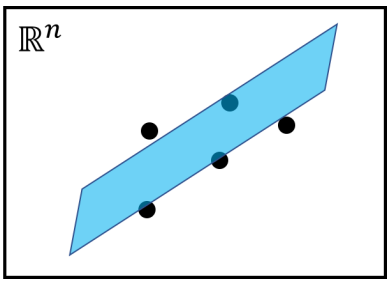

FIG. 2. *With each dot corresponding to a data vector with n features, PCA finds a linear subspace (in blue) that best represents the data vectors by capturing most of the energy of the dataset.*

PCA is also the building block of other dimensionality reduction techniques such as sparse PCA [28], kernel PCA [36, 37], multidimensional scaling [5], and nonnegative matrix factorization (NMF) [17]. For example, sparse PCA aims to find the important features of data by requiring the loading matrix to be sparse, namely, to have very few nonzero entries. Sparse PCA is useful, for instance, in studying gene expression data, where we are interested in singling out a small number of genes that are responsible for a certain trait or disease [1]. NMF, on the other hand, requires both AX and X to have nonnegative entries, which is valuable in recommender systems, for instance, where the data matrix A containing, say, film ratings is nonnegative and one would expect the same from the projected data matrix AX.

More formally, assume throughout this paper that the data matrix $A \in \mathbb{R}^{m\times n}$ is mean-centered. That is, $\sum_{i=1}^m a_i = 0$, where $a_i \in \mathbb{R}^n$ is the ith row of A, namely, the ith data vector. For $p \leq n$, let $\mathbb{R}_p^{n\times p}$ be the space of full-rank $n \times p$ matrices and consider the *trace inflation function*

$$f_{\mathrm{tr}} : \mathbb{R}_p^{n\times p} \to \mathbb{R}$$
$$X \mapsto \frac{\|AX\|_F^2}{\|X\|_F^2} = \frac{\mathrm{tr}(X^* A^* AX)}{\mathrm{tr}(X^* X)} \tag{1}$$

and the program

$$\arg\max \{f_{\mathrm{tr}}(X) : \ X \in \mathrm{St}(n,p)\}. \tag{2}$$

Above, $\|\cdot\|_F$ and $\mathrm{tr}(\cdot)$ return the Frobenius norm and trace of a matrix, respectively, and A^* is the transpose of matrix A. With $p \leq n$, $\mathrm{St}(n,p)$ above denotes the Stiefel manifold, the set of all $n \times p$ matrices with orthonormal columns. Note that when

$X \in \mathrm{St}(n,p)$, the denominator in the definition of $f_{\mathrm{tr}}(X)$ in (1) is constant, but we include this term to highlight the structural similarity with other inflation functions that we will study later.

It is a consequence of the celebrated Eckart–Young–Mirsky (EYM) theorem that a Stiefel matrix $X \in \mathrm{St}(n,p)$ is a global maximizer of program (2) if and only if it consists of p leading right singular vectors of A, namely, the right singular vectors of A corresponding to its p largest singular values [10, 32]. In other words, program (2) performs PCA on the data matrix A: the loading matrix, namely, the global maximizer of program (2), is a p-leading right singular factor $V_p \in \mathbb{R}^{n\times p}$ of A, and the projected data matrix $AV_p \in \mathbb{R}^{n\times p}$ contains the first p principal components of A.

Note also that program (2) is nonconvex because $\mathrm{St}(n,p) \subset \mathbb{R}^{n\times p}$ is a non-convex set. Even though nonconvex, program (2) behaves like a convex problem in the sense that any local maximizer of program (2) is also a global maximizer. Indeed, it is also a consequence of the EYM theorem that program (2) does not have any spurious local maximizers. Moreover, all saddle points of this program are *strict*, namely, have an ascent direction. Therefore the nonconvex program (2) can be efficiently solved (to global optimality) using (certain variants of) the stochastic gradient descent; see, for instance, [25, 33, 39]. Even more efficiently, solving program (2) or equivalently computing the loading matrix and the principal components of A can be done in $O(\max(m,n)p^2)$ operations using fast algorithms for singular value decomposition (SVD); see, for example, Algorithm 8.6.1 in [18].

Motivation. Our motivation for this work was the following simple observation. The interpretation of PCA as a dimensionality reduction tool suggests that it should suffice to find a matrix $X \in \mathbb{R}^{n\times p}_p$ whose columns span the optimal subspace, which corresponds to p leading right singular vectors of A. That is, one would expect f_{tr} in program (2) to be a function on the Grassmannian $\mathrm{Gr}(n,p)$, the set of all p-dimensional subspaces of $\mathbb{R}^n$. In other words, one would like f_{tr} to be invariant under an arbitrary change of basis in its argument.

That is of course not the case, as a quick inspection of (1) reveals. Generally, we have $f_{\mathrm{tr}}(X\Theta) = f_{\mathrm{tr}}(X)$, only when $\Theta \in \mathrm{Orth}(p)$, namely, when $\Theta \in \mathbb{R}^{p\times p}$ itself is an orthonormal matrix. Program (2) is thus inherently constrained to work with Stiefel matrices, a requirement that is not particularly onerous in the case of PCA but becomes a conceptual nuisance when considering structured dimensionality reduction, such as sparse PCA or NMF. Indeed, enforcing sparsity or nonnegativity in the columns of X in conjunction with orthogonality for the columns of X tends to be very restrictive and is perhaps a questionable objective in the first place.

Contributions. Motivated by the above observation, this paper introduces many other ways of performing PCA, with various geometric interpretations, and proves that the corresponding family of nonconvex programs have no spurious local optima, while possessing only strict saddle points. These new programs therefore loosely behave like convex problems and can be solved to global optimality in polynomial time with, for example, the variants of stochastic gradient ascent in [25, 33]. More specifically, replacing tr in f_{tr} with any elementary symmetric polynomial yields an equivalent formulation for PCA; see the family of problems in (14) and the even larger family of problems in (17).

Program (2) above is indeed a member of this large family. Another notable member of this family is program (6) below, which is effectively unconstrained and consequently does *not* require X to have orthonormal columns. This observation is of particular importance in practice. As we show in section 3, this unconstrained formulation of PCA in program (6) potentially allows for an elegant approach to

structured PCA, in which we wish to impose additional structure on the loading matrix, such as sparsity or nonnegativity.

Let us add that it is known already that program (6) is equivalent to PCA [21]; see [34] for an application to optimal design and [24] for an example in the context of independent component analysis. Of course, this equivalence does not guarantee that program (6), like program (2), can also be solved in polynomial time. In this sense, our contribution is that the nonconvex program (6) has no spurious local optima, has only strict saddle points, and can therefore be solved efficiently by certain variants of the stochastic gradient descent. Moreover, the introduction of the rest of this large family of equivalent formulations of PCA and their analysis in this work is the other novel aspect of this work.

Organization. The rest of this paper is organized as follows. To present this work in an increasing order of complexity, we first introduce in section 2 the unconstrained formulation of PCA, namely, program (6), and discuss in section 3 its potential application in structured dimensionality reduction. In section 4, we then present programs (14) and (17), a large family of equivalent formulations of PCA, of which both programs (2) and (6) are members. The claim that all these programs are indeed equivalent to PCA and can be efficiently solved is proven in sections 5 and 6 and the appendices.

2. PCA by determinant optimization. In analogy to $f_{\rm tr}$ in (1), let us define the *volume inflation function* by

$$\begin{aligned} f_{\rm det} : \mathbb{R}_p^{n\times p} &\to \mathbb{R}, \\ X &\mapsto \frac{\det(X^*A^*AX)}{\det(X^*X)}, \end{aligned} \tag{3}$$

where det stands for determinant and, in analogy to program (2), consider the program

$$\arg\max \left\{ f_{\rm det}(X) :\ X \in {\rm St}(n,p) \right\}. \tag{4}$$

Observe that programs (2) and (4) coincide for $p = 1$, namely, when we seek the leading principal component of the matrix A, in which case X^*A^*AX and X^*X are both positive scalars. Unlike $f_{\rm tr}$, note that $f_{\rm det}$ is invariant under an arbitrary change of basis. Indeed, for arbitrary $X \in \mathbb{R}_p^{n\times p}$ and $\Theta \in {\rm GL}(p)$, we have that

$$\begin{aligned} f_{\rm det}(X\Theta) &= \frac{\det(\Theta)^2 \det(X^*A^*AX)}{\det(\Theta)^2 \det(X^*X)} \\ &= f_{\rm det}(X), \end{aligned} \tag{5}$$

where ${\rm GL}(p)$ is the general linear group, the set of all invertible $p \times p$ matrices. That is, $f_{\rm det}$ is naturally defined on the Grassmannian ${\rm Gr}(n,p)$ and consequently Program (4) is equivalent to the program

$$\arg\max \left\{ f_{\rm det}(X) :\ X \in \mathbb{R}_p^{n\times p} \right\}. \tag{6}$$

Because $f_{\rm det}$ is invariant under any change of basis by (5), program (6) inherently constitutes an optimization over the Grassmannian ${\rm Gr}(n,p)$. Moreover, it is important that program (6) is effectively unconstrained because $\mathbb{R}_p^{n\times p}$ is an open subset of $\mathbb{R}^{n\times p}$ with nonempty interior. To summarize, the drawback of program (2) in section 1 which served as the motivation of this work is overcome by program (6), because it

is an unconstrained optimization program that involves an objective function defined naturally on the Grassmannian.

A key observation of this paper is that program (6) appears to be a good model for dimensionality reduction. Indeed, note that $X^*A^*AX \in \mathbb{R}^{p\times p}$ is the sample covariance matrix of the projected data AX. Consider the normal distribution $\mathcal{N}(0, X^*A^*AX)$ with zero mean and covariance matrix X^*A^*AX, which has ellipsoidal level sets of the form

$$\{z \in \mathbb{R}^p \,:\, z^*X^*A^*AXz = c\} \tag{7}$$

for arbitrary $c \geq 0$. Let B_c be the bounding box of this level set and note that the volume of B_c is $c^p\sqrt{\det(X^*A^*AX)}$. We can therefore interpret program (6) as maximizing the volume of this bounding box. That is, program (6), loosely speaking, finds the directions that maximize the volume of the projected dataset.

In contrast, program (2) maximizes the energy of the projected data. That is, program (2) maximizes the diameter of the above bounding box, namely, $c\sqrt{\mathrm{tr}(X^*A^*AX)}$, rather than its volume; see Figure 3. It is perhaps peculiar that $\mathrm{tr}(X^*A^*AX)$ is commonly referred to as the "total variance" of the dataset, for this quantity does not play any role in the normalizing constant of the normal distribution $\mathcal{N}(0, X^*A^*AX)$, whereas $\det(X^*A^*AX)$ does, in direct generalization of the role the variance plays in the one-dimensional case. At any rate, we see that programs (2) and (6) are both sensible approaches for linear dimensionality reduction but that their geometric justifications are very different. Somewhat surprisingly, we find that program (6) also

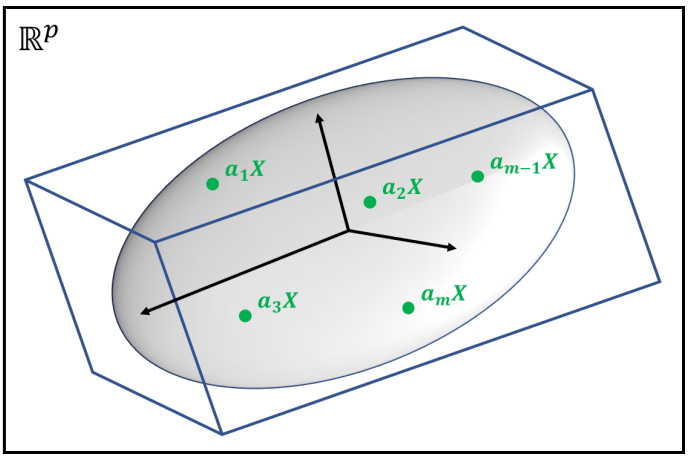

FIG. 3. *This figure illustrates the geometric intuition underlying this paper. Suppose that* $a_1, \ldots, a_m \in \mathbb{R}^n$ *are the rows of the data matrix* $A \in \mathbb{R}^{m\times n}$*, each representing a data vector. Then* $a_1X, \ldots, a_mX \in \mathbb{R}^p$ *are the projected data vectors, with reduced dimension of* $p \leq n$*. It is easy to see that the sample covariance matrix of these projected data vectors is* $X^*A^*AX \in \mathbb{R}^{p\times p}$*, with ellipsoidal level sets, one of which and its bounding box are displayed above. Then program* (2) *maximizes the diameter of this bounding box, which is proportional to* $\sqrt{\mathrm{tr}(X^*A^*AX)}$*. In contrast, program* (6) *maximizes the volume of this box, which is proportional to* $\sqrt{\det(X^*A^*AX)}$*. Remarkably, both programs* (2) *and* (6) *perform PCA of data matrix* A*; see sections* 1 *and* 2*. As discussed in section* 3*, program* (6) *is of particular importance in practice as it gives an elegant solution to the problem of structured linear dimensionality reduction. More generally, we also show that maximizing the sum of the volume squares of all* q*-dimensional facets of this bonding box is equivalent to PCA of matrix* A *for any* $1 \leq q \leq p$*; see section* 4*. In particular, programs* (2) *and* (6) *are special cases with* $q = 1$ *and* $q = p$*, respectively.*

performs PCA of A and has no spurious local optima, exactly like program (2). The next result is proven in section 5.

THEOREM 2.1 (determinant). *The following statements hold true:*

(i) $\widetilde{X} \in \mathbb{R}^{n\times p}$ *is a global maximizer of program* (6) *if and only if there exists a* p*-leading right singular factor* $V_p \in \mathbb{R}^{n\times p}$ *of* A *such that* $\mathrm{range}(\widetilde{X}) = \mathrm{range}(V_p)$.

(ii) *Program* (6) *does not have any spurious local optima, namely, any local maximum or minimum of program* (6) *is also a global maximum or minimum, respectively, and all other stationary points are strict saddle points. Moreover, if* $\sigma_p(A) > \sigma_{p+1}(A)$*, at any such strict saddle point* X_s*, there exists an ascent direction* $\Delta \in \mathbb{R}^{n\times p}$ *such that*

$$\nabla^2 f_{\mathrm{det}}(X_s)[\Delta,\Delta] \geq f_{\mathrm{det}}(X_s)\left(\frac{\sigma_p^2(A)}{\sigma_{p+1}^2(A)} - 1\right)\|\Delta\|_F^2, \tag{8}$$

where the bilinear operator $\nabla^2 f_{\mathrm{det}}(X_s)$ *is the Hessian of* f_{det} *at* X_s*. Above,* $\sigma_p(A)$ *is the* p*th largest singular value of* A*.*

In words, part (i) of Theorem 2.1 states that program (6) performs PCA on the data matrix A, and therefore programs (2) and (6) are equivalent in this sense. Note that program (6) provides a different geometric interpretation of PCA based on maximizing the "volume" of projected data rather than its "diameter," which was the case in program (2). Even though we present a new proof for the characterization of the global maximizers of program (6) in part (i) of Theorem 2.1, this result can also be proved using interlacing properties of singular values (see Corollary 3.2 in [22]) or via the Cauchy–Binet formula [29].

The main contribution of Theorem 2.1 is its part (ii) about the global landscape of the objective function f_{det}, stating that the nonconvex program (6) behaves like a convex problem in the sense that any local maximizer (minimizer) of program (6) is also a global maximizer (minimizer). Moreover, saddle points of program (6) are strict. In this way too, programs (2) and (6) are similar; see section 1. Note that part (ii) of Theorem 2.1 is crucial in the design of new dimensionality reduction algorithms: The instability of all stationary points except the global optima and the strictness of all saddle points establish that, for example, stochastic gradient ascent converges to the correct solution in polynomial time [25, 33]. That is, the nonconvex program (6) can be efficiently solved to global optimality. However, as discussed in section 1, computationally efficient algorithms for PCA are already available and application of, say, stochastic gradient ascent to program (6) is not intended to replace those algorithms. Instead, as discussed in section 3, the unconstrained program (6) potentially opens up a radically new approach to structured PCA.

We remark that Theorem 2.1 is in line with a recent trend in computational sciences to understand the geometry and performance of nonconvex programs and algorithms [31, 39, 12, 7, 6, 3, 4, 14, 26, 38, 15, 13, 16]. While the available results do not apply to our problem, the underlying phenomena are closely related. Perhaps the closest result to our work is [14], stating that the (nonconvex) matrix completion program has no spurious local optima when given access to randomly observed matrix entries. This result in a sense extends the EYM theorem [10, 32] to partially observed matrices.

From a computational perspective, we may consider the program

$$\arg\max \left\{\log(f_{\mathrm{det}}(X)) :\ X \in \mathbb{R}_p^{n\times p}\right\}, \tag{9}$$

which is equivalent to program (6) but has better numerical stability. As a numerical example, we generated generic $U, V \in \mathrm{Orth}(100)$ and random matrix $A \in \mathbb{R}^{100\times 100}$ with SVD $A = U\Sigma V^*$. The singular values of A, namely, the entries of the diagonal matrix $\Sigma \in \mathbb{R}^{100\times 100}$, were selected according to the power law. To be specific, we took $\sigma_i = i^{-1}$ to generate Figure 4 and $\sigma_i = i^{-2}$ to generate Figure 5, for every $i \in [100] = \{1, 2, \ldots, 100\}$. For $p = 5$, we let $V_p \in \mathbb{R}^{n\times p}$ denote the first p columns of V and, by Theorem 2.1, the unique maximizer of programs (6), and (9). (Note that V_p is also the unique maximizer of program (2) by the EYM theorem.) In order to find V_p, we then applied gradient ascent to program (9) with fixed step size of $\rho = 5$ and random initialization, producing a sequence of estimates $\{X_l\}_l \subset \mathbb{R}^{n\times p}$. We also recorded the error $\|X_l X_l^\dagger - V_p V_p^*\|$ in the lth iteration, namely, the sine of the principal angle between range(X_l) and range(V_p), which is plotted in Figures 4 and 5. As predicted by Theorem 2.1, the error vanishes in both examples as the algorithm progresses. We also refer the interested reader to [20] for a comparison between programs (2) and (9), as well as LAPACK's implementation of Lanczos' method [18] for performing PCA. While the numerical results presented in [20] are encouraging, a more comprehensive study is required to investigate the competitiveness of program (9) for PCA, as an alternative to more mainstream approaches [18].

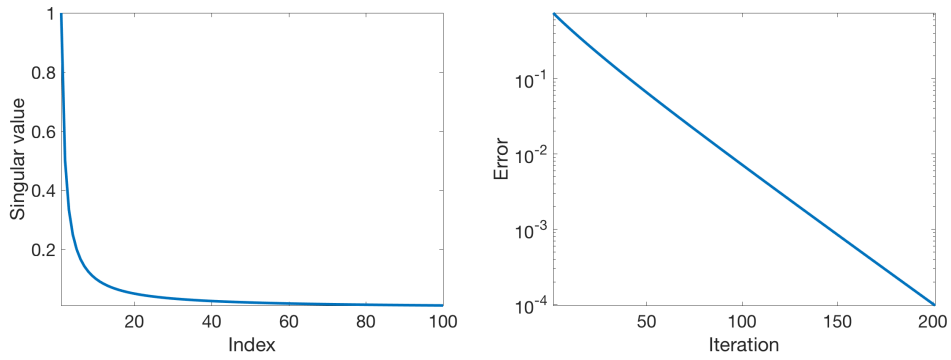

FIG. 4. *The left panel shows the spectrum* $\{\sigma_i\}_{i=1}^{100}$ *of a randomly generated matrix* $A \in \mathbb{R}^{100\times 100}$ *with* $\sigma_i = i^{-1}$*, and the right panel shows the progression of the gradient ascent algorithm with fixed step size, applied to program* (9)*; see section* 2 *for details.*

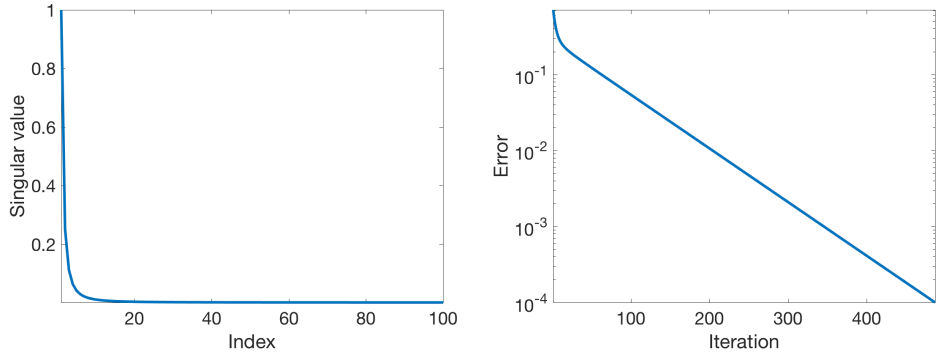

FIG. 5. *The left panel shows the spectrum* $\{\sigma_i\}_{i=1}^{100}$ *of a randomly generated matrix* $A \in \mathbb{R}^{100\times 100}$ *with* $\sigma_i = i^{-2}$*, and the right panel shows the progression of the gradient ascent algorithm with fixed step size, applied to program* (9)*; see section* 2 *for details.*

It might also be helpful to highlight the following practical consideration. Let $\widetilde{X} \in \mathbb{R}^{n\times p}$ denote a maximizer of program (6) or (9). Given $\widetilde{X}$, a few extra steps are required to compute the complete SVD of A, which we now list. Let $\widehat{X} \in \mathbb{R}^{n\times p}$ be an orthonormal basis for $\widetilde{X}$, which can be computed in $O(np^2)$ operations by SVD. Then computing the SVD of $A\widehat{X} = \widehat{U}\widehat{\Sigma}\widehat{V}^*$ can be performed in merely $O(mp^2)$ operations and yields the diagonal coefficients of $\widehat{\Sigma}$ as the p leading singular values of A, as well as $\widehat{U}$ and $\widehat{X}\widehat{V}$ as the corresponding p leading left and right singular vectors of A, respectively.

3. Structured PCA. As we will see in this section, the unconstrained formulation of PCA in program (6) might be of particular interest in practice, in contrast to program (2), which is restricted to the Stiefel manifold. Indeed, the determinant formulation of PCA in program (6) might allow for a more elegant approach to structured PCA, in which we wish to impose additional structure on the loading matrix, such as sparsity or nonnegativity.

For the purposes of this brief and informal discussion, let us focus on sparse PCA, the problem of finding a small number of features that best describe the data matrix $A \in \mathbb{R}^{m\times n}$. As one application, when working with gene expression data, we are interested in a small number of features (genes) which are responsible for certain traits or diseases [27, 9]. The "dual" of sparse PCA can also be interpreted as data clustering.

Loosely speaking, sparse PCA is the problem of finding a *sparse*[1] matrix $X \in \mathbb{R}^{n\times p}$ that retains, in the projected data $AX \in \mathbb{R}^{m\times p}$, as much as possible of the energy of A. More formally, sparse PCA might be formulated as a natural generalization of program (2), namely,

$$\arg\max \{f_{\mathrm{tr}}(X) : \ X \in \mathrm{St}(n,p) \text{ and } \|X\|_0 \leq k\}, \tag{10}$$

where $\|X\|_0$ is the number of nonzero entries of X, and the typically small integer k is the *sparsity level*. Note that program (10) forces X to have orthonormal columns *and* few nonzero entries, which tends to be restrictive and is also a somewhat questionable objective in the first place. With a few exceptions, particularly [28], this problem is often addressed by *deflating* A, namely, finding the sparse principal components sequentially, that is, one by one. Indeed, note that the Stiefel constraint from program (10) is redundant when $p = 1$, namely, when X is a column vector. One could therefore find the leading sparse principal component of A, say, $\widetilde{x}_1 \in \mathbb{R}^n$, by solving program (10) with $p = 1$, remove its contribution from A by forming $A_1 = A - A\widetilde{x}_1\widetilde{x}_1^*$, and then solve program (10) with A_1 in place of A to find the second sparse principal component $\widetilde{x}_2 \in \mathbb{R}^n$, and so on [8]. However, deflating A is believed to be inherently problematic when the problem is ill-posed [28].

The determinant formulation of PCA in program (9) might provide an elegant alternative to program (10). Recall that the feasible set of program (9) is an open subset of $\mathbb{R}^{n\times p}$ with nonempty interior, and thus program (9) is effectively unconstrained. We can therefore formulate sparse PCA by imposing a sparsity constraint on program (9), namely,

$$\arg\max \left\{\log(f_{\det}(X)) : \ X \in \mathbb{R}_p^{n\times p} \text{ and } \|X\|_0 \leq k\right\}, \tag{11}$$

which requires X to be full-rank and sparse, relaxing the far more restrictive requirement of being Stiefel and sparse in program (2). Note that removing the full-rank

[1] A sparse matrix has a small number of nonzero entries.

requirement in program (11) is impossible as that would mean A has fewer than p principal components and therefore the problem is ill-defined.

Similar ideas might be applied to NMF, in which X and AX are both required to be nonnegative. More generally, the unconstrained nature of program (9) might provide an entirely new approach to many structured dimensionality reduction problems, a research direction that remains to be explored in the future. In particular, what is the global geometry of program (11)? What is the precise relationship between programs (10) and (11)?

4. Generalization to positive symmetric polynomials. So far, we have seen that maximizing the trace objective function in program (2) and maximizing the determinant objective function in program (6) are equivalent, and both provide the leading principal components of the data matrix A. Moreover, both nonconvex programs can be solved to global optimality efficiently; see the discussion after Theorem 2.1, for example. Indeed, these claims for program (2) follow from the EYM theorem [10, 32] and the claims for program (6) follow from Theorem 2.1; see sections 1 and 2.

Note that both $\mathrm{tr}(X^*A^*AX)$ and $\det(X^*A^*AX)$ are *elementary symmetric polynomials*, namely, both are coefficients of the *characteristic polynomial* of X^*A^*AX. More specifically, let $X^*A^*AX = W_X \,\mathrm{diag}(\lambda_X) W_X^*$ be the eigendecomposition of X^*A^*AX, where $W_X \in \mathrm{Orth}(p)$ is an orthonormal matrix and the vector $\lambda_X \in \mathbb{R}^p$ contains the eigenvalues of X^*A^*AX. Here, $\mathrm{diag}(\lambda_X) \in \mathbb{R}^{p\times p}$ is the diagonal matrix formed by the vector λ_X. Then the characteristic polynomial associated with X^*A^*AX takes $t \in \mathbb{R}$ to

$$\begin{aligned}
\det(\mathrm{I}_p + tX^*A^*AX) &= \det(W_X(\mathrm{I}_p + t\,\mathrm{diag}(\lambda_X))W_X^*) \qquad (W_X \in \mathrm{Orth}(p)) \\
&= \det(W_X)\cdot\det(\mathrm{I}_p + t\,\mathrm{diag}(\lambda_X))\cdot\det(W_X^*) \\
&= \det\begin{bmatrix} 1+t\cdot\lambda_{X,1} & 0 & \cdots & 0 \\ 0 & 1+t\cdot\lambda_{X,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & 1+t\cdot\lambda_{X,p} \end{bmatrix} \quad (\det(W_X)=1) \\
&= \prod_{i=1}^{p}(1+t\cdot\lambda_{X,i}) \\
&=: \sum_{q=0}^{p} s_q(X^*A^*AX)\cdot t^q, \qquad (12)
\end{aligned}$$

where the qth elementary symmetric polynomial $s_q : \mathrm{Sym}(p) \to \mathbb{R}$ is the coefficient of t^q above, namely,

$$\begin{aligned}
&s_q : \mathrm{Sym}(p) \to \mathbb{R}, \\
&B \mapsto \sum_{1\le i_1<\cdots<i_q\le p} \prod_{j=1}^{q} \lambda_{B,i_j}, \qquad (13)
\end{aligned}$$

with the convention that $s_0(B) = 1$. Above, $\{\lambda_{B,i}\}_{i=1}^p$ are the eigenvalues of $B \in \mathrm{Sym}(p)$, and $\mathrm{Sym}(p)$ denotes the set of symmetric $p\times p$ matrices over the real numbers. We also remark that elementary symmetric polynomials are *spectral* functions in that they only depend on the eigenvalues of the input matrix. As mentioned earlier, $\mathrm{tr}(X^*A^*AX) = s_1(X^*A^*AX)$ and $\det(X^*A^*AX) = s_p(X^*A^*AX)$.

In analogy to trace and determinant objective functions (1), (3), let us define

$$f_{s_q} : \mathbb{R}^{n\times p} \to \mathbb{R}, \qquad X \mapsto \frac{s_q(X^*A^*AX)}{s_q(X^*X)} \tag{14}$$

for every $q \in [p] := \{1, \dots, p\}$. In particular, $f_{s_1} = f_{\mathrm{tr}}$ in (1) and $f_{s_p} = f_{\det}$ in (3) are two special cases. Last, in analogy to programs (2) and (4), consider the program

$$\arg\max \left\{ f_{s_q}(X) : \ X \in \mathrm{St}(n,p) \right\}. \tag{15}$$

Again note that programs (2) and (4) are special cases of program (15) for $q = 1$ and $q = p$, respectively. Revisiting the geometric interpretation discussed in section 2, we may also verify that $f_{s_q}(X)$ is proportional to the sum of volumes squared of all q-dimensional facets of the bounding box B_c; see right after (7) and also Figure 3. In this sense, program (15) finds the projected data AX that maximizes this geometric attribute.

Generalizing the EYM theorem for program (2) and Theorem 2.1 for program (4), the following result states that program (15) performs PCA of A and has no spurious local optima for every $q \in [p]$; see section 6 for the proof.

THEOREM 4.1 (elementary symmetric polynomials). *For every $q \in [p]$, the following statements hold true:*

(i) *$\widetilde{X} \in \mathbb{R}^{n\times p}$ is a global maximizer of program* (15) *if and only if there exists a p-leading right singular factor $V_p \in \mathbb{R}^{n\times p}$ of A such that* $\mathrm{range}(\widetilde{X}) = \mathrm{range}(V_p)$.

(ii) *Program* (15) *does not have any spurious local optima, namely, any local maximum or minimum of program* (15) *is also a global maximum, respectively, minimum, and all other stationary points are strict saddle points.*

In words, Theorem 4.1 introduces a family of equivalent formulations for PCA, namely, program (4.1) for every $q \in [p]$. This family includes PCA by trace optimization (program (2)) and PCA by determinant optimization (program (6)). In fact, maximizing *any* conic combination of elementary symmetric polynomials also performs PCA. To be specific, for nonnegative (but not all zero) coefficients $\{w_q\}_{q=0}^p$, consider the symmetric function

$$g_w : \mathrm{Sym}(p) \to \mathbb{R}, \qquad B \mapsto \sum_{q=0}^{p} w_q \cdot s_q(B), \tag{16}$$

where the elementary symmetric polynomial s_q was defined in (13). Also define

$$f_{g_w} : \mathbb{R}^{n\times p} \to \mathbb{R}, \qquad X \mapsto \frac{g_w(X^*A^*AX)}{g_w(X^*X)}, \tag{17}$$

and consider the program

$$\arg\max \left\{ f_{g_w}(X) : \ X \in \mathrm{St}(n,p) \right\}. \tag{18}$$

The following result is an immediate consequence of Theorem 4.1 and states that program (18) for *any* positive symmetric function performs PCA, thus providing a broad class of equivalent formulations for PCA.

COROLLARY 4.2 (positive symmetric functions). *Let $\{w_q\}_{q=0}^p$ be a set of nonnegative coefficients with at least one $w_q \neq 0$, and let f_{g_w} be the function defined in* (17)*. Then the following statements hold true:*

(i) *$\widetilde{X} \in \mathbb{R}^{n\times p}$ is a global maximizer of Program* (18) *if and only if there exists a p-leading right singular factor $V_p \in \mathbb{R}^{n\times p}$ of A such that* $\operatorname{range}(\widetilde{X}) = \operatorname{range}(V_p)$.

(ii) *Program* (18) *does not have any spurious local maximizers, namely, any local maximizer of program* (18) *is also a global maximizer, and all other stationary points are strict saddle points.*

Proof. Without loss of generality, we can assume that $w_0 = 0$. Indeed, since $s_0 = 1$ is constant by definition, setting $w_0 = 0$ does not change the optima and stationary points of program (18). Note that $X^*X = \mathrm{I}_p$ for every X feasible to programs (15) and (18). Therefore program (18) has the same optima and stationary points as

$$\arg\max \left\{ g_w(X^*A^*AX) : \; X \in \mathrm{St}(n,p) \right\}, \tag{19}$$

which, after recalling (16), has in turn the same optima and stationary points as

$$\arg\max \left\{ \sum_{q=0}^{p} w_q \cdot \frac{s_q(X^*A^*AX)}{s_q(X^*X)} \; X \in \mathrm{St}(n,p) \right\}. \tag{20}$$

Recall Theorem 4.1 about program (15) for every $q \in [p]$. Because the coefficients $\{w_q\}_q$ are nonnegative by assumption, the claims in Theorem 4.1 extend to program (20) and in turn to program (19) and then to program (18). This completes the proof of Corollary 4.2. ∎

In conclusion, this paper introduces a large family of equivalent interpretations of PCA which can all be solved to global optimality in polynomial time. One member of this family is an unconstrained formulation of PCA that might lead in the future to developing new algorithms and techniques for structured PCA.

5. Proof of Theorem 2.1. We first begin with a change of variables. Let $A = U\Sigma V^*$ be the SVD of A, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthonormal matrices, and the diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ is formed by the singular values of A in nonincreasing order, denoted by $\sigma_1 \geq \sigma_2 \geq \cdots$. Let us set $\Gamma := \Sigma^*\Sigma \in \mathbb{R}^{n\times n}$ for short and note that

$$\Gamma = \operatorname{diag}(\gamma), \qquad \text{where } \gamma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_r^2 & 0 & \cdots \end{bmatrix}^* \in \mathbb{R}^n, \tag{21}$$

where $\operatorname{diag}(\gamma)$ shapes the vector γ into a diagonal matrix, and $r = \operatorname{rank}(A)$ is the rank of A, namely, the number of positive singular values of A. Under the change of variables from X to $Y = V^*X$, Program (6) is equivalent to

$$\arg\max \left\{ \frac{\det(Y^*\Gamma Y)}{\det(Y^*Y)} \; : \; Y \in \mathbb{R}_p^{n\times p} \right\}. \tag{22}$$

Without loss of generality, we will therefore assume that $A^*A = \Gamma$ in (3), namely, we henceforth set

$$f_{\det}(X) = \frac{\det(X^*\Gamma X)}{\det(X^*X)}. \tag{23}$$

We will prove Theorem 2.1 by studying the stationary points of program (6). This program is unconstrained and therefore a stationary point $X_s \in \mathbb{R}_p^{n\times p}$ of program

(6) is characterized by $\nabla f_{\det}(X_s) = 0$. In light of the invariance in (5), we can also assume without loss of generality that $X_s \in \mathrm{St}(n,p)$. A stationary point $X_s \in \mathrm{St}(n,p)$ of Program (6) thus satisfies

$$\nabla f_{\det}(X_s) = 2\Gamma X_s \nabla \det(X_s^* \Gamma X_s) - 2\det(X_s^* \Gamma X_s) \cdot X_s \nabla \det(X_s^* X_s) = 0. \tag{24}$$

Note that the determinant is a *spectral function*, namely, it only depends on the eigenvalues of its input matrix. More precisely, for a matrix $Z \in \mathbb{R}^{p\times p}$, it holds that

$$\det(Z) = e_p(\lambda_Z) := \prod_{i=1}^{p} \lambda_{Z,i}, \tag{25}$$

where $\lambda_{Z,i}$ is the ith eigenvalue of Z and the vector $\lambda_Z = [\lambda_{Z,1}, \ldots, \lambda_{Z,p}]$ contains the eigenvalues of Z. The derivative of a spectral function is well-known. To be specific, let $Z = W_Z \cdot \mathrm{diag}(\lambda_Z) \cdot W_Z^*$ be the eigendecomposition of Z, where $W_Z \in \mathrm{Orth}(p)$ is an orthonormal matrix and, as before, $\mathrm{diag}(\lambda_Z) \in \mathbb{R}^{p\times p}$ is the diagonal matrix formed by the vector λ_Z. For a spectral function $\phi : \mathrm{Sym}(p) \to \mathbb{R}$, there exists a *symmetric function* $\psi : \mathbb{R}^p \to \mathbb{R}$ such that

$$\phi(Z) = \phi(\mathrm{diag}(\lambda_Z)) = \psi(\lambda_z), \tag{26}$$

where we recall that a symmetric function is a function that remains invariant after changing the order of its arguments. We then have from [30] that

$$\nabla\phi(Z) = W_Z \cdot \mathrm{diag}(\nabla\psi(\lambda_Z)) \cdot W_Z^*. \tag{27}$$

In our case, with $\phi = e_p$ and when $Z \in \mathrm{GL}(p)$, namely, when Z is nonsingular, we have that

$$\begin{aligned}
\nabla \det(Z) &= W_Z \cdot \mathrm{diag}(\nabla e_p(\lambda_Z)) \cdot W_Z^* \qquad \text{(see (27))} \\
&= W_Z \cdot \mathrm{diag}\left(e_p(\lambda_Z) \begin{bmatrix} \frac{1}{\lambda_{Z,1}} & \cdots & \frac{1}{\lambda_{Z,p}} \end{bmatrix}^* \right) \cdot W_Z^* \qquad \text{(see (25))} \\
&= \det(Z) W_Z \cdot \mathrm{diag}\left(\begin{bmatrix} \frac{1}{\lambda_{Z,1}} & \cdots & \frac{1}{\lambda_{Z,p}} \end{bmatrix}^* \right) \cdot W_Z^* \qquad \text{(see (25))}.
\end{aligned} \tag{28}$$

Returning to the proof and recalling that $X_s \in \mathrm{St}(n,p)$, (28) allows us to write that

$$\nabla \det(X_s^* X_s) = I_p. \tag{29}$$

Likewise, after recalling that $X_s^* \Gamma X_s \in \mathrm{GL}(p)$ by assumption, it follows from (28) that

$$\nabla \det(X_s^* \Gamma X_s) = \det(X_s^* \Gamma X_s) \cdot (X_s^* \Gamma X_s)^{-1}. \tag{30}$$

Substituting (29), (30), we find that (24) holds if and only if $\Gamma X_s = X_s X_s^* \Gamma X_s$, and since $X_s X_s^*$ is the projection into $\mathrm{range}(X_s)$, this is true if and only if $\mathrm{range}(\Gamma X_s) \subseteq \mathrm{range}(X_s)$. From this it follows that $X_s \in \mathrm{St}(n,p)$ such that $X_s^* \Gamma X_s \in \mathrm{GL}(p)$ is a stationary point of program (6), if and only if

$$\mathrm{range}(X_s) = \mathrm{range}(\Gamma X_s). \tag{31}$$

The following result, proved in Appendix A, characterizes the stationary points of program (6). That is, the next result characterizes matrices $X_s \in \mathrm{St}(n,p)$ that satisfy (31).

LEMMA 5.1. *For a singular value σ_j of A, let n_j denote the multiplicity of σ_j in A. Suppose that $X_s \in \mathrm{St}(n,p)$ satisfies $X_s^*\Gamma X_s \in \mathrm{GL}(p)$ and $\mathrm{range}(X_s) = \mathrm{range}(\Gamma X_s)$. Then there exists an index set $J \subset [n]$ such that*

1. *$\{\sigma_i\}_{i\in J}$ are distinct and positive, and*
2. *for every $i \notin J$, the rows of X_s corresponding to σ_i are zero, and*
3. *for every $i \in J$, the rows of X_s corresponding to σ_i span a $\dim(\sigma_i)$-dimensional subspace of $\mathbb{R}^p$, and*
4. *for every distinct pair $\{i,i'\} \subset J$, the corresponding subspaces are orthogonal, and finally*
5. *$\sum_{i\in J} \dim(\sigma_i) = p$.*

Above, for every $i \in J$, we set

$$\dim(\sigma_i) = \max\left[p - \sum_{i'\neq i} n_{i'}, 1\right].$$

Based on the characterization of stationary points in Lemma 5.1, we next calculate the Hessian of f_{det} at a stationary point of program (6), which will later help us determine the stability of these stationary points. The following result is in fact more general; see Appendix B for the proof.

LEMMA 5.2. *Consider a differentiable spectral function $\phi : \mathrm{Sym}(p) \to \mathbb{R}$ and the associated symmetric function $\psi : \mathbb{R}^p \to \mathbb{R}$; see* (26). *Consider also the function*

$$\begin{aligned} f_\phi : \mathbb{R}_p^{n\times p} &\to \mathbb{R}, \\ X &\mapsto \frac{\phi(X^*\Gamma X)}{\phi(X^*X)}. \end{aligned} \tag{32}$$

Consider last $X_s \in \mathrm{St}(n,p)$ such that $\mathrm{range}(X_s) = \mathrm{range}(\Gamma X_s)$ and the corresponding index set $J = \{i_1, i_2, \cdots\} \subseteq [n]$ in Lemma 5.1. *Then we can assume without loss of generality that X_s is block-diagonal. Under this assumption, it holds that*

$$X_s^*\Gamma X_s = \mathrm{diag}(\widetilde{\gamma}_1), \tag{33}$$

where

$$\widetilde{\gamma}_1 := \left[\ \overbrace{\sigma_{i_1}^2\cdots\sigma_{i_1}^2}^{\dim(\sigma_{i_1})}\ \ \overbrace{\sigma_{i_2}^2\cdots\sigma_{i_2}^2}^{\dim(\sigma_{i_2})}\ \ \cdots\ \right]^* \in \mathbb{R}^p. \tag{34}$$

Moreover, let $K \subset [n]$ denote the index set corresponding to the nonzero rows of X_s. Then it holds that

$$\nabla^2 f_\phi(X_s)[\Delta,\Delta] = \frac{1}{\phi(\mathrm{I}_p)} \sum_{i\in K^C} \sum_{j=1}^{p} \left(\sigma_i^2 \partial_j\psi(\widetilde{\gamma}_1) - f_\phi(X_s)\partial_1\psi(1_p)\right)\Delta_{i,j}^2 \tag{35}$$

for every $\Delta \in \mathbb{R}^{n\times p}$ that is zero on the rows indexed by K. Above, the bilinear operator $\nabla^2 f_\phi(X_s) : \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p} \to \mathbb{R}$ is the Hessian of f_ϕ at X_s. Also, K^C is the complement of the set K, $\partial_i\psi(\widetilde{\gamma}_1)$ is the ith entry of the gradient vector $\nabla\psi(\widetilde{\gamma}_1) \in \mathbb{R}^p$, and $1_p \in \mathbb{R}^p$ is the vector of all ones.

In particular, when $\phi = \det$ and, consequently, $\psi = e_p$ in Lemma 5.2, let us simplify the expression for the Hessian in (35) by noting that

$$\phi(\mathrm{I}_p) = \det(\mathrm{I}_p) = 1,$$

$$\begin{aligned}
\partial_j \psi(\widetilde{\gamma}_1) &= \partial_j e_p(\widetilde{\gamma}_1) \\
&= \prod_{i \neq j} \widetilde{\gamma}_{1,i} \\
&= \frac{\prod_{i=1}^{p} \widetilde{\gamma}_{1,i}}{\widetilde{\gamma}_{1,j}} \\
&= \frac{\det(\mathrm{diag}(\widetilde{\gamma}_1))}{\widetilde{\gamma}_{1,j}} \\
&= \frac{\det(X_s^* \Gamma X_s)}{\widetilde{\gamma}_{1,j} \det(X_s^* X_s)} \qquad (\Gamma = \mathrm{diac}(\widetilde{\gamma}_1) \text{ and } X_s \in \mathrm{St}(n,p)) \\
&= \frac{f_{\det}(X_s)}{\widetilde{\gamma}_{1,j}} \qquad \text{(see (32))}.
\end{aligned}$$

$$\partial_1 \psi(1_p) = \partial_1 e_p(1_p) = 1. \tag{36}$$

For any $\Delta \in \mathbb{R}^{n \times p}$ that is zero on the rows indexed by K, substituting the above values back into (35) in Lemma 5.2 with $\phi = \det$ yields that

$$\nabla^2 f_{\det}(X_s)[\Delta, \Delta] = f_{\det}(X_s) \sum_{i \in K^C} \sum_{j=1}^{p} \left(\frac{\sigma_i^2}{\widetilde{\gamma}_{1,j}} - 1 \right) \Delta_{i,j}^2. \tag{37}$$

If $\{\sigma_i\}_{i \in J}$ are *not* the unique numbers in the p leading singular values of A, then there exists $\Delta \in \mathbb{R}^{n \times p}$ such that $\nabla^2 f_{\det}(X_s)[\Delta, \Delta] > 0$, namely, Δ is an ascent direction at X_s. Indeed, let σ_{i_0} with $i_0 \in [n]$ be one of the p leading singular values of A, not listed in $\{\sigma_i\}_{i \in J}$. Then it holds that

$$\sigma_{i_0}^2 > \min_{i \in J} \sigma_i^2 = \min_{j \in [p]} \widetilde{\gamma}_{1,j} =: \widetilde{\gamma}_{1,j_0} \qquad \text{(see (34))} \tag{38}$$

and,[2] moreover,

$$\frac{\sigma_{i_0}^2}{\widetilde{\gamma}_{1,j_0}} \geq \frac{\sigma_p^2}{\sigma_{p+1}^2}. \tag{39}$$

Let $\Delta \in \mathbb{R}^{n \times p}$ be such that Δ_{i_0,j_0} is its only nonzero entry and note that this choice of Δ is indeed zero on the rows indexed by K. With this choice of Δ in (37), we find that

$$\nabla^2 f_{\det}(X_s)[\Delta, \Delta] = f_{\det}(X_s) \left(\frac{\sigma_{i_0}^2}{\widetilde{\gamma}_{1,j_0}} - 1 \right) \Delta_{i_0,j_0}^2 > 0 \qquad \text{(see (38))}. \tag{40}$$

That is, if $\{\sigma_i\}_{i \in J}$ are *not* the unique numbers in the p leading singular values of A, then there exists an ascent direction at X_s. Moreover, if there is a nontrivial spectral gap $\sigma_p > \sigma_{p+1}$, then it also holds that

$$\begin{aligned}
\nabla^2 f_{\det}(X_s)[\Delta, \Delta] &= f_{\det}(X_s) \left(\frac{\sigma_{i_0}^2}{\widetilde{\gamma}_{1,j_0}} - 1 \right) \Delta_{i_0,j_0}^2 \\
&\geq f_{\det}(X_s) \left(\frac{\sigma_p^2}{\sigma_{p+1}^2} - 1 \right) \Delta_{i_0,j_0}^2 \qquad \text{(see (39))} \\
&= f_{\det}(X_s) \left(\frac{\sigma_p^2}{\sigma_{p+1}^2} - 1 \right) \|\Delta\|_F^2, \qquad (41)
\end{aligned}$$

[2] In (38), the index j_0 might not be uniquely defined.

where the last line uses the fact that Δ_{i_0,j_0} is the only nonzero entry of Δ above. Likewise, we can establish that if $\{\sigma_i\}_{i\in J}$ are *not* the unique numbers in the p trailing singular values of A, then there exists a descent direction at X_s. We conclude that if $\{\sigma_i\}_{i\in J}$ are neither the unique numbers in the p leading nor the p trailing singular values of A, then X_s is a strict saddle point (because it has both an ascent and a descent direction).

On the other hand, if $\{\sigma_i\}_{i\in J}$ are the unique numbers in the p leading singular values of A, then all corresponding stationary points take the same objective value f_{det}, which must (globally) maximise the (continuous) objective f_{det} on the compact set $\text{St}(n,p)$. That is, every such stationary point X_s is in fact a global maximizer of program (6). Likewise, if $\{\sigma_i\}_{i\in J}$ are the unique numbers in the p trailing singular values of A, then all corresponding stationary points are global minimizers. This completes the proof of Theorem 2.1.

The advantage of Lemma 5.2 above is that it gives an explicit expression for the Hessian of f_ϕ, which will be used to prove Theorem 4.1. For the sake of completeness, however, let us show how Lemma 5.2 can be replaced with a simpler argument here, described next. In light of Lemma 5.1 and, if necessary, after a change of basis in (22), we can without loss of generality assume that a stationary point X_s of Program (6) is of the form

$$X_s = \begin{bmatrix} c_{i_1} & \cdots & c_{i_p} \end{bmatrix} \in \text{St}(n,p), \tag{42}$$

where $\sigma_{i_1} \geq \cdots \geq \sigma_{i_p}$, and $c_i \in \mathbb{R}^n$ is the ith canonical vector that takes one at index i and zero elsewhere. If X_s does not correspond to p leading singular values of A, then there exists $i_0 < i_1$ such that $\sigma_{i_0} > \min_{i_j} \sigma_{i_j}$. To simplify the presentation below, let us assume that in fact $\sigma_{i_0} > \sigma_{i_1}$. Now consider the trajectory $\theta \to X(\theta)$ specified as

$$X(\theta) = \begin{bmatrix} c_{i_0} & c_{i_1} & \cdots & c_{i_p} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta & & \\ -\sin\theta & \cos\theta & & \\ & & & \\ & & & \mathrm{I}_{p-1} \end{bmatrix} \begin{bmatrix} 0 \\ \mathrm{I}_p \end{bmatrix} \in \text{St}(n,p), \tag{43}$$

where the empty blocks in the square matrix above are filled with zeros. It is easy to verify that

$$\frac{d^2 f_{\text{det}}}{d\theta^2}(0) = \frac{2(\sigma_{i_0}^2 - \sigma_{i_1}^2)}{\sigma_{i_1}^2} > 0. \tag{44}$$

That is, there exists an ascent direction at any stationary point that does not correspond to p leading singular values of A. Likewise, one can verify that there exists a descent direction at any stationary point that does not correspond to p trailing singular values of A, and now the rest of the proof of Theorem 2.1 follows as before.

6. Proof of Theorem 4.1. The proof strategy is similar to that of Theorem 2.1 but with some technical subtleties. Without loss of generality, we assume again that $A^*A = \Gamma$ in (14), namely, we assume henceforth that

$$f_{s_q}(X) = \frac{s_q(X^*\Gamma X)}{s_q(X^*X)}. \tag{45}$$

As with Theorem 2.1, we will prove Theorem 4.1 by studying the stationary points of program (15), which we rewrite in the equivalent form

$$\max_{X\in\mathbb{R}^{n\times p}} \min_{\Lambda\in\mathbb{R}^{p\times p}} f_{s_q}(X) + \langle X^*X - I_p, \Lambda\rangle. \tag{46}$$

Therefore, $X_s \in \mathrm{St}(n,p)$ is a stationary point of programs (15) and (46) if and only if there exists $\Lambda_s \in \mathbb{R}^{p\times p}$ such that

$$\nabla f_{s_q}(X_s) + X_s(\Lambda_s + \Lambda_s^*) = 0, \tag{47}$$

namely, when $\nabla f_{s_q}(X_s)$ belongs to the normal space to the Stiefel manifold at X_s [11]. Without loss of generality, let us assume that $\Lambda_s = \Lambda_s^*$, so that the above condition simplifies to

$$\nabla f_{s_q}(X_s) + 2X_s\Lambda_s = 0. \tag{48}$$

In particular, since $X_s \in \mathrm{St}(n,p)$, we can multiply both sides above by X_s^* and solve for Λ_s above to obtain that

$$\Lambda_s = -\frac{1}{2}X_s^*\nabla f_{s_q}(X_s). \tag{49}$$

Next, from (45), it follows that

$$\nabla f_{s_q}(X_s) = \frac{2\Gamma X_s \nabla s_q(X_s^*\Gamma X_s)}{s_q(X_s^*X_s)} - \frac{2s_q(X_s^*\Gamma X_s)\cdot X_s\nabla s_q(X_s^*X_s)}{s_q(X_s^*X_s)^2}. \tag{50}$$

Let us examine the above expression more carefully. For $Z \in \mathbb{R}^{p\times p}$ with eigendecomposition $Z = U_Z \operatorname{diag}(\lambda_Z)U_Z^*$, note that the symmetric function corresponding to $\phi = s_q$ is

$$\psi(\lambda_Z) = e_q(\lambda_Z) := \sum_{1\le i_1<\cdots<i_q\le p}\ \prod_{j=1}^{q}\lambda_{Z,i_j}. \tag{51}$$

For a nonsingular matrix $Z \in \mathrm{GL}(p)$, it is then not difficult to verify that

$$\nabla e_q(\lambda_Z) = \begin{bmatrix} e_{q-1}(\lambda_Z^1) & \cdots & e_{q-1}(\lambda_Z^p) \end{bmatrix}^*, \tag{52}$$

where $\lambda_Z^i \in \mathbb{R}^{p-1}$ is formed from $\lambda_Z \in \mathbb{R}^p$ by removing its ith entry, namely, $\lambda_{Z,i}$. Using (27), we immediately find that

$$\nabla s_q(Z) = W_Z \operatorname{diag}\left(\begin{bmatrix} e_{q-1}(\lambda_Z^1) & \cdots & e_{q-1}(\lambda_Z^p) \end{bmatrix}^*\right) W_Z^*. \tag{53}$$

Recalling that $X_s \in \mathrm{St}(n,p)$ and using (53), we calculate the gradients involved in (50) as

$$\begin{aligned} \nabla s_q(X_s^*X_s) &= \binom{p-1}{q-1} X_s^*X_s \qquad \text{(see (51), (53))} \\ &= \binom{p-1}{q-1} I_p, \qquad (X_s \in \mathrm{St}(n,p)), \end{aligned} \tag{54}$$

$$\begin{aligned} \nabla s_q(X_s^*\Gamma X_s) &= X_s^* \operatorname{diag}\left(\begin{bmatrix} e_{q-1}(\widetilde{\gamma}_1^1) & \cdots & e_{q-1}(\widetilde{\gamma}_1^p) \end{bmatrix}^*\right) X_s \qquad \text{(see (33), (53))} \\ &= \operatorname{diag}\left(\begin{bmatrix} e_{q-1}(\widetilde{\gamma}_1^1) & \cdots & e_{q-1}(\widetilde{\gamma}_1^p) \end{bmatrix}^*\right) \qquad \text{(see Lemma 5.2)} \\ &=: \operatorname{diag}(\widehat{\gamma}_1), \end{aligned} \tag{55}$$

where $\widetilde{\gamma}_1^i \in \mathbb{R}^{p-1}$ is formed from $\widetilde{\gamma}_1$ by removing its ith entry; see (21). By substituting (54), (55) back into (50), we conclude that as in the proof of Theorem 2.1 $X_s \in \mathrm{St}(n,p)$ such that $X_s^*\Gamma X_s \in \mathrm{GL}(p)$ is a stationary point of Program (15) if and only if

$$\text{range}(X_s) = \text{range}(\Gamma X_s), \tag{56}$$

which is identical to (31) in the proof of Theorem 2.1, and consequently Lemmas 5.1 and 5.2 therein apply here too. Moreover, note that

$$\begin{aligned} s_q(X_s^* X_s) &= s_q(I_p) \qquad (X_s \in \text{St}(n,p)) \\ &= e_q(1_p) = \binom{p}{q} \qquad \text{(see (51))}, \end{aligned} \tag{57}$$

$$\begin{aligned} s_q(X_s^* \Gamma X_s) &= s_q(\text{diag}(\widetilde{\gamma}_1)) \qquad \text{(see Lemma 5.2)} \\ &= e_q(\widetilde{\gamma}_1) \qquad \text{(see (51))}. \end{aligned} \tag{58}$$

We can also revisit (49) to obtain that

$$\begin{aligned} \Lambda_s &= -\frac{1}{2} X_s^* \nabla f_{s_q}(X_s) \qquad \text{(see (49))} \\ &= -\frac{\text{diag}(\widetilde{\gamma}_1)\,\text{diag}(\widehat{\gamma}_1)}{\binom{p}{q}} + \frac{e_q(\widetilde{\gamma}_1)\binom{p-1}{q-1} I_p}{\binom{p}{q}^2} \qquad \text{(see (54), (55), (57), (58))} \\ &= -\frac{\binom{p-1}{q-1}}{\binom{p}{q}} \left(\frac{\text{diag}(\widetilde{\gamma}_1 \widehat{\gamma}_1)}{\binom{p-1}{q-1}} - \frac{e_q(\widetilde{\gamma}_1) I_p}{\binom{p}{q}} \right). \end{aligned} \tag{59}$$

In particular, when $\phi = s_q$ (and consequently $\psi = e_q$), we next simplify the expression for Hessian in (35) by noting that

$$\phi(\mathrm{I}_p) = s_q(\mathrm{I}_p) = e_q(1_p) = \binom{p}{q} \qquad \text{(see (51))},$$

$$\partial_j \psi(\widetilde{\gamma}_1) = \partial_j e_q(\widetilde{\gamma}_1) = e_{q-1}(\widetilde{\gamma}_1^j), \qquad j \in [p],$$

where $\widetilde{\gamma}_1^j \in \mathbb{R}^{p-1}$ is formed from $\widetilde{\gamma}_1 \in \mathbb{R}^p$ by removing its jth entry; see (34). Moreover,

$$\partial_1 \psi(1_p) = \partial_1 e_q(1_p) = e_{q-1}(1_{p-1}) = \binom{p-1}{q-1} \qquad \text{(see (51))},$$

$$\begin{aligned} f_{s_q}(X_s) &= \frac{s_q(X_s^* \Gamma X_s)}{s_q(X_s^* X_s)} \qquad \text{(see (32))} \\ &= \frac{s_q(\text{diag}(\widetilde{\gamma}_1))}{s_q(\mathrm{I}_p)} \qquad \text{((33) and } X_s \in \text{St}(n,p)) \\ &= \frac{e_q(\widetilde{\gamma}_1)}{e_q(1_p)} \\ &= \frac{e_q(\widetilde{\gamma}_1)}{\binom{p}{q}} \qquad \text{(see (51))}, \end{aligned} \tag{60}$$

$$\begin{aligned} \nabla^2 f_{s_q}(X_s)[\Delta,\Delta] &= \frac{1}{\phi(\mathrm{I}_p)} \sum_{i \in K^C} \sum_{j=1}^{p} \left(\sigma_i^2 \partial_j e_q(\widetilde{\gamma}_1) - f_{s_q}(X_s) \partial_1 e_q(1_p) \right) \Delta_{i,j}^2 \qquad \text{(see (35))} \\ &= \frac{1}{\binom{p}{q}} \sum_{i \in K^C} \sum_{j=1}^{p} \left(\sigma_i^2 e_{q-1}(\widetilde{\gamma}_1^j) - \frac{e_q(\widetilde{\gamma}_1)\binom{p-1}{q-1}}{\binom{p}{q}} \right) \Delta_{i,j}^2 \\ &= \frac{\binom{p-1}{q-1}}{\binom{p}{q}} \sum_{i \in K^C} \sum_{j=1}^{p} \left(\frac{\sigma_i^2 e_{q-1}(\widetilde{\gamma}_1^j)}{\binom{p-1}{q-1}} - \frac{e_q(\widetilde{\gamma}_1)}{\binom{p}{q}} \right) \Delta_{i,j}^2. \end{aligned} \tag{61}$$

In light of (46), (48), let us record for future reference that $\Delta \in \mathbb{R}^{n\times p}$ is an ascent direction at X_s if

$$\nabla^2 f_{s_q}(X_s)[\Delta,\Delta] + \langle \Delta^*\Delta, \Lambda_s\rangle \geq 0 \tag{62}$$

and

$$X_s^\top \Delta + \Delta^\top X_s = 0. \tag{63}$$

By definition in (34), $\{\sigma_i^2\}_{i\in J}$ are the distinct numbers appearing in $\widetilde{\gamma}_1 \in \mathbb{R}^p$. Suppose now that $\{\sigma_i\}_{i\in J}$ are *not* the unique numbers in the p leading singular values of A. Therefore there exist $i_0 \notin K$ and $j_0 \in [p]$ such that

$$\sigma_{i_0}^2 > \min_{i\in J} \sigma_i^2 = \min_{j\in[p]} \widetilde{\gamma}_{1,j} =: \widetilde{\gamma}_{1,j_0} \qquad \text{(see (34))}, \tag{64}$$

and,[3] moreover,

$$\frac{\sigma_{i_0}^2}{\widetilde{\gamma}_{1,j_0}} \geq \frac{\sigma_p^2}{\sigma_{p+1}^2}. \tag{65}$$

Let us set $\Delta \in \mathbb{R}^{n\times p}$ such that Δ_{i_0,j_0} is its only nonzero entry and note that (63) holds because $i_0 \notin K$. For this choice of Δ, we find that

$$\begin{aligned}
&\nabla^2 f_{s_q}(X_s)[\Delta,\Delta] + \langle \Delta^*\Delta, \Lambda_s\rangle \\
&\quad= \frac{\binom{p-1}{q-1}}{\binom{p}{q}} \left(\frac{\sigma_{i_0}^2 e_{q-1}(\widetilde{\gamma}_1^j)}{\binom{p-1}{q-1}} - \frac{e_q(\widetilde{\gamma}_1)}{\binom{p}{q}} \right) \Delta_{i_0,j_0}^2 + \Lambda_{s,j_0,j_0}^2 \Delta_{i_0,j_0}^2 \qquad \text{(see (61))} \\
&\quad= \frac{\binom{p-1}{q-1}}{\binom{p}{q}} \left(\frac{\sigma_{i_0}^2 e_{q-1}(\widetilde{\gamma}_1^{j_0})}{\binom{p-1}{q-1}} - \frac{e_q(\widetilde{\gamma}_1)}{\binom{p}{q}} \right) \Delta_{i_0,j_0}^2 \\
&\qquad - \frac{\binom{p-1}{q-1}}{\binom{p}{q}} \left(\frac{\widetilde{\gamma}_{1,j_0}\widehat{\gamma}_{1,j_0}}{\binom{p-1}{q-1}} - \frac{e_q(\widetilde{\gamma}_1)}{\binom{p}{q}} \right) \Delta_{i_0,j_0}^2 \qquad \text{(see (59))} \\
&\quad= \frac{\sigma_{i_0}^2 e_{q-1}(\widetilde{\gamma}_1^{j_0}) - \widetilde{\gamma}_{1,j_0}\widehat{\gamma}_{1,j_0}}{\binom{p}{q}} \Delta_{i_0,j_0}^2 \\
&\quad= \frac{(\sigma_{i_0}^2 - \widetilde{\gamma}_{1,j_0}) e_{q-1}(\widetilde{\gamma}_1^j)}{\binom{p}{q}} \Delta_{i_0,j_0}^2 \qquad \text{(see (55))} \\
&\quad> 0 \qquad \text{(see (64))}.
\end{aligned} \tag{66}$$

That is, if $\{\sigma_i\}_{i\in J}$ are *not* the unique members in the p leading singular values of A, then there exists an ascent direction at X_s. The rest of the proof of Theorem 4.1 is now the same as that of Theorem 2.1.

Appendix A. Proof of Lemma 5.1. Consider $X_s \in \mathrm{St}(n,p)$ such that $X_s^*\Gamma X_s \in \mathrm{GL}(p)$ and

$$\mathrm{range}(X_s) = \mathrm{range}(\Gamma X_s). \tag{67}$$

Each row of X_s naturally corresponds to a singular value of A, namely, the ith row corresponds to σ_i, where σ_i^2 is the ith diagonal entry of Γ. Let $J \subset [n]$ be the index

[3] In (64), the index j_0 might not be uniquely defined.

set such that $\Sigma_J := \{\sigma_i\}_{i\in J}$ is the set of *distinct* singular values corresponding to the nonzero rows of X_s.[4] For future reference, let us record that Σ_J contains only positive singular values, namely,

$$\Sigma_J \subset \mathbb{R}_+. \tag{68}$$

Indeed, if $0 \in \Sigma_J$, namely, if $\sigma_n = 0$, then the rows of X_s corresponding to σ_n are zero too thanks to (67) and consequently $0 \notin \Sigma_J$, which leads to a contradiction. Let us now set

$$\dim(\sigma_i) := \max\left[p - \sum_{j\in J,\, j\neq i} n_j, 1\right], \qquad i \in J, \tag{69}$$

for short, where n_i is the multiplicity of σ_i. Fix $i_0 \in J$. Consider $\mathcal{K}_{i_0}$, the collection of all index sets $K \subset [n]$ of size p such that

$$\sigma_{i_0} \in \text{unique}\left(\{\sigma_i\}_{i\in K}\right) \subseteq \Sigma_J, \tag{70}$$

where unique$(\{\sigma_i\}_{i\in K})$ returns the distinct members of the set $\{\sigma_i\}_{i\in K}$. In words, every index set $K \in \mathcal{K}_{i_0}$ contains σ_{i_0} and $p-1$ other (not necessarily distinct) singular values of A corresponding to nonzero rows of X_s. Consider an arbitrary $K \in \mathcal{K}_{i_0}$. It follows from (69) that

$$\{\sigma_i\}_{i\in K} \text{ contains at least } \dim(\sigma_{i_0}) \text{ copies of } \sigma_{i_0}. \tag{71}$$

On the other hand, (67) implies that there exists $B \in \mathbb{R}^{p\times p}$ such that

$$X_s B = \Gamma X_s. \tag{72}$$

By multiplying both sides above by X_s^* and using the fact that $X_s \in \mathrm{St}(n,p)$, we infer from (72) that

$$B = X_s^* \Gamma X_s \in \mathrm{GL}(p), \tag{73}$$

where the invertibility of B follows from the assumption of Lemma 5.1. In addition, (72) means that each row of X_s is an eigenvector of B. By restricting (72) to the index set K, we find that

$$X_s[K,:] \cdot B = \Gamma[K,K] \cdot X_s[K,:] \text{ is the eigendecomposition of } B, \tag{74}$$

because $K \in \mathcal{K}_{i_0}$ is a set of size p by the definition of $\mathcal{K}_{i_0}$ earlier. Above, we used the MATLAB matrix notation. For example, $X_s[K,:] \in \mathbb{R}^{p\times p}$ above is the row-submatrix of X_s corresponding to the rows indexed by K. It follows from (74) that

[4] Throughout, we treat $\{\sigma_i\}_{i\in J}$ and similar items as sequences (rather than sets) to allow for repetitions.

(75) σ_{i_0} is an eigenvalue of B with the multiplicity of at least $\dim(\sigma_{i_0})$,

because, by (71), $\{\sigma_i\}_{i\in K}$ contains at least $\dim(\sigma_{i_0})$ copies of σ_{i_0}. In fact, there exists an index set $K_0 \in \mathcal{K}_{i_0}$ such that $\{\sigma_i\}_{i\in K_0}$ contains *exactly* $\dim(\sigma_{i_0})$ copies of σ_{i_0}.[5] It follows that

(76) σ_{i_0} is an eigenvalue of B with the multiplicity of exactly $\dim(\sigma_{i_0})$.

By (73), B is full-rank and it follows from (76) that the corresponding eigenvectors of B span a $\dim(\sigma_{i_0})$-dimensional subspace of $\mathbb{R}^p$, namely, the geometric multiplicity of σ_{i_0} is $\dim(\sigma_{i_0})$. Since every row of X_s is an eigenvectors of B by (74), it follows that the

(77) rows of X_s that correspond to σ_{i_0} span a $\dim(\sigma_{i_0})$-dimensional subspace of $\mathbb{R}^p$.

Since the choice of $i_0 \in J$ was arbitrary above, we find for every $i \in J$ that

(78) the rows of X_s corresponding to σ_i span a $\dim(\sigma_i)$-dimensional subspace of $\mathbb{R}^p$, denoted by $S_i \in \mathrm{Gr}(p, \dim(\sigma_i))$.

Because B is symmetric by its definition in (73), these subspaces are orthogonal to one another, namely,

(79) $$S_i \perp S_j, \qquad i \neq j \text{ and } i, j \in J.$$

On the other hand, note that

$$\begin{aligned} B &= X_s^* \Sigma X_s \qquad \text{(see (72))} \\ &= \sum_{i\in J} \sigma_i \left(\sum_{\sigma_j = \sigma_i} X_s\,[j,:]^*\, X_s[j,:] \right) + \sum_{\sigma_j \notin \Sigma_J} \sigma_j \cdot X_s[j,:]^* X_s[j,:] \\ &=: \sum_{i\in J} \sigma_i Y_i + \sum_{\sigma_j \notin \Sigma_J} \sigma_j \cdot X_s[j,:]^* X_s[j,:] \\ &= \sum_{i\in J} \sigma_i Y_i, \end{aligned}$$

where the last line above follows from the definition of J, namely, any singular value $\sigma_i \notin \Sigma_J$ corresponds to a zero row of X_s. For every $i \in J$, note that

(80) $$\mathrm{range}(Y_i) = S_i$$

by definition of S_i in (78). It therefore follows from (79) that $\{Y_i\}_{i\in J}$ are pairwise orthogonal matrices, namely,

(81) $$Y_i^* Y_j = 0, \qquad i \neq j \text{ and } i, j \in J.$$

Therefore, $B = \sum_{i\in J} \sigma_i Y_i$ is the eigendecomposition of B and, because B is full-rank by (73), we find that

[5] Indeed, if $\dim(\sigma_{i_0}) > 1$, such an index set K_0 would include n_i copies of singular value σ_i for every $\sigma_i \neq \sigma_{i_0}$ with $i \in K$. The construction is similar if $\dim(\sigma_{i_0}) = 1$.

$$
\begin{aligned}
p &= \sum_{i\in J,\,\sigma_i\neq 0} \dim(\operatorname{span}(Y_i)) \\
&= \sum_{i\in J,\,\sigma_i\neq 0} \dim(S_i) \qquad \text{(see (80))} \\
&= \sum_{i\in J,\,\sigma_i\neq 0} \dim(\sigma_i) \qquad \text{(see (78))} \\
&= \sum_{i\in J} \dim(\sigma_i) \qquad \text{(see (68)).} \qquad (82)
\end{aligned}
$$

This completes the proof of Lemma 5.1.

Appendix B. Proof of Lemma 5.2. Suppose that $X_s \in \mathrm{St}(n,p)$ satisfies

$$\operatorname{range}(X_s) = \operatorname{range}(\Gamma X_s), \tag{83}$$

which implies that

$$X_s B = \Gamma X_s, \qquad \text{where } B = X_s^* \Gamma X_s. \tag{84}$$

Let $J \subseteq [n]$ be the corresponding index set prescribed in Lemma 5.1, and recall that $\{\sigma_i\}_{i\in J}$ are distinct by item 1 in Lemma 5.1. Consider an index set $K \supseteq J$ such that $\{\sigma_i\}_{i\in K}$ contains all available copies of the singular values listed in $\{\sigma_i\}_{i\in J}$. Let k denote the size of K. For convenience, we define

$$X_{s,1} := X_s\,[K,:] \in \mathbb{R}^{k\times p}, \qquad X_{s,2} := X_s\left[K^C,:\right] \in \mathbb{R}^{(n-k)\times p},$$

$$\Gamma_1 := \Gamma\,[K,K] \in \mathbb{R}^{k\times k}, \qquad \Gamma_2 := \Gamma\left[K^C,K^C\right] \in \mathbb{R}^{(n-k)\times(n-k)}, \tag{85}$$

where we used the MATLAB matrix notation above. For example, $X_s[K,:]$ is the restriction of X_s to the rows indexed in K. Also, K^C is the complement of index set K with respect to $[n]$. In particular, item 2 in Lemma 5.1 immediately implies that

$$X_{s,2} = 0, \tag{86}$$

and, consequently,

$$
\begin{aligned}
X_{s,1}^* X_{s,1} &= X_{s,1}^* X_{s,1} + X_{s,2}^* X_{s,2} \\
&= X_s^* X_s \\
&= \mathrm{I}_p \qquad (X_s \in \mathrm{St}(n,p))\,. \qquad (87)
\end{aligned}
$$

That is,

$$X_{s,1} \in \mathrm{St}(k,p). \tag{88}$$

Note that (83) holds also after a change of basis from X_s to $X_s\Theta$ for invertible $\Theta \in \mathrm{GL}(p)$. Therefore, thanks to (83) and item 4 in Lemma 5.1, we can assume without loss of generality that the supports of rows and also columns of $X_{s,1}$ are disjoint. More specifically, with the enumeration $J = \{i_1, i_2, \ldots\}$, we assume without loss of generality that

$$X_{s,1} = \begin{bmatrix} X_{s,1,1} & 0 & \cdots \\ 0 & X_{s,1,2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \in \mathbb{R}^{k \times p}, \tag{89}$$

where the rows of the block $X_{s,1,1}$ correspond to the singular value σ_{i_1} and has $\dim(\sigma_{i_1})$ columns, the block $X_{s,1,2}$ corresponds to σ_{i_2} and so on. In particular, (88) implies that

$$X_{s,1,1}^* X_{s,1,1} = \mathrm{I}_{\dim(\sigma_{i_1})}, \qquad X_{s,1,2}^* X_{s,1,2} = \mathrm{I}_{\dim(\sigma_{i_2})}, \qquad \cdots, \tag{90}$$

namely, $X_{s,1,1}$ has orthonormal columns, and so do $X_{s,1,2}$ and the rest of the diagonal blocks of $X_{s,1}$. Another necessary ingredient in our analysis below is the observation that

$$\begin{aligned} X_s^* \Gamma X_s &= X_{s,1}^* \Gamma_1 X_{s,1} + X_{s,2}^* \Gamma_2 X_{s,2} \\ &= X_{s,1}^* \Gamma_1 X_{s,1} \qquad \text{(see (86))} \\ &= \begin{bmatrix} \sigma_{i_1}^2 \cdot X_{s,1,1}^* X_{s,1,1} & 0 & \cdots \\ 0 & \sigma_{i_2}^2 \cdot X_{s,1,2}^* X_{s,1,2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{i_1}^2 \cdot \mathrm{I}_{\dim(\sigma_{i_1})} & 0 & \cdots \\ 0 & \sigma_{i_2}^2 \cdot \mathrm{I}_{\dim(\sigma_{i_1})} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \qquad \text{(see (90))} \\ &=: \widetilde{\Gamma}_1 \in \mathbb{R}^{p \times p} \\ &=: \mathrm{diag}(\widetilde{\gamma}_1), \end{aligned} \tag{91}$$

namely, the diagonal matrix $\widetilde{\Gamma}_1$ contains $\dim(\sigma_{i_1})$ copies of $\sigma_{i_1}^2$, $\dim(\sigma_{i_2})$ copies of $\sigma_{i_2}^2$, and so on. To compute the Hessian of f_ϕ, we make a small perturbation to its argument. To be specific, consider $\Delta \in \mathbb{R}^{n \times p}$ that is supported only on the rows indexed by K^C and let

$$\Delta_2 := \Delta \left[K^C, : \right] \in \mathbb{R}^{(n-k) \times p} \tag{92}$$

be the nonzero block of Δ. Note in particular that

$$X_s^* \Delta = 0, \tag{93}$$

because by construction X_s and Δ are supported on the rows indexed by K and K^C, respectively; see (86). Let $h_\Gamma(X) = \phi(X^* \Gamma X)$ for short and note that

$$\begin{aligned} h_\Gamma(X_s + \Delta) &= \phi((X_s + \Delta)^* \Gamma (X_s + \Delta)) \\ &= \phi(X_s^* \Gamma X_s + X_s^* \Gamma \Delta + \Delta^* \Gamma X_s + \Delta^* \Gamma \Delta) \\ &= \phi(X_s^* \Gamma X_s + B X_s^* \Delta + \Delta^* X_s B + \Delta^* \Gamma \Delta) \qquad \text{(see (84))} \\ &= \phi(X_s^* \Gamma X_s + \Delta^* \Gamma \Delta) \qquad \text{(see (93))} \\ &= \phi(\widetilde{\Gamma}_1 + \Delta_2^* \Gamma_2 \Delta_2) \qquad \text{(see (91), (92), (85))} \\ &= \phi(\widetilde{\Gamma}_1) + \langle \nabla \phi(\widetilde{\Gamma}_1), \Delta_2^* \Gamma_2 \Delta_2 \rangle + O(\|\Delta_2\|^3) \qquad \text{(Taylor expansion)} \\ &= \phi(X_s^* \Gamma X_s) + \langle \nabla \phi(\widetilde{\Gamma}_1), \Delta_2^* \Gamma_2 \Delta_2 \rangle + O(\|\Delta\|^3) \qquad \text{(see (91), (92))}, \end{aligned} \tag{94}$$

where we used the standard Big-O notation above. Recall from (91) that $\widetilde{\Gamma}_1 = \mathrm{diag}(\widetilde{\gamma}_1)$. Because ϕ is by assumption a spectral function with the corresponding symmetric function ψ, (27) implies that

$$\nabla\phi(\widetilde{\Gamma}_1) = \mathrm{diag}(\nabla\psi(\widetilde{\gamma}_1)) = \mathrm{diag}\left(\begin{bmatrix} \partial_1\psi(\widetilde{\gamma}_{1,1}) & \cdots \partial_p\psi(\widetilde{\gamma}_{1,p}) \end{bmatrix}^*\right), \tag{95}$$

which allows us to rewrite the last line above as

$$\begin{aligned} h_\Gamma(X_s+\Delta) &= \phi(X_s^*\Gamma X_s) + \langle \mathrm{diag}(\nabla\psi(\widetilde{\gamma}_1)), \Delta_2^*\Gamma_2\Delta_2\rangle + O(\|\Delta\|^3) \\ &= \phi(X_s^*\Gamma X_s) + \sum_{i\in K^C}\sum_{j=1}^{p} \sigma_i^2\cdot\partial_j\psi(\widetilde{\gamma}_1)\cdot\Delta[i,j]^2 + O(\|\Delta\|^3), \end{aligned} \tag{96}$$

where $\Delta[i,j]$ is the $[i,j]$th entry of Δ. Let $1_p \in \mathbb{R}^p$ be the vector of all ones. After setting $h_{\mathrm{I}}(X) = \phi(X^*X)$ and after replacing Γ with I_n above, we find that

$$\begin{aligned} h_{\mathrm{I}}(X_s+\Delta) &= \phi(X_s^*X_s) + \sum_{i\in K^C}\sum_{j=1}^{p} \partial_j\psi(1_p)\cdot\Delta[i,j]^2 + O(\|\Delta\|^3) \\ &= \phi(\mathrm{I}_p) + \partial_1\psi(1_p)\sum_{i\in K^C}\sum_{j=1}^{p} \Delta[i,j]^2 + O(\|\Delta\|^3), \end{aligned} \tag{97}$$

where in the last line above we used the fact that $X_s \in \mathrm{St}(n,p)$ and that ψ is a symmetric function, hence $\partial_j\psi(1_p) = \partial_1\psi(1_p)$ for every $j \in [p]$. Since $f_\phi = h_\Gamma/h_{\mathrm{I}}$ by definition, (96), (97) imply that

(98)

$$\begin{aligned} f_\phi(X_s+\Delta) &= \frac{h_\Gamma(X_s+\Delta)}{h_{\mathrm{I}}(X_s+\Delta)} \\ &= \frac{\phi(X_s^*\Gamma X_s) + \sum_{i\in K^C}\sum_{j=1}^{p}\sigma_i^2\cdot\partial_j\psi(\widetilde{\gamma}_1)\cdot\Delta[i,j]^2 + O(\|\Delta\|^3)}{\phi(\mathrm{I}_p) + \partial_1\psi(1_p)\sum_{i\in K^C}\sum_{j=1}^{p}\Delta_{i,j}^2 + O(\|\Delta\|^3)} \qquad \text{(see (96), (97))} \\ &= \left(\phi(X_s^*\Gamma X_s) + \sum_{i\in K^C}\sum_{j=1}^{p}\sigma_i^2\cdot\partial_j\psi(\widetilde{\gamma}_1)\cdot\Delta[i,j]^2 + O(\|\Delta\|^3)\right) \\ &\quad\cdot\frac{1}{\phi(\mathrm{I}_p)}\left(1 - \frac{\partial_1\psi(1_p)}{\phi(\mathrm{I}_p)}\sum_{i\in K^C}\sum_{j=1}^{p}\Delta[i,j]^2 + O(\|\Delta\|^3\right) \quad \left(\frac{1}{1+a} = 1-a+O(a^2)\right) \\ &= \frac{\phi(X_s^*\Gamma X_s)}{\phi(\mathrm{I}_p)} - \frac{\phi(X_s^*\Gamma X_s)}{\phi(\mathrm{I}_p)}\cdot\frac{\partial_1\psi(1_p)}{\phi(\mathrm{I}_p)}\sum_{i\in K^C}\sum_{j=1}^{p}\Delta[i,j]^2 \\ &\quad + \frac{1}{\phi(\mathrm{I}_p)}\sum_{i\in K^C}\sum_{j=1}^{p}\sigma_i^2\cdot\partial_j\psi(\widetilde{\gamma}_1)\cdot\Delta[i,j]^2 + O(\|\Delta\|^3) \\ &= f_\phi(X_s) - f_\phi(X_s)\frac{\partial_1\psi(1_p)}{\phi(\mathrm{I}_p)}\sum_{i\in K^C}\sum_{j=1}^{p}\Delta[i,j]^2 \\ &\quad + \frac{1}{\phi(\mathrm{I}_p)}\sum_{i\in K^C}\sum_{j=1}^{p}\sigma_i^2\cdot\partial_j\psi(\widetilde{\gamma}_1)\cdot\Delta[i,j]^2 + O(\|\Delta\|^3) \qquad \text{(see (32))}, \end{aligned}$$

and, consequently,

$$\nabla^2 f_\phi(X_s)[\Delta,\Delta] = \frac{1}{\phi(\mathrm{I}_p)} \sum_{i\in K^C} \sum_{j=1}^{p} \left(\sigma_i^2 \partial_j \psi(\widetilde{\gamma}_1) - f_\phi(X_s)\partial_1\psi(1_p)\right) \Delta_{i,j}^2, \tag{99}$$

for our particular choice of Δ that satisfies $\Delta[K,:] = 0$. Here, the bilinear operator $\nabla^2 f_\phi(X_s) : \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p} \to \mathbb{R}$ is the Hessian of f_ϕ at X_s. This completes the proof of Lemma 5.2.

Acknowledgments. AE would like to thank Stephen Becker and David Bortz for pointing out the connection to D-optimality in optimal design and Mike Davis for the connection to independent component analysis.

REFERENCES

[1] O. Alter, P. O. Brown, and D. Botstein, *Singular value decomposition for genome-wide expression data processing and modeling*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 10101–10106.

[2] O. Alter and G. Golub, *Singular value decomposition of genome-scale MRNA lengths distribution reveals asymmetry in RNA gel electrophoresis band broadening*, Proc. Natl. Acad. Sci. USA, 103 (2006), pp. 11828–11833, https://doi.org/10.1073/pnas.0604756103.

[3] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, *Dropping convexity for faster semi-definite optimization*, in Proceedings of the Conference on Learning Theory, 2016, pp. 530–582.

[4] S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Global optimality of local search for low rank matrix recovery*, in Proceedings of Advances in Neural Information Processing Systems, 2016, pp. 3873–3881.

[5] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer Ser. Statist., Springer New York, 2013, https://books.google.co.uk/books?id=NYDSBwAAQBAJ.

[6] N. Boumal, V. Voroninski, and A. Bandeira, *The non-convex Burer-Monteiro approach works on smooth semidefinite programs*, in Proceedings of Advances in Neural Information Processing Systems, 2016, pp. 2757–2765.

[7] S. Burer and R. D. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Math. Program., 95 (2003), pp. 329–357.

[8] A. d'Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming*, in Proceedings of Advances in Neural Information Processing Systems, 2005, pp. 41–48.

[9] Y. Deshpande and A. Montanari, *Information-theoretically optimal sparse pca*, in Proceedings of the IEEE International Symposium on Information Theory, IEEE, 2014, pp. 2197–2201.

[10] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218, https://doi.org/10.1007/BF02288367.

[11] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.

[12] A. Eftekhari, G. Ongie, L. Balzano, and M. B. Wakin, *Streaming principal component analysis from incomplete data*, J. Mach. Learn. Res., 20 (2019), pp. 1–62.

[13] R. Ge, F. Huang, C. Jin, and Y. Yuan, *Escaping from saddle points—online stochastic gradient for tensor decomposition*, in Proceedings of the Conference on Learning Theory, 2015, pp. 797–842.

[14] R. Ge, J. D. Lee, and T. Ma, *Matrix completion has no spurious local minimum*, in Proceedings of Advances in Neural Information Processing Systems, 2016, pp. 2973–2981.

[15] R. Ge, J. D. Lee, and T. Ma, *Learning One-Hidden-Layer Neural Networks with Landscape Design*, https://arxiv.org/abs/1711.00501, 2017.

[16] R. Ge and T. Ma, *On the optimization landscape of tensor decompositions*, in Proceedings of Advances in Neural Information Processing Systems, 2017, pp. 3653–3663.

[17] N. Gillis, *The why and how of nonnegative matrix factorization*, in Regularization, Optimization, Kernels, and Support Vector Machines, J. A. K. Suykens, M. Signoretto, and A. Argyriou, eds., Chapman & Hall/CRC Press, Boca Raton, FL, 2015, pp. 257–292.

[18] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, 1996, https://books.google.ch/books?id=mlOa7wPX6OYC.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Ser. Statist., Springer New York, 2013, https://books.google.co.uk/books?id=yPfZBwAAQBAJ.

[20] R. A. Hauser, A. Eftekhari, and H. F. Matzinger, *Pca by determinant optimization has no spurious local optima*, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 1504–1511.

[21] R. Horn, R. Horn, and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990, https://books.google.co.uk/books?id=PlYQN0ypTwEC.

[22] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1994, https://books.google.co.uk/books?id=ukd0AgAAQBAJ.

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

[24] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley Ser. Adaptive Learning Syst. Signal Process. Learning Commun. Control, Wiley, 2004, New York, https://books.google.co.uk/books?id=96D0ypDwAkkC.

[25] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, *How to escape saddle points efficiently*, in J. Mach. Learn. Res., 70 (2017), pp. 1724–1732.

[26] C. Jin, S. M. Kakade, and P. Netrapalli, *Provable efficient online matrix completion via non-convex stochastic gradient descent*, in Proceedings of Advances in Neural Information Processing Systems, 2016, pp. 4520–4528.

[27] I. M. Johnstone and A. Y. Lu, *On consistency and sparsity for principal components analysis in high dimensions*, J. Amer. Statist. Assoc., 104 (2009), pp. 682–693.

[28] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, *Generalized power method for sparse principal component analysis*, J. Mach. Learn. Res., 11 (2010), pp. 517–553.

[29] S. Karlin and Y. Rinott, *A generalized Cauchy Binet formula and applications to total positivity and majorization*, J. Multivariate Anal., 27 (1988), pp. 284–299.

[30] A. S. Lewis and H. S. Sendov, *Quadratic expansions of spectral functions*, Linear Algebra Appl., 340 (2002), pp. 97–121.

[31] Q. Li and G. Tang, *The Nonconvex Geometry of Low-Rank Matrix Optimizations with General Objective Functions*, https://arxiv.org/abs/1611.03060v1, 2016.

[32] L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Quart. J. Math. Oxford, (1966), pp. 1156–1159.

[33] A. Mokhtari, A. Ozdaglar, and A. Jadbabaie, *Escaping saddle points in constrained optimization*, in Proceedings of Advances in Neural Information Processing Systems, 2018, pp. 3633–3643.

[34] *NIST/SEMATECH Engineering Statistics Handbook*, https://books.google.co.uk/books?id=v-XXjwEACAAJ, 2002.

[35] K. Pearson, *On lines and planes of closest fit to systems of points in space*, Philos. Mag., 2 (1901), pp. 559–572, https://doi.org/10.1080/14786440109462720.

[36] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2002, https://books.google.co.uk/books?id=y8ORL3DWt4sC.

[37] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, UK, 2004, https://books.google.co.uk/books?id=MX0hAwAAQBAJ.

[38] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, *Theoretical Insights into the Optimization Landscape of Over-Parameterized Shallow Neural Networks*, https://arxiv.org/abs/1707.04926, 2017.

[39] J. Sun, Q. Qu, and J. Wright, *When Are Nonconvex Problems Not Scary?*, https://arxiv.org/abs/1510.06096, 2015.