Manifold Learning and its Applications:
Papers from the AAAI Fall Symposium (FS-09-04)

Multiscale Estimation of Intrinsic Dimensionality of Data Sets

Anna V. Little and **Yoon-Mo Jung** and **Mauro Maggioni**
Department of Mathematics, Duke University,
P.O. Box 90320, Durham, NC, 27708, U.S.A.

Abstract

We present a novel approach for estimating the intrinsic dimensionality of certain point clouds: we assume that the points are sampled from a manifold $\mathcal{M}$ of dimension k, with $k << D$, and corrupted by D-dimensional noise. When $\mathcal{M}$ is linear, one may analyze this situation by SVD: with no noise one would obtain a rank k matrix, and noise may be treated as a perturbation of the covariance matrix. When $\mathcal{M}$ is a nonlinear manifold, global SVD may dramatically overestimate the intrinsic dimensionality. We introduce a multiscale version SVD and discuss how one can extract estimators for the intrinsic dimensionality that are highly robust to noise, while require a smaller sample size than current estimators.

1. Introduction

The problem of estimating the intrinsic dimensionality of a point cloud is of interest in a wide variety of problems. To cite some important instances, it is equivalent to estimating the number of variables in a statistical linear model in statistics, the number of degrees of freedom in a dynamical system, the intrinsic dimensionality of a data set modeled by a probability distribution highly concentrated around a low-dimensional manifold. Many applications and algorithms crucially rely on the estimation of the number of components in the data, for example spectrometry, signal processing, genomics, economics, to name only a few. Finally, many manifold learning algorithms assume that the intrinsic dimensionality is given.

When the data is generated by a multivariate linear model, $x_i = \sum_{j=1}^{k} \alpha_j v_j$ for some random variables α_j, and the v_j's are fixed vectors in $\mathbb{R}^D$, the number of components is simply the rank of the data matrix X, where X_{ij} is the j-th coordinate of the i-th sample. In this case principal component analysis (PCA) or singular value decomposition (SVD) may be used to recover the rank and the subspace spanned by the v_j's, and this situation is well understood, at least when the number of samples n goes to infinity. The more interesting case in most applications assumes that we observe noise measurements $\tilde{x}_i = x_i + \eta_i$, where η represents noise. This case is also quite well understood, at least when n tends to infinity (see e.g., out of many works, (Johnstone 2001), (Paul 2007), (Silverstein 2007) and references therein).

The finite sample situation is less well understood. While we derive new results in this direction with a new approach, the situation in which we are really interested is that of data having a geometric structure more complicated than linear. In particular, an important trend in machine learning and analysis of high-dimensional data sets assumes that data lies on a nonlinear low-dimensional manifold. Several algorithms have been proposed to estimate intrinsic dimensionality in this setting; for lack of space we cite only (Levina and Bickel 2005) (Haro, Randall, and Sapiro 2008), (Carter and Hero 2008), (Carter, Hero, and Raich 2007), (Costa and Hero 2004), (Camastra and Vinciarelli 2002), (Cao and Haralick 2006), (Raginsky and Lazebnik 2005), (Takens 1985), (Hein and Audibert 2005), (Borovkova, Burton, and Dehling 1999), (Grassberger and Procaccia 1983), (Farahmand and Audibert 2007), (Fukunaga and Olsen 1971) and we refer the reader to the many references therein. One the most important take away messages from the papers above is that most methods are based on estimating volume growth rate of an approximate ball of radius r on the manifold, and that such techniques necessarily require a number of sample points exponential in the intrinsic dimension. Moreover, such methods do not consider the case when high-dimensional noise is added to the manifold. The methods of this paper, on a restricted model of "manifold+noise", only requires a number of points essentially proportional to the intrinsic dimensionality, as detailed in (Lee et al. 2009).

2. Multiscale Dimensionality Estimation

We start by describing a stochastic geometric model generating the point clouds we will study. Let $(\mathcal{M}, g)$ be a smooth k-dimensional Riemannian manifold with bounded curvature, isometrically embedded in $\mathbb{R}^D$. Let η be $\mathbb{R}^D$-valued with $\mathbb{E}[\eta] = 0$, $\mathrm{Var}[\eta] = \sigma^2$ (the "noise"), for example $\eta \sim \sigma\mathcal{N}(0, I_D)$. Let $X = \{x_i\}_{i=1}^n$ be a set of uniform (with respect to the natural volume measure on $\mathcal{M}$) [1] independent random samples on $\mathcal{M}$. Our observations $\tilde{X}$ are noisy samples: $\tilde{X} = \{x_i + \sigma\eta_i\}_{i=1}^n$, where η_i are i.i.d. sam-

[1] In fact, the same techniques work for a measure μ absolutely continuous w.r.t. the volume measure, with Radon-Nykodym derivative bounded above and below.

ples from η and $\sigma > 0$. These points may also be thought of as sampled from a probability distribution $\tilde{\mathcal{M}}$ supported in $\mathbb{R}^D$, concentrated around $\mathcal{M}$. Here and in what follows we represent a set of n points in $\mathbb{R}^D$ by an $n \times D$ matrix, whose (i, j) entry is the j-th coordinate of the i-th point. In particular, X and $\tilde{X}$ will be used to denote both the point cloud and the associated $n \times D$ matrices, and N is the noise matrix of the η_i's.

The problem we consider is to **estimate $k = \dim \mathcal{M}$, given $\tilde{X}$.** Let us first consider the case when $\mathcal{M}$ is a *linear manifold* (e.g. the image of a cube under a linear map), and η is Gaussian noise; the standard approach is to compute the principal components of $\tilde{X}$, and use the singular values to estimate k. Let $\mathrm{cov}(X) = \frac{1}{n} X^T X$ be the $D \times D$ covariance matrix of X, and $\Sigma(X) = (\sigma_i^2)_{i=1}^D$ its eigenvalues. These are the singular values (S.V.) squared of $n^{-1/2}X$, i.e. the first D diagonal entries of Σ in the singular value decomposition (SVD) $X = U\Sigma V^T$. At least for $n \sim k \log k$, it is easy to show that with high probability (w.h.p.) exactly k S.V.'s are nonzero, and the remaining $D - k$ are equal to 0.

Since we are given $\tilde{X}$ and not X, we think of $\tilde{X}$ as a random perturbation of X and expect that $\Sigma(\tilde{X})$ is close to $\Sigma(X)$, so that $\sigma_1, \ldots, \sigma_k \gg \sigma_{k+1}, \ldots, \sigma_D$, allowing us to estimate k correctly with high probability (w.h.p.). This is a very common procedure, applied in a wide variety of situations, and often generalized to kernelized versions of principal component analysis, such as widely used dimensionality reduction methods.

In the general case when $\mathcal{M}$ is a *nonlinear manifold*, there are several problems with the above line of reasoning. Curvature in general forces the dimensionality of a best-approximating hyperplane, such as that provided by a truncated SVD, to be much higher than necessary. As a first trivial example, consider a planar circle ($k = 1$) embedded in $\mathbb{R}^D$: $\mathrm{cov}(X)$ has exactly 2 nonzero eigenvalues equal to the radius squared. More generally, it is easy to construct a one-dimensional manifold ($k = 1$) in $\mathbb{R}^D$ such that $\mathrm{cov}(X)$ has full rank D: it is enough to pick a curve that spirals out in more and more dimensions. A simple construction (sometimes called Y. Meyer's staircase) is the following: let $\chi_{[0,1)}(x) = 1$ if $x \in [0, 1)$ and 0 otherwise. Then the set $\{x_t := \chi_{[0,1)}(\cdot - t)\}_{t \in \mathbb{R}} \subset L^1(\mathbb{R}) := \{f : \mathbb{R} \to \mathbb{R}, f \text{ meas. } s.t. \int_{\mathbb{R}} |f| < \infty\}$ is a one-dimensional manifold which is not contained in any subspace of dimension less than D. Notice that x_{t_1} and x_{t_2} are orthogonal whenever $|t_1 - t_2| > 1$, so this curve spirals into larger and larger subspaces as t increases. Similar considerations would hold after discretization of the space and restriction of t to a bounded interval. The above phenomena are a consequence of performing SVD globally: if one thinks locally, the matter seems once again easily resolved. Let $z \in \mathcal{M}$, r a small enough radius and consider only the points $X^{(z,r)}$ in $B_z(r) \cap \mathcal{M}$ (the ball is in $\mathbb{R}^D$). For r small enough, $X^{(z,r)}$ is well-approximated by a portion of k-dimensional tangent plane $T_z(\mathcal{M})$, and therefore we expect $\mathrm{cov}(X^{(z,r)})$ to have k large eigenvalues, growing like $O(r^2)$, and smaller eigenvalues caused by curvature, growing like $O(r^4)$. By letting $r_z \to 0$, i.e. *choosing r_z small enough dependent on curvature*, the curvature eigenvalues tend to 0 faster than the top k eigenvalues. Therefore, if we were given X, in the limit as $n \to \infty$ and $r_z \to 0$, this would give a consistent estimator of k. It is important to remark that these two limits are not unconditional: we need r_z to approach 0 slow enough and n to grow fast enough so that $B_z(r)$ contains enough points to accurately estimate the SVD in $B_z(r)$.

However, more interestingly, if we are given $\tilde{X}$, i.e. noise is added to the samples, we meet another constraint, that forces us to not take r_z too small. Let $\tilde{X}^{(z,r)}$ be $\tilde{\mathcal{M}} \cap B_z(r)$. If r_z is comparable to a quantity dependent on σ and D (for example, $r_z \sim \sigma\sqrt{D}$ when $\eta \sim \sigma\mathcal{N}(0, I_D)$), and $B_z(r)$ contains enough points, $\mathrm{cov}(\tilde{X}^{(z,r)})$ may approximate the covariance of η rather than that of points on $T_z(\mathcal{M})$. Only for r_z larger than a quantity dependent on σ, yet smaller than a quantity depending on curvature, conditioned on $B_z(r)$ containing enough points, will we expect $\mathrm{cov}(\tilde{X}^{(z,r)})$ to approximate a "noisy" version of $T_z(\mathcal{M})$.

Since we are not given σ, nor the curvature of $\mathcal{M}$, we cannot determine quantitatively the correct scale $r = r(z)$ qualitatively described above. This suggests that we take a *multiscale approach*. For every point $z \in \mathcal{M}$ and scale parameter $r > 0$, let $\mathrm{cov}(z, r) = \mathrm{cov}(\tilde{X}^{(z,r)})$ and let $\{\sigma_i^{(z,r)}\}_{i=1,\ldots,D}$ be the corresponding eigenvalues, as usual sorted in nonincreasing order. We will call them multiscale singular values (S.V.'s) (P.W.Jones 1990). What is the behavior of $\sigma_i^{(z,r)}$ as a function of i, z, r? How can they be used to detect k, and what can they tell us about $\mathcal{M}$? What is the effect of sampling and noise in estimating them? To build our intuition, we start with a simple yet surprising example.

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

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 $\mathbb{R}^D$ using the natural embedding of $\mathbb{R}^{k+1}$ in $\mathbb{R}^D$ via the first $k + 1$ coordinates. We obtain X by sampling n points uniformly at random from $\mathbb{S}^k$, and $\tilde{X}$ is obtain by adding D-dimensional white Gaussian noise of variance σ in every direction. We call this data set $\mathbb{S}^k(D, n, \sigma)$.

In Figure 1 we consider the multiscale S.V.'s of $\mathbb{S}^9(100, 1000, 0.1)$ as a function of r. Several observations are in order. First of all, notice that $\mathbb{R}^{10}$ is divided into 2^{10} sectors, and therefore by sampling 1000 points on $\mathbb{S}^9$ we obtain about 1 point per sector (!). Secondly, observe that the noise size, if measured by $||x_i - \tilde{x}_i||_2^2$, i.e. by how much each point is displaced, would be of order $\mathbb{E}[\sigma^2 \chi_D^2] \sim 1$ (where χ_D^2 is a χ^2 distribution with D degrees of freedom), which is comparable with the radius of the sphere itself (!). Therefore this data set may be described as randomly sampling one point per sector at distance 1 from the origin in the first $k + 1$ coordinates, then moving by 1 in a random direction in $\mathbb{R}^{100}$. The situation may seem hopeless.

In fact, we can reliably detect the intrinsic dimensionality of $\mathcal{M}$. At very small scales, $B_z(r)$ is empty or contains less than $O(k)$ points, and the rank of $\mathrm{cov}(z, r)$ is even less than k. At small scales, no gap among the $\sigma_i^{(z,r)}$ is visible: $B_z(r)$ contains too few points, scattered in all directions by the noise, and new increasing S.V.'s keep arising for several scales. At larger scales, the top $9 = k$ 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 (quadratically) growing 10th S.V., while the remaining smaller S.V.'s tend approximately to the *one-dimensional* noise variance σ: this is the size of the noise relevant in our procedure, rather than the much larger expected displacement measured in the full $\mathbb{R}^D$, which was of size $O(1)$.

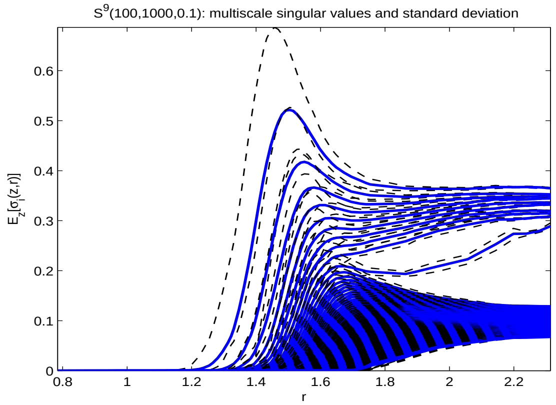


Figure 1: $\mathbb{S}^9(100, 1000, 0.1)$: plot of $\mathbb{E}_z[\sigma_i^{(z,r)}]$, and corresponding standard deviation bands (dotted), as a function of r. The top 9 S.V.'s dominate and correspond to the intrinsic dimensions; the 10-th S.V. corresponds to curvature, and slowly increases with scale (note that at large scale $\Delta_{10} > \Delta_9$); the remaining S.V.'s correspond to noise in the remaining 90 dimensions, and converge to the one-dimensional noise size σ.

3. Analysis

Motivated by applications to large data sets in high-dimensional spaces that are assumed to be intrinsically low-dimensional, we are interested in the regime where D is large, $k << D$, and will ask how small n needs to be in order to estimate k correctly w.h.p. In a classical statistical framework one may rather be interested in the regime where D, k are fixed and n tends to infinity, but in that case one would conduct the analysis as $r \to 0$ and this would lead essentially to the problem of consistency of PCA, and noise would be a relatively minor complication. In many applications D is large and n cannot be taken much larger than D itself: we will therefore be interested in the regime where D and n are large but $\frac{n}{D} = O(1)$ or, even more ambitiously, $\frac{n}{k} = O(1)$, independently of D.

Let us fix $z \in \mathcal{M}$ and a radius r, and focus our attention on $\mathcal{M} \cap B_z(r)$, the intersection of $\mathcal{M}$ with the Euclidean ball of radius r around z. Let us observe n points sampled from $\mathcal{M} \cap B_z(r)$ and corrupted by noise. With the same notation as in section 2,let $X, \tilde{X} \in \mathrm{Mat}(n, D)$ be the matrices representing the noiseless samples and the noisy samples. Without loss of generality we may assume that $T_z(\mathcal{M})$ is spanned by the first k coordinates in $\mathbb{R}^D$. Let $N \in \mathrm{Mat}(n, D)$ be the noise matrix (i.e. N_{ij} is the j-th coordinate of η_i), so that $\tilde{X} = X + N$. It will be handy to write $N = [N_1 | N_2]$, where $N_1 \in \mathrm{Mat}(n, k)$ and $N_2 \in \mathrm{Mat}(n, d)$, and $\tilde{X}_1 = X + N_1$. Here and in what follows, $d = D - k$ is the dimension of the normal bundle of $\mathcal{M}$. Finally, let $X_0 = n^{-1/2} X$ and $\sigma_{\max} = \mathbb{E}[\sigma_{\max}(X_0)]$.

3.1 Case A: Linear and single-scale

We start from the following special setting: we assume that $\mathcal{M}$ is a k-dimensional subspace of $\mathbb{R}^D \cap B_z(r)$. In this case everything is scale invariant, and we may as well focus on any scale r and location z: assume therefore that we have n points $X^{(z,r)}$ in $\mathcal{M} \cap B_z(r)$. We are interested in studying $\mathrm{cov}(\tilde{X}^{(z,r)}) = \mathrm{cov}(X^{(z,r)} + N)$, in particular deciding under which conditions we can detect the intrinsic dimensionality k. To simplify the notation, in this subsection we let $X = X^{(z,r)}$ and $\tilde{X} = \tilde{X}^{(z,r)}$.

Clearly the gaps between the S.V.'s of $\mathrm{cov}(X)$ will play a fundamental role.

Definition. Let $\Delta_i := \Delta_i(C) := \sigma_i(\mathrm{cov}(X)) - \sigma_{i+1}(\mathrm{cov}(X)) = \sigma_i^2(X_0) - \sigma_{i+1}^2(X_0)$, for $i = 1, \ldots, D-1$, $\Delta_D := \Delta_D(C) = \sigma_D(\mathrm{cov}(X))$, and let $\tilde{\Delta}_i$ be the analogous quantities for $\tilde{X}$.

The introduction of the noise η generates immediately random matrices. We shall work with noise such that the η_i's are Gaussian i.i.d. random variables, albeit all the results hold much more generally.

We recall the following properties of random matrices. We are interested in non-asymptotic results, that hold for finite m, n, since they will imply finite sample inequalities with high probability (w.h.p.). These results are therefore quite different from the usual asymptotic results, either "classical" ($n \to +\infty$, everything else being fixed) or from random matrix theory ($Dn^{-1} \to \gamma$, with both $D, n \to +\infty$, see e.g. (Johnstone 2001), (Silverstein 2007), (Baik and Silverstein 2006)). Another fundamental difference between our results is that they focus on the case where the structure of the data is intrinsically low-dimensional, and therefore the true covariance is very singular. Such singularity is often either avoided by assumption or considered as a consequence of lack of data (n too small), rather than as a fundamental assumption on the underlying geometric structure of the data as we do here.

The following is well known (Rudelson and Vershynin 2008):

Theorem 31. *Let $Y \in \mathrm{Mat}(m, n)$ with Y_{ij} i.i.d. standard Gaussian random variables. Then for $t \geq 0$,*

$$\sqrt{m} - \sqrt{n} - t \leq \sigma_{\min}(Y) \leq \sigma_{\max}(Y) \leq \sqrt{m} + \sqrt{n} + t \quad (1)$$

with probability at least $1 - 2e^{-t^2/2}$, and

$$\sqrt{m} - \sqrt{n} \leq \mathbb{E}[\sigma_{\min}(Y)] \leq \mathbb{E}[\sigma_{\max}(Y)] \leq \sqrt{m} + \sqrt{n} \quad (2)$$

In particular, an $n \times n$ random matrix as in the Theorem has norm $\sim \sqrt{n}$ on the complement of an event with very (exponentially) small probability. In our context we deduce the following bounds (we postpone the proofs to the Appendix):

Lemma 32. *With the notation and assumptions above,*

$$\begin{aligned} \mathbb{E}[n^{-1}||N_1^T N_1||] &\leq \sigma^2(1+\sqrt{kn^{-1}})^2 && (3)\\ \mathbb{E}[n^{-1}||N_2^T N_2||] &\leq \sigma^2(1+\sqrt{dn^{-1}})^2 && (4)\\ \mathbb{E}[n^{-1}\sigma_{\min}(N_2^T N_2)] &\geq \sigma^2(1-\sqrt{dn^{-1}})^2 && (5)\\ \mathbb{E}[n^{-1}||N_1^T X||] &\leq 2c_1\sigma\sqrt{kn^{-1}}\,\sigma_{\max} && (6)\\ \mathbb{E}[n^{-1}||N_2^T \tilde{X}_1||] &\leq \sigma(c_1\sigma_{\max}+c_2\sigma)\cdot \\ &\cdot \sqrt{n^{-1}}(\sqrt{k}+\sqrt{d}) && (7) \end{aligned}$$

Here c_1, c_2 are universal constants that depend only on the distribution of η.

The first two inequalities bound the size of the noise tangent and normal to $\mathcal{M}$, resp., and the last two inequalities bound the size of the correlations between $\mathcal{M}, \tilde{\mathcal{M}}$ and tangential and normal noise. We observe that (3), (4) may be replaced by exponential tail inequalities as in (1), and standard ϵ-net arguments (see e.g. (Vershynin 2008)) show that the same holds in (6).

To relate the singular values of $\tilde{C}$ to those of C, we split the perturbation $\tilde{C} - C$ as follows:

$$nC = \begin{pmatrix} X^T X & 0 \\ 0 & 0 \end{pmatrix} \rightarrow^{P_1} \begin{pmatrix} \tilde{X}_1^T \tilde{X}_1 & 0 \\ 0 & 0 \end{pmatrix}$$
$$\rightarrow^{P_2} \begin{pmatrix} \tilde{X}_1^T \tilde{X}_1 & 0 \\ 0 & N_2^T N_2 \end{pmatrix} \rightarrow^{P_3} n\tilde{C} = \begin{pmatrix} \tilde{X}_1^T \tilde{X}_1 & \tilde{X}_1^T N_2 \\ N_2^T \tilde{X}_1 & N_2^T N_2 \end{pmatrix}$$

We shall also let $C = n^{-1}X^T X$, $\tilde{C}_1 = n^{-1}\tilde{X}_1^T \tilde{X}_1$ and $\tilde{C}_2 = n^{-1}N_2^T N_2$. These perturbations are essentially independent, and therefore we expect sharp bounds by bounding the norm of their sum by the sum of their norms.

We could estimate accurately $\sigma_i(\tilde{C})$ (and the singular vectors), but for lack of space we discuss only $\tilde{\Delta}_i$. In particular we upper bound $\tilde{\Delta}_i$ for $i < k$ (these are the "manifold" S.V.'s, corresponding to perturbed tangent eigenvectors) and for $i > k$ (these are the "noise" S.V.'s, corresponding to perturbed normal eigenvectors), and we lower bound $\tilde{\Delta}_k$, which separates tangent S.V.'s from noise S.V.'s and will be used for intrinsic dimensionality estimation.

Proposition 33. *Conditioned on* $\Omega_1 := \{\sigma_{\min}(\tilde{C}_1) \geq \sigma_{\max}(\tilde{C}_2)\}$,

$$\begin{aligned} \mathbb{E}[\tilde{\Delta}_i] &\leq \Delta_i + 8c_1\sigma_{\max}\sigma\sqrt{kn^{-1}} + \sigma^2(1+\sqrt{kn^{-1}})^2 \\ &\quad + (c_1\sigma_{\max}+c_2\sigma)\sigma\sqrt{n^{-1}}(\sqrt{k}+\sqrt{d}) \ \ , i \leq k-1 \\ \mathbb{E}[\tilde{\Delta}_k] &\geq \sigma_k(C) - 4c_1\sigma_{\max}\sigma\sqrt{kn^{-1}} - \sigma^2(1+\sqrt{dn^{-1}})^2 \\ \mathbb{E}[\tilde{\Delta}_i] &\leq (c_1\sigma_{\max}+c_2\sigma)\sigma(\sqrt{kn^{-1}}+\sqrt{dn^{-1}}) \\ &\quad + 4\sigma\sqrt{dn^{-1}} \ \ , k+1 \leq i \leq D-1 \\ \mathbb{E}[\tilde{\Delta}_D] &\sim \sigma^2(1+\sqrt{dn^{-1}})^2 \end{aligned}$$

Deviation inequalities also hold. Since $\mathcal{M}$ is linear, $\sigma_{\max} = \sigma_1(C) = \cdots = \sigma_k(C)$, and $\Delta_i = 0$ for all $i \leq k-1$. Ω_1 has small probability as soon as $\sigma_k(C) > 8c_1\sigma_{\max}\sigma\sqrt{kn^{-1}} + \sigma^2(1+\sqrt{dn^{-1}})^2$.

Several observation are in order.

(1) In the language of signal processing, we can interpret the terms above in terms of signal and noise, where the signal here is dictated by the geometry of $\mathcal{M}$. In the lower bound for $\mathbb{E}[\tilde{\Delta}_k]$ we clearly see the interaction between the term representing the strength of the "geometric signal" $\sigma_k(C)$ and two noise terms: one is the projection of the noise along the tangent space of $\mathcal{M}$ (the "signal" space), which has strength that depends on the intrinsic dimension k of $\mathcal{M}$, and the other is the component of the noise normal to $\mathcal{M}$, of strength depending on the ambient dimension D.

(2) We see here quantified our previous empirical observation that the "size" of the noise that affects our algorithm is not, as one may expect for an algorithm based on distance computations, proportional to the distance distortion $\mathbb{E}[||\tilde{x}_i - \tilde{x}_j||^2]/||x_i - x_j||^2\mathbb{E}[\sigma^2\chi_d^2] \sim \sigma^2 D$, but has size only $\sim \sigma^2(1+\sqrt{Dn^{-1}})^2$. Therefore, as soon as $n = \gamma D$ and $\gamma > O(1)$, the size of the noise is essentially σ^2 which, among other things, is *independent of the ambient dimension*. This is apparent in Figure 1, as well as in all our experiments. Another regime in which things are dimension-independent is when $\sigma \sim D^{-\frac{1}{2}}$, so that $\mathbb{E}[||\eta||^2] = O(1)$ independently of D. This is a very natural geometric scaling that fixes the size of the D-dimensional noise when D varies by shrinking the size of the noise in each dimension. In this regime too, we show in (Lee et al. 2009) that we may estimate k w.h.p. with $n \sim k \log k$ points.

(3) This is also in sharp contrast with other methods (e.g. (Haro, Randall, and Sapiro 2008), (Carter and Hero 2008), (Carter, Hero, and Raich 2007)) based on volume considerations, i.e. counting the number of points in $B_z(r)$. While we are not aware of theoretical analyses of such algorithms, it seems clear upon inspection that they will require a number of points proportional to the volume of $B_z(r)$, i.e. at least exponential in k when there is no noise, and possibly in D in the noisy case.

3.2 Case B: Linear and multiscale

Suppose now that $\mathcal{M}$ is a linear submanifold of $\mathbb{R}^D$ (with uniform volume measure); without loss of generality we can translate and rescale and focus our attention on $X = \mathcal{M} \cap B_0(1)$; we assume we have n points in this intersection. We fix $z = 0$ and consider $\tilde{X}^{(0,r)}$. As r increases, $n(r) = |\tilde{X}^{(0,r)}|$ increases, in a way that depends on k but also on D and η. For example when η is Gaussian, $\mathbb{E}[||x_i - \tilde{x}_i||] = \mathbb{E}[||\eta_i||^2] = \sigma^2 D$, so that $\tilde{X}$ is obtained by scattering the points of X into $\mathbb{R}^D$ by moving each of them, w.h.p., by a quantity very close to $\sigma\sqrt{D}$. Moreover, at least when $D >> k$ (in fact D/k larger than a constant such as 5 will suffice), η_i is almost orthogonal to $\mathcal{M}$ w.h.p.. Then $n(r)$ will be close to 0 for r small, it will increase rapidly when $r \sim \mathbb{E}[||\eta_i||]$, and it will grow at a slower rate for $r \gtrsim \mathbb{E}[||\eta||]$.

The estimates in Proposition 33 can be applied effectively for $C(\eta, D, k) < r < 1 - C(\eta, D, k)$, where C is a constant (in r) depending on η, D, k, i.e. in the range when n is not too small nor too large for ∂X to have any effect. Therefore, provided that $n(r)$ is large enough in this range, the estimates in Proposition 33 will predict when k is identifiable w.h.p.

Some remarks about the effect of the boundary of $\mathcal{M}$ are in order. It is well known that when k is not small ($k > 5$ will suffice) the fraction of points close to $\partial\mathcal{M}$ may be very large. This can be seen in many ways, for example, in the isotropic case, by observing that the number of points in $B_0(r)$ (as above) is $|B_0(1)| \int_0^r \rho^{k-1} d\rho$ (integration in polar coordinates), and because ρ^{k-1} is very close to 0 unless ρ is very close to 1, very few points are close to 0. Algorithms for dimensionality estimation that are based on volume considerations are heavily impacted by this phenomenon, to the point of justifying extra work to "debias" their results (Carter, Hero, and Raich 2007). Our algorithm is not impacted by this issue: the crucial quantity measured by the S.V.'s is not volume- or density- related, but is the length of orthogonal vectors spanning the local subspace spanned by the data. Qualitatively speaking, in absence of noise, $O(1)$ points at roughly equal distance in each of k orthogonal directions are enough for our approach to declare dimensionality at least k. These points exist as soon as $B_z(r)$ contains $O(k)$ randomly sampled points from $\mathcal{M}$, regardless of whether z is in the interior of $\mathcal{M}$ or at the boundary of $\mathcal{M}$. In order to verify this empirically, we considered the k-dimensional unit cube centered at $0 \in \mathbb{R}^k$, looked at the behavior of $\sigma_i^{(z,r)}$ as a function of $||z||$, and noticed no changes impacting our ability to estimate k.

3.3 Case C: Nonlinear and multiscale

In the general case when $\mathcal{M}$ is a nonlinear manifold, it is not true anymore, as in the case when $\mathcal{M}$ is a linear subspace, that larger r's lead to the most useful $\sigma_i^{(z,r)}$ for detecting k. This is due to the curvature of $\mathcal{M}$ in $\mathbb{R}^D$, which forces $X^{(z,r)}$ to have extrinsic dimensionality possibly much higher than k. First of all notice that we are not talking about the intrinsic curvature of $\mathcal{M}$, but of the curvature of $\mathcal{M}$ inside $\mathbb{R}^D$. The same Riemannian manifold $(\mathcal{M}, g)$ may be isometrically embedded in $\mathbb{R}^D$ in different ways, without of course any change in intrinsic curvatures (for example Gauss curvature for 2-manifolds), but with very significant changes in the (extrinsic) curvatures relevant to our analysis.

Nevertheless, certain intrinsic properties of $\mathcal{M}$ may be recovered. For example if $\mathcal{M}$ is flat, an isometric parametrization may be recovered (at least in the limit $n \to \infty$) by Hessian eigenmaps (Donoho and Grimes 2003); more generally, bi-Lipschitz atlases may be found by heat kernel triangulation (Jones, Maggioni, and Schul 2008).

Fix a point $z \in \mathcal{M}$ and consider a chart $u : \mathcal{U} \subseteq \mathbb{R}^k \to \mathcal{M} \subseteq \mathbb{R}^D$, where $\mathcal{U}$ is an open set of $\mathbb{R}^k$ containing the origin, and $u(0) = z$. Consider a path $\gamma : (-\delta, \delta) \to \mathcal{U}$ such that $\gamma(0) = 0$. Then $u \circ \gamma$ is a path on $\mathcal{M}$ through z, and by a Taylor expansion it is easy to see that up to first-order the Jacobian of u governs the first order variations of the path, and the Hessian of u governs the second order variations of the path. There are therefore at most $k(k+1)/2$ extrinsic principal curvatures. At any rate, one may compute exactly $\{\sigma_i^{(z,r)}\}_{i=1,2}$ in the case of each path γ and see that the effect of curvature is, for small r, to generate a second nontrivial S.V. $\sigma_2^{(z,r)}$ with $\sigma_2^{(z,r)} \sim O(\kappa^2)\sigma_1^{(z,r)}$, where κ is the curvature of γ viewed as a path on $\mathcal{M} \subset \mathbb{R}^D$.

We therefore expect to see at most $k + k(k+1)/2$ "curvature" S.V.'s when running our algorithm on a general k-dimensional manifold $\mathcal{M}$. These $\sigma_i^{(z,r)}$'s, however, will exhibit a behavior different from that of the first k "tangential" $\sigma_i^{(z,r)}$; more precisely, we expect $\sigma_l^{(z,r)} \sim O({\sigma_i^{(z,r)}}^2)$, where $i \leq k$, and $l > k$ indexes the curvature S.V.'s.

Another issue that we need to consider is that we cannot expect in general to have $\sigma_1^{(z,r)} \approxeq \cdots \approxeq \sigma_k^{(z,r)}$. Recall that from the bounds in Proposition 33 it was useful to have $\sigma_k(C^{(z,r)})$ as large as possible, i.e. as close as possible to $\sigma_1(C^{(z,r)})$. Geometrically, this corresponds to assuming that $X^{(z,r)}$ is close to being locally round, i.e. locally the principal curvatures should be roughly the same. A prototypical example when this is not the case, is an ellipsoid with axes of very different lengths: in this situation the first k $\sigma_i(C^{(z,r)})$'s will have different sizes, depending on the lengths of the axes. Depending on these lengths, and on the scale r, our line of thinking leads to declare that the ellipsoid would have different dimensionalities at different scales. While this may sound strange, it is related to the ill-posedness of estimating the intrinsic dimensionality of a long and thin cigar-shaped manifold: at infinitesimally small scale (i.e. infinitely many points *and* no noise) we would recognize the true dimensionality, but at any given scale r, certain axes are too short to be visible, and the effective dimensionality is lower.

4. The algorithm

We assume that the data is indeed noisy, i.e. $\sigma > 0$, as well as that there exists a hyperplane of dimension $D' < D$ containing $\mathcal{M}$. The results above suggest the following algorithm

(1) **compute** $\sigma_i^{(z,r)}$, for each $z \in \mathcal{M}$, $r > 0$, $i = 1, \ldots, D$.

(2) **estimate the noise size** σ, obtained from the bottom S.V.'s which do not grow with r. Split the S.V.'s into noise S.V.'s and non-noise S.V.'s.

(3) **identify a range of scales** where the noise S.V.'s are small compared to the other S.V.'s.

(4) **estimate**, in the range of scales identified, which S.V.'s, among the non-noise S.V.'s, correspond to tangent directions and which ones correspond to curvatures, by comparing the growth rate as being linear or quadratic in r^2.

4.1 Algorithmic and computational considerations

(1) Instead of computing $\mathrm{cov}(z, r)$ for every z, r, we may perform a randomized subsampling as follows. We select an increasing sequence $0 \leq \delta_0 < \cdots < \delta_j < \ldots$ with $\delta_j \to \infty$, and for every j we construct a δ_j-net, called Γ_j[2]. In the current implementation, we choose the δ_j's as follows: we fix the number of desired scales J, and let $n_j = \frac{(j+1)N}{J}$, for $j = 0, \ldots, J-1$, and let δ_j be the

[2] A set $\Gamma \subset X$ is called a δ-net in X if $\{B_z(2\delta)\}_{z\in\Gamma}$ covers X and $\inf\{d_X(z_1, z_2), z_1, z_2 \in X, z_1 \neq z_2\} \geq \delta$.

smallest radius r such that in average (over $z \in X$) the ball $B_z(r)$ contains at least n_j points. This allows us to quickly go through the scales were there are few points, and have more scales around the distances where lots of points are being added to a ball $B_z(r)$ as r increases. We compute $\{\{\sigma_i^{(z,\delta_j)}\}_{z\in\Gamma_j}\}_{j=0,\ldots,J-1}$. Here i may range from 1 up to $\min\{D, |B_z(r_z)|\}$, the maximum rank of $\mathrm{cov}(z,r)$. In practice, if we have an upper bound K on the intrinsic dimension, we only compute the first K singular values.

This construction yields a discretization of the continuous (in space and scale space) quantities $\sigma_i^{(z,r)}$ [3]. The cost of computing $\{\sigma_i^{(z,r_j)}\}_{i=1,\ldots,I}$ is $O(K \cdot |B_z(r_j)| \cdot I \cdot C_{\mathrm{nn}})$, where C_{nn} is the cost of computing a nearest neighbor, and this is done $O(|\mathcal{M}|/|B_z(r_j)|)$ times (for each $z \in \Gamma_j$), and then across all scales $j = 0,\ldots,J$, with $J = O(\log|\mathcal{M}|)$, for a total cost of $O(K \cdot I \cdot n \log n C_{\mathrm{nn}})$. In the worst case, $I = \min\{K,n\}$, yielding $O(Kn\min\{K,n\}\log n)$.

(2) The noise estimation is performed by identifying the noise eigenvalues at large scales, which are the smallest eigenvalues, are lumped together, and do not grow with r. This identification is easy since we assumed $\sigma > 0$ and that $\mathcal{M}$ is contained in a D'-dimensional subspace, with $D' < D$. This identifies K largest S.V.'s that are not noise S.V.'s.

(3) We identify a range of scales $r > r_0$ where the noise is small compared to the other singular values. For example in Figure 1 we would have $r_0 \sim 1.7$. We also identify a largest r_1 beyond which no singular value grows. In Figure 1 we would have $r_1 \sim 2.1$. From this point on we focus on $r \in (r_0, r_1)$.

(4) We need to determine which of the K largest S.V.'s correspond to tangent singular vectors, and which ones are due to curvature. The tangential ones are expected to grow linearly in r^2, while the curvature ones are expected to grow quadratically and be concave. Therefore we do a linear and quadratic fit to each multiscale S.V. (as a function of the scale r^2), starting from the smallest; if the quadratic fit is significantly better than the linear fit, and is concave, then we declare such a S.V. to be due to curvature, and proceed to the next largest, with a similar analysis.

5. Experimental Results

We have already introduced the set $\mathbb{S}^k(D,n,\sigma)$. We let $\mathbb{Q}^k(D,n,\sigma)$ be n i.i.d. points uniformly sampled from the unit k-dimensional cube embedded in $\mathbb{R}^D$ ($D \geq k$), with added noise $\sigma\mathcal{N}(0,I_D)$. We also introduce $\mathbb{S}h^{k+2}(D,n,\sigma)$, consisting of n i.i.d. points uniformly sampled from a manifold obtained as tensor product of an S-shaped curve (of diameter $O(1)$) with $\mathbb{Q}^k$, embedded in $\mathbb{R}^D$ ($D \geq k+2$), with added noise $\sigma\mathcal{N}(0,I_D)$. Our test sets consists of: $\mathbb{Q}^k(d,n,\sigma)$ with $(k,d) = \{(5,10),(5,100)\}$ and for each such choice, any combination of $n = 500, 100$, $\sigma = 0, 0.01, 0.05, 0.1, 0.2$; $\mathbb{S}^k(d,n,\sigma)$ with $k = \{5,9\}$ and the other parameters as for the cube. We therefore have a portion of a linear manifold (the cube), a manifold with simple curvature (the sphere). We considered several other cases, with more complicated curvatures, with similar results, but we have no space to include and discuss them here. We compare our algorithm with the algorithms of (Haro, Randall, and Sapiro 2008), (Carter and Hero 2008), (Carter, Hero, and Raich 2007), see Figure 2. It is apparent that all the algorithms that we tested against are at the very least extremely sensitive to noise, and in fact they do not seem reliable even in the absence of noise. We are not surprised by such sensitivity, however we did try hard to optimize the parameters involved. First of all we note that our algorithm has no parameters besides $\tilde{X}$. The algorithms we are comparing it to have several parameters (from 4 to 7), which we tuned after rather extensive experimentation, relying on the corresponding papers and commentary in the code. In Figure 2 we report the mean, minimum, and maximum estimated dimensionality by each algorithm, upon varying the parameters of the algorithm. In most cases, this did non improve their performance significantly. Our algorithm, as described in the Algorithm section, is randomized (the multiscale nets are random), and we reported mean, minimum and maximum dimensionality estimates over 5 runs. The tiny variability certainly justifies the multiscale subsampling approach suggested in the Algorithm section. We also ran the TPMM algorithm from (Haro, Randall, and Sapiro 2008), but notwithstanding extensive experimentation we were unable to find any range of parameters (including the defaults in the code) that would give results comparable to the others.

We interpret these experiments, also in light of the theoretical results of (Lee et al. 2009), as a consequence that our algorithm, unlike the others considered, is not volume-based. The estimation of volumes is expected to require a number of samples exponential in the dimensionality of $\mathcal{M}$ (). The results in this paper, and their considerable refinements in (Lee et al. 2009) for the finite sample case, support the intuition that the presented technique only requires a number of points linear in the dimensionality of $\mathcal{M}$.

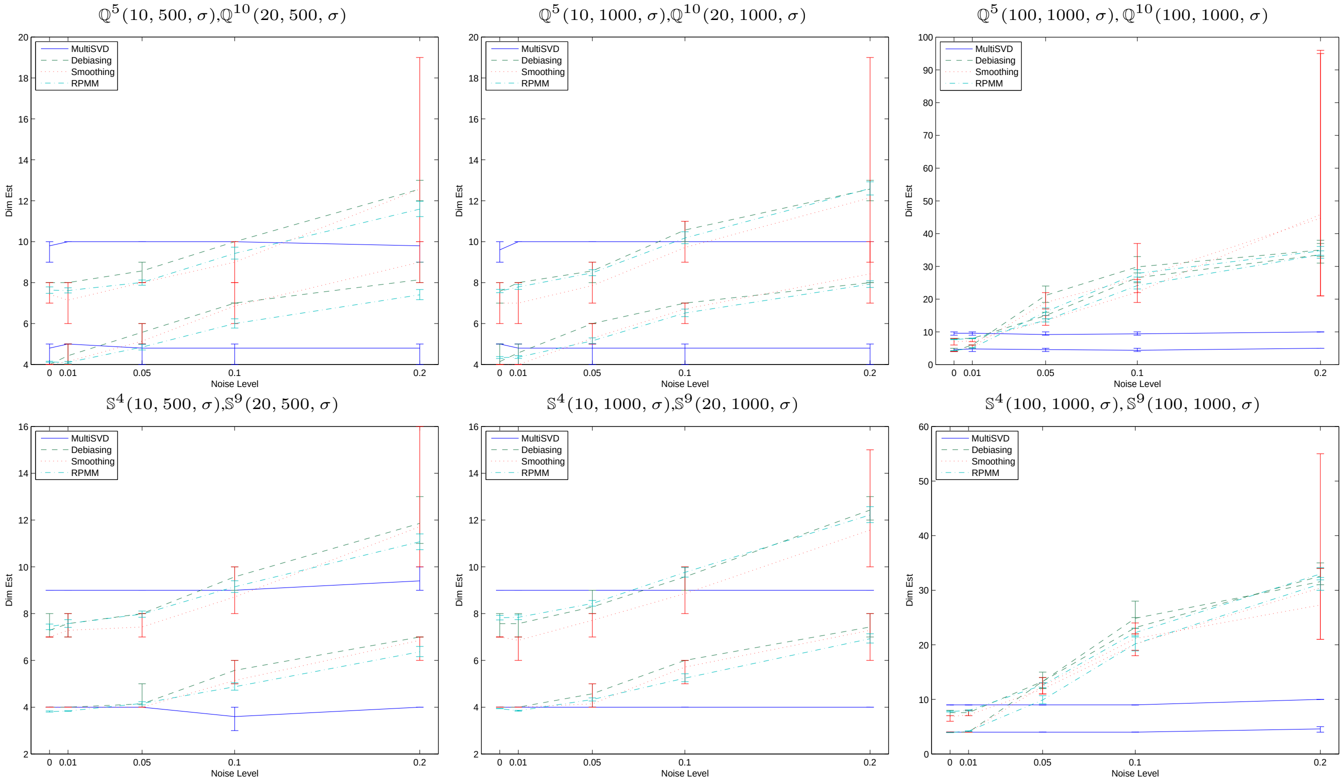
We also ran our algorithm on the MNIST data set: for each digit from 0 to 9 we extracted 2000 random points, applied the algorithm, and projected them onto their top 80 principal components to regularize the search for nearest neighbors. The estimated intrinsic dimensionalities where: $2, 3, 5, 3, 5, 3, 2, 4, 3$. However, the plots of multiscale singular values is not consistent with the hypothesis that the data is generated as a low-dimensional manifold plus noise [4]. In particular, for several digits the intrinsic dimensionality depends on scale.

6. Conclusion and future directions

We presented and analysed an algorithm based on multiscale geometric analysis via principal components, that can very effectively and in a stable way estimate the intrinsic

[3] In order to avoid artifacts due to the randomness of Γ_j, one may repeat this construction a few times and take the expectation, over the runs, of all the quantities we shall be interested in.

[4] Increasing the number of points (or decreasing it, for that matter!) did not produce appreciable changes.



$\mathbb{Q}^5(10, 500, \sigma)$,$\mathbb{Q}^{10}(20, 500, \sigma)$

$\mathbb{Q}^5(10, 1000, \sigma)$,$\mathbb{Q}^{10}(20, 1000, \sigma)$

$\mathbb{Q}^5(100, 1000, \sigma)$, $\mathbb{Q}^{10}(100, 1000, \sigma)$

$\mathbb{S}^4(10, 500, \sigma)$,$\mathbb{S}^9(20, 500, \sigma)$

$\mathbb{S}^4(10, 1000, \sigma)$,$\mathbb{S}^9(20, 1000, \sigma)$

$\mathbb{S}^4(100, 1000, \sigma)$, $\mathbb{S}^9(100, 1000, \sigma)$

Figure 2: Benchmark data sets, comparison between our algorithm and "Debiasing" (Carter, Hero, and Raich 2007), "Smoothing" (Carter and Hero 2008) and RPMM in (Haro, Randall, and Sapiro 2008) (with choice of the several parameters in those algorithm that seem optimal). $\mathbb{Q}^k(D, n, \sigma)$ consists of n points uniformly sampled from the k-dimensional cube embedded in D dimensions, and corrupted by $\eta \sim \sigma\mathcal{N}(0, I_D)$. $\mathbb{S}^k$ is a k-dimensional sphere. We report mean, minimum and maximum values of the output of the algorithms over several realizations of the data and noise. The horizontal axis is the size of the noise, the vertical is the estimated dimension. Even without noise current state-of-art algorithm do not work very well, and when noise is present they are unable to tell the noise from the intrinsic manifold structure. Our algorithm shows great robustness to noise.

dimensionality of point clouds generated by perturbing low-dimensional manifolds by high-dimensional noise. In (Lee et al. 2009) it is shown that the results in expectation in this paper can be transformed into finite sample guarantees of success with high probability. Moreover, under reasonable hypotheses on the manifold $\mathcal{M}$ and the size of the noise, the probability of success is high as soon as the number of samples n is proportional to the intrinsic dimension k of $\mathcal{M}$. It is therefore a manifold-adaptive technique. This is the case even for the pointwise estimation of intrinsic dimensionality.

Future research directions include kernelization, for example by using approximated heat kernels one would obtain an embedding in a feature space where the manifold would have maximally large almost flat pieces (Jones, Maggioni, and Schul 2008), thereby maximizing the probability of detection of the correct dimensionality. The results above seem robust under bi-Lipschitz perturbation, so in particular $\mathcal{M}$ does not need to be smooth. Besides studying this robustness, this suggests that the technique may be combined with random projections (of dimensionality essentially dependent only on k, the dimensionality of $\mathcal{M}$) such as those analyzed in (Baraniuk and Wakin 2009). Our preliminary experiments seem to suggest that this may not be as successful as theory predicts, because of loss of information about the curvature. We are also considering the extension to multiple manifolds of different dimensionality.

In general, we expect multiscale geometric methods to become important in the analysis of high-dimensional data sets.

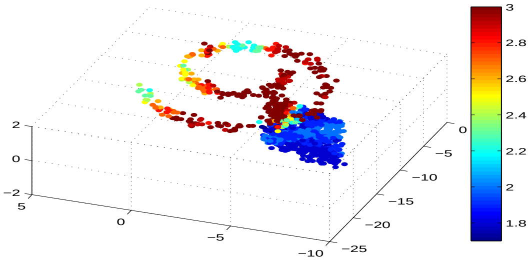


Figure 3: Our algorithm can produce pointwise estimates, without taking advantage of any "smoothness" or clustering property of the local dimension as a function of the point. The data is a very noisy 1-dimensional spiral intersecting a noisy two-dimensional plane. Our algorithm assigns dimensionality 3 to the spiral (because of the noise), dimension 2 to the plane, and again dimension 3 to points at the intersection between the spiral and the plane, as well as to a portion of the plane very close to the spiral. These results are worse, in the sense of classification error for the task of assigning points to the plane or the spiral, than those showed in (Haro, Randall, and Sapiro 2008).

Acknowledgements MM was partially supported by NSF-IIS 0803293, NSF-CHE 0835712, ONR N00014-07-1-0625 and the Sloan Foundation.

7. Appendix

Proof of Lemma 32. The first three inequalities follow from (2), the fourth follows from (Vershynin 2008):

$$\begin{aligned}\mathbb{E}[n^{-1}\sigma_{\max}(N_1^T X)] &= \mathbb{E}\big[\mathbb{E}[n^{-1}\sigma_{\max}(N_1^T X)\big|||X_0||]\big] \\ &\leq 2c_1\sigma\sqrt{kn^{-1}}\,\mathbb{E}[||X_0||]\,,\end{aligned}$$

where we used the fact that X_0 and N_1 are independent. The last inequality follows similarly. □

Proof of Prop. 33. First of all we observe that by the min-max properties of eigenvalues of symmetric matrices, under P_3 the first k S.V.'s increase and the last d S.V.'s decrease. Quite interestingly, this is to our advantage since it increases $\tilde{\Delta}_k$: $\tilde{\Delta}_k \geq \sigma_{\min}(\tilde{C}_1) - \sigma_{\max}(\tilde{C}_2)$, whenever the latter quantity is nonnegative (i.e. on Ω_1).

To prove the first inequality, observe that

$$\begin{aligned}\mathbb{E}\big[\tilde{\Delta}_i\big] &\leq \mathbb{E}\big[\sigma_i(\tilde{C}_1) - \sigma_{i+1}(\tilde{C}_1) + \sigma n^{-1}||\tilde{X}_1^T N_2||\big] \\ &\leq \mathbb{E}\big[\Delta_i + 4\sigma n^{-1}||N_1^T X|| + \sigma^2 n^{-1}||N_1^T N_1|| \\ &\quad + \sigma n^{-1}||\tilde{X}^T N_2||\big] \\ &\leq \Delta_i + 8c_1\sigma_{\max}\sigma\sqrt{kn^{-1}} + \sigma^2(1+\sqrt{kn^{-1}})^2 \\ &\quad + (c_1\sigma_{\max} + c_2\sigma)\cdot\sigma\sqrt{n^{-1}}(\sqrt{k}+\sqrt{d})\end{aligned}$$

where for the first inequality we used the observation above that $\sigma_i(\tilde{C}_1) \geq \sigma_i(\tilde{C})$, and for all the other inequalities we used Lemma 32. To prove the lower bound on $\mathbb{E}[\tilde{\Delta}_k]$ we estimate, using again the observation above about P_3 and Lemma 32:

$$\begin{aligned}\mathbb{E}[\tilde{\Delta}_k] &\geq \sigma_{\min}(\tilde{C}_1) - n^{-1}\sigma_{\max}(N_2^T N_2) \\ &\geq \sigma_k(X_0^T X_0 + n^{-1}N_1^T N_1) - 2\sigma n^{-1}\mathbb{E}\big[||N_1^T X||\big] \\ &\quad - \sigma^2\mathbb{E}\big[n^{-1}||N_2^T N_2||\big] \\ &\geq \sigma_k(C) - 4c_1\sigma_{\max}\sigma\sqrt{kn^{-1}} - \sigma^2(1+\sqrt{dn^{-1}})^2\end{aligned}$$

where the second inequality uses the monotonicity principle of eigenvalues for positive definite matrices, i.e. $\lambda_i(A+E) \geq \lambda_i(A)$ whenever A, E are positive definite (here applied to $A = X^T X$ and $E = n^{-1}N_1^T N_1$).

The estimates for the other gaps follow immediately from the observation above about P_3 and from Lemma 32. □

References

Baik, J., and Silverstein, J. W. 2006. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* 97(6):1382–1408.

Baraniuk, R. G., and Wakin, M. B. 2009. Random projections of smooth manifolds. *Foundations of Computational Mathematics* 9(1):51–77.

Borovkova, S.; Burton, R.; and Dehling, H. 1999. Consistency of the Takens estimator for the correlation dimension. *Ann. Appl. Probab.* 9(2):376–390.

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

Cao, W., and Haralick, R. 2006. Nonlinear manifold clustering by dimensionality. *icpr* 1:920–924.

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

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

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

Donoho, D., and Grimes, C. 2003. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences* 100(10):5591–5596.

Farahmand, A. M., and Audibert, C. S. J.-Y. 2007. Manifold-adaptive dimension estimation. *Proc. I.C.M.L.*

Fukunaga, K., and Olsen, D. 1971. An algorithm for finding intrinsic dimensionality of data. *I.E.E.E. Trans. on Computer* C-20(2).

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

Haro, G.; Randall, G.; and Sapiro, G. 2008. Translated poisson mixture model for stratification learning. *Int. J. Comput. Vision* 80(3):358–374.

Hein, M., and Audibert, Y. 2005. Intrinsic dimensionality estimation of submanifolds in euclidean space. In De Raedt, L., S. W., ed., *ICML Bonn*, 289 – 296.

Johnstone, I. M. 2001. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.*

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

Lee, J.; Little, A.; Maggioni, M.; and Rosasco, L. 2009. Multiscale estimation of intrinsic dimensionality of point clouds and data sets. *in preparation.*

Levina, E., and Bickel, P. 2005. Maximum likelihood estimation of intrinsic dimension. *In Advances in NIPS 17,Vancouver, Canada.*

Paul, D. 2007. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* 17:1617–1642.

P.W.Jones. 1990. Rectifiable sets and the traveling salesman problem. *Inventiones Mathematicae* 102:1–15.

Raginsky, M., and Lazebnik, S. 2005. Estimation of intrinsic dimensionality using high-rate vector quantization. *Proc. NIPS* 1105–1112.

Rudelson, M., and Vershynin, R. 2008. The least singular value of a random square matrix is $O(n^{-1/2})$. *Comptes rendus de l'Acadmie des sciences - Mathmatique* 346:893–896.

Silverstein, J. 2007. On the empirical distribution of eigenvalues of large dimensional information-plus-noise type matrices. *Journal of Multivariate Analysis* 98:678–694.

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

Vershynin, R. 2008. Spectral norm of products of random and deterministic matrices. Submitted.