Provable Dimension Detection using Principal Component Analysis*

Siu-Wing Cheng† Yajun Wang† Zhuangzhi Wu‡

February 8, 2006

Abstract

We analyze an algorithm based on principal component analysis (PCA) for detecting the dimension k of a smooth manifold $\mathcal{M} \subset \mathbb{R}^d$ from a set P of point samples. The best running time so far is $O(d\,2^{O(k^7 \log k)})$ by Giesen and Wagner after the adaptive neighborhood graph is constructed. Given the adaptive neighborhood graph, the PCA-based algorithm outputs the true dimension in $O(d2^{O(k)})$ time, provided that P satisfies a standard sampling condition as in previous results. Our experimental results validate the effectiveness of the approach. A further advantage is that both the algorithm and its analysis can be generalized to the noisy case, in which small perturbations of the samples and a small portion of outliers are allowed.

*Research supported by Research Grant Council, Hong Kong, China (612005). Part of this work was conducted while the third author visited the Department of Computer Science, HKUST, Hong Kong.

†Department of Computer Science, HKUST, Hong Kong. Email: {scheng,yalding}@cs.ust.hk

‡School of Computer Science and Engineering, Beihang University, Beijing, China. Email: zzwu@buaa.edu.cn

1 Introduction

Background. In applications such as speech recognition, weather forecasting and economic prediction, a large set P of point samples are generated by experiments or observations. The samples in P may reside in $\mathbb{R}^d$, but they are often postulated to lie on a manifold $\mathcal{M}$ of dimension $k < d$. The manifold learning problem is to compute a model for $\mathcal{M}$ and one important task is to compute its dimension. The challenge is to obtain an algorithm that is fast (even if d is very large compared with k) and robust against noise. Our main results are simple dimension detection algorithms for both the noiseless and the noisy cases.

Our results as well as the previous ones [9, 14] assume that P satisfies a standard sampling condition, which we review below. The *medial axis* of $\mathcal{M}$ is the set of centers of maximal empty d-dimensional balls that touch $\mathcal{M}$. For any point $x \in \mathcal{M}$, the *local feature size* $f(x)$ is the distance from x to the medial axis of $\mathcal{M}$. The local feature size satisfies the Lipschitz condition, i. e. , $f(x) \leqslant f(y) + ||x - y||$. The set P is an (ϵ, δ)-sampling of $\mathcal{M}$ for some constants $0 < \delta \leqslant \epsilon < 1$ if:

(i) $\epsilon/\delta \leqslant c_0$ for some constant c_0,

(ii) for any point $x \in \mathcal{M}$, there exists a sample $p \in P$ such that $\|p - x\| \leqslant \epsilon f(x)$, and

(iii) for any two samples $p, q \in P$, $\|p - q\| \geqslant \delta f(p)$.

Previous work. Dey et al. [9] gave the first provably correct algorithm. They first construct the d-dimensional Voronoi diagram of P. Then they analyze the shape of the Voronoi cell of a sample to determine the dimension of $\mathcal{M}$. If $\mathcal{M}$ has multiple components, this step can be repeated for all samples to yield the dimensions of all components. The worst-case complexity of the Voronoi diagram of n points in $\mathbb{R}^d$ is $\Theta(n^{\lceil d/2 \rceil})$ [11]. This is a huge quantity when d is large.

Giesen and Wagner [14] proposed to construct the *adaptive neighborhood graph* $G(P, c)$ by connecting each point sample p to other samples q such that $\|p - q\|$ is no more than c times the nearest neighbor distance of p, where c is a suitably chosen constant. The adaptive neighborhood graph $G(P, c)$ can be constructed in $O(dn^2)$ time by brute force. It is shown that $G(P, c)$ has the same connectivity as $\mathcal{M}$ and the shortest path distances in $G(P, c)$ approximate the geodesic distances on $\mathcal{M}$. For dimension detection, Giesen and Wagner fit a ℓ-dimensional affine subspace to a sample p and its neighbors in $G(P, c)$ so that the maximum distance from the samples to the subspace is approximately minimized. The fitting is done for $\ell = 1, 2, \ldots$ until the approximate maximum distance from the subspace is less than some threshold. Then they report the final value of ℓ as the dimension of the manifold component containing p. This takes $O(d2^{O(k^7 \log k)})$ worst-case time. Although this is no longer exponential in d, the exponential dependency on k^7 is prohibitive.

Tenenbaum et al. [20] proposed to construct a graph similar to the adaptive neighborhood graph. However, their method requires a globally uniform sampling, which is much stricter than the (ϵ, δ)-sampling. They did not give any combinatorial bound on the running time to detect the manifold dimension [20].

Our results. Given the adaptive neighborhood graph, we propose to apply *principal component analysis* to a sample p and its neighbors to detect the dimension of the manifold component containing p. We collect the vectors $q - p$ for all neighbors q of p and compute the eigenvalues of this set of vectors. If the true dimension is k, then the eigenvectors for the k largest eigenvalues should approximately span the tangent space at p. The eigenvector for the $(k+1)$th largest eigenvalue should be approximately normal to the tangent space. So the $(k+1)$th largest eigenvalue should be tiny when compared with the k largest eigenvalues. Our strategy is to detect the gap in the eigenvalues.

We assume the knowledge of the bound c_0 on ϵ/δ as in previous results [9, 14]. We give an algorithm that reports the true dimension in $O(d2^{O(k)})$ time. The proofs require a very high sampling density. We present experimental results to show that our approach is much more effective than the theory predicts. Next, we propose an $(\epsilon, \delta, \sigma)$-noisy-sampling model. Small perturbations of the samples as well as a small portion of outliers are allowed. The maximum noise magnitude of non-outliers is allowed to be ϵ^2 times the local feature sizes. The outliers can be arbitrarily placed and the fraction of outliers is allowed to be $O(2^{-\Theta(k)}(\log n)^{-1})$, where n is the total number of samples. Since the outliers may mess up the connectivity, we have to restrict the unknown manifold to be connected. We present an algorithm that reports the true dimension with probability at least $1 - 1/n^2$ in $O(d2^{O(k)} \log n)$ expected time. Our experimental results show that the solution quality is robust against noise.

Fukunaga and Olsen [13] also proposed to apply principal component analysis to detect the manifold dimension. However, they did not give any formal analysis. Our work gives the theoretical justification why principal component analysis works for dimension detection in both the noiseless and noisy cases.

Outline. We review the basics of principal component analysis in Section 2. Section 3 studies the variance in a ball which will be useful later. We present the algorithmic and experimental results on dimension detection for the noiseless case in Section 4 and the noisy case in Section 5. We conclude in Section 6.

2 Principal component analysis

Principal component analysis (PCA) is used to reduce the dimension of a set of vectors X by identifying the most significant directions [16]. Given two vectors u and v, we use $\langle u, v \rangle$ to denote their inner product. Let $X = \{x_1, x_2, \ldots, x_m\}$ be a set of d-dimensional vectors. For any unit vector $v \in \mathbb{R}^d$, the *variance of X in direction v* is

$$\mathrm{var}(X, v) = \sum_{i=1}^{m} \langle x_i, v \rangle^2.$$

The most significant direction corresponds to the unit vector v_1 such that $\mathrm{var}(X, v_1)$ is maximum. In general, after identifying the j most significant directions $B_j = \{v_1, v_2, \ldots, v_j\}$, the $(j+1)$th most significant direction corresponds to the unit vector v_{j+1} such that $\mathrm{var}(X, v_{j+1})$ is maximum among all unit vectors in $\mathrm{span}(B_j)^\perp$, where $\mathrm{span}(B_j)$ denotes the linear subspace spanned by B_j and $\mathrm{span}(B_j)^\perp$ denotes its orthogonal complement.

This above procedure can be formulated as an eigenvalue problem. For $1 \leqslant j \leqslant d$, we use x_{ij} to denote the jth coordinate of the vector x_i. Define the $d \times d$ *covariance matrix*

$$\mathrm{C} \;=\; \sum_{i=1}^{m} (x_{i1}, x_{i2}, \ldots, x_{id})^t \; (x_{i1}, x_{i2}, \ldots, x_{id}).$$

The covariance matrix C is symmetric and positive semi-definite. It can be verified that

$$\forall \text{ unit vector } v \in \mathbb{R}^d, \;\; \langle \mathrm{C}v, v \rangle \;=\; \mathrm{var}(X, v). \tag{1}$$

If $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_d$ are the eigenvalues of C, then the unit eigenvector v_j for λ_j is the jth most significant direction as defined by the previous procedure. The following result summarizes the above background knowledge on PCA. For any set S of orthonormal unit vectors in $\mathbb{R}^d$, we use $\mathrm{var}(X, S)$ to denote $\sum_{v \in S} \mathrm{var}(X, v)$.

Lemma 2.1 *For $1 \leqslant j \leqslant d$, let λ_j be the jth largest eigenvalue of* C *and let v_j denote the unit eigenvector for λ_j. Let $B_j = \{v_1, v_2, \ldots, v_j\}$. Then $\lambda_1 = \max\{\, \mathrm{var}(X, v) \;:$ unit vector v in $\mathbb{R}^d \,\}$. For $2 \leqslant j \leqslant d$, we have*

(i) $\lambda_j = \max\{\, \mathrm{var}(X, v) \;:$ *unit vector v in* $\mathrm{span}(B_{j-1})^{\perp} \,\}$,

(ii) $\lambda_j = \min\{\, \mathrm{var}(X, v) \;:$ *unit vector v in* $\mathrm{span}(B_j) \,\}$, *and*

(iii) $\mathrm{var}(X, B_j) \geqslant \mathrm{var}(X, S)$ *for any set S of j orthonormal vectors.*

3 Variance of a ball

To ease the subsequent analysis, we generalize the notion of variance to a continuous ball. Let $\mathcal{B}$ be an i-dimensional ball, $i \leqslant d$, centered at the origin with radius R. Each point $x \in \mathcal{B}$ is viewed as a position vector. We assign a positive weight W to $\mathcal{B}$. Take a unit vector v in $\mathbb{R}^d$. The *variance of* $\mathcal{B}$ in direction v is defined as

$$\mathrm{var}(\mathcal{B}, v) \;=\; W \int_{-R}^{R} \mathrm{vol}(H(v, r) \cap \mathcal{B}) \; r^2 \, dr,$$

where $H(v, r)$ is the $(i-1)$-dimensional hyperplane that is normal to v and at signed distance r from the origin.

Lemma 3.1 *Let $\mathcal{B}$ be an i-dimensional ball, $i \leqslant d$, centered at the origin with radius R and weight W. Let v be a unit vector in $\mathbb{R}^d$ such that $Rv \in \mathcal{B}$. Then* $\mathrm{var}(\mathcal{B}, v) = W R^{i+2} V_i / (i+2)$, *where V_i is the volume of the i-dimensional unit ball.*

Proof. We cut $\mathcal{B}$ with planes normal to v into slices with width dr. Take a slice S that is at distance r from the origin and subtends an angle $\pi - 2\theta$ at the origin. See Figure 1. Then $r = R \sin\theta$, $dr = R \cos\theta \, d\theta$, and the radius of S is $R \cos\theta$. Thus,

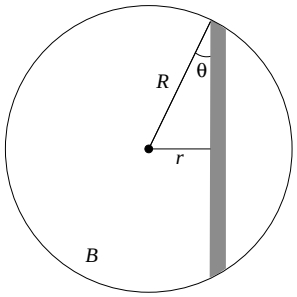

Figure 1: Lemma 3.1.

$$\begin{aligned}\mathrm{var}(\mathcal{B}, v) &= 2W \int_0^{\frac{\pi}{2}} (R\cos\theta)^{i-1} V_{i-1} \cdot (R\sin\theta)^2 \cdot R\cos\theta \, d\theta \\ &= 2W R^{i+2} V_{i-1} \int_0^{\frac{\pi}{2}} (\cos^i\theta - \cos^{i+2}\theta) \, d\theta.\end{aligned}$$

Rewriting $\int_0^{\frac{\pi}{2}} \cos^{i+2}\theta \;\; d\theta$ as $\int_0^{\frac{\pi}{2}} \cos^{i+1}\theta \;\; d\sin\theta$ and using integration by parts, one can show that $\int_0^{\frac{\pi}{2}} \cos^{i+2}\theta \, d\theta = \frac{i+1}{i+2} \int_0^{\frac{\pi}{2}} \cos^i\theta \, d\theta$. Therefore, $\mathrm{var}(\mathcal{B}, v) = \frac{2WR^{i+2}V_{i-1}}{i+2} \int_0^{\frac{\pi}{2}} \cos^i\theta \, d\theta$. The volume V_i can be evaluated by integrating the volume of $(i-1)$-dimensional slices orthogonal to a diameter as discussed in the above. This yields $V_i = 2V_{i-1} \int_0^{\frac{\pi}{2}} \cos^i\theta \, d\theta$. , i.e., $\int_0^{\frac{\pi}{2}} \cos^i\theta \, d\theta = V_i/(2V_{i-1})$. Substituting these into the above yields the result. □

4 Dimension detection

Let $\mathcal{M} \subset \mathbb{R}^d$ be the unknown k-dimensional manifold from which the point samples P are drawn. We assume that P satisfies the (ϵ, δ)-sampling conditions. In this section, we show that PCA can be used to efficiently and provably detect k.

Let $c \geqslant 1$ be some constant to be defined later. Let $\mathcal{B}_p(c)$ denote the d-dimensional ball centered at p with radius $c\epsilon f(p)$. Let $P_p(c)$ be the set of samples $P \cap \mathcal{B}_p(c)$. Let $X_p(c)$ be the set of vectors $\{\, q - p \,:\, q \in P_p(c) \setminus \{p\} \,\}$. Let $\mathcal{T}_p$ and $\mathcal{N}_p$ denote the tangent space and the normal space at p to the manifold $\mathcal{M}$, respectively.

We give an overview of our analysis. Let $k = \dim(\mathcal{T}_p)$. First, we show that the variance of $X_p(c)$ does not vary much over the directions in $\mathcal{T}_p$. Second, we show that the variance of $X_p(c)$ in any direction in $\mathcal{N}_p$ is tiny when compared the variance of $X_p(c)$ in any direction in $\mathcal{T}_p$. Therefore, the k largest eigenvalues are close to $\mathrm{var}(X_p(c), u)$ for any unit vector $u \in \mathcal{T}_p$. The $(k+1)$th largest eigenvalue is less than the upper bound of $\mathrm{var}(X_p(c), v)$ for any unit vector $v \in \mathcal{N}_p$. This establishes the gap between the kth and $(k+1)$th eigenvalues, which the algorithm will detect.

4.1 Projection onto the tangent space

For each point $q \in P_p(c)$, let $\widehat{q}$ denote its orthogonal projection onto $\mathcal{T}_p$. Let $\widehat{P}_p(c)$ denote the set of projected points. Similarly, let $\widehat{X}_p(c)$ be the projection of the vectors in $X_p(c)$

onto $\mathcal{T}_p$. We are interested in $\widehat{X}_p(c)$ because for any unit vector $u \in \mathcal{T}_p$, $\mathrm{var}(\widehat{X}_p(c), u) = \mathrm{var}(X_p(c), u)$. We need the following two results in [14] (rephrased to fit our presentation).

Lemma 4.1 *Assume that $\alpha\epsilon < 1$. For any points $a, b \in \mathcal{M}$, if $\|a - b\| \leqslant \alpha\epsilon f(a)$, the distance from b to $\mathcal{T}_a$ is at most $(\alpha\epsilon)^2 f(a)/2$.*

Lemma 4.2 *Assume that $\alpha\epsilon < 1/4$. Let a be a point in $\mathcal{M}$. Let b be a point in $\mathcal{T}_a$. Let $\bar{b} \in \mathcal{M}$ be the closest point to b. If $\|a - b\| \leqslant \alpha\epsilon f(a)$, then $\|b - \bar{b}\| \leqslant 2(\alpha\epsilon)^2 f(p)$.*

We show that $\widehat{P}_p(c)$ is fairly uniformly distributed. Recall that $\epsilon/\delta \leqslant c_0$ for some constant c_0.

Lemma 4.3 *Assume that $c > 2$ and $\epsilon \leqslant \min\{\frac{1}{(c-2)(4c-7)}, \frac{1}{2c^2c_0}\}$. There exist constants ξ and η such that $\xi \in [\epsilon, 2\epsilon]$, $\eta \in [\delta/4, \delta/2]$, and the following hold.*

(i) *For any point $a \in \mathcal{B}_p(c-2) \cap \mathcal{T}_p$, there exists $\widehat{s} \in \widehat{P}_p(c)$ such that $\|a - \widehat{s}\| \leqslant \xi f(p)$.*

(ii) *For any $q, s \in P_p(c)$, $\|\widehat{q} - \widehat{s}\| \geqslant 2\eta f(p)$.*

Proof. We prove the lemma for the constants $\xi = \epsilon(1 + (c-2)\epsilon + 4(c-2)^2\epsilon)$ and $\eta = \delta/2 - (c\epsilon)^2/2$. Since $\epsilon \leqslant \frac{1}{(c-2)(4c-7)}$, $(c-2)\epsilon + 4(c-2)^2\epsilon \leqslant 1$ and so $\xi \in [\epsilon, 2\epsilon]$. Since $\epsilon/\delta \leqslant c_0$ and $\epsilon \leqslant \frac{1}{2c^2c_0}$, $\eta \geqslant \delta/2 - (c^2\epsilon)/(4c^2c_0) = \delta/2 - \epsilon/(4c_0) \geqslant \delta/4$.

Consider (i). Let $\beta = c - 2$. Let $\bar{a} \in \mathcal{M}$ be the closest point to a. Since $c > 2$ and $\epsilon \leqslant 1/(2c^2c_0)$, we have $c\epsilon \leqslant 1/(2cc_0) < 1/4$. So Lemma 4.2 implies that

$$\|a - \bar{a}\| \;\leqslant\; 2\beta^2\epsilon^2 f(p).$$

The Lipschitz condition implies that $f(\bar{a}) \leqslant f(p) + \|p - \bar{a}\| \leqslant f(p) + \|p - a\| + \|a - \bar{a}\| \leqslant (1 + \beta\epsilon + 2\beta^2\epsilon^2) f(p)$. By the sampling condition, there is one point $s \in P$ such that

$$\|\bar{a} - s\| \;\leqslant\; \epsilon f(\bar{a}) \;\leqslant\; \epsilon(1 + \beta\epsilon + 2\beta^2\epsilon^2) f(p).$$

Therefore, $\|a - \widehat{s}\| \leqslant \|a - s\| \leqslant \|a - \bar{a}\| + \|\bar{a} - s\| \leqslant \epsilon(1 + \beta\epsilon + 4\beta^2\epsilon^2) f(p) = \xi f(p)$. Moreover, $\|p - s\| \leqslant \|p - a\| + \|a - s\| \leqslant \beta\epsilon f(p) + \xi f(p) \leqslant (\beta + 2)\epsilon f(p) = c\epsilon f(p)$. Thus $s \in P_p(c)$ and $\widehat{s} \in \widehat{P}_p(c)$.

Consider (ii). By Lemma 4.1, the distances of q and s from $\mathcal{T}_p$ are at most $(c\epsilon)^2 f(p)/2$. So $\|\widehat{q} - \widehat{s}\| \geqslant \|q - s\| - \|q - \widehat{q}\| - \|s - \widehat{s}\| \geqslant \delta f(p) - (c\epsilon)^2 f(p) = 2\eta f(p)$. ∎

Lemma 4.3(ii) also implies that no two points in $P_p(c)$ project to the same point in $\widehat{P}_p(c)$.

4.2 Upper and lower bounds of variance

First, we prove upper bounds for $|P_p(c)|$ and $\mathrm{var}(X_p(c), u)$ for any unit vector $u \in \mathcal{T}_p$.

Lemma 4.4 *Let $k = \dim(\mathcal{T}_p)$. Assume that $c > 2$ and $\epsilon \leqslant \min\{\frac{1}{(c-2)(4c-7)}, \frac{1}{2c^2c_0}\}$. Then*

$$|P_p(c)| \;\leqslant\; \frac{2(c\epsilon+\eta)^k}{\eta^k}$$

and for any unit vector $u \in \mathcal{T}_p$,

$$\mathrm{var}(X_p(c), u) \;\leqslant\; \frac{2f(p)^2}{k+2} \cdot \frac{(c\epsilon+2\eta)^{k+2}}{\eta^k}.$$

Proof. Let $\mathcal{B}_0$ be the k-dimensional ball $\mathcal{B}_p(c) \cap \mathcal{T}_p$. The ball $\mathcal{B}_0$ has radius $c\epsilon f(p)$ and the projected points $\widehat{P}_p(c)$ lie inside $\mathcal{B}_0$. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be the k-dimensional balls in $\mathcal{T}_p$ centered at p with radii $(c\epsilon+\eta)f(p)$ and $(c\epsilon+2\eta)f(p)$, respectively.

Take a unit vector $u \in \mathcal{T}_p$. Starting at p, we move along u and cut $\mathcal{B}_2$ into k-dimensional slices $S_0, S_1, S_2, \cdots$ with width $\eta f(p)$ and normal to u. Although the last slice may have width less than $\eta f(p)$, those that stab $\mathcal{B}_1$ have width exactly $\eta f(p)$. Then we repeat again in direction $-u$. Refer to Figure 2.

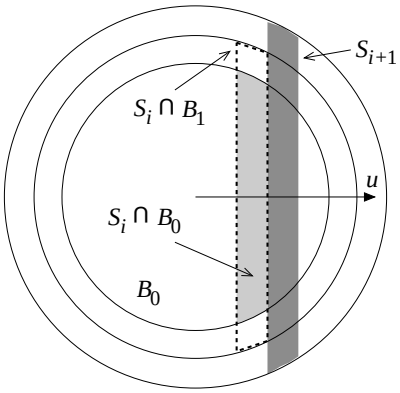

Figure 2: The balls shown are $\mathcal{B}_0$, $\mathcal{B}_1$ and $\mathcal{B}_2$. Their radii are $c\epsilon f(p)$, $(c\epsilon+\eta)f(p)$ and $(c\epsilon+2\eta)f(p)$, respectively. The width of the slices shown are $\eta f(p)$.

Let S_i be a slice that stabs $\mathcal{B}_0$. We bound the cardinality of $\widehat{P}_p(c) \cap S_i$. In Figure 2, the intersection $S_i \cap \mathcal{B}_0$ is lightly shaded and the intersection $S_i \cap \mathcal{B}_1$ has dashed border. By Lemma 4.3(ii), we can center disjoint empty k-dimensional balls with radii $\eta f(p)$ at the points in $\widehat{P}_p(c) \cap S_i$. Observe that at least half of each such ball lies inside $S_i \cap \mathcal{B}_1$. Recall that V_k denotes the volume of a k-dimensional ball with unit radius. Then the volume of a k-dimensional ball with radius $\eta f(p)$ is $V_k \eta^k f(p)^k$. It follows that

$$|\widehat{P}_p(c) \cap S_i| \;\leqslant\; \frac{2\ \mathrm{vol}(S_i \cap \mathcal{B}_1)}{V_k \eta^k f(p)^k}.$$

Summing up over all slices that stab $\mathcal{B}_0$ yields $|P_p(c)| = |\widehat{P}_p(c)| \leqslant \frac{2\ \mathrm{vol}(\mathcal{B}_1)}{V_k \eta^k f(p)^k} = \frac{2(c\epsilon+\eta)^k}{\eta^k}$.

Next, we prove the upper bound on $\mathrm{var}(X_p(c), u)$. Since S_i stabs $\mathcal{B}_0$, the next slice S_{i+1} must stab $\mathcal{B}_1$ and has width exactly $\eta f(p)$. The slice S_{i+1} is darkly shaded in Figure 2. It can be checked that $\mathrm{vol}(S_i \cap \mathcal{B}_1) \leqslant \mathrm{vol}(S_{i+1})$. We conclude that

$$|\widehat{P}_p(c) \cap S_i| \;\leqslant\; \frac{2\ \mathrm{vol}(S_i \cap \mathcal{B}_1)}{V_k \eta^k f(p)^k} \;\leqslant\; \frac{2\ \mathrm{vol}(S_{i+1})}{V_k \eta^k f(p)^k}.$$

Let r be the maximum distance of any point in S_i from p in direction u. The contribution of points in $\widehat{P}_p(c) \cap S_i$ to $\mathrm{var}(\widehat{X}_p(c), u)$ is upper bounded by $|\widehat{P}_p(c) \cap S_i| \cdot r^2 \leqslant \frac{2r^2 \text{ vol}(S_{i+1})}{V_k \eta^k f(p)^k}$. Note that r is a lower bound of the distance of any point in S_{i+1} from p in direction u. Thus, if we assign $\frac{2}{V_k \eta^k f(p)^k}$ as the weight of $\mathcal{B}_2$, then $\frac{2r^2 \text{ vol}(S_{i+1})}{V_k \eta^k f(p)^k}$ is a lower bound of the contribution of S_{i+1} to $\mathrm{var}(\mathcal{B}_2, u)$. Hence, Lemma 3.1 implies that $\mathrm{var}(X_p(c), u) = \mathrm{var}(\widehat{X}_p(c), u) \leqslant \mathrm{var}(\mathcal{B}_2, u) = \frac{2f(p)^2}{k+2} \cdot \frac{(c\epsilon+2\eta)^{k+2}}{\eta^k}$. □

Next, we prove lower bounds for $|P_p(c)|$ and $\mathrm{var}(X_p(c), u)$ for any unit vector $u \in \mathcal{T}_p$. This shows that the variance of $X_p(c)$ does not vary much over the directions in $\mathcal{T}_p$.

Lemma 4.5 *Let $k = \dim(\mathcal{T}_p)$. Assume that $c > 10$ and $\epsilon \leqslant \min\{\frac{1}{(c-2)(4c-7)}, \frac{1}{2c^2 c_0}\}$. Then*

$$|P_p(c)| \;\geqslant\; \frac{((c-2)\epsilon - 4\xi)^k}{4^k \xi^k}$$

and for any unit vector $u \in \mathcal{T}_p$,

$$\mathrm{var}(X_p(c), u) \;\geqslant\; \frac{f(p)^2}{k+2} \cdot \frac{((c-2)\epsilon - 4\xi)^{k+2}}{4^k \xi^k}.$$

Proof. Let $\mathcal{B}_0$ be the k-dimensional ball centered at p with radius $(c-2)\epsilon f(p)$ in $\mathcal{T}_p$. By Lemma 4.3(i), every point in $\mathcal{B}_0$ is close to some point in $\widehat{P}_p(c)$. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be the k-dimensional balls in $\mathcal{T}_p$ centered at p with radii $((c-2)\epsilon - 2\xi)f(p)$ and $((c-2)\epsilon - 4\xi)f(p)$, respectively. (We require $c > 10$ so that $\mathcal{B}_2$ has a positive radius.)

Starting at p, we move in direction u and cut $\mathcal{B}_0$ into k-dimensional slices $S_0, S_1, S_2, \cdots$ with width $2\xi f(p)$ and normal to u. The last slice may have width less than $2\xi f(p)$ but the slices that stab $\mathcal{B}_1$ have width exactly $2\xi f(p)$. Then we repeat again in direction $-u$. Refer to Figure 3. The rest of the argument is similar to that in the proof of Lemma 4.4.

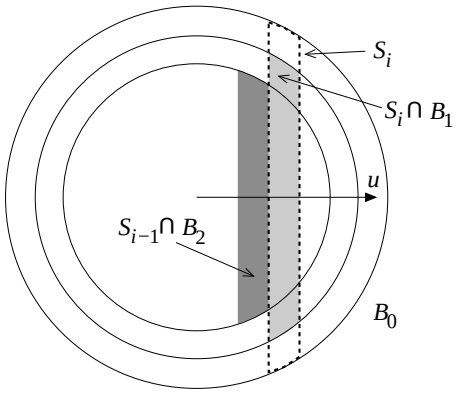

Figure 3: The balls shown are $\mathcal{B}_0$, $\mathcal{B}_1$ and $\mathcal{B}_2$. Their radii are $(c-2)\epsilon f(p)$, $((c-2)\epsilon - 2\xi)f(p)$ and $((c-2)\epsilon - 4\xi)f(p)$, respectively. The width of the slices shown are $2\xi f(p)$.

Let S_i be a slice that stabs $\mathcal{B}_1$. In Figure 3, S_i has dashed border and the intersection $S_i \cap \mathcal{B}_1$ is lightly shaded. We pack a maximal set of points into $S_i \cap \mathcal{B}_1$ such that the

distance between any two points is at least $4\xi f(p)$. So we can form a set Σ of disjoint balls centered at these points with radii $2\xi f(p)$. At least half of each ball in Σ lies inside S_i, which implies that each ball in Σ contains a ball of radius $\xi f(p)$ that lies inside S_i. Let Σ' denote this set of smaller disjoint balls. By Lemma 4.3(i), each ball in Σ' contains a point in $\widehat{P}_p(c)$. Thus $|\widehat{P}_p(c) \cap S_i| \geqslant |\Sigma'| = |\Sigma|$. By the construction of Σ, if we double the radius of each ball in Σ, the expanded balls cover $S_i \cap \mathcal{B}_1$. Hence,

$$|\widehat{P}_p(c) \cap S_i| \;\geqslant\; |\Sigma| \;\geqslant\; \frac{\mathrm{vol}(S_i \cap \mathcal{B}_1)}{V_k 4^k \xi^k f(p)^k}.$$

Summing over all slices that stab $\mathcal{B}_1$, we get $|P_p(c)| = |\widehat{P}_p(c)| \geqslant \frac{\mathrm{vol}(\mathcal{B}_1)}{V_k 4^k \xi^k f(p)^k} = \frac{((c-2)\epsilon - 2\xi)^k}{4^k \xi^k}$.

Next, we prove the lower bound on $\mathrm{var}(X_p(c), u)$. Take a slice S_{i-1} that stabs $\mathcal{B}_2$. So the next slice S_i must stab $\mathcal{B}_1$ and has width $2\xi f(p)$. The intersection $S_{i-1} \cap \mathcal{B}_2$ is shown darkly shaded in Figure 3. It can be checked that $\mathrm{vol}(S_{i-1} \cap \mathcal{B}_2) \leqslant \mathrm{vol}(S_i \cap \mathcal{B}_1)$. Thus

$$|\widehat{P}_p(c) \cap S_i| \;\geqslant\; \frac{\mathrm{vol}(S_i \cap \mathcal{B}_1)}{V_k 4^k \xi^k f(p)^k} \;\geqslant\; \frac{\mathrm{vol}(S_{i-1} \cap \mathcal{B}_2)}{V_k 4^k \xi^k f(p)^k}.$$

Let r be the minimum distance of points in S_i from p in direction u. The contribution of points in $\widehat{P}_p(c) \cap S_i$ to $\mathrm{var}(\widehat{X}_p(c), u)$ is lower bounded by $|\widehat{P}_p(c) \cap S_i| \cdot r^2 \geqslant \frac{r^2 \,\mathrm{vol}(S_{i-1} \cap \mathcal{B}_2)}{V_k 4^k \xi^k f(p)^k}$. Note that r is an upper bound of the distances of points in $S_{i-1} \cap \mathcal{B}_2$ from p in direction u. Thus, if we assign $\frac{1}{V_k 4^k \xi^k f(p)^k}$ as the weight of $\mathcal{B}_2$, then $\frac{r^2 \,\mathrm{vol}(S_{i-1} \cap \mathcal{B}_2)}{V_k 4^k \xi^k f(p)^k}$ is an upper bound of the contribution of $S_{i-1} \cap \mathcal{B}_2$ to $\mathrm{var}(\mathcal{B}_2, u)$. By Lemma 3.1, $\mathrm{var}(X_p(c), u) = \mathrm{var}(\widehat{X}_p(c), u) \geqslant \mathrm{var}(\mathcal{B}_2, u)$ which is $\frac{f(p)^2}{k+2} \cdot \frac{((c-2)\epsilon - 4\xi)^{k+2}}{4^k \xi^k}$. □

Finally, we prove an upper bound for $\mathrm{var}(X_p(c), v)$ for any unit vector $v \in \mathcal{N}_p$. This upper bound has an extra ϵ^2 factor when compared with the variance of $X_p(c)$ in any direction in $\mathcal{T}_p$. Thus, when ϵ is small, for any unit vector $v \in \mathcal{N}_p$, $\mathrm{var}(X_p(c), v)$ is small compared with the variance of $X_p(c)$ in any direction in $\mathcal{T}_p$.

Lemma 4.6 *Let $k = \dim(\mathcal{T}_p)$. Assume that $c > 2$ and $\epsilon \leqslant \min\{\frac{1}{(c-2)(4c-7)}, \frac{1}{2c^2 c_0}\}$. For any unit vector $v \in \mathcal{N}_p$,*

$$\mathrm{var}(X_p(c), v) \;\leqslant\; \frac{c^4 \epsilon^4 f(p)^2}{2} \cdot \frac{(c\epsilon + \eta)^k}{\eta^k}.$$

Proof. Take any unit vector $v \in \mathcal{N}_p$. By Lemma 4.1, every point $q \in P_p(c)$ is at distance $(c\epsilon)^2 f(p)/2$ or less from p in direction v. So $\mathrm{var}(X_p(c), v) \leqslant |P_p(c)| \cdot (c\epsilon)^4 f(p)^2/4$. Substituting the upper bound of $|P_p(c)|$ in Lemma 4.4 yields the result. □

4.3 Upper and lower bounds of eigenvalues

We apply the previous results to bound the eigenvalues of the covariance matrix for subsets and supersets of $X_p(c)$. Define the following functions:

$$\alpha_1(c) = \frac{(c-10)^{k+2}}{(k+2)2^{3k}}, \quad \alpha_2(c) = \frac{(c+1)^{k+2} 2^{2k+1} c_0^k}{k+2}, \quad \alpha_3(c) = \frac{c^4 (4cc_0 + 1)^k}{2}.$$

Lemma 4.7 *Let p be a sample in P. Let X be a set of d-dimensional vectors. Let $\lambda_1 \geqslant \lambda_2 \cdots \geqslant \lambda_d$ be the eigenvalues of the covariance matrix for X. Let $k = \dim(\mathcal{T}_p)$. Assume $c > 10$ and $\epsilon \leqslant \min\{\frac{2}{c\sqrt{k+2}}, \frac{1}{(c-2)(4c-7)}, \frac{1}{2c^2c_0}\}$.*

(i) *If $X_p(c) \subseteq X$, then $\lambda_j \geqslant \alpha_1(c)\epsilon^2 f(p)^2$ for $1 \leqslant j \leqslant k$.*

(ii) *If $X \subseteq X_p(c)$, then $\lambda_j \leqslant \alpha_2(c)\epsilon^2 f(p)^2$ for $1 \leqslant j \leqslant k$.*

(iii) *If $X \subseteq X_p(c)$, then $\lambda_j \leqslant \alpha_3(c)\epsilon^4 f(p)^2$ for $k+1 \leqslant j \leqslant d$.*

Proof. Let v_j be the unit eigenvector for λ_j for $1 \leqslant j \leqslant d$. Consider (i). For $1 \leqslant j \leqslant k$, the dimension of $\text{span}(\{v_1, \cdots, v_{j-1}\})^\perp$ is $d - j + 1 \geqslant d - k + 1$. (We define $\text{span}(\emptyset)^\perp$ to be $\mathbb{R}^d$.) So the dimensions of $\text{span}(\{v_1, \cdots, v_{j-1}\})^\perp$ and $\mathcal{T}_p$ sum to $\geqslant d+1$, meaning that they intersect at a subspace of dimension at least 1. Take a unit vector u in the intersection. Since $u \in \text{span}(\{v_1, \cdots, v_{j-1}\})^\perp$, we have $\lambda_j \geqslant \text{var}(X, u)$ by Lemma 2.1(i). Since $X_p(c) \subseteq X$, we have $\text{var}(X_p(c), u) \leqslant \text{var}(X, u) \leqslant \lambda_j$. Because $u \in \mathcal{T}_p$, the lower bound for $\text{var}(X_p(c), u)$ in Lemma 4.5 implies that $\lambda_j \geqslant \frac{f(p)^2}{k+2} \cdot \frac{((c-2)\epsilon - 4\xi)^{k+2}}{4^k \xi^k}$. Substituting the inequality $\xi \leqslant 2\epsilon$ into the bound yields $\lambda_j \geqslant \alpha_1(c)\epsilon^2 f(p)^2$.

Consider (ii). The eigenvector v_j can be written as $\sqrt{\kappa} \cdot u + \sqrt{1-\kappa} \cdot w$ for some constant $0 \leqslant \kappa \leqslant 1$ and some unit vectors $u \in \mathcal{T}_p$ and $w \in \mathcal{N}_p$. Since u and w are orthogonal, $\lambda_j = \text{var}(X, v_j) = \kappa \cdot \text{var}(X, u) + (1-\kappa) \cdot \text{var}(X, w)$. Since $X \subseteq X_p(c)$, we have $\text{var}(X, u) \leqslant \text{var}(X_p(c), u)$ and $\text{var}(X, w) \leqslant \text{var}(X_p(c), w)$. Let U_u be the upper bound of $\text{var}(X_p(c), u)$ shown in Lemma 4.4 and let U_w be the upper bound of $\text{var}(X_p(c), w)$ shown in Lemma 4.6. Then $\lambda_j \leqslant \kappa U_u + (1-\kappa)U_w$. By our assumption that $\epsilon \leqslant \frac{2}{c\sqrt{k+2}}$, we have $4 \geqslant c^2\epsilon^2(k+2)$ and hence $4(c\epsilon + 2\eta)^{k+2} \geqslant c^4\epsilon^4(c\epsilon + \eta)^k(k+2)$. Under this condition, $U_u \geqslant U_w$ which implies that $\lambda_j \leqslant U_u$. Then substituting the inequalities $\eta \leqslant \epsilon/2$, $\eta \geqslant \delta/4$ and $\epsilon/\delta \leqslant c_0$ into U_u yields $\lambda_j \leqslant \alpha_2(c)\epsilon^2 f(p)^2$.

Consider (iii). The dimensions of $\text{span}(\{v_1, \ldots, v_{k+1}\})$ and $\mathcal{N}_p$ sum to $d+1 > d$. So they intersect at a subspace of dimension at least 1. Take a unit vector w in the intersection. Since $w \in \text{span}(\{v_1, \ldots, v_{k+1}\})$, we have $\lambda_{k+1} \leqslant \text{var}(X, w)$ by Lemma 2.1(ii). Since $X \subseteq X_p(c)$, we have $\text{var}(X, w) \leqslant \text{var}(X_p(c), w)$ and so $\lambda_{k+1} \leqslant \text{var}(X_p(c), w)$. Then, as $w \in \mathcal{N}_p$, λ_{k+1} is no more than the upper bound of $\text{var}(X_p(c), w)$ in Lemma 4.6. Substituting the inequalities $\eta \geqslant \delta/4$ and $\epsilon/\delta \leqslant c_0$ yields into this upper bound yields $\lambda_{k+1} \leqslant \alpha_3(c)\epsilon^4 f(p)^2$. Lastly, $\lambda_j \leqslant \lambda_{k+1}$ for all $k+1 < j \leqslant d$. □

In all, if $X_p(c) \subseteq X \subseteq X_p(c')$ for some constants c and c', then the $d-k$ smallest eigenvalues of the covariance matrix for X differ from the k largest eigenvalues essentially by a factor of ϵ^2. The additional factor ϵ^2 creates a gap between these two groups of eigenvalues.

4.4 Algorithmic results

Recall that the $d \times d$ covariance matrix C is equal to $\sum_{i=1}^{m} (x_{i1}, x_{i2}, \ldots, x_{id})^t \ (x_{i1}, x_{i2}, \ldots, x_{id})$ for the vectors $(x_{i1}, x_{i2}, \ldots, x_{id})^t$, $1 \leqslant i \leqslant m$. Alternatively, if we use A to denote the $d \times m$ matrix in which the ith column is $(x_{i1}, x_{i2}, \ldots, x_{id})^t$, then $\text{C} = \text{A}\,\text{A}^t$.

Using the algorithm by Yau and Lu [21], the eigenvalues of a $N \times N$ symmetric matrix can be computed in $O(T(N) \log N)$ time, i.e., $O(T(d) \log d)$ for C, where $T(N)$ is the time for multiplying two $N \times N$ matrices. (It is known that $T(N) \log N = o(N^3)$.) This is time-consuming when d is huge, but a trick comes to our rescue. First, observe that when $m < d$, C does not have full rank and it has at most m non-zero eigenvalues. Second, if λ is a non-zero eigenvalue of C, λ is an eigenvalue of $\mathrm{A}^t\,\mathrm{A}$, which is a $m \times m$ matrix. We give a proof of this known fact below for completeness.

Lemma 4.8 *If $\lambda \neq 0$ is an eigenvalue of* C, *λ is an eigenvalue of* $\mathrm{A}^t\,\mathrm{A}$.

Proof. Let u be an eigenvector of C corresponding to λ. Note that $\mathrm{A}^t u \neq 0$; otherwise, $\mathrm{C}u = \mathrm{A}\,\mathrm{A}^t u = 0$ instead of λu, a contradiction. If we multiply $\mathrm{A}^t\,\mathrm{A}$ with $\mathrm{A}^t u$, we get $\mathrm{A}^t\,(\mathrm{A}\,\mathrm{A}^t u) = \lambda \mathrm{A}^t\, u$. This shows that $\mathrm{A}^t u$ is an eigenvector of $\mathrm{A}^t\,\mathrm{A}$ with eigenvalue λ. □

Therefore, it suffices to spend $O(T(m) \log m)$ time to compute the eigenvalues of $\mathrm{A}^t\,\mathrm{A}$ in order to obtain the non-zero eigenvalues of C. Forming the products $\mathrm{A}\,\mathrm{A}^t$ and $\mathrm{A}^t\,\mathrm{A}$ take time $O(d^2 m)$ and $O(dm^2)$, respectively. In all, the non-zero eigenvalues of C can be obtained in $O(\min\{T(d) \log d + d^2 m, T(m) \log m + dm^2\})$ time. The simpler upper bound $O(dm^2)$ will be sufficient for our purposes.

The following psuedo-code summarizes our algorithm for reporting the dimension of the manifold component containing a sample p. We assume the knowledge of the bound c_0 on ϵ/δ as in previous results [9, 14]. The parameter c is set to be at least $26c_0$. The setting of the threshold θ will be explained later. Let $G(P, c)$ denote the adaptive neighborhood graph as defined by Giesen and Wagner [14].

DIMENSION(p, c): /* $c \geqslant 26c_0$ */

1. Let $N(p, c)$ denote the set including p and its neighbors in $G(P, c)$. Let $Z(p, c)$ denote the set of vectors $\{q - p : q \in N(p, c) \wedge q \neq p\}$.
2. Compute the non-zero eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \ldots$ of the covariance matrix for $Z(p, c)$ in $O(d|Z(p, c)|^2)$ time.
3. Find the smallest j such that $\lambda_{j+1}/\lambda_1 < \theta$ and output j as the dimension.

In the following, we derive the setting of θ so that DIMENSION(p, c) reports the true dimension.

Since P is an (ϵ, δ)-sampling, the nearest neighbor distance of any sample $x \in P$ is at most $2\epsilon f(x)/(1 - \epsilon)$ [14]. Therefore, for any sample $q \in N(p, c)$, $\|p - q\| \leqslant 2c\epsilon \cdot \max\{f(p), f(q)\}/(1 - \epsilon) \leqslant 2c\epsilon f(p)/(1 - \epsilon) + 2c\epsilon\|p - q\|/(1 - \epsilon)$. This implies that $\|p - q\| \leqslant 2c\epsilon f(p)/(1 - (2c + 1)\epsilon)$, which is less than $3c\epsilon f(p)$ given our assumptions about ϵ in Lemma 4.7 and that $c \geqslant 26c_0$. On the other hand, the nearest neighbor distance of p is at least $\delta f(p)$ which is at least $\epsilon f(p)/c_0$ as $\epsilon/\delta \leqslant c_0$. We conclude that

$$P_p(26) \;\subseteq\; P_p(c/c_0) \;\subseteq\; N(p, c) \;\subseteq\; P_p(3c),$$

$$X_p(26) \subseteq X_p(c/c_0) \subseteq Z(p,c) \subseteq X_p(3c).$$

Let $\lambda_1 \geqslant \lambda_2 \ldots$ be the non-zero eigenvalues of the covariance matrix for $Z(p,c)$. Since $X_p(26) \subseteq Z(p,c) \subseteq X_p(3c)$, the lower and upper bounds in Lemma 4.7(i) and (ii) imply that

$$\forall\, 1 \leqslant j \leqslant k, \quad \alpha_1(26)\epsilon^2 f(p)^2 \leqslant \lambda_j \leqslant \alpha_2(3c)\epsilon^2 f(p)^2. \tag{2}$$

Similarly, the upper bound in Lemma 4.7(iii) implies that

$$\lambda_{k+1} \leqslant \alpha_3(3c)\epsilon^4 f(p)^2. \tag{3}$$

So the following inequalities hold:

$$\forall\, 2 \leqslant j \leqslant k, \quad \frac{\lambda_j}{\lambda_1} \geqslant \frac{\alpha_1(26)}{\alpha_2(3c)}. \qquad\qquad \frac{\lambda_{k+1}}{\lambda_1} \leqslant \frac{\epsilon^2\alpha_3(3c)}{\alpha_1(26)}$$

We want to enforce $\theta \leqslant \frac{\alpha_1(26)}{\alpha_2(3c)}$ so that DIMENSION(p,c) will not terminate before reaching λ_{k+1}. It can be checked that $\frac{\alpha_1(26)}{\alpha_2(3c)} = \frac{16^{k+2}}{2^{5k+1}c_0^k(3c+1)^{k+2}} = \left(\frac{16}{32c_0(3c+1)}\right)^k \cdot \frac{256}{2(3c+1)^2} \geqslant \left(\frac{1}{2c_0(3c+1)^3}\right)^k$. By Lemma 4.5, $|N(p,c)| \geqslant |P_p(26)| \geqslant ((24\epsilon - 4\xi)/(4\xi))^k$. As $\xi \leqslant 2\epsilon$, $|N(p,c)| \geqslant 2^k$. Therefore, we can compute the constant $h = \lceil \log_2 2c_0(3c+1)^3 \rceil$ and set $\theta = |N(p,c)|^{-h}$. Then $\theta \leqslant \left(\frac{1}{2c_0(3c+1)^3}\right)^k \leqslant \frac{\alpha_1(26)}{\alpha_2(3c)}$. Notice that θ is set without the knowledge of k.

As long as ϵ becomes sufficiently small, $\epsilon^2\alpha_3(3c)/\alpha_1(26) < \theta$ and so DIMENSION(p,c) will stop exactly at λ_{k+1}. This constrains ϵ to be $O(2^{-\Theta(k)})$.

In all, we have shown that θ can be chosen so that the true dimension is reported. The running time of DIMENSION(p,c) is dominated by the eigenvalue computation which takes $O(d|Z(p,c)|^2) = O(d2^{O(k)})$ time.

Theorem 4.1 *Let P be an (ϵ,δ)-sampling of a smooth manifold in $\mathbb{R}^d$, where $\epsilon/\delta \leqslant c_0$ for some constant c_0. Assume that $G(P,c)$ is given for some $c \geqslant 26c_0$. Let p be a sample in P and let k be the dimension of the manifold component containing p. There exists a value ϵ_0 depending on c, c_0, and k such that if $\epsilon \leqslant \epsilon_0$,* DIMENSION$(p,c)$ *outputs k in $O(d2^{O(k)})$ time.*

4.5 Experimental results

We developed an implementation and collected some experimental results. The value of the threshold θ predicted by theory is pessimistically small and it only works when the sampling density is extremely high. We try to be more optimistic in the experiments and set θ arbitrarily to $1/4$. Our program measures some statistics that reflect the trustworthiness of the output dimension. We will discuss this later.

Our data include a torus and unit k-spheres for $k = 3, 4, 5$. The torus is in the form of a triangular mesh and we use the vertices as samples. Each unit k-sphere is centered at the origin. The point samples are generated on the sphere as follows. We

Data	Size	Avg. no. of neighbors	Avg. % of correct votes	Avg. cutoff
Torus	4096	19.92	100%	0.0361
3-sphere	150000	35.57	93.2%	0.036
4-sphere	150000	64.98	100%	0.039
5-sphere	150000	128.68	98.9%	0.125

Table 1: Experimental results.

first randomly generate k-dimensional points with coordinates uniformly picked between -1 and 1. Then we project each point onto the k-sphere.

For each set of data, we vary d to see its effect on the running time. We append zeros to the coordinates of the samples to lift them to $\mathbb{R}^d$. Although the zeros give no advantage to our dimension detection algorithm, we reflect the samples about a random hyperplane to remove this structure in the coordinates of the point samples. Since the data is imperfect, we perform dimension detection using 11 randomly picked point samples and their neighbors. Then we take a majority vote to decide the dimension of the data. The neighbors of each point sample p is decided as follows. We compute the fourth nearest neighbor q of p and collect all samples within a distance of $2\|p-q\|$ from p. We use the fourth nearest neighbor instead of the nearest neighbor of p just in case the data is imperfect.

The experiments were run on a Pentium 4 with a 3.2GHz CPU and 1GB RAM. The eigenvalues were computed using Matlab. Table 1 shows the total number of samples in the data set, the average number of neighbors per sample used in dimension detection, the voting statistics, and the average *cutoff* λ_{k+1}/λ_k. For each data set, the averages are taken over all the trials for all values of d experimented (11 trials per value of d). The voting statistics is the average percentage of correct votes. The voting statistics and the average cutoff reflect how trustworthy the output dimension is.

Our algorithm reports the true dimension after the majority vote in all cases. The average cutoff shows that λ_{k+1} is roughly an order of magnitude less than λ_k (and hence $\lambda_1, \lambda_2, \ldots, \lambda_{k-1}$ too). In the sphere experiments, the average cutoff increases steadily from the 3-sphere case to the 5-sphere case. This reflects the relative decrease in sampling density as k increases. Figure 4 shows the total running time of the 11 trials as d is varied from 10 to 1000. We do not include the time to compute the neighbors of the samples. As predicted by the theory, the running time does not grow quickly with d. We believe that the fluctuation in the running time for each data set is just experimental abberration and it does not have any significance.

5 Noisy case

In this section, we show that our approach can handle noise to a certain extent. We use three parameters ϵ, δ, and σ to describe our noise model. All three parameters ϵ, δ, and σ are from the interval $[0, 1)$. The set P of noisy samples consists of *non-outliers* and *outliers*. We use Q to denote the subset of non-outliers. For each point $q \in Q$, we use

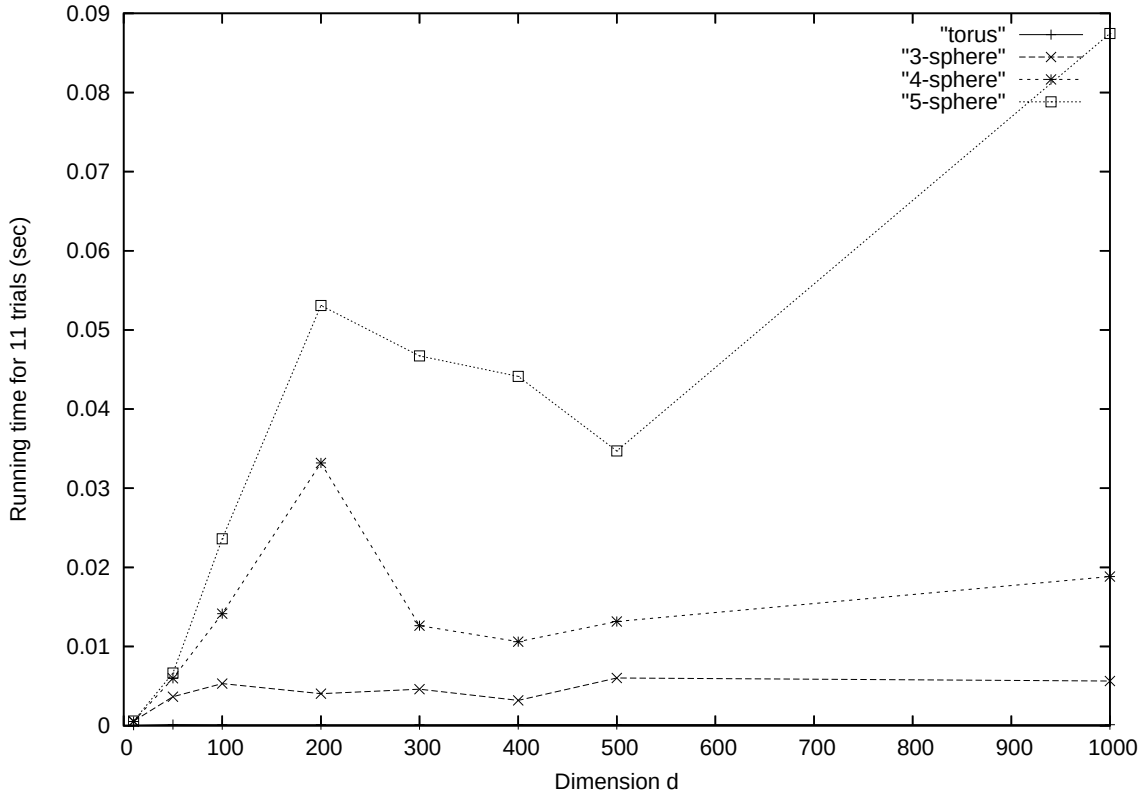

Figure 4: The running time of the 11 trials against the dimension d. The time to compute the neighbors is not included.

$\bar{q}$ to denote its closest point in $\mathcal{M}$. We say that P is an $(\epsilon, \delta, \sigma)$-noisy-sampling if the following conditions are satisfied.

(i) $1 \leqslant \epsilon/\delta \leqslant c_0$ for some constant c_0.

(ii) For any sample $p \in Q$, $\|p - \bar{p}\| \leqslant \epsilon^2 f(\bar{p})$.

(iii) For any point $a \in \mathcal{M}$, there exists $q \in Q$ such that $\|a - \bar{q}\| \leqslant \epsilon f(a)$.

(iv) For any samples $p, q \in Q$, $\|p - q\| \geqslant \delta f(\bar{p})$.

(v) $|P \setminus Q| \leqslant \sigma |P|$.

Conditions similar to (ii) have been used in the literature [6, 9, 17] to model small perturbations of samples in the problems of reconstructing curves and surfaces from noisy samples. But outliers have not been allowed previously. Our proofs will require the parameter ϵ to be $O(2^{-\Theta(k)})$ and the parameter σ to be $O(2^{-\Theta(k)}(\log n)^{-1})$, where $n = |P|$. The condition (v) for outliers is motivated by the usual observation that the positions of outliers are arbitrary but they are sparse. We emphasize that although our model and analysis distinguish between non-outliers and outliers, our algorithm does not require or guess this information.

Our basic strategy for the noiseless case can be carried over to the noisy case. The intuition is that, given the sparsity of the outliers, there is a good chance to pick a neighborhood free of outliers. Then the small noise magnitude will not interfere much with our approximate eigenvalue computation. So our previous approach will still detect the manifold dimension with high probability. However, when the algorithm encounters an outlier, its running time may be very high. Therefore, we need to modify the algorithm in order to lower the probability of such an event.

The detailed analysis is given in the rest of this section. We model it mostly after the one for the noiseless case so as to reuse the previous results. The following result will be useful.

Lemma 5.1 *Assume $\epsilon \leqslant 1/(6c_0)$. Let $\bar{Q} = \{\bar{q} : q \in Q\}$. Then $\bar{Q}$ is an (ϵ, δ')-sampling of $\mathcal{M}$, where $\delta' = \delta - 2\epsilon^2 \geqslant 2\delta/3$.*

Proof. It follows from the definition of $(\epsilon, \delta, \sigma)$-noisy-sampling that for any point $x \in \mathcal{M}$, there is a point $\bar{p} \in \bar{Q}$ such that that $\|\bar{p} - x\| \leqslant \epsilon f(x)$. Take two points $\bar{p}, \bar{q} \in \bar{Q}$. Assume that $f(\bar{p}) \geqslant f(\bar{q})$. We have $\|\bar{p} - \bar{q}\| \geqslant \|p - q\| - \|p - \bar{p}\| - \|q - \bar{q}\|$. By our noisy sampling model, we have $\|p - q\| \geqslant \delta f(\bar{p})$, $\|p - \bar{p}\| \leqslant \epsilon^2 f(\bar{p})$, and $\|q - \bar{q}\| \leqslant \epsilon^2 f(\bar{q})$. It follows that $\|\bar{p} - \bar{q}\| \geqslant (\delta - 2\epsilon^2) f(\bar{p}) = \delta' f(\bar{p})$. So $\|\bar{p} - \bar{q}\| \geqslant \delta' f(\bar{q})$ too. □

5.1 Projection of the non-outliers onto the tangent space

Take a point p in the set Q of non-outliers. Define

$$\mathcal{T}_p = \{x + p - \bar{p} : x \in \mathcal{T}_{\bar{p}}\}, \qquad \mathcal{N}_p = \{x + p - \bar{p} : x \in \mathcal{N}_{\bar{p}}\}.$$

That is, $\mathcal{T}_p$ and $\mathcal{N}_p$ are copies of $\mathcal{T}_{\bar{p}}$ and $\mathcal{N}_{\bar{p}}$, respectively, translated to contain p. Note that $\mathcal{N}_p = \mathcal{N}_{\bar{p}}$ because $p \in \mathcal{N}_{\bar{p}}$ by definition.

Let $\mathcal{C}_p(c)$ denote the ball centered at p with radius $c\epsilon f(\bar{p})$. We use $Q_p(c)$ to denote the set of points $Q \cap \mathcal{C}_p(c)$. Define the following:

- For each $\bar{q} \in \bar{Q}$, let $\widehat{q}$ denote the projection of $\bar{q}$ onto $\mathcal{T}_{\bar{p}}$. Let $\widehat{Q}_p(c)$ denote the set of points $\{\widehat{q} : \bar{q} \in \bar{Q} \cap \mathcal{B}_{\bar{p}}(c)\}$.
- For each $q \in Q$, let q^* denote the projection of q onto $\mathcal{T}_p$. Let $Q_p^*(c)$ denote the set of points $\{q^* : q \in Q_p(c)\}$.

We show that $Q_p^*(c)$ is fairly uniformly distributed.

Lemma 5.2 *Assume that $c > 3$ and $\epsilon \leqslant \frac{1}{60(c+2)^2 c_0}$. Let p be a sample in Q. There exist ψ and ω such that $\psi \in [\epsilon, 2\epsilon]$, $\omega \in [\delta/4, \delta/2]$, and the following hold.*

(i) *For any point $a \in \mathcal{C}_p(c-3) \cap \mathcal{T}_p$, there exists $q^* \in Q_p^*(c)$ such that $\|a - q^*\| \leqslant \psi f(\bar{p})$.*

(ii) *For any points $q, s \in Q_p(c)$, $\|q^* - s^*\| \geqslant 2\omega f(\bar{p})$.*

Proof. Consider (i). By Lemma 5.1, $\bar{Q}$ is an (ϵ, δ')-sampling. So we can apply Lemma 4.3 to $\bar{Q}$ and $\bar{p}$ with the constant c. It follows from Lemma 4.3(i) that

$$\forall \text{ point } x \in \mathcal{B}_{\bar{p}}(c-2) \cap \mathcal{T}_{\bar{p}}, \ \ \exists\, \widehat{s} \in \widehat{Q}_p(c) \text{ such that } \|x - \widehat{s}\| \leqslant \xi f(\bar{p}), \tag{4}$$

where $\xi = \epsilon(1 + (c-2)\epsilon + 4(c-2)^2\epsilon)$. We prove (i) for the constant $\psi = \xi + \epsilon^2 + c\epsilon^3$. Note that $\psi \leqslant \xi + (c+1)\epsilon^2 \leqslant \epsilon(1 + (2c-1)\epsilon + 4(c-2)^2\epsilon) \leqslant (1 + 4c^2\epsilon)\epsilon$. Thus $\psi \in [\epsilon, 2\epsilon]$ as $\epsilon < 1/(4c^2)$ by assumption.

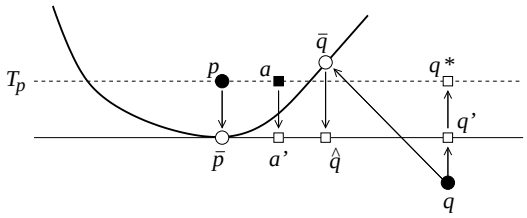

Figure 5: The bold curve denotes $\mathcal{M}$. The solid line denotes $\mathcal{T}_{\bar{p}}$. The dashed line denotes $\mathcal{T}_p$. Each arrow represents a projection.

Let $\beta = c - 3$. Refer to Figure 5. Take a point $a \in \mathcal{C}_p(\beta) \cap \mathcal{T}_p$. Let a' be the projection of a onto $\mathcal{T}_{\bar{p}}$. Since $\mathcal{T}_p$ and $\mathcal{T}_{\bar{p}}$ are parallel, $a' \in \mathcal{B}_{\bar{p}}(\beta) \cap \mathcal{T}_{\bar{p}}$. Since $\beta < c - 2$, by (4), there is a point $\widehat{q} \in \widehat{Q}_p(c)$ such that

$$\|a' - \widehat{q}\| \ \leqslant \ \xi f(\bar{p}). \tag{5}$$

We are to show that q^* satisfies (i). Figure 5 illustrates the situation.

We first show that $q^* \in Q_p^*(c)$. By our sampling condition, $\|p - q\| \leqslant \|\bar{p} - \bar{q}\| + \epsilon^2 f(\bar{p}) + \epsilon^2 f(\bar{q})$. Since $\widehat{q} \in \widehat{Q}_p(c)$, we have $\bar{q} \in \mathcal{B}_{\bar{p}}(c)$ by definition. By the Lipschitz condition,

$$f(\bar{q}) \ \leqslant \ f(\bar{p}) + \|\bar{p} - \bar{q}\| \ \leqslant \ (1 + c\epsilon) f(\bar{p}). \tag{6}$$

Thus we have

$$\|p-q\| \;\leqslant\; \|\bar{p}-\bar{q}\| + (2\epsilon^2 + c\epsilon^3) f(\bar{p}). \tag{7}$$

By the triangle inequality, $\|\bar{p}-\bar{q}\| \leqslant \|\bar{p}-a'\| + \|a'-\widehat{q}\| + \|\widehat{q}-\bar{q}\|$. We already know the bound of $\|a'-\widehat{q}\|$ from (5). Since $a \in \mathcal{C}_p(\beta)$ by assumption, $\|\bar{p}-a'\| = \|p-a\| \leqslant \beta\epsilon f(\bar{p})$. Since $\bar{q} \in \mathcal{B}_{\bar{p}}(c)$, Lemma 4.1 implies that $\|\widehat{q}-\bar{q}\| \leqslant (c^2\epsilon^2/2)\cdot f(\bar{p})$. In all, we have

$$\|\bar{p}-\bar{q}\| \;\leqslant\; (\beta\epsilon + \xi + c^2\epsilon^2/2) f(\bar{p}).$$

Substituting this into (7) yields (note that $\xi \leqslant 2\epsilon$)

$$\begin{aligned} \|p-q\| &\leqslant (\beta\epsilon + \xi + c^2\epsilon^2/2 + 2\epsilon^2 + c\epsilon^3) f(\bar{p}) \\ &\leqslant ((\beta+2)\epsilon + (c^2+2c+4)\epsilon^2/2) f(\bar{p}) \\ &\leqslant (\beta+3)\epsilon f(\bar{p}), \end{aligned}$$

because $\epsilon \leqslant \frac{1}{60(c+2)^2 c_0}$ implies that $(c^2+2c+4)\epsilon/2 \leqslant 1$. Thus $\|p-q\| \leqslant c\epsilon f(\bar{p})$ as $\beta = c-3$. It follows that $q \in Q_p(c)$ and so $q^* \in Q_p^*(c)$.

Next, we show that $\|a-q^*\| \leqslant \psi f(\bar{p})$. Let q' be the projection of q onto $\mathcal{T}_{\bar{p}}$. Observe that $\|a-q^*\| = \|a'-q'\| \leqslant \|a'-\widehat{q}\| + \|\widehat{q}-q'\|$. We already know that $\|a'-\widehat{q}\| \leqslant \xi f(\bar{p})$ from (5). We have $\|\widehat{q}-q'\| \leqslant \|q-\bar{q}\| \leqslant \epsilon^2 f(\bar{q})$. Replacing $f(\bar{q})$ using equation (6), we get

$$\|\widehat{q}-q'\| \;\leqslant\; (\epsilon^2 + c\epsilon^3) f(\bar{p}). \tag{8}$$

Hence, $\|a-q^*\| \leqslant (\xi + \epsilon^2 + c\epsilon^3) f(\bar{p}) = \psi f(\bar{p})$. This finishes the proof of (i).

Consider (ii). By Lemma 5.1, $\bar{Q}$ is an (ϵ, δ')-sampling. So we can apply Lemma 4.3 to $\bar{Q}$ and $\bar{p}$ with the constant $2c+4$. It follows from Lemma 4.3(ii) that

$$\forall\, \bar{q}, \bar{s} \in \bar{Q} \cap \mathcal{B}_{\bar{p}}(2c+4), \;\; \|\widehat{q}-\widehat{s}\| \geqslant 2\eta f(\bar{p}), \tag{9}$$

where $\eta = \delta'/2 - (2c+4)^2\epsilon^2/2$. We prove (ii) for constant $\omega = \eta - \epsilon^2 - (2c+4)\epsilon^3$. Since $\delta' \geqslant 2\delta/3$ by Lemma 5.1, we have $\omega \geqslant \delta/3 - (2c+4)^2\epsilon^2/2 - \epsilon^2 - (2c+4)\epsilon^3 \geqslant \delta/3 - 5(c+2)^2\epsilon^2$. Since $\epsilon \leqslant \frac{1}{60(c+2)^2 c_0}$ by assumption, we have $\omega \geqslant \delta/3 - \epsilon/(12c_0) \geqslant \delta/4$. In all, $\omega \in [\delta/4, \delta/3] \subset [\delta/4, \delta/2]$ as desired.

Take two points $q, s \in Q_p(c)$. Let q' be the projection of q onto $\mathcal{T}_{\bar{p}}$ and let s' be the projection of s onto $\mathcal{T}_{\bar{p}}$. Figure 6 illustrates the situation.

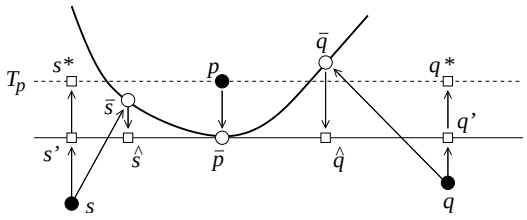

Figure 6: The bold curve denotes $\mathcal{M}$. The solid line denotes $\mathcal{T}_{\bar{p}}$. The dashed line denotes $\mathcal{T}_p$. Each arrow represents a projection.

By the triangle inequality, we have $\|\bar{p}-\bar{q}\| \leqslant \|p-q\| + \epsilon^2 f(\bar{p}) + \epsilon^2 f(\bar{q})$. Since $q \in Q_p(c)$, we have $\|p-q\| \leqslant c\epsilon f(\bar{p})$. Replacing $f(\bar{q})$ using the inequality $f(\bar{q}) \leqslant f(\bar{p}) + \|\bar{p}-\bar{q}\|$

yields

$$\|\bar{p} - \bar{q}\| \;\leqslant\; \frac{c\epsilon + 2\epsilon^2}{1-\epsilon^2} f(\bar{p}) \;\leqslant\; (2c+4)\epsilon f(\bar{p}), \tag{10}$$

as $2\epsilon^2 \leqslant 2\epsilon$ and $1-\epsilon^2 \geqslant 1/2$. So $\bar{q} \in \bar{Q} \cap \mathcal{B}_{\bar{p}}(2c+4)$. Similarly, $\bar{s} \in \bar{Q} \cap \mathcal{B}_{\bar{p}}(2c+4)$. Then by (9), we conclude that $\|\widehat{q} - \widehat{s}\| \geqslant 2\eta f(\bar{p})$. It follows that

$$\begin{aligned} \|q^* - s^*\| &= \|q' - s'\| && (11)\\ &\geqslant \|\widehat{q} - \widehat{s}\| - \|\widehat{q} - q'\| - \|\widehat{s} - s'\| && (12)\\ &\geqslant 2\eta f(\bar{p}) - \|\widehat{q} - q'\| - \|\widehat{s} - s'\|. && (13) \end{aligned}$$

We have $\|\widehat{q} - q'\| \leqslant \|q - \bar{q}\| \leqslant \epsilon^2 f(\bar{q})$. By the Lipschitz condition, $f(\bar{q}) \leqslant f(\bar{p}) + \|\bar{p} - \bar{q}\| \leqslant f(\bar{p}) + (2c+4)\epsilon f(\bar{p})$. This implies that

$$\|\widehat{q} - q'\| \;\leqslant\; (\epsilon^2 + (2c+4)\epsilon^3) f(\bar{p}).$$

Similarly, $\|\widehat{s} - s'\| \leqslant (\epsilon^2 + (2c+4)\epsilon^3) f(\bar{p})$. Substituting these two inequalities into (13) yields $\|q^* - s^*\| \geqslant 2(\eta - \epsilon^2 - (2c+4)\epsilon^3) f(\bar{p}) = 2\omega f(\bar{p})$. This proves (ii). □

5.2 Bounds of Eigenvalues

Let p be a sample in Q. Define $Y_p(c)$ to be the set of vectors $q - p$ for all sample points q in $Q_p(c) \setminus \{p\}$. Lemma 5.2 allows us to obtain the following results.

Lemma 5.3 *Let p be a sample in Q. Assume that $c > 11$ and $\epsilon \leqslant \frac{1}{60(c+2)^2 c_0}$. Then*

(i) $\frac{((c-3)\epsilon - 4\psi)^k}{4^k \psi^k} \leqslant |Q_p(c)| \leqslant \frac{2(c\epsilon+\omega)^k}{\omega^k}$.

(ii) *For any unit vector $u \in \mathcal{T}_p$,* $\frac{f(\bar{p})^2}{k+2} \cdot \frac{((c-3)\epsilon - 4\psi)^{k+2}}{4^k \psi^k} \leqslant \mathrm{var}(Y_p(c), u) \leqslant \frac{2f(\bar{p})^2}{k+2} \cdot \frac{(c\epsilon+2\omega)^{k+2}}{\omega^k}$.

(iii) *For any unit vector $v \in \mathcal{N}_p$,* $\mathrm{var}(Y_p(c), v) \leqslant \frac{18(c+2)^4 \epsilon^4 f(\bar{p})^2 (c\epsilon+\omega)^k}{\omega^k}$.

Proof. Lemma 5.2 allows us to derive the bounds in (i) and (ii) in exactly the same ways as in the proofs of Lemmas 4.4 and 4.5. We just have to substitute ξ by ψ and η by ω. Also, $c-2$ in the lower bounds in Lemma 4.5 becomes $c-3$ in the lower bound of $|Q_p(c)|$ in (i) and $\mathrm{var}(Y_p(c), u)$ in (ii). (This also explains why we require $c > 11$ when compared with the requirement of $c > 10$ in Lemma 4.5.)

Consider (iii). We first bound the distance from q to $\mathcal{T}_p$ for any sample $q \in Q_p(c)$. By the triangle inequality, the distance from q to $\mathcal{T}_p$ is no more than the sum of the distances between q and $\bar{q}$, between $\bar{q}$ and $\mathcal{T}_{\bar{p}}$, and between $\mathcal{T}_{\bar{p}}$ and $\mathcal{T}_p$.

The distance between $\mathcal{T}_{\bar{p}}$ and $\mathcal{T}_p$ is at most $\epsilon^2 f(\bar{p})$. The distance $\|q - \bar{q}\|$ is at most $\epsilon^2 f(\bar{q}) \leqslant \epsilon^2 f(\bar{p}) + \epsilon^2 \|\bar{p} - \bar{q}\|$ by the Lipschitz condition. By equation (10) in the proof of Lemma 5.2, we have $\|\bar{p} - \bar{q}\| \leqslant (2c+4)\epsilon f(\bar{p})$. Since $\epsilon \leqslant \frac{1}{60(c+2)^2 c_0}$, we get $\|\bar{p} - \bar{q}\| < f(\bar{p})$. So $\|q - \bar{q}\| \leqslant 2\epsilon^2 f(\bar{p})$. Because $\|\bar{p} - \bar{q}\| \leqslant (2c+4)\epsilon f(\bar{p})$, Lemma 4.1 implies that the distance from $\bar{q}$ to $\mathcal{T}_{\bar{p}}$ is at most $(2c+4)^2 \epsilon^2 f(\bar{p})/2 = 2(c+2)^2 \epsilon^2 f(\bar{p})$.

Hence, the distance from q to $\mathcal{T}_p$ is at most $(3\epsilon^2 + 2(c+2)^2\epsilon^2)f(\bar{p}) \leqslant 3(c+2)^2\epsilon^2 f(\bar{p})$. Therefore, for any unit vector $v \in \mathcal{N}_p$, $\mathrm{var}(Y_p(c), v) \leqslant |Q_p(c)| \cdot 9(c+2)^4\epsilon^4 f(\bar{p})^2$. Substituting the upper bound of $|Q_p(c)|$ in (i) proves that $\mathrm{var}(Y_p(c), v) \leqslant \frac{18(c+2)^4\epsilon^4 f(\bar{p})^2(c\epsilon+\omega)^k}{\omega^k}$. □

Define the following functions:

$$\gamma_1(c) = \frac{(c-11)^{k+2}}{(k+2)2^{3k}}, \quad \gamma_2(c) = \frac{(c+1)^{k+2}2^{2k+1}c_0^k}{k+2}, \quad \gamma_3(c) = 18(c+2)^4(4cc_0+1)^k.$$

By substituting the inequalities $\psi \leqslant 2\epsilon$, $\delta/4 \leqslant \omega \leqslant \delta/2 \leqslant \epsilon/2$, and $\epsilon/\delta \leqslant c_0$ into the bounds in Lemma 5.3(ii) and (iii), we conclude that $\gamma_1(c)\epsilon^2 f(\bar{p})^2$ is at most the lower bound of $\mathrm{var}(Y_p(c), u)$ in Lemma 5.3(ii), $\gamma_2(c)\epsilon^2 f(\bar{p})^2$ is at least the upper bound of $\mathrm{var}(Y_p(c), u)$ in Lemma 5.3(ii), and $\gamma_3(c)\epsilon^4 f(\bar{p})^2$ is at least the upper bound of $\mathrm{var}(Y_p(c), v)$ in Lemma 5.3(iii). Given a set Y of d-dimensional vectors, Lemma 5.3(ii) and (iii) allow us to bound the eigenvalues of the covariance matrix for Y in the same way as in Lemma 4.7, when $Y \subseteq Y_p(c)$ or $Y_p(c) \subseteq Y$. In all, we have the following result.

Lemma 5.4 *Let p be a sample in Q. Let Y be a set of d-dimensional vectors. Let $\lambda_1 \geqslant \lambda_2 \ldots \geqslant \lambda_d$ be the eigenvalues of the covariance matrix for Y. Let $k = \dim(\mathcal{T}_p)$. Assume that $c > 11$. There exists ϵ_0 depending on c, c_0 and k such that if $\epsilon \leqslant \epsilon_0$, then*

(i) *if $Y_p(c) \subseteq Y$, then $\lambda_j \geqslant \gamma_1(c)\epsilon^2 f(p)^2$ for $1 \leqslant j \leqslant k$,*

(ii) *if $Y \subseteq Y_p(c)$, then $\lambda_j \leqslant \gamma_2(c)\epsilon^2 f(p)^2$ for $1 \leqslant j \leqslant k$, and*

(iii) *if $Y \subseteq Y_p(c)$, then $\lambda_j \leqslant \gamma_3(c)\epsilon^4 f(p)^2$ for $k+1 \leqslant j \leqslant d$.*

5.3 Sparsity of outliers

The adaptive neighborhood graph is an undirected graph in the noiseless case. But this does not work in the presence of outliers. An outlier far away could become the neighbor of all other vertices, and in this case, the neighbors of a non-outlier are not necessarily in close proximity as desired. This calls for a *directed version* of the adaptive neighborhood graph. For every sample $p \in P$, we assign an arc from p to other samples $q \in P$ such that $\|p - q\|$ is no more than c times the nearest neighbor distance of p. We denote this set of neighbors by $DN(p, c)$ for every $p \in P$ and the resulting directed adaptive neighborhood graph by $DG(P, c)$. Comparing with the undirected adaptive neighborhood graph, $DN(p, c)$ is a subset of $N(p, c)$ but it suffices for our purposes.

We quantify the sparsity of outliers by showing that $DN(p, c)$ is free of outliers for many choices of p. We first show a technical result. Recall that for a sample $p \in Q$, $\mathcal{C}_p(c)$ is the ball centered at p with radius $c\epsilon f(\bar{p})$.

Lemma 5.5 *Assume that $\epsilon \leqslant 1/(6c_0)$. Then $DN(p, c) \subseteq \mathcal{C}_p(3c)$ for any sample $p \in Q$.*

Proof. Take a sample $p \in Q$. Since $\bar{Q}$ is an (ϵ, δ')-sampling by Lemma 5.1, it has been shown [14] that the distance between $\bar{p}$ and the nearest point $\bar{q} \in \bar{Q}$ is at most $2\epsilon f(\bar{p})/(1-\epsilon)$. The Lipschitz condition implies that $f(\bar{q}) \leqslant (1+\epsilon)f(\bar{p})/(1-\epsilon)$. Therefore,

$\|p-q\| \leqslant \frac{2\epsilon}{(1-\epsilon)} f(\bar{p}) + \epsilon^2 f(\bar{p}) + \epsilon^2 f(\bar{q}) \leqslant \frac{2\epsilon}{(1-\epsilon)} f(\bar{p}) + \epsilon^2 f(\bar{p}) + \frac{\epsilon^2(1+\epsilon)}{1-\epsilon} f(\bar{p})$. It can be checked that this bound is no more than $3\epsilon f(\bar{p})$ for $\epsilon \leqslant 1/5$. Since $\|p-q\|$ is at least the distance from p to its nearest neighbor in $DN(p,c)$, we conclude that $DN(p,c) \subseteq \mathcal{C}_p(3c)$. □

Lemma 5.6 *Assume that* $\epsilon \leqslant \frac{1}{60(c+2)^2 c_0}$. *Let* $l \geqslant 1$ *and let* $\sigma \leqslant \frac{1}{24l(24cc_0+1)^k}$. *If a sample* p *is picked uniformly randomly from* P, $DN(p,c)$ *contains an outlier with probability less than* $1/(8l)$.

Proof. Let $o \in P \setminus Q$ be an outlier. Let S be the set of non-outliers $p \in Q$ such that $o \in \mathcal{C}_p(3c)$. We first prove an upper bound of $|S|$. Take the non-outlier $q \in S$ with the maximum $f(\bar{q})$. For all $p \in S$, since $\mathcal{C}_p(3c)$ and $\mathcal{C}_q(3c)$ overlap, we have $\|p-q\| \leqslant 3c\epsilon f(\bar{p}) + 3c\epsilon f(\bar{q}) \leqslant 6c\epsilon f(\bar{q})$. Therefore, S lies inside $\mathcal{C}_q(6c)$ which implies that $S \subseteq Q_q(6c)$. By Lemma 5.3(i), we have $|S| \leqslant |Q_q(6c)| \leqslant 2(6c\epsilon/\omega + 1)^k$. Substituting the inequalities $\omega \geqslant \delta/4$ and $\epsilon/\delta \leqslant c_0$ yields the bound $2(24cc_0+1)^k$.

There are at most $\sigma|P|$ outliers by our noisy sampling condition. It follows that there at most $2\sigma(24cc_0+1)^k|P| \leqslant |P|/(12l)$ non-outliers p such that $\mathcal{C}_p(3c)$ contains an outlier. By Lemma 5.5, we conclude that there are at most $|P|/(12l)$ non-outliers p whose $DN(p,c)$ contains an outlier. Hence, if we pick a sample p from P uniformly randomly, the probability of picking a sample p whose $DN(p,c)$ contains an outlier is at most $1/(12l) + \sigma < 1/(12l) + 1/(24l) = 1/(8l)$. □

5.4 Algorithmic results

We assume the knowledge of the bound c_0 on ϵ/δ as in previous results [9, 14]. We also assume the directed adaptive neighborhood graph $DG(P,c)$ is given. The constant c is set to be at least $27c_0$. The basic idea is to pick $4l-1$ random examples for some $l \geqslant 1$, run PCA, and then take a majority vote to determine the manifold dimension. There are slight complications in order to get a good expected running time. The details are summarized in the following pseudocode.

NoisyDim(c, l): /* $c \geqslant 27c_0$ and $l \geqslant 1$ */

1. Pick $4l-1$ random samples $p_1, \ldots, p_{4l-1}$ from P. For $1 \leqslant i \leqslant 4l-1$, let $Z(p_i, c)$ denote the set of vectors $\{q - p_i : q \in DN(p_i, c) \wedge q \neq p_i\}$.
2. Throw away the p_i's with the $2l$ largest neighborhood sizes $|DN(p_i, c)|$.
3. Let $h = \lceil \log_2 2c_0(3c+1)^3 \rceil$. For each surviving p_i, set $\theta_i = |DN(p_i, c)|^{-h}$.
4. For each surviving p_i, do the following:
   (a) Compute the non-zero eigenvalues $\lambda_{i,1} \geqslant \lambda_{i,2} \geqslant \cdots$ of the covariance matrix for $Z(p_i, c)$ in $O(d|Z(p_i, c)|^2)$ time.
   (b) Find the smallest j such that $\lambda_{i,j}/\lambda_{i,1} < \theta_i$ and record $j-1$ as p_i's vote.
5. Output the majority vote as the manifold dimension.

If we accidentally pick an outlier p_i far away from other points in P in step 1, $DN(p_i, c)$ may contain a huge number of samples in P. In this case, it will be very time-consuming to run PCA on $Z(p_i, c)$. Therefore, we introduce step 2 to lower the probability of this event. In the rest of this section, we analyze the performance of NoisyDim.

Lemma 5.7 *Assume that $l \geqslant 1$ and $c \geqslant 27c_0$. Let k be the manifold dimension. Let S be the set of surviving sample points after step 2, and let $K \subseteq S$ be the subset of samples whose $DN(p_i, c)$ contains an outlier. There exists ϵ_0 depending on c, c_0 and k such that if $\epsilon \leqslant \epsilon_0$ and $\sigma \leqslant \frac{1}{24l(24cc_0+1)^k}$, the following hold.*

(i) *The probability that $|K| \geqslant l$ is at most 2^{-l}.*

(ii) *For any sample $p \in S \setminus K$, $Q_p(27) \subseteq Q_p(c/c_0) \subseteq DN(p, c) \subseteq Q_p(3c)$. The cardinality of $Q_p(27)$ is at least 2^k.*

(iii) *Let $m = \max\{|DN(p_i, c)| : p_i \in S\}$. It holds with probability at least $1 - 2^{-2l}$ that $m \leqslant 2(12cc_0 + 1)^k$.*

(iv) *For each $p_i \in S \setminus K$, $\lambda_{i,j}/\lambda_{i,1} \geqslant \theta_i$ for $1 \leqslant j \leqslant k$ and $\lambda_{i,k+1}/\lambda_{i,1} < \theta_i$.*

Proof. Consider (i). Let K' be the subset of the $4l-1$ samples picked in step 1 whose $DN(p_i, c)$ contains an outlier. It suffices to show that $\text{Prob}(|K'| \geqslant l) \leqslant 2^{-l}$. By Lemma 5.6, for any $p_i \in K'$, the probability of $DN(p_i, c)$ containing an outlier is at most $1/(8l)$. So $\text{Prob}(|K'| \geqslant l) \leqslant \binom{4l-1}{l}(8l)^{-l} \leqslant (4l)^l(8l)^{-l} = 2^{-l}$.

Consider (ii). By Lemma 5.5, $DN(p, c) \subseteq \mathcal{C}_p(3c)$ for any non-outlier p. Since $DN(p, c)$ is free of outliers, we have $DN(p, c) \subseteq Q_p(3c)$. By our noisy sampling condition, the distance from p to the nearest non-outlier is at least $\delta f(\bar{p}) \geqslant \epsilon f(\bar{p})/c_0$ as $\epsilon/\delta \leqslant c_0$. Thus

$Q_p(27) \subseteq Q_p(c/c_0) \subseteq DN(p,c)$. By Lemma 5.3(i) and the fact that $\psi \leqslant 2\epsilon$, we have $|Q_p(27)| \geqslant 2^k$.

Consider (iii). For a sample $p \in S \setminus K$, by (ii), we have $|DN(p,c)| \leqslant |Q_p(3c)|$ for any non-outlier p. By Lemma 5.3(i) and the fact that $\epsilon/\delta \leqslant c_0$ and $\omega \geqslant \delta/4$, we get $|DN(p,c)| \leqslant |Q_p(3c)| \leqslant 2(12cc_0+1)^k$. Let $p_i \in S$ be the sample with largest neighborhood size $|DN(p_i,c)|$, i.e., $|DN(p_i,c)| = m$. If $m > 2(12cc_0+1)^k$, $DN(p_i,c)$ must contain an outlier. Since all the samples thrown away in step 2 have neighborhood size no less than m, their neighborhood contain outliers as well. Thus $\mathrm{Prob}(m > 2(12cc_0+1)^k)$ is at most the probability that $2l$ or more samples have outliers in their neighborhood. By Lemma 5.6, this probability is at most $\binom{4l-1}{2l}(8l)^{-2l} \leqslant (4l)^{2l}(8l)^{-2l} = 2^{-2l}$.

The arguments for proving (iv) is similar to the analysis in Section 4.4 in the noiseless case. The result in (ii) and Lemma 5.4 imply that

- $\forall\, 1 \leqslant j \leqslant k,\ \ \gamma_1(27)\epsilon^2 f(\bar{p}_i)^2 \leqslant \lambda_{i,j} \leqslant \gamma_2(3c)\epsilon^2 f(\bar{p}_i)^2$.
- $\lambda_{i,k+1} \ \leqslant\ \gamma_3(3c)\epsilon^4 f(\bar{p}_i)^2$.

We have $\frac{\gamma_1(27)}{\gamma_2(3c)} = \frac{16^{k+2}}{2^{5k+1}c_0^k(3c+1)^{k+2}} = \left(\frac{16}{32c_0(3c+1)}\right)^k \cdot \frac{256}{2(3c+1)^2} \geqslant \left(\frac{1}{2c_0(3c+1)^3}\right)^k$. By (ii), $|DN(p_i,c)| \geqslant 2^k$. Since step 3 enforces that $\theta_i \leqslant |DN(p_i,c)|^{-h} \leqslant 2^{-hk} \leqslant \left(\frac{1}{2c_0(3c+1)^3}\right)^k \leqslant \frac{\gamma_1(27)}{\gamma_2(3c)}$, NoisyDim will not terminate before reaching $\lambda_{i,k+1}$. As long as ϵ becomes sufficiently small, $\epsilon^2\gamma_3(3c)/\gamma_1(27)$ is less than θ_i, so NoisyDim will stop at $\lambda_{i,k+1}$. □

We are ready to show that NoisyDim outputs the correct manifold dimension with high probability.

Lemma 5.8 *Assume that $c \geqslant 27c_0$. Let k be the manifold dimension. There exists ϵ_0 depending on c, c_0 and k such that if $\epsilon \leqslant \epsilon_0$ and $\sigma \leqslant \frac{1}{24l(24cc_0+1)^k}$, NoisyDim$(c,l)$ outputs the correct manifold dimension with probability at least $1-2^{-l}$.*

Proof. Let L be the subset of surviving p_i's after step 2 whose $DN(p_i,c)$ is free of outliers. By Lemma 5.7(i), L consists of the majority of the surviving p_i's with probability at least $1-2^{-l}$. By Lemma 5.7(iv), the vote of a $p_i \in L$ is the correct manifold dimension. Thus, NoisyDim reports the correct manifold dimension with probability at least $1-2^{-l}$. □

Next, we analyze the expected running time of NoisyDim. In the following, we set $l = \lceil \log_2 n \rceil$ to get a good expected running time.

Lemma 5.9 *Assume that $c \geqslant 27c_0$. Let $l = \lceil \log_2 n \rceil$. Let k be the manifold dimension. There exists ϵ_0 depending on c, c_0 and k such that if $\epsilon \leqslant \epsilon_0$ and $\sigma \leqslant \frac{1}{24l(24cc_0+1)^k}$, NoisyDim$(c,l)$ runs in $O(d2^{O(k)}\log n)$ expected time.*

Proof. We assume that the graph $DG(P,c)$ is given in the adjacency lists representation. If each vertex in $DG(P,c)$ stores the number of its neighbors, step 2 can be done in $O(l)$ time. Otherwise, we scan the adjacency lists of the p_i's one neighbor at a time in a round-robin fashion. We stop the round-robin scanning as soon as we have exhausted the

adjacency list of $2l-1$ p_i's. They form the set of surviving p_i's. Recall that m is the maximum cardinality of neighborhood for surviving samples, as defined in Lemma 5.7(iii). So steps 1 and 2 take $O(ml)$ time, where m is the size of the largest neighborhood of the surviving samples after step 2. By Lemma 5.7(iii), $m \leqslant 2(12cc_0+1)^k = 2^{O(k)}$ with probability at least $1-2^{-2l}$ and we always have $m \leqslant n$. Therefore, the expected running time of steps 1 and 2 is bounded by $O(l2^{O(k)} + 2^{-2l}nl)$. Since $l = \lceil \log_2 n \rceil$, this simplifies to $O(2^{O(k)} \log n)$. The running time to compute the non-zero eigenvalues for all surviving p_i's is $O(ldm^2)$. So the expected running time is bounded by $O(ld2^{O(k)} + 2^{-2l}ldn^2)$. Since $l = \lceil \log_2 n \rceil$, the second term becomes $O(ld)$. So the expected time of step (iv) is $O(d2^{O(k)} \log n)$. □

When we substitute $l = \lceil \log_2 n \rceil$, the probability bound in Lemma 5.8 is at least $1 - 1/n$. The following theorem summarizes our result for the noisy case.

Theorem 5.1 *Let P be an $(\epsilon, \delta, \sigma)$-noisy-sampling of a smooth connected k-dimensional manifold in $\mathbb{R}^d$, where $\epsilon/\delta \leqslant c_0$ for some constant c_0. Define n to be the size of P, c to be a constant at least $27c_0$, and $l = \lceil \log_2 n \rceil$. Assume that $DG(P, c)$ is given. Then there exists a value ϵ_0 depending on c, c_0, and k such if $\epsilon \leqslant \epsilon_0$ and $\sigma \leqslant \frac{1}{24l(24cc_0+1)^k}$, NoisyDim$(c, l)$ reports the manifold dimension with probability $1 - 1/n$ in $O(d2^{O(k)} \log n)$ expected time.*

5.5 Experimental results

We add outliers and perturb the data used in the experiments for the noiseless case. For each set of data, we take its smallest bounding box, expand it by 20%, and then generate outliers in the box uniformly. Each non-outlier sample p is perturbed by displacing it in a random direction with an amount randomly chosen between 0 and half of the nearest neighbor distance of p. We use the same parameters for the experiments as in the noiseless case. Table 2 shows the results. Figure 7 shows the plot of the total running time against d for $l = 6$. (We used a more optimistic setting of l than what the theory predicts.) That is, NoisyDim draws $4l - 1 = 23$ random samples, keeps the eleven with the smallest number of neighbors, and runs PCA on them. We do not include the time to compute the neighbors. As shown in Table 2, our implementation correctly detects the dimension for all the data. The average cutoff is rather large for the 5-sphere data, when compared with the threshold $\theta = 1/4$ that we used, to have high confidence in the result. A higher sampling density is needed for the 5-sphere data to counteract the noise perturbations and outliers. As predicted by the theory, the total running time does not grow quickly with d. If one compares Figure 7 with Figure 4 in the noiseless case, one notices that the running time in the noisy case is smaller. The cause is that, for each data set, the average number of neighbors per sample decreases after pruning away the 12 samples with the largest number of neighbors.

6 Conclusion

We presented dimension detection algorithms based on principal component analysis and their analysis. Our experiments show that the solution quality is robust against outliers

Data	Size	# Outliers	Avg. % of correct votes	Avg. cutoff
Torus	4096	409	98.9%	0.079
3-sphere	150000	15000	96.6%	0.069
4-sphere	150000	15000	94.3%	0.127
5-sphere	150000	15000	98.9%	0.186

Table 2: Experimental Results.

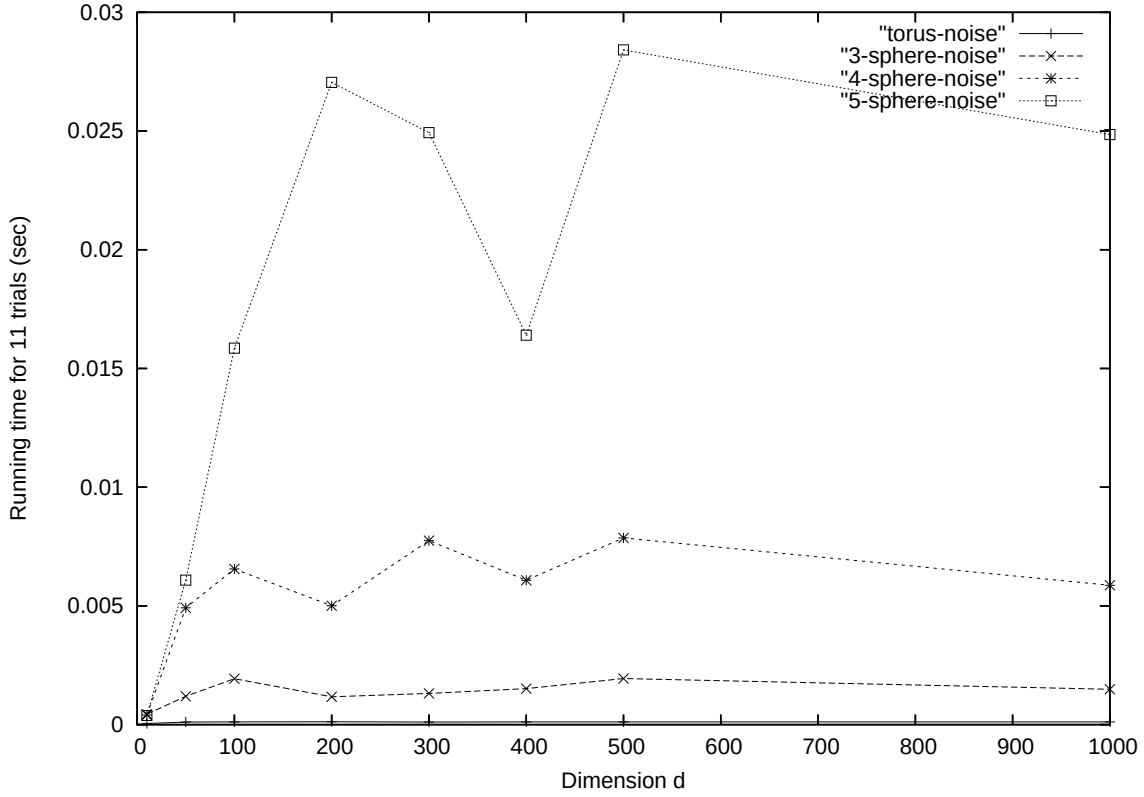

Figure 7: The running time of NoisyDim against the dimension d with $l = 6$. NoisyDim draws 23 random samples, keeps the 11 with the smallest number of neighbors, and runs PCA on them. The time to compute the neighbors is not included.

and small perturbation of samples. In the conference version of this paper [7], we claimed that the adaptive neighborhood graph also allows a good approximation of the geodesic distances in the special noisy case in which there is no outlier. We believe that a routine adaptation of the original proofs in Giesen and Wagner [14] should work, so we decided not to include this claim in the journal version.

Acknowledgment

We thank Beifang Chen, Mordecai Golin, Sheung-Hung Poon, and Jieping Ye for helpful discussions. We also thank an anonymous referee for very helpful comments.

References

[1] N. Amenta and M. Bern. Surface reconstruction by Voronoi filtering. *Discrete and Computational Geometry*, 22, 1999, 481–504.

[2] N. Amenta, S. Choi, T. Dey and N. Leekha. A simple algorithm for homeomorphic surface reconstruction. *International Journal on Computational Geometry & Applications*, 12, 2002, 125–141.

[3] N. Amenta, S. Choi and R.K. Kolluri. The power crust, unions of balls, and the medial axis transform. *Computational Geometry: Theory and Applications*, 19, 2001, 127–153.

[4] D. Attali, J.-D. Boissonnat, and A. Lieutier. Complexity of the Delaunay triangulation of points on surfaces: the smooth case. *Proc. 19th Annu. ACM Sympos. Comput. Geom.*, 2003, 237–246.

[5] S.-W. Cheng, T.K. Dey, and E.A. Ramos. Manifold reconstruction from point samples. *Proc. 16th Annu. ACM-SIAM Sympos. Discrete Alg.*, 2005, 1018–1027.

[6] S.-W. Cheng, S. Funke, M. Golin, P. Kumar, S.-H. Poon, and E. A. Ramos. Curve Reconstruction from Noisy Samples. *Computational Geometry: Theory and Applications*, 31 (2005), 63–100.

[7] S.-W. Cheng, Y. Wang, and Z. Wu. Provable dimension detection using principal component analysis. *Proc. 21st Annu. Sympos. Comput. Geom.*, 2005, 208–217.

[8] T. K. Dey. Curve and surface reconstruction. *Handbook of Discrete and Computational Geometry*, Goodman and O' Rourke eds., CRC press, 2nd edition, 2004.

[9] T. K. Dey, J. Giesen, S. Goswami and W. Zhao. Shape dimension and approximation from samples. *Discrete and Computational Geometry*, 29 (2003), 419–434.

[10] T.K. Dey and S. Goswami. Provable surface reconstruction from noisy samples. *Proc. 21th Annu. ACM Sympos. Comput. Geom.*, 2004, 330–339.

[11] H. Edelsbrunner. *Algorithms in combinatorial geometry*, Springer-Verlag, 1987.

[12] J. Friedman. Computing Betti Numbers via Combinatorial Laplacians. *Algorithmica*, 21 (1998), 331–346.

[13] K. Fukunaga, D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transaction on Computers*, 20 (1971), 176–183.

[14] J. Giesen and U. Wagner. Shape dimension and intrinsic metric from samples of manifolds with high co-dimension. *Proc. 19th Annu. ACM Sympos. Comput. Geom.*, 2003, 329–337.

[15] G.H. Golub and C.F. van Loan. *Matrix Computations*, The Johns Hopkins University Press, 1996.

[16] I. Jolliffe. *Principal Component Analysis*, Springer-Verlag, New York, 1986.

[17] R.K. Kolluri. Provable moving least squares surfaces. *Proc. 16th Annu. ACM-SIAM Sympos. Discrete Alg.*, 2005, 1008–1017.

[18] B. N. Parlett. *The Symmetric Eigenvalue Problem*, SIAM, 1998.

[19] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, Cambridge University Press, 1997.

[20] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (2000), 2319–2323.

[21] S.-T. Yau and Y.Y. Lu. Reducing the symmetric matrix eigenvalue problem to matrix multiplications. *SIAM J. Scientific Computing*, 14 (1993), 121–136.