The generalized orthogonal Procrustes problem in the high noise regime

Thomas Pumir* Amit Singer† Nicolas Boumal‡

July 3, 2019

Abstract

We consider the problem of estimating a cloud of points from numerous noisy observations of that cloud after unknown rotations, and possibly reflections. This is an instance of the general problem of estimation under group action, originally inspired by applications in 3-D imaging and computer vision. We focus on a regime where the noise level is larger than the magnitude of the signal, so much so that the rotations cannot be estimated reliably. We propose a simple and efficient procedure based on invariant polynomials (effectively: the Gram matrices) to recover the signal, and we assess it against fundamental limits of the problem that we derive. We show our approach adapts to the noise level and is statistically optimal (up to constants) for both the low and high noise regimes. In studying the variance of our estimator, we encounter the question of the sensivity of a type of thin Cholesky factorization, for which we provide an improved bound which may be of independent interest.

1 Introduction

We consider the problem of estimating k labeled points in $\mathbb{R}^d$, with $k \geq d$. This cloud of points, which we call the parameter, is represented as a matrix X of size $d \times k$. We restrict ourselves to the case where the smallest singular value of X is bounded away from zero, that is, the cloud spans all d dimensions. We observe N independent measurements $Y_1, \ldots, Y_N$ of X, following the model

$$Y_i = Q_i X + \sigma E_i, \tag{1}$$

where $\sigma > 0$ is the standard deviation of the noise, $E_1, \ldots, E_N$ are independent noise matrices in $\mathbb{R}^{d\times k}$ with independent, standard Gaussian entries, and $Q_1, \ldots, Q_N$ are drawn uniformly and independently at random from the orthogonal group,

$$\mathcal{O}(d) = \{Q \in \mathbb{R}^{d\times d} : Q^T Q = I_d\}, \tag{2}$$

*Department of Operations Research and Financial Engineering, Princeton University
†Mathematics Department and PACM, Princeton University
‡Mathematics Department, Princeton University

where I_d is the identity matrix of size d. In what follows, we refer to orthogonal matrices as rotations, bearing in mind that (in our meaning) they may also include a reflection. The method we propose in Section 2.2 applies under relaxed assumptions on the distributions of Q_i and E_i: our assumptions here serve to streamline exposition. This problem we investigate belongs to a larger class of estimation problems under group actions [Bandeira et al., 2017b].

Notice that the distribution of the observations Y_i is unchanged if X is replaced by QX, for any orthogonal Q. As a result, we can only hope to recover the cloud X up to a global rotation. Accordingly, we define an equivalence relation $\sim$ over $\mathbb{R}^{d\times k}$:

$$X_1 \sim X_2 \iff X_1 = QX_2 \text{ for some } Q \in \mathcal{O}(d). \tag{3}$$

This equivalence relation partitions the parameter space into equivalence classes

$$[X] = \{QX : Q \in \mathcal{O}(d)\}.$$

The set of equivalence classes is the quotient space $\mathbb{R}^{d\times k}/\sim$. The distribution of the measurements Y_i is parameterized by $[X]$, which we aim to estimate.

A natural approach to estimate $[X]$ would be to estimate the rotations Q_i first (seen as latent or nuisance variables), align the observations Y_i using the estimated rotations, and average. Typical of those approaches, the Maximum Likelihood Estimator (MLE) is a solution of the following non-convex optimization problem

$$\min_{\hat{X}\in\mathbb{R}^{d\times k},\hat{Q}_1,\ldots,\hat{Q}_N\in\mathcal{O}(d)} \sum_{i=1}^{N} \|\hat{Q}_i^T Y_i - \hat{X}\|_{\mathrm{F}}^2. \tag{4}$$

For any fixed choice of estimators $\hat{Q}_1, \ldots, \hat{Q}_N$, the optimal estimator for $\hat{X}$ according to the above is

$$\hat{X} = \frac{1}{N}\sum_{i=1}^{N} \hat{Q}_i^T Y_i. \tag{5}$$

This estimator can be plugged into the cost function of (4), reducing the problem to that of estimating only the rotations, upon which $\hat{X}$ can be deduced from (5). A number of papers focus on the resulting problem, called synchronization of rotations [Singer, 2011].

Such approaches, however, necessarily fail at low signal-to-noise ratio (SNR). Specifically, we argue that if the noise level is too large, no procedure can reliably determine the latent rotations Q_i. Essentially, this is because, when σ is too large, the distribution of $QX + \sigma E$ is indistinguishable from that of $Q'X + \sigma E$, where Q and Q' are two rotations.

To see this, consider the following strictly simpler problem: we observe $Y = sQX + \sigma E$, where $Q \in \mathcal{O}(d)$, $X \in \mathbb{R}^{d\times k}$ and $\sigma > 0$ are *known*, while $E \in \mathbb{R}^{d\times k}$ has independent standard Gaussian entries and s is uniformly sampled from $\{+1, -1\}$, both unknown. An estimator ψ (deterministic) assigns an estimate of s to an observed Y. Estimating s in this context is strictly simpler than estimating Q_i's to any reasonable accuracy in our model (where furthermore X is unknown), as we are only asked to determine whether Q_i is some given rotation, or its opposite. Yet, even this simpler problem is hopeless at low SNR, directly implying the impossibility of estimating the rotations in problem (4):

Proposition 1.1. *For any tolerance $\tau \in (0, 1/2)$, there exists a critical noise level σ_0 such that, for any estimator ψ, if $\sigma > \sigma_0$, then the probability of error $\mathbb{P}(\psi(Y) \neq s)$ exceeds τ.*

(The proof relies on the optimality of the likelihood ratio test for Gaussian distributions, see Appendix B.)

Because of this fundamental obstruction, in this paper, we aim to estimate $[X]$ directly from observations Y_i, bypassing any estimation (even implicit) of the latent Q_i's. To do so, we follow a trend in signal processing that consists in estimating $[X]$ from features of the observations that are *invariant* under the group action of $\mathcal{O}(d)$ [Tukey, 1984], [Sadler and Giannakis, 1992], [Giannakis, 1989], [Abbe et al., 2018a], [Bandeira et al., 2017b], [Perry et al., 2017], [Boumal et al., 2017].

Specifically, consider the Gram matrix of X, that is, X^TX: it is invariant under orthogonal transformations since $(QX)^T(QX) = X^TX$ for any $Q \in \mathcal{O}(d)$. Thus, up to noise terms that we will handle, the Gram matrices of the observations Y_i reveal information about the Gram matrix of X without the need to estimate the Q_i's. In Section 2, we call upon invariant theory to argue that no other polynomial invariant features are necessary, in the sense that (a) they would be redundant with the Gram matrix, and (b) the Gram matrix is sufficient to fully characterize the equivalence class $[X]$.

Based on these observations, we proceed (still in Section 2) to derive an estimator for the Gram matrix of X from the given observations (1)—this also requires estimating the noise level σ, which we discuss. From the estimated Gram matrix, we construct an estimator for the equivalence class $[X]$. Since we only have access to a finite number N of observations, we can only hope to recover an approximation of the Gram matrix. Accordingly, in Section 3 we study the sensitivity of the mapping from the Gram matrix to the sought equivalence class $[X]$. This reduces to showing stability of the factorization of a rank-d positive semidefinite matrix. To address this question, we propose a new analysis of such matrix factorization, with a geometric proof.

In Section 4, we show that the proposed approach is statistically optimal, that is, it makes the best use of available samples. Moreover, we show that the mean squared error (MSE) of our estimator behaves like $O\left(\sigma^2/N + \sigma^4/N\right)$, which is shown to be adaptively optimal. Indeed, this highlights the existence of two regimes: at low SNR, our estimator's MSE behaves like $O(\sigma^4/N)$, while at high SNR we obtain an MSE of order $O(\sigma^2/N)$ which matches the regime we would get if the rotations were known. In particular, the MSE can be driven to zero at *any* noise level, provided the number of observations N is sufficiently large. We give an explicit characterization of those two regimes and we support our claim with numerical simulations in Section 5.

To make sense of MSE in estimating $[X]$, we need a distance on the quotient space. A natural choice consists in computing the Frobenius distance between two aligned representatives, that is,

$$\rho([X_1], [X_2]) = \min_{Q \in \mathcal{O}(d)} \|X_1 - QX_2\|_{\mathrm{F}}, \tag{6}$$

where $\|X\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(X^T X)}$. We sometimes write simply $\rho(X_1, X_2)$, where it is clear that we mean the distance between $[X_1]$ and $[X_2]$. Computing this distance is known as the orthogonal Procrustes problem: it can be done efficiently via singular value decomposition (SVD) [Schönemann, 1966] (see also Appendix A).

Related work

Procrustes problems consist in finding correspondences between shapes that have been transformed through translation, rotation or dilation [Schönemann, 1966, Ten Berge, 1977]. They notably find application in multivariate analysis [Hurley and Cattell, 1962, Green and Carroll, 1976], multidimensional scaling [Borg and Groenen, 2005], computer vision [Zhang, 2000] and natural language processing [Xing et al., 2015, Smith and Hammerla, 2017, Grave et al., 2018]. An extensive survey on applications of Procrustes problem can be found in [Gower and Dijksterhuis, 2005].

A particular line of work on such problems has focused on estimating each rotation before estimating the orbit (that is, equivalence class) of the matrix X. In particular, the MLE (4) is cast into the following, non-convex, optimization problem

$$\min_{\hat{Q}_1,\ldots,\hat{Q}_N \in \mathcal{O}(d)} \sum_{i \neq j} \|\hat{Q}_i Y_i - \hat{Q}_j Y_j\|_{\mathrm{F}}^2, \tag{7}$$

or equivalently

$$\max_{\hat{Q}_1,\ldots,\hat{Q}_N \in \mathcal{O}(d)} \sum_{i \neq j} \langle \hat{Q}_i Y_i, \hat{Q}_j Y_j \rangle, \tag{8}$$

where $\langle A, B \rangle = \mathrm{Tr}(A^T B)$ is the inner product we use throughout. Among many, this problem was investigated by Nemirovski [2007] and Man-Cho So [2010]. A particularization of this problem amounts to the little Grothendieck problem [Khot and Naor, 2012, Naor et al., 2013]. A natural way to deal with the non-convexity and ensuing computational complexity is to study a convex relaxation of (8). Following the line initiated by the seminal work of Goemans and Williamson [1995], Bandeira et al. [2016] and many others investigated such semidefinite relaxation. In contrast, we focus on a noise regime where Q_i's cannot be estimated, hence these approaches cannot succeed.

The problem we consider finds its original motivation in cryo-electron microscopy (cryo-EM), an imaging technique used in structural biology to estimate the 3D shape of a molecule from 2D projections of that molecule under random and unknown orientations—see [Singer, 2018] for a review of mathematical aspects of this task. Our setting and approach is also closely connected to the multi-reference alignment (MRA) problem and recent literature on the topic, where noisy realizations of a randomly, cyclically shifted version of a vector are observed [Bandeira et al., 2014].

Close to the approach taken in the present work, Giannakis [1989], Sadler and Giannakis [1992] and more recently Bendory et al. [2018] and Abbe et al. [2018a] among others, consider estimating moments of the signal that are invariant under cyclic shifts. The advantage of such formulation is its validity for any SNR regime, provided sufficiently many sample are available. In particular, such approach has been shown to be

optimal both in terms of rate of estimation [Bandeira et al., 2017b] and sample complexity [Perry et al., 2017].

2 Invariant features approach

We focus on estimating $[X]$ from samples drawn according to (1), disregarding latent variables $Q_1, \ldots, Q_N$. Based on the discussion above, it is apparent that any method for estimating $[X]$ which (explicitly on implicitly) relies on estimating the rotations reliably must fail beyond a certain noise level. Hence, we take a different approach. For each observation Y_i, we compute polynomial functions of Y_i that, aside from the noise E_i, are invariant under the rotation Q_i. Such functions are called *invariant features*. Then, our estimation problem reduces to that of estimating $[X]$ from the estimated invariant features. This approach completely bypasses estimation of the latent variables. We first explain the method with some background below, before showing that it does not break down at high noise levels.

2.1 Invariant polynomials

Invariant theory is concerned with polynomials that are invariant under some group action. In our case, this specializes to the following central definition.

Definition 2.1. *A multivariate polynomial p is said to be invariant under the action of the orthogonal group if, for all $X \in \mathbb{R}^{d\times k}$ and for all $Q \in \mathcal{O}(d)$,*

$$p(QX) = p(X).$$

The goal of this section is to identify all invariant polynomials in our specific setting. A standard observation in invariant theory is that it is sufficient to consider *homogeneous* invariant polynomials.

Proposition 2.2. *Consider a multivariate polynomial p of degree r, decomposed into a sum of homogeneous parts $p_1, \ldots, p_r$ where p_i has degree i:*

$$p(X) = p_0 + p_1(X) + \cdots + p_r(X).$$

If p is invariant under the action of the orthogonal group, that is, if for all $Q \in \mathcal{O}(d)$, $p(QX) \equiv p(X)$, then each homogeneous part p_i is itself invariant under that action.

Proof. By invariance of p, we obtain for all X and orthogonal Q that

$$p_0 + p_1(X) + \cdots + p_r(X) = p(X) = p(QX) = p_0 + p_1(QX) + \cdots + p_r(QX).$$

By identifying the homogeneous parts of the polynomials, we obtain for all i that $p_i(X) \equiv p_i(QX)$ for all $Q \in \mathcal{O}(d)$, hence each p_i is itself invariant. □

Homogeneous invariant polynomials necessarily have even degree. In particular, there are no interesting invariant polynomials of degree one or less.

Proposition 2.3. *No homogeneous polynomial of odd degree is invariant under orthogonal group action.*

Proof. By contradiction, let $p(X)$ be a homogeneous polynomial of odd degree, invariant under orthogonal group action. Then, $p(X) = p(QX)$ for all orthogonal Q. This holds in particular for $Q = -I_d$, so that $p(X) = p(-X) = -p(X)$ (we used that the degree is odd in the last equality). Thus, $p(X) = 0$ for all X, which contradicts the fact that p has odd degree. □

We now give an elementary statement of a key property of degree-two invariant polynomials: they can be expressed as a linear function of the *Gram matrix* of X, namely X^TX. We use the notation $[k] = \{1, \dots, k\}$ and $\langle A, B\rangle = \mathrm{Tr}(A^TB)$.

Proposition 2.4. *Any homogeneous polynomial $p(X)$ of degree two that is invariant under the action of the orthogonal group is a linear combination of scalar products between the column vectors of X, denoted by $x_1, \dots, x_k$. In other words, there exist coefficients m_{ij} for all $i, j \in [k]$ such that*

$$p(X) = \sum_{i,j} m_{ij}\langle x_i, x_j\rangle = \langle M, X^TX\rangle. \tag{9}$$

Proof. Any homogeneous polynomial of degree two in X can be written as

$$p(X) = \langle \mathrm{vec}(X)\mathrm{vec}(X)^T, A\rangle,$$

where $\mathrm{vec}(X)$ vectorizes a matrix by stacking its columns, and A is a coefficient matrix of size $dk \times dk$. Using the property $\mathrm{vec}(ABC) = (C^T \otimes A)\mathrm{vec}(B)$ where $\otimes$ denotes the Kronecker product, we find that

$$\begin{aligned} p(QX) &= \langle (I_k \otimes Q)\mathrm{vec}(X)\mathrm{vec}(X)^T(I_k \otimes Q)^T, A\rangle \\ &= \langle \mathrm{vec}(X)\mathrm{vec}(X)^T, (I_k \otimes Q)^TA(I_k \otimes Q)\rangle. \end{aligned}$$

If p is invariant, then $p(X) = p(QX)$ for all X and $Q \in \mathcal{O}(d)$. Hence, by identification:

$$A = (I_k \otimes Q)^TA(I_k \otimes Q), \quad \text{for all } Q \in \mathcal{O}(d).$$

Let A be a block matrix with blocks $A_{ij} \in \mathbb{R}^{d\times d}$ for i, j ranging in $[k]$. The above states:

$$A_{ij} = Q^TA_{ij}Q,$$

for all $i, j \in [k]$ and $Q \in \mathcal{O}(d)$. This has the following consequences:

1. Considering each Q in the set of diagonal matrices with diagonal entries in $\{-1, +1\}$ shows that all the elements off the diagonal of A_{ij} are equal to their opposite, implying that A_{ij} is diagonal.

2. Considering each Q in the set of permutation matrices implies that all the diagonal elements are equal.

Thus, $A_{ij} = m_{ij} I_d$ and $A = M \otimes I_d$, where $M \in \mathbb{R}^{k \times k}$. As a result:

$$\begin{aligned} p(X) &= \langle \mathrm{vec}(X), (M \otimes I_d)\mathrm{vec}(X) \rangle \\ &= \langle \mathrm{vec}(X), \mathrm{vec}(XM^T) \rangle \\ &= \langle X, XM^T \rangle \\ &= \langle X^T X, M \rangle. \end{aligned}$$

In words, $p(X)$ is a linear combination of the Gram matrix entries. □

A question that naturally arises is: what are the other invariant polynomials? The first fundamental theorem of the orthogonal group (see [Kăc, 1994, Thm. 14-1.2] for instance) provides an answer to this question.

Theorem 2.5. *The functions $g_{ij} \colon (x_1, \ldots, x_k) \mapsto \langle x_i, x_j \rangle, 1 \leq i \leq j \leq k$, generate the ring of invariant polynomials, that is: any invariant polynomial is a polynomial combination of the degree two invariants g_{ij}.*

Thus, invariant polynomials of degree higher than two do not carry further information about $[X]$. The next natural question is: are the invariants sufficient to fully characterize the equivalence classes? The following classical theorem from invariant theory provides a positive answer to this question in the case of compact groups (see [Kăc, 1994, Thm. 6-2.2] for instance).

Theorem 2.6. *(Informal) The full invariant ring characterizes the equivalence classes.*

Theorems 2.5 and 2.6 combined imply that $X_1^T X_1 = X_2^T X_2$ if and only if $[X_1] = [X_2]$, which is a well-known fact here derived through the prism of invariant features.

2.2 Estimation algorithm

Above, we have shown that the problem of recovering $[X]$ can be reduced (without loss) to that of estimating the Gram matrix $G = X^T X$, with the advantage that the latter is invariant under orthogonal transformations. We build on this observation to propose a concrete algorithm. Consider the Gram matrix of an observation as in eq. (1):

$$Y_i^T Y_i = X^T X + \sigma \left(X^T Q_i^T E_i + E_i^T Q_i X \right) + \sigma^2 E_i^T E_i. \tag{10}$$

By the strong law of large numbers, their empirical mean converges almost surely to their expectation (a characterization of the fluctuations for finite N follows)

$$\hat{M}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i^T Y_i \underset{N \to \infty}{\longrightarrow} X^T X + d\sigma^2 I_k. \tag{11}$$

Here, we used independence of the Y_i's, independence of Q_i and E_i for each i, and the fact that individual entries of each E_i are independent with mean zero and unit variance.

Algorithm 1 Estimation algorithm with σ given

1: Compute the sample mean of the Gram matrices: $\hat{M}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i^T Y_i$.
2: Compute d top eigenvalues $\lambda_1, \ldots, \lambda_d$ of $\hat{M}_N$ with associated orthonormal eigenvectors $v_1, \ldots, v_d \in \mathbb{R}^k$.
3: Define the scaling factors $\alpha_i = \sqrt{\max(0, \lambda_i - d\sigma^2)}$ for $i = 1, \ldots, d$.
4: Form $\tilde{X}_N \in \mathbb{R}^{d\times k}$ with rows $\alpha_i v_i^T$ for $i = 1, \ldots, d$. Our estimator is $[\tilde{X}_N]$.

If the noise level σ is known, we can get an unbiased estimator for $X^T X$ as

$$\hat{G}_N = \hat{M}_N - d\sigma^2 I_k. \tag{12}$$

Since $\hat{G}_N$ is expected to be close to $X^T X$ for large N, it is reasonable to consider an estimator $[\tilde{X}_N]$ for the equivalence class $[X]$ where $\tilde{X}_N$ is a solution of the optimization problem

$$\min_{\hat{X}\in\mathbb{R}^{d\times k}} \|\hat{G}_N - \hat{X}^T\hat{X}\|_{\mathrm{F}}. \tag{13}$$

Well-known extremal properties of the eigenvalue decomposition of a symmetric matrix tell us that an optimizer $\tilde{X}_N$ can be obtained by computing d dominant, orthonormal eigenvectors of $\hat{G}_N$, and scaling them by the square root of their corresponding eigenvalues. Since $\hat{G}_N$ and $\hat{M}_N$ share the same eigenvectors, with eigenvalues related through $\lambda_\ell(\hat{G}_N) = \lambda_\ell(\hat{M}_N) - d\sigma^2$, this computation can equivalently be executed from $\hat{M}_N$ directly. This procedure is summarized in Algorithm 1. Provided $\hat{M}_N$ has an eigengap separating its dth and $(d+1)$st largest eigenvalues, this procedure uniquely defines $[\tilde{X}_N]$. In Section 3, we argue that such an eigengap exists with high probability if N is sufficiently large, and we bound the error $\rho([X], [\tilde{X}_N])$.

We note that the Gram matrix estimator $\hat{G}_N$ could be replaced by a more robust estimator, for example based on median-of-means as favored in [Bandeira et al., 2017a], leveraging work by Nemirovsky and Yudin [1983] and more recently by Joly et al. [2017]. Such refinements are not necessary under our assumption of Gaussian noise but could be useful for heavy-tailed noise.

2.3 Estimation when σ is unknown

If σ is unknown, it can be estimated from the eigenvalues of $\hat{M}_N$ (11). Indeed, in the limit of N going to infinity, the $k-d$ smallest eigenvalues are equal to $d\sigma^2$. For finite N, they fluctuate around that value. Thus, with $\lambda_1 \geq \cdots \geq \lambda_k$ the eigenvalues of $\hat{M}_N$, a possible estimator $\hat{\sigma}_N \geq 0$ for σ is defined by

$$d\hat{\sigma}_N^2 = \frac{1}{k-d}\sum_{\ell=d+1}^{k} \lambda_\ell = \frac{1}{k-d}\Big(\mathrm{Tr}(\hat{M}_N) - (\lambda_1 + \cdots + \lambda_d)\Big), \tag{14}$$

Algorithm 2 Estimation algorithm with σ unknown

1: Execute steps 1 and 2 of Algorithm 1.
2: Define $\hat{\sigma}_N \geq 0$ such that $\hat{\sigma}_N^2 = \dfrac{1}{d(k-d)}\Big(\mathrm{Tr}(\hat{M}_N) - (\lambda_1 + \cdots + \lambda_d)\Big)$.
3: Define the scaling factors $\alpha_i = \sqrt{\max(0, \lambda_i - d\hat{\sigma}_N^2)}$ for $i = 1, \ldots, d$.
4: Execute step 4 of Algorithm 1.

where the second form is computationally favorable. The resulting procedure is summarized as Algorithm 2.

We mention that in the case where X is centered, that is, when $X\mathbf{1} = 0$, σ can be reliably estimated by computing the empirical variance of the dN i.i.d. entries of samples $\frac{1}{\sqrt{k}} Y_i \mathbf{1} \sim \mathcal{N}(0, \sigma^2 I_d)$ for $i = 1, \ldots, N$.

3 Stability of the estimator

The consistency of $\hat{G}_N$ (12) as an estimator of the Gram matrix is guaranteed by the law of large numbers. However, for finite N, we can only hope to estimate the Gram matrix approximately: we characterize the expected errors here. From this approximate Gram matrix, we obtain our estimator $[\tilde{X}_N]$ by solving the optimization problem (13) (which is akin to forming a type of thin Cholesky factorization of $\hat{G}_N$ after projecting the latter to the positive semidefinite matrices): we call this step the *Gram inversion*. To understand the final error on our estimator, we need to study the sensitivity of Gram inversion: we start with this.

3.1 Sensitivity of Gram inversion

Consider the function $f([X]) = X^T X$. This is a map from the quotient space

$$\mathcal{M} = \mathbb{R}_*^{d \times k} / \sim, \tag{15}$$

where $\sim$ is the equivalence relation defined in (3) and $\mathbb{R}_*^{d \times k}$ is the set of matrices in $\mathbb{R}^{d \times k}$ of full rank d, to the set

$$\mathcal{N} = \{G \in \mathrm{Sym}_k : \mathrm{rank}(G) = d, X \succeq 0\}, \tag{16}$$

where Sym_k is the set of symmetric matrices of size k and $G \succeq 0$ means G is positive semidefinite. In Section 2, we argued that $[X]$ can be recovered uniquely from $X^T X$, meaning that f is globally invertible. In this section, we are concerned with the sensitivity of the inverse, f^{-1}.

A bound on the sensitivity appears in [Tu et al., 2016, Lem. 5.4]. We repeat it here.

Lemma 3.1. *For any* $[X], [\tilde{X}] \in \mathcal{M}$,

$$\rho([X], [\tilde{X}]) \leq \frac{L}{\sigma_d(X)} \|X^T X - \tilde{X}^T \tilde{X}\|_{\mathrm{F}}, \qquad \textit{with} \qquad L = \frac{1}{\sqrt{2(\sqrt{2} - 1)}}.$$

This result establishes a Lipschitz constant for f^{-1} in the vicinity of $[X]$, with respect to the distance ρ on $\mathcal{M}$ (6) and the Frobenius distance on $\mathcal{N}$. Through a geometric argument, we confirm that the coefficient $\sigma_d(X)$ in the denominator cannot be avoided, and we show L must be at least $1/\sqrt{2}$ (that is 0.71.. compared to 1.10.. above). Then, we lean on Lemma 3.1 to obtain a new bound with (essentially) that optimal constant. To do so, we use the proof mechanics proposed by Chang and Stehlé [2010] in their study of the stability of the Cholesky decomposition of (strictly) positive definite matrices.

We equip the quotient space $\mathcal{M}$ with a smooth structure as a Riemannian quotient manifold of $\mathbb{R}_*^{d\times k}$ with the standard inner product $\langle\cdot,\cdot\rangle$ [Absil et al., 2008, §3.4].[1] Likewise, we endow $\mathcal{N}$ with a smooth structure as a Riemannian submanifold of Sym_k with the standard inner product $\langle\cdot,\cdot\rangle$, as in [Vandereycken et al., 2009]. With these smooth structures, $f\colon \mathcal{M}\to\mathcal{N}$ is a smooth function. Detailed background on both geometries and their relations can be found in [Massart and Absil, 2018].

The distance ρ happens to be the geodesic distance on $\mathcal{M}$ [Massart and Absil, 2018]. Furthermore, since $\mathcal{N}$ is a Riemannian submanifold of Sym_k equipped with the trace inner product, the Frobenius distance between close-by points of $\mathcal{N}$ is an excellent approximation for the geodesic distance between them. As a result, locally around X^TX, the operator norm (the largest singular value) of the differential of f^{-1} at X^TX reveals the local Lipschitz constant of f^{-1} with respect to these distances.

The inverse function theorem [Lee, 2012, Thm. 4.5] states that the differential of f^{-1} at X^TX is the inverse of the differential of f at $[X]$. Accordingly, we first study the singular values of the differential of f at $[X]$.

As a preliminary step, for $X\in\mathbb{R}_*^{d\times k}$, consider the differential $\mathcal{L}_X$ of the map $X\mapsto X^TX$ (its relation to f is elucidated below):

$$\begin{array}{rccc} \mathcal{L}_X & : & \mathbb{R}^{d\times k} & \to & \mathrm{Sym}_k \\ & & \dot{X} & \mapsto & X^T\dot{X}+\dot{X}^TX. \end{array} \tag{17}$$

Clearly, the following subspace is included in the kernel of $\mathcal{L}_X$:

$$\mathrm{V}_X = \{\Omega X : \Omega+\Omega^T=0\}. \tag{18}$$

We can thus restrict our attention to the orthogonal complement of V_X, which we denote by $\mathrm{H}_X=(\mathrm{V}_X)^\perp$. By definition, $\dot{X}\in\mathbb{R}^{d\times k}$ is orthogonal to V_X if and only if $\langle\dot{X},\Omega X\rangle=0$ for all skew-symmetric matrices Ω, hence:

$$\mathrm{H}_X = \{\dot{X}\in\mathbb{R}^{d\times k} : \dot{X}X^T = X\dot{X}^T\}. \tag{19}$$

Restricted to H_X, the smallest singular value of $\mathcal{L}_X$ is positive.

Proposition 3.2. *Given $X\in\mathbb{R}_*^{d\times k}$, with notation as above, consider the restriction of $\mathcal{L}_X$ as an operator from H_X to $\mathcal{L}_X(\mathrm{H}_X)$. That operator is invertible and its smallest singular value is $\sqrt{2}\sigma_d(X)$.*

[1] In contrast, the quotient space $\mathbb{R}^{d\times k}/\sim$ (without rank restriction) does not admit such a smooth structure, because not all its equivalence classes have the same dimension as submanifolds of $\mathbb{R}^{d\times k}$.

Proof. See appendix C. □

Using tools from differential geometry [Absil et al., 2008, §3.5.8][Massart and Absil, 2018], it can be shown that H_X (equipped with the standard inner product) is isometric to the tangent space of $\mathcal{M}$ at $[X]$, so that the singular values of $\mathcal{L}_X$ on H_X are equal to the singular values of the differential of f at $[X]$. Thus, calling upon the inverse function theorem, we conclude that the largest singular value of the differential of f^{-1} at X^TX is $1/\sqrt{2}\sigma_d(X)$. In turn, this shows the constant L in Lemma 3.1 must be at least $\sqrt{2}$, and the coefficient $\sigma_d(X)$ cannot be removed.

Lemma 3.1 and Proposition 3.2 together allow us to show our main result regarding the local sensitivity of Gram inversion, using a technique by Chang and Stehlé [2010].

Theorem 3.3. *Consider two matrices $X, \tilde{X} \in \mathbb{R}_*^{d\times k}$. If their Gram matrices $G = X^TX$ and $\tilde{G} = \tilde{X}^T\tilde{X}$ are close, specifically, if*

$$\|G - \tilde{G}\|_{\mathrm{F}} \leq \frac{\sigma_d^2(X)}{2},$$

then the equivalence classes must be close too:

$$\rho([X], [\tilde{X}]) \leq \frac{\sigma_d(X)}{\sqrt{2}} \left(1 - \sqrt{1 - \frac{2\|G - \tilde{G}\|_{\mathrm{F}}}{\sigma_d^2(X)}}\right), \tag{20}$$

where ρ is the distance between equivalence classes defined in (6).

Crucially, notice that for small $\|G - \tilde{G}\|_{\mathrm{F}}$, we have $\sqrt{1 - \frac{2\|G-\tilde{G}\|_{\mathrm{F}}}{\sigma_d^2(X)}} \approx \frac{1}{\sigma_d^2(X)}\|G - \tilde{G}\|_{\mathrm{F}}$. Hence, the right-hand side of (20) behaves like $\frac{1}{\sqrt{2}\sigma_d(X)}\|G-\tilde{G}\|_{\mathrm{F}}$, which by the discussion above cannot be improved.

Proof. Let $U\Sigma V^T$ be the SVD of $X\tilde{X}^T$. The orthogonal matrix $Q = UV^T$ optimally aligns X and $\tilde{X}$, in the sense that $\rho([X], [\tilde{X}]) = \|X - Q\tilde{X}\|_{\mathrm{F}}$. Then,

$$X\tilde{X}^TQ^T = U\Sigma U^T = Q\tilde{X}X^T.$$

Since the theorem statement depends on X and $\tilde{X}$ only through $[X]$ and $[\tilde{X}]$, without loss of generality, suppose that X and $\tilde{X}$ are already rotationally aligned, that is, $Q = I$. Then, $X\tilde{X}^T$ is symmetric, positive semidefinite. Define $\Delta X = \tilde{X} - X$: notice that $\Delta X X^T$ is also symmetric, and

$$\tilde{X}^T\tilde{X} = (X + \Delta X)^T(X + \Delta X) = X^TX + (\Delta X^TX + X^T\Delta X) + \Delta X^T\Delta X.$$

Rearranging, we get:

$$\tilde{X}^T\tilde{X} - X^TX - \Delta X^T\Delta X = \mathcal{L}_X(\Delta X),$$

where $\mathcal{L}_X$ is the operator defined in (17). Since ΔX is in the subspace H_X defined in (19), it is also orthogonal to the null space V_X of $\mathcal{L}_X$. As a result, we may write

$$\Delta X = \mathcal{L}_X^{\dagger}\left(\tilde{X}^T\tilde{X} - X^TX - \Delta X^T\Delta X\right),$$

where $\mathcal{L}_X^{\dagger}$ is the Moore–Penrose pseudo-inverse of $\mathcal{L}_X$. Proposition 3.2 then implies:

$$\|\Delta X\|_{\mathrm{F}} \leq \frac{1}{\sqrt{2}\cdot\sigma_d(X)}\left[\|\tilde{X}^T\tilde{X} - X^TX\|_{\mathrm{F}} + \|\Delta X\|_{\mathrm{F}}^2\right].$$

Reorganizing, we get the following inequality:

$$0 \leq \|\Delta X\|_{\mathrm{F}}^2 - \sqrt{2}\cdot\sigma_d(X)\|\Delta X\|_{\mathrm{F}} + \|\tilde{X}^T\tilde{X} - X^TX\|_{\mathrm{F}}. \tag{21}$$

The right-hand side of this inequality is a quadratic in $\|\Delta X\|_{\mathrm{F}}$. Under our assumptions, the two roots of this quadratic, ξ_- and ξ_+, are real and nonnegative:

$$\xi_{\pm} = \frac{\sqrt{2}\cdot\sigma_d(X)}{2} \pm \frac{\sqrt{2\cdot\sigma_d(X)^2 - 4\cdot\|\tilde{X}^T\tilde{X} - X^TX\|_{\mathrm{F}}}}{2}.$$

Since $\|\Delta X\|_{\mathrm{F}}$ satisfies (21), it lies outside the open interval defined by (ξ_-, ξ_+). This means there are two possibilities: either $\|\Delta X\|_{\mathrm{F}} \in [0, \xi_-]$, or $\|\Delta X\|_{\mathrm{F}} \geq \xi_+$. We aim to exclude the latter. To do so, notice that Lemma 3.1 together with our proximity assumption on the Gram matrices implies:

$$\|\Delta X\|_{\mathrm{F}} = \rho([X],[\tilde{X}]) \leq \frac{1}{\sigma_d(X)}\frac{1}{\sqrt{2(\sqrt{2}-1)}}\frac{\sigma_d^2(X)}{2} < \frac{\sqrt{2}\cdot\sigma_d(X)}{2} \leq \xi_+.$$

This allows to conclude that $\rho([X],[\tilde{X}]) = \|\Delta X\|_{\mathrm{F}} \leq \xi_-$. Upon factoring out $\frac{\sigma_d(X)}{\sqrt{2}}$ in the expression for ξ_-, this completes the proof. □

3.2 Upper bounds on cloud estimation error

Theorem 3.3 quantifies how a good estimator for the Gram matrix of $[X]$ can be turned into a good estimator for $[X]$ itself. In this part, we first show quantitatively that, with high probability, we can indeed have a good estimator for the Gram matrix. Afterwards, we connect this result with the above theorem to produce a bound on the estimation error of $[X]$.

The first result relies on standard concentration bounds for quadratic forms of Gaussian random variables. (We remark that it is possible to relax the assumptions to require subgaussian noise rather than Gaussian noise.) We assume σ is known, and we use the notation $\|X\|_{\mathrm{op}} = \sigma_1(X)$ for the operator norm. The projection to the set of positive semidefinite matrices of rank at most d is necessary to apply Theorem 3.3, and causes no difficulties in practice.

Theorem 3.4. *Let $Y_1, \dots, Y_N$ be i.i.d. observations drawn from model (1) and let $\hat{G}_N$ be the Gram estimator as defined in (12). Since the true Gram matrix $G = X^T X$ is positive semidefinite with rank at most d, project $\hat{G}_N$ to that set; in the notation of Algorithm 1:*

$$\tilde{G}_N = \sum_{i=1}^{d} \alpha_i^2 v_i v_i^T \in \operatorname*{argmin}_{H \succeq 0 : \operatorname{rank}(H) \leq d} \|\hat{G}_N - H\|_{\mathrm{op}}.$$

Then, for any $\delta \in (0,1)$, with probability at least $1 - \delta$:

$$\|\tilde{G}_N - G\|_{\mathrm{F}} \leq 8\sqrt{2d} \left[\sqrt{\left(\frac{2\|X\|_{\mathrm{op}}^2 \sigma^2 + d\sigma^4}{N} \right) k \log\left(\frac{10}{\delta}\right)} + \frac{\sigma^2}{N} k \log\left(\frac{10}{\delta}\right) \right].$$

Proof. For all $u \in \mathbb{S}^{k-1}$ (the unit sphere in $\mathbb{R}^k$), consider:

$$\begin{aligned} u^T(\hat{G}_N - G)u &= u^T \left(\frac{1}{N} \sum_{i=1}^{N} Y_i^T Y_i - d\sigma^2 I_k - X^T X \right) u \\ &= \frac{1}{N} \sum_{i=1}^{N} \left[\|(Q_i X + \sigma E_i)u\|^2 - (\|Xu\|^2 + d\sigma^2) \right]. \end{aligned} \tag{22}$$

Notice that $\|(Q_i X + \sigma E_i)u\|^2 = \|(X + \sigma Q_i^T E_i)u\|^2$ is equal in distribution to $\|(X + \sigma E_i)u\|^2$ since entries of E_i are standard Gaussian. Thus, the first part of (22) is distributed like a sum of squared norms of i.i.d. non-centered Gaussian vectors. Reorganizing standard concentration bounds for noncentral χ^2 (see for instance [Birgé, 2001, Lem. 8.1]) gives, for all u in $\mathbb{S}^{k-1}$,

$$\mathbb{P}\left[u^T(\hat{G}_N - G)u \geq 2\sqrt{\left(2\|Xu\|^2 + d\sigma^2\right)\frac{t}{N}\sigma^2} + 2\frac{t}{N}\sigma^2 \right] \leq e^{-t}, \text{ and} \tag{23}$$

$$\mathbb{P}\left[u^T(\hat{G}_N - G)u \leq -2\sqrt{\left(2\|Xu\|^2 + d\sigma^2\right)\frac{t}{N}\sigma^2} \right] \leq e^{-t}. \tag{24}$$

We then cover the sphere $\mathbb{S}^{k-1}$ with an ε-net. A union bound over all the elements of the net (see for instance [Vershynin, 2010, §5.2.2]) yields:

$$\mathbb{P}\left[\|\hat{G}_N - G\|_{\mathrm{op}} \leq 4\sqrt{(2\|X\|_{\mathrm{op}}^2 + d\sigma^2)\frac{t}{N}\sigma^2} + 4\frac{t}{N}\sigma^2 \right] \geq 1 - 2 \cdot 5^k \cdot e^{-t}.$$

Since G is positive semidefinite and has rank d, the projection $\tilde{G}_N$ satisfies:

$$\|\hat{G}_N - \tilde{G}_N\|_{\mathrm{op}} \leq \|\hat{G}_N - G\|_{\mathrm{op}}.$$

By the triangle inequality, we obtain:

$$\|G - \tilde{G}_N\|_{\mathrm{op}} \leq \|G - \hat{G}_N\|_{\mathrm{op}} + \|\hat{G}_N - \tilde{G}_N\|_{\mathrm{op}} \leq 2\|G - \hat{G}_N\|_{\mathrm{op}}.$$

Since $\operatorname{rank}(G), \operatorname{rank}(\tilde{G}_N) \leq d$, it follows that $\operatorname{rank}(G - \tilde{G}_N) \leq 2d$ and we get:

$$\|G - \tilde{G}_N\|_{\mathrm{F}} \leq 2\sqrt{2d}\|G - \hat{G}_N\|_{\mathrm{op}}.$$

As a result,

$$\mathbb{P}\left[\|\tilde{G}_N - G\|_{\mathrm{F}} \leq 8\sqrt{2d(2\|X\|_{\mathrm{op}}^2 + d\sigma^2)\frac{t}{N}\sigma^2} + 8\sqrt{2d}\frac{t}{N}\sigma^2\right] \geq 1 - 2 \cdot 5^k \cdot e^{-t}.$$

Taking $t = \log\left(\frac{2 \cdot 5^k}{\delta}\right)$ implies the final result:

$$\|\tilde{G}_N - G\|_{\mathrm{F}} \leq 8\sqrt{2d}\left[\sqrt{\left(2\|X\|_{\mathrm{op}}^2 + d\sigma^2\right)\frac{\sigma^2}{N}k\log\left(\frac{10}{\delta}\right)} + \frac{\sigma^2}{N}k\log\left(\frac{10}{\delta}\right)\right]$$

with probability at least $1 - \delta$. □

Combining Theorem 3.4 with a stability result on the thin Cholesky decomposition gives the main result on the stability of the proposed estimator, as measured with the distance ρ (6).

Corollary 3.5. *Let $Y_1, \ldots, Y_N$ be N i.i.d. samples drawn according to (1). Let $[\tilde{X}]$ be the estimator returned by Algorithm 1. Then, for large N, with probability at least $1 - \delta$,*

$$\rho([X], [\tilde{X}]) \leq \frac{8L\sqrt{2d}}{\sigma_d(X)}\left[\sqrt{\left(2\|X\|_{\mathrm{op}}^2 + d\sigma^2\right)\frac{\sigma^2}{N}k\log\left(\frac{10}{\delta}\right)} + \frac{\sigma^2}{N}k\log\left(\frac{10}{\delta}\right)\right],$$

where L can be taken as $1/\sqrt{2(\sqrt{2}-1)}$ or, for large enough N, arbitrarily close to $1/\sqrt{2}$.

Proof. The first claim follows from Theorem 3.4 and Lemma 3.1. For N large enough, Theorem 3.4 shows that the assumption of Theorem 3.3, namely, $\|\tilde{G}_N - G\|_{\mathrm{F}} \leq \frac{\sigma_d^2(X)}{2}$, is satisfied with high probability. In that scenario, combining the two theorems yields the second result. □

4 Statistical optimality of the estimator

Estimating the equivalence class of X from samples of the form (1) is a particular instance of an estimation problem under a group action. Bandeira et al. [2017a] showed that the statistical complexity of estimation problems under a group action is connected to the structure of the group acting on the parameter. In particular, it is shown that the minimum number of samples required to reliably estimate the parameter grows as $O(\sigma^{2p})$ where p is the smallest degree of invariant polynomials required to fully characterize the equivalence classes. We have shown in Section 2 that, in our case, $p = 2$. In this section, we build on those results to show minimax lower bounds on the estimation of

the equivalence class $[X]$. We also provide matching upper bounds, hence showing the statistical optimality of our estimator in the low SNR regimes.

To fix scale and to avoid pathological cases, throughout this section we assume that X belongs to the space

$$\mathcal{X} = \{X \in \mathbb{R}^{d\times k} : \|X\|_{\mathrm{F}}^2 \leq d \text{ and } \sigma_d(X) \geq \eta\}, \tag{25}$$

where $\sigma_d(X)$ is the dth (that is, smallest) singular value of X, and $\eta > 0$ is fixed.

4.1 Lower bound on the estimation error for high noise regimes

In the presence of large noise (that is, for large σ), the MSE of *any* estimator of the equivalence class of X scales with σ as σ^4: we make this precise in the following theorem.

Theorem 4.1. *Suppose we observe N samples $Y_1, \ldots, Y_N$ drawn independently according to (1). Then the so-called minimax risk for estimating the orbit of X satisfies:*

$$\inf_{\hat{X}} \sup_{X\in\mathcal{X}} \mathbb{E}[\rho^2(X, \hat{X})] \asymp \frac{\sigma^4}{N},$$

for sufficiently large σ, where ρ is as in (6) and $\mathcal{X}$ is defined by (25), and the infimum is taken over all possible estimators, random or deterministic.

The general study of lower bounds for estimation under a group action has been addressed by Abbe et al. [2018b], using Chapman–Robbins bounds. Here, we take a different approach, proving minimax rates using two ingredients: a *tight* bound on the Kullback–Leibler (KL) divergence, and a bound on the packing number of a particular metric space. However, since we are not aware of any result on the packing number of our parameter space for the metric ρ, we consider a strict subset of this parameter space for which tight bounds on the packing number are known. Specifically, we start by noticing that the Grassmannian $\mathrm{G}(k, d)$—the set of d dimensional subspaces of $\mathbb{R}^k$—is in correspondence with a subset of the parameter space. Building on this observation, we restrict ourselves to the strictly simpler problem where the matrix X satisfies the condition $XX^T = I_d$. In this restricted setting, there is a one-to-one correspondence between an equivalence class and an element of the Grassmannian. We then use a result on the covering number of the Grassmannian to control its *local* packing number. This result was used by Cai et al. [2013] in the context of optimal rates of estimation for the principal subspace of a covariance matrix under a sparsity assumption. Then, we show a tight bound of the KL divergence. This bound is a particular case of a more general result on the KL divergence of samples observed under the action of a group.

1. Packing number of the Grassmannian: We start by defining a metric on $\mathrm{G}(k, d)$. This metric on the Grassmanian is shown to be equivalent to the distance between equivalence classes ρ. We then show tight lower and upper bounds on the covering number of $\mathrm{G}(k, d)$ for the aforementioned metric. This result is due to Szarek [1982]. We then use a result on the local packing number of the Grassmannian:

leveraging the previous result, for any $\alpha \in (0,1)$ and for any ε small enough, we give a lower bound on the $\alpha\varepsilon$-packing number of a ball of radius ε. This follows the technique proposed by Yang and Barron [1999].

2. Tight control of the KL divergence: we state a lemma giving a bound on the KL divergence of the distribution $\mathbb{P}_X$ of samples drawn according to (1). This lemma follows from a result by Bandeira et al. [2017a].

We start by giving the result on the covering number of the Grassmannian.

Lemma 4.2. *[Cai et al. [2013], Lemma 1] Define the metric on* $\mathrm{G}(k,d)$ *by* $\tilde{\rho}(V,U) = \|V^TV - U^TU\|_{\mathrm{F}}$. *Then, for any* $\varepsilon \in \left(0, \sqrt{2\min(d,k-d)}\right]$, *we have*

$$\left(\frac{c_0}{\varepsilon}\right)^{d(k-d)} \leq \mathfrak{N}(\mathrm{G}(k,d),\varepsilon) \leq \left(\frac{c_1}{\varepsilon}\right)^{d(k-d)},$$

where $\mathfrak{N}(\mathrm{G}(k,d),\varepsilon)$ *is the* ε*-covering number of* $\mathrm{G}(k,d)$ *with respect to the metric* $\tilde{\rho}$ *and* c_0, c_1 *are absolute constants.*

Building on the previous result, we can state a result on the *local* packing of $G(k,d)$.

Lemma 4.3. *Let* $B(V,\varepsilon) = \{U \in \mathrm{G}(k,d) : \tilde{\rho}(U,V) \leq \varepsilon\}$, $\alpha \in (0,1)$ *and* $\varepsilon \in (0,\varepsilon_0]$. *Then, there exists* $V^* \in \mathrm{G}(k,d)$ *such that:*

$$\mathfrak{M}(B(V^*,\varepsilon),\alpha\varepsilon) \geq \left(\frac{c_0}{\alpha c_1}\right)^{d(k-d)},$$

where $\mathfrak{M}(\mathrm{G}(k,d),\varepsilon)$ *is the* ε*-packing number of* $\mathrm{G}(k,d)$ *with respect to the metric* $\tilde{\rho}$ *and* c_0, c_1 *are the absolute constants in Lemma 4.2.*

Proof. The original proof can be found in [Yang and Barron, 1999]. For convenience, we provide the proof in Appendix D. □

Finally, we state a bound on the KL divergence of the distribution $\mathbb{P}_X$ of samples drawn according to (1).

Lemma 4.4. *For given dimensions* d *and* k, *there exists a universal constant* C *such that, for any* X_1, $X_2 \in \mathbb{R}^{d\times k}$ *with* $\rho(X_1,X_2) \leq \dfrac{\|X_1\|_{\mathrm{F}}}{3}$ *and for any* $\sigma > 1$, *we have:*

$$\mathrm{KL}(\mathbb{P}_{X_1}||\mathbb{P}_{X_2}) \leq C\sigma^{-4}\rho^2(X_1,X_2).$$

Proof. See Appendix E. □

Combining these three results, we obtain Theorem 4.1. The proof technique is inspired from [Yang and Barron, 1999] and [Cai et al., 2013]. The full proof can be found in Appendix D. The constant in the lower bound exhibits a dependence in the dimension, through the metric entropy of the Grassman manifold. As we focus on the dependence in σ and N (noise level and number of samples), we leave the question of whether this dependence in the dimension is tight for future research.

4.2 Lower bound on the sample complexity for high noise regimes

We now show that, still when σ is large, the number of samples drawn according to (1) necessary to reliably estimate the equivalence class of $[X]$ grows as $O(\sigma^4)$. For that matter, we provide matching upper and lower bounds: the upper bound on the KL divergence obtained in Lemma 4.4 in combination with Neyman–Pearson's Lemma (see Lemma 4.3 in [Rigollet and Hütter, 2017] for instance) yields the lower bound, while properties of the estimator obtained with Algorithm 1 give the upper bound. This result is similar in its message to, but technically different from, Theorem 4.1.

Theorem 4.5. *Let $\tau \in (0, 1/2)$ be arbitrary. There exists a constant $\tilde{c}$ (possibly function of τ) such that the following holds: Let $X_1, X_2 \in \mathcal{X}$ belong to two distinct equivalence classes, and let $Y_1, \dots, Y_N$ be drawn according to (1), where X is either X_1 or X_2, with probability 1/2 each. Any test ψ whose task it is to decide whether X is X_1 or X_2 has probability of error at least $1/2 - \tau$ whenever $N < \tilde{c}\sigma^4$.*

Moreover, there exists a constant $\tilde{C}$ such that Algorithm 1 outputs $\hat{X}_N$ satisfying:

$$\mathbb{P}\left(\rho(\hat{X}_N, X) \leq \varepsilon\right) \geq 1 - \delta$$

whenever $N \geq \tilde{C}\dfrac{\sigma^4}{\delta \cdot \varepsilon^2}$.

Proof. We start by proving that no statistical procedure can reliably distinguish between two equivalence classes whenever the number of samples is less than $O(\sigma^4)$. In particular, for any test ψ using N samples we have:

$$\begin{aligned}
\mathbb{P}_{X_1}(\psi = 2) + \mathbb{P}_{X_2}(\psi = 1) &\geq 1 - \mathrm{TV}(\mathbb{P}_{X_1}^N || \mathbb{P}_{X_1}^N) \\
&\geq 1 - \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_{X_1}^N || \mathbb{P}_{X_2}^N)} \\
&\geq 1 - \sqrt{\frac{1}{2}N \cdot \mathrm{KL}(\mathbb{P}_{X_1} || \mathbb{P}_{X_2})} \\
&\geq 1 - \sqrt{\frac{CN\rho^2(X_1, X_2)}{2\sigma^4}},
\end{aligned}$$

where the first inequality follows from Neyman–Pearson's Lemma while the second follows from Pinsker's inequality. Therefore, if $N \leq \dfrac{8}{C\rho^2(X_1, X_2)}\tau^2\sigma^4$, we get that:

$$\frac{1}{2}\mathbb{P}_{X_1}(\psi = 2) + \frac{1}{2}\mathbb{P}_{X_2}(\psi = 1) \geq \frac{1}{2} - \tau,$$

yielding the sought lower bound on the probability of error. The constant $\tilde{c}$ can be set uniformly against the choice of X_1, X_2 since $\mathcal{X}$ is bounded.

For the upper bound, we simply notice that, for large σ, by Markov's inequality,

$$\mathbb{P}\left(\rho(X, \hat{X}) \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathbb{E}\left[\rho^2(X, \hat{X})\right] \leq \frac{\tilde{C}}{\varepsilon^2}\frac{\sigma^4}{N},$$

where the first inequality follows from Markov's inequality while the second inequality follows from the computations in Appendix F.2. □

4.3 A comment regarding the special orthogonal group

In our model, we consider the case where the orthogonal transformations acting on the cloud are in $\mathcal{O}(d)$. In some applications, it may be more appropriate to assume rotations $Q_1, \ldots, Q_N$ in the special orthogonal group $\mathcal{SO}(d)$, that is, with determinant $+1$ (no reflections). However, the problem in $\mathcal{SO}(d)$ is harder. Indeed, by the first fundamental theorem for the special orthogonal group, the entries of the Gram matrix $X^T X$ and the $d \times d$ minors of X generate the ring of invariant polynomials (see [Kraft and Procesi, 2000, Prop. 10.2] for instance). Since our algorithm only recovers the Gram matrix, that is, the orbits of matrix X up to a reflection, it falls short of estimating such reflection. Estimating the reflection can be done by estimating the determinants of the submatrices of X of size $d \times d$, which requires $O(\sigma^{2d})$ samples at high noise level, as further discussed in [Bandeira et al., 2017a].

4.4 Noise regimes

We here summarize the adaptive optimality property of our estimator, characterized by a phase transition around a critical noise level.

Proposition 4.6. *The estimator in Algorithm 1 exhibits a phase transition around a critical noise level, namely:*

- *If $\sigma^2 \gg 1$, the MSE of the estimator behaves like $O\left(\sigma^4/N\right)$, which is optimal by Theorem 4.1.*
- *If $\sigma^2 \ll 1$, the MSE of the estimator behaves like $O\left(\sigma^2/N\right)$, which is optimal given the fact that the problem is strictly harder than estimating the matrix X provided the rotations are available, in which case the lower bound on the estimation is $O(\sigma^2/N)$.*

Proof. The proof of the upper bound relies on a computation of bounds on the MSE of the estimator proposed in Algorithm 1: see Appendix F. □

5 Numerical Experiments

In this section, we provide numerical support for our theoretical predictions.[2] In particular, we highlight the adaptative optimality of Algorithm 1 and validate the heuristic proposed for the estimation of σ^2.

5.1 Noise regimes

Figure 1 illustrates the accuracy of Algorithm 1 over a large range of values of N (number of observations) and σ (noise level), with $k = 100$ and $d = 3$ fixed. A clear phase transition is visible, delineating a regime where our estimator is accurate, and one where

[2] Code to generate the figures: https://github.com/thomaspdl/GeneralizedProcrustes.

it fails. Crucially, the phase transition illustrates the adaptiveness of the estimator. Specifically, as predicted, for small σ, the number of observations needs to grow as $N \sim \sigma^2$ in order to preserve a constant MSE while compensating for noise. For large σ, this relationship deteriorates and we require $N \sim \sigma^4$. As explained in Section 4, this is not merely a requirement of our estimator: any estimator is subject to the same needs, and it is a positive feature of Algorithm 1 that it adapts automatically to both regimes.

To illustrate these two noise regimes, the graph is overlaid with two lines. The red line of slope $1/2$ represents a noise level varying as $O(\sqrt{N})$. Its intercept is chosen as follows: consider an oracle which knows the rotations Q_i affecting the measurements Y_i. This oracle can compute the maximum likelihood estimator simply by undoing the rotations, then averaging. It is easy to see that the MSE of that oracle is $\frac{\sigma^2 dk}{N}$ in Frobenius distance (it may be slightly less in ρ distance). The red line shows the relationship between N and σ when that oracle has a relative MSE of 0.95. It is interesting to see that, for low noise levels, the invariants-based estimator has performance similar to that oracle. The blue line of slope $1/4$ represents a noise level varying as $O(N^{\frac{1}{4}})$: its intercept is chosen manually, for illustration.

5.2 Estimation of σ^2

To illustrate the accuracy of the approximation made in (14), we compute the mean empirical error of our estimator. In particular, for different number of observations N, we estimate the average relative error of the estimator in Figures 2. We observe a small empirical relative error decreasing with the number of samples.

6 Perspectives

Throughout this paper, we only consider the case where one cloud of points must be estimated. An interesting variant is *heterogeneous* Procrustes, where each observation consists of a noisy, rotated version of a cloud of points picked at random among K possibilities, $X_1, \ldots, X_K$; for example:

$$Y = QX_s + \sigma E, \text{ with } \mathbb{P}(s = j) = w_j,$$

where $w_1, \ldots, w_K \geq 0$ sum to 1. Several approaches to solve this problem are possible: taking a moments-based method similar to the one proposed by Hsu and Kakade [2013] or inverting invariant features using a nonconvex optimization approach as in [Boumal et al., 2017, Ma et al., 2018] could be two interesting directions. Regarding the latter, one idea is to consider these fourth-order $\mathcal{O}(d)$-invariants: $Y^TY \otimes Y^TY$, where $\otimes$ is the Kronecker product.

Another possible extension is to consider observations with *unlabeled* clouds of points, that is: we observe several copies of X after an unknown rotation, and also after an unknown permutation of the points to model the fact that we do not know which point is which across various observations. This can be modeled as

$$Y = QXP + \sigma E,$$

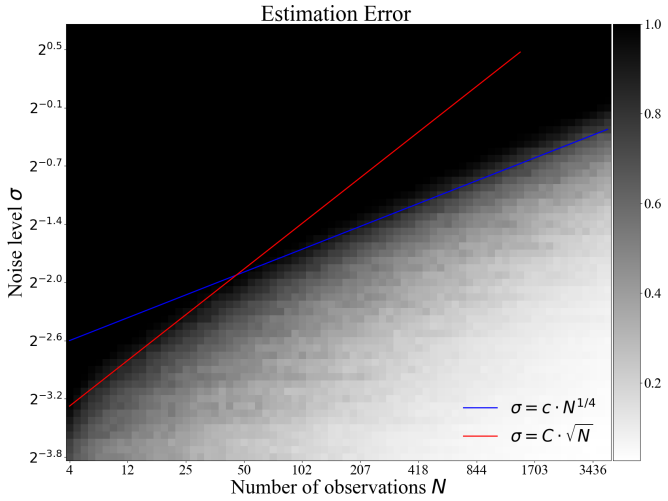

Figure 1: We generate a random cloud $X \in \mathbb{R}^{d\times k}$ with $d = 3$ and $k = 100$ and i.i.d. Gaussian entries of variance 1. The cloud is then renormalized to have unit Frobenius norm. For each pair (N, σ) on a log-log grid, N observations of X are produced with noise level σ (known), and the equivalence class $[X]$ is estimated using Algorithm 1. Each pixel's brightness indicates the relative estimation error $\rho([X], [\hat{X}])/\|X\|_{\mathrm{F}}$ (6), capped at one and then averaged over 50 independent repetitions. The blue and red lines illustrate how Algorithm 1 is adaptive to noise levels: see Section 5.1 for details.

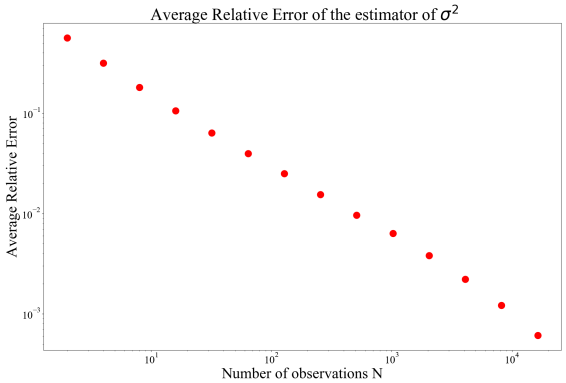

Figure 2: We generate a random cloud $X \in \mathbb{R}^{d\times k}$ with $d = 3$ and $k = 100$ and i.i.d. Gaussian columns. For different number of samples, we compute the relative error $\frac{|\sigma^2-\hat{\sigma}^2|}{\sigma^2}$ on the estimation of σ^2. This relative error is then averaged over 250 i.i.d. realizations. The results are displayed in log-log scale. In practice, we notice that the empirical variance of the estimator is dominated by the relative error of estimation.

where Q is an unknown orthogonal matrix of size $d \times d$ as usual, and P is an unknown permutation matrix of size $k \times k$. Assuming i.i.d. Gaussian noise, in distribution, $Y = Q(X + \sigma E)P$, so that the singular values of Y are equal to those of $X + \sigma E$: up to the noise, the singular values are (non-polynomial) invariants. Studying the distribution of the singular values of Y as a function of those of X may allow one to estimate the general shape of the cloud X (specifically, its singular values) without the need to register clouds (no rotation estimation, and no point correspondence estimation). This connects to principal component analysis.

Yet another extension is to consider *projected* observations. For example, X is a cloud of points in 3-D ($d = 3$), but observations are of the form $Y = P(QX + \sigma E)$, where P is a projector to a 2-D plane (for example, the camera plane). Equivalently,

$$Y = U^T X + \sigma E,$$

where U is an unknown matrix of size 3×2 with orthonormal columns, and E has size $2 \times k$. Assuming U is uniformly distributed, $\mathbb{E}[UU^T] = \frac{2}{3} I_3$, so that $Y^T Y$, while not an invariant, does reveal information about X in expectation since $\mathbb{E}[Y^T Y] = \frac{2}{3} X^T X + d\sigma^2 I_k$. Much of the machinery developed in the present paper applies directly to this extended setting: it would be interesting to study it further.

In all of these, one could also include the possibility that clouds are not centered, that is: they are observed only after an unknown translation, on top of other transformations.

Acknowledgments

We thank Afonso Bandeira, Adrien Bilal, Jianqing Fan, Joe Kileel and Joao Pereira for insightful discussions. NB is partially supported by NSF award DMS-1719558.

References

E. Abbe, T. Bendory, W. Leeb, J. Pereira, N. Sharon, and A. Singer. Multireference Alignment is Easier with an Aperiodic Translation Distribution. 2018a. URL https://arxiv.org/pdf/1710.02793.pdf.

E. Abbe, J. Pereira, and S. A. Estimation in the group action channel. In *International Symposium on Information Theory (ISIT)*. 2018b. URL https://arxiv.org/abs/1801.04366.

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3.

A. Bandeira, B. Blum-Smith, J. Kileel, A. Perry, J. Weed, and A. Wein. Estimation under group actions: recovering orbits from invariants. 2017a. URL https://arxiv.org/abs/1712.10163.

A. Bandeira, P. Rigollet, and J. Weed. Optimal rates of estimation for multi-reference alignment. 2017b. URL https://arxiv.org/abs/1702.08546.

A. S. Bandeira, M. Charikar, A. Singer, and A. Zhu. Multireference alignment using semidefinite programming. In *Innovations in Theoretical Computer Science, ITCS'14, Princeton, NJ, USA, January 12-14, 2014*, pages 459–470, 2014. doi: 10.1145/2554797.2554839. URL https://doi.org/10.1145/2554797.2554839.

A. S. Bandeira, C. Kennedy, and A. Singer. Approximating the Little Grothendieck Problem over the Orthogonal and Unitary Groups. *Mathematical Programming SERIES A*, 160:433–475, 2016.

T. Bendory, N. Boumal, C. Ma, Z. Zhao, and A. Singer. Bispectrum Inversion with Application to Multireference Alignment. *IEEE Transactions on Signal Processing*, 66(4):1037–1050, 2018. doi: 10.1109/TSP.2017.2775591.

L. Birgé. An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, 36:113–133, 2001.

I. Borg and P. J. Groenen. *Modern Multidimensional Scaling: Theory and Applications (Second Edition)*. Springer, New York, NY, 2005. ISBN 0-387-25150-2.

N. Boumal, T. Bendory, R. Lederman, and A. Singer. Heterogeneous multireference alignment: a single pass approach. 2017. URL arXivpreprintarXiv:1710.02590.

T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. *Annals of Statistics*, 41(6):3074–3110, 2013.

X.-W. Chang and D. Stehlé. Rigorous Perturbation Bounds of Some Matrix Factorizations. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2841–2859, 2010.

R. M. Fano. Class notes for MIT course 6.574: Transmission of Information, 1952.

G. B. Giannakis. Signal reconstruction from multiple correlations: frequency-and time-domain approaches. *JOSA A*, 6(5):682–697, 1989.

M. Goemans and D. Williamson. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995. doi: 10.1145/227683.227684.

J. Gower and G. Dijksterhuis. *Procrustes Problem*, volume 30. Oxford Statistical Science Series, 2005. ISBN 978-0198510581.

E. Grave, A. Joulin, and Q. Berthet. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. 2018. URL https://arxiv.org/abs/1805.11222.

P. E. Green and R. D. Carroll. Mathematical Tools for Applied Multivariate Analysis. *New York: Academic*, 1976.

D. Hsu and S. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science (ITCS)*, pages 11–20. ACM, 2013.

J. R. Hurley and R. B. Cattell. The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 7:258–262, 1962.

E. Joly, G. Lugosi, and R. Oliveira. On the estimation of the mean of a random vector. *Electron. J. Statist.*, 11(1):440–451, 2017. doi: 10.1214/17-EJS1228. URL https://doi.org/10.1214/17-EJS1228.

S. Khot and A. Naor. Grothendieck-Type Inequalities in Combinatorial Optimization. *Communications on Pure and Applied Mathematics*, 65(7):992–1035, 2012. doi: 10.1002/cpa.21398.

H. Kraft and C. Procesi. Classical invariant theory, a primer. *Lecture Notes, Version*, 2000.

V. Kăc. Invariant theory, 1994. URL https://people.kth.se/~laksov/notes/invariant.pdf.

J. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer-Verlag New York, 2 edition, 2012. ISBN 978-1-4419-9981-8. doi: 10.1007/978-1-4419-9982-5.

C. Ma, T. Bendory, N. Boumal, F. Sigworth, and A. Singer. Heterogeneous multireference alignment for images with application to 2-D classification in single particle reconstruction. 2018. URL arXivpreprintarXiv:1811.10382.

A. Man-Cho So. Moment inequalities for sums of random matrices and their applications in optimization. *Mathematical programming*, 130(1):125–151, 2010.

E. Massart and P.-A. Absil. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. 2018. URL https://sites.uclouvain.be/absil/2018-06/quotient_tech_report.pdf.

A. Naor, O. Regev, and T. Vidick. Efficient Rounding for the Noncommutative Grothendieck inequality. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 71–80. ACM, 2013.

A. Nemirovski. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Mathematical Programming*, 109(2–3):283–317, 2007. doi: 10.1007/s10107-006-0033-0.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics, 1983.

J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

A. Perry, J. Weed, A. Bandeira, P. Rigollet, and A. Singer. The sample complexity of multi-reference alignment. 2017. URL https://arxiv.org/abs/1707.00943.

P. Rigollet and J.-C. Hütter. High Dimensional Statistics, 2017. URL http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf.

B. M. Sadler and G. B. Giannakis. Shift and Rotation Invariant Object Reconstruction Using The Bispectrum. *JOSA A*, 9(1):57–69, 1992.

P. H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011. doi: 10.1016/j.acha.2010.02.001.

A. Singer. Mathematics for cryo-electron microscopy. volume 3, pages 3981–4000, 2018. URL https://eta.impa.br/dl/035.pdf.

T. D. H. H. S. Smith, S. L. and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. 2017. URL https://arxiv.org/abs/1702.03859.

S. Szarek. Nets of Grassmann manifold and orthogonal groups. In *Proceedings of Banach Spaces Workshop, University of Iowa Press*, pages 169–185. 1982.

J. Ten Berge. Orthogonal Procrustes rotation for two or more matrices. *Psychometrika*, 42(2):267–276, 1977. doi: 10.1007/BF02294053.

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag, 2009.

S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank Solutions of Linear Matrix Equations via Procrustes Flow. *ICML 16 Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 48:964–973, 2016.

J. W. Tukey. *The spectral representation and transformation properties of the higher moments of stationary time series.*, volume 1. Wadsworth, 1984.

B. Vandereycken, P.-A. Absil, and S. Vandewalle. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 389–392, Aug 2009. doi: 10.1109/SSP.2009.5278558.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. 2010. URL arXivpreprintarXiv:1011.3027.

C. Xing, D. Wang, C. Liu, and Y. Lin. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL HLT)*, pages 1006–1011, 2015.

Y. Yang and A. Barron. Information-Theoretic Determination of Minimax Rates of Convergence. *Annals of Statistics*, 27:1564–1599, 1999.

Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1330–1334, December 2000. URL https://www.microsoft.com/en-us/research/publication/a-flexible-new-technique-for-camera-calibration/. MSR-TR-98-71, Updated March 25, 1999.

A Procrustes distance

Here, we show how to explicitly compute the distance ρ (6). Since:

$$\|X_1 - QX_2\|_{\mathrm{F}}^2 = \|X_1\|_{\mathrm{F}}^2 + \|X_2\|_{\mathrm{F}}^2 - 2\langle X_1, QX_2\rangle,$$

finding the optimal Q is equivalent to solving:

$$\max_{Q\in\mathcal{O}(d)} \langle X_1, QX_1\rangle = \max_{Q\in\mathcal{O}(d)} \mathrm{Tr}(QX_2X_1^T).$$

By singular value decomposition we can write $X_1X_2^T = U\Sigma V^T$ and the solution of the problem is given by the polar factor of $X_1X_2^T$:

$$Q = UV^T. \tag{26}$$

B Impossibility of estimation for the rotations

Here, we give a proof of Proposition (1.1). We start by reminding the statement of the proposition below.

Proposition B.1. *Let us consider the following hypothesis testing problem on the distribution of the unknown distribution $\mathbb{P}$ of the sample:*

$$H_1 : \mathbb{P}_1 \sim \mathcal{N}(QX, \sigma^2 I_{dk}),$$
$$H_2 : \mathbb{P}_2 \sim \mathcal{N}(-QX, \sigma^2 I_{dk}).$$

Then for any test ψ and for any precision $\delta > 0$, it is possible to find a sufficiently large noise level σ_0 such that, for any noise level σ larger than σ_0, the sum of type I and type II errors is large:

$$\mathbb{P}_1(\psi(Y) = 2) + \mathbb{P}_2(\psi(Y) = 1) \geq 1 - 2\delta.$$

Hence if the two hypotheses are equally likely, the probability of error is at least $1/2 - \delta$.

Proof. The likelihood ratio test consists of studying the ration of the densities $L(Y) = \dfrac{f_1(Y)}{f_2(Y)}$ given observation Y. If $L(Y) > 1$, then H_1 is kept, otherwise H_2 is kept. Hence $\mathbb{P}_1(\psi^* = 2) = \mathbb{P}_1(L(Y) < 1)$.

$$\begin{aligned} L(Y) = \frac{f_1(Y)}{f_2(Y)} &= \frac{e^{-\frac{1}{2\sigma^2}\|Y-QX\|_{\mathrm{F}}^2}}{e^{-\frac{1}{2\sigma^2}\|Y+QX\|_{\mathrm{F}}^2}} \\ &= \frac{e^{-\frac{1}{2\sigma^2}(\|Y\|_{\mathrm{F}}^2+\|X\|_{\mathrm{F}}^2-2\langle Y,QX\rangle)}}{e^{-\frac{1}{2\sigma^2}(\|Y\|_{\mathrm{F}}^2+\|X\|_{\mathrm{F}}^2+2\langle Y,QX\rangle)}} \\ &= e^{\frac{2}{\sigma^2}\langle Y,QX\rangle} = e^{\frac{2}{\sigma^2}\langle QX+\sigma E,QX\rangle} \\ &= e^{\frac{2}{\sigma^2}(\|X\|_{\mathrm{F}}^2+\sigma\langle E,QX\rangle)}. \end{aligned}$$

This yields:

$$L(Y) < 1 \iff \log(L(Y)) < 0 \iff \|X\|_{\mathrm{F}}^2 + \sigma\langle E, QX\rangle < 0.$$

Yet,

$$\mathbb{P}(\|X\|_{\mathrm{F}}^2 + \sigma\langle E, QX\rangle < 0) = \mathbb{P}(\mu + \tilde{\sigma}\xi < 0),$$

with $\mu = \|X\|_{\mathrm{F}}^2$, $\tilde{\sigma} = \sigma\|X\|_{\mathrm{F}}$ and $\xi \sim \mathcal{N}(0,1)$. Hence:

$$\begin{aligned}
\mathbb{P}_1(\psi^* = 2) &= \mathbb{P}(\mu + \tilde{\sigma}\xi < 0) \\
&= \mathbb{P}(\xi < -\frac{\mu}{\tilde{\sigma}}) \\
&= \mathbb{P}\left(\xi < -\frac{\|X\|_{\mathrm{F}}}{\sigma}\right) \\
&= F\left(-\frac{\|X\|_{\mathrm{F}}}{\sigma}\right).
\end{aligned}$$

Where F is the cumulative density function of a standard Gaussian random variable. Similarly, we get:

$$\mathbb{P}_2\left(\psi^* = 1\right) = F\left(-\frac{\|X\|_{\mathrm{F}}}{\sigma}\right).$$

Hence, by Neyman–Pearson's lemma [Neyman and Pearson, 1933], it is sufficient to take σ such that $F\left(-\frac{\|X\|_{\mathrm{F}}}{\sigma}\right) = \frac{1-2\delta}{2}$ to get that for any test ψ:

$$\mathbb{P}_2(\psi = 1) + \mathbb{P}_1(\psi = 2) \geq 1 - 2\delta,$$

where F is the cumulative density function of a standard normal distribution. □

C Stability of the Cholesky decomposition

We give a proof of Proposition 3.2 which is used in the proof of Theorem (3.3) about the stability of estimator (1).

Proof. Let $X = U\Sigma V^T$ be the thin SVD of X. Here, $V \in \mathbb{R}^{k\times d}$ has orthonormal columns, U is an orthogonal matrix of size $d \times d$ and Σ is a diagonal matrix of size d with entries $\sigma_1 \geq \cdots \geq \sigma_d > 0$. It is always possible to pick $V_\perp$ (a complement of the orthonormal basis V) such that $\begin{bmatrix} V & V_\perp \end{bmatrix}$ is an orthogonal matrix of size $k \times k$. Hence, for any $\dot{X} \in \mathbb{R}^{d\times k}$, there exist matrices A, B of appropriate size such that

$$\dot{X} = UAV^T + UBV_\perp^T.$$

Using this parameterization,

$$\begin{aligned}\mathcal{L}_X(\dot{X}) &= V\Sigma(AV^T + BV_\perp^T) + (VA^T + V_\perp B^T)\Sigma V^T \\ &= \begin{bmatrix} V & V_\perp \end{bmatrix} \begin{bmatrix} \Sigma A + A^T\Sigma & \Sigma B \\ B^T\Sigma & 0 \end{bmatrix} \begin{bmatrix} V & V_\perp \end{bmatrix}^T.\end{aligned}$$

We use this to derive an SVD of $\mathcal{L}_X$ restricted to H_X, that is, for $\dot{X}$ such that $A\Sigma = \Sigma A^T$. To this end, consider the orthonormal basis of H_X composed of the following elements, and its image through $\mathcal{L}_X$ (below, e_i denotes the ith column of the identity matrix of appropriate dimension as indicated by context):

1. With $A = e_i e_i^T, B = 0$ for $1 \le i \le d$,
   $\mathcal{L}_X(\dot{X}) = 2\sigma_i \begin{bmatrix} V & V_\perp \end{bmatrix} \begin{bmatrix} e_i e_i^T & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V & V_\perp \end{bmatrix}^T;$

2. With $A = \frac{\sigma_i e_i e_j^T + \sigma_j e_j e_i^T}{\sqrt{\sigma_i^2 + \sigma_j^2}}, B = 0$ for $1 \le i < j \le d$,
   $\mathcal{L}_X(\dot{X}) = \sqrt{2(\sigma_i^2 + \sigma_j^2)} \begin{bmatrix} V & V_\perp \end{bmatrix} \begin{bmatrix} \frac{e_i e_j^T + e_j e_i^T}{\sqrt{2}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V & V_\perp \end{bmatrix}^T;$

3. With $A = 0, B = e_i e_j^T$ for $1 \le i \le d$ and $1 \le j \le k - d$,
   $\mathcal{L}_X(\dot{X}) = \sqrt{2}\sigma_i \begin{bmatrix} V & V_\perp \end{bmatrix} \begin{bmatrix} 0 & \frac{e_i e_j^T}{\sqrt{2}} \\ \frac{e_j e_i^T}{\sqrt{2}} & 0 \end{bmatrix} \begin{bmatrix} V & V_\perp \end{bmatrix}^T.$

To verify that the inputs yield an orthonormal basis of H_X as announced, check that each of these choices yields a matrix $\dot{X}$ in H_X; they are indeed orthonormal; and they are in sufficient number to cover $\dim \mathrm{H}_X = dk - \dim \mathrm{V}_X = dk - \frac{d(d-1)}{2}$. As the outputs of $\mathcal{L}_X$ applied to the basis elements are also orthogonal, it is clear that the singular values of $\mathcal{L}_X$ restricted to H_X are:

1. $2\sigma_1, \ldots, 2\sigma_d$;
2. $\sqrt{2(\sigma_i^2 + \sigma_d^2)}$ for $1 \le i < j \le d$; and
3. $\sqrt{2}\sigma_1, \ldots, \sqrt{2}\sigma_d$, each repeated $k - d$ times.

(And of course, the singular values of $\mathcal{L}_X$ on V_X are zero, $d(d-1)/2$ times.) □

D Minimax lower bound

D.1 Proof of Lemma 4.3

Proof. Let us consider G_ε, a minimal ε-cover of $\mathrm{G}(k,d)$ for the semi metric $\tilde{\rho}$, i.e. such that:

$$\mathrm{G}(k,d) = \bigcup_{U \in \mathrm{G}_\varepsilon} B(U,\varepsilon).$$

Then:

$$\mathfrak{N}(\mathrm{G}(k,d),\alpha\varepsilon) = \mathfrak{N}\left(\bigcup_{U\in G_\varepsilon} B(U,\varepsilon),\alpha\varepsilon\right) \leq \sum_{U\in\mathrm{G}_\varepsilon} \mathfrak{N}(B(U,\varepsilon),\alpha\varepsilon). \tag{27}$$

For contradiction, assume that for all U in G_ε we have

$$\mathfrak{N}(B(U,\varepsilon),\alpha\varepsilon) < \frac{\mathfrak{N}(\mathrm{G}(k,d),\alpha\varepsilon)}{\mathfrak{N}(\mathrm{G}(k,d),\varepsilon)} = \frac{\mathfrak{N}(\mathrm{G}(k,d),\alpha\varepsilon)}{|\mathrm{G}_\varepsilon|}.$$

This implies:

$$\mathfrak{N}(\mathrm{G}(k,d),\alpha\varepsilon) = \mathfrak{N}\left(\bigcup_{U\in G_\varepsilon} B(U,\varepsilon),\alpha\varepsilon\right) > \sum_{U\in\mathrm{G}_\varepsilon} \mathfrak{N}(B(U,\varepsilon),\alpha\varepsilon),$$

which contradicts (27). Hence, there exists $U^* \in \mathrm{G}_\varepsilon$ such that:

$$\mathfrak{N}(B(U^*,\varepsilon),\alpha\varepsilon) \geq \frac{\mathfrak{N}(\mathrm{G}(k,d),\alpha\varepsilon)}{\mathfrak{N}(\mathrm{G}(k,d),\varepsilon)} \geq \left(\frac{c_0}{\alpha c_1}\right)^{d(k-d)},$$

where the second inequality follows from Lemma 4.2. The fact that $\mathfrak{M}(E,\varepsilon) \geq \mathfrak{N}(E,\varepsilon)$ for any E allows to conclude the proof. □

D.2 Proof of Theorem 4.1

We will use the following lemma, originally given in [Tu et al., 2016].

Lemma D.1. *For any any $X_2 \in \mathbb{R}^{d\times k}$ obeying $\rho(X_2,X_1) \leq \frac{1}{4}\|X_1\|_{\mathrm{op}}$, we have:*

$$\|X_2^T X_2 - X_1^T X_1\|_{\mathrm{F}} \leq \frac{9}{4}\|X_1\|_{\mathrm{op}}\rho(X_2,X_1).$$

Proof. For all Q in $\mathcal{O}(d)$ we have

$$\begin{aligned}
\|X_2^T X_2 - X_1^T X_1\|_{\mathrm{F}} &= \|X_2^T X_2 - X_2^T Q X_1 + X_2^T Q X_1 - (QX_1)^T Q X_1\|_{\mathrm{F}} \\
&= \|X_2^T(X_2 - QX_1) + (X_2^T - (QX_1)^T)QX_1\|_{\mathrm{F}} \\
&\leq \left(\|X_2\|_{\mathrm{op}} + \|X_1\|_{\mathrm{op}}\right)\|X_2 - QX_1\|_{\mathrm{F}} \\
&\leq \frac{9}{4}\|X_1\|_{\mathrm{op}}\|X_2 - QX_1\|_{\mathrm{F}}.
\end{aligned}$$

Taking the infimum of the right-hand side over $\mathcal{O}(d)$ gives the result. □

Proof of Theorem 4.1. We restrict ourself to equivalence classes $[X]$ such that $XX^T = I_d$. There is a one-to-one mapping between such equivalence classes and elements of $\mathrm{G}(k,d)$, since X^TX is then the orthogonal projector to the space spanned by the columns of X^T (see [Absil et al., 2008, §3.4.4] for instance). We get that:

$$\begin{aligned}\inf_{\hat{X}} \sup_{X\in\mathcal{X}} \mathbb{E}[\rho^2(X,\hat{X})] &\geq \inf_{\hat{X}} \sup_{X\in\mathrm{G}(k,d)} \mathbb{E}[\rho^2(X,\hat{X})] \\ \text{(by Lemma D.1)} &\geq \left(\frac{4}{9\|X\|_{\mathrm{op}}}\right)^2 \inf_{\hat{X}} \sup_{X\in\mathrm{G}(k,d)} \mathbb{E}[\tilde{\rho}^2(X,\hat{X})] \\ &\geq \frac{16}{81} \inf_{\hat{X}} \sup_{X\in\mathrm{G}(k,d)} \mathbb{E}[\tilde{\rho}^2(X,\hat{X})],\end{aligned}$$

where $\tilde{\rho}$ is the metric defined in Lemma 4.2. Lemma 4.3, allows us to consider a point U^* and m points $\{U_1,\dots,U_m\} \subset B(U^*,\varepsilon)$, such that $\alpha\varepsilon \leq \tilde{\rho}(U_i,U_j) \leq 2\varepsilon$ for all $i \neq j$ and $m \geq \left(\frac{c_0}{\alpha c_1}\right)^{d(k-d)}$. Without loss of generality, we can assume that $\varepsilon \leq \frac{1}{2}$. By standard arguments (see for instance chapter 4 of [Rigollet and Hütter, 2017]) we can lower bound the minimax risk by the probability of error for testing a finite number of hypothesis:

$$\inf_{\hat{X}} \sup_{X\in\mathrm{G}(k,d)} \mathbb{E}[\tilde{\rho}^2(X,\hat{X})] \geq (\alpha\varepsilon)^2 \inf_{\psi} \max_{1\leq j\leq m} \mathbb{P}_{U_j}(\psi \neq j).$$

Moreover, by Fano's inequality [Fano, 1952] (see Theorem 4.19 in [Rigollet and Hütter, 2017]) we can lower bound the probability of error for the multiple hypothesis testing problem $\{\mathbb{P}_{U_i} : i \in [m]\}$ by:

$$\inf_{\psi} \max_{1\leq j\leq m} \mathbb{P}_{U_j}(\psi \neq j) \geq 1 - \frac{\min_{i\neq j} \mathrm{KL}(\mathbb{P}_{U_i}||\mathbb{P}_{U_j}) + \log(2)}{\log(m)}.$$

By Lemma 4.4, for all pair i,j we have that:

$$\mathrm{KL}(\mathbb{P}_{U_i}||\mathbb{P}_{U_j}) \leq C\sigma^{-4}\rho^2(U_i,U_j),$$

where by hypothesis $\tilde{\rho}(U_i,U_j) \leq 2\varepsilon$. Theorem 3.3 implies that for small enough ε, $\rho(U_i,U_j) \leq \sqrt{2}\varepsilon$. Hence for small enough ε and large enough σ we have that:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{U_i}||\mathbb{P}_{U_j}) &\leq C\sigma^{-4}\rho^2(U_i,U_j) \\ &\leq C\sigma^{-4}\left(\frac{1}{\sqrt{2}\sigma_d(X)}\right)^2 \tilde{\rho}^2(U_i,U_j) \\ &\leq C'\sigma^{-4}\varepsilon^2.\end{aligned}$$

Hence, the minimax error admits the following lower bound:

$$\inf_{\hat{X}} \sup_{X\in\mathcal{X}} \mathbb{E}[\rho^2(X,\hat{X})] \geq \frac{16}{81}\alpha^2\varepsilon^2\left(1 - \frac{2NC'\sigma^{-4}\varepsilon^2 + \log(2)}{d(k-d)\log(\frac{c_0}{\alpha c_1})}\right),$$

for any $\varepsilon \in (0, \varepsilon_0]$ and $\alpha \in (0,1)$. By picking

$$\alpha = \frac{c_0}{4c_1}, \qquad \text{and} \qquad \varepsilon^2 = \frac{\sigma^4}{N} \cdot \frac{d(k-d)\log(2)}{6C'}$$

if $\varepsilon_0^2 \geq \frac{\sigma^4}{N} \cdot \frac{d(k-d)\log(2)}{6C'}$, and $\varepsilon^2 = \frac{\varepsilon_0^2}{2}$ otherwise, we get the following inequality:

$$\inf_{\hat{X}} \sup_{X \in \mathcal{X}} \mathbb{E}[\rho^2(X, \hat{X})] \geq \left(\frac{4c_0}{9c_1}\right)^2 \min\left(\frac{\sigma^4}{N}\frac{d(k-d)\log(2)}{576C'}, \frac{\varepsilon_0^2}{96}\right),$$

which yields the lower bound. The upper bound directly follows from the upper bound on the MSE of the estimator, which can be found in Appendix F.2 . □

E Bound on the KL divergence

In this section, we prove the tight bound on the KL divergence of the distribution of samples (1) stated in Lemma 4.4. Our result is a direct consequence of Proposition 7.8 in [Bandeira et al., 2017b].

Proof. By the properties of the orthogonal group and the Haar measure, we have:

$$\mathbb{E}_Q[QX_1] = \mathbb{E}_Q[QX_2] = 0.$$

We now bound KL divergence between $\mathbb{P}_{X_1}$ and $\mathbb{P}_{X_2}$ by bounding the χ^2-divergence between $\mathbb{P}_{X_1}$ and $\mathbb{P}_{X_2}$. The density of $\mathbb{P}_X$, f_X can be written as:

$$f_X(Y) = \mathbb{E}_Q[\sigma^{-dk} f(\sigma^{-1}(Y - QX))] = \sigma^{-dk} f(\sigma^{-1}Y)\mathbb{E}_Q[e^{-\frac{1}{2\sigma^2}(\|X\|_{\mathrm{F}}^2 - 2\langle Y, QX\rangle)}],$$

where f is the density of a standard dk-dimensional Gaussian. Since $\|X\|_{\mathrm{F}}^2 \leq d$ by hypothesis, we obtain by Jensen's inequality:

$$f_X(Y) \geq \sigma^{-dk} f(\sigma^{-1}Y) e^{-\frac{1}{2\sigma^2}(d - 2\mathbb{E}_Q[\langle Y, QX\rangle])} = \sigma^{-dk} f(\sigma^{-1}Y) e^{-\frac{d}{2\sigma^2}}. \tag{28}$$

Hence the χ^2 divergence can then be bounded as:

$$\begin{aligned}
\chi^2(\mathbb{P}_{X_1}, \mathbb{P}_{X_2}) &= \int \frac{(f_{X_1}(Y) - f_{X_2}(Y))^2}{f_{X_1}(Y)} dY \\
&\leq e^{\frac{d}{2\sigma^2}} \int \left(\mathbb{E}_{Q_1}\left[e^{-\frac{\|X_1\|_{\mathrm{F}}^2 - 2\langle Y, Q_1 X_1\rangle}{2\sigma^2}} \right] \right. \\
&\qquad \left. - \mathbb{E}_{Q_2}\left[e^{-\frac{\|X_2\|_{\mathrm{F}}^2 - 2\langle Y, Q_2 X_2\rangle}{2\sigma^2}} \right] \right)^2 \sigma^{-dk} f(\sigma^{-1} Y) dY \\
&= e^{\frac{d}{2\sigma^2}} \int \left(\mathbb{E}_{Q_1}\left[e^{\langle Y, \sigma^{-1} Q_1 X_1\rangle - \frac{1}{2}\|\sigma^{-1} X_1\|_{\mathrm{F}}^2} \right] \right. \\
&\qquad \left. - \mathbb{E}_{Q_2}\left[e^{\langle Y, \sigma^{-1} Q_2 X_2\rangle - \frac{1}{2}\|\sigma^{-1} X_2\|_{\mathrm{F}}^2} \right] \right)^2 f(Y) dY \\
&= e^{\frac{d}{2\sigma^2}} \mathbb{E}_E \left[\left(\mathbb{E}_{Q_1}\left[e^{\langle E, \sigma^{-1} Q_1 X_1\rangle - \frac{1}{2}\|\sigma^{-1} X_1\|_{\mathrm{F}}^2} \right] \right. \right. \\
&\qquad \left. \left. - \mathbb{E}_{Q_2}\left[e^{\langle E, \sigma^{-1} Q_2 X_2\rangle - \frac{1}{2}\|\sigma^{-1} X_2\|_{\mathrm{F}}^2} \right] \right)^2 \right],
\end{aligned}$$

where E is a $d \times k$ matrix with i.i.d. entries $\sim \mathcal{N}(0,1)$. The first inequality comes from applying inequality (28) to the denominator, and the second equality comes from a change of variables. By Fubini's theorem we can change the order of expectations. Using the fact that for a standard Gaussian U, for any λ, we have $\mathbb{E}[e^{\lambda U}] = e^{\frac{\lambda^2}{2}}$, we obtain that:

$$\mathbb{E}_{Q_1,Q_2}\mathbb{E}_E\left[e^{\langle E, \sigma^{-1}(Q_1 X_1 + Q_2 X_2)\rangle - \frac{1}{2}(\|\sigma^{-1} X_1\|_{\mathrm{F}}^2 + \|\sigma^{-1} X_2\|_{\mathrm{F}}^2)} \right] = \mathbb{E}_{Q_1,Q_2}\left[e^{\frac{\langle Q_1 X_1, Q_2 X_2\rangle}{\sigma^2}} \right].$$

We use this to expand the square in the above expression for $\chi^2(\mathbb{P}_{X_1}, \mathbb{P}_{X_2})$:

$$\begin{aligned}
\chi^2(\mathbb{P}_{X_1}, \mathbb{P}_{X_2}) &\leq e^{\frac{d}{2\sigma^2}} \left(\mathbb{E}_{Q_1,\tilde{Q}_1}\left[e^{\frac{1}{\sigma^2}\langle Q_1 X_1, \tilde{Q}_1 X_1\rangle} \right] - 2\mathbb{E}_{Q_1,Q_2}\left[e^{\frac{1}{\sigma^2}\langle Q_1 X_1, Q_2 X_2\rangle} \right] \right. \\
&\qquad \left. + \mathbb{E}_{Q_2,\tilde{Q}_2}\left[e^{\frac{1}{\sigma^2}\langle Q_2 X_2, \tilde{Q}_2 X_2\rangle} \right] \right) \\
&= e^{\frac{d}{2\sigma^2}} \left(\mathbb{E}_{Q_1,\tilde{Q}_1}\left[e^{\frac{1}{\sigma^2}\langle X_1, Q_1^T \tilde{Q}_1 X_1\rangle} \right] - 2\mathbb{E}_{Q_1,Q_2}\left[e^{\frac{1}{\sigma^2}\langle X_1, Q_1^T Q_2 X_2\rangle} \right] \right. \\
&\qquad \left. + \mathbb{E}_{Q_2,\tilde{Q}_2}\left[e^{\frac{1}{\sigma^2}\langle X_2, Q_2^T \tilde{Q}_2 X_2\rangle} \right] \right) \\
&= e^{\frac{d}{2\sigma^2}} \mathbb{E}_Q\left[e^{\frac{1}{\sigma^2}\langle X_1, Q X_1\rangle} - 2e^{\frac{1}{\sigma^2}\langle X_1, Q X_2\rangle} + e^{\frac{1}{\sigma^2}\langle X_2, Q X_2\rangle} \right].
\end{aligned}$$

Since Q_1,$\tilde{Q}_1$, Q_2 and $\tilde{Q}_2$ are drawn according to the Haar measure on the orthogonal group, according to Fubini's theorem we can expand each term as a power series and exchange summation and expectation. Using the fact that $\langle u, v\rangle^l = \langle u^{\otimes l}, v^{\otimes l}\rangle$ and that

for all i.i.d. vectors x and y we have that

$$\mathbb{E}_{x,y}[\langle x,y\rangle^l] = \mathbb{E}_{x,y}[\langle x^{\otimes l}, y^{\otimes l}\rangle] = \langle \mathbb{E}_x[x^{\otimes l}], \mathbb{E}_y[y^{\otimes l}]\rangle = \|\mathbb{E}_x[x^{\otimes l}]\|_{\mathrm{F}}^2,$$

we obtain :

$$\begin{aligned}
\chi^2(\mathbb{P}_{X_1}, \mathbb{P}_{X_2}) &\leq e^{\frac{d}{2\sigma^2}} \sum_{l=0}^{\infty} \frac{\sigma^{-2l}}{l!} \Big(\mathbb{E}_Q[\langle X_1, QX_1\rangle^l - 2\langle X_1, QX_2\rangle^l + \langle X_2, QX_2\rangle^l] \Big) \\
&= e^{\frac{d}{2\sigma^2}} \sum_{l=0}^{\infty} \frac{\sigma^{-2l}}{l!} \Big(\mathbb{E}_Q[\langle \mathrm{vec}(X_1), \mathrm{vec}(QX_1)\rangle^l - 2\langle \mathrm{vec}(X_1), \mathrm{vec}(QX_2)\rangle^l \\
&\qquad + \langle \mathrm{vec}(X_2), \mathrm{vec}(QX_2)\rangle^l] \Big) \\
&= e^{\frac{d}{2\sigma^2}} \sum_{l=0}^{\infty} \frac{\sigma^{-2l}}{l!} \Big(\mathbb{E}_{Q,\tilde{Q}}[\langle \mathrm{vec}(\tilde{Q}X_1), \mathrm{vec}(QX_1)\rangle^l - 2\langle \mathrm{vec}(\tilde{Q}X_1), \mathrm{vec}(QX_2)\rangle^l \\
&\qquad + \langle \mathrm{vec}(\tilde{Q}X_2), \mathrm{vec}(QX_2)\rangle^l] \Big) \\
&= e^{\frac{d}{2\sigma^2}} \sum_{l=0}^{\infty} \frac{\sigma^{-2l}}{l!} \Big(\mathbb{E}_{Q,\tilde{Q}}[\langle \mathrm{vec}(\tilde{Q}X_1)^{\otimes l}, \mathrm{vec}(QX_1)^{\otimes l}\rangle \\
&\qquad - 2\langle \mathrm{vec}(\tilde{Q}X_1)^{\otimes l}, \mathrm{vec}(QX_2)^{\otimes l}\rangle + \langle \mathrm{vec}(\tilde{Q}X_2)^{\otimes l}, \mathrm{vec}(QX_2)^{\otimes l}\rangle] \Big) \\
&= e^{\frac{d}{2\sigma^2}} \sum_{l=0}^{\infty} \frac{\sigma^{-2l}}{l!} \Big(\langle \mathbb{E}_{\tilde{Q}}[\mathrm{vec}(\tilde{Q}X_1)^{\otimes l}], \mathbb{E}_Q[\mathrm{vec}(QX_1)^{\otimes l}]\rangle \\
&\qquad - 2\langle \mathbb{E}_{\tilde{Q}}[\mathrm{vec}(\tilde{Q}X_1)^{\otimes l}], \mathbb{E}_Q[\mathrm{vec}(QX_2)^{\otimes l}]\rangle \\
&\qquad + \langle \mathbb{E}_{\tilde{Q}}[\mathrm{vec}(\tilde{Q}X_2)^{\otimes l}], \mathbb{E}_Q[\mathrm{vec}(QX_2)^{\otimes l}]\rangle \Big) \\
&= e^{\frac{d}{2\sigma^2}} \sum_{l=0}^{\infty} \frac{\sigma^{-2l}}{l!} \Big(\|\mathbb{E}_{\tilde{Q}}[\mathrm{vec}(\tilde{Q}X_1)^{\otimes l}]\|_{\mathrm{F}}^2 \\
&\qquad - 2\langle \mathbb{E}_{\tilde{Q}}[\mathrm{vec}(\tilde{Q}X_1)^{\otimes l}], \mathbb{E}_{\tilde{Q}}[\mathrm{vec}(\tilde{Q}X_2)^{\otimes l}]\rangle \\
&\qquad + \|\mathbb{E}_{\tilde{Q}}[\mathrm{vec}(\tilde{Q}X_2)^{\otimes l}]\|_{\mathrm{F}}^2 \Big) \\
&= e^{\frac{d}{2\sigma^2}} \sum_{l=0}^{\infty} \frac{\sigma^{-2l}}{l!} \|\mathbb{E}_{\tilde{Q}}[\mathrm{vec}(\tilde{Q}X_1)^{\otimes l}] - \mathbb{E}_{\tilde{Q}}[\mathrm{vec}(\tilde{Q}X_2)^{\otimes l}]\|_{\mathrm{F}}^2 \\
&= e^{\frac{d}{2\sigma^2}} \sum_{l=0}^{\infty} \frac{\sigma^{-2l}}{l!} \|\Delta_l\|_{\mathrm{F}}^2 \\
&= e^{\frac{d}{2\sigma^2}} \Big(\sigma^{-2}\|\Delta_1\|_{\mathrm{F}}^2 + \frac{\sigma^{-4}}{2!}\|\Delta_2\|_{\mathrm{F}}^2 + \sum_{l=3}^{\infty} \frac{\sigma^{-2l}}{l!}\|\Delta_l\|_{\mathrm{F}}^2 \Big),
\end{aligned}$$

with $\Delta_l = \mathbb{E}_Q[\mathrm{vec}(QX_1)^{\otimes l} - \mathrm{vec}(QX_2)^{\otimes l}]$, where the expectation is taken over the Haar measure. In particular,

$$\Delta_1 = \mathbb{E}_Q[\mathrm{vec}(QX_1) - \mathrm{vec}(QX_2)] = \mathrm{vec}(\mathbb{E}_Q[Q(X_1 - X_2)]) = 0.$$

Moreover, by Lemma E.1 below, for all $l \geq 2$ we have:

$$\|\Delta_l\|_{\mathrm{F}}^2 \leq 12(2d)^l \rho^2(X_1, X_2).$$

Hence, for any $\sigma > 1$ we obtain:

$$\begin{aligned}
\chi^2(\mathbb{P}_{X_1}, \mathbb{P}_{X_2}) &\leq e^{\frac{d}{2\sigma^2}} \left(\frac{\sigma^{-4}}{2} \|\Delta_2\|_{\mathrm{F}}^2 + \sum_{l=3}^{\infty} \frac{\sigma^{-2l}}{l!} 12 \cdot (2d)^l \rho^2(X_1, X_2) \right) \\
&\leq e^{\frac{d}{2\sigma^2}} \left(c_1 \rho^2(X_1, X_2) \sigma^{-4} + c_2 \rho^2(X_1, X_2) \sigma^{-4} \right) \\
&\leq C \rho^2(X_1, X_2) \sigma^{-4}.
\end{aligned}$$

The inequality $\mathrm{KL}(\mathbb{P}_{X_1} || \mathbb{P}_{X_2}) \leq \chi^2(\mathbb{P}_{X_1}, \mathbb{P}_{X_2})$ (see Lemma 2.2 in [Tsybakov, 2009] for instance) finishes the proof of the claim. □

The general structure of the proof of the lemma below is as in Lemma B.12 of [Bandeira et al., 2017b], with adaptations as needed.

Lemma E.1. *Given $X_1, X_2 \in \mathbb{R}^{d \times k}$ such that $\rho(X_1, X_2) < \dfrac{\|X_1\|_{\mathrm{F}}}{3}$, for all $l \geq 1$,*

$$\|\Delta_l\|_{\mathrm{F}}^2 = \|\mathbb{E}_Q[\mathrm{vec}(QX_1)^{\otimes l} - \mathrm{vec}(QX_2)^{\otimes l}]\|_{\mathrm{F}}^2 \leq 12(2d)^l \rho^2(X_1, X_2).$$

Proof. Without loss of generality we assume that X_1 and X_2 are rotationally aligned, i.e., $\rho(X_1, X_2) = \|X_1 - X_2\|_{\mathrm{F}} = \varepsilon \|X_1\|_{\mathrm{F}}$, with $\varepsilon < \dfrac{1}{3}$. By Jensen's inequality,

$$\begin{aligned}
\|\mathbb{E}_Q[\mathrm{vec}(QX_1)^{\otimes l} - \mathrm{vec}(QX_2)^{\otimes l}]\|_{\mathrm{F}}^2 &\leq \mathbb{E}_Q[\|\mathrm{vec}(QX_1)^{\otimes l} - \mathrm{vec}(QX_2)^{\otimes l}\|_{\mathrm{F}}^2] \\
&\leq \|\mathrm{vec}(X_1)^{\otimes l} - \mathrm{vec}(X_2)^{\otimes l}\|_{\mathrm{F}}^2.
\end{aligned}$$

Expanding the norm, it is easy to check that

$$\begin{aligned}
\|\mathrm{vec}(X_1)^{\otimes l} - \mathrm{vec}(X_2)^{\otimes l}\|_{\mathrm{F}}^2 &= \|\mathrm{vec}(X_1)\|_{\mathrm{F}}^{2l} - 2\langle X_1, X_2 \rangle^l + \|\mathrm{vec}(X_2)\|_{\mathrm{F}}^{2l} \\
&= \|X_1\|_{\mathrm{F}}^{2l} \left(1 - 2(1+\gamma)^l + (1 + 2\gamma + \varepsilon^2)^l \right),
\end{aligned}$$

where $\gamma = \dfrac{\langle X_1, X_2 - X_1 \rangle}{\|X_1\|_{\mathrm{F}}^2}$. By Cauchy–Schwarz we have $|\gamma| \leq \varepsilon < \dfrac{1}{3}$. Moreover, $2\gamma + \varepsilon^2 \leq 3\varepsilon < 1$. By the binomial theorem, for all x such that $|x| < 1$, there exists an $r_l \leq 2^l x^2$ such that

$$(1+x)^l = \sum_{k=0}^{l} \binom{l}{k} x^k = 1 + lx + r_l.$$

Hence,

$$1-2(1+\gamma)^l+(1+2\gamma+\varepsilon^2)^l \leq 1-2-2l\gamma+2^{l+1}\varepsilon^2+1+2l\gamma+l\varepsilon^2+2^l\cdot 9\varepsilon^2$$
$$\leq (l+11\cdot 2^l)\varepsilon^2 \leq 12\cdot 2^l\varepsilon^2,$$

and $\|\Delta_l\|_{\mathrm{F}}^2 \leq \|X_1\|_{\mathrm{F}}^{2l}12(2)^l\varepsilon^2 = \|X_1\|_{\mathrm{F}}^{2l-2}12(2)^l(\|X_1\|_{\mathrm{F}}\varepsilon)^2 \leq 12(2d)^l\rho^2(X_1,X_2)$. □

F Statistical properties of the Estimator

We now compute an upper bound on the MSE of our estimator in order to prove the second part of Proposition 4.6. We start by computing the covariance matrix of the estimator in part F.1, before proving the upper bound in part F.2.

F.1 Covariance matrix of the estimator $\hat{M}_N$

It is straightforward to get that

$$\mathbb{E}[\hat{M}_N] = X^TX + \sigma^2 dI_k.$$

We now compute the covariance matrix Σ of $\hat{M}_N$ i.e.

$$\Sigma = \mathbb{E}[\mathrm{vec}(\hat{M}_N-\mathbb{E}[\hat{M}_N])\mathrm{vec}(\hat{M}_N-\mathbb{E}[\hat{M}_N])^T] \tag{29}$$
$$= \mathbb{E}[\mathrm{vec}(\hat{M}_N)\mathrm{vec}(\hat{M}_N)^T] - \mathrm{vec}(\mathbb{E}[\hat{M}_N])\mathrm{vec}(\mathbb{E}[\hat{M}_N])^T. \tag{30}$$

For that purpose, we define F_N as

$$F_N = \hat{M}_N - \mathbb{E}[\hat{M}_N]$$
$$= \sigma\left(X^T\left(\frac{1}{N}\sum_{i=1}^N E_i\right) + \left(\frac{1}{N}\sum_{i=1}^N E_i\right)^T X\right) + \sigma^2\left(\frac{1}{N}\sum_{i=1}^N E_i^TE_i - dI_k\right).$$

Each of the entry of F_N can be written as

$$(F_N)_{s,t} = \sigma\left(\langle x_s,h_t\rangle + \langle x_t,h_s\rangle\right) + \sigma^2\left(\frac{1}{N}\sum_{i=1}^N\langle e_s^{(i)},e_t^{(i)}\rangle - d\delta_{s,t}\right),$$

where $X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_k \\ | & | & & | \end{bmatrix}$, $E_i = \begin{bmatrix} | & | & & | \\ e_1^{(i)} & e_2^{(i)} & \dots & e_k^{(i)} \\ | & | & & | \end{bmatrix}$ and $H_N = \frac{1}{N}\sum_{i=1}^N E_i = \begin{bmatrix} | & | & & | \\ h_1 & h_2 & \dots & h_k \\ | & | & & | \end{bmatrix}$. Since the entries of F_N have zero mean, to

compute their variance we start by computing

$$
\begin{aligned}
(F_N)_{s,t}(F_N)_{u,v} = \sigma^2 \left(\langle x_s, h_t\rangle\langle x_u, h_v\rangle + \langle x_t, h_s\rangle\langle x_u, h_v\rangle + \langle x_s, h_t\rangle\langle x_v, h_u\rangle + \langle x_t, h_s\rangle\langle x_v, h_u\rangle\right) \\
+ \sigma^3 \left((\langle x_s, h_t\rangle + \langle x_t, h_s\rangle (\frac{1}{N}\sum_{i=1}^{N}\langle e_u^{(i)}, e_v^{(i)}\rangle - d\delta_{u,v})) \right) \\
+ \sigma^3 \left((\langle x_u, h_v\rangle + \langle x_v, h_u\rangle (\frac{1}{N}\sum_{i=1}^{N}\langle e_s^{(i)}, e_t^{(i)}\rangle - d\delta_{s,t})) \right) \\
+ \sigma^4 \left(\frac{1}{N^2}\sum_{i,j}\langle e_s^{(i)}, e_t^{(i)}\rangle\langle e_u^{(i)}, e_v^{(i)}\rangle + d^2\delta_{s,t}\delta_{u,v} \right) \\
- \sigma^4 \left(d\frac{1}{N}\sum_{i}[\langle e_s^{(i)}, e_t^{(i)}\rangle\delta_{u,v} + \langle e_u^{(i)}, e_v^{(i)}\rangle\delta_{s,t}] \right).
\end{aligned}
$$

Hence, taking the expectation we get

$$
\begin{aligned}
\mathbb{E}[(F_N)_{s,t}(F_N)_{u,v}] = \frac{\sigma^2}{N}\left(\delta_{t,v}G_{s,u} + \delta_{s,v}G_{t,u} + \delta_{t,u}G_{s,v} + \delta_{s,u}G_{t,v}\right) \\
+ \sigma^4 d^2\delta_{s,t}\delta_{u,v} - \frac{\sigma^4 d^2}{N}\sum_{i=1}^{N}[\delta_{s,t}\delta_{u,v} + \delta_{u,v}\delta_{s,t}] \\
+ \frac{\sigma^4}{N^2}\sum_{i,j}\left[(1-\delta_{i,j})(d^2\delta_{s,t}\delta_{u,v}) + \delta_{i,j}\mathbb{E}[\langle e_s^{(i)}, e_t^{(i)}\rangle\langle e_u^{(i)}, e_v^{(i)}\rangle]\right].
\end{aligned}
$$

Equivalently,

$$
\begin{aligned}
\mathbb{E}[(F_N)_{s,t}(F_N)_{u,v}] = \frac{\sigma^2}{N}\left(\delta_{t,v}G_{s,u} + \delta_{s,v}G_{t,u} + \delta_{t,u}G_{s,v} + \delta_{s,u}G_{t,v}\right) \\
+ \sigma^4 d^2\delta_{s,t}\delta_{u,v}\left(1 - \frac{2N}{N} + \frac{N^2-N}{N^2}\right) + \frac{\sigma^4}{N}A_{s,t,u,v},
\end{aligned}
$$

where $E = \begin{bmatrix} | & | & & | \\ e_1 & e_2 & \dots & e_k \\ | & | & & | \end{bmatrix}$ is a $d \times k$ random matrix with i.i.d. Gaussian coefficients and $A_{s,t,u,v} = \mathbb{E}[\langle e_s, e_t\rangle\langle e_u, e_v\rangle]$. By expanding $A_{s,t,u,v}$, we can write

$$\begin{aligned} A_{s,t,u,v} &= \mathbb{E}\left[\sum_{\alpha} e_{\alpha,s}e_{\alpha,t}\sum_{\beta} e_{\beta,u}e_{\beta,v}\right] \\ &= \sum_{\alpha,\beta}\mathbb{E}\left[e_{\alpha,s}e_{\alpha,t}e_{\beta,u}e_{\beta,v}\right] \\ &= \sum_{\alpha}\mathbb{E}\left[e_{\alpha,s}e_{\alpha,t}e_{\alpha,u}e_{\alpha,v}\right] + \sum_{\alpha\neq\beta}\mathbb{E}\left[e_{\alpha,s}e_{\alpha,t}\right]\mathbb{E}[e_{\beta,u}e_{\beta,v}\right] \\ &= d\cdot\mathbb{E}\left[z_s z_t z_u z_v\right] + (d^2-d)\delta_{s,t}\delta_{u,v}. \end{aligned}$$

In order to compute $\mathbb{E}\left[z_s z_t z_u z_v\right]$, we distinguish seven cases below:

$$\begin{aligned} s=t=u=v &: \mathbb{E}[z_s z_t z_u z_v] = 3 \\ s=t\neq u=v &: \mathbb{E}[z_s z_t z_u z_v] = 1 \\ s=t \text{ and } u\neq v &: \mathbb{E}[z_s z_t z_u z_v] = 0 \\ s=u\neq t=v &: \mathbb{E}[z_s z_t z_u z_v] = 1 \\ s=u \text{ and } t\neq v &: \mathbb{E}[z_s z_t z_u z_v] = 0 \\ s=v\neq t=u &: \mathbb{E}[z_s z_t z_u z_v] = 1 \\ s=v \text{ and } t\neq u &: \mathbb{E}[z_s z_t z_u z_v] = 0. \end{aligned}$$

In other words,

$$\mathbb{E}[z_s z_t z_u z_v] = \delta_{s,t}\delta_{u,v} + \delta_{s,u}\delta_{t,v} + \delta_{s,v}\delta_{t,u} = \delta_{s,u}\delta_{t,v} + \delta_{s,v}\delta_{t,u}.$$

We therefore get that the entry of matrix Σ defined in (29) corresponding to the covariance between elements (s,t) and (u,v) of $\hat{M}_N$ is

$$\begin{aligned} \Sigma_{(s,t),(u,v)} &= \mathbb{E}[(F_N)_{s,t}(F_N)_{u,v}] \\ &= \frac{1}{N}\left[\sigma^2(\delta_{t,v}G_{s,u} + \delta_{s,v}G_{t,u} + \delta_{s,u}G_{t,v}) + \sigma^4 d(\delta_{s,u}\delta_{t,v} + \delta_{s,v}\delta_{t,u})\right], \end{aligned}$$

where $G = X^T X$.

F.2 Mean Squared Error of $\hat{G}_N$

Using Lemma 3.1 for example [Tu et al., 2016, Lem. 5.4], we find

$$\rho^2(X,\hat{X}) \leq \frac{1}{2(\sqrt{2}-1)\sigma_d^2(X)}\|G - \tilde{G}_N\|_{\mathrm{F}}^2,$$

where $\tilde{G}_N = \underset{H \succeq 0 : \mathrm{rank}(H) \leq d}{\mathrm{argmin}} \|\hat{G}_N - H\|_{\mathrm{F}}$ is the projection of $\hat{G}_N$ on the cone of positive semidefinite matrices of rank at most d. Since $\mathrm{rank}(G) = d$, $\tilde{G}_N$ satisfies

$$\|\hat{G}_N - \tilde{G}_N\|_{\mathrm{F}} \leq \|\hat{G}_N - G\|_{\mathrm{F}}.$$

By triangle inequality we have

$$\|G - \tilde{G}_N\|_{\mathrm{F}} \leq \|G - \hat{G}_N\|_{\mathrm{F}} + \|\hat{G}_N - \tilde{G}_N\|_{\mathrm{F}} \leq 2\|G - \hat{G}_N\|_{\mathrm{F}}.$$

Hence

$$\rho^2(X, \hat{X}) \leq \frac{2}{(\sqrt{2}-1)\sigma_d^2(X)} \|G - \hat{G}_N\|_{\mathrm{F}}^2. \tag{31}$$

Building on the computations from subsection F.1, we get that

$$\begin{aligned}
\mathbb{E}[\|G - \hat{G}_N\|_{\mathrm{F}}^2] &= \mathbb{E}[\|F_N\|_{\mathrm{F}}^2] \\
&= \sum_{s=1}^{k} \sum_{t=1}^{k} \mathbb{E}[(F_N)_{s,t}^2] \\
&= \frac{\sigma^2}{N} \left[\sum_{s} (4G_{s,s} + 2\sigma^2 d) + \sum_{s \neq t} (G_{s,s} + G_{t,t} + \sigma^2 d) \right] \\
&= \frac{\sigma^2}{N} \left[k(k+1)\sigma^2 d + 4\mathrm{Tr}(G) + 2(k-1)\mathrm{Tr}(G) \right] \\
&= \frac{(k+1)\sigma^2}{N} \left[k\sigma^2 d + \|X\|_{\mathrm{F}}^2 \right].
\end{aligned} \tag{32}$$

Finally, (31) and (32) imply that

$$\mathbb{E}[\rho^2(X, \hat{X})] \leq \frac{2(k+1)\sigma^2}{N(\sqrt{2}-1)\sigma_d^2(X)} \left[k\sigma^2 d + \|X\|_{\mathrm{F}}^2 \right] = O\left(\frac{\sigma^2 + \sigma^4}{N} \right).$$

This directly implies that, in the case where $\sigma \ll 1$, we have $\mathbb{E}[\rho^2(X, \hat{X})] = O\left(\frac{\sigma^2}{N}\right)$, while in the case where $\sigma \gg 1$, we have $\mathbb{E}[\rho^2(X, \hat{X})] = O\left(\frac{\sigma^4}{N}\right)$.