ESTIMATION OF INTRINSIC DIMENSIONALITY OF SAMPLES FROM NOISY LOW-DIMENSIONAL MANIFOLDS IN HIGH DIMENSIONS WITH MULTISCALE SVD

Anna V. Little, Jason Lee, Yoon-Mo Jung, Mauro Maggioni

Department of Mathematics, Duke University, P.O. Box 90320, Durham, NC, 27708, U.S.A.

ABSTRACT

The problem of estimating the intrinsic dimensionality of certain point clouds is of interest in many applications in statistics and analysis of high-dimensional data sets. Our setting is the following: the points are sampled from a manifold $\mathcal{M}$ of dimension k, embedded in $\mathbb{R}^D$, with $k \ll D$, and corrupted by D-dimensional noise. When $\mathcal{M}$ is a linear manifold (hyperplane), one may analyse this situation by SVD, hoping the noise would perturb the rank k covariance matrix. When $\mathcal{M}$ is a nonlinear manifold, SVD performed globally may dramatically overestimate the intrinsic dimensionality. We discuss a multiscale version SVD that is useful in estimating the intrinsic dimensionality of nonlinear manifolds.

***Index Terms*—** Multiscale analysis, intrinsic dimensionality, high dimensional data, manifolds, point clouds, sample covariance, SVD, PCA.

1. INTRODUCTION

The problem of estimating the intrinsic dimensionality of a point cloud is of interest in a wide variety of situations, for example in determining the number of variables in a linear model in statistics, or in determining the number of degrees of freedom of a dynamical system. Moreover, many algorithms in machine learning (and manifold learning in particular), require the intrinsic dimensionality of the data as a crucial input parameter. When data $\{\mathbf{x}_i\}_{i=1}^n$ is assumed to lie on a k-dimensional hyperplane in $\mathbb{R}^D$, a standard technique is to use the Singular Value Decomposition (SVD, called Principal Component Analysis, or PCA, in the statistical literature), since the number of non-zero singular values equals k, the intrinsic dimension of the data (and the rank of the covariance matrix). When D-dimensional noise is added to the data, so that we observe $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \eta_i$, where η represents noise, independent of $\mathbf{x}$, the noise will induce perturbation of the covariance matrix of the data, which will not have rank k anymore. It may be expected, nevertheless, that if the noise is small, to have a k-th singular value significantly larger than the $(k+1)$-st. At least when n, the number of samples, tends to infinity, the behavior of this estimator is quite well understood (see e.g., out of many works, [1], [**?**] and references therein). However, the finite-sample situation is less well understood, and so is the the situation in which we are really interested in, which is that of data having geometric structures more complicate than linear. A recent 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 [2] [3], [4], [5], [6], [7], and references therein.

MM is grateful for partial support from NSF (DMS 0650413, CCF 0808847, IIS 0803293), ONR N00014-07-1-0625, the Sloan Foundation and Duke. YMJ was supported by ONR. AVL was supported by ONR and DMS IIS. JL was supported by the PRUV program at Duke.

2. MULTISCALE SVD FOR DIMENSIONALITY ESTIMATION

2.1. Setup

We consider the following generative stochastic geometric model. Let $(\mathcal{M}, g)$ be a compact k-dimensional Riemannian manifold (isometrically embedded) in $\mathbb{R}^D$. Let the noise be $\eta \sim \sigma\mathcal{N}(0, I_D)$ (because of certain universality phenoma, the model for the noise may be much more general [10, 11]). Let $X_n = \{\mathbf{x}_i\}_{i=1}^n$ be a set of independent random samples on $\mathcal{M}$, uniform with respect to the natural volume measure on $\mathcal{M}$, or a measure having smooth density with respect to the volume measure, bounded above and below from 0. We observe $\tilde{X_n} = \{\mathbf{x}_i + \sigma\eta_i\}_{i=1}^n$, where η_i are i.i.d. samples from η, and $\sigma > 0$. Alternatively, we may of these points as being 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 a $n \times D$ matrix X_n, whose (i,j) entry is the j-th coordinate of the i-th point. In particular X_n and $\tilde{X_n}$ 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 concern ourselves with is to **estimate** $k = \dim \mathcal{M}$**, given** $\tilde{X}$**.**

2.2. Singular Value Decomposition for Linear Manifolds

When $\mathcal{M}$ is a hyperplane and η is Gaussian noise, the standard approach is to compute the SVD of X_n, with the singular values yielding k. Assuming centered data, $\mathrm{cov}(X_n) = \frac{1}{n}X_n^T X_n$ is the $D \times D$ covariance matrix of X_n, and $\Sigma(X_n) = (\sigma_i)_{i=1}^D$ are the singular values (s.v.) squared of $n^{-1/2}X_n$. At least for $n \gtrsim k$, with high probability (w.h.p.) exactly

the first k S.V. are non-zero. We can think of $\tilde{X}_n$ as a random perturbation of X_n and expect $\Sigma(\tilde{X}_n) \cong \Sigma(X_n)$, so that $\sigma_1, \ldots, \sigma_k \gg \sigma_{k+1}, \ldots, \sigma_D$, allowing 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*, the above approach will not work. Curvature will in general force the dimensionality of a best-approximating hyperplane to the whole manifold to be much larger than the intrinsic dimensionality: for example for a planar circle ($k = 1$) rigidly embedded in $\mathbb{R}^D$, $\mathrm{cov}(X_n)$ has exactly 2 nonzero eigenvalues equal to the radius squared; more generally, one may construct a one-dimensional manifold ($k = 1$) such that $\mathrm{cov}(X_n)$ has full rank (w.h.p.): it is enough to pick a curve that spirals out in all dimensions.

2.3. Multiscale SVD, curvature, and noise

The above failures are a consequence of performing PCA globally - so we shall think locally. Let $z \in \mathcal{M}$, r a radius (scale parameter), and consider $X^{(z,r)} := B_z(r) \cap \mathcal{M}$ (the ball is in $\mathbb{R}^D$). For r small enough, $X^{(z,r)}$ is well-approximated (up to second order in r) by a portion of the k-dimensional tangent plane $T_z(\mathcal{M})$: therefore $\mathrm{cov}(X^{(z,r)})$ to have k large eigenvalues and possibly other smaller eigenvalues caused by curvature. As $r \to 0$, i.e. *choosing r_z small enough dependent on curvature*, these smaller eigenvalues will tend to 0 faster than the top k eigenvalues of size $O(r)$. Therefore, if we were given X_n, in the limit as $n \to \infty$ and $r_z \to 0$, this would give a consistent estimator of k. Because of sampling, we need r_z to approach 0 slow enough so that $B_z(r)$ contains, with high enough probability, at least (conservatively) $O(k)$ points. This can be made rigorous by using results on covariance matrix approximation (see [10, 11] for details). It is remarkable already, though, that we only require a number of samples *linear* (not exponential, as we would with volume estimators such as those in [2] [3], [4], [5], [6], [7], [8], [9]) in the intrinsic dimension. In practice, since the curvature is not given to us, we shall need to select many values of r and perform the SVD analysis on $X^{(z,r)}$ for all these values, from which the name Multiscale SVD. It is related to constructions used in deep results in geometric measure theory [12].

Since we observe $\tilde{X}$, we meet another constraint, that forces to take r_z not 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 η (for example, $r \sim \sigma\sqrt{D}$ when $\eta \sim \sigma\mathcal{N}(0, I_D)$), and $B_z(r)$ contains enough points (say, $O(k)$), $\mathrm{cov}(\tilde{X}^{(z,r)})$ 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* $T_z(\mathcal{M})$. Once again, σ and the curvature of $\mathcal{M}$ are unknown, which 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) [12]. We shall use the behavior of $\sigma_i^{(z,r)}$ as a function of i, z, r to estimate the intrinsic dimensionality k.

3. THE ALGORITHM

The reasonings above suggest the following algorithm: for each $z \in \mathcal{M}$, $r > 0$, $i = 1, \ldots, D$, we compute $\sigma_i^{(z,r)}$. 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, in the regime for the values of K, D, σ, n suggested by our results. Clearly, $k \leq K$. Moreover, $\{\sigma_i^{(z,r)}\}_{i=K+1,\ldots,D}$ will be highly concentrated and we use them to estimate σ, which is useful per se. By viewing $\{\sigma_i^{(z,r)}\}_{i=K+1,\ldots,D}$, we identify an interval in r where the noise is almost flat, i.e. we remove the small scales where the distortion due to noise dominates.

We look at the first $\{\sigma_i^{(z,r)}\}_{i=1,\ldots,K}$, and the goal is to decide how many of them are due to the extrinsic curvature of $\mathcal{M}$. But the curvature S.V.'s grow quadratically w.r.t. the "tangential" (non-curvature) S.V.'s: a best least-square linear and quadratic fit to $\sigma_i^{(z,r)}$, as a function of r, is enough to tell the curvature S.V.'s from the tangential S.V.'s. The analysis in [10] shows that for succeeding w.h.p. the algorithm only requires a $n \gtrsim O(k \log k)$ and a certain natural "geometric signal to noise ratio" condition, involving the curvature of the embedding of $\mathcal{M}$ and the size of the noise, to be satisfied.

4. EXAMPLES AND EXPERIMENTS

4.1. Example: k-dimensional sphere in $\mathbb{R}^d$, with noise

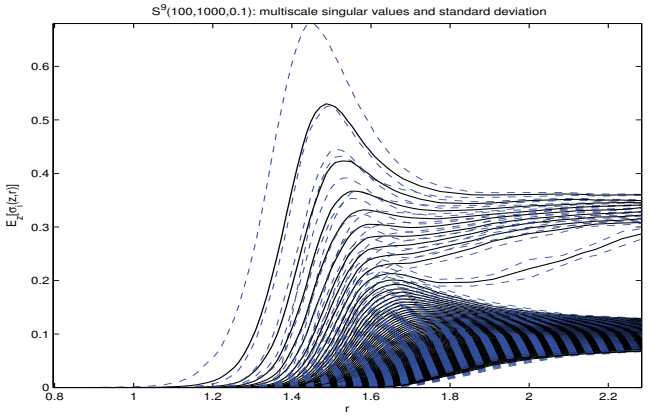

Fig. 1. $\mathbb{S}^9(100, 1000, 0.1)$: $\mathbb{E}_z[\sigma_i^{(z,r)}]$ (and std bands) as a function of r. Top 9 s.v.'s correspond to the intrinsic dimensions; the 10-th s.v.(increasing with scale) corresponds to curvature; the remaining s.v.'s correspond to noise in the remaining 90 dimensions, tending to σ.

We start with a simple yet perhaps surprising example, in order to build our intuition. Let $\mathbb{S}^k = \{x \in \mathbb{R}^{k+1} : ||x||_2 = 1\}$ be the unit sphere in $\mathbb{R}^{k+1}$, with $\dim(\mathbb{S}^k) = k$. We embed $\mathbb{S}^k$ in $\mathbb{R}^{k+1}$ and then in $\mathbb{R}^D$ in $\mathbb{R}^D$ by fixing the first $k+1$ coordinates. We sample n points uniformly at random on $\mathbb{S}^k$ to obtain X_n, and $\tilde{X}_n$ is obtained by adding D-dimensional white Gaussian noise of variance σ in every direction. We denote this data set by $\mathbb{S}^k(D, N, \sigma)$. In Figure 1 we consider the multiscale s.v.'s values of $\mathbb{S}^9(100, 1000, 0.1)$, as a function of r. Since $\mathbb{R}^{10}$ is partitioned 2^{10} orthants, by sampling 1000 points on $\mathbb{S}^9$ we obtain about 1 point per sector (!). Moreover, observe that the noise size, if measured by $||x_i - \tilde{x}_i||_2^2$, i.e. by how much each point is displaced, is of order $\mathbb{E}[\sigma^2 \chi^2_D] \sim 1$, which is comparable with the radius of the sphere itself (!). In light of these observations, estimating the dimensionality of this data set may seem hopeless.

In fact, we can detect reliably the intrinsic dimensionality of $\mathcal{M}$. From Figure 1 we see that 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 slightly larger, but still small scales, no gap among the $\sigma_i^{(z,r)}$ is visible: $B_z(r_j)$ contains too few points, scattered in all directions by the noise, and new increasing S.V.'s keep arising for several scales. At even larger scales, the top $9 = k$ S.V.'s start to separate from the others: we interpret this as the noisy tangent space being detected. Finally, at even larger scales, the curvature starts affecting the covariance, as indicated by the slowly growing 10th S.V., while the remaining smaller S.V.'s tend approximately to the *one-dimensional* noise variance: this is the size of the noise relevant in our procedure, rather than the much large expected displacement measured in the full $\mathbb{R}^D$, which was of size $O(1)$.

4.2. Increasing the ambient dimensionality D

We consider the following scaling limit as $D \to +\infty$. We fix the intrinsic dimensionality k and the number of sampled points n, and we scale the noise η as a function of D by letting $\sigma = \sigma(D) = \sigma_0 D^{-\frac{1}{2}}$. This is the natural scaling since in this way $\mathbb{E}[||\eta||_D^2] = \sigma_0^2$, independently of D, and it is also consistent in the limit with infinite dimensional Brownian motion.

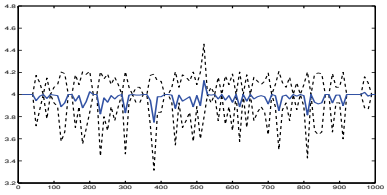

Fig. 2. $\mathbb{S}^5(D, 200, 0.05D^{-\frac{1}{2}})$, for D as in the horizontal axis. We estimate the dimensionality pointwise, and return the mean estimate (without rounding) and its standard deviation. The estimate is correct for all dimensions, with similar standard deviations.

In this scaling limit, our algorithm is able to recover the correct dimensionality independently of D. In fact, in order to estimate accurately the covariance of the data it needs a number of points proportional to k (in this case, fixed), and geometrical assumptions on the size of the noise in the ambient space compared to the k-th singular value of the covariance of the data. These two latter components are independent of D in this scaling limit, and therefore the success of the algorithm is independent of the ambient dimension.

4.3. Increasing the intrinsic dimensionality k

In order to probe the dependency of our algorithm on the intrinsic dimensionality k, we perform the following experiment. We fix the ambient dimension $D = 100$, the size of the ambient noise $\sigma = 0.05$, so that the noise is $0.05\mathcal{N}(0, I_{100})$, and consider the k-dimensional sphere $\mathbb{S}^9$ embedded in $\mathbb{R}^D$, for $k = 5, \ldots, 20$. Since our algorithm requires a number of points n roughly linear in the intrinsic dimensionality, for each k we let $n = n(k) = 50 \cdot k$. The results of the experiment are in Figure 3. We see that the algorithm performs perfectly till $k = 13$ (this number depends on the ratio $n(k)/k$), and the starts underestimating the intrinsic dimensionality: albeit the true intrinsic dimensionality keeps growing linearly, our algorithm starts estimating roughly the same dimensionality. For even larger k (nor shown) the estimated dimensionality actually starts decreasing (!). The same phenomenon happens if we considered $\mathbb{B}^{k-1}$ instead of $\mathbb{S}^k$, it is independent of possible issues due to curvature. The root cause of this is a phenomenon has to do with the fact that the covariance matrix of the points on $\mathbb{S}^k$ (or $\mathbb{B}^{k-1}$ for that matter) has norm that scales in k as k^{-1}, and since the singular values are all equal (and approximately equal due to sampling), the k-th singular value of the data behaves as $k^{-\frac{1}{2}}$. But then, it becomes harder and harder, as k increases, to tell apart the k-th singular value of the data from the first singular value of the noise, which is constant and equal to σ. This is related to the idea of observable diameter introduced by Gromov (see [13] and references therein).

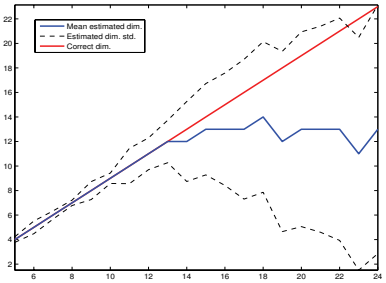

Fig. 3. $\mathbb{S}^k(100, 50k, 0.05)$, for k as in the horizontal axis. We estimate the dimensionality pointwise, and return the mean estimate and its standard deviation, as well as the true intrinsic dimensionality.

4.4. Short comparison with other algorithms

We have already introduced the set $\mathbb{S}^k(D, N, \sigma)$. Our test sets consists of $\mathbb{S}^k(d, 1000, \sigma)$ with $(k, d) = \{(5, 10), (5, 100)\}$ and $\sigma = 0, 0.01, 0.05, 0.1, 0.2$. We considered several other cases, with more complicate curvatures, but we have no space to include and discuss them here [10, 11]. We compare our

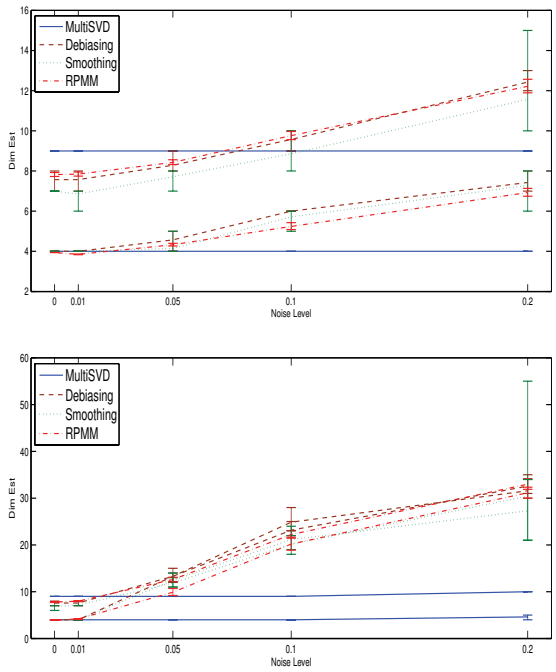

Fig. 4. Benchmark data sets, comparison between our algorithm and "Debiasing" [5], "Smoothing" [4] and RPMM in [3] (with choice of the several parameters in those algorithm that seem optimal). Top: $\mathbb{S}^4(10,1000,\sigma)$,$\mathbb{S}^9(20,1000,\sigma)$; bottom: $\mathbb{S}^4(100,1000,\sigma)$, $\mathbb{S}^9(100,1000,\sigma)$. 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; the correct dimension is 4 and 9 respectively. 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 (blue curve) shows great robustness to noise.

algorithm with the algorithms of [3], [4], [5], see Figure 4. 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 (often many!) parameters involved. In Figure 4 we report the mean, mininum and maximum estimated dimensionality by each algorithm, upon varying the parameters of the algorithm. In most cases, this did non improve their performance significantly. For more extensive comparisons, that include different manifolds, see [10, 11].

5. CONCLUSION AND FUTURE DIRECTIONS

We presented a promising algorithm based on multiscale geometric analysis via singular value decompositions at different scales. The crucial observation is that our algorithm requires only $O(k)$ points for a noiseless manifold since it is based on fitting an approximate plane to the manifold, which even intuitively should only required $O(1)$ points per dimension, compared to the $O(2^k)$ points required by any algorithm based on volume considerations. When noise is added, volume can behave very differently from r^k, and the multiscale approximate plane fitting procedure we use is much less sensitive to noise than volume-based algorithms. Future research directions include kernelization, in particular the use of heat kernels [14] and pointwise estimates of dimensionality, which are important in several applications where multiple manifolds of different dimensions arise.

6. REFERENCES

[1] Iain M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Stat.*, vol. 29, no. 2, pp. 295–327, April 2001.

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

[3] G. Haro, G. Randall, and G. Sapiro, "Translated poisson mixture model for stratification learning," *Int. J. Comput. Vision*, vol. 80, no. 3, pp. 358–374, 2008.

[4] K.M. Carter and A.O. Hero, "Variance reduction with neighborhood smoothing for local intrinsic dimension estimation," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3917–3920, 31 2008-April 4 2008.

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

[6] J.A. Costa and A.O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2210–2221, Aug. 2004.

[7] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE P.A.M.I.*, vol. 24, no. 10, pp. 1404–10, 2002.

[8] Wenbo Cao and Robert Haralick, "Nonlinear manifold clustering by dimensionality," *icpr*, vol. 1, pp. 920–924, 2006.

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

[10] Y.-M. Jung, J. Lee, A.V. Little, M. Maggioni, and L. Rosasco, "Multiscale estimation of intrinsic dimensionality of point clouds ands data sets," *in preparation*, 2009.

[11] A.V. Little, Y.-M. Jung, and M. Maggioni, "Multiscale estimation of intrinsic dimensionality of data sets," *to appear in Proc. A.A.A.I.*, 2009.

[12] P.W.Jones, "Rectifiable sets and the traveling salesman problem," *Inventiones Mathematicae*, vol. 102, pp. 1–15, 1990.

[13] Ledoux, *The Concentration of Measure Phenomenon*, Amer. Math. Soc., 2005.

[14] P.W. Jones, M. Maggioni, and R. Schul, "Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels," *Proc. Nat. Acad. Sci.*, vol. 105, no. 6, pp. 1803–1808, Feb. 2008.