MULTISCALE GEOMETRIC METHODS FOR ESTIMATING INTRINSIC DIMENSION

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

[1,2]Department of Mathematics and [2]Computer Science, Duke University
[3]Center for Biological and Computational Learning, MIT
Emails: {avl, mauro}@math.duke.edu, lrosasco@mit.edu

ABSTRACT

We present a novel approach for estimating the intrinsic dimension 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 PCA: 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 PCA may dramatically overestimate the intrinsic dimension. We discuss a multiscale version of PCA and how one can extract estimators for the intrinsic dimension that are highly robust to noise, and we derive some of their finite-sample-size properties.

***Keywords*—** Dimension estimation, multiscale analysis, geometric measure theory, point cloud data

1. INTRODUCTION

We are interested in developing tools for the quantitative analysis of the geometry of point cloud data, in the situation where such point clouds arise as random samples from a low-dimensional set embedded in a high-dimensional space and corrupted by high-dimensional noise. Our main motivation arises from the need to analyze large, high-dimensional data sets that arise in a wide variety of applications. These data sets are often modeled as low-dimensional sets embedded in high-dimensional, typically Euclidean, space. Our approach for estimating the intrinsic dimension of a point cloud is a variation of classical ideas in multiscale geometric measure theory [8], especially at its intersection with harmonic analysis.

The problem of estimating the intrinsic dimension of a point cloud is of interest in a wide variety of situations. In fact, to cite some important instances, it is equivalent to estimating: the number of variables in a linear model in statistics (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, for example spectrometry, signal processing, genomics and economics, to name only a few. Moreover, many manifold learning algorithms assume that the intrinsic dimension is given.

When the data is generated by a multivariate linear model, principal component analysis (PCA) may be used to recover both the dimension of the data and the subspace of $\mathbb{R}^D$ which contains the data, and it requires a number of samples essentially linear in the intrinsic dimension. This situation is well understood, even when the data is corrupted by noise. However, when the data is nonlinear, PCA fails, curvature causing PCA to overestimate the intrinsic dimension. Several volume-based algorithms have been proposed to estimate intrinsic dimension of nonlinear data when the data lies on a low-dimensional manifold; see [13] for a complete description; [2] also provides a survey of many proposed estimation techniques. As these volume-based algorithms are based on empirical estimates of the number of samples obtained in a local neighborhood, in general they require a number of (local) sample points exponential in the intrinsic dimension and tend to be highly sensitive to noise.

The papers [9, 1] first attempted to estimate the intrinsic dimension by partitioning the data into small neighborhoods and applying PCA locally within each neighborhood; although this technique takes advantage of the essentially linear sample size requirements of PCA, it has met with limited success in comparison with the volume-based techniques [15]. We build on this technique of applying PCA locally, but instead of the fixed-scale approach of [9, 1], we propose a multiscale approach, since determining a "good" range of scales at which the estimate is reliable is a key aspect to the problem. Our results, in addition to being highly robust to noise, show that as soon as some rather mild geometric regularity is assumed on the data, a number of (local) samples essentially linear in the intrinsic dimension suffices.

2. SETTING AND MAIN RESULT

Definitions. We define the covariance and the empirical covariance of a random variable X from which we have drawn n

MM is greteful for support from NSF (DMS 0650413, CCF 0808847, IIS 0803293), ONR N00014-07-1-0625 and the Sloan Foundation. AVL is partially supported by NSF DMS 0650413 and ONR N00014-07-1-0625.

samples $x_1, \ldots, x_n$ as:

$$\mathrm{cov}(X) = \mathbb{E}[(X - \mathbb{E}[X]) \otimes (X - \mathbb{E}[X])]$$
$$\mathrm{cov}(X_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}_n[X]) \otimes (x_i - \mathbb{E}_n[X])\,,$$

where $\mathbb{E}_n[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$. We let $\lambda_i(\mathrm{cov}(X))$ denote the i^{th} eigenvalue of the operator $\mathrm{cov}(X)$, sorted in decreasing order, and define the k^{th} gap to be:

$$\Delta_k(\mathrm{cov}(X)) = \lambda_k(\mathrm{cov}(X)) - \lambda_{k+1}(\mathrm{cov}(X))\,.$$

The i^{th} singular value (S.V.) of X is $\sqrt{\lambda_i(\mathrm{cov}(X))}$.

Set-up. Let X and N be two random variables, defined on a probability space $(\Omega, \mathbb{P})$, taking values in $\mathbb{R}^D$. The random variable N represents noise: for example $N \sim \sigma\mathcal{N}(0, I_D)$, where $\mathcal{N}(\mu, \Sigma)$ denotes the Gaussian distribution with mean μ and covariance Σ. Let μ_X be the distribution of X, and $\mathcal{M} = \operatorname{supp}\mu_X$. Our observations will be points $\tilde{x}_1, \ldots, \tilde{x}_n$ drawn in the form

$$\tilde{x}_i = \mathbf{x}_i + \sigma\eta_i, \quad \sigma > 0, \tag{1}$$

where $x_1, \ldots, x_n$ are drawn i.i.d. from X and $\sigma\eta_1, \ldots, \sigma\eta_n$ are drawn i.i.d. from N. If we let $\tilde{X} = X + N$, we can think of the data as random draws from n identical and independent copies of $\tilde{X}$. With some abuse of notation, we let X_n denote both the set of samples and the $n \times D$ matrix whose rows are the n samples $x_1, \ldots, x_n$; similarly for $\tilde{X}_n$. We fix a center $\tilde{z} \in \{\tilde{x}_1, \ldots, \tilde{x}_n\}$ and let $B_{\tilde{z}}(r)$ be the Euclidean ball (in $\mathbb{R}^D$) centered at $\tilde{z}$ with radius r. We let $\widetilde{X_{n,\tilde{z},r}} = \tilde{X}_n \cap B_{\tilde{z}}(r)$ be the noisy data intersected with a local ball centered at one of the noisy data points. Finally, we define $X_{z,r}$ to be the random variable X conditioned on taking values in $\mathcal{M} \cap B_z(r)$.

Our goal is to estimate k, the intrinsic dimension of X at z, given the noisy samples $\{\tilde{x}_1, \ldots, \tilde{x}_n\}$. We wish to determine a good range of scales for r, so that $\mathrm{cov}(X_{z,r})$ will have k large eigenvalues and $D - k$ small eigenvalues, allowing us to correctly estimate k. However, we meet several constraints:

(i) **curvature**: for r small enough, $X_{z,r}$ is well-approximated in the least squares sense by a portion of the k-dimensional tangent plane $T_z(\mathcal{M})$: $\mathrm{cov}(X_{z,r})$ will have k large eigenvalues and smaller eigenvalues caused by curvature. By letting $r \to 0$, i.e. *choosing r small enough dependent on curvature*, these smaller 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 to have enough samples in $X_n \cap B_z(r)$ 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 $\widetilde{X_{n,\tilde{z},r}}$, we meet a *noise* constraint, that forces us to take r above the "scale" of the noise, i.e. not too small, otherwise $\mathrm{cov}(X_{z,r})$ would be too affected by the covariance structure of the noise, instead of that of $X_{z,r}$.

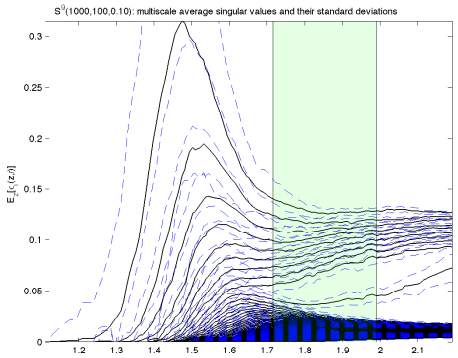

Fig. 1. Plot of $\mathbb{E}_z[\sqrt{\lambda_i^{(z,r)}}]$ (the S.V.'s averaged over the samples), as a function of r, for 1000 noisy samples ($\sigma = .1$) of $\mathbb{S}^9$.

2.1. Assumptions on the Geometry

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

$$\begin{aligned} {X_{z,r}}^{||} &= P^{(z,r)}X_{z,r} &&=: P^{(z,r)}_{\mathcal{M}}(X_{z,r}) \\ {X_{z,r}}^{\perp} &= (I - P^{(z,r)})X_{z,r} &&=: P^{(z,r)}_{\mathcal{M}^\perp}(X_{z,r}) \end{aligned}$$

then the following conditions hold, for all $r \in (R_{\min}, R_{\max})$ and for some choice of positive $\lambda_{\min}, v_{\min}, v_{\max}, \kappa$:

$$\lambda(\mathrm{cov}({X_{z,r}}^{||})) \subseteq k^{-1}[\lambda^2_{\min}, \lambda^2_{\max}]r^2$$
$$||||{X_{z,r}}^{\perp}||_{\mathbb{R}^D}||^2_{\psi_2} \leq k^{-1}\kappa^2 r^4$$
$$\mu_X(B_z(r)) \in v_z(r)\mu_{\mathbb{R}^k}(\mathbb{B}^k)\left(r^2 - ||z - z_{\mathcal{M}}||^2\right)^{\frac{k}{2}}\,,$$

for $z \in \mathbb{R}^D$ such that $||z - z_{\mathcal{M}}||^2 \leq \sigma^2 D$, where $z_{\mathcal{M}}$ is the closest point on $\mathcal{M}$ to z, with $v_z(r)$ smooth and $v_z(r) \in [v_{\min}, v_{\max}]$ for $r \in (R_{\min}, R_{\max})$. Here $\lambda(A)$ is the spectrum of A, $\mu_{\mathbb{R}^k}$ is k-dimensional Lebesgue measure, and $||Z||_{\psi_2}$ denotes the 2-Orlicz norm defined as $||Z||_{\psi_2} = \inf\{c > 0 : \mathbb{E}\left[\exp\left(\frac{|Z|^2}{c^2}\right)\right] < 2\}$.

2.2. Assumptions on the Noise

(i) N has mean 0 and is independent of X and (ii) for $\sigma\eta \sim N$, the random vector $\eta = (\eta_1, \ldots, \eta_D)$ has independent coordinates with subgaussian moment 1, that is, $\mathbb{P}(|\eta_i| > t) \leq 2e^{-t^2}$.

2.3. Main result

Theorem 2.1 ($\mathbf{D \to \infty}$, $\sigma = \frac{\sigma_0}{\sqrt{\mathbf{D}}}$)**.** *Under the above assumptions, for $\lambda = \lambda_{\min} = \lambda_{\max}$ and D large enough, if*

$$\sigma = \frac{\sigma_0}{\sqrt{D}}\,, \text{ with } \frac{\sigma}{r} \in \{0\} \cup \left[\frac{\sqrt{\log D}}{D}, \infty\right),$$

then $\Delta_k(\mathrm{cov}(\widetilde{X_{n,\tilde{z},r}}))$ is the largest gap of $\mathrm{cov}(\widetilde{X_{n,\tilde{z},r}})$ w.h.p. in the range of scales

$$R^2_{\min} \vee \frac{{\sigma_0}^2}{\log k} \vee \left(\frac{c_1 k \log k}{\lambda^2 n v_{min}\mu_{\mathbb{R}^k}(\mathbb{B}^k)}\right)^{\frac{2}{k}} < r^2 - 2{\sigma_0}^2 \leq \frac{\lambda^2}{c_2\kappa^2} \wedge (R_{\max} - 2\sigma_0)^2\,.$$

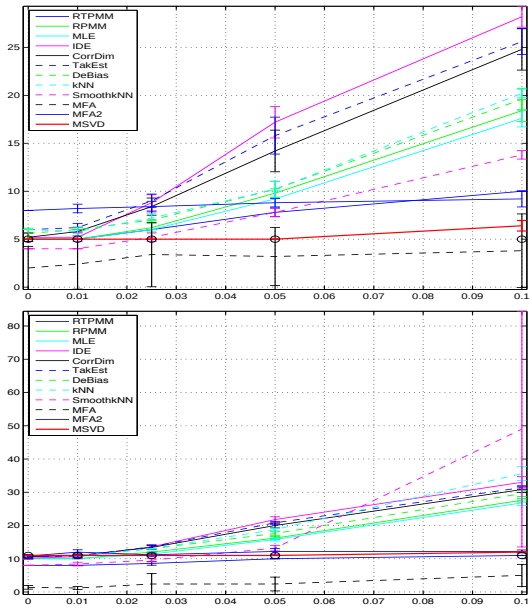

Fig. 2. Estimated intrinsic dimension as a function of the noise level σ; top: $\mathbb{S}^5$ with $n = 250$ and $D = 100$; bottom: $\mathbb{S}^{11}$ with $n = 1000$ and $D = 100$.

For $R_{\min}$ (resp., $R_{\max}$) smaller (resp., larger) than the terms it is compared to, this interval is non-empty as soon as

$$n \gtrsim \frac{k \log k}{\lambda^2} \left(\frac{c_2 \kappa}{\lambda}\right)^k \frac{1}{v_{\min} \mu_{\mathbb{R}^k}(\mathbb{B}^k)} \quad \textit{and} \quad \sigma_0 \lesssim \frac{\lambda \sqrt{\log k}}{\kappa}.$$

3. MSVD ALGORITHM

The above results suggest the following algorithm: for each $z \in \mathcal{M}$, $r > 0$, $i = 1, \ldots, D$, we compute $\lambda_i^{(z,r)} := \lambda_i(\widetilde{\mathrm{cov}(X_{n,z,r})})$. 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, $\{\lambda_i^{(z,r)}\}_{i=K+1,\ldots,D}$ will be highly concentrated and we use them to estimate σ, which is useful per se. Viewing $\{\lambda_i^{(z,r)}\}_{i=K+1,\ldots,D}$ 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 $\{\lambda_i^{(z,r)}\}_{i=1,\ldots,K}$, and the goal is to decide how many of them are due to the extrinsic curvature of $\mathcal{M}$. But the curvature squared S.V.'s grow with rate at most r^4, while the "tangential" (non-curvature) squared S.V.'s grow with rate r^2: a least-square fit to $\lambda_i^{(z,r)}$, as a function of r, is used to tell the curvature squared 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's in which $\Delta_{\hat{k}}^{(z,r)}$ is the largest gap. See Fig. 1 for an example plot of how the S.V.'s grow as a function of scale.

The many details and available options are documented in the MATLAB code available at `www.math.duke.edu/~mauro`; the code includes a User Interface for navigating the multiscale S.V.'s. The computational cost of the algorithm is no worse than $O(C_{\mathrm{nn}} D n \min\{D, n, K\} \log n)$, where C_{nn} is the cost of computing a nearest neighbor.

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

Fig. 3. Dimension estimates for various manifolds; $n = 1000$ and $\sigma = 0$.

4. EXPERIMENTS

Manifold data. 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 $\mathbb{Q}^k$, $\mathbb{S}^k$, $\mathcal{S}$, $\mathcal{Z}^k$ 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 Meyer's staircase $\{\chi_{0,k}(\cdot - l)\}_{l=0,\ldots,D}$. 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 $\eta \sim \sigma \mathcal{N}(0, I_D)$ is added.

We consider a range of values for k, σ and n. We compare our algorithm against "Debiasing" [3], "Smoothing" [4] and RPMM in [10], "MLE" [12], "kNN" [7], "MFA" [6], and "MFA2" [5]. For each combination of the parameters, we generate 5 realizations of the data set and report the most frequent (integral) dimension returned by each algorithm, as well as the standard deviation of the estimated dimension. We attempted to optimize the parameters of the competing algorithms by running them on several training examples and then fixing the required parameters. See [13] for a complete comparison. Fig. 2 shows the results for $\mathbb{S}^5$ (resp. $\mathbb{S}^{11}$) with ambient dimension $D = 100$ and $n = 250$ (resp. $n = 1000$) samples, as the noise level σ is increased. All other results look qualitatively similar to these. Fig. 3 contains the dimension estimates for various manifolds in the 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. All algorithms exhibit a similar behavior, both with and without noise, except for "MFA" and "MFA2," which perform uniformly poorly but are insensitive to noise.

Real world data sets. In this section we describe experiments on publicly available real data sets and compare with previously reported results. We first of all consider the well known MNIST database (http://yann.lecun.com/exdb/mnist) containing several

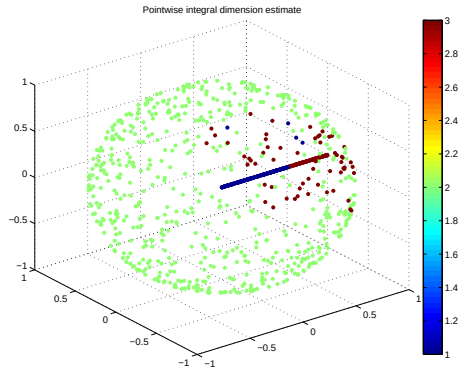

Fig. 4. Pointwise intrinsic dimensionality estimates distinguish the 1-dimensional line (blue) from the 2-dimensional sphere (green); points near the intersection are classified as 3-dimensional.

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

Fig. 5. MNIST database: intrinsic dimension estimate for each digit obtained with our method (MSVD), the smoothed Grassberg-Procaccia estimator from [11] (IDE), and the high rate vector quantization methods in [14] (HRVQ).

thousands images of hand written digits. Each image is 28 times 28 pixels. In Table 4 we report the dimension estimated for each individual digit and compare with the smoothed Grassberg Procaccia estimator from [11] and the high rate vector quantization approach in [14]. We also consider the IsoMap faces database (http://isomap.stanford.edu/dataset.html) consisting of 698 images of size 64 times 64 pixels. We find an average intrinsic dimension $k = 2$, see Figure 6. The different methods based on volume estimates return similar results, with $k \in [3, 4.65]$. See [13] for more extensive experiments on real-world data sets.

Speed. We compared the speed of the algorithms (see [13]): MSVD has mild linear growth as a function of n and of k, with speed comparable with that of most other algorithms, except MFA and MFA2 (3 orders of magnitude slower), and RTPMM and "Smoothing" that could not be run in reasonable time (less than several hours) for more than $n = 16,000$ points.

5. CONCLUSION

This work introduces a multiscale, geometric approach to intrinsic dimension estimation. By using PCA locally, we require a sample size essentially linear in the intrinsic dimension, and by analyzing the singular values over a range of scales, we are able to distinguish the underlying k-dimensional structure of the data from the effects of noise and curvature. The MSVD method, which was tested on both manifold and real-world data sets, frequently out-performs competing algorithms, particularly in the presence of noise. By applying this technique, there is great potential for improvement in current classification and dimensionality reduction algorithms.

6. REFERENCES

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

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

[3] 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*, pages 601–605, Aug. 2007.

[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*, pages 3917–3920, 31 2008-April 4 2008.

[5] H. Chen, J. Silva, D. Dunson, and L. Carin. Hierarchical bayesian embeddings for analysis and synthesis of dynamic data. *submitted*, 2010.

[6] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans. Signal Processing*, 2010.

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

[8] Guy David. *Wavelets and Singular Integrals on Curves and Surfaces*. Springer-Verlag, 1991.

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

[10] Gloria Haro, Gregory Randall, and Guillermo Sapiro. Translated poisson mixture model for stratification learning. *Int. J. Comput. Vision*, 80(3):358–374, 2008.

[11] M. Hein and Y. Audibert. Intrinsic dimensionality estimation of submanifolds in euclidean space. In S. Wrobel De Raedt, L., editor, *ICML Bonn*, pages 289 – 296, 2005.

[12] Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 777–784. MIT Press, Cambridge, MA, 2005.

[13] A.V. Little, M. Maggioni, and L. Rosasco. Multiscale geometric methods for data sets I: Estimation of intrinsic dimension. *in preparation*, 2010.

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

[15] Peter J. Verveer and Robert P.W. Duin. An evaluation of intrinsic dimensionality estimators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1), 1995.

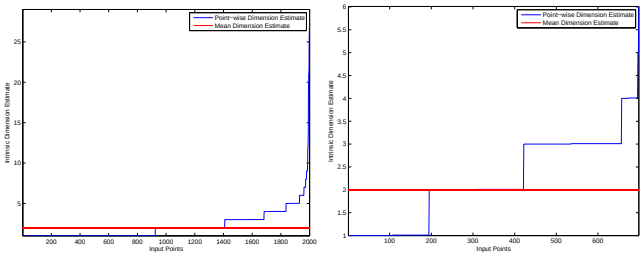

Fig. 6. Left: MNIST database; right: IsoMap face database; point-wise estimates for a subset of the points (blue) and the average across points (red).

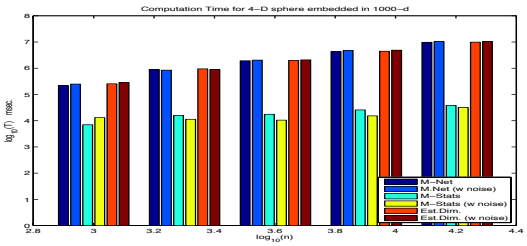

Fig. 7. Time to construct the multiscale nets ('M-Net'), calculation of multiscale statistics ('M-Stats') and the total time ('Est.Dim.').