Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2012-029
CBCL-310

September 8, 2012

Multiscale Geometric Methods for Data Sets I: Multiscale SVD, Noise and Curvature

Anna V. Little, Mauro Maggioni, and Lorenzo Rosasco

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Multiscale Geometric Methods for Data Sets I: Multiscale SVD, Noise and Curvature

Anna V. Little[1], Mauro Maggioni[1,2], Lorenzo Rosasco[3]

[1]*Department of Mathematics and* [2]*Computer Science, Duke University*

[3] *Laboratory for Computational and Statistical Learning, Massachusetts Institute of Technology and Istituto Italiano di Tecnologia*

avl@math.duke.edu,mauro.maggioni@duke.edu, lrosasco@mit.edu

September 5, 2012

Abstract

Large data sets are often modeled as being noisy samples from probability distributions μ in $\mathbb{R}^D$, with D large. It has been noticed that oftentimes the support $\mathcal{M}$ of these probability distributions seems to be well-approximated by low-dimensional sets, perhaps even by manifolds. We shall consider sets that are locally well approximated by k-dimensional planes, with $k \ll D$, with k-dimensional manifolds isometrically embedded in $\mathbb{R}^D$ being a special case. Samples from μ are furthermore corrupted by D-dimensional noise. Certain tools from multiscale geometric measure theory and harmonic analysis seem well-suited to be adapted to the study of samples from such probability distributions, in order to yield quantitative geometric information about them. In this paper we introduce and study multiscale covariance matrices, i.e. covariances corresponding to the distribution restricted to a ball of radius r, with a fixed center and varying r, and under rather general geometric assumptions we study how their empirical, noisy counterparts behave. We prove that in the range of scales where these covariance matrices are most informative, the empirical, noisy covariances are close to their expected, noiseless counterparts. In fact, this is true as soon as the number of samples in the balls where the covariance matrices are computed is linear in the intrinsic dimension of $\mathcal{M}$. As an application, we present an algorithm for estimating the intrinsic dimension of $\mathcal{M}$.

1 Introduction

We are interested in developing tools for the quantitative analysis of the geometry of samples from a probability distribution in a high-dimensional Euclidean space, which is approximately supported on a low-dimensional set, and is corrupted by high-dimensional noise. Our main motivation arises from the need to analyze large, high dimensional data sets arising in a wide variety of applications. These data sets are often modeled as samples from a probability measure μ concentrated on or around a low-dimensional set embedded in high dimensional space (see for example [1, 2, 3, 4, 5, 6, 7]). While it is often assumed that such low-dimensional sets are in fact low-dimensional smooth manifolds, empirical evidence suggests that this is only a idealized situation: these sets may be not be smooth [8, 4, 9], they may have a non-differentiable metric tensor, self-intersections, and changes in dimensionality (see [10, 11, 12, 13] and references therein).

Principal components or the singular value decomposition is one of the most basic and yet generally used tools in statistics and data analysis. In this work we consider the local singular value decomposition of samples of μ in a ball $B_z(r)$ of radius r (the scale) centered at a data point z, and we are interested in inferring geometric properties of the underlying distribution from the behavior of all the singular values as a function of r, i.e. across scales. We investigate properties of these singular values and vectors when the data lies close to a rather general class of low-dimensional sets, and is perturbed by high-dimensional noise. We show that key properties hold with high probability as soon as the number of samples in a ball of radius r of interest is essentially linear in the intrinsic dimension. The usefulness of the multi-scale singular values is demonstrated in the context of the classical problem of estimating the intrinsic dimension of a distribution from random samples.

The analysis of this fundamental problem will require us to develop an analysis of these tools in the setting of random samples from a probability distribution in high dimensional spaces (sometimes referred to as a "point cloud"). The problem of estimating the intrinsic dimension of point clouds is of interest in a wide variety of situations. In fact, to cite some important instances, is related to estimating: the number of latent variables in a statistical model (points are samples from the model), the number of degrees of freedom in a dynamical system (points are configurations in the state space of the system sampled from trajectories), the intrinsic dimension of a data set modeled by a probability distribution highly concentrated around a low-dimensional manifold (samples are data points). Many applications and algorithms crucially rely on the estimation of the number of components in the data.

Beyond dimension estimation, the quantities studied in this paper are extremely useful in a variety of contexts:

(i) in [14, 15] they are used to explore the geometry of trajectories of very high-dimensional dynamical systems arising in molecular dynamics simulations, and to construct robust dimensionality reduction approximations to such dynamical systems;

(ii) in [16] to construct a novel multiscale representation and "transform" of point clouds, yielding fast algorithms for constructing data-driven dictionaries and obtaining sparse representation of data, for which an analogue of compressive sensing may be developed [17];

(iii) in [18] to construct estimators for μ itself, bringing approximation theory into the space of measures;

(iv) in [19, 20, 21, 22] to attack the problem of estimating the support of μ when it is a union of an unknown small number of unknown low-dimensional hyperplanes.

The inspiration for the current work originates from ideas in classical statistics (principal component analysis), dimension estimation of point clouds (see Section 2.1, 7 and references therein) and attractors of dynamical systems [23, 24, 25], and geometric measure theory [26, 27, 28], especially at its intersection with harmonic analysis. The ability of these tools to quantify and characterize geometric properties of rough sets of interest in harmonic analysis, suggests that they may be successfully adapted to the analysis of sampled noisy point clouds, where sampling and noise may be thought of as new types of (stochastic) perturbations not considered in the classical theory. In this paper we amplify and provide full proofs and extensions of the ideas originally presented in the reports [29, 30, 31], in the summary [19], and fully laid out in generality in the thesis [32].

2 Multiscale Geometric Analysis and Dimension Estimation

In the seminal paper [33] [1] multiscale quantities that measure geometric quantities of k-dimensional sets in $\mathbb{R}^D$ were introduced. These quantities could be used to characterized rectifiability and construct near-optimal solutions to the analyst's traveling salesman problem. We consider the L^2 version of these quantities, called Jones' β-numbers: for a probability measure μ and a cube Q in $\mathbb{R}^D$,

$$\beta_{2,k}(Q) := \frac{1}{\operatorname{diam}(Q)} \left(\inf_{\substack{\pi \text{ a } k-\text{dim.} \\ \text{affine hyperplane}}} \frac{1}{\mu(Q)} \int_Q \|y - P_\pi y\|^2 d\mu(y) \right)^{\frac{1}{2}} ,$$

with P_π the orthogonal projection onto π and $\|\cdot\|$ denotes the euclidean norm in $\mathbb{R}^D$. This dimensionless quantity measures the deviation (in the least-squares sense) of the measure in Q from a best-fitting k-dimensional plane. If we consider the probability measure $\mu_{|Q}(A) := \mu(A)/\mu(Q)$, obtained by localizing μ on Q, and let X_Q be a random variable with distribution $\mu_{|Q}$, then we have

$$\beta_{2,k}(Q) := \frac{1}{\operatorname{diam}(Q)} \left(\sum_{i=k+1}^{D} \lambda_i^2(\operatorname{cov}(X_Q)) \right)^{\frac{1}{2}}$$

[1] see also, among many others, [34, 35] and the pleasant short survey [36]

where $\mathrm{cov}(X_Q) = \mathbb{E}[(X_Q - \mathbb{E}[X_Q]) \otimes (X_Q - \mathbb{E}[X_Q])]$ is the covariance matrix of X_Q and $(\lambda_i^2(\mathrm{cov}(X_Q)))_{i=1,\dots,D}$ are its eigenvalues sorted in decreasing order.

In practice one may observe n random samples drawn from μ, and often such samples may be corrupted by noise in $\mathbb{R}^D$. If for simplicity we fix Q, we may formalize the above as follows: let $X_1, \dots, X_n$ be i.i.d. copies of X_Q, and $N_1, \dots, N_n$ be i.i.d. random variables representing noise, for example let them have distribution $\mathcal{N}(0, I_D)$. Given n_Q realizations $\tilde{\mathbf{X}}_{n_Q}$ of $\tilde{X}_1 := X_1 + \sigma N_1, \dots, \tilde{X}_n := X_{n_Q} + \sigma N_{n_Q}$ lying in Q, we may construct empirical versions of the quantities above:

$$\beta_{n_Q,2,k}(Q) := \frac{1}{\mathrm{diam}(Q)} \left(\sum_{i=k+1}^{D} \tilde{\lambda}_i^2 \left(\mathrm{cov}(\tilde{\mathbf{X}}_{n_Q}) \right) \right)^{\frac{1}{2}} ,$$

where $\mathrm{cov}(\mathbf{X}_n) := \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}_n[X])^T (X_i - \mathbb{E}_n[X])$ is the $D \times D$ empirical covariance matrix of a sample $\mathbf{X}_n$, $\mathbb{E}_n[X] := \frac{1}{n} \sum_{i=1}^n X_i$, $(\lambda_i^2(\mathrm{cov}(\mathbf{X}_n)))_{i=1,\dots,D}$ are its eigenvalues sorted in decreasing order.

Here $\mathrm{cov}(\tilde{\mathbf{X}}_{n_Q})$ and its eigenvalues are random variables, and it is natural to ask how close these empirical quantities are to the expected quantities above as a function of sample size, how they depend on k and the ambient dimension D, and how much noise affects the above, depending on the scale of the cube Q. For example changing the scale of Q affects n_Q, and therefore the variance of the above random variables, as well as the relative size of the noise.

In this paper we investigate these questions, and their relevance to the analysis of digital data sets that, while lying in high-dimensional spaces, may be concentrated along low-dimensional structures.

Here and in what follows $\|\cdot\|$ denotes the euclidean norm in $\mathbb{R}^D$. A set of n points in $\mathbb{R}^D$ is often thought of as an $n \times D$ matrix, whose (i,j) entry is the j-th coordinate of the i-th point. For example $\mathbf{X}_n$ and $\tilde{\mathbf{X}}_n$ will be used to denote both the point clouds corresponding to a sample of $(X_i)_{i=1}^n$ and $(\tilde{X}_i)_{i=1}^n$ and the associated $n \times D$ matrices. Similarly $\mathbf{N}_n$ may denote the matrix corresponding to a sample of $(N_i)_{i=1}^n$.

2.1 Manifolds, Local PCA and intrinsic dimension estimation

Consider random variables X, N in $\mathbb{R}^D$ with distribution μ and $\mathcal{N}(0, I_D)$, respectively. When the support of μ, which we denote by $\mathcal{M}$, has low-dimensional structure, a natural question is to estimate the unknown $k = \dim \mathcal{M}$, from random noisy data, that is from a sample $\tilde{\mathbf{X}}_n$ of $\tilde{X}_1 = X_1 + \sigma N_1, \dots, \tilde{X}_n = X_n + \sigma N_n$, where $(X_i)_i, (N_i)_i$ are i.i.d. copies of X, N and $\sigma \geq 0$ is the noise standard deviation. When $\mathcal{M}$ is linear, e.g. the image of a cube under a well-conditioned affine map, the standard approach is to perform principal components analysis (PCA) and threshold the singular values of $\tilde{\mathbf{X}}_n$ to estimate k. Let $\mathrm{cov}(\mathbf{X}_n)$ be the $D \times D$ empirical covariance matrix of the samples $\mathbf{X}_n$, with eigenvalues $(\lambda_i^2)_{i=1}^D$ ordered in decreasing order. At least for $n \gtrsim k \log k$ (with a constant that may depend on the "aspect ratio" of $\mathcal{M}$), Rudelson's Lemma [37] (see also the review [38]) implies that with high probability (w.h.p.) the empirical covariance matrix is close to the true covariance matrix. In particular, exactly k singular values will be well-separated from 0, and the remaining $D-k$ will be equal to 0. Since we observe $\tilde{\mathbf{X}}_n = \mathbf{X}_n + \sigma \mathbf{N}_n$ and not $\mathbf{X}_n$, one may consider the covariance $\mathrm{cov}(\tilde{\mathbf{X}}_n)$ as a random perturbation of $\mathrm{cov}(\mathbf{X}_n)$ and expect $\Sigma(n^{-\frac{1}{2}} \tilde{\mathbf{X}}_n)$, the set of singular values of the matrix $n^{-\frac{1}{2}} \tilde{\mathbf{X}}_n$, to be close to $\Sigma(n^{-\frac{1}{2}} \mathbf{X}_n)$, so that $\lambda_1^2, \dots, \lambda_k^2 \gg \lambda_{k+1}^2, \dots, \lambda_D^2$, allowing one to estimate k correctly w.h.p..

When $\mathcal{M}$ is a manifold, several problems arise when one tries to generalize the above line of thinking. The curvature of $\mathcal{M}$ in $\mathbb{R}^D$ in general forces the dimension of a global approximating hyperplane to be much higher than necessary. For example, consider a planar circle ($k=1$) embedded in $\mathbb{R}^D$: the *true* covariance $\mathrm{cov}(X)$ of X has exactly $2 \neq k = 1$ nonzero eigenvalues equal to half of the radius squared. In fact, it is easy to construct a one-dimensional manifold ($k=1$) such that $\mathrm{cov}(X)$ has rank equal to the *ambient* dimension: it is enough to pick a curve that spirals out in more and more dimensions. A simple example (sometimes referred to as the Y. Meyer's staircase) is the following: let $\chi_{[0,1)}(x) = 1$ if $x \in [0,1)$ and 0 otherwise. Then the set $\{x_t := \chi_{[0,2)}(\cdot - t)\}_{t=0,\dots,d-1} \subset L^2(\mathbb{R})$ is a one-dimensional (non-smooth) manifold, which is not contained in any finite-dimensional subspace of dimension. It is clear how to discretize this example and make it finite-dimensional. Notice that x_{t_1} and x_{t_2} are orthogonal whenever $|t_1 - t_2| > 2$, so this curve spirals into new directions on the unit sphere of $L^2(\mathbb{R})$ as t increases. Similar considerations would hold after discretization of the space and restriction of t to a bounded interval.

The failure of PCA in this situation can be seen as a consequence of performing PCA globally. It has been attempted to localize PCA to small neighborhoods [39, 40, 41, 42], without much success [43], at least compared to what we may call volume-based methods [44, 45, 46, 47, 48, 12, 13, 49, 50, 51, 52, 53, 54], which we discuss at length in Section 7. These methods, roughly speaking, are based on empirical estimates of the volume of $\mathcal{M} \cap B_z(r)$, for $z \in \mathcal{M}$ and $r > 0$: such volume grows like r^k when $\mathcal{M}$ has dimension k, and k is estimated by fitting the empirical volume estimates for different values of r. We expect such methods, at least when naively implemented, to both require a number of samples exponential in k (if $O(1)$ samples exist in $\mathcal{M} \cap B_z(r_0)$, for some $r_0 > 0$, these algorithms require $O(2^k)$ points in $\mathcal{M} \cap B_z(2r_0)$), and to be highly sensitive to noise, which affects the density in high dimensions. The results of our experiments (Section 5.2.1) are consistent with these observations.

The approach we propose here is quite different: we do not give up on linear approximations, with their promise of needing a number of samples essentially linear in k, but instead of a local, fixed-scale approach as in [39, 40], we propose a *multiscale approach*, since determining an appropriate range of scales at which the estimate is reliable is a key aspect to the problem. Let $z \in \mathcal{M}$, r a radius and consider the random variable $X_{z,r}$ corresponding to X conditioned to take values in $\mathcal{M} \cap B_z(r)$, where $B_z(r)$ is the Euclidean ball (in $\mathbb{R}^D$) centered at z with radius r. We will be varying r (the "scale"). We encounter 3 constraints:

(i) **curvature**: for r small enough, $\mathcal{M} \cap B_z(r)$ is well-approximated in the least squares sense by a portion of the k-dimensional tangent plane $T_z(\mathcal{M})$, and therefore we expect the covariance $\mathrm{cov}(X_{z,r})$ of $X_{z,r}$ to have k large eigenvalues and possibly other smaller eigenvalues caused by curvature. *Choosing r small enough depending on curvature*, the eigenvalues will tend to 0 faster than the top k eigenvalues of size $O(r^2)$. Therefore we would like to choose r *small*.

(ii) **sampling**: we need the number $n_{z,r}$ of samples of $X_{z,r}$ to be sufficiently high in order to estimate $\mathrm{cov}(X_{z,r})$. Therefore, for n fixed, we would like to choose r *large*.

(iii) **noise**: since we are given points corrupted by *noise*, say Gaussian with variance $\sigma^2 I_D$, we will be forced to consider r above the "scale" of the noise, i.e. not too small, since at smaller scales the estimation of the covariance of the data is completely corrupted by noise.

To summarize, only for r larger than a quantity dependent on σ^2, the variance of the noise, yet smaller than a quantity depending on curvature, conditioned on $B_z(r)$ containing enough points, will we expect local covariances to be able to detect a "noisy" version of $T_z(\mathcal{M})$. For every point $z \in \mathcal{M}$ and scale parameter $r > 0$, we let $\{(\tilde{\lambda}_i^{[z,r]})^2\}_{i=1,\ldots,D}$ be the Square Singular Values (S.S.V.'s) of $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$, where $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}$ are noisy samples in a ball of radius r centered at $\tilde{Z} := z + N$, where $N \sim \mathcal{N}(0, \sigma^2 I_D)$, sorted in nonincreasing order. We will call them the multiscale squared singular values (S.S.V.'s) of $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}$.

During the redaction of this manuscript, we were made aware by M. Davies and K. Vixie of the works [41, 42, 55, 56] where a similar approach is suggested, in the spirit of exploratory data analysis. The effects of sampling, noise, and possibly very high ambient dimension, which we think all are at the heart of the matter, are not analyzed, nor are fast multi scale algorithms for the necessary computations, also crucial in view of applications to large data sets.

2.2 Example: k-dimensional sphere in $\mathbb{R}^D$, with noise

To build our intuition, we start with a simple, yet perhaps surprising, example. Let $\mathbb{S}^k = \{x \in \mathbb{R}^{k+1} : ||x||_2 = 1\}$ be the unit sphere in $\mathbb{R}^{k+1}$, so $\dim(\mathbb{S}^k) = k$. We embed $\mathbb{S}^k$ in $\mathbb{R}^D$ via the natural embedding of $\mathbb{R}^{k+1}$ in $\mathbb{R}^D$ via the first $k+1$ coordinates. We obtain $\mathbf{X}_n$ by sampling n points uniformly at random from $\mathbb{S}^k$, and $\tilde{\mathbf{X}}_n$ is obtain by adding D-dimensional Gaussian noise of standard deviation σ in every direction. We call this model $\mathbb{S}^k(n, D, \sigma)$.

In Figure 1 we consider the multiscale S.S.V.'s corresponding to $\mathbb{S}^9(1000, 100, 0.1)$ as a function of r. Several observations are in order. First of all, notice that $\mathbb{R}^{10}$ is divided into $2^{10} = 1024$ sectors, and therefore by sampling 1000 points on $\mathbb{S}^9$ we obtain "in average" 1 point per sector (!) - of course we have so few points that we are typically far from this expected value. Secondly, observe that the noise size, if measured by $||X_i - \tilde{X}_i||^2$, i.e. by how much each point is displaced, would be order $\mathbb{E}[||X_i - \tilde{X}_i||^2] \sim 1$, where $x_i - \tilde{X}_i \sim \sigma\mathcal{N}(0, I_D) = \mathcal{N}(0, \sigma^2 I_D)$.

By concentration of measure, in fact $||x_i - \tilde{X}_i|| \sim 1$ with high probability, a length comparable with the radius of the sphere itself.

Notwithstanding the considerations above, we can in fact reliably detect the intrinsic dimension of $\mathcal{M}$. At very small scales, $B_{\tilde{Z}}(r)$ is empty or contains $o(k)$ points, and the rank of $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}$ is $o(k)$. From Figure 1, we see that at small scales, no gap among the $(\tilde{\lambda}_i^{[z,r]})^2$ is visible: $B_z(r)$ contains too few points, scattered in all directions by the noise. At larger scales, the top $9 = k$ S.S.V.'s start to separate from the others: at these scales the noisy tangent space is detected. At even larger scales, the curvature starts affecting the covariance, as indicated by the slowly growing 10th S.S.V., while the remaining smaller S.S.V.'s tend approximately to the *one-dimensional* noise variance σ^2.

3 Setting and Main Results

3.1 Notation

Random variables are denoted with capital letters, e.g. $X : (\Omega, P) \to \mathbb{R}^D$, and samples are denoted with lowercase letters, $x = X(\omega)$, $\omega \in \Omega$. Covariance matrices are denoted by

$$\mathrm{cov}(X) = \mathbb{E}[(X - \mathbb{E}[X]) \otimes (X - \mathbb{E}[X])] \tag{3.1}$$

and cross-covariance between two random variables Y, X by $\mathrm{cov}(Y, X) = \mathbb{E}[(Y - \mathbb{E}[Y]) \otimes (X - \mathbb{E}[X])]$. We will use bold letters to denote sets of random variables, in particular n i.i.d copies of a random variable X are denoted by $\mathbf{X}_n$. Given n i.i.d copies of a random variable and a subset $B \subset \mathbb{R}^D$, we let define the random set of indices

$$I_{B,\mathbf{X}_n} : (\Omega, P) \to 2^{\{1,\dots,n\}}, \qquad I_{B,\mathbf{X}_n}(\omega) = \{i = 1, \dots, n \mid X_i(\omega) \in B, X_i \in \mathbf{X}_n\}, \tag{3.2}$$

and

$$n_{B,\mathbf{X}_n} : (\Omega, P) \to \{1, \dots, n\}, \qquad n_{B,\mathbf{X}_n} = |I_{B,\mathbf{X}_n}|. \tag{3.3}$$

Note that $n_{B,\mathbf{X}_n}$ can be equivalently defined as $n_{B,\mathbf{X}_n} = \sum_{i=1}^n \mathbf{1}_B(X_i)$, the sum of binomial random variables $\mathrm{Bin}(\mu(B), n)$, where μ is the law of X. When clear from the context we might write I_B, n_B in place of $I_{B,\mathbf{X}_n}, n_{B,\mathbf{X}_n}$. We further define the random set

$$\mathbf{X}_n^B = \{X_i \in \mathbf{X}_n \mid i \in I_B\}, \tag{3.4}$$

and an associated random matrix

$$\mathrm{cov}(\mathbf{X}_n^B) = \frac{1}{n_B} \sum_{i \in I_B} \left(X_i - \left(\frac{1}{n_B} \sum_{i \in I_B} X_i \right) \right) \otimes \left(X_i - \left(\frac{1}{n_B} \sum_{i \in I_B} X_i \right) \right). \tag{3.5}$$

Given two sets of random variables $\mathbf{Y}^B, \mathbf{X}^A$, $A, B \subset X$, $\mathrm{cov}(\mathbf{Y}^B, \mathbf{X}^A)$ is defined analogously. Note that if B contains the support of X then $\mathrm{cov}(\mathbf{X}_n^B) = \mathrm{cov}(\mathbf{X}_n)$ is the empirical covariance matrix for X. If $B = B_z(r) = \{x \in \mathbb{R}^D \mid \|x - z\| \leq r\}$ for $z \in R^D$, we simplify the above notation writing $\mathbf{X}_n^{z,r}$ for $\mathbf{X}_n^{B_z(r)}$ and similarly $I_{z,r}$ for $I_{B_z(r)}$ We often view a random set in $\mathbb{R}^D$ as a random matrix, e.g. $\mathbf{X}_n$ can be thought of as a n by D matrix. For example, viewing $\mathbf{Y}_n, \mathbf{X}_n$ as matrices, we will write $\mathrm{cov}(\mathbf{Y}_n, \mathbf{X}_n) = \frac{1}{n} \overline{\mathbf{Y}}_n^T \overline{\mathbf{X}}_n$, where $\overline{\mathbf{Y}}_n, \overline{\mathbf{X}}_n$ denote the matrices obtained centering the rows of $\mathbf{Y}_n, \mathbf{X}_n$ with respect to the centers of mass of the corresponding sets.

Definition 1. *We let $\{\lambda_i^2(\mathrm{cov}(X))\}$ be the Squared Singular Values of X, i.e. the eigenvalues of $\mathrm{cov}(X)$ (possibly up to a set of 0 eigenvalues), sorted in decreasing order. We let $\Delta_i(\mathrm{cov}(X)) := \lambda_i^2(\mathrm{cov}(X)) - \lambda_{i+1}^2(\mathrm{cov}(X))$, for $i = 1, \dots, D-1$, $\Delta_D(\mathrm{cov}(X)) = \lambda_D^2(\mathrm{cov}(X))$, $\Delta_{\max} := \max_{i=1,\dots,D} \Delta_i$.*

We denote by $\|\cdot\|$ the euclidean norm for vectors and the operator norm for matrices. We let $\mathbb{S}^k$ be the unit k-dimensional sphere and $\mathbb{B}^k$ the unit k-dimensional ball. We let $\mu_{\mathbb{R}^k}$ be the Lebsegue measure in $\mathbb{R}^k$.
Finally, in what follows C, C_1, C_2 will denote numeric constants independent of all parameters, and their values may change from line to line. We will write $f(x) \lesssim g(x)$ if there exists a numerical constant C such that $f(x) \leq Cg(x)$ for all x, and $f(x) \approx g(x)$ if there exist two numerical constants C_1, C_2 such that $C_1 g(x) \leq f(x) \leq C_2 g(x)$ for all x.

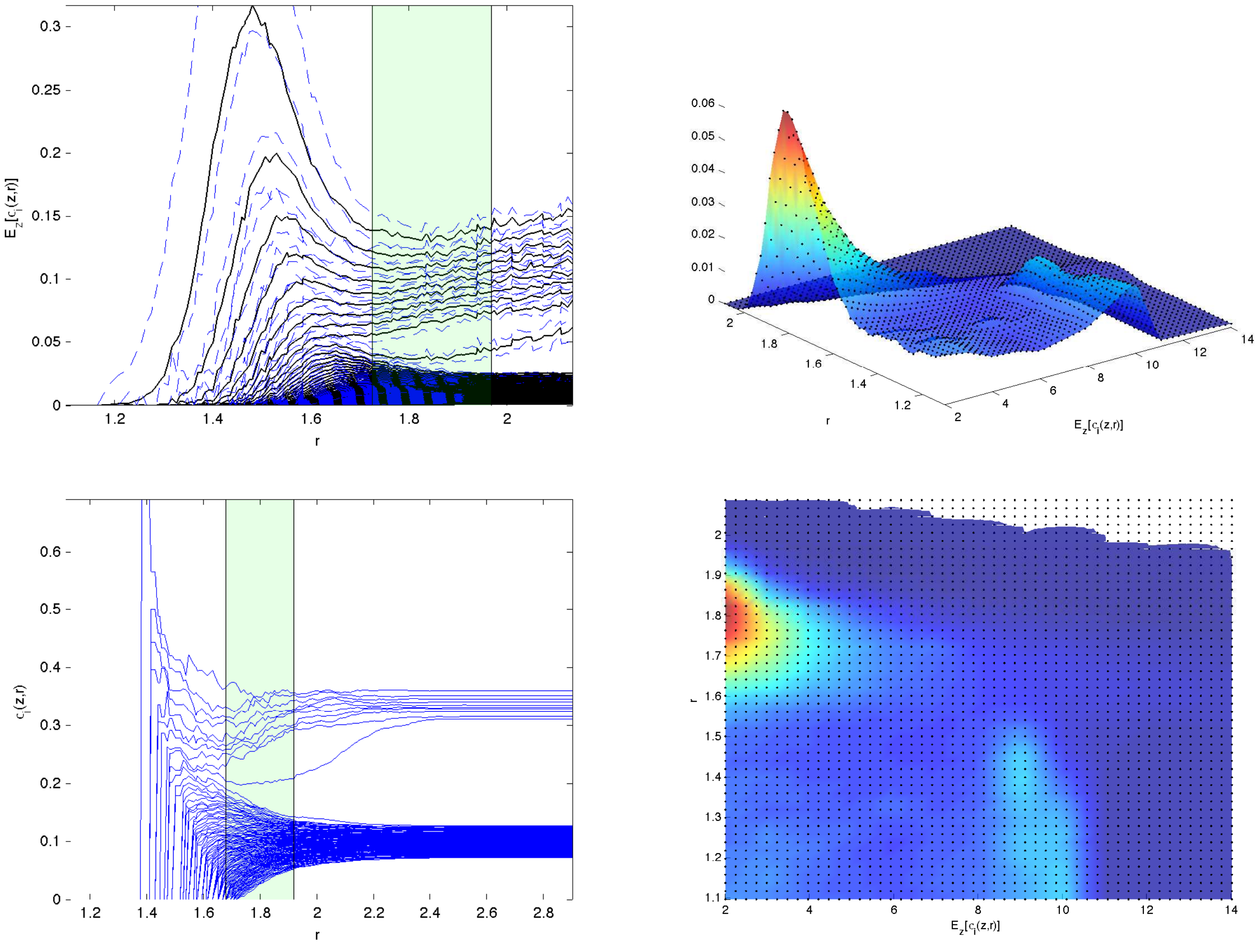

Figure 1: $\mathbb{S}^9(1000, 100, 0.1)$. Top left: plot of $\mathbb{E}_z[(\tilde{\lambda}_i^{[z,r]})^2]$, and corresponding standard deviation bands (dotted), as a function of r. The top 9 S.S.V.'s dominate and correspond to the intrinsic dimensions; the 10-th S.S.V. corresponds to curvature, and slowly increases with scale (note that at large scale $\Delta_{10} > \Delta_9$, where $\Delta_i = (\tilde{\lambda}_i^{[z,r]})^2 - (\tilde{\lambda}_{i+1}^{[z,r]})^2$); the remaining S.S.V.'s correspond to noise in the remaining 90 dimensions, and converge to the one-dimensional noise size σ^2. Top right: smoothed plot of the gaps $(\tilde{\lambda}_k^{[z,r]})^2 - (\tilde{\lambda}_{k+1}^{[z,r]})^2$ of the multiscale singular values on a portion the "scale-frequency" plane (where "frequency" is index of the singular value): note the 10-th gap passing the 9-th gap at large scales. At smaller scales (not shown), noisy singular values create large random gaps. Bottom left: the multiscale S.S.V. $(\tilde{\lambda}_i^{[z,r]})^2$ for a fixed (randomly chosen) point z: the algorithm is run at only that point, and both the global range of scale and the correct range of "good scale" are detected automatically. Bottom right: a view of the surface top right from above.

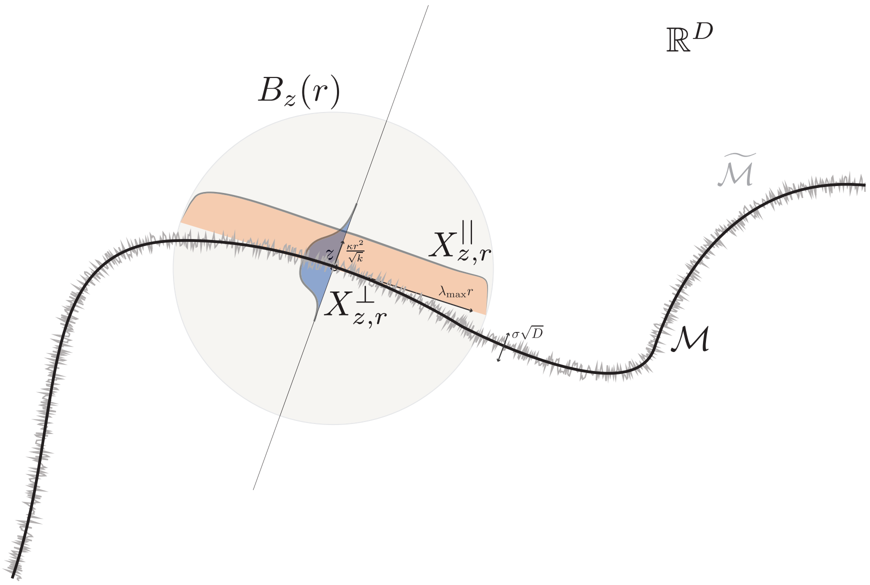

Figure 2: A pictorial representation of some of our geometric assumptions.

3.2 Problem Setting

Let X be a random variable in $\mathbb{R}^D$ with distribution μ_X and $\mathcal{M} := \mathrm{supp}\{\mu_X\}$. We will be interested in the case when $\mathcal{M}$ is low-dimensional, for example a k dimensional manifold, or a k Ahlfors regular k-rectifiable set [34, 35], with $k \ll D$. More generally, $\mathcal{M}$ may be just approximated by a low-dimensional set, in a sense that our assumptions below will make precise. Let N be a random variable, for example $N \sim \mathcal{N}(0, I_D)$, that will think of as as noise, and let $\tilde{X} = X + \sigma N$.

Roughly speaking, we are interested into the properties of *local* covariance matrices and in how they can be estimated from random noisy samples. More precisely, fix $z \in \mathcal{M}$, and consider the random variable $X_{z,r}$ with values in $B_z(r)$ and distribution $\mu_{z,r}$, where $\mu_{z,r}(A) := \mu_X(A \cap B_z(r))/\mu_X(B_z(r))$ is the restriction of μ_X to $B_z(r)$. We are interested into estimating the multiscale family of matrices $\mathrm{cov}(X_{z,r})$, and in particular in the behavior of its eigenvalues as a function of r, for fixed $z \in \mathrm{supp}\{\mu_X\}$, since it contains useful geometric information.

Towards this end, we have at disposal a sample of $\tilde{\mathbf{X}}_n$ obtained from n i.i.d. copies of $\tilde{X}$ and have access to a sample of the random variable $\tilde{Z} = z + \sigma N$. Then we can consider the random matrix $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$. Indeed, we will show that $\mathrm{cov}(X_{z,r})$, and its spectral properties, can be estimated by $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$ whenever r is in a suitable range of scales depending on the geometry of the data distribution and the noise.

It is then crucial to understand how close $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$ is to $\mathrm{cov}(X_{z,r})$. Towards this end, we make use of a few intermediate (theoretical) quantities that are not accessible in practice. In particular, we will consider the random sets

$$\widetilde{\mathbf{X}_n^{[\tilde{Z},r]}} = \mathbf{X}_n^{[\tilde{Z},r]} + \sigma \mathbf{N}_n, \qquad \widetilde{\mathbf{X}_n^{[z,r]}} = \mathbf{X}_n^{[z,r]} + \sigma \mathbf{N}_n$$

where in the first set the noise is added only after localization and in the second set we assume to have access to the *noiseless center* $z \in \mathcal{M}$. The above sets are not observable and can be contrasted to $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}$ which is available in practice.

3.3 Assumptions

We make the following assumptions, which we call "usual assumptions" from this point onwards.

I. Assumptions on the Geometry. We assume that for every $z \in \mathcal{M}$ there exists a range of scales $r \in (R_{\min}, R_{\max})$, an integer k and an orthogonal projection $P^{[z,r]}$ onto an affine subspace of dimension k such that if we let

$$X_{z,r}{}^{||} \;\; = P^{[z,r]} X_{z,r} \quad , \quad X_{z,r}{}^{\perp} \;\; = (I - P^{[z,r]}) X_{z,r}$$

then the following conditions hold almost surely, for all $r \in (R_{\min}, R_{\max})$ and $\tilde{Z} \in \mathbb{R}^D$ satisfying $||\tilde{Z} - z|| \leq R_{\max}$, and for some $1 \leq \lambda_{\max} \leq \sqrt{k}$, $\lambda_{\min}, \delta, v_{\min} > 0$, $\kappa \geq 0$, and $v_{\tilde{Z}}(r)$, called the *geometric parameters*:

$$\boxed{\begin{array}{lll} \lambda_i^2(\mathrm{cov}(X_{z,r}{}^{||})) \subseteq \dfrac{[\lambda_{\min}^2, \lambda_{\max}^2]}{k} r^2 & , \max\limits_{i<k} \Delta_i(\mathrm{cov}(X_{z,r}{}^{||})) \leq \dfrac{\delta}{k} r^2 & \\ \left\|X^{\perp}\right\| \leq \sqrt{k}\kappa r^2 \;\; \text{a.s.} & , ||\mathrm{cov}(X_{z,r}{}^{\perp})|| \leq \dfrac{\kappa^2}{k} r^4 \, , & \dfrac{\mathrm{tr}(\mathrm{cov}(X_{z,r}{}^{\perp}))}{||\mathrm{cov}(X_{z,r}{}^{\perp})||} \leq 2k^2 \\ \mu_X(B_{\tilde{Z}}(r)) = \mu_{\mathbb{R}^k}(\mathbb{B}^k) v_{\tilde{Z}}(\rho)\rho^k & , \rho^2 := r^2 - d(\tilde{Z}, \mathcal{M}) & \\ \dfrac{v_{\tilde{Z}}(r(1+h))}{v_{\tilde{Z}}(r)} \leq (1+h)^k \;, h > 0 & , \dfrac{v_{\tilde{Z}}(r)}{v_z(r)} \leq 1 + \dfrac{||z - \tilde{Z}||}{r} \, , & v_{\tilde{Z}}(r) \geq v_{\min} \end{array}} \tag{3.6}$$

where $\mu_{\mathbb{R}^k}$ is k-dimensional Lebesgue measure. We think of $\lambda_{\min}, \lambda_{\max}$ as being of order 1.

II. Assumptions on the Noise. We assume that N is independent of X, and has a standard multivariate normal distribution, i.e. independent, centered Gaussian coordinates with variance 1.

Finally, we shall assume that there exists a constant C_ξ that depends (continuously) only on $\xi := \frac{\sigma\sqrt{D}}{r}$, such that for any $z \in \mathcal{M}$ and for $\xi < 1/3$

$$\sum_{l=1}^{\infty} e^{-l^2} \mu_X\left(\left(B_z(\sqrt{r^2 + (l+1)^2\sigma^2 D})\right) \setminus B_z(\sqrt{r^2 + l^2\sigma^2 D})\right) \leq C_\xi \mu_{\mathbb{R}^k}(\mathbb{B}^k) r^k \, ,$$

which ensures that $\mathcal{M}$ does not come close to self-intersecting too often, in a rather weak, measure-theoretic sense. We make of course all μ_X-measurability assumptions needed for the above to make sense.

We interpret $X_{z,r}{}^{||}$ and $X_{z,r}{}^{\perp}$ as the projections of $X_{z,r}$ onto a local approximating plane and its orthogonal complement, respectively (see also Figure 3). The first condition in (3.6) roughly determines the elongation of $X_{z,r}$ projected onto the approximate tangent plane. Note that, after subtracting the means, $X_{z,r}{}^{||}, X_{z,r}{}^{\perp}$ are almost surely bounded by r, but tighter conditions on $X_{z,r}{}^{\perp}$ are possible depending on the curvature. Indeed, the second condition enforces second-order closeness (in the least-squares sense) of $X_{z,r}$ to the tangent plane. It also renormalizes $X_{z,r}{}^{\perp}$ so that κ is a measure of extrinsic curvature that takes into account the distribution in the normal directions where $X_{z,r}{}^{\perp}$ is supported. The condition on the effective rank $\mathrm{tr}(A)/||A||$ for $A = \mathrm{cov}(X_{z,r}{}^{\perp})$ is motivated by the fact that locally $\mathcal{M}$ may only curve in $\binom{k+1}{2}$ dimensions (see remark in Section 4.2), so that $X_{z,r}{}^{\perp}$ is effectively $O(k^2)$-dimensional. This condition may also be generalized (or even removed) depending on a priori knowledge on properties of the curvature, with obvious changes in our results (essentially only the value of κ', introduced later, is affected). The number k is what we will call the intrinsic dimension of $\mathcal{M}$, at least in the range of scales $(R_{\min}, R_{\max})$. It may change with the range of scales. In particular cases, it coincides with classical notions of intrinsic dimension, notably when $\mathcal{M}$ is an embedded manifold and the range of scales considered is small enough. So on the one hand it overlaps with standard notions of dimension on a reasonably large class of sets, on the other hand it is a notion robust to perturbations, and is scale-dependent, therefore removing part of the ill-posedness of the estimation problem. The volume growth condition in (3.6) is similar to the notion of Ahlfors-David k-regular sets, but localized in both space and scale.

One may consider more general models of noise, with i.i.d. strictly subgaussian coordinates and an approximate spherical symmetry, but we postpone these technical issues here. We refer the reader to Appendix 11 for a review of the definition and basic properties of subgaussian random variables, and to section 9.5 (and equations (9.24) in particular) for a discussion of approximate spherical symmetry.

The assumptions above are local in a neighborhood of z, and $\lambda_{\min}, \lambda_{\max}, \kappa, v_{\min}, v_{\max}$ may depend on z, as well as on k. We introduced factors of k in our conditions because they are the natural scalings for certain manifolds (see Section 4), in the sense that in these particular cases the remaining parameters become independent of k.

Example 1. *We specialize the general hypotheses above to various settings of interest:*

(i) *the "manifold case": μ_X the normalized volume measure on a k-dimensional smooth compact Riemannian manifold $\mathcal{M}$. Such a manifold has positive* reach, *guaranteeing the existence of a nontrivial interval $[R_{\min}, R_{\max}]$. In fact, typically $R_{\min} = 0$. More generally, μ_X may be a measure on $\mathcal{M}$ which is absolutely continuous with respect to the volume measure, with Radon-Nykodym derivative uniformly bounded above and below. Certain non-compact manifolds are also possible, since our conditions are local. The "curvature" κ in general is not determined by any intrinsic metric property that $\mathcal{M}$ may have, but in general depends on the embedding of $\mathcal{M}$ in $\mathbb{R}^D$ (see Section 4).*

(ii) *As a special case of the above, consider the k-dimensional unit sphere $\mathbb{S}^k$ in $\mathbb{R}^{k+1}$. This example helps identify some natural scaling laws for the parameters, as a function of the intrinsic dimension k. For $\mathbb{S}^k$ (and a whole class of manifolds) we show in Section 4 that $\lambda_{\max} = \lambda_{\min} = 1$, $\kappa^2 \sim k^{-1}$ and $v_{\min}\mu_{\mathbb{R}^k}(\mathbb{B}^k) \sim k^{-\frac{1}{2}}$, and $R_{\max} \sim 1$ where $\sim$ subsumes universal constants independent of k.*

(iii) *For a finite union of k-dimensional manifolds as in (i), the assumptions are satisfied except when z is in the intersection of at least two manifolds. The manifolds do not need to have the same dimension k, in which case the assumptions hold for different values of k depending on z. A particular case is finite unions of planes. Note that the conditions hold also for certain infinite unions of manifolds. All that is needed is that intersections are isolated and, for the problem to be well-conditioned in the sense that $R_{\max}$ is not too small on sets of large measure, that the regions at a certain distance from intersections are not too large.*

3.4 Main results

We are interested in understanding the relationships between $\mathrm{cov}(X_{z,r})$ and $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$. The former is the true local covariance of μ_X restricted to $B_z(r)$, where $z \in \mathcal{M}$ while the second is the observed empirical noisy covariance of the sample points $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}$ that lie in $B_{\tilde{Z}}(r)$, where $\tilde{Z} = z + \sigma N$ is a (random) noisy center. The latter is the quantity observable by an algorithm. The covariance $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$ is a "corrupted" version of $\mathrm{cov}(X_{z,r})$ because of sampling and noise: sampling creates random fluctuations around the expected covariance, noise corrupts the center (from z to $z + \sigma N$) and the points, causing points in $B_z(r)$ to exit the ball, and points from outside $B_z(r)$ to enter the ball. We are interested in non-asymptotic results, for D large, that hold for finite n and for nonvanishing noise size, guaranteeing that $\mathrm{cov}(X_{z,r})$ and $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$ are close. In fact, it turns out that since the noise is typically not negligible, it is natural to allow for a change in scale, and compare instead $\mathrm{cov}(X_{z,r_=})$ with $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$, where $r_=^2 = r^2 - 2\sigma^2 D$.

Our results show that, for a fixed point z, as soon as the noise has "size" smaller than "the scale of the curvature", there is a range of scales such that if $O(k \log k)$ points are available in $B_z(r_=)$, then indeed $\mathrm{cov}(X_{z,r_=})$ and $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$ are essentially as close as it may be expected given the noise size; in particular the top k eigenvalues (which are much larger than the remaining ones) of the two matrices are comparable, and so are the corresponding eigenspaces, that represent approximate tangent planes. The "size of the noise" is measured by $\mathbb{E}[\sigma\|N\|] \sim \sigma\sqrt{D}$, and the "the scale of the curvature" is measured roughly by $\frac{1}{\kappa\sqrt{k}}$ (see Theorem 1 for a more precise statement).

We shall restrict our attention to the range of scales

$$r \in \left(R_{\min} + 4\sigma\sqrt{D} + \frac{1}{6k}, R_{\max} - \sigma\sqrt{D} - \frac{1}{6k}\right). \tag{3.7}$$

We would no need to restrict it now, but it would be imposed on us later anyway. Scales in this range are above the scale of noise, and below the scale at which curvature of $\mathcal{M}$ affects too severely the multi scale singular values. We introduce some natural parameters: t will tune the probability of success, which will be in the form $1 - ce^{-ct^2}$.

Define

$$\begin{aligned} \overline{n} := \mathbb{E}[n_{z,r_=}] = n\mu_X(B_z(r_=)) \quad , \quad \epsilon^2 = \epsilon^2_{r_=,n,t} := \frac{t^2\, k \log k}{\overline{n}} \\ (\epsilon^\perp)^2 = (\epsilon^\perp_{r_=,n,t})^2 = \frac{t^2 k^2 \log(D \wedge \overline{n})}{\overline{n}} = \epsilon^2 k \log_k(D \wedge \overline{n}) \quad , \quad \kappa' := \kappa((1+\epsilon^\perp) \wedge k) \\ \sigma_0 := \sigma\sqrt{D} \quad . \end{aligned} \tag{3.8}$$

These quantities represent, respectively, the expected number of (noiseless) points in $B_z(r_=)$, and the reciprocal of a "local oversampling factor" for a k-dimensional covariance estimation, since $O(t^2 k \log k)$ points in $B_z(r_=)$ suffice to estimate the leading portion of the covariance matrix in $B_z(r_=)$ with high confidence. In the normal direction $\epsilon^\perp$ is the smallest of two terms, the first one coming from the covariance having effective rank k^2, and the second coming from the standard concentration rate $\sqrt{D/\overline{n}}$. The latter kicks in only for $\overline{n} \geq D$, which is not the case of interest here, but it is helpful to show consistency (the limit $n \to +\infty$) as a simple corollary of our results.

Theorem 1 (D large)**.** *Fix $z \in \mathcal{M}$. Let the assumptions in Section 3.3 be satisfied. For $D \geq k^2$, σ_0 constant, $t \in (C, C_{\lambda_{\max},\lambda_{\min},\delta,\epsilon}\frac{\sqrt{D}}{k})$, and $\epsilon = \epsilon_{r_=,n,t} \leq \frac{1}{2\lambda_{\max}}$, for r in the range of scales* (3.7) *intersected with*

$$r \in \left(\frac{4\sigma_0}{\lambda^2_{\min} - \delta^2\lambda_{\max}\epsilon - \frac{\epsilon^2}{\lambda^2_{\min}}\left(\frac{C\sigma_0 k}{r} \vee \frac{1}{\overline{n}}\right) - \frac{\sigma_0\kappa'}{t}}, \frac{\frac{\lambda_{\max}}{4} \wedge \sqrt{k}}{\kappa'} \right) ,$$

the following hold, with probability at least $1 - Ce^{-Ct^2}$:

(i) $\Delta_k(\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}))$ *is the largest gap of* $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$*;*

(ii) $||\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}) - \mathrm{cov}(X_{z,r_=}) - \sigma^2 I_D|| \leq \left(\sigma_0^2\epsilon + \lambda_{\max}\sigma_0 r + \left(\lambda_{\max} + 2\sigma_0\kappa' + \frac{\epsilon}{\overline{n}}\right) r^2 + O\left(\frac{r^3}{\epsilon}\right)\right)\frac{\epsilon}{k}$.

(iii) if we let Π_k and $\tilde{\Pi}_k$ be the spaces spanned by the top k singular vectors of $\mathrm{cov}(X_{z,r_=})$ *and* $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$*, we have*

$$|\sin\Theta(\Pi_k,\tilde{\Pi}_k)| \leq \frac{\frac{\sigma_0^2\epsilon}{\sqrt{kD}} + \frac{2\sqrt{\lambda_{\max}}\sigma_0\epsilon t}{k} + \frac{\epsilon\lambda_{\max}+\sigma_0\sqrt{\frac{k}{D}}\kappa'}{k}r^2 + \frac{\lambda_{\max}\kappa'}{k}r^3 + \frac{\frac{\kappa'^2}{k}\lambda^2_{\max}}{\lambda^2_{\min}-\kappa'^2r^2}r^4 + \frac{\epsilon^2r^2}{k}\left(\frac{C\sigma_0 k}{r} \vee \frac{1}{\overline{n}}\right)}{\frac{\lambda^2_{\min}-\kappa'^2r^2}{k}r^2 - \frac{\sigma_0^2}{D} - \frac{\sigma_0\epsilon(2\kappa' r^2+\sigma_0\epsilon)}{k} - \frac{\epsilon^2r^2}{k}\left(\frac{C\sigma_0 k}{r} \vee \frac{1}{\overline{n}}\right)}$$

This Theorem is a special case of Theorem 2 when D is large and $\sigma\sqrt{D} =: \sigma_0$ fixed. Here we are mostly interested in the case $t \approx 1$ and $\overline{n} \approx k \log k$:

Corollary 1. *Under the assumptions of Theorem 1, if $\delta \ll \lambda_{\min} \approx \lambda_{\max} \approx 1$, ϵ small enough and D large enough depending on $\lambda_{\min}, \lambda_{\max}, \kappa, \sigma$, then Δ_k is the largest gap with high probability for r in the range* (3.7) *intersected with*

$$5\sigma_0 \leq r_= \leq \frac{\lambda_{\max}}{4\kappa k} \,.$$

If in addition $\overline{n} \gtrsim k^2$ so that $\epsilon^\perp \leq \frac{1}{2}$, the upper bound of this interval may be increased to $\frac{\lambda_{\max}}{6\kappa}$.

The lower bound is comparable to the length of noise vector, the upper bound is comparable to the largest radius where the curvature is not too large (see the proof of Proposition 1, and Figure 3). The geometry is that of a rather hollow "tube" or radius $\sqrt{D}$ around the support of μ, curving at scale roughly $1/\kappa'$, and the Theorem guarantees, among other things, that for r larger than the scale of the noise and smaller than the radius of curvature, with only $k \log k$ samples in a ball of radius r from noisy samples we obtain a faithful empirical estimate of the local covariance matrix.

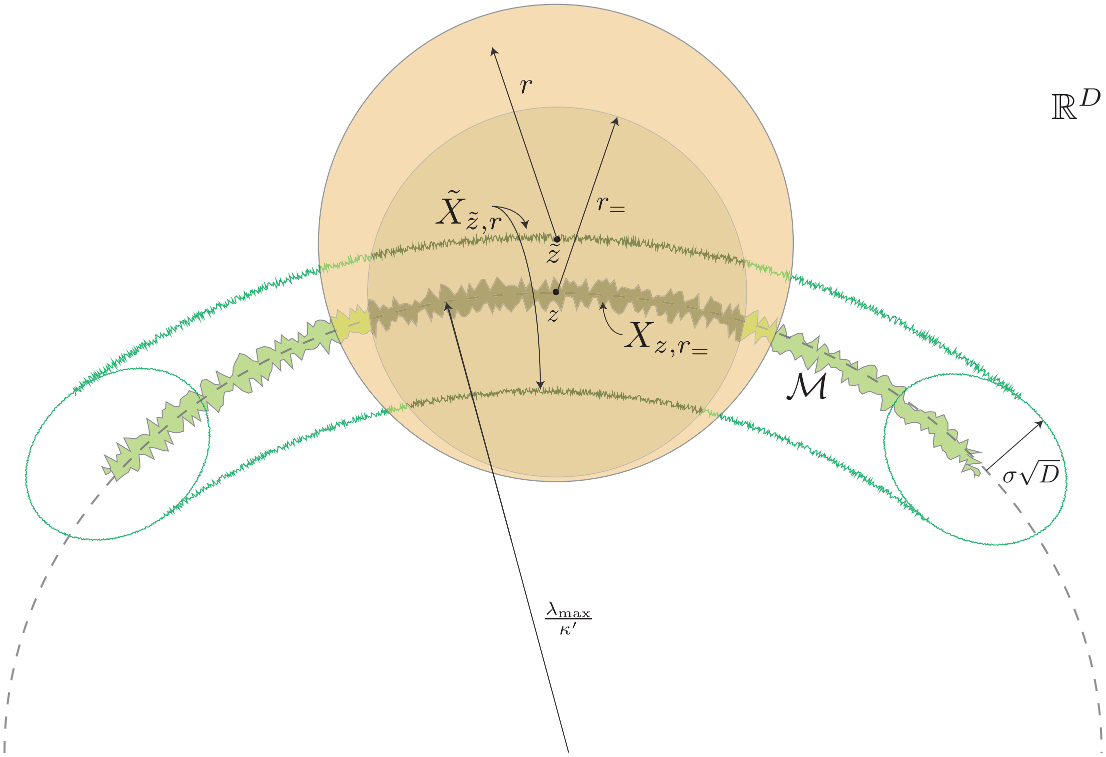

Figure 3: A pictorial representation of the natural scalings, as well as the corresponding local random variables for r in the good range of scales. The noise pushes points on $\mathcal{M}$ at distance roughly $\sigma\sqrt{D}$ (w.h.p.), mostly in the normal direction. Therefore we expect that good scales will correspond to $r \gtrsim \sigma\sqrt{D}$. Also, r needs to be below the "radius of curvature" of $\mathcal{M}$, which turns out to be comparable to $\lambda_{\max}/\kappa'$. Since we only have access to $\tilde{z}$ and the noisy data, we need to compare the covariance noisy data in $B_{\tilde{Z}}(r)$ with that of the clean data in a slightly smaller ball, $B_z(r_=)$. In the scaling limit $D \to +\infty$, we impose that this picture is invariant, which is achieved by scaling σ so that $\sigma\sqrt{D} =: \sigma_0$ independently of D.

3.4.1 Technical Results

The main Theorems above are consequences of Theorem 2 below, which builds upon Propositions 1 and 2.

- In Proposition 1 we study the perturbation $\mathrm{cov}(X_{z,r_=}) \to \mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})$. However, we do not have access to z, which is not an observed data point, but only to a sample of $\tilde{Z}$. Likewise, for a fixed scale r and center z, we do not have access to $\widetilde{\mathbf{X}_n^{[z,r_=]}}$ but only to $\tilde{\mathbf{X}}_n^{[z,r_=]}$.
- Proposition 2 then shows that with high probability, up to a small change in scale from r to $r_=$, the covariance computed from $\widetilde{\mathbf{X}_n^{[z,r_=]}}$ and $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}$ are close, allowing us to translate our analysis above of $\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})$ to $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$.

Our main Theorem combines these two perturbations.

Proposition 1. *Let the assumptions in Section 3.3 hold for a fixed $z \in \mathcal{M}$, and choose $r_= \in (R_{\min}, R_{\max})$. Let $\overline{n} = \overline{n}_{r_=,n}$, $\epsilon = \epsilon_{r_=,n,t}$ be as in* (3.8)*, $t \geq C_1$, $\epsilon \leq \frac{1}{C_2} \leq \frac{1}{2}$ and choose $\gamma = \gamma_{r_=,n,t}$ and $\varphi = \varphi_{r_=,n,t}$ as follows:*

$$2\gamma^2 := \lambda_{\min}^2 - \delta^2 - \lambda_{\max}\epsilon \quad , \quad \varphi^2 := \frac{\gamma^2}{1+\epsilon} - \sigma\kappa'\sqrt{k}\left(\sqrt{\frac{D}{\overline{n}}} + \epsilon\right) \tag{3.9}$$

Then, with probability at least $1 - ce^{-ct^2}$, for r as above and such that

$$\epsilon\frac{\sigma\sqrt{D}}{\varphi}\left[\frac{\lambda_{\max}}{\varphi} \vee \left(\mathbf{1}_{\overline{n}\leq CD} + \sqrt[4]{\frac{\overline{n}}{D}}\mathbf{1}_{\overline{n}\geq CD}\right)\right] \leq r \leq \frac{\lambda_{\max}}{4\kappa'}\left(1 + \frac{6\gamma^2}{\lambda_{\max}^2}\right) \tag{3.10}$$

we have

(i) $\Delta_k(\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}}))$ *is the largest gap of* $\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})$*;*

(ii) the following bound holds

$$\begin{aligned}\left\|\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}}) - \mathrm{cov}(X_{z,r_=}) - \sigma^2 I_D\right\| \leq \Bigg(&2\sigma^2\sqrt{\frac{D}{k\log k}}\epsilon\left(1 + \sqrt{\frac{D}{k\log k}}\epsilon\mathbf{1}_{\overline{n}\leq CD}\right) \\ &+ \frac{\lambda_{\max}\sigma}{\sqrt{k}}\epsilon\left(1 + \sqrt{\frac{D}{k\log k}}\right)r_= + \left(\frac{\epsilon\lambda_{\max}}{\sqrt{k}} + \sigma\kappa'\left(2\sqrt{\frac{D}{k\log k}}\epsilon + 1\right)\right)\frac{{r_=}^2}{\sqrt{k}} \\ &+ \frac{2\lambda_{\max}\kappa'}{k}{r_=}^3 + \frac{2\kappa'}{k}{r_=}^4\Bigg)(1+\epsilon) =: E_{1,r_=}\,.\end{aligned}$$

(iii) if we let Π_k (respectively $\tilde{\Pi}_k$) be the spaces spanned by the top k singular vectors of $\mathrm{cov}(X_{z,r_=})$ *(respectively* $\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})$*), for r in the range above we have*

$$|\sin\Theta(\Pi_k, \tilde{\Pi}_k)| \leq \frac{\sigma^2\sqrt{\frac{D}{\overline{n}}}t + \frac{\sqrt{\lambda}_{\max}\sigma}{\sqrt{k}}\left(\sqrt{\frac{D}{\overline{n}}}t + \epsilon\right)t + \frac{\epsilon\lambda_{\max}+\sigma\sqrt{k}\kappa'}{k}{r_=}^2 + \frac{\lambda_{\max}\kappa'}{k}{r_=}^3 + \frac{\frac{\kappa'^2}{k}\lambda_{\max}^2}{\lambda_{\min}^2 - \kappa'^2{r_=}^2}{r_=}^4}{\frac{\lambda_{\min}^2 - \kappa'^2{r_=}^2}{k}{r_=}^2 - \sigma^2\mathbf{1}_{\overline{n}\leq CD} - \sigma\sqrt{\frac{D}{\overline{n}}}t\left[\frac{2\kappa'{r_=}^2}{\sqrt{k}} + \sigma\mathbf{1}_{\overline{n}\geq CD} + \sigma\sqrt{\frac{D}{\overline{n}}}\mathbf{1}_{\overline{n}\leq CD}\right]}$$

Proof. See Appendix 8. □

Corollary 2 (D large)**.** *With the same assumptions and notation as in Proposition 1. For $\overline{n} \geq C, t \geq C$, and D large compared to k, let $\sigma_0 := \sigma\sqrt{D}$ be independent of D, and assume $\epsilon \leq \frac{1}{2}$. Then with probability at least $1 - Ce^{-Ct^2}$, in the range*

$$\frac{3\epsilon\sigma_0\lambda_{\max}}{\lambda_{\min}^2 - \delta^2 - \lambda_{\max}\epsilon - 3\sigma_0\kappa'\sqrt{k}} \leq r_= \leq \frac{\lambda_{\max}}{4\kappa'}\left(1 + \frac{\lambda_{\min}^2 - \delta^2 - \lambda_{\max}\epsilon}{\lambda_{\max}^2}\right)$$

we have that

(i) the largest gap of $\widetilde{\mathrm{cov}(\mathbf{X}_n^{[z,r_=]}})$ *is the k-th gap;*

(ii) $\left\|\widetilde{\mathrm{cov}(\mathbf{X}_n^{[z,r_=]}})-\mathrm{cov}(X_{z,r_=})-\sigma^2 I_D\right\| \leq \frac{3\left(\sigma_0^2\epsilon^2+\lambda_{\max}\sigma_0\epsilon r+(\lambda_{\max}+\sigma_0)\epsilon r^2+\kappa\sqrt{k}r^3(\lambda_{\max}+r)\right)}{k\log k}$

We have neither access to a point $z \in \mathcal{M}$, nor to $B_z(r_=) \cap \mathcal{M}$ since our observations are points perturbed by noise. We show that the effect of this perturbation may be offset by a change in scale, from $r_=$ to r, up to the appearance of terms depending on the "geometric signal to noise ratio" r/σ.

Proposition 2. *Let $D \geq C$, and*

$$r \in \left(R_{\min} + 4\sigma\sqrt{D} + \frac{1}{6\kappa'}, R_{\max} - \sigma\sqrt{D} - \frac{1}{6\kappa'}\right) \cap \left(3\sigma\left(\sqrt{D} \vee k\right), \frac{\sqrt{k}}{\kappa'}\right) \tag{3.11}$$

where C is a universal constant. Then, for $t, v \geq C$, $\overline{n} = \overline{n}_{r_=,n} \geq t^2$, $s^2 < \frac{r^2/k}{12\sigma^2 D}\sqrt{D}$

$$||\widetilde{\mathrm{cov}(\mathbf{X}_n^{[z,r_=]}}) - \mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})|| \leq v^2\left(\beta_s \vee \frac{1}{\overline{n}}\right) r^2 =: E_{2,r} \tag{3.12}$$

holds with

$$\beta_s := \left(1 + \frac{s^2\sigma\sqrt{D}}{r} + \left(1 \vee \frac{\sigma^2 D}{r^2/k}\right)\sqrt{\log\frac{r}{3\sigma k}}\right)\frac{\sigma k}{r}$$

and with probability at least $1 - ce^{-c((v^2\overline{n})\wedge s^4 \wedge t^2)}$.

Proof. See Appendix 9. □

We combine the two perturbations above to obtain the following

Theorem 2. *Fix $z \in \mathcal{M}$ and let the assumptions in Section 3.3. Choose r in the range* (3.11) *intersected with $(3\sigma(\sqrt{D} \vee k), \frac{\sqrt{k}}{\kappa'})$. With $\overline{n}$ and ϵ defined as in* (3.8)*, $t, v \geq C_1$, $\epsilon \leq \frac{1}{C_2} \leq 1$, $1 \leq s^2 \leq \frac{{r_=}^2/k}{12\sigma^2 D}\sqrt{D}$. Then, with probability at least $1 - ce^{-c((v^2\overline{n})\wedge s^4\wedge t^2)}$, if*

$$\epsilon\frac{\sigma\sqrt{D}}{\varphi}\left[\frac{\lambda_{\max}}{\varphi} \vee \left(\mathbf{1}_{\overline{n}\leq CD} + \sqrt[4]{\frac{\overline{n}}{D}}\mathbf{1}_{\overline{n}\geq CD}\right)\right] \leq r_= \leq \frac{\lambda_{\max}}{4\kappa'}\left(1 + \frac{6\gamma^2}{\lambda_{\max}^2}\right) \tag{3.13}$$

where $\varphi = \varphi_{r,n,v,s}$ is obtained from (3.9) *by replacing $\gamma_{r,n,t}$ with*

$$2\gamma_{r,n,v,s}^2 := \lambda_{\min}^2 - \delta^2 - \lambda_{\max}\epsilon_{r_=,n,t} - \frac{v^2 k}{\lambda_{\min}^2}\left(\beta_s \vee \frac{1}{\overline{n}}\right).$$

we have:

(i) $\Delta_k(\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}))$ *is the largest spectral gap of* $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$*;*

(ii) $||\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}) - \mathrm{cov}(X_{z,r_=}) - \sigma^2 I_D|| \leq E_{1,r_=} + E_{2,r}$*, where $E_{1,r_=}$, $E_{2,r}$ are given in Proposition 2 and 1, respectively;*

(iii) if we let Π_k and $\tilde{\Pi}_k$ be the spaces spanned by the top k singular vectors of $\mathrm{cov}(X_{z,r_=})$ *and* $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$*, we have*

$$|\sin\Theta(\Pi_k,\tilde{\Pi}_k)| \leq \frac{\sigma^2\sqrt{\frac{D}{\overline{n}}}t + \frac{\sqrt{\lambda}_{\max}\sigma}{\sqrt{k}}\left(\sqrt{\frac{D}{\overline{n}}}t+\epsilon\right)t + \frac{\epsilon\lambda_{\max}+\sigma\sqrt{k}\kappa'}{k}r^2 + \frac{\lambda_{\max}\kappa'}{k}r^3 + \frac{\frac{\kappa'^2}{k}\lambda_{\max}^2}{\lambda_{\min}^2-\kappa'^2r^2}r^4 + E_{2,r}}{\frac{\lambda_{\min}^2-\kappa'^2r^2}{k}r^2 - \sigma^2\mathbf{1}_{\overline{n}\leq CD} - \sigma\sqrt{\frac{D}{\overline{n}}}t\left[\frac{2\kappa'r^2}{\sqrt{k}}+\sigma\mathbf{1}_{\overline{n}\geq CD}+\sigma\sqrt{\frac{D}{\overline{n}}}\mathbf{1}_{\overline{n}\leq CD}\right] - E_{2,r}} \tag{3.14}$$

This result implies Theorem 1, which explores the regime we are most interested in, specifically $\overline{n} \approxeq k\cdot\log k\cdot\log D \ll D$ with $\sigma\sqrt{D} = O(1)$. It is trivial to obtain other results in other interesting regimes, for example for $n \to +\infty$ with k, D fixed (albeit in this case our results are not sharp, in the sense that when $\overline{n} \gg D$ the terms $O(r^3)$ and $O(r^4)$ would start decreasing with rate $\overline{n}^{-\frac{1}{2}}$). In random matrix theory and free probability a regime of interest is when $D, \overline{n} \to +\infty$ with fixed ratio $D/\overline{n} = \phi$. In our context that would correspond to fixing the ratio between D and $\overline{n}$ to ϕ, e.g. setting $\overline{n} \sim \frac{D}{\phi\frac{t^2 k\log k}{\lambda_{\max}^2}}$.

Proof of Theorem 2. The proof follows from modifying the proof of Prop. 1 to include one final perturbation, the perturbation given in Prop. 2, which may be upper-bounded by

$$v^2\left(\underbrace{\left(1+\frac{s^2\sigma\sqrt{D}}{r}+\left(1\vee\frac{\sigma^2 D}{r^2/k}\right)\sqrt{\log\frac{r}{3\sigma k}}\right)\frac{\sigma k}{r}}_{=:\beta_s}\vee\frac{1}{\overline{n}}\right)r^2 \leq v^2\left(\frac{k\beta_s}{\lambda_{\min}^2}\vee\frac{k}{\lambda_{\min}^2\overline{n}}\right)\frac{\lambda_{\min}^2 r_=^{\ 2}}{k},$$

with probability at least as in Proposition 2. From the proof of Proposition 1, one obtains that under the conditions of Propositions 1 and 2, if $r_=$ satisfies (3.13) then $\Delta_k(\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}))$ is the largest gap with probability as claimed. Solving the above for $r_=$ completes the proof. The proof of (iii) follows similarly, by reasoning as in the proof of (iii) in Proposition 1 and adding the perturbation of Proposition 2. □

Proof of Theorem 1. This follows directly from Theorem 2 for D large, and choosing $v^2\mu_X(B_z(r_=))n = s^4 = t^2$. One then sees that

$$\beta_s \leq C\frac{\sigma_0 k}{r} \quad , \quad \varphi^2 \geq \left(\lambda_{\min}^2 - \delta^2\lambda_{\max}\epsilon - \frac{\epsilon^2}{\lambda_{\min}^2}\left(\frac{C\sigma_0 k}{r}\vee\frac{1}{\overline{n}}\right) - \frac{\sigma_0\kappa'}{t}\right)\frac{1}{2(1+\epsilon)}.$$

In order to prove (iii), we start from (3.14) and use the assumptions to simplify and upper bound various terms, and in particular notice that $E_{2,r} \leq \frac{C\epsilon^2 r^2}{k}\left(\frac{\sigma_0 k}{r}\vee\frac{1}{\overline{n}}\right)$. □

Example 2. *For the unit sphere $\mathbb{S}^k$, we have $\lambda_{\min} = \lambda_{\max} \sim 1$, $\kappa \sim k^{-\frac{1}{2}}$, $\kappa' \sim 1$, $R_{\min} = 0$, and $R_{\max} \sim 1$.*

3.5 Varying the scale

For a fixed point z, one may discretize the range of good scales in the results above at multiple values $\{r_j\}$ of r, and consider the behavior of $\lambda_i(\mathrm{cov}(X_{z,r_j}))$ and its empirical and noisy versions. One may then apply the results above for each $r = r_j$ and by taking union bounds derive bounds on the behavior of $\mathrm{cov}(X_{z,r_j})$ for fixed z, as a function of j.

In practice, in our application to estimation of intrinsic dimension, we do the above and determine the intrinsic dimension by detecting which eigenvalues of $\sqrt{\mathrm{cov}(X_{z,r})}$ grows linearly in r (those corresponding to the intrinsic dimension), quadratically (those corresponding to curvature directions), and which ones do not grow (those corresponding to noise), and in which range of scales this holds.

4 The manifold case

Consider a smooth compact Riemannian manifold $\mathcal{M}$ of dimension $\dim(\mathcal{M})$ isometrically embedded in $\mathbb{R}^D$, endowed with its volume measure denoted by vol. We let $\mu_X = \mathrm{vol}$ in $\mathbb{R}^D$, normalized to be a probability measure. The usual assumptions are satisfied, with $k = k(z) = \dim(\mathcal{M})$, $v_{\min}$ dependent on $\mathrm{vol}(\mathcal{M})$ and upper bounds on the curvature of $\mathcal{M}$, under rather general conditions on $\mathcal{M}$. In this case $P^{[z,r]}$ may be chosen to be the projection on the k-dimensional tangent plane to $\mathcal{M}$ at z, translated along the normal direction to $\mathcal{M}$ at z to ensure that $\mathbb{E}[X_{z,r}{}^{\perp}] = 0$. Following H. Federer, let

$$\mathrm{reach}(\mathcal{M}) = \sup\{r \geq 0 \,:\, \mathrm{tub}_r(\mathcal{M}) \subset D(\mathcal{M})\}\,,$$

where $D(\mathcal{M}) = \{y \in \mathbb{R}^D \,:\, \exists!\; x \in \mathcal{M} : ||x - y|| = \min_{x' \in \mathcal{M}} ||x' - y||\}$ and $\mathrm{tub}_r(\mathcal{M}) = \{y \in \mathbb{R}^D \,:\, d(y, \mathcal{M}) < r\}$. If $\mathrm{reach}(\mathcal{M}) > 0$ then we may choose $R_{\min} = 0$ and $R_{\max} = \mathrm{reach}(\mathcal{M})$ for every $z \in \mathcal{M}$.

Remark 1. *Since our constructions and results are of a local nature (with the only assumption of global character being on the* reach*), it is clear how to generalize the setting above to the case of non-compact manifolds, manifolds with boundaries, and measures different from the volume measure.*

Remark 2. *We may choose a measure μ_X on $\mathcal{M}$ which is mutually absolutely continuous with respect to* vol*, and the usual assumptions will still be satisfied, at least locally, depending on the bounds on the Radon-Nykodim derivative $\frac{d\mu_X}{d\mathrm{vol}}$.*

Remark 3. *The usual assumptions on $X_{z,r}{}^{\perp}$ allow for a lot of flexibility in the model: for example we could have a manifold $\mathcal{M}$ as above, "corrupted" by complicated structures in the normal directions, which are small in the sense of our usual assumptions on $X_{z,r}{}^{\perp}$.*

Finally, we observe that the eigenvalues of $\mathrm{cov}(X_{z,r})$ and the corresponding eigenspaces vary smoothly as a function of r (and z!), and we may therefore smooth the empirical S.S.V.'s $\lambda_i^2(\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{z},r]}))$, for fixed i and $\tilde{z}$, as a function of r. Such denoising of the S.S.V.'s, by taking into account that their expected growth as a function of r is r^2 (for the top k) and r^4 (for the curvature ones), is justified by the near-independence of the covariance matrices across well-separated scales.

4.1 The case of a manifold with co-dimension 1

Let $\mathcal{M}$ be a k-dimensional manifold embedded in $\mathbb{R}^{k+1}$. Fix a point $z \in \mathcal{M}$. Let $\kappa_1, ...\kappa_k$ be the principal curvatures of $\mathcal{M}$ at z. In appropriate coordinates $(x_1, x_2, ..., x_k, y)$, $\mathcal{M}$ is locally given by $y = f(x)$, where:

$$f(x) = \frac{1}{2}(\kappa_1 x_1^2 + ... + \kappa_k x_k^2) + O(||x||^3), \tag{4.1}$$

that is, the second order Taylor expansion of f is quadratic with coefficients given by the principal curvatures [57]. We start by approximating $\mathcal{M} \cap B_z(r)$ by a set over which integration will be simpler: for small r, $X_{0,r} := \{(x, f(x)) : ||(x, f(x))||_{\mathbb{R}^{k+1}} \leq r\}$ satisfies

$$\{(x, f(x)) : x \in B^k\,(r_{\min})\} \subseteq X_{0,r} \subseteq \{(x, f(x)) : x \in B^k\,(r_{\max})\}$$

where $r_{\{\min,\max\}} := r\sqrt{1 - 4^{-1}\kappa^2_{\{\max,\min\}} r^2}$, $\kappa_{\min} = \min_i \kappa_i$ and $\kappa_{\max} = \max_i \kappa_i$. The difference between the sets involved is small and will be disregarded, since it would only produce terms which have higher order in $||x||$ than those we are estimating. The volume element is given by

$$d\mathrm{vol}(x) = \sqrt{1 + ||\nabla f||^2} = 1 + \frac{1}{2}\sum_{i=1}^{k} \kappa_i^2 X_i^2 + O(||x||^4)\,,$$

so that, up to higher order terms in r^2, denoting the Lebesgue measure $\mu_{\mathbb{R}^k}$ by $|\cdot|$,

$$|B^k(r)|\left(1 + \frac{k}{2(k+2)}\kappa^2_{\min} r^2\right) \leq \mathrm{vol}(X_{0,r}) \leq |B^k(r)|\left(1 + \frac{k}{2(k+2)}\kappa^2_{\max} r^2\right).$$

Therefore $\frac{\mathrm{vol}(X_{0,r})}{|B^k(r)|} = 1 + O(r^2)$ and $\frac{|B^k(r)|}{\mathrm{vol}(X_{0,r})} = 1 - O(r^2)$, and we discard the higher order factors as they will not affect the calculations that follow. The first k squared singular values corresponding to of $X_{0,r}$ are computed as follows: by symmetry considerations the matrix of second moments is diagonal, up to second order in $||x||$, in the chosen coordinates. For r small, disregarding $O(||x||^4)$ terms, and for $l = 1, \ldots, k$,

$$\begin{aligned}(\lambda_l^{[0,r]})^2(f) &= \frac{1}{\mathrm{vol}(X_{0,r})} \int_{B^k(r)} x_l^2 \sqrt{1 + ||\nabla f||^2} dx = \frac{|B^k(r)|}{\mathrm{vol}(X_{0,r})} \frac{1}{|\mathbb{B}^k|} \int_{\mathbb{B}^k} x_l^2 \left(1 + \frac{1}{2} \sum_i \kappa_i^2 X_i^2\right) dx \\ &= \sigma_i^2(\mathbb{B}^k) r^2 + O(r^4) = \sigma_1^2(\mathbb{B}^k) r^2 + O(r^4)\end{aligned}$$

with $\mathbb{B}^k$ the unit k-dimensional ball. Similarly, for $(\lambda_{k+1}^{[0,r]})^2(f)$, we have

$$\begin{aligned}(\lambda_{k+1}^{[0,r]})^2(f) &= \frac{|B^k(r)|}{\mathrm{vol}(X_{0,r})} \left[\frac{1}{|B^k(r)|} \int_{B^k(r)} f(x)^2 d\mathrm{vol}(x) - \frac{|B^k(r)|}{\mathrm{vol}(X_{0,r})} \left(\frac{1}{|B^k(r)|} \int_{B^k(r)} f(x) d\mathrm{vol}(x) \right)^2 \right] \\ &= \left[\frac{1}{4} \sum_i \kappa_i^2 \left(\xi_{ii}^4(\mathbb{B}^k) - \lambda_i^4(\mathbb{B}^k) \right) + \frac{1}{2} \sum_{i<j} \kappa_i \kappa_j \left(\xi_{ij}^4(\mathbb{B}^k) - \lambda_i^2 \lambda_j^2(\mathbb{B}^k) \right) \right] r^4 + O(r^6)\end{aligned} \tag{4.2}$$

where $\xi_{ij}^4(\mathbb{B}^k) = \frac{1}{|\mathbb{B}^k|} \int_{\mathbb{B}^k} X_i^2 x_j^2 dx$. Since the second and fourth moments of $\mathbb{S}^{k-1}$, the $(k-1)$-dimensional unit sphere, are

$$\lambda_l^2(\mathbb{S}^{k-1}) = \frac{1}{k} \quad , \quad \xi_{lj}^4(\mathbb{S}^{k-1}) = \frac{1 + 2\delta_{lj}}{k(k+2)} \quad \text{for } l, j = 1, \ldots, k\,,$$

the corresponding moments of the unit ball are

$$\lambda_l^2(\mathbb{B}^k) = \frac{1}{k+2} \quad , \quad \xi_{lj}^4(\mathbb{B}^k) = \frac{1 + 2\delta_{lj}}{(k+2)(k+4)} \quad \text{for } l, j = 1, \ldots, k\,.$$

These may be compared, for large k, with the moments of $X \sim \mathcal{N}(0, \frac{1}{k} I_k)$ in $\mathbb{R}^k$, which are $\mathbb{E}[(X)_l^2] = \frac{1}{k}$, $\mathbb{E}[(X)_l^2 (X)_j^2] = \frac{1+2\delta_{lj}}{k^2}$. We may simplify (4.2):

$$(\lambda_{k+1}^{[0,r]})^2(f) = \frac{1}{(k+2)^2(k+4)} \left[\frac{k+1}{2} \sum_{i=1}^k \kappa_i^2 - \sum_{1 \le i < j \le k} \kappa_i \kappa_j \right] r^4\,. \tag{4.3}$$

The gap between $(\lambda_l^{[0,r]})^2(f)$, for $l = 1, \ldots, k$ and $(\lambda_{k+1}^{[0,r]})^2(f)$ is large when this last expression is small. Considering the scaling as a function of k, we see that $(\lambda_l^{[0,r]})^2(f)$ always has the natural scaling $k^{-1} r^4$, as in pur usual geometric assumptions, while for $(\lambda_{k+1}^{[0,r]})^2(f)$ we observe that:

(i) In this context, the constant κ in our geometric assumptions may be chosen equal to $\kappa_{\max} := \max_i |\kappa_i|$. If this is independent of k, equation (4.3) implies that $(\lambda_{k+1}^{[0,r]})^2(f)$ scales at most like $k^{-1} r^4$ (as in geometric assumptions), since the term in square brackets scales at most like k^2. This guarantees a spectral gap in the covariance of size independent of k.

(ii) There are cases where a combination of sizes and signs of the κ_i's cause the term in square brackets in (4.3) to be $O(k)$, and thus κ will scale like $k^{-\frac{1}{2}}$. This happens for example for the unit sphere $\mathbb{S}^{k-1}$, or for the hyperbolic paraboloid with $\kappa_1, \ldots, \kappa_{k-1} = 1$ and $\kappa_k = -1$, as discussed in the next section.

4.1.1 Example: the sphere $\mathbb{S}^k$ and hyperbolic paraboloids

The expression (4.3), in the case of the k-dimensional unit sphere $\mathbb{S}^k$, yields that in a small neighborhood of any point, for $l = 1, \ldots, k$

$$(\lambda_l^{[z,r]})^2(\mathbb{S}^k) = \frac{1}{k+2} r^2 \sim \frac{r^2}{k} \quad , \quad (\lambda_{k+1}^{[z,r]})^2(\mathbb{S}^k) = \frac{k}{(k+2)^2(k+4)} r^4 \sim \frac{r^4}{k^2}$$

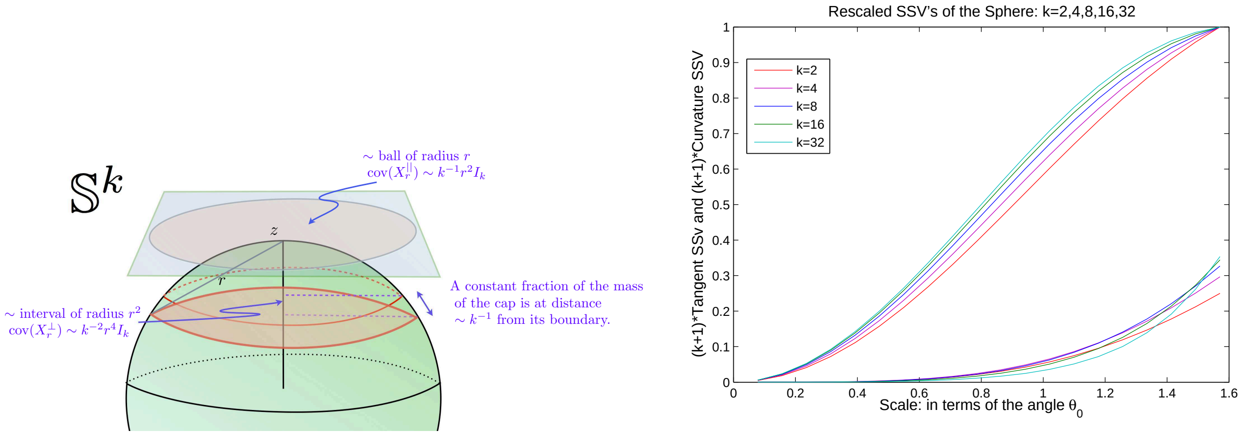

Figure 4: Left: Because of concentration of measure phenomena, in the case of $\mathbb{S}^k$ the size of our notion of curvature κ^2 is small, both in the sense that it scales like k^{-1} as a function of k, and it stays small on large neighborhoods (of size $O(1)$). We take advantage of the same phenomenon when estimating the effect of the noise in the proof of Prop. 2 in Appendix 9. Right: plot of $(k+1)\lambda^2_{l,z,r}(\mathbb{S}^k)$ for z equal to the north pole, as a function of the angle subsumed by the cap $B_z(r) \cap \mathbb{S}^k$, for different values of k: we see that up to a scale $O(1)$ independent of k the $k+1$-st S.S.V. is much smaller than the top k.

In particular, this implies that the curvature quantity κ^2 in (3.6) scales like k^{-1}, as a function of k. In the case of a hyperbolic paraboloid with $\kappa_1, \dots, \kappa_{k-1} = 1$ and $\kappa_k = -1$, we obtain from (4.3):

$$(\lambda_{k+1}^{[z,r]})^2(f) = \frac{(3k-2)}{(k+2)^2(k+4)} r^4 \sim \frac{3r^4}{k^2}.$$

Again, this implies the curvature quantity κ in (3.6) scales like $k^{-\frac{1}{2}}$.

If k is even and we have a hyperbolic paraboloid with $\kappa_1, \dots, \kappa_{\frac{k}{2}} = 1$ and $\kappa_{\frac{k}{2}+1}, \dots, \kappa_k = -1$, we obtain from (4.3):

$$(\lambda_{k+1}^{[z,r]})^2(f) = \frac{k(k+1)}{2(k+2)^2(k+4)} r^4 \sim \frac{r^4}{2k}$$

Here we have κ bounded independently of k, which is sufficient for the number of samples to be linear in k.

Regarding $v_{\min}$, we have $R_{\min} = 0$ and we may choose $v_{\min} = (\mu_{\mathbb{R}^k}(\mathbb{S}^k))^{-1}$; therefore

$$v_{\min}\mu_{\mathbb{R}^k}(\mathbb{B}^k) = \frac{\mu_{\mathbb{R}^k}(\mathbb{B}^k)}{\mu_{\mathbb{R}^k}(\mathbb{S}^k)} = \frac{\mu_{\mathbb{R}^k}(\mathbb{B}^k)}{(k+1)\mu_{\mathbb{R}^k}(\mathbb{B}^{k+1})} \sim \frac{1}{\sqrt{k}}$$

A more intuitive way (which may be readily formalized using the ideas underlying Lemma 9) to find the scaling, with respect to k, of the squared singular values of the sphere $\mathbb{S}^k$ is the following (see Figure 4). We start by observing that $1-\delta$ of the mass of V_r^k is concentrated in a ring of thickness $\sim_\delta k^{-1}$ at the boundary of the cap $V_r^k = B_{(0,\dots,0,1)}(r) \cap \mathbb{S}^k$ (since the volume of V_r^k, as a function of r, grows like r^k). Therefore the projection of V_r^k onto the tangent plane at z will have covariance comparable to that of an annulus of thickness $\sim rk^{-1}\cos\theta_0$ and dimension k and radius $r\cos\frac{\theta_0}{2}$, which behaves like $k^{-1}r^2 I_k$ for $\theta_0 \lesssim 1$. This determines the scaling for $(\lambda_1^{[z,r]})^2, \dots, (\lambda_k^{[z,r]})^2$. As for the scaling of $(\lambda_{k+1}^{[z,r]})^2$, it is the variance of the projection of V_r^k onto the axis normal to the tangent plane at z: at least for r not to small, this is a measure concentrated on an interval of size $\sim k^{-1}r\sin\theta_0$, which has variance $\sim k^{-2}r^4$. Observe that this reasoning implies that the "curvature" κ we use may be small in a neighborhood much larger than one may expect.

4.2 The case of a manifold with general codimension

The case of a manifold of general co-dimension could be treated in similar fashion: let $\mathcal{M}$ be a k-dimensional manifold embedded in $\mathbb{R}^D$. In appropriate coordinates $(x_1, \dots, x_k, y_1, \dots, y_{D-k})$ in $\mathbb{R}^k \oplus \mathbb{R}^{D-k}$, $\mathcal{M}$ is locally given by $y = f(x) + O(||x||^3)$, where:

$$f(x) = \frac{1}{2}\left(x^T H_1 x, \dots, x^T H_{D-k} x\right) \tag{4.4}$$

where $H_1, \dots, H_{D-k} \in \mathbb{R}^{k \times k}$ are the Hessians of $f = (f_1, \dots, f_{D-k}) : \mathbb{R}^k \to \mathbb{R}^{D-k}$. The computation for the first k multiscale singular values proceeds as above, yielding once again that they do not depend on the curvatures of $\mathcal{M}$. For the remaining multiscale singular values, we proceed as follows. Let us consider the $(k+1)$-st multiscale singular value: it corresponds to an eigenvector v_{k+1} orthogonal to $\langle x_1, \dots, x_k \rangle$ (which is the span of the first k multiscale singular vectors) and it is a direction of maximal variance for $\mathcal{M}$ in $\langle y_1, \dots, y_{D-k} \rangle$. In other words, v_{k+1} is the direction of a unit vector w maximizing the variance of

$$f_w(x) := \langle f(x), w \rangle = \langle (x^T H_1 x, \dots, x^T H_{D-k} x), w \rangle = x^T = x^T \left(\sum_{l=1}^{D-k} w_l H_l \right) x \,,$$

which, by the calculations for the case of codimension 1, is given by

$$\frac{1}{2(k+2)(k+4)} \left[\left\| \sum_{l=1}^{D-k} w_l H_l \right\|_F^2 - \frac{1}{k+2} \left(\sum_{l=1}^{D-k} w_l \mathrm{Tr}(H_l) \right)^2 \right] r^4 \,. \tag{4.5}$$

Also, observe that while the codimension of $\mathcal{M}$ is as large as $D-k$, the range of f above is no more than $k(k+1)/2$-dimensional, since $\dim(\mathrm{span}\{H_1, \dots, H_{D-k}\}) \leq k(k+1)/2$. This implies that the rank of f as above is in fact at most $\binom{k+1}{2}$. Therefore, we expect at most $\binom{k+1}{2}$ squared singular values of size $O(r^4)$, due to the various curvatures, obtained by maximizing (4.5) over increasingly smaller subspaces orthogonal to the already constructed tangent and curvature directions.

Finally, we observe that similar calculations may be extended to classes of manifolds which are less than $\mathcal{C}^2$, for example to manifolds that are locally graphs of $\mathcal{C}^\alpha$ functions, by replacing Taylor expansions by residuals that are Hölder of the appropriate order. This is because our notions of tangent approximations and curvatures are L^2 notions, which are well-defined and stable even in situations were there is a lack of smoothness.

5 An algorithm for estimating intrinsic dimension

The results above suggest the following algorithm: for each $\tilde{z}$ in the training set and $r > 0$, we compute the eigenvalues $(\tilde{\lambda}_i^{[\tilde{z},r]})^2$, $i = 1, \dots, D$, of $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{z},r]})$. When r is large, if $\mathcal{M}$ is contained in a linear subspace of dimension K ($K \geq k$) we will observe K large eigenvalues and $D-K$ smaller noise eigenvalues (we will assume that $K < D$). Clearly, $k \leq K$. Moreover, $\{(\tilde{\lambda}_i^{[\tilde{Z},r]})^2\}_{i=K+1,\dots,D}$ will be highly concentrated and we use them to estimate σ, which is useful per se. Viewing $\{(\tilde{\lambda}_i^{[\tilde{Z},r]})^2\}_{i=K+1,\dots,D}$ as a function of r, we identify an interval in r where the noise is almost flat, thereby removing the small scales where the distortion due to noise dominates. From this point onwards the algorithm will work on this restricted interval. We look at the first $\{(\tilde{\lambda}_i^{[\tilde{Z},r]})^2\}_{i=1,\dots,K}$, and the goal is to decide how many of them are due to the extrinsic curvature of $\mathcal{M}$. But the curvature S.S.V.'s grow with rate at most r^4, while the "tangential" (non-curvature) S.S.V.'s grow with rate r: a least-square fit to $\{(\tilde{\lambda}_i^{[\tilde{Z},r]})^2\}$, as a function of r, is used to tell the curvature S.S.V.'s from the tangential ones, yielding our estimate for k. Finally, we estimate $[\hat{R_{\min}}, \hat{R_{\max}}]$ as the largest interval of r^2's in which $\tilde{\Delta}_{\hat{k}}^{[z,r]} := (\tilde{\lambda}_{\hat{k}}^{[\tilde{Z},r]})^2 - (\tilde{\lambda}_{\hat{k}+1}^{[\tilde{Z},r]})^2$ is the largest gap.

The many details and available options are documented in the code[2].

[2] `www.math.duke.edu/~mauro`

```
$[\hat{k}, \hat{R_{\min}}, \hat{R_{\max}}]$ = EstDimMSVD ($\tilde{\mathbf{X}}_n, \tilde{z}, K_{\max}$)

// **Input:**
// $\tilde{\mathbf{X}}_n$ : an $n \times D$ set of noisy samples
// $\tilde{z}$ : a point in $\tilde{\mathbf{X}}_n$
// $K_{\max}$ : upper bound on the intrinsic dimension $k$
// **Output:**
// $\hat{k}$ : estimated intrinsic dimension at $\tilde{z}$
// $(\hat{R_{\min}}, \hat{R_{\max}})$ : estimated interval of good scales

$\{\hat{k}_1, (\tilde{\lambda}^{[z,r]}_{\hat{k}_1+1})^2\}$ ← FindLargestNoiseSingularValue($\tilde{\mathbf{X}}_n$, $\tilde{z}$)
$\hat{R_{\min}}$ ← Smallest scale for which $(\tilde{\lambda}^{[\tilde{z},r]}_{\hat{k}_1+1})^2$ is decreasing and $|B_z(\hat{R_{\min}})| \gtrsim K_{\max} \log K_{\max}$
$\hat{R_{\max}}$ ← Largest scale $> R_{\min}$ for which $(\tilde{\lambda}^{[\tilde{z},r]}_1)^2$ is nonincreasing
$\hat{k}$ ← Largest $i$ such that:
    · for $r \in (\hat{R_{\min}}, \hat{R_{\max}})$, $(\tilde{\lambda}^{[\tilde{z},r]}_i)^2$ is linear and $(\tilde{\lambda}^{[\tilde{z},r]}_{i+1})^2$ is quadratic in $r$, and
    · $\Delta^{[\tilde{z},r]}_i$ is largest gap for $r$ in a large fraction of $(\hat{R_{\min}}, \hat{R_{\max}})$
$(\hat{R_{\min}}, \hat{R_{\max}})$ ← Largest interval in which $\Delta^{[\tilde{z},r]}_{\hat{k}}$ is the largest gap
```

Figure 5: Pseudo-code for the Intrinsic Dimension Estimator based on multiscale SVD.

5.1 Algorithmic and Computational Considerations

Instead of computing $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{z},r]})$ for every $\tilde{z}, r$, we perform a subsampling in scale and in space, as follows. A set $\Gamma \subset \tilde{\mathbf{X}}_n$ is called a δ-net in $\tilde{\mathbf{X}}_n$ if $\{B_z(2\delta)\}_{z\in\Gamma}$ covers $\tilde{\mathbf{X}}_n$ and any pair of points in Γ has distance larger than δ. We select an increasing sequence $0 \le \delta_0 < \cdots < \delta_j < \ldots$ with $\delta_j \to \infty$, and for every j we construct a δ_j-net, called Γ_j. The construction of multiscale nets is of general interest, we refer the interested reader to [58, 59] and references therein. For example, we may choose $\delta_j = 2^j\delta_0$, or in such a away that $\mathbb{E}_X[|B_x(\delta_j)|]$ grows linearly in j, and stop at the smallest level J s.t. $|\Gamma_J| < 10$, say. Here and in what follows $|B_x(\delta_j)|$ is the number of samples in $B_x(\delta_j)$. We compute $(\tilde{\lambda}^{[z,r]}_i)^2$ for $r = \delta_0, \ldots, \delta_J$ and, for $r = \delta_j$, $z \in \Gamma_j$. Here i may range from 1 up to $I := \min\{D, n_r, K\}$, the maximum rank of $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{z},r]})$, where K is a pre-specified parameter (that may be $D \wedge n$). We therefore obtain a discretization of the continuous (in space and scale space) quantities $(\tilde{\lambda}^{[z,r]}_i)^2$. Note that we still get an estimate of the intrinsic dimension *at every point*: given an $x \in \tilde{\mathbf{X}}_n$, at each scale j, we associate x with the $z_{x,j} \in \Gamma_j$ that is closest to it, and we approximate $(\tilde{\lambda}^{[x,\delta_j]}_i)^2$ by $(\tilde{\lambda}^{[z_{x,j},\delta_j]}_i)^2$. In order to avoid artifacts due to the randomness of Γ_j, one may repeat this construction a few times and take the expectation (or vote), over the runs, of all the quantities we are interested in. The cost of computing $\{(\tilde{\lambda}^{[z,r_j]}_i)^2\}_{i=1,\ldots,I}$ by [60] is $O(D \cdot |B_z(r_j)| \cdot (I + C_{\mathrm{nn}}))$, where C_{nn} is the (normalized) cost of computing a nearest neighbor, which, after the preprocessing step of constructing the multiscale nets, is $O(2^{ck}\log n)$, where c is a universal constant (e.g. [59] and references therein). The procedure is repeated $O(n/|B_z(r_j)|)$ times at each scale (for each $z \in \Gamma_j$), and then across all scales $j = 0, \ldots, J$, with $J = O(\log n)$, for a total cost of $O(D \cdot n\log n \cdot (I + C_{\mathrm{nn}}))$. In the worst case, $I = \min\{D, n, K\}$, yielding $O(D \cdot n\log n \cdot (\min\{D, n, K\} + C_{\mathrm{nn}}))$. Finally, we observe that our algorithm is very highly parallelizable and easily distributable.

We have run rather extensive comparisons with other algorithms, see Figure 6.

5.2 Experimental Results

5.2.1 Manifolds

We test our algorithm on several data sets obtained by sampling manifolds, and compare it with existing algorithms. The test is conducted as follows. We fix the ambient space dimension to $D = 100$. We let $\mathbb{Q}^k$, $\mathbb{S}^k$, $\mathcal{S}$,

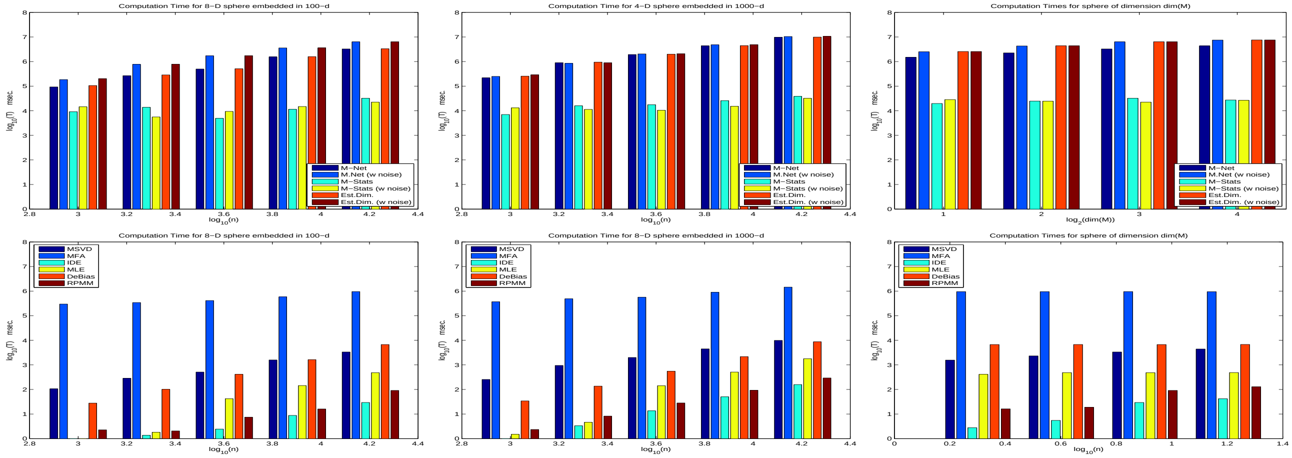

Figure 6: Top: Timing experiments for our algorithm: time to construct the multiscale nets ('M-Net'), calculation of multiscale statistics ('M-Stats') and the total time ('Est.Dim.'). All plots are in log-log scale. Left: time vs. n for $\mathbb{S}^k(n, D, \sigma)$, for $n = 1000, 2000, 4000, 8000, 16000$, $k = 8$, $D = 100$, and $\sigma = 0, \frac{0.25}{\sqrt{D}}$. Times grow linearly in n, with the noise slightly increasing the computational time of each sub-computation. Center: same as left, but with $D = 1000$: the increased ambient dimensionality does not cause, in this instance, almost any increase in time, not even by the meager factor of 10, which one would expect from the cost handling vectors which are 10 times larger in distance computations. In particular, no curse of ambient dimensionality appears. Right: computation times as a function of intrinsic dimension $k = 2, 4, 8, 16$, and notice a mild increase in computation time. Tests were run on an Apple® Mac Pro with 2×2.93Ghz Quad-Core Intel Xeon® processors, 32 Gb of RAM, and Matlab® 7.10 with parallel mode enabled (the time reported is total CPU time across all CPU's). Absolute values on the y-axis should not be taken too seriously. Bottom:Comparison of running time between our algorithm and competitors (with the parameters set as in all other experiments). "RTPMM" and "Smoothing" had complexity that grew too quickly in n to make their inclusion practical. The same applies to "MFA" (3 orders of magnitude slower than "MSVD"), so we ran 500 times faster by reducing the number of iterations/initializations (with respect to the default value of these parameters), and, assuming a constant cost per iteration, multiplied the running time back by 500.

$\mathcal{Z}^k$ be, respectively, the unit k-dimensional cube, the k-dimensional sphere of unit radius, a manifold product of an S-shaped curve of roughly unit diameter and a unit interval, and the Meyer's staircase $\{\chi_{0,k}(\cdot - l)\}_{l=0,\dots,D}$. Each of these manifolds is embedded isometrically in $\mathbb{R}^K$, where $K = k$ for $\mathbb{Q}^k$, $K = k+1$ for $\mathbb{S}^k$, $K = 3$ for $\mathcal{S}$ and $K = D$ for $\mathcal{Z}^k$, and $\mathbb{R}^K$ is embedded naturally in $\mathbb{R}^D$. Finally, a random rotation is applied (this should be irrelevant since all the algorithms considered are supposed to be invariant under isometries); n samples are drawn uniformly (with respect to the volume measure) at random from each manifold, and noise $N \sim \sigma\mathcal{N}(0, I_D)$ is added. We incorporate these parameters in the notation by denoting $\mathbb{Q}^k(n, \sigma)$ the set of n samples obtained as above, where the manifold is the k-dimensional unit cube and the noise has variance σ (and analogously for the other manifold considered). We also consider a variant of these sets, where we dilate $\mathbb{R}^K$, after embedding the manifold, but before any other operation, by a diagonal dilation with factors drawn uniformly at random in the multiset $\{1, 1, 1, 1, 0.9, 0.9, 0.9, 0.8, 0.8\}$.

We consider here $k = 6, 12, 24, 48$ for $\mathbb{Q}^k$, $k = 5, 11, 23, 47$ for $\mathbb{S}^k$, and $l = 20$ for $\mathcal{Z}^l$. The samples size is set as $n = 250, 500, 1000, 2000$. We let the noise parameter $\sigma = 0, 0.01, 0.025, 0.05, 0.1$. For each combination of these parameters we generate 5 realizations of the data set and report the most frequent (integral) dimension returned by the set of algorithms specified below, as well as the standard deviation of such estimated dimension. We test the following algorithms, which include volume-based methods, TSP-based methods, and state-of-art Bayesian techniques: "Debiasing" [47], "Smoothing" [46] and RPMM in [61], "MLE" [62], "kNN" [63], "SmoothKNN" [64], "IDE", "TakEst", "CorrDim" [51], "MFA" [65], "MFA2" [66]. It is difficult to make a fair comparison, as several of these algorithms have one or more parameters, and the choice of such parameters is in general not obvious. We attempted to optimize the parameters of the algorithms by running them on several training examples, and we then fixed such parameters. The Bayesian algorithm "MFA" of [65], which implicitly estimates intrinsic dimension, was run on the test set by the authors of [65], given the knowledge that no data set would have intrinsic dimension larger than $K = 100$, which is the input to our algorithm. For "MFA2", the authors of [66] were also given access to the the code that we used to generate the manifolds used in order to fine tune the algorithm from [65] (but not the knowledge of the manifolds in the test set), and therefore were able to somewhat tune and modify their algorithms accordingly. While both "MFA" and "MFA2" were therefore given an advantage compared to the other methods, the results show that no advantage in terms of performance was achieved.

5.2.2 Varifolds

We consider a few simple examples of how the analysis and algorithm apply to the case when the dimension of the point cloud varies at different locations. We apply the analysis and the algorithm pointwise, to some toy examples: to a data set used as benchmark for several algorithms in [45], and to the data sets that we analyze, where the dimensionality is expected to change from point to point.

5.2.3 Real Data sets

In this section we describe experiments on publicly available real data sets and compare to previously reported results. We consider the MNIST database[3]containing several thousands images of hand written digits. Each image is 28 times 28 pixels. A mixture of ones and twos is considered in [44] and [63] who find $k = 11.26$ and $k = 9$ respectively. In Figure 13 we show the plot of the point-wise estimates at different points and the average. Figure 13 shows the same plot for different digits. In Table 5.2.3 we report the dimension estimated for each individual digit and compare with the smoothed Grassberger Procaccia estimator from [51] and the high rate vector quantization approach in [49].

Next we consider the IsoMap faces database[4] consisting of 698 images of size 64 times 64 pixels. We find an average intrinsic dimension $k = 2$ (Figure 13). [67] finds k between 3 and 4 (smaller values at large scales), [68] find $k \in [3.65, 4.65]$, [51] find an intrinsic dimension $k = 3$ using either Takens, Grassberger Procaccia or the Smoothed Grassberger Procaccia estimators, [69] find $k = 4$ and $k = 3$ depending on the way the point-wise estimates are combined (average and voting, respectively), and finally [44] find $k = 4.3$.

[3] `http://yann.lecun.com/exdb/mnist`

[4] `http://isomap.stanford.edu/dataset.html`

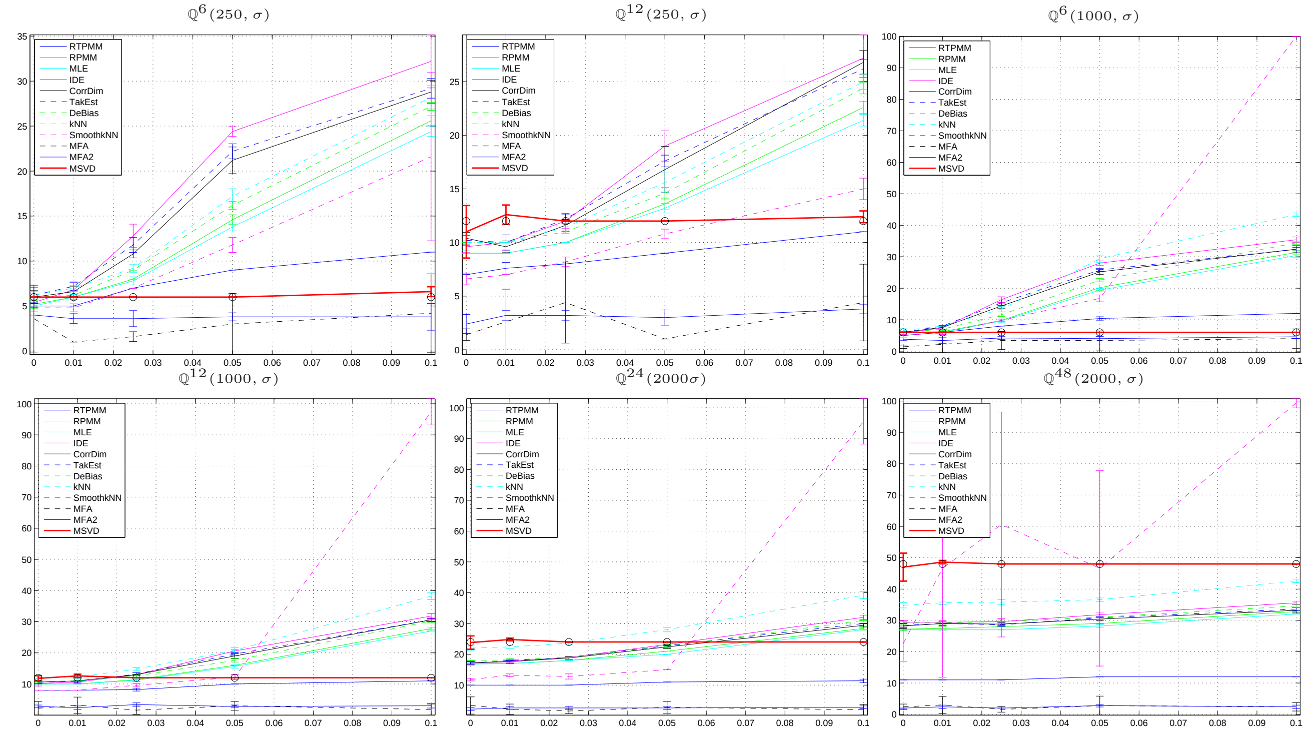

Figure 7: Benchmark data sets: cube.

The face video database[5] consists of 1965 images of size 20 by 28 pixels. We find an intrinsic dimension $k = 2$, see Figure 13. [49] estimate $k \in [4.25, 8.30]$.

Finally, we consider some data-sets whose intrinsic dimension has not been previously analyzed. The CBCL faces database (`http://cbcl.mit.edu`) contains 472 images of size 19 times 19 pixels. We find an intrinsic dimension $k = 2$, see Figure 13. The 20 news group is a set of 1161 documents represented as vectors in 1153 dimensions, and we find an intrinsic dimension $k = 9$, see Figure 13.

5.3 Linear Bi-Lipschitz perturbations

The following lemma characterize the effect of a linear Bi-Lipschitz perturbation of the data.

Lemma 1 (Bi-Lipschitz perturbations). *Suppose $\mathbf{X}_n = \{x_i\}_{i=1}^n$ is a (deterministic) set of n points in $\mathbb{R}^D$ with $\max_i ||x_i|| \leq r$. Let $\Phi : \mathbb{R}^D \to \mathbb{R}^d$ a linear map of $\mathbf{X}_n$ into $\mathbb{R}^d$ satisfying, for every x_i, x_j in $\mathbf{X}_n$, the bi-Lipschitz condition*

$$(1-\epsilon)||x_i - x_j||^2 \leq ||\Phi x_i - \Phi x_j||^2 \leq (1+\epsilon)||x_i - x_j||^2 \,. \tag{5.1}$$

Then:

$$|\lambda_i^2(\mathrm{cov}(\mathbf{X}_n)) - \lambda_i^2(\mathrm{cov}(\Phi(\mathbf{X}_n)))| \leq 4\epsilon r^2 \,.$$

The above result is straightforward, and we report a short proof for the sake of completeness.

Proof. Let $m = \frac{1}{n}\sum_{i=1}^n X_i$. The eigenvalues of $\mathrm{cov}(\mathbf{X}_n)$ are the same as those of the $n \times n$ matrix $\frac{1}{n}(\mathbf{X}_n - m \otimes 1)(\mathbf{X}_n - m \otimes 1)^T$, where $\mathbf{X}_n$ is the $n \times D$ matrix representing the point cloud, and similarly for $\mathrm{cov}(\Phi\mathbf{X}_n)$. Note

[5] `http://www.cs.toronto.edu/~roweis/data.html`

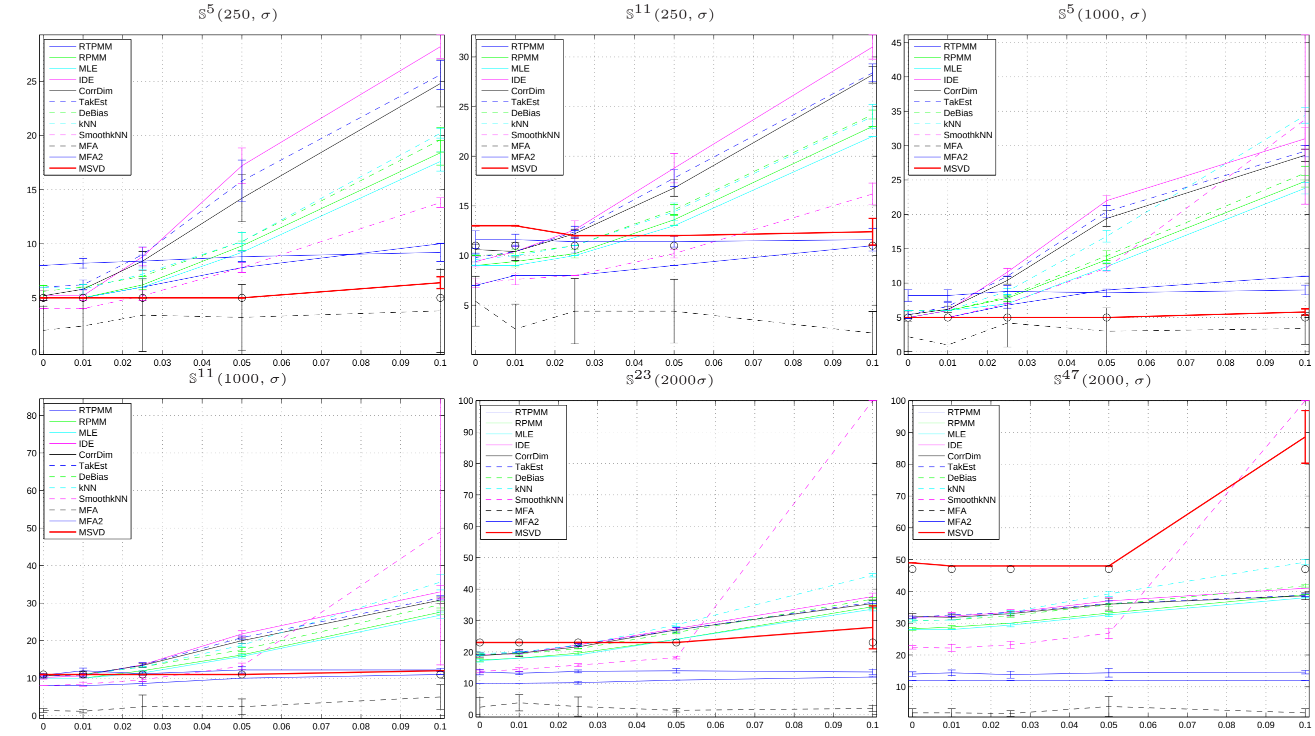

Figure 8: Benchmark data sets: sphere. Note the failures of our algorithm: at very high noise, at very small sample size compared to intrinsic dimension ($\mathbb{S}^{11}(250,\sigma), \mathbb{S}^{47}(2000,\sigma)$).

that $(\frac{1}{n}(\mathbf{X}_n - m \otimes 1)(\mathbf{X}_n - m \otimes 1)^T)_{i,j} = \frac{1}{n}\langle X_i - m, X_j - m\rangle$. Let $\mathcal{D} = \{x_i - x_j : x_i, x_j \in \mathbf{X}_n\}$ be the set of all differences between the points. $\Phi^T\Phi$ is close to the identity on the set $\mathcal{D}$:

$$\langle \Phi^T\Phi(x_i - x_j), x_i - x_j\rangle = \langle (I+E)(x_i - x_j), x_i - x_j\rangle = ||x_i - x_j||^2 \left(1 + \frac{\langle E(x_i - x_j), x_i - x_j\rangle}{||x_i - x_j||^2}\right)$$

Our bi-Lipschitz condition implies $\frac{|\langle E(x_i - x_j), x_i - x_j\rangle|}{||x_i - x_j||^2} \le \epsilon$ for all $x_i - x_j \in \mathcal{D}$. Because E is symmetric, $||E|_{\mathcal{D}}|| = \max_{z \in \mathcal{D}} \frac{|\langle Ez, z\rangle|}{||z||^2} \le \epsilon$. We may write $x_1 - \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}\sum_{i=1}^n (x_1 - x_i)$, and then estimate

$$\begin{aligned}\langle \Phi^T\Phi(x_1 - m), x_2 - m\rangle &= \langle \frac{1}{n}\sum_{i=1}^n \Phi^T\Phi(x_1 - x_i), x_2 - m\rangle = \langle \frac{1}{n}\sum_{i=1}^n (I+E)(x_1 - x_i), x_2 - m\rangle \\ &= \langle x_1 - m, x_2 - m\rangle + \frac{1}{n}\sum_{i=1}^n \langle E(x_1 - x_i), x_2 - m\rangle\end{aligned}$$

Since $|\frac{1}{n}\sum_{i=1}^n \langle E(x_1 - x_i), x_2 - m\rangle| \le ||E||_D(2r)(2r) = 4\epsilon r^2$, we have $\mathrm{cov}(\Phi(\mathbf{X}_n)) = \mathrm{cov}(\mathbf{X}_n) + \frac{1}{n}R$, where $|R_{i,j}| \le 4\epsilon r^2$ and R is $n \times n$. The upper bound $||R|| \le n||R||_{\max} \le 4\epsilon r^2 n$, where $||R||_{\max} = \max_{i,j}|R_{i,j}|$, implies

$$|\lambda_i^2(\mathrm{cov}(\mathbf{X}_n)) - \lambda_i^2(\mathrm{cov}(\Phi(\mathbf{X}_n)))| \le 4\epsilon r^2\,.$$

□

Example 3 (Johnson-Lindenstrauss [70]). *We can consider taking Φ to be a multiple of a random projection. In particular, let $P : \mathbb{R}^D \to \mathbb{R}^d$ be a projection onto a random (in the sense of [70]) d dimensional subspace of $\mathbb{R}^D$, and let $\Phi = \sqrt{\frac{D}{d}}P$.*

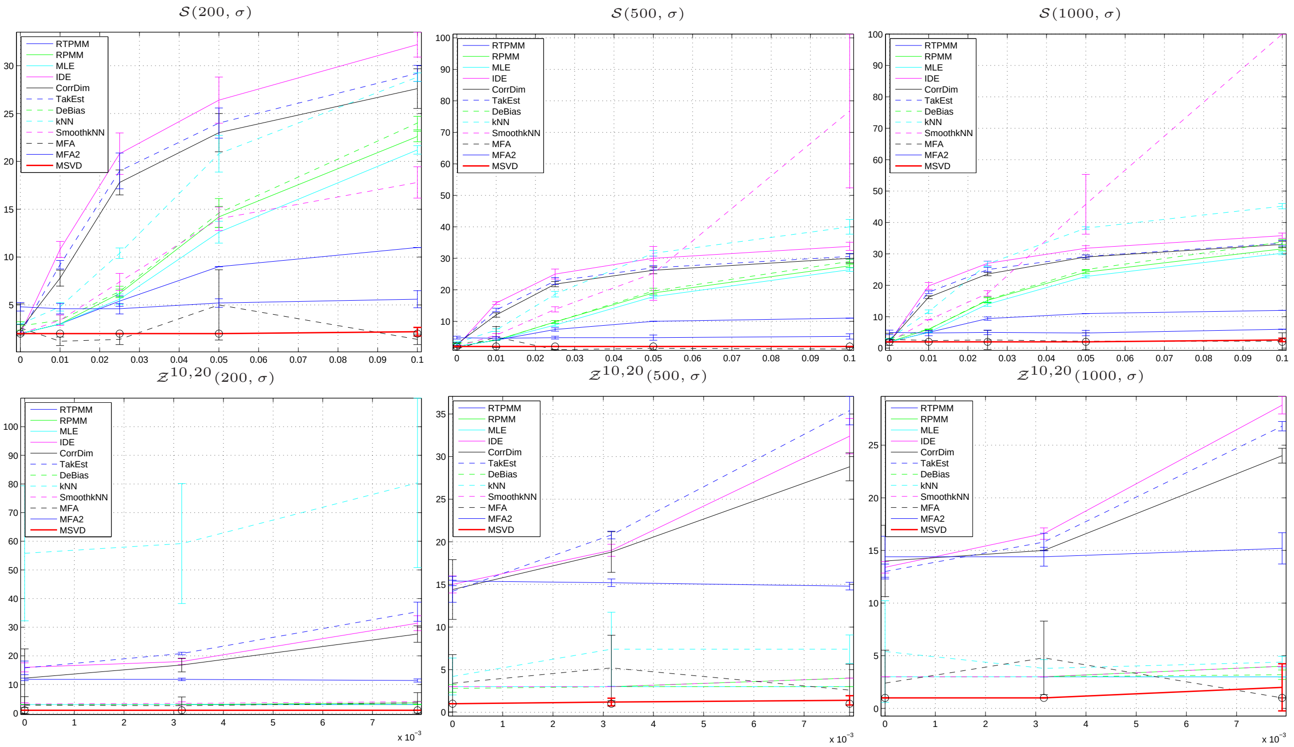

Figure 9: Benchmark data sets: S-shaped manifold $\mathcal{S}$ and Meyer's staircase $\mathcal{Z}$. The results for $\mathcal{Z}^{20}$ are consistently better than those for $\mathcal{Z}^{10}$, once fixed the number of points and the level of noise. This is consistent with the fact that $\mathcal{Z}^{20}$ has a smaller effective curvature than $\mathcal{Z}^{10}$.

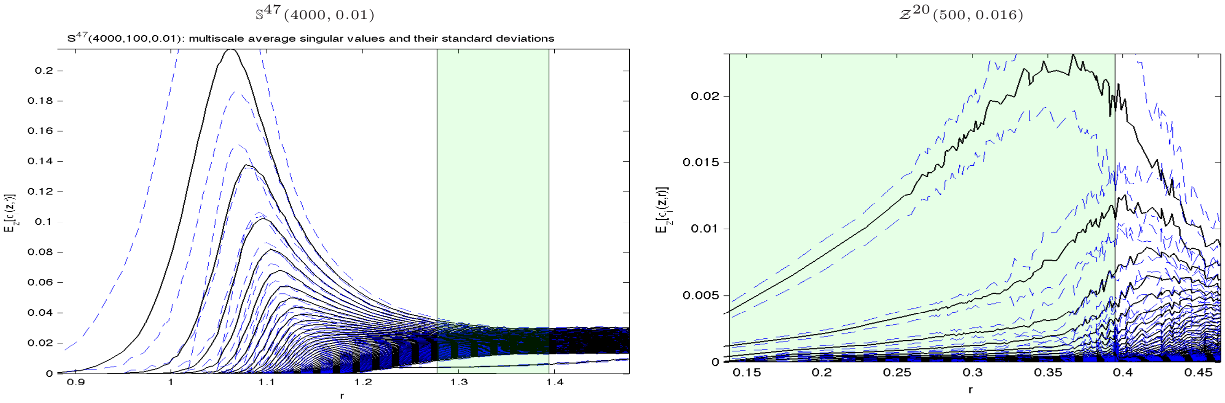

Figure 10: Two examples that pose difficulties. Left: $\mathbb{S}^{47}(4000,100,0.01)$ has large intrinsic dimension, even so, with only 4000 samples the curvature is visible even in presence of (small) noise, albeit hard to automatically detect. In this case the algorithm fails to identify a range of good scales. Right: $\mathcal{Z}^{20}(1000,1000,0.016)$ has multiple curvatures at multiple scales, and looks like a high-dimensional ball at large scale.

Then if

$$d \geq \frac{4\log n + \log(\frac{4}{\delta^2})}{\epsilon^2},$$

	RTPMM	RPMM	MLE	IDE	CorrDim	TakEst	DeBias	kNN	SmoothkNN	MFA	MFA2	MSVD
$\mathbb{Q}^6(1000, 0.00)$	5	5	5	**6**	5	5	**6**	**6**	4	1	4	**6**
$\mathbb{Q}^{12}(1000, 0.00)$	7	9	9	10	10	10	10	**12**	7	1	3	**12**
$\mathbb{Q}^{24}(1000, 0.00)$	9	16	16	17	17	17	17	20	11	1	2	**24**
$\mathbb{Q}^{48}(1000, 0.00)$	11	26	25	29	28	28	28	32	19	1	2	**48**
$\mathbb{S}^5(1000, 0.00)$	4	**5**	**5**	**5**	**5**	**5**	**5**	**5**	4	1	9	**5**
$\mathbb{S}^{11}(1000, 0.00)$	7	9	9	10	10	10	10	10	8	1	12	**11**
$\mathbb{S}^{23}(1000, 0.00)$	10	17	16	18	18	18	18	18	13	1	14	**23**
$\mathbb{S}^{47}(1000, 0.00)$	11	27	26	31	30	31	29	29	21	1	14	48
$\mathcal{S}(1000, 0.00)$	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	1	5	**2**
$\mathcal{Z}^1(1000, 0.00)$	NaN	NaN	2	93	0	14	2	68	3	**1**	15	**1**

Figure 11: This table contains the dimension estimates for a quite benign regime with 1000 samples and no noise. Even in this setting, and for the simplest manifolds, the estimation of dimension is challenging for most methods. Our algorithm fails with nonnegligible probability on $\mathbb{S}^{47}(1000, 0.00)$ because of the curse of intrinsic dimensionality (see Figure 9).

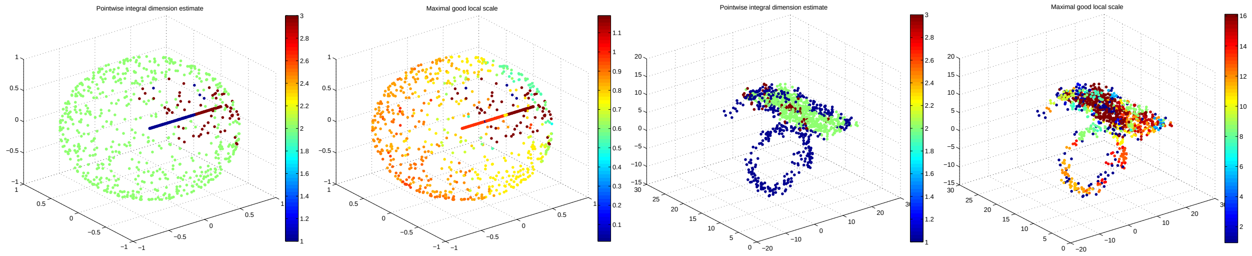

Figure 12: Our algorithm can produce pointwise estimates, albeit it is not designed to take advantage of any "smoothness" or clustering property of the local dimension as a function of the point. Top left: a 2-sphere and a segment. Top right: for every point we plot the estimated maximal good scale: it is large when sphere and segment are far away, and small close to the intersection. Bottom left: The data is a very noisy 1-dimensional spiral intersecting a noisy two-dimensional plane from [45]. Our algorithm assigns the correct dimension 1 to the spiral (because of the noise), and dimension 2 to the plane. 86% of the points on the spiral are assigned a dimension smaller than 2, and 77% of the points on the plane are assigned dimension 2 (or greater). Overall, clustering by dimension gives an accuracy of 86%, which is not as good as the 97% reported in [45], the present state-of-art to our knowledge (that uses knowledge about the number of clusters, and that each cluster is a smooth manifold, to gather strength across multiple neighborhoods).

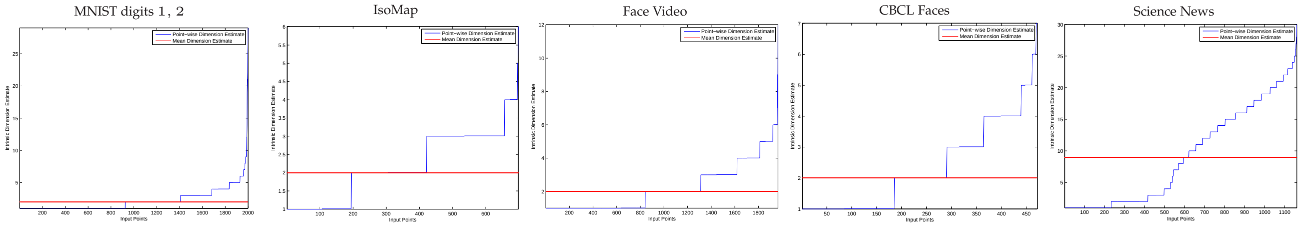

Figure 13: For each data sets we plot the point-wise estimate for a subset of the points (blue) and the average across the different points (red).

Digit	0	1	2	3	4	5	6	7	8	9
MSVD	2	2	3	2	2	2	2	2	3	3
IDE	11	7	13	13	12	11	10	13	11	11
HRVQ ($r = 2$)	16	7	21	20	17	20	16	16	20	17

Figure 14: This table contains the intrinsic dimension estimate for each digit obtained with our method (MSVD), with the smoothed Grassberger-Procaccia estimator from [51] and with high rate vector quantization methods in [49].

equation (5.1) *will be satisfied with probability larger than* $1 - \delta$.

Example 4 (Johnson-Lindenstrauss for manifolds [71]). *If $\mathcal{M}$ is a well-conditioned manifold, i.e. $\mathcal{M}$ is smooth and has positive* reach, *then by approximating $\mathcal{M}$ by a finite number of tangent planes, applying the Johnson-Lindenstrauss Lemma to the portion of $\mathcal{M}$ associated to each chosen tangent plane, and taking union bounds, one sees that in this case a (rescaled by $\sqrt{\frac{D}{d}}$) random projection onto a $d = O(k \log D)$-dimensional subspace will satisfy* (5.1) *w.h.p.. See [71] for the precise conditions and statements.*

5.4 Kernel methods

It seems common practice to estimate the intrinsic dimension of a data set by applying a manifold learning algorithm, that produces a map $\Phi_l : \mathcal{M} \to \mathbb{R}^l$, for a given l considered a parameter. The map Φ_l is usually sought to have small distortion in some sense. For example, ISOMAP [1], one of the first and most popular algorithms, tries to minimize the distortion between geodesic distances on $\mathcal{M}$ and Euclidean distances between the image points $\Phi_l(\mathcal{M}) \subseteq \mathbb{R}^l$. It returns a residual variance defined as

$$\mathrm{resVar}_l := 1 - \frac{\sum_{x_i,x_j} d_{\mathcal{M}}(x_i,x_j) \cdot ||\Phi_l(x_i) - \Phi_l(x_j)||}{\left(\sum_{x_i,x_j} d_{\mathcal{M}}(x_i,x_j)^2\right)^{\frac{1}{2}} \left(\sum_{x_i,x_j} ||\Phi_l(x_i) - \Phi_l(x_j)||^2\right)^{\frac{1}{2}}} \in [0,1]\,,$$

which is minimized when $d_{\mathcal{M}}(x_i,x_j) = ||\Phi_l(x_i) - \Phi_l(x_j)||$ for every i,j. The vector $(\mathrm{resVar}_l)_{l \geq 0}$ is often used in practice as a spectrum (related to that of a MultiDimensional Scaling operator associated with the matrix of geodesic distances on $\mathcal{M}$) from which the intrinsic dimension of $\mathcal{M}$ may be inferred. However, there exist few and weak guarantees for when this may indeed yield the correct dimension, which motivated the search for better algorithms (e.g. [4, 72, 73]). The few simple experiments that follow suggest that the use of such algorithms to infer intrinsic dimension is potentially misleading (see Figure 5.4). Moreover, we ran our algorithm on $\Phi_l(\mathcal{M})$, with Φ_l computed by ISOMAP (for $l = 50$ and $l = \dim(\mathcal{M}) + 1$), and the results consistently underestimated the true intrinsic dimension, except for $\mathcal{S}(1000, 0)$. We expect similar phenomena to be common to other manifold learning algorithms, and leave a complete investigation to future work.

In [74] it is suggested that diffusion maps [7] may be used in order to estimate intrinsic dimension as well as a scale parameter, for example in the context of dynamical systems where a small number of slow variables are present. Rather than an automatic algorithm for dimension estimation, [74] suggests a criterion that involves eyeballing the function $\sum_{i,j} e^{-\frac{||x_i - x_j||^2}{\epsilon^2}}$, as a function of ϵ, to find a region of linear growth, whose slope is an estimate of the intrinsic dimension. Figure 5.4 shows that this technique may fail even with rather small amounts of noise.

A promising approach, for which guarantees may be derived, would be to apply the eigenfunction or heat kernels maps described in [72, 73]. Such mappings provide a $1+\epsilon$ bi-Lipschitz embedding of a ball centered at a point z which has a near maximal radius R_z in the sense that balls of larger radius would not admit any $1+\epsilon$ bi-Lipschitz embedding to Euclidean space. Combining these results together with Lemma 1 we deduce that the present algorithm could be run in the range of, say, the heat kernel map of [72, 73], where balls around a point z have been "maximally flattened". In this way, the present algorithm becomes independent of the embedding of the manifold in the ambient space, because such is the embedding of [72, 73]. Moreover, such independence is achieved with the most favorable possible parameters, for example for $R_{\max}$ essentially as large as possible.

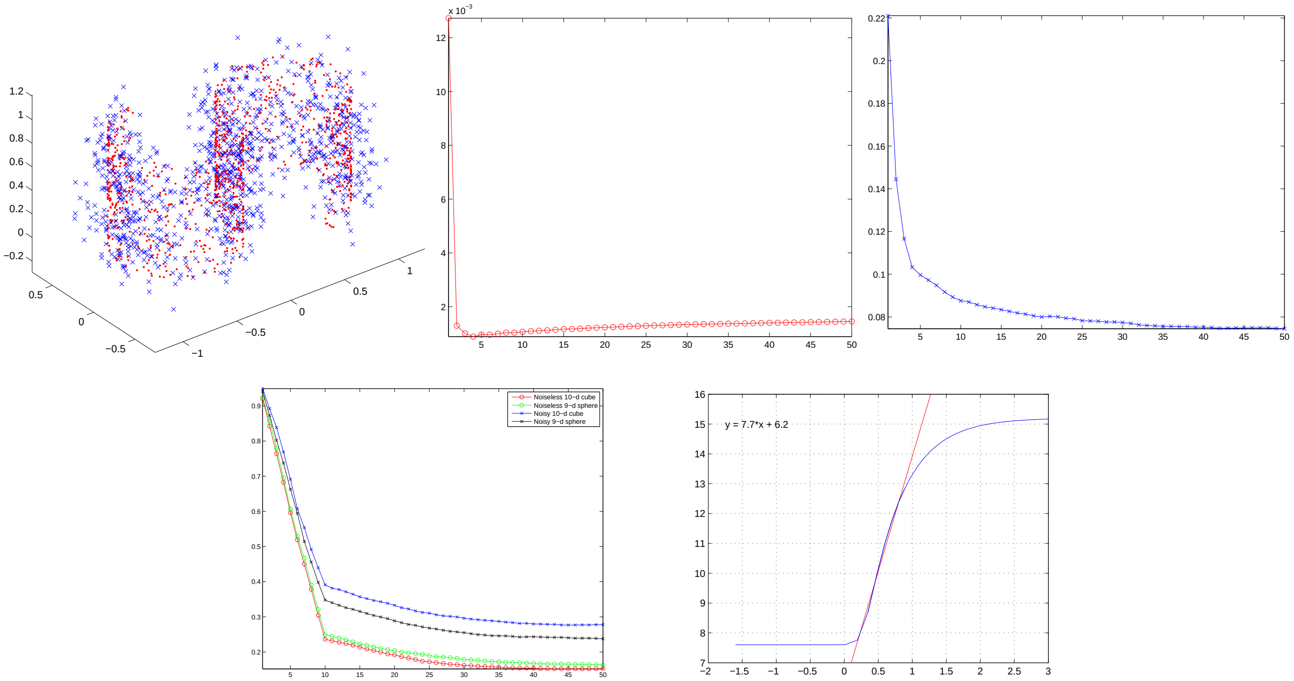

Figure 15: From left to right, top to bottom: (a) A realization of $\mathcal{S}(1000, 0)$ (red circles) and $\mathcal{S}(1000, 0.1)$ as in section 5.2.1. (b) $(\mathrm{resVar}_l)_l$ for $\mathcal{S}(1000, 0)$, from which the intrinsic dimension 2 may be inferred. (c) $(\mathrm{resVar}_l)_l$ for $\mathcal{S}(1000, 0.1)$, from which the intrinsic dimension seems hard to infer. Our algorithm, as shown in section 5.2.1, handles these cases correctly (w.h.p.). (d) the vectors of $(\mathrm{resVar}_l)_l$ for $\mathbb{Q}^{10}(1000, 0)$, $\mathbb{S}^9(1000, 0)$, $\mathbb{Q}^{10}(1000, 0.1)$, $\mathbb{S}^9(1000, 0.1)$: it seems hard to see a difference between the intrinsic dimensions 10 and 9, in both the noiseless and noisy cases. (d) the vectors of $(\mathrm{resVar}_l)_l$ for $\mathbb{Q}^{10}(1000, 0)$, $\mathbb{S}^9(1000, 0)$, $\mathbb{Q}^{10}(1000, 0.1)$, $\mathbb{S}^9(1000, 0.1)$: it seems hard to see a difference between the intrinsic dimensions 10 and 9, in both the noiseless and noisy cases. (e) The dimension of $\mathbb{S}^9(2000, 0.01)$ in $\mathbb{R}^{100}$ estimated according to the heuristic in [74] yields the wrong dimension (~ 8) even for small amounts of noise; this is of course not a rigorous test, and it is a heuristic procedure, not an algorithm, as described in [74].

6 Extensions

The work presented here may be extended in several directions, with minor modifications to the proofs. For example, the usual assumptions may be weakened or changed in different directions without substantially changing the conclusions. We mention a few cases here that may be of interest:

(i) The scaling of tangent and normal singular values may be changed, for example by assuming that there exist $0 < \alpha_T < \alpha_N$ such that

$$\lambda_i^2(\mathrm{cov}(X_{z,r}{}^{||})) \subseteq k^{-1}r^{\alpha_T}[\lambda^2_{\min},\lambda^2_{\max}] \quad , \quad \max_{1\le i<k} \Delta_i(\mathrm{cov}(X_{z,r}{}^{||})) \le k^{-1} r^{\alpha_T}\delta^2$$

$$\left\|X^\perp\right\| \le \sqrt{k}\kappa r^{\frac{\alpha_N}{2}} \text{ for a.s. } , \; ||\mathrm{cov}(X_{z,r}{}^\perp)|| \le \frac{\kappa^2}{k} r^{\alpha_N} \; , \quad \frac{\mathrm{tr}(\mathrm{cov}(X_{z,r}{}^\perp))}{||\mathrm{cov}(X_{z,r}{}^\perp)||} \le 2k^2$$

The main results still hold with simple modifications. Moreover, if α_T, α_N are not known a priori, they may be inferred from data: one may estimate α_T from $(\tilde\lambda_1^{[z,r]})^2$ and α_N from $(\tilde\lambda_K^{[z,r]})^2$ (as functions of r).

(ii) The exponent in the scaling of the volume does not need to be exactly equal to k: this would only change the "entropy" terms in the Theorems, used to lower bound, w.h.p., the number of points in a ball, but everything else stays the same. This shows how the geometry and scaling of the covariance matrix is crucial to our approach, rather than the scaling of the volume.

(iii) The $\log k$ factors are not needed if $X_{z,r}{}^{||}$ is assumed to be subgaussian, i.e. satisfies $\mathbb{P}(|\langle X_{z,r}{}^{||},\theta\rangle| > t) \le 2e^{-\frac{t}{\lambda_{\max} r}}$ for any $\theta \in \mathbb{S}^{k-1}$, and $r \in [R_{\min}, R_{\max}]$. Then $\mathrm{cov}(X_{z,r}{}^{||})$ may be approximated with a number of samples that depends only linearly on k, without the extra $\log k$ factor [75]. In this (natural, in our opinion) setting, all other assumptions being the same, the $\log k$ would disappear from all our results. Extending to the case $\sigma N = \sigma^{||} N^{\|} + \sigma^\perp N^\perp$ with $\sigma^{||} \neq \sigma^\perp$ is possible, where $\sigma^{||} N^{\|} = P^{(z,r)}\sigma N$ and $\sigma^\perp N^\perp = (I - P^{(z,r)})\sigma N$, by modifying the part of the proof in Appendix 9, in particular the upper bounds on the set I_2. We believe that robustness with respect to much larger classes of noises holds, in particular noise without spherical symmetry.

(iv) The methods presented may of course be applied after *kernelization*, leading to estimates of the dimension in the image of the kernel map. In particular, this may be combined with the maps introduced in [72, 73], which provide provably bi-Lipschitz embeddings of large portions of a manifold. See also section 5.3 about bi-Lipschitz perturbations.

7 Overview of previous work on dimension estimation

The problem of estimating intrinsic dimension has been considered in physics – for example to study attractors of dynamical systems– and in statistics/signal processing– for the analysis of high dimensional data/signals. Different definitions of "intrinsic" dimension have been proposed in the different domains, in particular there exist several notions of fractals dimensions: correlation dimension, Hausdorff dimension, box counting, packing dimension [76]. When the data are modeled as a manifold, possibly corrupted by noise, we can define the intrinsic dimension, at least locally, as the dimension of the manifold.

There is a large body of literature at the intersection between harmonic analysis and geometric measure theory ([77, 26, 27, 28] and references therein) that explores and connect the behavior of multiscale quantities, such as Jones' β-numbers [77], with quantitative notions of rectifiability. This body of work has been our major inspiration. The multiscale singular values we use are related to the multiscale β-numbers of Jones, but instead of fixing a dimension and exploring the behavior of the appropriate β-number for that dimension, we explore all the dimensions simultaneously and deduce the suitable dimension from the behavior of this ensemble. While the past analysis was done on continuous sets, the techniques in this paper allow to analyze what happens when we only have a finite number of random samples from a measure, and, additionally, a perturbation by noise in the ambient space. This is expected to be useful in other situations when sampled noisy data sets are analyzed by such tools.

A survey of many proposed estimation techniques can be found in [12]. We focus on those methods for which implementations are available and that we considered in our experiments. Some classical estimators originally proposed in the physics literature are based on the notion of correlation integral and correlation dimension [53]: the correlation integral is defined as

$$C(r) := \mathbb{P}\left(\|X - X'\| \leq r\right), \tag{7.1}$$

where X, X' are independent copies of X, and the correlation dimension is defined by $k_{cd} = \lim_{r \to 0} \frac{\log C(r)}{\log r}$. In practice one assumes that, for some constant c and r sufficiently small,

$$C(r) \sim cr^{k_{cd}}, \tag{7.2}$$

and approximates the correlation integral from a finite number of observations $\{x_i\}_{i=1}^n$. Grassberger and Procaccia [53] considered the following empirical proxy to $C(r)$

$$C_n(r) = \frac{2}{n(n-1)} \sum_{i<j}^{n} \chi_{\|x_i - x_j\| \leq r}.$$

The above quantity can be computed for different values of r and the correlation dimension can be estimated from $\log C_n(r) \sim \log c + k_{GB} \log r$ using linear regression. A different estimator was proposed by Takens [50]: it is based on assumption (7.2) and on a maximum likelihood estimate of the distribution of the distances among points. The Takens estimator of the correlation dimension is given by

$$k_T = -\left(\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \log\left(\frac{D_{ij}}{r}\right)\right)^{-1}$$

where $D_{ij} = \|X_i - X_j\|$ are all the distances smaller than r. A smooth version of the Grassberger-Procaccia estimator is proposed in [51] where, in the definition of correlation integral, the indicator function is replaced by a suitable kernel function, that is

$$U_n(h, m) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n}^{n} K_h(\|x_i - x_j\|^2) \tag{7.3}$$

with $K_{h,\ell}(\|x_i - x_j\|^2) = \frac{1}{h^\ell} K(\|x_i - x_j\|^2 / h^2)$, where $K : \mathbb{R}_+ \to \mathbb{R}_+$ is a suitable compactly supported kernel function [51]. If $\ell = k$ it is shown in [51] that $U_n(h, \ell)$ converges with high probability to

$$C \int_{\mathcal{M}} p^2(x) dx$$

for $n \to \infty$, provided that we choose h so that $h \to 0$ and $nh^\ell \to \infty$ (C is a suitable constant). With the same choice of h, $U_n(h, \ell)$ converges to 0 if $\ell > k$ and to ∞ if $\ell < k$.
Note that, in all the above algorithms, rather than using different values of r we can consider (and vary) the number of K nearest neighbors, letting $r = r_K$, the distance to the K nearest neighbor, which is done in [47, 46], where they consider estimators based on

$$L_n(r_K) = \sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - x_j\|^2 \chi_{\|x_i - x_j\| \leq r_K}(x_j).$$

It is possible to prove [48] that for $n \to \infty$, $L_n(r_K) n^{-((k-2)/k)}$ converges to

$$b_{k,K} C \int_{\mathcal{M}} p^{(k-2)/k}(x) dx$$

where C depend both on k and K. For *large* n one could consider the approximate equation

$$L_n(r_K) = c_{k,K} n^{(k-2)/k} + \varepsilon_n.$$

Similarly to the Grassberger-Procaccia estimator, one can compute $L_n(r_K)$ for different n and use linear regression to estimate k, but one has to further optimize w.r.t. $c_{k,K}$. The estimators in [47, 46] are based on refinements of the above idea.

Approaches similar to those described above have been proposed to approximate the intrinsic estimator at a given point, essentially by fixing one of the two variables in (7.1) at some point x and considering $C(x,r) = \mathbb{P}\left(\|X' - x\| \le r\right)$. Similarly to (7.2), if $C(x,r) \sim N(x,r)r^k$ and $N(x,r) \sim N_0$ then $\log C(x,r) = \log N_0 + k \log r$. If we estimate $C(x,r)$ for different values of r, then we can estimate the local intrinsic dimension by linear regression. Levina and Bickel [44] propose a method based on the above intuition, approximating the process that counts the number of points falling into a ball of radius r can be approximated by a Poisson process. In this way they derive the maximum likelihood estimator

$$k_{LB} = \left(\frac{1}{n_r} \sum_{i=1}^{n_r} \log \left(\frac{r}{\|x - x_j\|} \right) \right)^{-1}$$

where x_j, $j = 1, \ldots, n_r$, are points contained in the ball of radius r centered at x (in practice the nearest neighbor version is used). This estimator is a local version of the Takens estimator. In [45] translated Poisson processes are considered to improve the robustness of the above estimator towards noise and outliers, with a regularization parameter α and a noise parameter σ (which are likely to be related) control the robustness levels. The authors also propose a non local version of the above estimator, based on a mixture of processes with different parameters: such a model can describe data that is a combination of a small number of regions with different dimension.

An intrinsic dimension estimator based on local SVD was first discussed in [39]– see also [43, 40]. The main difference between the above algorithm and the one proposed in this paper is that instead of considering the SVD at one scale, we proceed in a multiscale fashion. As we have discussed in the introduction this is related to the choice of the size r of the local neighborhood which is guided by the following trade-off: if r is *small* the linear approximation of the manifold is good but we are prone to see the effect of the noise, viceversa when r is *large* there is a better robustness to noise but curvature might lead to an overestimation of the dimension (the so called *noise/curvature dilemma* in [40]). The methods in [39, 43, 40] are restricted to work in a range of very small scales where the linear approximation is exact, whereas, as a by product of our theoretical analysis, we have a whole range of *larger* scales where we can clearly distinguish the eigenvalues due to the curvature from those yielding the intrinsic dimension. This immediately translates into better noise robustness and sample complexity properties, and is the likely explanation the the poor experimental results in [43], which could lead one to prematurely conclude that eigenvalue methods are sensitive to noise and small sample size. Our empirical analysis indicates that this is not the case.

There are several works studying the statistical properties of intrinisic dimension estimators in terms of finite sample bounds and consistency. The consistency (without rates) of the Takens estimator is studied under very general conditions in [52]. Sample bounds for the U-statistics (7.3) used in the smoothed Grassberger Procaccia estimator are given in [51], where it is shown that the sample complexity is exponential in the ambient dimension. The statistical properties of a local nearest neighbor estimator are studied in [69]; in this case the rates are still exponential but only in the intrinsic dimension. It is worth noting that none of the previous theoretical analyses consider the case where the data are corrupted with (possibly high dimensional) noise.

While working on this manuscript we were made aware by K. Vixie of the work in [41] (and some references therein, in particular [55, 56]), where it is observed that the growth rate of singular values of data computed at multiple scales may be used for detecting intrinsic dimension. While the local data whitening technique considered there is not well-suited with noisy data in high-dimension, in those references the authors do consider some noisy data sets, mostly in low-dimensions. Our analysis extends those works by carefully analyzing the effects of sampling and noise on the multiscale singular values, by providing finite sample size bounds, and by extending these techniques to a setting far more general than the setting of smooth manifolds, which may be better suited for the analysis of data sets in truly high dimensional spaces. Also, M. Davies pointed us to work in the dynamical systems community, in particular [23, 24, 25] where local linear approximations to the trajectories of dynamic systems are considered in order to construct reduced models for the dynamics, and local singular values are used to decide on the dimension of the reduced system and/or on Lyapunov exponents. Finally, [78] discusses various tree constructions for fast nearest neighbor computations and studies how they adapt to the intrinsic dimension of the support of the probability measure from which points are drawn, and discusses and presents some examples of the behavior of the smallest k such that (in the notation of this paper) $\sum_{i=1}^{k} \lambda(\mathrm{cov}(X_{z,r})) \geq (1-\epsilon) \sum_{i=1}^{D} \lambda(\mathrm{cov}(X_{z,r}))$, where ϵ is some fixed parameter (e.g. $\epsilon = 10^{-2}$). This should be contrasted with our approach, where all the $\lambda(\mathrm{cov}(X_{z,r}))$ are studied jointly as a function of r, rather than fixing a threshold ϵ; this also has the advantage, when the samples are corrupted by noise with covariance of size σ^2, of requiring the largest eigenvalues of the covariance to be larger than σ^2, rather than $\sigma^2 D$ as required by the condition above.

The estimates we obtain on the multiscale covariance matrices, generalize somewhat a variety of estimates in the literature, as they combine a nonlinear geometric model with noise, instead of focusing on a linear model with noise, whether they are asymptotic in character (see e.g. [79, 80, 81, 82] and references therein) or in the finite-sample size regime (see e.g. [83, 84]).

We recently became of aware of the work [85], which was written after [32, 19, 31] and cites at least some of these results. The main result of [85], on perturbation of approximation of tangent planes, follows immediately for the results of [32, 19, 31], and of course also as a very particular case of this work. The main differences between [85] and [32] and this work are that only the smooth manifold case is considered there and the observation model in [85] is not clearly specified, but our understanding is that the noisy matrices considered there are not naturally observable, being those that in our notation correspond to $\widetilde{\mathbf{X}_n^{[z,r]}}$, rather than $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}$, which is really all that can be observed.

Finally, recent work [65, 66] uses Bayesian methods to construct certain models for data, and as part of such models the intrinsic dimension of data is estimated (while this occurs in both the papers cited, [66] explicitly discusses this aspect). The comparisons we presented suggest that, for the particular task of estimating dimension, such methods are less sensitive than other methods when noise corrupts the data (as they explicitly take noise into account in the underlying model), but they may not be reliable dimension estimators. They also lack theoretical guarantees, in particular finite sample size bounds. We conjecture that the difference in our experience with MFA, compared to the results in [66], is due to the fact that all manifolds considered in [66] have very small curvature, except for one example of a very low-dimensional manifold (a 2-d torus). We suspect that the priors imposed in the models make them susceptible to gross errors in estimation, at least when the sample size is not extremely large, and hyper parameters are not optimally set.

Acknowledgements

MM is greteful for support from NSF (DMS 0847388, CCF 0808847, IIS 0803293), ONR N00014-07-1-0625 and the Sloan Foundation. AVL is partially supported by NSF DMS 0650413, 0847388 and ONR N00014-07-1-0625. LR is on leave of absence from the Università di Genova, Italy. MM is thankful for useful discussions with B. Allard, L. Carin, M. Davies, J. Harer, P.W. Jones, J. Tropp, A. Szlam, R. Vershynin, K. Vixie, as well as to G. Haro et al. for sharing their code and to J.G. Silva for providing results for the algorithms [65, 66] on test sets. LR is thankful to G. Canas for useful discussions.

8 Appendix: Proof of Proposition 1

8.1 Some Preliminary Observartions

In this section we prove Proposition 1 comparing $\mathrm{cov}(X_{z,r_=})$ with $\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})$, and establishing conditions that guarantee that the k-th gap of $\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})$ is the largest gap, notwithstanding the high-dimensional noise in the ambient space. We will use the following observation. If we condition on $n_{z,r}$, then the samples $\mathbf{X}_n^{[z,r]}$ have the same distribution as $n_{z,r}$ copies of a random variable $X_{z,r}$, where we recall that $X_{z,r}$ has distribution $\mu_{z,r}$, where $\mu_{z,r}(A) := \mu_X(A \cap B_z(r))/\mu_X(B_z(r))$ is the restriction of μ_X to $B_z(r)$. Similarly, we can see $\widetilde{\mathbf{X}_n^{[z,r]}}$ as m i.i.d sample from $(\mu_{z,r})^{n_{z,r}} * (\mu_{\sigma N})$, where $\mu_{\sigma N}$ is the distribution of σN. Note that, in the absence of noise, we have $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} = \mathbf{X}_n^{[z,r]}$ so that conditioning on $n_{z,r} = m$

$$\mathrm{cov}(\mathbf{X}_n^{[z,r]}) = \frac{1}{m}\overline{\mathbf{X}_n^{[z,r]}}^T \overline{\mathbf{X}_n^{[z,r]}}$$

and moreover

$$\mathbb{E}\left[\mathrm{cov}(\mathbf{X}_n^{[z,r]})\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{1}{n_{z,r}}\overline{\mathbf{X}_n^{[z,r]}}^T \overline{\mathbf{X}_n^{[z,r]}} | n_{z,r}\right]\right] = \mathbb{E}\left[\mathrm{cov}(X_{z,r})|n_{z,r}\right] = \mathrm{cov}(X_{z,r})\,.$$

This suggests that we can proceed by first conditioning on $n_{z,r}$ and then removing the conditioning.

Proof of Proposition 1. We stated the Proposition in terms of $r_=$ since in that form it is more easily combined with Proposition 2 to yield 2; here, to ease the notation, we shall work with r instead of $r_=$. Moreover, through out this section, we will fix $z \in \mathcal{M}$ and drop it from the notation.
We dispose the random variable n_r as follows: first of all we have, if we let $\Omega_{t,0} = \{n_r > \frac{\overline{n}}{2} = \frac{1}{2}\mu_X(B_z(r))n\}$, then using a Chernoff bound [86] and $n \geq \frac{2t^2 k \log k}{\mu_X(B_z(r))}$:

$$\mathbb{P}(\Omega_{t,0}) = \mathbb{P}\left(n_r \geq \frac{\overline{n}}{2}\right) \geq 1 - e^{-\frac{t^2 k}{4}}\,.$$

Then, we have

$$\begin{aligned}&\mathbb{P}\left(||\mathrm{cov}(X_{z,r}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r]}})|| \leq \epsilon||\mathrm{cov}(X_{z,r})||\right)\\ &\geq \mathbb{P}\left(||\mathrm{cov}(X_{z,r}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r]}})|| \leq \epsilon||\mathrm{cov}(X_{z,r})||\mid \Omega_{t,0}\right)\mathbb{P}\left(\Omega_{t,0}\right)\\ &\geq \mathbb{P}\left(||\mathrm{cov}(X_{z,r}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r]}})|| \leq \epsilon||\mathrm{cov}(X_{z,r})||\mid \Omega_{t,0}\right)\left(1 - e^{-\frac{t^2 k}{4}}\right)\end{aligned} \tag{8.1}$$

and therefore we proceed by bounding the interesting event conditioned on $\Omega_{t,0}$.

We split the perturbation from the true covariance matrix $\mathrm{cov}(X^{[r]})$ to the sampled noisy covariance matrix $\mathrm{cov}(\widetilde{\mathbf{X}_n}^{[r]})$ into the following steps:

$$\begin{aligned}\mathrm{cov}(X^{[r]}) \quad &= \quad \begin{bmatrix}\mathrm{cov}(X^{[r]^{||}}) & \mathrm{cov}(X^{[r]^{||}}, X^{[r]^{\perp}})\\ \mathrm{cov}(X^{[r]^{\perp}}, X^{[r]^{||}}) & \mathrm{cov}(X^{[r]^{\perp}})\end{bmatrix} \underbrace{\rightarrow^{P_1}}_{\substack{\textit{Wielandt's}\\ \textit{Lemma}}} \begin{bmatrix}\mathrm{cov}(X^{[r]^{||}}) & 0\\ 0 & \mathrm{cov}(X^{[r]^{\perp}})\end{bmatrix}\\ &\underbrace{\rightarrow^{P_2}}_{\textit{Sampling}} \begin{bmatrix}\mathrm{cov}(\mathbf{X}_n^{[r]^{||}}) & 0\\ 0 & \mathrm{cov}(\mathbf{X}_n^{[r]^{\perp}})\end{bmatrix} \underbrace{\rightarrow^{P_3}}_{\substack{\textit{Diagonal}\\ \textit{noise}}} \begin{bmatrix}\mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{||}}}) & 0\\ 0 & \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{\perp}}})\end{bmatrix}\\ &\underbrace{\rightarrow^{P_4}}_{\substack{\textit{Wielandt's}\\ \textit{Lemma}}} \begin{bmatrix}\mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{||}}}) & \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{||}}}, \widetilde{\mathbf{X}_n^{[r]^{\perp}}})\\ \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{\perp}}}, \widetilde{\mathbf{X}_n^{[r]^{||}}}) & \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{\perp}}})\end{bmatrix} = \mathrm{cov}(\widetilde{\mathbf{X}_n}^{[r]})\end{aligned}$$

where without loss of generality we assumed that $\mathrm{range}(P^{[r]}) = \langle\{e_i\}_{i=1}^k\rangle$, and with abuse of notation we considered $\mathrm{cov}(X^{[r]^{||}})$ as a $k \times k$ matrix instead of a $D \times D$ matrix. The eigenvalues (sorted in decreasing order, as usual) of the 5 matrices above will be denoted by $\{\lambda_1^2, \ldots, \lambda_D^2\}$, $\{(\lambda_1^{||})^2, \ldots, (\lambda_k^{||})^2, (\lambda_{k+1}^{\perp})^2, \ldots, (\lambda_D^{\perp})^2\}$,

$\{(\lambda^{||}_{n_r,1})^2, \ldots, (\lambda^{||}_{n_r,k})^2, (\lambda^{\perp}_{n_r,k+1})^2, \ldots, (\lambda^{\perp}_{n_r,D})^2\}$,
$\{(\tilde\lambda^{||}_{n_r,1})^2, \ldots, (\tilde\lambda^{||}_{n_r,k})^2, (\tilde\lambda^{\perp}_{n_r,k+1})^2, \ldots, (\tilde\lambda^{\perp}_{n_r,D})^2\}$, $\{\tilde\lambda^2_{n_r,1}, \ldots, \tilde\lambda^2_{n_r,D}\}$, respectively. Except when specified otherwise, we shall say that an event Ω_t has high probability if $\mathbb{P}(\Omega_t) \geq 1 - c_1 e^{-c_2 t^2}$ for all $t \geq c_3$, with c_1, c_2, c_3 universal constants.

$\mathbf{P_1}$ [Geometric cross-terms]: We bound the error in approximating $\mathrm{cov}(X^{[r]})$ by $\mathrm{cov}(X^{[r]^{||}})$ and $\mathrm{cov}(X^{[r]^{\perp}})$, thereby showing that our usual Assumptions are equivalent to hypotheses on the spectrum of $\mathrm{cov}(X^{[r]})$: for $r \in (R_{\min}, R_{\max})$

$$\left\|\mathrm{cov}(X^{[r]}) - P^{[r]}(\mathrm{cov}(X^{[r]}))\right\| \leq \frac{\lambda_{\max}\kappa}{k} r^3 \left(\frac{\kappa\lambda_{\max} r}{\lambda^2_{\min} - \kappa^2 r^2} \wedge 1\right) \tag{8.2}$$

In particular, the characteristic scale $r \sim \frac{\lambda_{\min}}{\kappa}$ is where the effect of "curvature" correction switches from $O(r^4)$ to $O(r^3)$. We prove this bound by observing that $\mathrm{cov}(X^{[r]}) = \begin{pmatrix} \mathrm{cov}(X^{[r]^{||}}) & (X^{[r]^{||}})^T X^{[r]^{\perp}} \\ (X^{[r]^{\perp}})^T X^{[r]^{||}} & \mathrm{cov}(X^{[r]^{\perp}}) \end{pmatrix}$, and therefore the result follows from Wielandt's Lemma 10: for example for $i = 1, \ldots, k$ we have

$$\begin{aligned} 0 \leq \lambda^2_i(\mathrm{cov}(X^{[r]})) - \lambda^2_i(\mathrm{cov}(X^{[r]^{||}})) &\leq \frac{\|\mathrm{cov}(X^{[r]^{||}}, X^{[r]^{\perp}})\|^2}{\frac{\lambda^2_{\min} r^2}{k} - \frac{\kappa^2 r^4}{k}} \wedge \|\mathrm{cov}(X^{[r]^{||}}, X^{[r]^{\perp}})\| \\ &\leq \frac{\frac{\kappa^2}{k}\lambda^2_{\max}}{\lambda^2_{\min} - \kappa^2 r^2} r^4 \wedge \frac{\lambda_{\max}\kappa}{k} r^3 \,. \end{aligned}$$

The bounds for $i = k+1, \ldots, D$ follow in the same manner.

We start by conditioning on $n_r = m$. In the rest of the proof, we let $\epsilon := \epsilon(k, m, t) = \sqrt{\frac{t^2 k \log k}{m}}$ and assume $\epsilon \leq 1$.

$\mathbf{P_2}$ [Tangent and normal sampling] By Proposition 7 we have, on an event $\Omega_{t,1}$ having high probability,

$$||\mathrm{cov}(\mathbf{X}^{[r]^{||}}_n) - \mathrm{cov}(X^{[r]^{||}})|| \leq \frac{\lambda_{\max} r^2}{k}\sqrt{\frac{k\log k}{m}}t + \frac{r^2}{m}t^2 \leq \frac{\lambda_{\max} r^2}{k}\epsilon\,(1+\epsilon) := P^{||}_2\,, \tag{8.3}$$

and $\frac{1}{\sqrt{m}}||\overline{\mathbf{X}^{[r]}_n}^{||}|| \leq \frac{\sqrt{\lambda_{\max}}r}{\sqrt{k}}\sqrt{\lambda_{\max}+\epsilon} \leq \frac{\lambda_{\max} r}{\sqrt{k}}(1+\epsilon)$. As for $X_{z,r}{}^{\perp}$, since it is bounded, and since centering only reduces the norm of the matrix, we have (recalling that the bar notation indicates *centering with respect to the empirical mean*)

$$\begin{aligned} ||\mathrm{cov}(\mathbf{X}^{[r]^{\perp}}_n)|| &\leq \frac{\kappa^2 r^4}{k}\left((1+\epsilon')\wedge k\right)^2 =: \frac{\kappa'^2 r^4}{k}, & \frac{||\overline{\mathbf{X}^{[r]}_n}^{\perp}|}{\sqrt{m}}| &\leq \frac{\kappa' r^2}{\sqrt{k}} \\ ||\mathrm{cov}(\mathbf{X}^{[r]^{\perp}}_n) - \mathrm{cov}(X^{[r]^{\perp}})|| &\leq \frac{\kappa^2 r^4}{k}\left(\epsilon'(1+\epsilon')\wedge k^2\right) =: \frac{\kappa''^2 r^4}{k}, & \frac{||\overline{\mathbf{X}^{[r]}_n}^{\perp} - \overline{X^{[r]}}^{\perp}||}{\sqrt{m}} &\leq \frac{\kappa'' r^2}{\sqrt{k}} \end{aligned} \tag{8.4}$$

where $\epsilon' = \sqrt{\frac{t^2 k^2 \log(D\wedge m)}{m}}$, $\kappa' := \kappa((1+\epsilon')\wedge k)$, $\kappa'' := \kappa(\sqrt{\epsilon(1+\epsilon')}\wedge k)$.

$\mathbf{P_3}$ [Tangential and normal noise]: We start by considering the perturbation $\mathrm{cov}(\mathbf{X}^{[r]^{||}}_n) \to \mathrm{cov}(\mathbf{X}^{[r]^{||}}_n + \sigma\mathbf{N}^{[r]^{||}}_m)$. Since $\mathbb{E}[\mathrm{cov}(\mathbf{N}^{[r]^{||}}_m)] = I_k$,

$$||\mathrm{cov}(\mathbf{X}^{[r]^{||}}_n + \sigma\mathbf{N}^{[r]^{||}}_m) - \mathrm{cov}(\mathbf{X}^{[r]^{||}}_n) - \sigma^2 I_k|| \leq \frac{2\sigma}{m}||\overline{\mathbf{X}^{[r]}_n}^{||^T}\overline{\mathbf{N}}^{[r]^{||}}_m|| + \sigma^2\left\|\mathrm{cov}(\mathbf{N}^{[r]^{||}}_m) - \mathrm{cov}(N^{[r]^{||}})\right\|\,.$$

Since, by Proposition 8, on an event of high probability $\Omega_{t,2}$ we have

$$||\mathrm{cov}(\mathbf{N}^{[r]^{||}}_m) - I_k|| \leq \sqrt{\frac{k}{m}}t \quad , \quad \frac{1}{\sqrt{m}}||\overline{\mathbf{N}}^{[r]^{||}}_m|| \leq 1 + \sqrt{\frac{k}{m}}t\,.$$

Therefore $\frac{1}{m}||\overline{\mathbf{X}_n^{[r]}}^{||^T}\overline{\mathbf{N}}_m^{[r]^{||}}|| \le \frac{\lambda_{\max}r}{\sqrt{k}}\,(1+\epsilon)\left(1+\sqrt{\frac{k}{m}t}\right) \le \frac{\lambda_{\max}r}{\sqrt{k}}\,(1+3\epsilon)$, so that on $\Omega_{t,1}\cap\Omega_{t,2}$

$$||\mathrm{cov}(\mathbf{X}_n^{[r]^{||}}+\sigma\mathbf{N}_m^{[r]^{||}})-\mathrm{cov}(\mathbf{X}_n^{[r]^{||}})-\sigma^2 I_k|| \le \left(\frac{2\lambda_{\max}(1+3\epsilon)r}{\sqrt{k}}+\sigma\right)\sigma\epsilon\,, \tag{8.5}$$

which implies

$$(\tilde{\lambda}_{m,i}^{||})^2 \in (\lambda_{m,i}^{||})^2+\sigma^2+\underbrace{\left(\frac{2\lambda_{\max}(1+3\epsilon)r}{\sqrt{k}}+\sigma\right)\sigma\epsilon}_{P_3^{||}}\cdot[-1,+1]\,. \tag{8.6}$$

We record the following estimate, obtained by combining (8.3) and (8.5) and replacing 4ϵ by ϵ:

$$||\widetilde{\mathrm{cov}(\mathbf{X}_n^{[r]^{||}})}-\sigma^2 I_k-\mathrm{cov}(X^{[r]^{||}})|| \le \epsilon\frac{\lambda_{\max}r}{\sqrt{k}}\left(\frac{r}{\sqrt{k}}+\sigma\right) \tag{8.7}$$

Now we consider the perturbation $\mathrm{cov}(\mathbf{X}_n^{[r]^\perp}) \to \mathrm{cov}(\mathbf{X}_n^{[r]^\perp}+\sigma\mathbf{N}_m^{[r]^\perp})$. When $m \ge CD$, by (11.2) and Propositions 8 and 10, on an event $\Omega_{t,3}$ of high probability we have

$$\mathrm{cov}(\mathbf{X}_n^{[r]^\perp},\mathbf{N}_m^{[r]^\perp}) = \frac{1}{m}||\overline{\mathbf{X}_n^{[r]}}^{\perp^T}\overline{\mathbf{N}}_m^{[r]^\perp}|| \le \frac{\kappa' r^2}{\sqrt{k}}\sqrt{\frac{D}{m}}t$$

Moreover by Proposition 8, $||\mathrm{cov}(\mathbf{N}_m^{[r]^\perp})-I_{D-k}|| \le \sqrt{\frac{D}{m}}t$ on an event $\Omega_{t,4,1}$ having high probability. Therefore on $\Omega_{t,2}\cap\Omega_{t,3}\cap\Omega_{t,4,1}$ we have

$$\begin{aligned}||\mathrm{cov}(\mathbf{X}_n^{[r]^\perp}+\sigma\mathbf{N}_m^{[r]^\perp})-\mathrm{cov}(\mathbf{X}_n^{[r]^\perp})-\sigma^2 I_{D-k}|| &\le \frac{2\sigma}{m}||\overline{\mathbf{X}_n^{[r]}}^{\perp^T}\overline{\mathbf{N}}_m^{[r]^\perp}||+\sigma^2||\mathrm{cov}(\mathbf{N}_m^{[r]^\perp})-I_{D-k}||\\ &\le \left(\frac{2\kappa' r^2}{\sqrt{k}}+\sigma\right)\sigma\sqrt{\frac{D}{m}}t\,(1+\epsilon)\ ,\end{aligned}$$

and hence

$$(\tilde{\lambda}_{m,i}^{\perp})^2 \in (\lambda_{m,i}^{\perp})^2+\sigma^2+\underbrace{\left(\frac{2\kappa' r^2}{\sqrt{k}}+\sigma\right)\sigma\sqrt{\frac{D}{m}}t\,(1+\epsilon)}_{P_{3,1}^{\perp}}\cdot[-1,1]\,.$$

When $m < CD$, we use $||\mathrm{cov}(\mathbf{X}_n^{[r]^\perp}+\sigma\mathbf{N}_m^{[r]^\perp})-\mathrm{cov}(\mathbf{X}_n^{[r]^\perp})|| \le \frac{2\sigma}{m}||\overline{\mathbf{X}_n^{[r]}}^{\perp^T}\overline{\mathbf{N}}_m^{[r]^\perp}||+\sigma^2||\mathrm{cov}(\mathbf{N}_m^{[r]^\perp})||$. By Proposition 10, on an event $\Omega_{t,4,2}$ of high probability, letting $\delta_1 := \delta_1(m,D,t) := C\sqrt{m/D}+t/\sqrt{D}$,

$$||\mathrm{cov}(\mathbf{N}_m^{[r]^\perp})|| \le \frac{D}{m}(1+\delta_1)^2 \quad , \quad \frac{1}{\sqrt{m}}||\overline{\mathbf{N}}_m^{[r]^\perp}|| \le \sqrt{\frac{D}{m}}(1+\delta_1)$$

and on $\Omega_{t,2}\cap\Omega_{t,3}\cap\Omega_{t,4,2}$

$$||\mathrm{cov}(\mathbf{X}_n^{[r]^\perp}+\sigma\mathbf{N}_m^{[r]^\perp})-\mathrm{cov}(\mathbf{X}_n^{[r]^\perp})|| \le \left(\frac{2\kappa' r^2}{\sqrt{k}}+\sigma\sqrt{\frac{D}{m}}(1+\delta_1)^2\right)\sigma\sqrt{\frac{D}{m}}t\,,$$

so that

$$(\tilde{\lambda}_{m,i}^{\perp})^2 \in (\lambda_{m,i}^{\perp})^2+\underbrace{\left(\frac{2\kappa' r^2}{\sqrt{k}}+\sigma\sqrt{\frac{D}{m}}(1+\delta_1)^2\right)\sigma\sqrt{\frac{D}{m}}t(1+\epsilon)}_{P_{3,2}^{\perp}}\cdot[-1,1]\,. \tag{8.8}$$

Letting $\Omega_{t,4} = \Omega_{t,4,1}$ for $m \geq CD$ and $\Omega_{t,4} = \Omega_{t,4,2}$ for $m < CD$, we have on an event $\Omega_{t,3} \cap \Omega_{t,4}$ of high probability, that:

$$(\tilde{\lambda}^{\perp}_{m,i})^2 \in (\lambda^{\perp}_{m,i})^2 + \sigma^2 \mathbf{1}_{(m \geq CD)} + \big(\underbrace{P^{\perp}_{3,1} \mathbf{1}_{(m \geq CD)} + P^{\perp}_{3,2} \mathbf{1}_{(m < CD)}}_{=: P^{\perp}_3} \big) \cdot [-1, +1] \,. \tag{8.9}$$

$\mathbf{P_4}$ [Noisy cross-terms]: Assuming that $(\tilde{\lambda}^{||}_{m,k})^2 > (\tilde{\lambda}^{\perp}_{m,k+1})^2$, by Wielandt's Lemma 10, $(\tilde{\lambda}^{||}_{m,i})^2 < \tilde{\lambda}^2_{m,i}$ for $i = 1, \ldots, k$ and $(\tilde{\lambda}^{\perp}_{m,i})^2 > \tilde{\lambda}^2_{m,i}$ for $i = k+1, \ldots, D$. Moreover, again by Wielandt's lemma, the size of each perturbation is bounded by $||B|| \wedge \frac{||B||^2}{\Delta}$, where $\Delta = (\tilde{\lambda}^{||}_{m,k})^2 - (\tilde{\lambda}^{\perp}_{m,k+1})^2$, and

$$\begin{aligned} B &:= \mathrm{cov}({\mathbf{X}^{[r]}_n}^{||} + \sigma {\mathbf{N}^{[r]}_m}^{||}, {\mathbf{X}^{[r]}_n}^{\perp} + \sigma {\mathbf{N}^{[r]}_m}^{\perp}) \\ &= \frac{{\overline{\mathbf{X}^{[r]}_n}^{||}}^T \overline{\mathbf{X}^{[r]}_n}^{\perp}}{m} + \frac{\sigma {\overline{\mathbf{X}^{[r]}_n}^{||}}^T \overline{\mathbf{N}}^{[r]}_m{}^{\perp}}{m} + \frac{\sigma {\overline{\mathbf{N}}^{[r]}_m{}^{||}}^T \overline{\mathbf{X}^{[r]}_n}^{\perp}}{m} + \frac{\sigma^2 {\overline{\mathbf{N}}^{[r]}_m{}^{||}}^T \overline{\mathbf{N}}^{[r]}_m{}^{\perp}}{m} \,. \end{aligned}$$

Since $\overline{\mathbf{X}^{[r]}_n}^{||}$ and $\overline{\mathbf{X}^{[r]}_n}^{\perp}$ are not necessarily independent, on $\Omega_{t,1}$ we use the bound

$$\left\| \frac{1}{m} {\overline{\mathbf{X}^{[r]}_n}^{||}}^T \overline{\mathbf{X}^{[r]}_n}^{\perp} \right\| \leq \frac{\lambda_{\max} \kappa' r^3}{k} (1 + \epsilon) \ ,$$

which holds w.h.p.; by Proposition 8 and 10, on $\Omega_{t,1}$

$$\begin{aligned} \frac{1}{m} \left\| {\overline{\mathbf{X}^{[r]}_n}^{||}}^T \overline{\mathbf{N}}^{[r]}_m{}^{\perp} \right\| &\leq \frac{\lambda_{\max} r}{\sqrt{k}} (1 + \epsilon) \left(\left(\sqrt{\frac{k}{m}} + \sqrt{\frac{D}{m}} \right) t + \sqrt{\frac{D}{m}} \sqrt{1 + \frac{t}{\sqrt{D}}} \right) \\ &\leq \frac{\lambda_{\max} r}{\sqrt{k}} \sqrt{\frac{D}{m}} t \, (1 + 2\epsilon) \ . \end{aligned}$$

On $\Omega_{t,3}$ we have

$$\frac{1}{m} \left\| {\overline{\mathbf{N}}^{[r]}_m{}^{||}}^T \overline{\mathbf{X}^{[r]}_n}^{\perp} \right\| \leq \frac{\kappa' r^2}{\sqrt{k}} \left(1 + \sqrt{\frac{k}{m}} t \right) \leq \frac{\kappa' r^2}{\sqrt{k}} (1 + \epsilon) \,.$$

Finally, by (11.2) and Propositions 8 and 9, for $m \geq CD$

$$\frac{1}{m} \left\| {\overline{\mathbf{N}}^{[r]}_m{}^{||}}^T \overline{\mathbf{N}}^{[r]}_m{}^{\perp} \right\| \leq \left(\sqrt{\frac{k}{m}} + \sqrt{\frac{D}{m}} \right) \tilde{t} \leq \sqrt{\frac{D}{m}} t$$

on an event $\Omega_{t,5}$ with high probability, by changing the universal constants in Proposition 9. For $m \leq CD$ we have w.h.p. that

$$\frac{1}{m} \left\| {\overline{\mathbf{N}}^{[r]}_m{}^{||}}^T \overline{\mathbf{N}}^{[r]}_m{}^{\perp} \right\| \leq \left(1 + \sqrt{\frac{D}{m}} \right) \left(1 + \sqrt{\frac{k}{m}} \right) \tilde{t} \leq \left(1 + \sqrt{\frac{D}{m}} \right) (\tilde{t} + \epsilon) \leq \sqrt{\frac{D}{m}} t \,.$$

Summarizing, on a high probability event we have

$$||B|| \leq \left(\frac{\lambda_{\max} \kappa' r^3}{k} + \frac{\sigma \kappa' r^2}{\sqrt{k}} + \frac{\sqrt{\lambda}_{\max} \sigma r}{\sqrt{k}} \sqrt{\frac{D}{m}} t + \sigma^2 \sqrt{\frac{D}{m}} t \right) (1 + 2\epsilon) =: P_4 \,.$$

The largest gap, part I. Let $\tilde{\Delta}_i = \tilde{\lambda}^2_{m,i} - \tilde{\lambda}^2_{m,i+1}$ for $i = 1, \ldots, D-1$, $\tilde{\Delta}_D = \tilde{\lambda}^2_{m,D}$. We want to lower bound the probability that $\tilde{\Delta}_k = \max_{i=1,\ldots,D} \tilde{\Delta}_i$. For $1 \leq i < k$:

$$\begin{aligned} \tilde{\Delta}_i = \tilde{\lambda}^2_{m,i} - \tilde{\lambda}^2_{m,i+1} &\leq (\tilde{\lambda}^{||}_{m,i})^2 - (\tilde{\lambda}^{||}_{m,i+1})^2 + P_4 \\ &\leq (\lambda^{||}_{m,i})^2 - (\lambda^{||}_{m,i+1})^2 + 2P^{||}_3 + P_4 \leq (\lambda^{||}_i)^2 - (\lambda^{||}_{i+1})^2 + 2P^{||}_2 + 2P^{||}_3 + P_4 \\ &\leq \frac{\delta^2 r^2}{k} + 2P^{||}_2 + 2P^{||}_3 + P_4 \,. \end{aligned}$$

Object	Bound
P_1	$\frac{\lambda_{\max}\kappa r^3}{k}\left(\frac{\lambda_{\max}\kappa r^3}{\lambda_{\min}^2 r^2-\kappa^2 r^4}\wedge 1\right)$
$P_2^{\vert\vert}$	$\frac{\lambda_{\max} r^2}{k}\epsilon\left(1+\epsilon\right)$
$P_3^{\vert\vert}$	$\sigma\left(\frac{2\lambda_{\max} r}{\sqrt{k}}+\sigma\right)\epsilon(1+3\epsilon)$
$P_3^{\perp}$	$\sigma\sqrt{\frac{D}{m}}t\left(\frac{2\kappa' r^2}{\sqrt{k}}+\sigma\left(\mathbf{1}_{m\geq CD}+\sqrt{\frac{D}{m}}\left(1+\delta_1\right)^2\mathbf{1}_{m\leq CD}\right)\right)(1+\epsilon)$
P_4	$\left(\frac{\lambda_{\max}\kappa' r^3}{k}+\frac{\sigma\kappa' r^2}{\sqrt{k}}+\frac{\lambda_{\max}\sigma r}{\sqrt{k}}\sqrt{\frac{D}{m}}t+\sigma^2\sqrt{\frac{D}{m}}t\right)(1+2\epsilon)$

Figure 16: Bounding the P_i's; recall that we have let $\epsilon=\sqrt{\frac{t^2 k\log k}{m}}<1$, $\delta_1:=C\sqrt{m/D}+t/\sqrt{D}$, $\kappa':=\kappa((1+\epsilon')\wedge k)$, where $\epsilon'=\sqrt{\frac{t^2k^2\log(D\wedge m)}{m}}$.

For $i=k$, using (8.6), (8.8) and (8.9):

$$\begin{aligned}
\tilde{\Delta}_k &= \tilde{\lambda}_{m,k}^2-\tilde{\lambda}_{m,k+1}^2 \geq (\tilde{\lambda}_{m,k}^{||})^2-(\tilde{\lambda}_{m,k+1}^{\perp})^2 \\
&\geq (\lambda_{m,k}^{||})^2+\sigma^2-P_3^{||}-(\lambda_{m,k+1}^{\perp})^2-\sigma^2\mathbf{1}_{(m\geq CD)}-\left(P_3^{\perp}+\sigma^2\mathbf{1}_{(m<CD)}\right) \\
&\geq (\lambda_k^{||})^2-\frac{\kappa'^2 r^4}{k}+\sigma^2-\sigma^2-P_2^{||}-P_3^{||}-P_3^{\perp} \\
&\geq \frac{\lambda_{\min}^2 r^2}{k}-\frac{\kappa'^2 r^4}{k}-P_2^{||}-P_3^{||}-P_3^{\perp}\,.
\end{aligned}$$

For $k<i\leq D$:

$$\tilde{\Delta}_i\leq\tilde{\lambda}_{m,k+1}^2\leq(\tilde{\lambda}_{m,k+1}^{\perp})^2\leq(\lambda_{m,k+1}^{\perp})^2+P_3^{\perp}+\sigma^2\leq\frac{\kappa'^2r^4}{k}+P_3^{\perp}+\sigma^2\,.$$

Therefore, in order for $\tilde{\Delta}_k$ to be the largest gap, we have the sufficient condition:

$$\frac{\lambda_{\min}^2 r^2}{k}-\frac{\kappa'^2 r^4}{k}-P_2^{||}-P_3^{||}-P_3^{\perp}\geq\left(\frac{\delta^2r^2}{k}+2P_2^{||}+2P_3^{||}+P_4\right)\vee\left(\frac{\kappa'^2r^4}{k}+P_3^{\perp}+\sigma^2\right). \tag{8.10}$$

Remark 4. *Note the "inflation effect" of noise: the increase in the bottom singular values by σ^2 is somewhat mitigated by the tangent singular values being "inflated" by σ^2. This phenomenon had been noticed before in the literature (see e.g. [80, 81]).*

Observe that (8.10) implies $(\tilde{\lambda}_{m,k}^{||})^2>(\tilde{\lambda}_{m,k+1}^{\perp})^2$, which we had assumed in P_4.

The largest gap, part II. We now put together the bounds above in order to determine a range for r that guarantees $\tilde{\Delta}_k$ is the largest gap with high probability. Restricting ourselves, following (8.1), for the rest of the proof to $\Omega_{t,0}$, we have to consider only the case $m\geq\overline{n}$. But for each such m, $\cap_{i=1}^5\Omega_{t,i}$ has high probability (uniformly in m) and the bounds in Table 16 hold with m replaced by the smaller quantity $\overline{n}$, on an event of high probability intersected with $\Omega_{t,6}$. Combining those with equation (8.10), upon letting $2\gamma_{\overline{n}}^2:=\lambda_{\min}^2-\delta^2-3\lambda_{\max}\epsilon(1+\epsilon)$, and replacing Ct by t, increasing $\overline{n}$ as needed in order to maintain $\epsilon\leq 1$, inequality (8.10) is implied by

$$\begin{cases} r\leq\frac{\lambda_{\max}}{4\kappa'}\left(1+\frac{6\gamma_{\overline{n}}^2}{\lambda_{\max}^2}\right) \\ \left[\frac{\gamma_{\overline{n}}^2}{k(1+\epsilon)}-\frac{2\sigma\kappa'}{\sqrt{k}}\left(\sqrt{\frac{D}{\overline{n}}}t+1\right)\right]r^2-\left[\frac{\lambda_{\max}\sigma}{\sqrt{k}}\left(\sqrt{\frac{D}{\overline{n}}}t+\epsilon\right)\right]r-\sigma^2\sqrt{\frac{D}{\overline{n}}}t\left(\mathbf{1}_{\overline{n}\geq CD}+\sqrt{\frac{D}{\overline{n}}}\mathbf{1}_{\overline{n}\leq CD}\right)\geq 0\end{cases}$$

where the first which inequality comes from curvature terms, and the second one from noise terms. Upon letting $\varphi_{\overline{n},t}^2:=\frac{\gamma_{\overline{n}}^2}{1+\epsilon}-\sigma\kappa'\sqrt{k}\left(\sqrt{\frac{D}{\overline{n}}}+\epsilon\right)$, and using the assumption $\epsilon\leq 1$ to simplify various terms, we obtain that for r in the range

$$\epsilon\frac{\sigma\sqrt{D}}{\varphi_{\overline{n},t}}\left[\frac{\lambda_{\max}}{\varphi_{\overline{n},t}}\vee\left(\mathbf{1}_{\overline{n}\leq CD}+\sqrt[4]{\frac{\overline{n}}{D}}\mathbf{1}_{\overline{n}\geq CD}\right)\right]\leq r\leq\frac{\lambda_{\max}}{4\kappa'}\left(1+\frac{6\gamma_{\overline{n}}^2}{\lambda_{\max}^2}\right)$$

the gap $\Delta_k(\mathrm{cov}(\widetilde{\mathbf{X}_n{}^{[r]}}))$ is the largest gap with probability at least $1 - ce^{-ct}$. Furthermore, since

$$\mathrm{cov}(X^{[r]}r) - \mathrm{cov}(\widetilde{\mathbf{X}_n{}^{[r]}}r) =$$

$$= \begin{bmatrix} \mathrm{cov}(X^{[r]^{||}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{||}}}) + \sigma^2 I_k & \mathrm{cov}(X^{[r]^{||}}, X^{[r]^{\perp}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{||}}}, \widetilde{\mathbf{X}_n^{[r]^{\perp}}}) \\ \mathrm{cov}(X^{[r]^{\perp}}, X^{[r]^{||}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{\perp}}}, \widetilde{\mathbf{X}_n^{[r]^{||}}}) & \mathrm{cov}(X^{[r]^{\perp}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{\perp}}}) + \sigma^2 I_{D-k} \end{bmatrix} - \sigma^2 I_D$$

combining all the bounds above we obtain

$$\left\| \mathrm{cov}(X^{[r]}r) - \mathrm{cov}(\widetilde{\mathbf{X}_n{}^{[r]}}r) + \sigma^2 I_D \right\| \leq \left(2\sigma^2 \sqrt{\frac{D}{\overline{n}}} t \left(1 + \sqrt{\frac{D}{\overline{n}}} t \mathbf{1}_{\overline{n} \leq CD} \right) \right.$$
$$\left. + \frac{\lambda_{\max}\sigma}{\sqrt{k}} \left(\epsilon + \sqrt{\frac{D}{\overline{n}}} t \right) r + \left(\frac{\epsilon \lambda_{\max}}{\sqrt{k}} + \sigma\kappa' \left(2\sqrt{\frac{D}{\overline{n}}} t + 1 \right) \right) \frac{r^2}{\sqrt{k}} + \frac{2\lambda_{\max}\kappa'}{k} r^3 + \frac{2\kappa'}{k} r^4 \right) (1 + \epsilon) \,.$$

Finally, recall that Π_k (respectively $\tilde{\Pi}_k$) is the space spanned by the top k singular vectors of $\mathrm{cov}(X^{[r]})$ (respectively $\mathrm{cov}(\widetilde{\mathbf{X}_n{}^{[r]}}) - \sigma^2 I$). Then in order to prove the bound in (iii), we use the classical Davis-Kahan "$\sin\theta$" Theorem, which gives

$$|\sin\Theta(\Pi_k, \tilde{\Pi}_k)| \leq \frac{||(\mathrm{cov}(X^{[r]}) - \mathrm{cov}(\widetilde{\mathbf{X}_n{}^{[r]}}) + \sigma^2 I)\Pi_k||}{\lambda_k^2(\mathrm{cov}(X^{[r]})) - \lambda_{k+1}^2(\mathrm{cov}(\widetilde{\mathbf{X}_n{}^{[r]}}) - \sigma^2 I)}$$
$$\leq \frac{||\mathrm{cov}(X^{[r]^{||}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{||}}}) + \sigma^2 I^{||}|| + ||\mathrm{cov}(X^{[r]^{||}}, X^{[r]^{\perp}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[r]^{||}}}, \widetilde{\mathbf{X}_n^{[r]^{\perp}}})||}{|\lambda_k^2(\mathrm{cov}(X^{[r]}) - \lambda_{k+1}^2(\mathrm{cov}(X^{[r]}))| - |\lambda_{k+1}^2(\mathrm{cov}(X^{[r]})) - \lambda_{k+1}^2(\mathrm{cov}(\widetilde{\mathbf{X}_n{}^{[r]}}) - \sigma^2 I)|}$$
$$\leq \frac{\sigma^2 \sqrt{\frac{D}{\overline{n}}} t + \frac{\sqrt{\lambda}_{\max}\sigma}{\sqrt{k}} \left(\sqrt{\frac{D}{\overline{n}}} t + \epsilon \right) t + \frac{\epsilon\lambda_{\max} + \sigma\sqrt{k}\kappa'}{k} r^2 + \frac{\lambda_{\max}\kappa'}{k} r^3 + \frac{\frac{\kappa'^2}{k}\lambda_{\max}^2}{\lambda_{\min}^2 - \kappa'^2 r^2} r^4}{\frac{\lambda_{\min}^2 - \kappa'^2 r^2}{k} r^2 - \sigma^2 \mathbf{1}_{\overline{n} \leq CD} - \sigma\sqrt{\frac{D}{\overline{n}}} t \left[\frac{2\kappa' r^2}{\sqrt{k}} + \sigma \mathbf{1}_{\overline{n} \geq CD} + \sigma\sqrt{\frac{D}{\overline{n}}} \mathbf{1}_{\overline{n} \leq CD} \right]}.$$

□

9 Appendix: Proof of Proposition 2

In all that will follow we shall fix z and r and prove a bound on $||\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}}) - \mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})||$.

9.1 Some Preliminary definition

Recall that $r_=^2 := r^2 - 2\sigma^2 D$ and let $r_-^2 := r^2 - \sigma^2 D$. Let $\tilde{Z}_{\mathcal{M}}$ be a closest point to $\tilde{Z}$ on $\mathcal{M}$: $\tilde{Z}_{\mathcal{M}} \in \mathrm{argmin}_{y \in \mathcal{M}} \left\| \tilde{Z} - y \right\|$. We let, in what follows,

$$\begin{aligned} \xi &:= \frac{\sigma\sqrt{D}}{r}\,, & d &:= D - k \quad, & \rho(x) &:= ||x - \tilde{Z}_{\mathcal{M}}|| \\ f(r) &:= \sqrt{r^2 + \sigma^2 d}\,, & s(r) &:= \sqrt{r^2 - \sigma^2 d}\,, & q &:= s^2\sigma^2\sqrt{D} + 4t_0\sigma r_-(1 + 2\kappa k^{-\frac{1}{2}} r_-) \end{aligned} \tag{9.1}$$

where we will choose the parameters s, t_0 later, and with the following constraints:

$$\frac{r}{\sqrt{k}} \in \left(3\sigma\sqrt{k \vee \frac{D}{k}}, \frac{1}{\kappa} \right)\,,\; s^2 \leq \sqrt{D}\,,\; t_0^2 \in \left(0, \log\frac{r/\sqrt{k}}{3\sigma\sqrt{k}} \right] \,. \tag{9.2}$$

where σ is small enough so that the first interval is not empty. The lower bound on r is required in order to work at scales larger than the scale of the noise, and the upper bound is motivated by (8.2), which shows that at larger scales the curvature terms dominate. The interval is not empty if the "scale of noise" $\sigma\sqrt{D}$ is below the "scale of the curvature", here $\sqrt{k}/\kappa$. For these values of the parameters we have the the following bound on "noise to signal ratio": $\xi < \frac{1}{3}\left(\frac{\sqrt{D}}{k} \wedge 1\right)$, i.e. $r > 3\sigma(\sqrt{D} \vee k)$, and

$$q \leq s^2 D^{-\frac{1}{2}}\xi^2 r^2 + 4t_0\sigma r_- \left(1 + \frac{2r_-}{r}\right) \leq (\xi s^2 + 12t_0)\frac{\xi r^2}{\sqrt{D}} \leq c_{4,\xi,s,t_0}\frac{\xi r^2}{\sqrt{D}} = c_{4,\xi,s,t_0}\sigma r\,. \tag{9.3}$$

with $c_{4,\xi,s,t_0} := \xi s^2 + 12t_0$.

Numerical constants will be denoted by C and their value may change at every instance.

Remark 5. *We may assume all the realizations of the noise $\{N_i\}_{i=1}^n$, are such that $\{\sigma N_i\}_{i=1}^n$ are bounded by $\sigma^2 D(1 + C\frac{\ln(n \wedge D}{\sqrt{D}}) \approx \sigma^2 D$. Thus by a tiny reduction in the number of samples, we may assume we have i.i.d., bounded noise vectors. We do not address here how this outlier detection step is implemented algorithmically (see for example [87]). A result showing that the removal of a small number of outliers as described here has a negligible effect on the covariance matrices we considered may be found in Appendix E of [32].*

9.2 General Proof Strategy

To derive a bound on on $||\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r=]}}) - \mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})||$, we consider some intermediate, objects obtained perturbing sets of points, and relating the corresponding covariance matrices. A basic idea that we use throughout this section is that if two (possibly random) nested sets contains roughly the same number of points, then the corresponding covariance matrices are also close (in operator norm and with high probability). This intuition is made precise in the two following Lemmata of which we shall make extensive use.

Lemma 2 (Covariance perturbation, worst case)**.** *Let $\mathbf{Y}_n$ be any (deterministic) set of n points in $\mathbb{R}^D$ and $\mathbf{E} \subseteq \mathbf{Y}_n$ such that $|\mathbf{E}| \leq \lfloor \epsilon \cdot n \rfloor$, $0 < \epsilon \leq 1$. Then*

$$||\mathrm{cov}(\mathbf{Y}_n) - \mathrm{cov}(\mathbf{Y}_n \setminus \mathbf{E})|| \leq 6\epsilon M^2. \tag{9.4}$$

Proof. For a set $\mathbf{A} \subset \mathbb{R}^D$ let $m(\mathbf{A}) = \frac{1}{|\mathbf{A}|}\sum_{x\in\mathbf{A}} x$, $C(\mathbf{A}) = \frac{1}{|\mathbf{A}|}\sum_{x\in\mathbf{A}} x \otimes x$, so that $\mathrm{cov}(\mathbf{A}) = C(\mathbf{A}) - m(\mathbf{A}) \otimes m(\mathbf{A})$. Then

$$\begin{aligned} ||\mathrm{cov}(\mathbf{Y}_n) - &\mathrm{cov}(\mathbf{Y}_n \setminus \mathbf{E})|| \\ &\leq ||C(\mathbf{Y}_n) - C(\mathbf{Y}_n \setminus \mathbf{E})|| + ||m(\mathbf{Y}_n) \otimes m(\mathbf{Y}_n) - m(\mathbf{Y}_n \setminus \mathbf{E}) \otimes m(\mathbf{Y}_n \setminus \mathbf{E})|| \\ &\leq ||C(\mathbf{Y}_n) - C(\mathbf{Y}_n \setminus \mathbf{E})|| + 2M||m(\mathbf{Y}_n) - m(\mathbf{Y}_n \setminus \mathbf{E})|| \end{aligned}$$

Let $e = |\mathbf{Y}_n \setminus \mathbf{E}|$, then we have

$$\begin{aligned} &||m(\mathbf{Y}_n) - m(\mathbf{Y}_n \setminus \mathbf{E})|| \leq ||\frac{1}{n}\sum_{x\in\mathbf{Y}_n} x - \frac{1}{e}\sum_{x\in\mathbf{Y}_n\setminus\mathbf{E}} x|| = \\ &||\frac{1}{n}\sum_{x\in\mathbf{E}} x + (1 - \frac{n}{e})\frac{1}{n}\sum_{x\in\mathbf{Y}_n\setminus\mathbf{E}} x|| \leq 2\epsilon M. \end{aligned}$$

The same reasoning gives $||C(\mathbf{Y}_n) - C(\mathbf{Y}_n \setminus \mathbf{E})|| \leq 2M^2\epsilon$, simply replacing x with $x \otimes x$ and noting that $||x \otimes x|| \leq M^2$. Then (2) easily follows combining the above inequalities. □

Recalling definitions (3.2), (3.3),(3.4),(3.5), in the next lemma we extend the above result allowing to random sets.

Lemma 3 (Covariance Perturbation, random case). *Let $\mathbf{X}_n$ be n i.i.d. copies of a random variable X with distribution μ_X. Let A, B be two (μ_X-measurable) sets in $\mathbb{R}^D$, with $B \subseteq A$, $\mu_X(B) \leq \delta\mu_X(A)$, and A bounded by M, and write $n_B = n_B(X)$, $n_A = n_A(X)$. Then for $s \geq 1$, $t > 0$, $n \geq t^2/\mu_X(A)$:*

$$\mathbb{P}\left(n_B \leq 4s^2\left(\delta \vee \frac{1}{\mu_X(A)n}\right)n_A\right) \geq 1 - e^{-\frac{1}{8}t^2} - 2e^{-\frac{1}{3}s^2(\delta\mu_X(A)n\vee 1)},$$

and with the same probability and conditions,

$$||\mathrm{cov}(\mathbf{A}_n) - \mathrm{cov}(\mathbf{A}_n \setminus \mathbf{B}_n)|| \leq Cs^2\left(\delta \vee \frac{1}{\mu_X(A)n}\right)M^2, \tag{9.5}$$

where $\mathbf{A}_n \setminus \mathbf{B}_n = \{X_i \in \mathbf{X}_n \mid i \in I_A(X) \setminus I_B(X)\}$. The same conclusion, with M^2 replaced by $M^2 + \sigma^2 D$, holds if $\mathbf{A}_n$ (resp. $\mathbf{B}_n$) is replaced by $\widetilde{\mathbf{A}_n} := \mathbf{A}_n + \mathbf{Y}_n$ (resp. $\tilde{\mathbf{B}}_n$) where $\mathbf{Y}_n$ are n i.i.d. copies of a random variable Y which is independent of X and bounded, $||Y|| \leq \sigma\sqrt{D}$.

Proof. Let $\Omega := \{n_A \leq \frac{1}{2}\mu_X(A)n\}$: the assumption on n and Chernoff's bound [86] imply $\mathbb{P}(\Omega) \leq e^{-\frac{1}{8}t^2}$. Now n_B is $\mathrm{Bin}(n, \mu_X(B))$; let $\tilde{n}_B$ be $\mathrm{Bin}(n, \delta\mu_X(A) \vee \frac{1}{n})$. Then on Ω:

$$\begin{aligned}
\mathbb{P}\left(\left\{n_B > 2(1+s^2)\left(\delta \vee \frac{1}{\mu_X(A)n}\right)n_A\right\} \cap \Omega^c\right) &\leq \mathbb{P}\left(n_B > (1+s^2)\left(\delta \vee \frac{1}{\mu_X(A)n}\right)\mu_X(A)n\right) \\
&\leq \mathbb{P}\left(\tilde{n}_B > (1+s^2)\left(\delta\mu_X(A) \vee \frac{1}{n}\right)n\right) \\
&= \mathbb{P}\left(\tilde{n}_B > (1+s^2)\mathbb{E}[\tilde{n}_B]\right) \\
&\leq e^{-\frac{s^2}{3}(\delta\mu_X(A)n\vee 1)}
\end{aligned}$$

for any $s \geq 1$, the last line also following from a Chernoff inequality [86]. Thus for any $s \geq 1$, $n \geq t^2/\mu_X(A)$:

$$n_B \leq 4s^2\left(\delta \vee \frac{1}{\mu_X(A)n}\right)n_A$$

with probability at least $1 - e^{-\frac{1}{8}t^2} - 2e^{-\frac{1}{3}s^2(\delta\mu_X(A)n\vee 1)}$. An application of Lemma 2 yields the desired bound. The case when noise is added follows in a similar fashion. □

Given the above results the proof of Prop. 2, develops in two steps:

(i) *recentering*: we first show that $\widetilde{\mathbf{X}_n^{[z,r_=]}}$ and $\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}$ are close w.h.p., in the sense that the set of sample points within distance $r_=$ of $z \in \mathcal{M}$ is roughly equivalent to the set of points within distance r_- of a noisy center $\tilde{Z} = z + \sigma N \notin \mathcal{M}$; thus by a change in scale, we can move from a center $z \in \mathcal{M}$ to a noisy center $\tilde{Z} \notin \mathcal{M}$. We prove this by bounding the following perturbations:

$$\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \to \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \cup \mathbf{A}_{1,t_0} \setminus \mathbf{A}_{2,t_0} \to \widetilde{\mathbf{X}_n^{[z,r_=]}},$$

where

$$\begin{aligned}
\mathbf{A}_{1,t_0} &:= \widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2-q}]}} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}^{c} = \{\tilde{X}_i : ||X_i - z|| < \sqrt{r_=^2 - q} \wedge ||X_i - \tilde{Z}|| > r_-\} \\
\mathbf{A}_{2,t_0} &:= \widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2+q}]}}^{c} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} = \{\tilde{X}_i : ||X_i - z|| > \sqrt{r_=^2 + q} \wedge ||X_i - \tilde{Z}|| < r_-\}
\end{aligned} \tag{9.6}$$

where s^2, t_0 are parameters to be chosen later. The first perturbation is small once we show that $|\mathbf{A}_{1,t_0}|$ and $|\mathbf{A}_{2,t_0}|$ are small relative to $|\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}|$; the second perturbation is small once we prove that $\widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2+q}]}} \setminus \widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2-q}]}}$, which contains the set where $\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \cup \mathbf{A}_{1,t_0} \setminus \mathbf{A}_{2,t_0}$ and $\widetilde{\mathbf{X}_n^{[z,r_=]}}$ differ, has small cardinality relative to $|\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}|$. Lemma 3 below then implies that $||\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}})||$ is small.

(ii) *Bringing in the noise:* the second step is to show that the sets $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}$ and $\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}$ are close w.h.p.: the set of noisy points that were within distance r_- of $\tilde{Z}$ before they were corrupted by noise is roughly equivalent to the set of noisy points within distance r of $\tilde{Z}$. In other words, intersecting with a ball and then adding noise is equivalent to adding noise and then intersecting with a ball of slightly different radius. To this end we bound the following perturbations:

$$\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} = (\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}) \cup \mathbf{I} \to \tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}} = (\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \setminus \mathbf{Q}_1) \cup \mathbf{Q}_2 \to \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}\,,$$

where

$$\begin{aligned}
\mathbf{I} &:= \tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}^{\,c} && =\{\tilde{X}_i : ||\tilde{X}_i - \tilde{Z}|| < r \wedge ||X_i - \tilde{Z}|| > r\} \\
\mathbf{Q}_1 &:= \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}^{\,c} \cap \left(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}\right)^c && =\{\tilde{X}_i : ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, r_-) \wedge ||\tilde{X}_i - \tilde{Z}|| > r\} \\
\mathbf{Q}_2 &:= \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}^{\,c} \cap \tilde{\mathbf{X}}_n^{[\tilde{Z},r]} && =\{\tilde{X}_i : ||X_i - \tilde{Z}|| \in [r_-, r] \wedge ||\tilde{X}_i - \tilde{Z}|| < r\}
\end{aligned} \tag{9.7}$$

The first perturbation is small if $|\mathbf{I}|$ is small relative to $|\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}|$, and the second perturbation is small if both $|\mathbf{Q}_1|$ and $|\mathbf{Q}_2|$ are small relative to $|\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}|$. Once this is established, Lemma 3 allows us to conclude that $||\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}})||$ is small.

Table 17 summarizes the bounds we will give on the perturbations above (and the probabilities with which such bounds hold). We define the following event characterizing when $\tilde{Z} = z + \sigma N$ is not an "outlier":

$$\Omega_{s,0} := \{\omega : |\,||N(\omega)||^2 - \sigma^2 D| \le s^2\sigma^2\sqrt{D}\}\,, \tag{9.8}$$

which has probability at least $1 - 2e^{-Cs^4}$ for $s^2 \le \sqrt{D}$.

9.3 Basic estimates and Lemmata

Lemma 4 (Distance of $\tilde{Z}$ to $\mathcal{M}$)**.** *Let $\tilde{Z}_{\mathcal{M}}$ be a closest point to $\tilde{Z}$ in $\mathcal{M}$. Under our usual assumptions, with probability at least $1 - 6e^{-cs^4}$*

$$\sigma^2 D\left(1 - (8\sqrt{2}+1)s^2 D^{-\frac{1}{2}}\right) \le \left\|\tilde{Z} - \tilde{Z}_{\mathcal{M}}\right\|^2 \le \sigma^2 D\left(1 + s^2 D^{-\frac{1}{2}}\right)\,. \tag{9.9}$$

Proof. By definition of $\tilde{Z}_{\mathcal{M}}$, on $\Omega_{s,0}$: $\left\|\tilde{Z} - \tilde{Z}_{\mathcal{M}}\right\|^2 \le \left\|\tilde{Z} - z\right\|^2 \le \sigma^2 D\left(1 + s^2 D^{-\frac{1}{2}}\right)$. Furthermore, $\left\|z - \tilde{Z}_{\mathcal{M}}\right\| \le \left\|z - \tilde{Z}\right\| + \left\|\tilde{Z} - \tilde{Z}_{\mathcal{M}}\right\| \le 2\sigma\sqrt{D}\left(1 + s^2 D^{-\frac{1}{2}}\right)^{\frac{1}{2}}$, so that
$\tilde{Z}_{\mathcal{M}} \in B_z\left(2\sigma\sqrt{D}\left(1 + s^2 D^{-\frac{1}{2}}\right)^{\frac{1}{2}}\right)$. Letting $P^{(z,2\sigma\sqrt{D}(1+s^2D^{-\frac{1}{2}})^{\frac{1}{2}})}$ be the approximate tangent plane projection as in our usual assumptions, and writing $\tilde{Z}_{\mathcal{M}} - z = \tilde{Z}_{\mathcal{M}}^{||} + \tilde{Z}_{\mathcal{M}}^{\perp}$ and $\tilde{Z} - z = \sigma(N^{||} + N^{\perp})$, the subgaussian condition on the noise gives that, on $\Omega_{s,0}$, with probability at least $1 - 4e^{-cs^4}$:

$$|\langle\tilde{Z}_{\mathcal{M}}^{||}, \sigma N^{||}\rangle| \le s^2\sigma\left(2\sigma\sqrt{D}\left(1 + s^2 D^{-\frac{1}{2}}\right)^{\frac{1}{2}}\right)\ ,\ |\langle\tilde{Z}_{\mathcal{M}}^{\perp}, \sigma N^{\perp}\rangle| \le s^2\sigma\kappa k^{-\frac{1}{2}}\left(4\sigma^2 D\left(1 + s^2 D^{-\frac{1}{2}}\right)\right)$$

for $s^2 \le D$. Since $\sigma\sqrt{D} \le \frac{\sqrt{k}}{2\sqrt{2}\kappa}$ and $s^2 \le \sqrt{D}$, by our usual assumptions:

$$\begin{aligned}
\left\|\tilde{Z}_{\mathcal{M}} - \tilde{Z}\right\|^2 &= \left\|\tilde{Z}_{\mathcal{M}} - z\right\|^2 + \left\|z - \tilde{Z}\right\|^2 + 2\langle\tilde{Z}_{\mathcal{M}}^{||}, \sigma N^{||}\rangle + 2\langle\tilde{Z}_{\mathcal{M}}^{\perp}, \sigma N^{\perp}\rangle \\
&\ge \left\|z - \tilde{Z}\right\|^2 + 2\langle\tilde{Z}_{\mathcal{M}}^{||}, \sigma N^{||}\rangle + 2\langle\tilde{Z}_{\mathcal{M}}^{\perp}, \sigma N^{\perp}\rangle \\
&\ge \sigma^2 D\left(1 - s^2 D^{-\frac{1}{2}}\right) - 8\sqrt{2}s^2\sigma^2\sqrt{D} = \sigma^2 D\left(1 - (8\sqrt{2}+1)s^2 D^{-\frac{1}{2}}\right)\,.
\end{aligned}$$

Removing the conditioning on $\Omega_{s,0}$, the bound (9.9) is obtained, with the desired probability. □

Define, for $r_1 > r_2$,

$$V_{\tilde{Z}}(r_1, r_2) := \frac{\mu(B_{\tilde{Z}}(r_1) \setminus B_{\tilde{Z}}(r_2))}{\mu(B_{\tilde{Z}}(r_2)} = \frac{\mu(B_{\tilde{Z}}(r_1))}{\mu(B_{\tilde{Z}}(r_2)} - 1\,. \tag{9.10}$$

Lemma 5 (Volume estimates). *With our usual assumptions, $r_1 \geq r_2 > d(\tilde{Z}, \mathcal{M})$,*

$$\frac{\mu_X(B_{\tilde{Z}}(r_1))}{\mu_X(B_{\tilde{Z}}(r_2))} \leq \left(\frac{r_1}{r_2}\right)^{2k} \left(\frac{1 - \frac{d(\tilde{Z},\mathcal{M})^2}{r_1^2}}{1 - \frac{d(\tilde{Z},\mathcal{M})^2}{r_2^2}}\right)^k \quad , \quad \frac{\mu_X(B_{\tilde{Z}}(r_2))}{\mu_X(B_{\tilde{Z}}(r_1))} \geq \left(\frac{r_2}{r_1}\right)^{2k} \left(\frac{1 - \frac{d(\tilde{Z},\mathcal{M})^2}{r_2^2}}{1 - \frac{d(\tilde{Z},\mathcal{M})^2}{r_1^2}}\right)^k \tag{9.11}$$

If in addition $0 < s^2 \leq \sqrt{D}$, on $\Omega_{s,0}$ (as in (9.8)), we have

$$V_{\tilde{Z}}(r_1, r_2) \leq e^{2k\frac{r_1-r_2}{r_2}\left(1+\left(1+\frac{r_1-r_2}{r_2}\right)\left(1-\frac{d(\tilde{Z},\mathcal{M})^2}{r_2^2}\right)^{-1}\right)} - 1 \tag{9.12}$$

and if furthermore $\frac{r_1-r_2}{r_2} \leq (2k)^{-1}$ and $r_2 \geq r_=$, we also have

$$V_{\tilde{Z}}(r_1, r_2) \leq 10k\frac{r_1 - r_2}{r_2}\,. \tag{9.13}$$

Finally, always on $\Omega_{s,0}$, for $c_{5,s,\xi} \leq 1$, $c_{6,\xi,s} \geq 1$, both tending to 1 as either $s^2 D^{-\frac{1}{2}}$ or ξ tend to 0,

$$c_{5,s,\xi}\mu_X(B_{\tilde{Z}}(r_-)) \;\leq\; \mu_X(B_z(r_=)) \;\leq c_{6,\xi,s}\,\mu_X(B_{\tilde{Z}}(r_-)) \tag{9.14}$$

Proof. By the usual assumptions, if we let $\rho_i^2 = r_i^2 - d(\tilde{Z}, \mathcal{M})^2$, $i = 1, 2$, we have the following estimates:

$$\frac{\mu_X(B_{\tilde{Z}}(r_1))}{\mu_X(B_{\tilde{Z}}(r_2))} = \frac{v_{\tilde{Z}}(\rho_1)}{v_{\tilde{Z}}(\rho_2)}\left(\frac{\rho_1}{\rho_2}\right)^k = \frac{v_{\tilde{Z}}(\rho_2 + (\rho_1 - \rho_2))}{v_{\tilde{Z}}(\rho_2)}\left(\frac{\rho_1}{\rho_2}\right)^k \leq \left(1 + \frac{\rho_1 - \rho_2}{\rho_2}\right)^k \left(\frac{\rho_1}{\rho_2}\right)^k \leq \left(\frac{\rho_1}{\rho_2}\right)^{2k}$$

from which the first inequality in (9.11) follows. The other bounds are proved similarly. To prove inequality (9.13), letting $d^2 = d^2(\tilde{Z}, \mathcal{M})$ and $\Delta r = r_1 - r_2$ for notational convenience, we have

$$V_{\tilde{Z}}(r_1, r_2) \leq \left(\frac{r_2 + \Delta r}{r_2}\right)^{2k} \left(\frac{(r_2 + \Delta r)^2 - d^2}{r_2^2 - d^2}\right)^k - 1 \leq \left(1 + \frac{\Delta r}{r_2}\right)^{2k} \left(1 + \frac{\Delta r}{r_2}\frac{2 + \frac{\Delta r}{r_2}}{1 - \frac{d^2}{r_2^2}}\right)^k - 1$$

and (9.12) follows by using the inequality $(1 + x)^\alpha \leq e^{\alpha x}$. In order to obtain (9.13) we proceed as follows:

$$\begin{aligned} V_{\tilde{Z}}(r_1, r_2) &\leq \left(1 + 2k\frac{\Delta r}{r_2}\right)\left(1 + k\frac{\Delta r}{r_2}\frac{2 + \frac{\Delta r}{r_2}}{1 - \frac{d^2}{r_2^2}}\right) - 1 \\ &\leq \left(1 + 2k\frac{\Delta r}{r_2}\right)\left(1 + k\frac{\Delta r}{r_2}\left(2 + \frac{2}{k}\right)\right) - 1 \quad \leq 10k\frac{\Delta r}{r_2}\,, \end{aligned}$$

where we used the inequality $(1 + x)^\alpha \leq 1 + \alpha x$ for $x \in [0, 1/\alpha]$, applied to $x = \Delta r/r_2$ and $\alpha = 2k$ for the first term in the product above, and $x = \frac{\Delta r}{r_2}\left(2 + \frac{\Delta r}{r_2}\right)\left(1 - \frac{d^2}{r_2^2}\right)^{-1}$ and $\alpha = k$ for the second term, and observed that our assumptions guarantee that $x \leq 1/\alpha$ in both cases.

We now prove the volume comparison estimate (9.14). Let $\zeta_{\tilde{Z}} = \left\|\tilde{Z} - \tilde{Z}_{\mathcal{M}}\right\|$.

$$\mu_X(B_z(r_=)) = v_z(r_=)\mu_{\mathbb{R}^k}(\mathbb{B}^k)r_=^k\,,\ \mu_X(B_{\tilde{Z}}(r_-)) = v_{\tilde{Z}}(\sqrt{r_-^2 - \zeta_{\tilde{Z}}^2})\mu_{\mathbb{R}^k}(\mathbb{B}^k)(r_-^2 - \zeta_{\tilde{Z}}^2)^{\frac{k}{2}}\,.$$

Assume $\zeta_{\tilde{Z}}^2 \in \sigma^2 D[1 - \frac{13s^2}{\sqrt{D}}, 1 + s^2 D^{-\frac{1}{2}}]$ and $\left\|z - \tilde{Z}\right\|^2 \in \sigma^2 D[1 - s^2 D^{-\frac{1}{2}}, 1 + s^2 D^{-\frac{1}{2}}]$, which by Lemma 4 and (9.8) happens with probability at least $1 - ce^{-cs^4}$ for $s^2 \leq \sqrt{D}$. Then, since $\frac{\xi^2}{1-2\xi^2} \leq 1/4$,

$$r_{=}^k \left(1 - s^2 C_\xi D^{-\frac{1}{2}}\right)^{\frac{k}{2}} \leq (r_-^2 - \zeta_{\tilde{Z}}^2)^{\frac{k}{2}} \leq r_{=}^k \left(1 + \frac{13}{4} s^2 D^{-\frac{1}{2}}\right)^{\frac{k}{2}} .$$

By the smoothness of $v_{\tilde{Z}}(\rho)$ in ρ:

$$v_{\tilde{Z}}(r_{=}) \left(1 - s^2 C_\xi D^{-\frac{1}{2}}\right)^{\frac{k}{2}} \leq v_{\tilde{Z}}(\sqrt{r_-^2 - \zeta_{\tilde{Z}}^2}) \leq v_{\tilde{Z}}(r_{=}) \left(1 + \frac{13}{4} s^2 D^{-\frac{1}{2}}\right)^{\frac{k}{2}} \tag{9.15}$$

Finally, the smoothness of $v_{\tilde{Z}}(\rho)$ in $\tilde{Z}$ gives:

$$v_z(r_{=}) \left(1 - \left(1 + s^2 D^{-\frac{1}{2}}\right)^{\frac{1}{2}} \sqrt{C_\xi}\right) \leq v_{\tilde{Z}}(r_{=}) \leq v_z(r_{=}) \left(1 + \left(1 + s^2 D^{-\frac{1}{2}}\right)^{\frac{1}{2}} /2\right) \tag{9.16}$$

Combining the estimates above we obtain that (9.14) holds with probability at least $1 - ce^{-cs^4}$ for $s^2 \leq \sqrt{D}$, where

$$\begin{aligned} c_{5,s,\xi} &= \left(1 - \frac{s^2}{4\sqrt{D}}\right)^k \left(1 - \frac{1}{2}\left(1 + s^2 D^{-\frac{1}{2}}\right)^{\frac{1}{2}}\right) \\ c_{6,\xi,s} &= \left(1 + \frac{13s^2}{4\sqrt{D}}\right)^k \left(1 + \frac{1}{2}\left(1 + s^2 D^{-\frac{1}{2}}\right)^{\frac{1}{2}}\right) . \end{aligned}$$

□

9.4 Recentering

The goal of this section is to prove the following result.

Proposition 3. *Let the usual bounds* (9.2) *hold. Conditioning on* $\Omega_{s,0}$ *defined in* (9.8)*, for* $v \geq 1, t > 0$ *and* $n \geq Ct^2/\mu_X(B_z(r_{=}))$*, let* $s^2 < \frac{r^2/k}{12\sigma^2 D}\sqrt{D}$ *and set* $t_0^2 := C(1 \vee \log \frac{r/\sqrt{k}}{3\sigma\sqrt{k}})$*: then on an event* $\Omega_{v,t,1}$ *having probability as in Table 17, we have*

$$||\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_{=}]}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}})|| \leq Cv^2 \left((c_{4,\xi,s,t_0} \vee 1)\frac{\sigma k}{r} \vee \frac{1}{\mu_X(B_z(r_{=}))n}\right) r_-^2 .$$

Proof. It follows by combining the results of Lemmata 6 and 8 that

$$\begin{aligned} ||\mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_{=}]}})|| &\leq Cv^2 \left(\left(c_{6,\xi,s} e^{-t_0^2} + c_{4,\xi,s,t_0}\frac{\xi k}{\sqrt{D}}\right) \vee \frac{1}{\mu_X(B_z(r_{=}))n}\right) r_-^2 \\ &\leq Cv^2 \left((c_{6,\xi,s} \vee c_{4,\xi,s,t_0})\frac{\sigma k}{r} \vee \frac{1}{\mu_X(B_z(r_{=}))n}\right) r_-^2 \end{aligned}$$

by setting $t_0^2 := C(1 \vee \log \frac{r/\sqrt{k}}{3\sigma\sqrt{k}})$. Since $c_{6,\xi,s} \leq C$ by our usual bounds (9.2), we obtain the desired estimate. □

Lemma 6. *Let the usual bounds* (9.2) *hold and* t_0 *be as defined there. Define the random sets* $\mathbf{A}_{1,t_0}, \mathbf{A}_{2,t_0}$ *as in* (9.6)*:*

$$\begin{aligned} \mathbf{A}_{1,t_0} &:= \widetilde{\mathbf{X}_n^{[z,\sqrt{r_{=}^2 - q}]}} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}^c = \{\tilde{X}_i : ||X_i - z|| < \sqrt{r_{=}^2 - q} \wedge ||X_i - \tilde{Z}|| > r_-\} \\ \mathbf{A}_{2,t_0} &:= \widetilde{\mathbf{X}_n^{[z,\sqrt{r_{=}^2 + q}]}}^c \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} = \{\tilde{X}_i : ||X_i - z|| > \sqrt{r_{=}^2 + q} \wedge ||X_i - \tilde{Z}|| < r_-\} \end{aligned}$$

Conditioning on a given sample $\mathbf{X}_n(\omega)$, $\omega \in \Omega$ and $\Omega_{s,0} := \{\omega : | \, ||N(\omega)||^2 - \sigma^2 D| \leq s^2\sigma^2\sqrt{D}\}$ as in (9.8)*, for $v \geq 1, t > 0$, and $n \geq Ct^2/\mu_X(B_z(r_=))$, on an event having probability at least as in Table 17, we have*

$$||\mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \cup \mathbf{A}_{1,t_0} \setminus \mathbf{A}_{2,t_0})|| \leq Cv^2 \left(c_{6,\xi,s} e^{-t_0^2} \vee \frac{1}{\mu_X(B_z(r_=))n} \right) r_-^2 \, . \tag{9.17}$$

Proof. This Lemma is a consequence of the following two Lemmata, that estimate the cardinality of $\mathbf{A}_{1,t_0}, \mathbf{A}_{2,t_0}$ relative to that of $\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}$, and of an application of the covariance perturbation Lemma 3:

$$\begin{aligned}||\mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \cup \mathbf{A}_{1,t_0} \setminus \mathbf{A}_{2,t_0})|| &\leq Cv^2 \left(c_{6,\xi,s} e^{-t_0^2} \vee \frac{1}{\mu_X(B_z(r_=))n} \right) (r_-^2 + \sigma^2 D) \\ &\leq Cv^2 \left(c_{6,\xi,s} e^{-t_0^2} \vee \frac{1}{\mu_X(B_z(r_=))n} \right) \frac{1+\xi^2}{1-\xi^2} r_-^2 \, .\end{aligned}$$

This implies the desired bounds after recalling our usual assumptions on ξ. □

Lemma 7. *Let the usual bounds* (9.2) *hold and t_0 be as defined there. Conditioning on $\Omega_{s,0}$ defined in* (9.8)*, and with $c_{6,\xi,s} \geq 1$ as in Lemma 5, we have*

$$\mathbb{E}[|\mathbf{A}_{1,t_0}|] \leq ne^{-t_0^2}\mu_X(B_z(r_=)) \quad , \quad \mathbb{E}[|\mathbf{A}_{2,t_0}|] \leq c_{6,\xi,s} \, ne^{-t_0^2}\mu_X(B_z(r_=)) \, ,$$

Proof. We work in $B_z(\sqrt{r_=^2 - q})$, and with the associated projection $P^{||}$ as in our usual assumptions. We have $X - z = X^{||} + X^{\perp}$ and $z - \tilde{Z} = \sigma(N^{||} + N^{\perp})$. Then:

$$||X - \tilde{Z}||^2 = ||X - z||^2 + \sigma^2||N||^2 + 2\sigma\langle X^{||}, N^{||}\rangle + 2\sigma\langle X^{\perp}, N^{\perp}\rangle \tag{9.18}$$

and $\mathbb{E}_N||X - \tilde{Z}||^2 = \mathbb{E}_N||X - z||^2 + \sigma^2 D$. Fix $x \in B_z(\sqrt{r_=^2 - q}) \subseteq B_z(r_-)$: the subgaussian condition on the noise implies:

$$\mathbb{P}_N\left(|\sigma\langle X^{||}, N^{||}\rangle| > t_0\sigma r_-\right) \leq 2e^{-t_0^2} \quad , \qquad \mathbb{P}_N\left(|\sigma\langle X^{\perp}, N^{\perp}\rangle| > t_0\sigma\kappa r_-^2 k^{-\frac{1}{2}}\right) \leq 2e^{-t_0^2} \, , \tag{9.19}$$

i.e. the event $\Omega_{t_0,x} := \{|\sigma\langle X^{||}, N^{||}\rangle| > t_0\sigma r_-\} \cap \{|\sigma\langle X^{\perp}, N^{\perp}\rangle| > t_0\sigma\kappa k^{-\frac{1}{2}} r_-^2\}$ has probability at most $4e^{-t_0^2}$. On such an event (hence with the same probability) $X \in B_{\tilde{Z}}(r_-)$, since

$$||X - \tilde{Z}||^2 \leq ||X - z||^2 + \sigma^2 D + s^2\sigma^2\sqrt{D} + 2t_0\sigma(r_- + \kappa k^{-\frac{1}{2}} r_-^2) \leq r_=^2 - q + \sigma^2 D + q = r_-^2 \, .$$

Therefore Lemma 5 implies

$$\begin{aligned}\mathbb{E}[|\mathbf{A}_{1,t_0}|] &= \sum_{i=1}^{n} \mathbb{P}\left(||X_i - z|| < \sqrt{r_=^2 - q}\, , ||X_i - \tilde{Z}|| > r_-\right) \\ &= \sum_{i=1}^{n} \mathbb{P}\left(||X_i - \tilde{Z}|| > r_- \, \big| \, ||X_i - z|| < \sqrt{r_=^2 - q}\right) \cdot \mathbb{P}\left(||X_i - z|| < \sqrt{r_=^2 - q}\right) \\ &\leq ne^{-t_0^2}\, \mu_X(B_z(\sqrt{r_=^2 - q})) \leq ne^{-t_0^2}\, \mu_X(B_z(r_=)) \, .\end{aligned}$$

To prove the second bound in (9.19), on $\Omega_{s,0}$ we let $\tilde{r_-}^2 := r_-{}^2 + \sigma^2 D(1 + s^2 D^{-\frac{1}{2}})$: then $X_i \in B_{\tilde{Z}}(r_-)$ implies $X_i \in B_z(\tilde{r_-})$. Working on $B_z(\tilde{r_-})$ and using the associated projection $P^{||}$ as in our usual assumptions, for $X_i - z = X_i^{||} + X_i^{\perp}$ and $\tilde{Z} - z = \sigma(N^{||} + N^{\perp})$, the bounds (9.19) hold with $\tilde{r_-}$ replacing r_-, on an event Ω'_{t_0} having probability at least $1 - 4e^{-t_0^2}$. But then $X_i \in B_z(\sqrt{r_=^2 + q})$, since from (9.18) we have

$$\begin{aligned}||X_i - z||^2 &= ||X_i - \tilde{Z}||^2 - \sigma^2||N||^2 - 2\sigma\langle X_i^{||}, N^{||}\rangle - 2\sigma\langle X_i^{\perp}, N^{\perp}\rangle \\ &\leq r_-^2 - \sigma^2 D + s^2\sigma^2\sqrt{D} + 4t_0\sigma\left(r_- + 2\kappa k^{-\frac{1}{2}} r_-^2\right) = r_=^2 + q \, ,\end{aligned}$$

This implies, by applying Lemma 5 as above,

$$\mathbb{E}[|\mathbf{A}_{2,t_0}|] = \sum_{i=1}^{n} \mathbb{P}(\,||X_i - z|| > \sqrt{r_=^2 + q}\,,\, ||X_i - \tilde{Z}|| < r_-\,) \le c_{6,\xi,s}\, ne^{-t_0^2}\, \mu_X(B_z(r_=))\,.$$

□

Lemma 8. *Let the usual bounds* (9.2) *hold. Additionally, assume that*

$$s^2 < \frac{r^2/k}{12\sigma^2 D}\sqrt{D}, \qquad\qquad t_0 < \frac{r/\sqrt{k}}{144\sigma\sqrt{k}}\,. \tag{9.20}$$

Conditioning on $\Omega_{s,0}$ *defined in* (9.8)*, let* $\mathbf{A}_{1,t_0}, \mathbf{A}_{2,t_0}$ *be as in* (9.6)*. For* $v \ge 1, t > 0$ *and* $n \ge t^2/\mu_X(B_z(r_=))$*, on an event of probability at least* $1 - 2e^{-\frac{1}{3}v^2((\delta_2\overline{n})\vee 1)} - e^{-\frac{1}{8}t^2}$*, we have*

$$||\mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \cup \mathbf{A}_{1,t_0} \setminus \mathbf{A}_{2,t_0}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})|| \le Cv^2\left(c_{4,\xi,s,t_0}\frac{\sigma k}{r} \vee \frac{1}{\mu_X(B_z(r_=))n}\right) r_-^2\,. \tag{9.21}$$

Proof. On $\Omega_{s,0}$ we have the inclusions $\widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2-q}]}} \subseteq \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \cup \mathbf{A}_{1,t_0} \setminus \mathbf{A}_{2,t_0} \subseteq \widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2+q}]}}$ and $\widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2-q}]}} \subseteq \widetilde{\mathbf{X}_n^{[z,r_=]}} \subseteq \widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2+q}]}}$, so the set where $\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \cup \mathbf{A}_{1,t_0} \setminus \mathbf{A}_{2,t_0}$ and $\widetilde{\mathbf{X}_n^{[z,r_=]}}$ differ is contained in $\widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2+q}]}} \setminus \widetilde{\mathbf{X}_n^{[z,\sqrt{r_=^2-q}]}}$. In order to use (9.13) and (9.3) we observe that with the conditions (9.20) we have

$$\begin{aligned}\frac{\sqrt{r_=^2+q}-\sqrt{r_=^2-q}}{\sqrt{r_=^2-q}} &\le \frac{q}{r_=^2-q} \le \frac{c_{4,\xi,s,t_0}\sigma r}{r_=^2 - c_{4,\xi,s,t_0}\sigma r} \le \frac{c_{4,\xi,s,t_0}\sigma/r}{1-2\xi^2-c_{4,\xi,s,t_0}\sigma/r} \\ &\le_{2\xi^2\le\frac{1}{2}} \frac{2c_{4,\xi,s,t_0}\sigma/r}{1-2c_{4,\xi,s,t_0}\sigma/r} < \frac{1}{2k}\,,\end{aligned}$$

since the last inequality is equivalent to asking $2c_{4,\xi,s,t_0}\sigma/r < 1/(2k+1)$ which is implied by (9.20). Then:

$$\begin{aligned}&\mu_X(B_z(\sqrt{r_=^2+q}) \setminus B_z(\sqrt{r_=^2-q})) = V_z(\sqrt{r_=^2+q}, \sqrt{r_=^2-q})\mu(B_z(\sqrt{r_=^2-q})) \\ &\quad\le 10k\left(\sqrt{r_=^2+q}-\sqrt{r_=^2-q}\right)\left(r_=^2-q\right)^{-\frac{1}{2}}\mu(B_z(r_=)) \le 40c_{4,\xi,s,t_0}\sigma\sqrt{k}/(r/\sqrt{k})\cdot\mu(B_z(r_=))\,.\end{aligned}$$

The bound (9.21) follows by an application of Lemma 3 (and recalling that $r_=^2 + \sigma^2 D = r_-^2$). □

9.5 Bringing in the noise

We will show that the following perturbations are small (in the sense of cardinality of sets):

$$\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} = (\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}) \cup \mathbf{I} \to \tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}} = (\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \setminus \mathbf{Q}_1) \cup \mathbf{Q}_2 \to \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}$$

and conclude that the eigenvalues of the associated covariance matrices are close. As we compare the analysis and algorithm centered at a noisy point $\tilde{Z}$, and not a point on $\mathcal{M}$, we will be dealing with two natural distances:

. distance of a point from $\tilde{Z}_{\mathcal{M}}$; this will be our variable of integration in what follows and will determine the *volume* of the sets we will be considering.

. distance of a point from $\tilde{Z}$, which determines the *probability* of entering or exiting $B_{\tilde{Z}}(r)$ when noise is added.

$\mathbf{\Omega_{v,t,i}}$	**Event definition**	**Upper bound for δ_i**	**Probability**
$\Omega_{v,t,1}$	$\lvert\lvert\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}})\rvert\rvert$ $\leq Cv^2\left(\delta_1 \vee \frac{1}{\overline{n}_{r_=}}\right) r_-^2$	$\delta_1 := \left(c_{4,\xi,s,t_0} \vee 1\right) \frac{\sigma k}{r}$	$1 - 4e^{-\frac{1}{3}v^2(\delta_1 \overline{n}_{r_=} \vee 1)} - 2e^{-\frac{1}{8}t^2}$
$\Omega_{v,t,2}$	$\lvert\lvert\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}) - \mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}})\rvert\rvert$ $\leq Cv^2\left(\delta_2 \vee \frac{1}{\overline{n}_{r_=}}\right) r^2$	$\delta_2 := c_{8,\xi,s}\frac{\sigma k}{r}$	$1 - 2e^{-\frac{1}{3}v^2(\delta_2 \overline{n}_{r_=} \vee 1)} - e^{-\frac{1}{8}t^2}$
$\Omega_{v,t,3}$	$\lvert\lvert\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}})\rvert\rvert$ $\leq Cv^2\left(\delta_3 \vee \frac{1}{\overline{n}_{r_=}}\right) r^2$	$\delta_3 := \frac{\sigma k}{r}\left(1 \vee \frac{\sigma^2 D}{r^2/k}\right)\sqrt{\log \frac{r/\sqrt{k}}{3\sigma\sqrt{k}}}$	$1 - 4e^{-\frac{1}{3}v^2(\delta_3 \overline{n}_{r_=} \vee 1)} - 2e^{-\frac{1}{8}t^2}$

Figure 17: Events $\Omega_{v,t,i}$, $i = 1, \ldots, 3$, their definitions and lower bounds on their probabilities; here $t > 0$, $v \geq 1$, $n \geq t^2/\mu_X(\utilde{B_z(r_=)})$, our usual assumptions hold, and we have conditioned on $\Omega_{s,0}$ defined in (9.8). We conclude that $\mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})$ and $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$ are close when all of the above are small; each δ_i may be replaced with an upper bound, in particular for each δ_i we may substitute $\delta = \max_i \delta_i$. $\Omega_{s,t,1}$ and $\Omega_{s,t,2}$ are from recentering; $\Omega_{s,t,3}$,$\Omega_{s,t,4}$,$\Omega_{s,t,5}$ from noise.

Lemma 9 (Concentration of measure on spherical caps)**.** *For $0 \leq \theta \leq \pi$, we define V_θ^{D-1} to be the spherical cap of $\mathbb{S}^{D-1}$ centered at the north pole and subsuming an angle θ. Let $\mu_{\mathbb{S}^{D-1}}$ denotes the normalized (i.e. $\mu_{\mathbb{S}^{D-1}}(\mathbb{S}^{D-1}) = 1$) Hausdorff measure on $\mathbb{S}^{D-1}$. The function*

$$h(\theta) := \mu_{\mathbb{S}^{D-1}}(V_\theta^{D-1}),$$

satisfies the following properties:

1. *$0 = h(0) \leq h(\theta) \leq h(\pi) = 1$ for every $0 \leq \theta \leq \pi$, and $h(\theta)$ is strictly increasing.*
2. *If $\theta = \frac{\pi}{2} - t$ for any $0 \leq t \leq \frac{\pi}{2}$, $h(\theta) \leq e^{-\frac{1}{2}t^2 D}$.*

For a proof of these facts, see Lec. 19 of [88]. The angle subsumed by the spherical cap $\left(x + ||N|| \cdot \mathbb{S}^D\right) \cap B_{\tilde{Z}}(r)$ is

$$\theta_0(r, R, ||N||) := \arccos\left(\left(R^2 + ||N||^2 - r^2\right) / \left(2R||N||\right)\right) \tag{9.22}$$

for values of $r, R, ||N||$ for which the argument of $\arccos$ is in $[-1, 1]$. If a point x is at distance R from $\tilde{Z}$, if N has a spherically symmetric distribution we would have

$$\begin{aligned} \mathbb{P}_N\left(x + N \in B_{\tilde{Z}}(r) \middle| ||N|| = l\right) &\approx h(\theta_0(r, ||x - \tilde{Z}||, l)) \\ \mathbb{P}_N\left(x + N \notin B_{\tilde{Z}}(r) \middle| ||N|| = l\right) &\approx h(\pi - \theta_0(r, ||x - \tilde{Z}||, l)) \end{aligned} \tag{9.23}$$

All we shall need, in fact, is that the above relations hold approximately, with universal constants (independent of k, D, x, z, r, l):

$$\begin{aligned} \mathbb{P}_N\left(x + N \in B_{\tilde{Z}}(r) \middle| ||N|| = l\right) &\approx h(\theta_0(r, ||x - \tilde{Z}||, l)) \\ \mathbb{P}_N\left(x + N \notin B_{\tilde{Z}}(r) \middle| ||N|| = l\right) &\approx h(\pi - \theta_0(r, ||x - \tilde{Z}||, l)) \end{aligned} \tag{9.24}$$

In what follows, in order to ease the notation, we will actually assume the equalities (9.23), i.e. that the distribution of N is exactly spherically symmetric. The arguments are readily generalized to distributions which are only approximately spherical in the sense of (9.24). A simple computation shows that θ_0 is decreasing in $||N||$ for $R^2 < ||N||^2 + r^2$ and decreasing in R for $R^2 > ||N||^2 - r^2$. Finally, the following simple observations will be useful: if $\theta_0(r, R, ||N||) = \pi/2 \pm \epsilon$ implies

$$R = r\left(1 - \frac{1}{2}\left(||N||/r\right)^2 \mp \epsilon||N||/r + O\left(\epsilon^2\left(||N||/r\right)^2 + \left(||N||/r\right)^4\right)\right). \tag{9.25}$$

9.5.1 I is negligible: comparing $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} = (\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}) \cup \mathbf{I}$ with $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}$

The goal of this section is to show that $\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]})$ and $\mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}})$ are close. We write $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} = (\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}) \cup \mathbf{I}$, where $\mathbf{I}$ is the (random) set of points that enter $B_{\tilde{Z}}(r)$ when noise is added, see (9.7).

Proposition 4. *Let the usual bounds* (9.2) *hold. Conditioning on $\Omega_{s,0}$ defined in* (9.8)*, for $t > 0$, $v \geq 1$ and $n \geq t^2/\mu_X(B_z(r_=))$, on an event $\Omega_{v,t,2}$ having probability as in Table 17, we have*

$$||\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}) - \mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}})|| \leq Cv^2 \left(c_{8,\xi,s} \frac{\sigma k}{r} \vee \frac{1}{\mu_X(B_{\tilde{Z}}(r_-))n} \right) r^2 \,, \tag{9.26}$$

where $c_{8,\xi,s} := C(1 + C_\xi c_{6,\xi,s} v_{\min}^{-1})$, with $c_{6,\xi,s}$ and C_ξ defined in (9.14) *and* (3.3).

Proof. We estimate $\mathbb{E}[|\mathbf{I}|]$ relative to $\mathbb{E}[|\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}|]$ and then apply Lemma 3. The points in $B_{\tilde{Z}}(r_-)$ have a probability larger than $1/2$ of staying in $B_{\tilde{Z}}(r)$ when noise is added:

$$\mathbb{E}[|\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}|] \geq \mathbb{E}[|\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}|] \geq \frac{1}{2}\mathbb{E}[|\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}|] = \frac{1}{2}\mu_X(B_{\tilde{Z}}(r_-))n,$$

therefore it will be enough to compute the expected cardinalities of $\mathbf{I}$ relative to $\overline{n}_{r_-} = \mu_X(B_{\tilde{Z}}(r_-))n$. For $r_+ > r$ to be chosen later, we partition $\mathbf{I}$ into the sets

$$\mathbf{I}_1 = \{\tilde{X}_i : ||\tilde{X}_i - \tilde{Z}|| < r \wedge ||X_i - \tilde{Z}|| \in [r, r_+]\} \,,\ \mathbf{I}_2 = \{\tilde{X}_i : ||\tilde{X}_i - \tilde{Z}|| < r \wedge ||X_i - \tilde{Z}|| \geq r_+\} \,.$$

Since $\sigma||N|| \sim \sigma\sqrt{D}$, we expect the majority of $\mathbf{I}$ to be from $\mathbf{I}_1$.

Step 1: bounding $|\mathbf{I}_1|$. Conditioning on $\Omega_{s,0}$, and with our usual assumptions, we prove

$$(\mu_X(B_{\tilde{Z}}(r_-))n)^{-1}\, \mathbb{E}[|\mathbf{I}_1|] \leq Ce^{-C\xi^2 D}\left(e^{C\xi^2 k} - 1\right) \,. \tag{9.27}$$

For each i define the events

$$\Omega_{t_1,i} = \left\{ |\,\sigma||N_i||^2 - \sigma^2 D| \leq t_1 \sigma^2 \sqrt{D} \right\} \,,\ \Omega_{2,i} := \Omega_{t_1,i} \cap \{||X_i - \tilde{Z}|| \in [r, r_+]\} \tag{9.28}$$

Clearly $\mathbb{P}(\Omega_{t_1,i}) \geq 1 - 2e^{-c(t_1^2 \wedge t_1\sqrt{D})}$. We estimate

$$\begin{aligned}
&(\mu_X(B_{\tilde{Z}}(r_-))n)^{-1}\, \mathbb{E}[|\mathbf{I}_1|] = (\mu_X(B_{\tilde{Z}}(r_-))n)^{-1} \sum_{i=1}^n \mathbb{P}(||\tilde{X}_i - \tilde{Z}|| < r \wedge ||X_i - \tilde{Z}|| \in [r, r_+]) \\
&= (\mu_X(B_{\tilde{Z}}(r_-))n)^{-1} \left(\sum_{i=1}^n \mathbb{P}(||\tilde{X}_i - \tilde{Z}|| < r \wedge \underbrace{||X_i - \tilde{Z}|| \in [r, r_+] \wedge \Omega_{t_1,i}}_{\Omega_{2,i}}) + \right. \\
&\qquad \left. \mathbb{P}(||\tilde{X}_i - \tilde{Z}|| < r \wedge ||X_i - \tilde{Z}|| \in [r, r_+] \wedge \Omega_{t_1,i}^c) \right) \\
&\leq (\mu_X(B_{\tilde{Z}}(r_-))n)^{-1} \left(\sum_{i=1}^n \mathbb{P}(||\tilde{X}_i - \tilde{Z}|| < r \,\big|\Omega_{2,i})\mathbb{P}(\Omega_{2,i}) + \mathbb{P}(\Omega_{t_1,i}^c)\mathbb{P}(||X_i - \tilde{Z}|| \in [r, r_+)) \right) \\
&\leq \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{\mathbb{P}(\Omega_{2,i})} \int_{\Omega_{2,i}} \mathbb{P}(||\tilde{X}_i - \tilde{Z}|| < r \,|\, ||X_i - \tilde{Z}||, ||N_{X_i}||\,)\, dP + 2e^{-c(t_1^2 \wedge t_1\sqrt{D})} \right) \\
&\qquad \cdot \frac{\mu_X(B_{\tilde{Z}}(r_+) \setminus B_{\tilde{Z}}(r))}{\mu_X(B_{\tilde{Z}}(r_-))} \\
&\leq \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{\mathbb{P}(\Omega_{2,i})} \int_{\Omega_{2,i}} h(\,\theta_0(r,\, ||X_i - \tilde{Z}||, ||N_{X_i}||\,)\,)\, dP + 2e^{-c(t_1^2 \wedge t_1\sqrt{D})} \right) V_{\tilde{Z}}(r_+, r_-) \\
&\leq \left(h(\,\theta_0(\,r, r, \sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}}))) + 2e^{-C(t_1^2 \wedge t_1\sqrt{D})} \right) V_{\tilde{Z}}(r_+, r_-)\,,
\end{aligned} \tag{9.29}$$

since on $\Omega_{2,i}$, $\theta_0(\,r, ||X_i - \tilde{Z}||, ||N_{X_i}||\,) \le \theta_0(\,r, r, \sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}})\,)$. We have

$$\cos(\theta_0(r,r,\sigma\sqrt{D}(1-t_1D^{-\frac{1}{2}}))) = \left(\sigma\sqrt{D}(1-t_1D^{-\frac{1}{2}})\right)/(2r) = \xi\left(1-t_1D^{-\frac{1}{2}}\right)/2\,;$$

thus if $\theta_0 := \frac{\pi}{2} - t$, we obtain $t = \arcsin(\xi\left(1-t_1D^{-\frac{1}{2}}\right)/2) \ge \xi\left(1-t_1D^{-\frac{1}{2}}\right)/2$ and by Lemma 9

$$h(\theta_0(r,r,\sigma\sqrt{D}(1-t_1D^{-\frac{1}{2}}))) \le e^{-\frac{1}{2}t^2D} \le e^{-\frac{1}{32}\xi^2D\left(1-t_1D^{-\frac{1}{2}}\right)^2}\,.$$

We choose $t_1 = \xi\sqrt{D}$; by our usual assumptions we have

$$(r_+ - r_-)/r_- \le \xi^2/\left(1-\xi^2\right) \le 2\xi^2 \qquad , \qquad d(\tilde{Z},\mathcal{M})^2/r_-^2 \le 2\xi^2/\left(1-\xi^2\right) \le 4\xi^2\,,$$

so Lemma 5 (in particular, estimate (9.12)) implies:

$$(\mu_X(B_{\tilde{Z}}(r_-))n)^{-1}\,\mathbb{E}[|\mathbf{I}_1|] \le \left(e^{-C\xi^2D(1-\xi)^2} + 4e^{-C(\xi^2\wedge\xi)D)}\right)V_{\tilde{Z}}(r_+,r_-) \le Ce^{-C\xi^2D}\left(e^{C\xi^2k}-1\right)\,.$$

Step 2: bounding $|\mathbf{I}_2|$. Conditioning on $\Omega_{s,0}$, and with our usual assumptions, we prove that:

$$(\mu_X(B_{\tilde{Z}}(r_-))n)^{-1}\,\mathbb{E}[|\mathbf{I}_2|] \le c_{8,\xi,s}e^{-C\xi^2D}\left(e^{C\xi^2k}-1\right)\,, \tag{9.30}$$

with $c_{8,\xi,s} = C_{\xi,k}c_{6,\xi,s}v_{\min}^{-1}$. To see this, let:

$$A_l := B_{\tilde{Z}}(\sqrt{r^2+(l+1)^2\sigma^2D}) \setminus B_{\tilde{Z}}(\sqrt{r^2+l^2\sigma^2D})$$
$$p_l := \mathbb{P}\left(||X-\tilde{Z}|| \in \left(\sqrt{r^2+l^2\sigma^2D}, \sqrt{r^2+(l+1)^2\sigma^2D}\right] \wedge ||X+\sigma N-\tilde{Z}|| < r\right)$$

and observe that $\mathbb{E}[|\mathbf{I}_2|] \le \sum_{l=1}^{\infty} p_l\mu_X(A_l)n \le C_\xi\mu_{\mathbb{R}^k}(\mathbb{B}^k)r^kn$ by condition (3.3) in our usual assumptions and the bounds (9.14), provided that $p_l \le Ce^{-Cl^2}$ (observe that the condition $\xi^2k < 1/2$ required for (3.3) to hold follows from our standing assumptions on ξ). To see that this bound in fact holds, observe that for a point x to enter $B_{\tilde{Z}}(r)$, two independent conditions must be met: $\sigma||N|| \ge \frac{l}{2}\sigma\sqrt{D}$ and N must point in the right direction. The subgaussian condition on the noise gives $\mathbb{P}(\sigma||N|| \ge \frac{l}{2}\sigma\sqrt{D}) \le 2e^{-\frac{1}{4}l^2}$. To upper bound the probability that N points in the appropriate direction, fix x such that $||x-\tilde{Z}||^2 \ge r^2 + l^2\sigma^2D$; let ϕ be the angle formed by the line segment connecting x and $\tilde{Z}$ and a tangent line to $\mathbb{S}_{\tilde{Z}}^{D-1}(r)$ passing through x, so that $\sin(\phi) = r/\sqrt{r^2+l^2\sigma^2D}$. The probability that N points in the appropriate direction is upper bounded by $\mu_{\mathbb{S}^{D-1}}(V_\phi^{D-1})$. Letting $t = \frac{\pi}{2} - \phi$, we obtain $t = \arccos(r/\sqrt{r^2+l^2\sigma^2D}) \ge \pi/2 \cdot (1 - r/\sqrt{r^2+l^2\sigma^2D}) \ge 1/2 \cdot l\sigma\sqrt{D}/\sqrt{r^2+l^2\sigma^2D}$. By Lemma 9 the probability of pointing in the right direction is bounded by $e^{-C\frac{l^2\xi^2}{1+l^2\xi^2}D}$ and therefore $p_l \le e^{-l^2-C\frac{l^2\xi^2}{1+l^2\xi^2}D}$. Using our usual assumptions on ξ, we now prove (9.30):

$$\begin{aligned}(\mu_X(B_{\tilde{Z}}(r_-))n)^{-1}\,\mathbb{E}[|\mathbf{I}_2|] &\le \frac{\sum_{l=1}^{\infty} p_l\mu_X(A_l)}{\mu_X(B_{\tilde{Z}}(r_-))} \le \frac{e^{-C\frac{\xi^2}{1+\xi^2}D}}{\mu_X\left(B_{\tilde{Z}}(r_-)\right)}\sum_{l=1}^{\infty} e^{-l^2}\mu_X(A_l) \le \frac{C_\xi e^{-C\xi^2D}\mu_{\mathbb{R}^k}(\mathbb{B}^k)r^k}{\mu_X\left(B_{\tilde{Z}}(r_-)\right)} \\ &\le C_\xi c_{6,\xi,s}v_{\min}^{-1}\left(\frac{1}{1-\xi^2}\right)^{\frac{k}{2}}e^{-C\xi^2D} \le c_{8,\xi,s}e^{-C\xi^2D}\left(e^{\frac{C\xi^2k}{1-\xi^2}}-1\right) \\ &\le c_{8,\xi,s}e^{-C\xi^2D}\left(e^{C\xi^2k}-1\right)\,.\end{aligned}$$

To complete the proof, an application of Lemma 3 yields the estimate

$$||\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}) - \mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}})|| \le Cv^2\left(c_{8,\xi,s}e^{-C\xi^2D}\left(e^{C\xi^2k}-1\right) \vee \frac{1}{\mu_X(B_{\tilde{Z}}(r_-))n}\right)r^2\,.$$

The desired estimate (9.26) is obtained by observing that on the one hand, if $\xi \le 1/(3\sqrt{k})$ then $e^{-C\xi^2D}\left(e^{C\xi^2k}-1\right) \le Ce^{-C\xi^2D}\xi^2ke^{C/9} \le C\xi kD^{-\frac{1}{2}}\left(\xi\sqrt{D}\right)e^{-C\left(\xi\sqrt{D}\right)^2} \le C\xi kD^{-\frac{1}{2}}$, and when $\xi \ge \frac{1}{3\sqrt{k}}$, $e^{-C\xi^2D}\left(e^{C\xi^2k}-1\right) \le e^{-C\xi^2(D-k)} \le Ce^{-CD/k} \le C\xi kD^{-1/2}$.

□

9.5.2 Comparing $\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}} = (\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \cup \mathbf{Q}_2) \setminus \mathbf{Q}_1$ with $\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}}$

Proposition 5. *Let our usual assumptions hold and, furthermore, let $D \geq C$. Conditioning on $\Omega_{s,0}$, for $t > 0, v \geq 1$ and $n \geq t^2/\mu(B_z(r_=))$, on an event $\Omega_{v,t,3}$ having probability as in Table 17, we have*

$$||\text{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]} \cap \widetilde{\mathbf{X}_n^{[\tilde{Z},r]}}) - \text{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}})|| \leq Cv^2 \left(\beta \vee \frac{1}{\mu_X(B_{\tilde{Z}}(r_-))n}\right) r^2 \,. \tag{9.31}$$

where

$$\beta := \frac{\sigma k}{r} \left(1 \vee \frac{\sigma^2 D}{r^2/k}\right) \sqrt{\log \frac{r}{3\sigma k}} \,.$$

Proof. Recall the definitions (9.7):

$$\mathbf{Q}_1 = \{\tilde{X}_i \in B_{\tilde{Z}}(r) : ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, r_-)\} \;\; , \; \mathbf{Q}_2 = \{\tilde{X}_i \in B_{\tilde{Z}}(r) : ||X_i - \tilde{Z}|| \in [r_-, r]\}$$

The bound (9.31) is proved by combining the bounds (9.32) and (9.35) below for $\mathbb{E}[|\mathbf{Q}_1|]$ and $\mathbb{E}[|\mathbf{Q}_2|]$ respectively, followed by an application of Lemma 3.

Bounding $|\mathbf{Q}_1|$. We will prove that, as soon as $D \geq C$

$$(\mu_X(B_{\tilde{Z}}(r_-))n)^{-1} \, \mathbb{E}[|\mathbf{Q}_1|] \leq C \frac{\sigma k}{r} \sqrt{\log \frac{r}{3\sigma k}} \,. \tag{9.32}$$

To see this, for any $\tilde{r}_- \in [\sigma\sqrt{d}, r_-)$ (to be chosen later), we have

$$\begin{aligned}
\mathbb{E}[|\mathbf{Q}_1|] &= \sum_{i=1}^n \mathbb{P}(\, ||\tilde{X}_i - \tilde{Z}|| > r \wedge ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, r_-)\,) \\
&= \sum_{i=1}^n \mathbb{P}(\, ||\tilde{X}_i - \tilde{Z}|| > r \,\big|\, ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, \tilde{r}_-\,) \cdot \mathbb{P}(\, ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, \tilde{r}_-\,) &(9.33)\\
&\quad + \sum_{i=1}^n \mathbb{P}(\, ||\tilde{X}_i - \tilde{Z}|| > r \,\big|\, ||X_i - \tilde{Z}|| \in [\tilde{r}_-, r_-)\,) \cdot \mathbb{P}(\, ||X_i - \tilde{Z}|| \in [\tilde{r}_-, r_-)\,)\,. &(9.34)
\end{aligned}$$

By (9.23), if X_i is at distance R from $\tilde{Z}$, the probability that $X_i + N_i \notin B_{\tilde{Z}}(r)$ is given by $h(\pi - \theta_0(r, R, ||N_i||))$. Note that $\pi - \theta_0(r, R, ||N||)$ is increasing in both R and $||N||$. By an identical argument to that in 9.5.1 (and (9.29) in particular), with $\Omega_{t_1,i}$ as in (9.28), we obtain the following bound on (9.33):

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{P}(\, ||\tilde{X}_i - \tilde{Z}|| > r \mid ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, \tilde{r}_-)\,) \cdot \mathbb{P}(\, ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, \tilde{r}_-\,)) \\
&= \sum_{i=1}^n \Bigg(\mathbb{P}(\, ||\tilde{X}_i - \tilde{Z}|| > r \mid ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, \tilde{r}_-), \Omega_{t_1,i}\,) \cdot \mathbb{P}(\Omega_{t_1,i}) \\
&\quad + \sum_{i=1}^n \mathbb{P}(\, ||\tilde{X}_i - \tilde{Z}|| > r \mid ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, \tilde{r}_-), \Omega_{t_1,i}^c\,) \cdot \mathbb{P}(\Omega_{t_1,i}^c)\Bigg) \mathbb{P}(\, ||X_i - \tilde{Z}|| \in [\sigma\sqrt{d}, \tilde{r}_-\,) \\
&\leq \left(h(\pi - \theta_0(r, \tilde{r}_-, \sigma\sqrt{D}(1 + t_1 D^{-\frac{1}{2}}))) + 2e^{-C(t_1^2 \wedge t_1\sqrt{D})}\right) \mu_X(B_{\tilde{Z}}(\tilde{r}_-) \setminus B_{\tilde{Z}}(\sigma\sqrt{d}))n \\
&\leq \left(e^{-\frac{1}{2}\epsilon^2 D} + 2e^{-Ct_1^2}\right) \mu_X(B_{\tilde{Z}}(\tilde{r}_-))n \leq 4e^{-Ct_1^2} \mu_X(B_{\tilde{Z}}(r_-))n
\end{aligned}$$

where the bound in the line before the last follows by applying Lemma 9 after choosing $\tilde{r}_-$ so that $\pi-\theta_0(r,\tilde{r}_-,\sigma\sqrt{D}(1+t_1D^{-\frac{1}{2}}))=\frac{\pi}{2}-\epsilon$, i.e., by (9.25),

$$\tilde{r}_-=r\left(\left(1-\frac{1}{2}\xi^2\left(1+t_1D^{-\frac{1}{2}}\right)^2\right)-\epsilon\sigma\sqrt{D}\left(1+t_1D^{-\frac{1}{2}}\right)+O\left(\left(\epsilon^2\xi^2+\xi^4\right)(1+t_1D^{-\frac{1}{2}})^2\right)\right),$$

and the last bound is a consequence of imposing $t_1\leq\sqrt{D}$ and choosing $\epsilon=Ct_1/\sqrt{D}$ to balance the two exponential terms.

In order to bound (9.34) we will need the following estimates, which hold as soon as our usual assumptions on ξ are satisfied, $t_1\leq\sqrt{D}$ as above: first of all $\frac{d(\tilde{Z},\mathcal{M})^2}{r_-^2}\leq\frac{2\xi^2}{1-\xi^2}$, and moreover

$$\begin{aligned}\frac{r_--\tilde{r}_-}{\tilde{r}_-}&\leq\frac{r\sqrt{1-\xi^2}-r(1-\frac{\xi^2}{2}(1+t_1D^{-\frac{1}{2}})^2)+\epsilon\sigma\sqrt{D}(1+t_1D^{-\frac{1}{2}})+rO(\xi^4+\epsilon^2\xi^2)}{1-r(1-\frac{\xi^2}{2}(1+t_1D^{-\frac{1}{2}})^2)+\epsilon\sigma\sqrt{D}(1+t_1D^{-\frac{1}{2}})+rO(\xi^4+\epsilon^2\xi^2)}\\&\leq\frac{2\xi^2t_1D^{-\frac{1}{2}}+2\epsilon\xi+O(\xi^4+\epsilon^2\xi^2)}{1-2\xi^2-2\epsilon\xi+O(\xi^4+\epsilon^2\xi^2)}\leq\frac{C\xi t_1}{\sqrt{D}}\end{aligned}$$

as soon as $t_1<\sqrt{D}$ and choosing ϵ as above. By Lemma 5

$$\begin{aligned}&\sum_{i=1}^n\mathbb{P}(\,||\tilde{X}_i-\tilde{Z}||>r\,\big|\,||X_i-\tilde{Z}||\in[\tilde{r}_-,r_-)\,)\cdot\mathbb{P}(\,||X_i-\tilde{Z}||\in[\tilde{r}_-,r_-)\,)\leq\mu_X\left(B_{\tilde{Z}}(r_-)\setminus B_{\tilde{Z}}(r_{\sigma_-})\right)n\\&\quad\leq V_{\tilde{Z}}(r_-,\tilde{r}_-)\mu_X(B_{\tilde{Z}}(r_-))n\leq\left(e^{Ckt_1\xi D^{-\frac{1}{2}}\left(1+\left(1+Ct_1\xi D^{-\frac{1}{2}}\right)\left(1-\frac{2\xi^2}{1-\xi^2}\right)^{-1}\right)}-1\right)\mu_X(B_{\tilde{Z}}(r_-))n\\&\quad\leq\left(e^{Ct_1\xi kD^{-\frac{1}{2}}}-1\right)\mu_X(B_{\tilde{Z}}(r_-))n\leq C\frac{t_1\xi k}{\sqrt{D}}\mu_X(B_{\tilde{Z}}(r_-))n\,.\end{aligned}$$

as soon as $t_1<C\sqrt{D}/(\xi k)$. Combining our bounds for (9.33) and (9.34), we obtain:

$$\frac{\mathbb{E}[|\mathbf{Q}_1|]}{\mu_X(B_{\tilde{Z}}(r_-))n}\leq4e^{-Ct_1^2}+Ct_1\frac{\xi k}{\sqrt{D}}\leq C\frac{\xi k}{\sqrt{D}}\sqrt{\log\frac{\sqrt{D}}{3\xi k}}\,,$$

by choosing $t_1^2=1\vee\log\frac{\sqrt{D}}{3\xi k}$, proving (9.32). Note that the conditions we imposed above on t_1 are satisfied under our usual assumptions and $D\geq C$. By Lemma 3 we have, with probability as in Table 17, the bound (recall that $r^2+\sigma^2D)\leq(1+\xi^2)r^2\leq\frac{4}{3}r^2)$

$$||\widetilde{\mathrm{cov}(\mathbf{X}_n^{[\tilde{Z},r_-]}}\cup\mathbf{Q}_2)-\widetilde{\mathrm{cov}(\mathbf{X}_n^{[\tilde{Z},r_-]}}\cup\mathbf{Q}_2\setminus\mathbf{Q}_1)||\leq Cv^2\left(\frac{\xi k}{\sqrt{D}}\sqrt{\log\frac{\sqrt{D}}{3\xi k}}\vee\frac{1}{\mu_X(B_{\tilde{Z}}(r_-))n}\right)r^2\,.$$

Bounding $|\mathbf{Q}_2|$. We now estimate $|\mathbf{Q}_2|$, and prove that as soon as $D\geq C$

$$(\mu_X(B_{\tilde{Z}}(r_-))n)^{-1}\,\mathbb{E}[|\mathbf{Q}_2|]\leq C\frac{\xi k}{\sqrt{D}}(1\vee\xi^2k)\sqrt{\log\frac{\sqrt{D}}{3\xi k}}\,.\tag{9.35}$$

By an argument similar to that in 9.5.1, we choose $r_{\sigma_+}\in[r_-,r]$ so that $\theta_0(r,r_{\sigma_+},\sigma\sqrt{D}(1-t_1D^{-\frac{1}{2}}))=\frac{\pi}{2}-\epsilon$, i.e., by (9.25),

$$r_{\sigma_+}=r\left(\left(1-\frac{1}{2}\xi^2\left(1-t_1D^{-\frac{1}{2}}\right)^2\right)+\epsilon\sigma\sqrt{D}\left(1-t_1D^{-\frac{1}{2}}\right)+O\left(\left(\epsilon^2\xi^2+\xi^4\right)(1-t_1D^{-\frac{1}{2}})^2\right)\right),$$

which implies $\frac{r_{\sigma_+}-r_-}{r_-} \leq \frac{C\xi t_1}{\sqrt{D}}$. For this choice of r_{σ_+}, we have the bounds $\frac{\sigma^2 D}{r_{\sigma_+}^2} \leq \frac{\sigma^2 D}{r_-^2} \leq \frac{\xi^2}{1-\xi^2} \leq 2\xi^2$, and, by Lemma 5, $\mu(B_{\tilde{Z}}(r_{\sigma_+}))/\mu(B_{\tilde{Z}}(r_-)) \leq (1+\epsilon\xi(1+t_1 D^{-\frac{1}{2}}))^{2k}(1-2\xi^2)^k$ and

$$\frac{\mu(B_{\tilde{Z}}(r))}{\mu(B_{\tilde{Z}}(r_{\sigma_+}))} \leq \frac{\mu(B_{\tilde{Z}}(r))}{\mu(B_{\tilde{Z}}(r_-))} \leq \left(\frac{1}{1-\xi^2}\right)^k \left(\frac{1-\xi^2(1+s^2D^{-\frac{1}{2}})}{1-\xi^2(1+s^2D^{-\frac{1}{2}})/\left(1-\xi^2\right)}\right)^k \leq e^{C\xi^2 k}\,.$$

Then

$$\begin{aligned}
\frac{\mathbb{E}[|\mathbf{Q}_2|]}{\mu_X(B_{\tilde{Z}}(r_-))n} &= \frac{1}{\mu_X(B_{\tilde{Z}}(r_-))n}\sum_{i=1}^n \mathbb{P}(\,||\tilde{X}_i-\tilde{Z}||<r \wedge ||X_i-\tilde{Z}|| \in [r_-,r]\,) \\
&= \frac{1}{\mu_X(B_{\tilde{Z}}(r_-))n}\sum_{i=1}^n \mathbb{P}(\,||\tilde{X}_i-\tilde{Z}||<r \,\big|\, ||X_i-\tilde{Z}|| \in [r_-,r_{\sigma_+}]\,)\cdot \mathbb{P}(\,||X_i-\tilde{Z}|| \in [r_-,r_{\sigma_+}]\,) \\
&\quad + \frac{1}{\mu_X(B_{\tilde{Z}}(r_-))n}\sum_{i=1}^n \mathbb{P}(\,||\tilde{X}_i-\tilde{Z}||<r \,\big|\, ||X_i-\tilde{Z}|| \in [r_{\sigma_+},r]\,)\cdot \mathbb{P}(\,||X_i-\tilde{Z}|| \in [r_{\sigma_+},r]\,) \\
&\leq V_{\tilde{Z}}(r_{\sigma_+},r_-) + \left(h(\theta_0(r,r_{\sigma_+},\sigma\sqrt{D}(1-t_1D^{-\frac{1}{2}}))) + 2e^{-C(t_1^2\wedge t_1\sqrt{D})}\right)\frac{\mu(B_{\tilde{Z}}(r))}{\mu(B_{\tilde{Z}}(r_-))} \\
&\leq \left(e^{C\frac{\xi k}{\sqrt{D}}t_1}-1\right) + \left(e^{-\frac{1}{2}\epsilon^2 D}+2e^{-Ct_1^2}\right)e^{C\xi^2 k} \\
&\leq_{t_1 \leq \frac{\sqrt{D}}{k\xi},\epsilon=\frac{2t_1}{\sqrt{D}}} C\frac{\xi k}{\sqrt{D}}t_1 + e^{-C(t_1^2-\xi^2 k)} \\
&\leq C\frac{\xi k}{\sqrt{D}}(1\vee\xi^2 k)\sqrt{\log\frac{\sqrt{D}}{3\xi k}}\,,
\end{aligned}$$

where we chose $\epsilon = 2t_1 D^{-\frac{1}{2}}$, $t_1^2 = \xi^2 k + \frac{1}{C}\log\frac{\sqrt{D}}{3\xi k}$. Lemma 3 implies that with probability as in Table 17

$$||\mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}} \cup \mathbf{Q}_2) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[\tilde{Z},r_-]}})|| \leq Cv^2\left(\frac{\xi k(1\vee\xi^2 k)}{\sqrt{D}}\sqrt{\log\frac{\sqrt{D}}{3\xi k}} \vee \frac{1}{\mu_X(B_{\tilde{Z}}(r_-))n}\right)r^2\,.$$

□

9.6 Putting it all together

We finally recall, and prove, the following Proposition 2:

Proposition 6. *Let $D \geq C$,*

$$r \in \left(R_{\min}+4\sigma\sqrt{D}+\frac{1}{6\kappa}, R_{\max}-\sigma\sqrt{D}-\frac{1}{6\kappa}\right) \cap \left(3\sigma\left(\sqrt{D}\vee k\right), \frac{\sqrt{k}}{\kappa}\right) \tag{9.36}$$

where C is a universal constant and σ is small enough so that the interval for r is not empty. Then, for $t, v \geq 1$, $n \geq t^2/\mu(B_z(r_=))$, $s^2 < \frac{r^2/k}{12\sigma^2 D}\sqrt{D}$

$$||\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})|| \leq Cv^2\left(\beta_s \vee \frac{1}{\mu_X(B_z(r_=))n}\right)r^2 \tag{9.37}$$

holds with

$$\beta_s := \left(1+\frac{s^2\sigma\sqrt{D}}{r}+\left(1\vee\frac{\sigma^2 D}{r^2/k}\right)\sqrt{\log\frac{r}{3\sigma k}}\right)\frac{\sigma k}{r}$$

and with probability at least $1-Ce^{-C(v^2 n\mu(B_{\tilde{Z}}(r_=))\wedge s^4\wedge t^2)}$.

Proof. This follows from combining the perturbations in Propositions 3,4 and 5, whose bounds are summarized in Table 17: we have that for $t, v \geq 1$, $1 \leq s^2 \leq \sqrt{D}$, $n \geq t^2/\mu_X(B_z(r_=))$, and conditioned on $\Omega_{s,0}$:

$$\begin{aligned}
&||\mathrm{cov}(\tilde{\mathbf{X}}_n^{[\tilde{Z},r]}) - \mathrm{cov}(\widetilde{\mathbf{X}_n^{[z,r_=]}})|| \\
&\leq Cv^2 \left(\left(c_{4,\xi,s,t_0} + c_{8,\xi,s} + \left(1 \vee \frac{\sigma^2 D}{r^2/k} \right) \sqrt{\log \frac{r}{3\sigma k}} \right) \frac{\sigma k}{r} \vee \frac{1}{\mu_X(B_z(r_=))n} \right) r^2 \\
&\leq Cv^2 \left(\underbrace{\left(\frac{s^2 \sigma \sqrt{D}}{r} + 1 + \left(1 \vee \frac{\sigma^2 D}{r^2/k} \right) \sqrt{\log \frac{r}{3\sigma k}} \right) \frac{\sigma k}{r}}_{\beta_s} \vee \frac{1}{\mu_X(B_z(r_=))n} \right) r^2 ,
\end{aligned} \tag{9.38}$$

by recalling that $t_0^2 = \log \frac{r}{3\sigma k}$ (as in Proposition 3), so that

$$c_{4,\xi,s,t_0} \leq C \left(\frac{s^2 \sigma \sqrt{D}}{r} + \sqrt{\log \frac{r}{3\sigma k}} \right) ,$$

and noting that $c_{8,\xi,s} \leq C$ since $\sigma\sqrt{D}/r \leq 1/3$. The bound (9.38) holds with probability at least $1 - Ce^{-Cv^2 \beta_s \mathbb{E}[n_{r_=}]} - Ce^{-Ct^2}$, conditioned on $\Omega_{s,0}$. Removing the conditioning on $\Omega_{s,0}$, whose complement had probability at most $2e^{-Cs^4}$, we obtain that (9.37) holds with probability at least $1 - Ce^{-C(v^2 n \mu(B_{\tilde{Z}}(r_=)) \wedge s^4 \wedge t^2)}$.

Finally, we determine the restrictions on $r_=$ in terms of $R_{\min}, R_{\max}$, the parameters that determined the range where our volume growth and covariance estimation assumptions hold. In Sections 9.4-9.5 we have assumed that all the radii involved lied in $[R_{\min}, R_{\max}]$, so we need to impose $r_= \pm (2\sigma^2 D + q(r)) \in [R_{\min}, R_{\max}]$, which is implied, upon noting that $q \leq \frac{1}{6}\frac{r^2}{k}$, by the restriction in (9.36). □

10 Appendix: Results from linear algebra and perturbation theory

Lemma 10 (Wielandt's inequality [89]). *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix in the form*

$$A = \begin{pmatrix} B & C \\ C^T & D \end{pmatrix}$$

with $B \in \mathbb{R}^{r \times r}$, $C \in \mathbb{R}^{r \times s}$ and $D \in \mathbb{R}^{s \times s}$, with $n = r + s$. Let $\lambda_i(E)$ denote the i-th largest eigenvalue of a matrix E. If $\lambda_r(B) > \lambda_1(D)$, then

$$\begin{aligned}
0 \leq \lambda_i(A) - \lambda_i(B) &\leq \frac{\lambda_1(C^T C)}{\lambda_i(B) - \lambda_1(D)} \wedge ||C|| \qquad \text{for } 1 \leq i \leq r \,, \\
0 \leq \lambda_j(D) - \lambda_{r+j}(A) &\leq \frac{\lambda_1(C^T C)}{\lambda_r(B) - \lambda_j(D)} \wedge ||C|| \quad \text{for } 1 \leq j \leq s \,.
\end{aligned}$$

The statement in [89] is only for positive definite matrices, but the result for general symmetric matrices follows easily by adding a multiple of the identity matrix that is large enough to make the matrices positive definite.

11 Random vectors, random matrices and covariances

We briefly recall some notations. We define the covariance and the empirical covariance of a random variable Y as

$$\begin{aligned}
\mathrm{cov}(Y) &= \mathbb{E}[(Y - \mathbb{E}[Y]) \otimes (Y - \mathbb{E}[Y])] \\
\mathrm{cov}(\mathbf{Y}_n) &= \frac{1}{n}\sum_{i=1}^{n} (Y_i - \mathbb{E}_n[Y]) \otimes (Y_i - \mathbb{E}_n[Y]) \quad , \mathbb{E}_n[Y] \quad = \frac{1}{n}\sum_{i=1}^{n} Y_i
\end{aligned} \tag{11.1}$$

where $\overline{Y} := Y - \mathbb{E}[Y]$, and $Y_1, \dots, Y_n$ are i.i.d. copies of Y. Moreover, $\mathrm{cov}(Y, X) = \mathbb{E}[(Y - \mathbb{E}[Y]) \otimes (X - \mathbb{E}[X])]$ is the cross-covariance between two random variables Y, X, and $\mathrm{cov}(\mathbf{Y}_n, \mathbf{X}_n)$ its empirical counterpart. Note that, we often view a sample $\mathbf{X}_n$ as a matrix, so that for example we can write $\mathrm{cov}(\mathbf{Y}_n, \mathbf{X}_n) = \frac{1}{n}\overline{\mathbf{Y}}_n^T \overline{\mathbf{X}}_n$, where $\overline{\mathbf{Y}}_n$ denotes a sample centered with respect to its empirical mean.

We are interested in the concentration properties of the empirical covariance and cross-covariance operators under different assumptions on Y and X. In particular, we are interested in the case when X, Y are bounded or subgaussian. We note the following elementary identity:

$$\mathrm{cov}(\mathbf{Y}_n, \mathbf{X}_n) = \mathbb{E}_n[(X - \mathbb{E}[X]) \otimes (Y - \mathbb{E}[Y])] - (\mathbb{E}[X] - \mathbb{E}_n[X]) \otimes (\mathbb{E}[Y] - \mathbb{E}_n[Y]) \,. \tag{11.2}$$

As usual, in all that follows C, c will denote a universal constant whose value may change from line to line. For bounded random vectors we have the following known results:

Proposition 7. *Let Z be a random variable in $\mathbb{R}^d$ with $\mathbb{E}[Z] = 0$ and $||Z|| \leq \sqrt{M}$ a.s. Let $Z_1, \dots, Z_n$ be i.i.d. copies of Z.*

(i) for any $t > 0$ we have

$$\mathbb{P}\left(||\mathbb{E}_n[Z]|| > \sqrt{\frac{M}{n}t}\right) \leq 2e^{-ct^2} \,. \tag{11.3}$$

(ii) for any $t > 0$ and $n \geq C\frac{t^2 M \log(d \wedge n)}{||\mathrm{cov}(Z)||}$,

$$\mathbb{P}\left(||\mathrm{cov}(\mathbf{Z}_n) - \mathrm{cov}(Z)|| > ||\mathrm{cov}(Z)||\sqrt{\frac{M \log(d \wedge n)}{||\mathrm{cov}(Z)||n}}t + \frac{M}{n}t^2\right) \leq 4e^{-ct^2} \,. \tag{11.4}$$

Proof. (i) follows from [90, 91]. (ii) follows from Corollary 5.52 in [38], together with (11.2) and (i). □

We remark, as it is done in [38] after Corollary 5.52, that the crucial quantity in determining the sampling requirement in (ii) above is not the ratio $M/||\mathrm{cov}(Z)||$ but the effective rank $\mathrm{tr}(\mathrm{cov}(Z))/||\mathrm{cov}(Z)||$. We use this observation for example to obtain the bounds in (8.4).

Definition 2. *A real-valued random variable Z is called strictly subgaussian if for all $t > 0$*

$$\mathbb{E}\left[e^{tZ}\right] \leq e^{\frac{\mathbb{E}[Z^2]t^2}{2}} \,.$$

We will write $Z \sim \mathrm{SSub}(\sigma^2)$, where $\sigma^2 = \mathbb{E}[Z^2]$.

We summarize in the following Proposition some well-known properties of strictly subgaussian random variables:

Proposition 8. *Let $Z^1, \dots, Z^d \in \mathbb{R}$ be i.i.d., $Z^i \sim \mathrm{SSub}(1)$, and $Z = (Z^1, \dots, Z^d) \in \mathbb{R}^d$.*
Then

(i) $\mathbb{E}[Z_i] = 0$ and for every $t > 0$, $\mathbb{P}(|Z_i| > t) \leq 2e^{-t^2/2}$.

(ii) For every $v \in \mathbb{R}^d$ we have $\langle Z, v\rangle \sim \mathrm{SSub}(||v||_2^2)$.

(iii) $\mathbb{E}[||Z||^2] = d$, and there exists a universal constant c such that for all $t > 0$

$$\mathbb{P}\left(\left|||Z||^2 - d\right| > t\sqrt{d}\right) \leq 2e^{-ct^2} \quad , \quad \mathbb{P}\left(||Z|| > \sqrt{d} + \sqrt{t}\sqrt[4]{d}\right) \leq 2e^{-ct^2} \,.$$

(iv) If $Z_1, \dots, Z_n$ are i.i.d. copies of Z, then

$$\mathbb{P}\left(||\mathbb{E}_n[Z]||^2 > \frac{d}{n} + \frac{\sqrt{d}}{n}t\right) \leq 2e^{-ct^2} \quad , \quad \mathbb{P}\left(||\mathbb{E}_n[Z]|| > \sqrt{\frac{d}{n}}\sqrt{1 + \frac{t}{\sqrt{d}}}\right) \leq 2e^{-ct^2}$$

(v) If $Z_1, \dots, Z_n$ are i.i.d. copies of Z, then with probability at least $1 - 2e^{-ct^2}$

$$(\sqrt{d} - C\sqrt{n} - t) \vee (\sqrt{n} - C\sqrt{d} - t) \leq \sigma_{\min}([Z_1|\dots|Z_n]) \leq$$
$$\sigma_{\max}([Z_1|\dots|Z_n]) \leq (\sqrt{d} + C\sqrt{n} + t) \wedge (\sqrt{n} + C\sqrt{d} + t)\,.$$

(vi) Let $Z_1, \dots, Z_n$ be i.i.d. copies of Z. Then for $t \geq C$, $n \geq Ct^2 d$, we have

$$\mathbb{P}\left(\|\mathrm{cov}(\mathbf{Z}_n) - I_d\| > \sqrt{\frac{d}{n}t} \right) \leq 4e^{-ct^2}$$

and for $n \leq Ct^2 d$ we have

$$\mathbb{P}\left(\|\mathrm{cov}(\mathbf{Z}_n)\| > \frac{d}{n}\left(1 + C\sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}}\right)^2 \right) \leq 4e^{-ct^2}\,.$$

Proof. These results are combinations of standard facts [92], so we only sketch the proofs. (i) and (ii) are proved by using the definition of strictly subgaussian and using the moment generating function. For (iii), the computation of $\mathbb{E}[||Z||^2]$ is trivial, and to prove the concentration inequality one can either use the moment generating function again, or (ii) together with an ϵ-net argument for discretizing $\mathbb{S}^{d-1}$ and a union bound. In order to prove (iv) we simply use (ii) and (iii). In order to prove (v) one uses standard ϵ-net arguments to discretize the unit sphere, together with (iii) and a union bound. Finally, (vi) follows from $\mathbb{E}[Z] = 0$, $n \geq Ct^2 d$, so that by a Corollary to Theorem 39 in [38], with probability at least $1 - 4e^{-ct^2}$

$$\begin{aligned} ||\mathrm{cov}(\mathbf{Z}_n) - \sigma^2 I_d|| &\leq \left\| \frac{1}{n}\sum_{l=1}^{n} Z_l \otimes Z_l - \mathbb{E}[Z \otimes Z] \right\| + ||\mathbb{E}_n[Z]||^2 \\ &\leq C\sqrt{\frac{d}{n}}t + \frac{d}{n} + \frac{\sqrt{d}}{n}t \leq C\sqrt{\frac{d}{n}}t\,. \end{aligned}$$

For $n \leq Ct^2 d$, since $||\mathrm{cov}(\mathbf{Z}_n)|| \leq ||1/n\sum_{i=1}^{n} Z_i \otimes Z_i||$ (since the centering by the empirical mean decreases the norm of the matrix) we have, by (v), with probability at least $1 - 4e^{-ct^2}$

$$||\mathrm{cov}(\mathbf{Z}_n)|| \leq\leq \left(\sqrt{\frac{d}{n}} + C + \frac{t}{\sqrt{n}}\right)^2 \leq \frac{d}{n}\left(1 + C\sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}}\right)^2$$

□

The following result is useful in controlling the norm of cross covariance operators when the range of X, Y is of dimension k, d respectively. These types of bounds are quite well-known, and we report here the version we need for the reader's convenience; the techniques are also used in results that follow.

Proposition 9 (Norm of product of random matrices)**.** *Let $\mathbf{N}_1 \in \mathbb{R}^{n\times k}$, $\mathbf{N}_2 \in \mathbb{R}^{n\times d}$ have i.i.d. subgaussian entries with mean 0 and subgaussian moment 1. Then for c, C universal constants,*

$$\mathbb{P}\left(\frac{1}{n}||\mathbf{N}_1^T\mathbf{N}_2|| > \sqrt{\frac{k+d}{n}}t\right) \leq \begin{cases} ce^{-c(k+d)t^2} & ,t \in C\cdot\left(1, \sqrt{\frac{n}{k+d}}\right) \\ ce^{-c\sqrt{n(k+d)}t} & ,t \geq C\max\left\{\sqrt{\frac{n}{k+d}}, \sqrt{\frac{k+d}{n}}\right\} \end{cases}. \tag{11.5}$$

and otherwise

$$\mathbb{P}\left(\frac{1}{n}||\mathbf{N}_1^T\mathbf{N}_2|| > \left(1 + \sqrt{\frac{d}{n}}\right)\left(1 + \sqrt{\frac{k}{n}}\right)t\right) \leq 4e^{-c(n+d\wedge k)t}\,.. \tag{11.6}$$

This result implies that $||\mathbf{N}_1^T\mathbf{N}_2||$, as a function of the inner dimension n, grows only like $\sqrt{n}$, thanks to the cancellations due to the independence of $\mathbf{N}_1$ and $\mathbf{N}_2$. It is easy to see that a similar estimate would hold for $\mathrm{cov}(\mathbf{N}_1, \mathbf{N}_2)$ as soon as $n \geq Ck$, by using (11.2) and (iv) in Proposition 8 in the cases covered by (11.5), and by observing that $||\overline{\mathbf{N}}_{\{1,2\}}|| \leq ||\mathbf{N}_{\{1,2\}}||$ in the case covered by (11.6).

Proof of Proposition 9. The format of the argument is standard: we discretize the unit sphere in the domain and range to finite nets, and estimate the size of the operator by restricting to the net and taking union bounds over the high probability events on which we can control the norm of the operator applied to each vector in the net. Let $\mathcal{N}^d$ be an ϵ_1-net for $\mathbb{S}^{d-1}$ and let $\mathcal{N}^k$ be an ϵ_2-net for $\mathbb{S}^{k-1}$. Observe that by a standard volume estimate we can choose the nets so that $|\mathcal{N}^d| \leq (1+2/\epsilon_1)^d$ and $|\mathcal{N}^k| \leq (1+2/\epsilon_2)^k$. Then

$$||\mathbf{N}_1^T\mathbf{N}_2|| = \max_{x\in\mathbb{S}^{d-1},y\in\mathbb{S}^{k-1}} \langle \mathbf{N}_1^T\mathbf{N}_2 x, y\rangle \leq (1-\epsilon_1)^{-1}(1-\epsilon_2)^{-1} \max_{x\in\mathcal{N}^d,y\in\mathcal{N}^k} \langle N_1^T N_2 x, y\rangle\,.$$

Therefore:

$$\begin{aligned}\mathbb{P}(||\mathbf{N}_1^T\mathbf{N}_2|| > t) &\leq \mathbb{P}(\max_{x\in\mathcal{N}^d,y\in\mathcal{N}^k} |\langle \mathbf{N}_1^T\mathbf{N}_2 x, y\rangle| > t(1-\epsilon_1)(1-\epsilon_2))\\ &\leq \sum_{x\in\mathcal{N}^d,y\in\mathcal{N}^k} \mathbb{P}(|\langle \mathbf{N}_1^T\mathbf{N}_2 x, y\rangle| > t(1-\epsilon_1)(1-\epsilon_2))\\ &\leq |\mathcal{N}^d|\,|\mathcal{N}^k| \mathbb{P}(|\langle \mathbf{N}_2 x, \mathbf{N}_1 y\rangle| > t(1-\epsilon_1)(1-\epsilon_2))\\ &\leq 5^{k+d}\,\mathbb{P}(|\langle \mathbf{N}_2 x, \mathbf{N}_1 y\rangle| > t/4)\,.\end{aligned}$$

by choosing $\epsilon_1 = \epsilon_2 = 1/2$. Since the entries of N_1, N_2 are i.i.d subgaussian, and $||x||_2 = ||y||_2 = 1$, $\mathbf{N}_2 x$ has i.i.d. subgaussian entries and so does $\mathbf{N}_1 y$, with the same subgaussian moments as the entries of $\mathbf{N}_2$ and $\mathbf{N}_1$ respectively. Moreover $\mathbf{N}_2 x$ and $\mathbf{N}_1 y$ are independent, so $\langle \mathbf{N}_2 x, \mathbf{N}_1 y\rangle$ is the sum of n independent subexponential random variables, and therefore (e.g. Cor. 17 in [38])

$$\mathbb{P}(||\mathbf{N}_1^T\mathbf{N}_2|| > t) \leq ce^{c_1(k+d)-c_2\min\{\frac{(t/4)^2}{n},t/4\}}\,. \tag{11.7}$$

If $t \leq 4n$, the last upper bound is nontrivial for, say, $c_1(k+d) < \frac{c_2}{2}\frac{t^2}{16n}$. Substituting t by $t\sqrt{n(k+d)}$, we obtain

$$\mathbb{P}\left(||\mathbf{N}_1^T\mathbf{N}_2|| > \sqrt{n(k+d)}t\right) \leq ce^{-c(k+d)t^2} \qquad ,t \in C\cdot\left(1,\sqrt{\frac{n}{k+d}}\right)\,.$$

On the other hand, if $t \geq 4n$, the upper bound in (11.7) is nontrivial for, say, $c_1(k+d) < \frac{c_2 t}{8}$, and letting substituting t with $t\sqrt{n(k+d)}$, we obtain

$$\mathbb{P}\left(||\mathbf{N}_1^T\mathbf{N}_2|| > \sqrt{n(k+d)}t\right) \leq ce^{-c\sqrt{n(k+d)}t} \qquad ,t \geq C\cdot\max\left\{\sqrt{\frac{n}{k+d}},\sqrt{\frac{k+d}{n}}\right\}\,.$$

The second inequality follows from the trivial bound $||\mathbf{N}_1^T\mathbf{N}_2|| \leq ||\mathbf{N}_1||||\mathbf{N}_2||$ and bound (11.9) in the next Proposition. □

Proposition 10. *Let $\mathbf{B} \in \mathbb{R}^{k\times n}$ and $\mathbf{A} \in \mathbb{R}^{n\times d}$, with $\mathbf{A}$ and $\mathbf{B}$ independent random matrices. Also suppose that $\mathbf{A}$ has i.i.d. subgaussian entries with subgaussian moment 1. Then for $t \geq C$*

$$\mathbb{P}\left(||\mathbf{BA}|| > ||\mathbf{B}||\sqrt{d+k}\,t\right) \leq 2e^{-c(d+k)t^2}\,. \tag{11.8}$$

In particular, when $\mathbf{B} = I_n$ and $d \geq n$, then for $t \geq C$

$$\mathbb{P}\left(||\mathbf{A}|| > \sqrt{d+n}\,t\right) \leq 2e^{-c(d+n)t^2}\,, \tag{11.9}$$

which may be simplified, when $d \geq n$ and for $t \geq C$, to

$$\mathbb{P}\left(||\mathbf{A}|| > \sqrt{d}\,t\right) \leq 2e^{-cdt^2}\,.$$

Proof. When B is deterministic, the proof of this Proposition is analogous (in fact, easier) to that of Proposition 9, the only difference being that all the r.v.'s in sight are subgaussian (instead of subexponential). An even simpler proof of (11.9) may be found in [93]. To extend the result to include non-deterministic B, note:

$$\begin{aligned}
\mathbb{P}(||\mathbf{BA}|| > t||\mathbf{B}||(\sqrt{d}+\sqrt{k})) &= \mathbb{E}_{\mathbf{A},\mathbf{B}}\left[1_{||\mathbf{BA}||>t||\mathbf{B}||(\sqrt{d}+\sqrt{k})} \right] \\
&= \mathbb{E}_{\mathbf{B}}\left[\mathbb{E}_{\mathbf{A},\mathbf{B}}\left[1_{||\mathbf{BA}||>t||\mathbf{B}||(\sqrt{d}+\sqrt{k})} \,|\, \mathbf{B} \right] \right] \\
&= \mathbb{E}_{\mathbf{B}}\left[\mathbb{E}_{\mathbf{A}}\left[1_{||\mathbf{BA}||>t||\mathbf{B}||(\sqrt{d}+\sqrt{k})} \,|\, \mathbf{B} \right] \right] \\
&\leq \mathbb{E}_{\mathbf{B}}\left[2e^{-c(d+k)t^2} \right] = 2e^{-c(d+k)t^2} \,.
\end{aligned}$$

Here we use the fact that due to independence $p_{\mathbf{A}|\mathbf{B}} = p_{\mathbf{A}}$. □

References

[1] J. B. Tenenbaum, V. D. Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.

[2] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[3] M. Belkin, P. Niyogi, Using manifold structure for partially labelled classification, Advances in NIPS 15.

[4] D. L. Donoho, C. Grimes, When does isomap recover natural parameterization of families of articulated images?, Tech. Rep. 2002-27, Department of Statistics, Stanford University (August 2002).

[5] D. L. Donoho, C. Grimes, Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data, Proc. Nat. Acad. Sciences (2003) 5591–5596.

[6] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, SIAM Journal of Scientific Computing 26 (2002) 313–338.

[7] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, PNAS 102 (21) (2005) 7426–7431.

[8] M. B. Wakin, D. L. Donoho, H. Choi, R. G. Baraniuk, The multiscale structure of non-differentiable image manifolds, in: SPIE Wavelets XI, San Diego, 2005.

[9] D. L. Donoho, O. Levi, J.-L. Starck, V. J. Martinez, Multiscale geometric analysis for 3-d catalogues, Tech. rep., Stanford Univ. (2002).

[10] J. Costa, A. Hero, Learning intrinsic dimension and intrinsic entropy of high dimensional datasets, in: Proc. of EUSIPCO, Vienna, 2004.

[11] F. Camastra, A. Vinciarelli, Intrinsic dimension estimation of data: An approach based on grassberger-procaccia's algorithm, Neural Processing Letters 14 (1) (2001) 27–34.

[12] F. Camastra, A. Vinciarelli, Estimating the intrinsic dimension of data with a fractal-based method, IEEE P.A.M.I. 24 (10) (2002) 1404–10.

[13] W. Cao, R. Haralick, Nonlinear manifold clustering by dimensionality, ICPR 1 (2006) 920–924.

[14] M. A. Rohrdanz, W. Zheng, M. Maggioni, C. Clementi, Determination of reaction coordinates via locally scaled diffusion map, J. Chem. Phys. (134) (2011) 124116.

[15] W. Zheng, M. A. Rohrdanz, M. Maggioni, C. Clementi, Polymer reversal rate calculated via locally scaled diffusion map, J. Chem. Phys. (134) (2011) 144108.

[16] W. Allard, G. Chen, M. Maggioni, Multiscale geometric methods for data sets II: Geometric wavelets, Appl. Comp. Harm. Anal., accepted.

[17] Approximation of points on low-dimensional manifolds via compressive measurements, preprintArxiv.

[18] G. Chen, M.Iwen, M.Maggioni, in preparation.

[19] G. Chen, A. Little, M. Maggioni, L. Rosasco, Wavelets and Multiscale Analysis: Theory and Applications, Springer Verlag, 2011, submitted March 12th, 2010.

[20] G. Chen, M. Maggioni, Multiscale geometric and spectral analysis of plane arrangements, in: Proc. CVPR, 2011, to appear.

[21] G. Chen, M. Maggioni, Multiscale geometric methods for data sets III: multiple planes, in preparation.

[22] T. Zhang, A. Szlam, Y. Wang, G. Lerman, Hybrid Linear Modeling via Local Best-fit Flats, ArXiv e-prints, and CVPR 2010.

[23] M. Muldoon, R. MacKay, J. Huke, D. Broomhead, Topolgy from time series, Physica D 65 (1993) 1–16.

[24] D. Broomhead, R. Indik, A. Newell, D. Rand, Local adaptive galerkin bases for large dimensional dynamical systems, Nonlinearity 4 (1991) 159–197.

[25] J. Farmer, J. Sidorowich, Predicting chaotic time series, Phys. Rev. Lett. 59(8) (1987) 845–848.

[26] P. W. Jones, The traveling salesman problem and harmonic analysis, Publ. Mat. 35 (1) (1991) 259–267, conference on Mathematical Analysis (El Escorial, 1989).

[27] G. David, S. Semmes, Uniform Rectifiability and Quasiminimizing Sets of Arbitrary Codimension, AMS.

[28] G. David, Wavelets and Singular Integrals on Curves and Surfaces, Springer-Verlag, 1991.

[29] A. Little, Y.-M. Jung, M. Maggioni, Multiscale estimation of intrinsic dimensionality of data sets, in: Proc. A.A.A.I., 2009.

[30] A. Little, J. Lee, Y.-M. Jung, M. Maggioni, Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD, in: Proc. S.S.P., 2009.

[31] G. Chen, A. Little, M. Maggioni, Multi-resolution geometric analysis for data in high dimensions, Proc. FFT 2011.

[32] A. V. Little, Estimating the Intrinsic Dimension of High-Dimensional Data Sets: A Multiscale, Geometric Approach (April 2011).

[33] P. W. Jones, Rectifiable sets and the traveling salesman problem, Invent. Math. 102 (1) (1990) 1–15.

[34] G. David, J. Journé, A boundedness criterion for generalized Calderón-Zygmund operators, Annals of Mathematics.

[35] G. David, S. Semmes, Analysis of and on uniformly rectifiable sets, Vol. 38 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 1993.

[36] R. Schul, Analyst's traveling salesman theorems. a survey., http://www.math.sunysb.edu/ schul/math/survey.pdf.

[37] M. Rudelson, Random vectors in the isotropic position, J. of Functional Analysis 164 (1) (1999) 60–67.

[38] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices (Aug. 2010).

[39] K. Fukunaga, D. Olsen, An algorithm for finding intrinsic dimensionality of data, IEEE Trans. Computer 20 (2) (1976) 165–171.

[40] J. Bruske, G. Sommer, Intrinsic dimensionality estimation with optimally topology preserving maps, IEEE Trans. Computer 20 (5) (1998) 572–575.

[41] D. Hundley, M. Kirby, Estimation of topological dimension, in: D. Barbara, C. Kamath (Eds.), Proc. Third SIAM Int. Conf. Data Mining, 2003, pp. 194–202.

[42] M. Kirby, Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns, John Wiley & Sons, Inc., New York, NY, USA, 2000.

[43] P. J. Verveer, R. P. Duin, An evaluation of intrinsic dimensionality estimators, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (1).

[44] E. Levina, P. Bickel, Maximum likelihood estimation of intrinsic dimension, In Advances in NIPS 17,Vancouver, Canada.

[45] G. Haro, G. Randall, G. Sapiro, Translated poisson mixture model for stratification learning, Int. J. Comput. Vision 80 (3) (2008) 358–374.

[46] K. Carter, A. Hero, Variance reduction with neighborhood smoothing for local intrinsic dimension estimation, Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on (2008) 3917–3920.

[47] K. Carter, A. O. Hero, R. Raich, De-biasing for intrinsic dimension estimation, Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on (2007) 601–605.

[48] J. Costa, A. Hero, Geodesic entropic graphs for dimension and entropy estimation in manifold learning, Signal Processing, IEEE Transactions on 52 (8) (2004) 2210–2221.

[49] M. Raginsky, S. Lazebnik, Estimation of intrinsic dimensionality using high-rate vector quantization, Proc. NIPS (2005) 1105–1112.

[50] F. Takens, On the numerical determination of the dimension of an attractor, in: Dynamical systems and bifurcations (Groningen, 1984), Vol. 1125 of Lecture Notes in Math., Springer, Berlin, 1985, pp. 99–106.

[51] M. Hein, Y. Audibert, Intrinsic dimensionality estimation of submanifolds in euclidean space, in: S. W. De Raedt, L. (Ed.), ICML Bonn, 2005, pp. 289 – 296.

[52] S. Borovkova, R. Burton, H. Dehling, Consistency of the Takens estimator for the correlation dimension, Ann. Appl. Probab. 9 (2) (1999) 376–390.

[53] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Phys. D 9 (1-2) (1983) 189–208.

[54] A. M. Farahmand, C. S. J.-Y. Audibert, Manifold-adaptive dimension estimation, Proc. I.C.M.L.

[55] R. J. D. S. Broomhead, G. P. King, Topological dimension and local coordinates from time series data, J. Phys. A: Math. Gen. 20 (1987) L563–L569.

[56] A. N. D.S. Broomhead, R. Indik, D. Rand, Local adaptive galerkin bases for large-dimensional dynamical systems, Nonlinearity 4 (1991) 159–197.

[57] J. Lee, Riemannian manifolds: An introduction to curvature, Springer, 1997.

[58] S. Har-Peled, M. Mendel, Fast construction of nets in low-dimensional metrics and their applications, SIAM J. Comput. 35 (5) (2006) 1148–1184.

[59] A. Beygelzimer, S. Kakade, J. Langford, Cover trees for nearest neighbor, in: ICML '06: Proceedings of the 23rd international conference on Machine learning, ACM, New York, NY, USA, 2006, pp. 97–104.

[60] V. Rokhlin, A. Szlam, M. Tygert, A randomized algorithm for principal component analysis, SIAM Jour. Mat. Anal. Appl. 31 (3) (2009) 1100–1124.

[61] G. Haro, G. Randall, G. Sapiro, Translated poisson mixture model for stratification learning, Int. J. Comput. Vision 80 (3) (2008) 358–374.

[62] E. Levina, P. J. Bickel, Maximum likelihood estimation of intrinsic dimension, in: L. K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA, 2005, pp. 777–784.

[63] J. Costa, A. Hero, Geodesic entropic graphs for dimension and entropy estimation in manifold learning, Signal Processing, IEEE Transactions on 52 (8) (2004) 2210–2221.

[64] K. Carter, A. Hero, Variance reduction with neighborhood smoothing for local intrinsic dimension estimation, Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on (2008) 3917–3920.

[65] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, L. Carin, Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds, IEEE Trans. Signal Processing.

[66] H. Chen, J. Silva, D. Dunson, L. Carin, Hierarchical bayesian embeddings for analysis and synthesis of dynamic data, submitted.

[67] B. Kegl, Intrinsic dimension estimation using packing numbers, 2002, pp. 681–688.

[68] M. Fan, H. Qiao, B. Zhang, Intrinsic dimension estimation of manifolds by incising balls, Pattern Recogn. 42 (5) (2009) 780–787.

[69] A. M. Farahmand, C. Szepesvári, J.-Y. Audibert, Manifold-adaptive dimension estimation., in: Proceedings of the 24th international conference on Machine learning, 2007.

[70] W. Johnson, J. Lindenstrauss, Extension of lipschitz maps into a hilbert space, Contemp. Math. 26 (1984) 189–206.

[71] R. Baraniuk, M. Wakin, Random projections of smooth manifolds, preprint.

[72] P. Jones, M. Maggioni, R. Schul, Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels, Proc. Nat. Acad. Sci. 105 (6) (2008) 1803–1808.

[73] P. Jones, M. Maggioni, R. Schul, Universal local manifold parametrizations via heat kernels and eigenfunctions of the Laplacian, Ann. Acad. Scient. Fen. 35 (2010) 1–44, http://arxiv.org/abs/0709.1975.

[74] A. Singer, R. Erban, I. G. Kevrekidis, R. R. Coifman, Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps, Proc. Natl. Acad. Sci. 106 (38) (2009) 16090–16095.

[75] R. Vershynin, How close is the sample covariance matrix to the actual covariance matrix?Submitted.

[76] B. B. Mandelbrot, R. L. Hudson, The (mis)behavior of markets, Basic Books, New York, 2004, a fractal view of risk, ruin, and reward.

[77] P.W.Jones, Rectifiable sets and the traveling salesman problem, Inventiones Mathematicae 102 (1990) 1–15.

[78] N. Verma, S. Kpotufe, S. Dasgupta, Which spatial partition trees are adaptive to intrinsic dimension?, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, AUAI Press, Arlington, Virginia, United States, 2009, pp. 565–574.
URL http://dl.acm.org/citation.cfm?id=1795114.1795180

[79] I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Stat. 29 (2) (2001) 295–327.
URL http://ProjectEuclid.org/getRecord?id=euclid.aos/1009210544

[80] J. Baik, J. W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, Journal of Multivariate Analysis 97 (6) (2006) 1382–1408.
URL http://arxiv.org/abs/math/0408165

[81] J. Silverstein, On the empirical distribution of eigenvalues of large dimensional information-plus-noise type matrices, Journal of Multivariate Analysis 98 (2007) 678–694.
URL http://www4.ncsu.edu/~jack/pub.html

[82] V. I. Koltchinskii, Empirical geometry of multivariate data: a deconvolution approach., Ann. Stat. 28 (2) (2000) 591–629.

[83] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statistica Sinica 17 (2007) 1617–1642.

[84] B. Nadler, Finite sample approximation results for principal component analysis: a matrix perturbation approach, Ann. Stat. 36 (6) 2791–2817.

[85] D. N. Kaslovsky, F. G. Meyer, Optimal Tangent Plane Recovery From Noisy Manifold Samples, ArXiv e-prints.

[86] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann. Stat. 23 (4) (1952) 493–507.
URL http://www.jstor.org/stable/2236576

[87] P. Niyogi, S. Smale, S. Weinberger, Finding the homology of submanifolds with high confidence from random samples, Discrete and Computational Geometry 39 (2008) 419–441, 10.1007/s00454-008-9053-2.
URL http://dx.doi.org/10.1007/s00454-008-9053-2

[88] A. Barvinok, Measure concentration (2005).
URL http://www.math.lsa.umich.edu/~barvinok/total710.pdf

[89] H. Wielandt, Topics in the Analytic Theory of Matrices, Univ. Wisconsin Press, Madison, 1967.

[90] I. Pinelis, An approach to inequalities for the distributions of infinite-dimensional martingales, Probability in Banach Spaces, 8, Proceedings of the 8th International Conference (1992) 128–134.

[91] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, Ann. Probab. 22 (4) (1994) 1679–1706.

[92] V. Buldygin, Y. Kozachenko, Metric Characterization of Random Variables and Random Processes, American Mathematical Society, 2000.

[93] M. Rudelson, R. Vershynin, The smallest singular value of a random rectangular matrix, submitted (Nov. 2008).
URL http://arxiv.org/abs/0802.3956