Dimensionality estimation without distances

Matthäus Kleindessner
University of Hamburg
kleindessner@informatik.uni-hamburg.de

Ulrike von Luxburg
University of Hamburg
luxburg@informatik.uni-hamburg.de

Abstract

While existing methods for estimating the intrinsic dimension of datasets require to know distances between data points, we consider a situation where one has considerably less information. Given a sample of points, all we get to see is who are the k nearest neighbors of every point. In other words, we get the adjacency matrix of the directed, unweighted k-nearest neighbor graph on the sample, but do not know any point coordinates or distances between the points. We provide two estimators for this situation, a naive one and a more elaborate one. Both of them can be proved to be statistically consistent. However, further theoretical and experimental evidence shows that the naive estimator performs rather poorly, whereas the elaborate one achieves results comparable to those of estimators based on distance information.

1 INTRODUCTION

In this paper we consider the problem of estimating the intrinsic dimension of a dataset observed in a high-dimensional space. This is a well-studied problem in the context of dimensionality reduction, with many publications around the time of the development of multidimensional scaling and with renewed attention after the invention of manifold learning algorithms (Trunk, 1968; Bennett, 1969; Fukunaga and Olsen, 1971; Pettis et al., 1979; Grassberger and Procaccia, 1983; Camastra and Vinciarelli, 2002; Kégl, 2002; Costa and Hero, 2004; Costa et al., 2005; Hein and Audibert, 2005; Levina and Bickel, 2005; Farahmand et al., 2007; Sricharan et al., 2010; Eriksson and Crovella, 2012). All these existing methods require the distance matrix of a data sample, or at least parts of it, as input. We are interested in a situation where we have considerably less information. Instead of knowing actual distance values, we just assume to have some knowledge about their relative ordering. Specifically, we assume to only know who are the k nearest neighbors of each data point, but we do not know anything more than that (like distances to the neighbors).

Why would anyone be interested in such a scenario? It turns out that in many modern applications of machine learning it is relatively easy to gather information about comparisons between objects, but considerably more difficult to estimate accurate distance or similarity scores. For example, humans are much better in comparing objects ("Movie A is more similar to movie B than to movie C") than in assigning scores ("The similarity between A and B is 0.9 and the similarity between A and C is 0.5"). There is a whole branch of machine learning that deals with data in ordinal (rather than cardinal) form (see, e.g., Rosales and Fung (2006), Shaw and Jebara (2009), McFee and Lanckriet (2009, 2011), Tamuz et al. (2011), Ailon (2012), von Luxburg and Alamgir (2013), Kleindessner and von Luxburg (2014), and references therein). Note that knowing the k nearest neighbors is one kind of such ordinal data though it contains different information than a collection of similarity triplets like in the motivating example with the movies.

In this paper, we provide two dimension estimators which just take information about the sets of k nearest neighbors as input. The first one is a straightforward estimator based on the doubling property of the Lebesgue measure. Even though we prove that it converges to the true dimension as the sample size increases, it turns out to severely underestimate the true dimension and needs a ridiculously high amount of sample points before it gives accurate results. Our second estimator is also based on geometric ideas, can also be proved to be statistically consistent, but is much more well-behaved in practice. Our experiments show that it can even compete with standard estima-

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

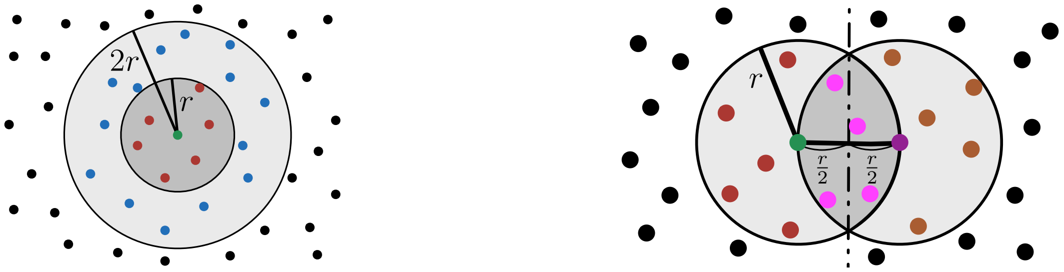


Figure 1: The idea behind our estimators E_{DP} (left) and E_{CAP} (right).

tors based on distance information.

2 OUR ESTIMATORS

Setup and problem statement. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a low-dimensional set, $\varphi : \mathcal{X} \to \mathcal{M} \subseteq \mathbb{R}^D$ a smooth embedding of $\mathcal{X}$ in a high-dimensional space and f a continuous probability density function supported on $\mathcal{X}$. Sample points $\{\tilde{x}_1, \ldots, \tilde{x}_n\} \subseteq \mathcal{X}$ are drawn from f and embedded into the observation space $\mathbb{R}^D$ via φ, possibly disturbed by noise $\eta_i \in \mathbb{R}^D$, resulting in the sample $\mathcal{D} = \{x_1, \ldots, x_n\}$ with $x_i = \varphi(\tilde{x}_i) + \eta_i$. Given some information about $\mathcal{D}$, our task is to infer d.

Note that although this problem is mathematically well-defined, it comes along with an inherent problem in practice: Data "looks" different at different scales, on the one hand due to the presence of noise, on the other hand due to the curvature of the manifold $\mathcal{M}$. For example, if our data lives in a small ε-tube around a one-dimensional sphere in $\mathbb{R}^2$, we will only be able to identify its one-dimensionality if we look at the data on a proper scale. If we "zoom in too closely", say we consider an ε-ball of the data, it appears to have dimension 2. If we "zoom out very far", the data will even look like a single point and thus may be considered as zero-dimensional.

While existing methods assume to know coordinates $(x_i^1, \ldots, x_i^D)$ or distance values $\|x_i - x_j\|_{\mathbb{R}^D}$, we assume to only know about memberships to the sets of k nearest neighbors. This information can be conveniently encoded by the directed, unweighted kNN-graph G built on $\mathcal{D}$. It has the vertex set $V = \{1, \ldots, n\}$ and a directed, unweighted edge from i to j (written as $i \to j$) if and only if x_j is among the k nearest sample points to x_i with respect to $\|\cdot\|_{\mathbb{R}^D}$. The parameter $k \ll n$ controls the scale as discussed in the previous paragraph: the larger k, the more we "zoom out".

In the following, we denote by $B_{\text{SP}}(i, r)$ the closed ball with center $i \in V$ and radius $r > 0$ in the graph G with respect to the (directed) shortest path distance d_{SP}, that is $B_{\text{SP}}(i, r) = \{j \in V : d_{\text{SP}}(i, j) \leq r\}$. A closed ball in $\mathbb{R}^d$ with center $x \in \mathbb{R}^d$ and radius $r > 0$ is denoted by $B(x, r)$. By λ_d we denote the d-dimensional Lebesgue measure and by $\eta_d = \lambda_d(B(0,1))$ the volume of the d-dimensional unit ball.

2.1 Estimator based on Doubling Property

Recall the doubling property of λ_d: for any $x \in \mathbb{R}^d$ and $r > 0$ we have $\lambda_d(B(x, 2r)) = 2^d \lambda_d(B(x, r))$. Consequently, we can determine the dimension d by

$$d = -\log_2(\lambda_d(B(x, r)) / \lambda_d(B(x, 2r))).$$

This is the property we are going to exploit in our first naive estimator. To carry it over to the finite sample setting, fix any sample point x_i (sufficiently far from the boundary of $\mathcal{M}$) and consider $B_{\text{SP}}(i, 1)$ and $B_{\text{SP}}(i, 2)$. If the sample size n is large enough and k is relatively small, points x_j with $j \in B_{\text{SP}}(i, 2)$ lie in such a small neighborhood of x_i on $\mathcal{M}$ that we can actually think of $\mathcal{M}$ as flat and identify it with $\mathbb{R}^d$ (here we do not take the noise into account). Then, as we will see in the proofs in Section 3, the balls $B_{\text{SP}}(i, 1)$ and $B_{\text{SP}}(i, 2)$ in G approximately correspond to balls $B(x_i, r)$ and $B(x_i, 2r)$ in $\mathbb{R}^d$ for some small radius r (see the left side of Figure 1 for an illustration: in green we see the point x_i, in red points x_j with $j \in B_{\text{SP}}(i, 1)$ and in blue points x_j with $j \in B_{\text{SP}}(i, 2) \setminus B_{\text{SP}}(i, 1)$). On these small balls we can consider the density f as roughly constant and obtain

$$L_{\text{DP}}(i) := \frac{|B_{\text{SP}}(i, 1)|}{|B_{\text{SP}}(i, 2)|} \approx \frac{n \ f(x_i) \ \lambda_d(B(x_i, r))}{n \ f(x_i) \ \lambda_d(B(x_i, 2r))} = \frac{1}{2^d}.$$

Note that $|B_{\text{SP}}(i, 1)|$ always equals $k + 1$.

Hence, an estimate of d is given by $-\log_2 L_{\text{DP}}(i)$. However, in order to obtain a more robust estimator we average over $L_{\text{DP}}(i)$ for various vertices $i \in A \subseteq V$. With $L_{\text{DP}}(A) := \frac{1}{|A|} \sum_{i \in A} L_{\text{DP}}(i)$ this leads to our first dimension estimator

$$E_{\text{DP}}(A) := -\log_2 L_{\text{DP}}(A).$$

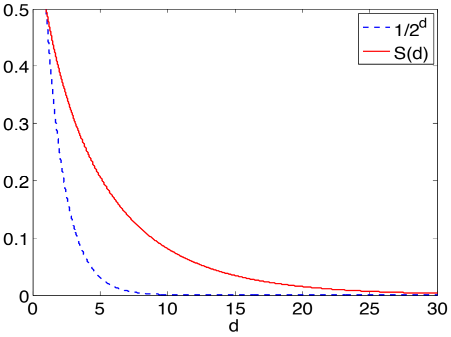


Figure 2: The functions $T(d) = 1/2^d$ and S. The latter is more well-behaved in terms of inversion.

2.2 Estimator based on Spherical Caps

Our second estimator relies on a different geometric idea. Fix $x, y \in \mathbb{R}^d$ with $\|x - y\|_{\mathbb{R}^d} = r$ and consider the set $B(x,r) \cap B(y,r)$. This set is the union of two congruent and disjoint spherical caps with height $r/2$ of a ball with radius r. An illustration is shown in the right figure of Figure 1 (dark grey area). According to Li (2011), the volume of such a cap is given by

$$\frac{1}{2}\eta_d r^d I_{\frac{3}{4}}\left(\frac{d+1}{2}, \frac{1}{2}\right),$$

where $I_x(a,b)$ is the regularized incomplete beta function. Consequently,

$$\frac{\lambda_d(B(x,r) \cap B(y,r))}{\lambda_d(B(x,r))} = I_{\frac{3}{4}}\left(\frac{d+1}{2}, \frac{1}{2}\right) =: S(d), \quad (1)$$

a quantity injectively depending on $d > 0$. A plot of the function S can be seen in Figure 2. Hence, the dimension d can be retrieved by inverting S.

Our goal is now to follow this idea in the finite sample setting. As in the previous section, we fix a sample point x_i and replace $B(x_i, r)$ by $B_{\mathrm{SP}}(i,1)$. We need to find a vertex j_0 such that x_{j_0} sits on the boundary of $B(x_i, r)$ and then consider $|B_{\mathrm{SP}}(i,1) \cap B_{\mathrm{SP}}(j_0,1)|$. Because $|B_{\mathrm{SP}}(i,1) \cap B_{\mathrm{SP}}(j,1)|$ tends to decrease with increasing distance between x_i and x_j, we can find such a vertex j_0 as the minimizer of the term $|B_{\mathrm{SP}}(i,1) \cap B_{\mathrm{SP}}(j,1)|$ over vertices j connected to i. This leads to

$$L_{\mathrm{CAP}}(i) := \frac{\min_{j \in V: i \to j} |B_{\mathrm{SP}}(i,1) \cap B_{\mathrm{SP}}(j,1)|}{k+1} \approx S(d).$$

An estimate for d is then given by $S^{-1}(L_{\mathrm{CAP}}(i))$. As in the previous section, we make the estimator more robust by averaging over $L_{\mathrm{CAP}}(i)$ for various vertices $i \in A \subseteq V$. With $L_{\mathrm{CAP}}(A) := \frac{1}{|A|}\sum_{i \in A} L_{\mathrm{CAP}}(i)$, our second dimension estimator is given by

$$E_{\mathrm{CAP}}(A) := S^{-1}(L_{\mathrm{CAP}}(A)).$$

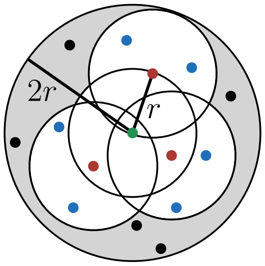


Figure 3: Explanation for the bias of E_{DP}: the union of the small balls approximates the large ball, but ignores a substantial part close to its boundary (shaded area).

2.3 First Comparison of E_{DP} and E_{CAP}

A closer look at the construction of our two estimators reveals two reasons why E_{CAP} might perform better than E_{DP}. This theoretical finding will later be confirmed in the experiments of Section 4.

The rationale of E_{DP} is to find an expression L_{DP} that approximates $T(d) := 1/2^d$, whereas in E_{CAP} we find an expression L_{CAP} approximating $S(d)$ as given in (1). In both cases, the final estimate is obtained by inverting T or S, respectively, to retrieve d. Inverting a function f is easy and robust in areas where the function is reasonably steep, but is difficult in areas where it is flat. In flat areas of f, small deviations in $f(x)$ lead to large deviations in $x = f^{-1}(f(x))$. Now consider the plot of the functions T and S in Figure 2. It is plain to see that S has a much larger range where it is well-behaved (say, from $d = 1$ to 20) than T (say, from $d = 1$ to 8). Consequently, in the range of $d = 9$ to $d = 20$ the estimator E_{CAP} is still rather robust against deviations of $L_{\mathrm{CAP}}(A)$ from $S(d)$, while small deviations of $L_{\mathrm{DP}}(A)$ from $1/2^d$ lead to large deviations of $E_{\mathrm{DP}}(A) = -\log_2(L_{\mathrm{DP}}(A))$ from d.

Our second insight is that E_{DP} systematically underestimates the true dimension, in particular if the latter is high. The estimator E_{DP} is based on approximating $B(x_i, 2r)$ by $B_{\mathrm{SP}}(i,2)$. However, as Figure 3 shows, there is a bias in this approximation: $B_{\mathrm{SP}}(i,2)$ is the union of $B_{\mathrm{SP}}(i,1)$ and balls $B_{\mathrm{SP}}(j,1)$ for vertices j with $i \to j$. Hence, $B_{\mathrm{SP}}(i,2)$ actually corresponds to points in the union of $B(x_i, r)$ and balls $B(x_j, r)$. In the limit, as n and k go to infinity, this union approximates $B(x_i, 2r)$ up to arbitrary precision, but for finite values of n and k this union is only a poor approximation of $B(x_i, 2r)$, just filling it partially and ignoring a substantial part close to the boundary of $B(x_i, 2r)$. As a consequence, we systematically underestimate $\lambda_d(B(x_i, 2r))$ and thus underestimate d. This effect is increased if d is high by the fact that in high-dimensional spaces almost all of the volume of a ball is concentrated in a thin shell close to the ball's boundary.

3 CONSISTENCY

In this section we prove that both our estimators $E_{\mathrm{DP}}(\{i\})$ and $E_{\mathrm{CAP}}(\{i\})$ converge to the true dimension d in probability if $n \to \infty$, $k = k(n)$ is chosen reasonably, and $i \in \{1, \ldots, n\}$ is chosen uniformly at random. For simplicity, we focus on the case when the manifold $\mathcal{M}$ is flat (that is φ is a global isometry) and there is no noise. The consistency of E_{DP} and E_{CAP} will hold in the general case as well, but the proofs will require a technical overhead that we want to avoid.

We make the following **regularity assumptions**: The domain $\mathcal{X} \subseteq \mathbb{R}^d$ is compact and has boundary of measure 0, i.e. $\lambda_d(\partial\mathcal{X}) = 0$. Furthermore, the boundary is nice in the sense that there exist constants $\alpha, \varepsilon_0 > 0$ such that

$$\lambda_d(B(x,\varepsilon) \cap \mathcal{X}) \geq \alpha \cdot \lambda_d(B(x,\varepsilon)), \quad x \in \mathcal{X}, \varepsilon < \varepsilon_0.$$

The density $f : \mathcal{X} \to \mathbb{R}$ is lower and upper bounded by $0 < f_{min} \leq f(x) \leq f_{max} < \infty$ for all $x \in \mathcal{X}$ and is Lipschitz continuous with constant L.

Main Theorem (Consistency of $E_{\mathbf{DP}}$ and $E_{\mathbf{CAP}}$) *Let the regularity assumptions hold. Let $\mathcal{D} = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ be an i.i.d. sample from f and G be the directed, unweighted kNN-graph on $\mathcal{D}$. Given G as input and a vertex $i \in \{1, \ldots, n\}$ chosen uniformly at random, both $E_{DP}(\{i\})$ and $E_{CAP}(\{i\})$ converge to the true dimension d in probability as $n \to \infty$ if $k = k(n)$ satisfies $k \in o(n)$, $\log n \in o(k)$, and there exists $k' = k'(n)$ with $k' \in o(k)$ and $\log n \in o(k')$.*

The growth conditions on k are the ones to be expected for random kNN-graphs. There are several ways of choosing k and k' in order to satisfy them. For example, we could choose $k = (\log n)^{1+\tau}$ and $k' = (\log n)^{1+\tau/2}$ for some $\tau > 0$.

Note that for proving the theorem it is sufficient to prove convergence of $L_{\mathrm{DP}}(\{i\})$ and $L_{\mathrm{CAP}}(\{i\})$ to $1/2^d$ and $S(d)$, respectively, since both $\log_2$ and S^{-1} are continuous functions. Furthermore, note that for showing convergence in probability of random variables U_n to a constant U it is sufficient to prove

$$f(\delta,n)U - e(\delta,n) \leq U_n \leq F(\delta,n)U + E(\delta,n) \tag{2}$$

with probability at least $P(\delta,n)$ for all $0 < \delta < \delta_0$ and $n \geq N(\delta)$, such that $f(\delta,n), F(\delta,n) \to 1$, $e(\delta,n), E(\delta,n) \to 0$ as $n \to \infty$, $\delta \to 0$ and $P(\delta,n) \to 1$ as $n \to \infty$ (for fixed δ).

Showing an inequality of the type of (2) for $L_{\mathrm{DP}}(\{i\})$ and $L_{\mathrm{CAP}}(\{i\})$ consists of a number of steps, which we formulate as separate propositions and lemmas. Due to space constraints we can only provide compact versions. A central role will be played by the kNN-radius

$$r_{k,n}(x_i) = \max\{\|x_i - x_j\| : i \to j \text{ in } G\}.$$

Furthermore, we will repeatedly encounter a quantity $u_{k,n}$ given by $u_{k,n} = (k/(n\eta_d))^{1/d}$. Note that the conditions on k imply that $u_{k,n} \to 0$.

All following statements hold for n sufficiently large and δ sufficiently small. The constants c_i in Propositions 1 and 3 depend on $d, \alpha, f_{min}, f_{max}$ and L.

Proposition 1 (kNN-radius is concentrated) *There exist $c_1, c_2 > 0$ such that with probability at least $1 - n\exp(-c_1\delta^2 k)$ we have*

$$\underline{R}_{k,n}(x_i,\delta) \leq r_{k,n}(x_i) \leq \overline{R}_{k,n}(x_i,\delta) \tag{3}$$

for all $x_i \in \mathcal{D}$ sufficiently distant to $\partial\mathcal{X}$, where

$$\underline{R}_{k,n}(x_i,\delta) = \frac{1}{(1+\delta)(1+c_2(k/n)^{1/d})} \cdot \frac{u_{k,n}}{f(x_i)^{1/d}},$$
$$\overline{R}_{k,n}(x_i,\delta) = \frac{1}{(1-\delta)(1-c_2(k/n)^{1/d})} \cdot \frac{u_{k,n}}{f(x_i)^{1/d}}.$$

This can be shown by standard concentration arguments. For example, a proof for a closely related statement can be found in von Luxburg et al. (2014). Note that $\underline{R}_{k,n}(x_i,\delta), \overline{R}_{k,n}(x_i,\delta) \to 0$ under the conditions on k.

Lemma 2 (Locally $r_{k,n}$ varies only slightly) *Assume the event considered in Proposition 1 holds. Then we have for a sample point $x_i \in \mathcal{D}$ sufficiently distant to $\partial\mathcal{X}$ and all $y \in \mathcal{D} \cap B(x_i, \overline{R}_{k,n}(x_i,\delta))$*

$$r_{k,n}(y) \geq \underline{R}_{k,n}(x_i,\delta) - a_{k,n}(\delta)u_{k,n}^{1+1/d}, \tag{4}$$
$$r_{k,n}(y) \leq \overline{R}_{k,n}(x_i,\delta) + a_{k,n}(\delta)u_{k,n}^{1+1/d}, \tag{5}$$

where $a_{k,n}(\delta) > 0$ converges to a positive constant as $n \to \infty$, $\delta \to 0$, assuming the conditions on k hold.

Lemma 2 immediately follows from the Lipschitz continuity of f. Due to the growth conditions on k we have $a_{k,n}(\delta)u_{k,n}^{1+1/d} \in o(\underline{R}_{k,n}(x_i,\delta))$ and $a_{k,n}(\delta)u_{k,n}^{1+1/d} \in o(\overline{R}_{k,n}(x_i,\delta))$.

Proposition 3 (Dense sampling lemma) *There exist $c_3, c_4 > 0$ such that for all $\gamma \leq \varepsilon_0$ we have*

$$\forall x \in \mathcal{X} \;\; \exists x_i \in \mathcal{D} : \; \|x_i - x\| \leq \gamma \tag{6}$$

with probability at least $1 - c_3\gamma^{-d}\exp(-c_4\gamma^d n)$.

The proof of Proposition 3 uses a simple covering argument and can be found in the supplementary material of Tenenbaum et al. (2000).

The next two lemmas are specific to $L_{\mathrm{DP}}(\{i\})$ and $L_{\mathrm{CAP}}(\{i\})$, respectively. Recall that $i \to j$ in G if and only if $\|x_i - x_j\| \leq r_{k,n}(x_i)$.

Lemma 4 (Geometric argument for $L_{\mathbf{DP}}(\{i\})$: $B_{\mathbf{SP}}(i,2)$ approximates $B(x_i, 2r_{k,n}(x_i))$) *Assume the events considered in Proposition 1 and Proposition 3 (with γ replaced by $\varepsilon_{k,n}$) hold. Then we have for a sample point $x_i \in \mathcal{D}$ sufficiently distant to $\partial\mathcal{X}$ and all $x_j \in \mathcal{D}$ the following implications:*

$$\begin{aligned} \|x_i - x_j\| \leq 2\underline{R}_{k,n}(x_i,\delta) - a_{k,n}(\delta)u_{k,n}^{1+1/d} - 2\varepsilon_{k,n} \\ \Longrightarrow d_{SP}(i,j) \leq 2, \\ \|x_i - x_j\| > 2\overline{R}_{k,n}(x_i,\delta) + a_{k,n}(\delta)u_{k,n}^{1+1/d} \\ \Longrightarrow d_{SP}(i,j) > 2. \end{aligned}$$

Proof: If $\|x_i - x_j\| \leq 2\underline{R}_{k,n}(x_i,\delta) - a_{k,n}(\delta)u_{k,n}^{1+1/d} - 2\varepsilon_{k,n}$, consider the point z on the line segment from x_i to x_j such that $\|x_i - z\| = \underline{R}_{k,n}(x_i,\delta) - \varepsilon_{k,n}$. Due to the assumption that (6) holds (with γ replaced by $\varepsilon_{k,n}$) there exists a sample point $x_l \in \mathcal{D}$ with $\|z - x_l\| \leq \varepsilon_{k,n}$. Then $\|x_i - x_l\| \leq \|x_i - z\| + \|z - x_l\| \leq \underline{R}_{k,n}(x_i,\delta)$ and hence $i \to l$ according to (3). Similarly, $\|x_j - x_l\| \leq \underline{R}_{k,n}(x_i,\delta) - a_{k,n}(\delta)u_{k,n}^{1+1/d}$ such that $l \to j$ because of $x_l \in B(x_i, \overline{R}_{k,n}(x_i,\delta))$ and (4). If $\|x_i - x_j\| > 2\overline{R}_{k,n}(x_i,\delta) + a_{k,n}(\delta)u_{k,n}^{1+1/d}$, then (3) and (5) immediately imply $d_{\mathrm{SP}}(i,j) > 2$. □

Lemma 5 (Geometric argument for $L_{\mathbf{CAP}}(\{i\})$: $B_{\mathbf{SP}}(i,1) \cap B_{\mathbf{SP}}(j_0,1)$ approximates union of spherical caps) *Assume the events considered in Proposition 1 and Proposition 3 (with γ replaced by $\varepsilon_{k,n}$) hold. Then we have for a sample point $x_i \in \mathcal{D}$ sufficiently distant to $\partial\mathcal{X}$ and all $j \in \{1,\dots,n\}$ with $i \to j$*

$$\begin{aligned} \|x_i - x_j\| \leq a_{k,n}(\delta)u_{k,n}^{1+1/d} \Longrightarrow \\ B(x_i,\underline{R}) \cap B(x_j,\underline{T}) = B(x_j,\underline{T}), \\ \|x_i - x_j\| > a_{k,n}(\delta)u_{k,n}^{1+1/d} \Longrightarrow \\ B(x_i,\underline{R}) \cap B(x_j,\underline{T}) \supseteq \\ C(x_i,\underline{R},h_1,x_j) \cup C(x_j,\underline{T},h_2,x_i). \end{aligned}$$

Here we abbreviate

$$\begin{aligned} \underline{R} &= \underline{R}_{k,n}(x_i,\delta), \qquad \underline{T} = \underline{R} - a_{k,n}(\delta)u_{k,n}^{1+1/d}, \\ \overline{R} &= \overline{R}_{k,n}(x_i,\delta), \qquad \overline{T} = \overline{R} + a_{k,n}(\delta)u_{k,n}^{1+1/d}, \end{aligned}$$

$$h_1 = \underline{R} - \frac{1}{2}\overline{R} + \frac{\underline{T}^2 - \underline{R}^2}{2\overline{R}}, \quad h_2 = \underline{T} - \frac{1}{2}\overline{R} + \frac{\underline{R}^2 - \underline{T}^2}{2\overline{R}}$$

and $C(z,r,h,w)$ denotes a spherical cap of a ball $B(z,r)$ with height h ($0 \leq h \leq 2r$) and apex on the half-line from z to w. Furthermore, there exists a sample point x_l with $i \to l$ such that

$$\begin{aligned} B(x_i,\overline{R}) \cap B(x_l,\overline{T}) \subseteq \\ C(x_i,\overline{R},h_1',x_l) \cup C(x_l,\overline{T},h_2',x_i) \end{aligned}$$

with

$$\begin{aligned} h_1' &= \overline{R} - \frac{1}{2}(\underline{R} - 2\varepsilon_{k,n}) + \frac{\overline{T}^2 - \overline{R}^2}{2(\underline{R} - 2\varepsilon_{k,n})}, \\ h_2' &= \overline{T} - \frac{1}{2}(\underline{R} - 2\varepsilon_{k,n}) + \frac{\overline{R}^2 - \overline{T}^2}{2(\underline{R} - 2\varepsilon_{k,n})}. \end{aligned}$$

The proof of Lemma 5 mainly consists of determining the heights of the two spherical caps which arise when one intersects two balls. Once one has a closed formula for these heights (depending on the radii of the balls and the distance of their centers) the proof is straightforward and as simple as the one of Lemma 4.

Proof of the Main Theorem (sketch): We choose $\varepsilon_{k,n} = (k'/n)^{1/d}$ for some $k' = k'(n)$ satisfying $k' \in o(k)$ and $\log n \in o(k')$. Then both the events of Propositions 1 and 3 (with γ replaced by $\varepsilon_{k,n}$) hold jointly with probability at least $1 - n\exp(-c_1\delta^2 k) - c_3(n/k')\exp(-c_4 k')$. This probability tends to 1. Furthermore, due to the assumption that $\lambda_d(\partial\mathcal{X}) = 0$, the probability of choosing a vertex i such that x_i is too close to $\partial\mathcal{X}$ tends to 0 (note that in the statements above "sufficiently distant" is defined in terms of $r_{k,n}(x_i)$, which uniformly tends to 0). So with high probability the implications in Lemmas 4 and 5 hold for x_i with the chosen i. By standard concentration arguments we can lower and upper bound $|B_{\mathrm{SP}}(i,2)|$ and $\min_{j\in V: i\to j} |B_{\mathrm{SP}}(i,1) \cap B_{\mathrm{SP}}(j,1)|$, which yields estimates of the form (2) for $L_{\mathrm{DP}}(\{i\})$ and $L_{\mathrm{CAP}}(\{i\})$, respectively. □

4 EXPERIMENTS

4.1 Implementation of our Estimators

There is no closed form for the inverse of the function S as given in (1). If one is merely interested in an integer estimate, the simplest procedure is to set $E_{\mathrm{CAP}}(A)$ to $d^* = \arg\min_{d\in\mathbb{N}} |S(d) - L_{\mathrm{CAP}}(A)|$. In case one would rather like a real-valued estimate, the simplest way is to create a fine-meshed lookup table.

Assuming that G is given by its unsorted adjacency lists, it is easy to see that both $L_{\mathrm{DP}}(i)$ and $L_{\mathrm{CAP}}(i)$ can be computed with worst case running time $\mathcal{O}(k^2 \log k)$. Hence, the computation of both $E_{\mathrm{DP}}(A)$ and $E_{\mathrm{CAP}}(A)$ can be performed in $\mathcal{O}(|A|\, k^2 \log k)$, assuming that the

Table 1: Estimated dimensions for several datasets. n denotes the size of the dataset and d its true dimension.

				Our estimators	**(kNN-graph)**	**Standard estimators**	**(distance information)**	
	n	**Distribution / Dataset**	d	$\boldsymbol{E_{\mathrm{CAP}}(V)}$	$\boldsymbol{E_{\mathrm{DP}}(V)}$	MLE	***CorrDim***	***RegDim***
	Artificial datasets (results averaged over 100 runs, $\pm STD$)							
1	1000	uniform on a helix in $\mathbb{R}^3$	1	1.00 ($\pm$0.05)	0.88 ($\pm$0.01)	1.00 ($\pm$0.01)	1.00 ($\pm$0.11)	0.99 ($\pm$0.01)
2	1000	Swiss roll in $\mathbb{R}^3$	2	2.14 ($\pm$0.05)	1.44 ($\pm$0.01)	1.94 ($\pm$0.02)	1.99 ($\pm$0.23)	1.87 ($\pm$0.04)
3	1000	$N_5(0, I)$	5	5.33 ($\pm$0.07)	2.47 ($\pm$0.01)	5.00 ($\pm$0.04)	4.91 ($\pm$0.56)	4.86 ($\pm$0.05)
4	1000	uniform on sphere $\mathbb{S}^7 \subseteq \mathbb{R}^8$	7	5.88 ($\pm$0.06)	2.82 ($\pm$0.01)	6.53 ($\pm$0.07)	6.85 ($\pm$0.66)	6.23 ($\pm$0.09)
5	5000	uniform on sphere $\mathbb{S}^7 \subseteq \mathbb{R}^8$	7	6.85 ($\pm$0.03)	3.21 ($\pm$0.00)	6.72 ($\pm$0.03)	6.95 ($\pm$0.98)	6.46 ($\pm$0.04)
6	1000	uniform on $[0,1]^{12}$	12	7.74 ($\pm$0.08)	3.04 ($\pm$0.01)	9.32 ($\pm$0.10)	10.66 ($\pm$1.18)	8.78 ($\pm$0.10)
7	5000	uniform on $[0,1]^{12}$	12	9.24 ($\pm$0.04)	3.50 ($\pm$0.00)	9.76 ($\pm$0.05)	10.83 ($\pm$1.49)	9.26 ($\pm$0.05)
	Real datasets (D = dimension of observation space)							
8	698	Isomap faces, $D = 4096 = 64^2$	?	3.04	1.73	3.99	3.53	4.22
9	481	Hands, $D = 245760$	?	1.27	0.95	2.88	3.92	2.56
10	7141	MNIST digit 3, $D = 784 = 28^2$	?	8.92	3.21	15.95	14.17	14.75
11	6824	MNIST digit 4, $D = 784 = 28^2$	?	8.13	3.07	14.44	9.54	13.16
12	6313	MNIST digit 5, $D = 784 = 28^2$	?	8.40	3.12	15.55	18	14.28

inversion of S as addressed above can be done in constant time. When G is given by its adjacency matrix J (that is $J_{ij} = 1$ if $i \to j$ and 0 otherwise), it is usually faster to exploit the following observations, in particular if $|A|$ is large (e.g., $A = V$):

$$\left(\tilde{J} \cdot \tilde{J}\right)_{ij} > 0 \Leftrightarrow j \in B_{\mathrm{SP}}(i, 2),$$
$$\left(\tilde{J} \cdot \tilde{J}^T\right)_{ij} = \left|B_{\mathrm{SP}}(i, 1) \cap B_{\mathrm{SP}}(j, 1)\right|.$$

Here, $\tilde{J}$ is the matrix J with the diagonal entries set to 1. Note that J and $\tilde{J}$ are sparse.

Both our estimators are statistically consistent for any random choice of A. The variance of $E_{\mathrm{DP}}(A)$ and $E_{\mathrm{CAP}}(A)$ decreases approximately like $1/|A|$ if A is chosen uniformly at random without replacement (which is hard to prove theoretically due to dependency issues, but can be easily verified in simulations), so we suggest to choose $|A|$ as large as one can afford due to computational reasons.

4.2 First Comparison with Estimators from the Literature

To get a first feeling, we compare our estimators E_{DP} and E_{CAP} to three standard estimators from the literature, all of them relying on distance information: The recent estimator MLE of Levina and Bickel (2005), which seems to be state of the art, and two widely used classical estimators: the correlation dimension-estimator *CorrDim* (Grassberger and Procaccia, 1983) and the estimator *RegDim*. MLE is the average of estimators $\hat{m}_k$ for several values of k where $\hat{m}_k$ in turn is the average of local maximum likelihood estimators $\hat{m}_k(x_i)$, which estimate the dimension of the data around a sample point x_i based on the distances between x_i and its k nearest neighbors. *CorrDim* estimates the dimension by regressing $\log C_r$ on $\log r$ over a suitable range of r, where

$$C_r = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{1}\left(\|x_i - x_j\| \leq r\right)$$

is the normalized number of pairs of sample points with distance not more than r from each other. Similarly, *RegDim* works by regressing $\log R_k$ on $\log k$, where $R_k = (1/n) \sum_{i=1}^{n} r_{k,n}(x_i)$ and $r_{k,n}(x_i)$ is the kNN-radius of sample point x_i. This is a slightly simplified version of the algorithm suggested by Pettis et al. (1979).

For several artificial and real datasets, Table 1 shows the estimated dimensions for the various estimators. All estimators require to set some parameters: a single parameter k for E_{DP} and E_{CAP} and two parameters k_1, k_2 for MLE, *CorrDim* and *RegDim*. For MLE these parameters determine the range for averaging over $\hat{m}_k$ and for *CorrDim* and *RegDim* the range for regressing (for *CorrDim* the range is given as $[r_{k_1}, \ldots, r_{k_2}]$ where r_i denotes the i-th smallest entry in the distance matrix of the data sample). For all experiments except (9) we set the parameters for MLE and *RegDim* as $k_1 = 10$, $k_2 = 20$ and for *CorrDim* as $k_1 = 10$, $k_2 = 100$ like Levina and Bickel, who also performed the experiments (2), (8) and (9). In (9), like Levina and Bickel, we changed the parameters for *CorrDim* to $k_1 = 500$, $k_2 = 1000$ since the original choice leads to the obviously wrong result of an estimated dimension of 19.7. For E_{DP} and E_{CAP} we simply set $k = 15$ if the size of the dataset is less than or equal to 1000 (for the real datasets this is (8) and (9)) and $k = 20$ in all other experiments (which deal with datasets of size 5000 or slightly greater).

For the artificial datasets the interpretation of the results is straightforward. The naive estimator $E_{\mathrm{DP}}(V)$ only gives reasonable results for the experiments (1) and (2), where the dimension is small. It is highly biased in the higher-dimensional cases. This confirms our arguments of Section 2.3. Our estimator $E_{\mathrm{CAP}}(V)$ performs comparably to the three distance-based estimators MLE, *CorrDim* and *RegDim*. This is quite surprising, given that it gets considerably less information about the data (no distance information, just nearest neighbor memberships).

For the real datasets the interpretation is not so obvious since the true intrinsic dimensions are unknown. Although the Isomap faces dataset, consisting of images of the face of a sculpture observed under different pose and lighting conditions, is usually considered to be three-dimensional, Levina and Bickel argue that its dimension should be higher because of the fact that we only deal with 2D-projections of the face. Similarly, according to them the intrinsic dimension of the Hands dataset, which is a sequence of snapshots of a hand moving along a one-dimensional curve, should be higher than one. In any case, the results of $E_{\mathrm{CAP}}(V)$ do not seem to be unreasonable, in particular if one additionally compares them to the results obtained by Hein and Audibert (2005). In their experiments the authors provide a dimension estimate of 3 for the Isomap faces dataset and estimates of 14, 13 and 12 for the sets of MNIST digits 3, 4 and 5, respectively.

4.3 Our Estimators in Detail

In the following we investigate in artificial datasets how our estimators behave with respect to choice of the parameter k, sample size n, intrinsic and ambient dimension and the presence of noise. As competitor we choose the state-of-the-art estimator $\hat{m}_k$: it is also based on the k nearest neighbors of sample points, but explicitly uses distances. Because our estimators do not get any distance information, we cannot expect E_{DP} and E_{CAP} to perform as well as $\hat{m}_k$, but consider the latter as a benchmark. In their paper Levina and Bickel suggest to average over $\hat{m}_k$ for a range of k (yielding the estimator MLE) in order to reduce the risk of choosing a bad value for it. In principle, this could also be done with our estimators, but in our setting this would require additional input information and so we do not want to pursue this idea any further.

Table 2: Estimated dimensions for data from $N_7(0, I)$ (averaged over 10 datasets, $\pm STD$). R denotes a random choice (without replacement) of 10 vertices.

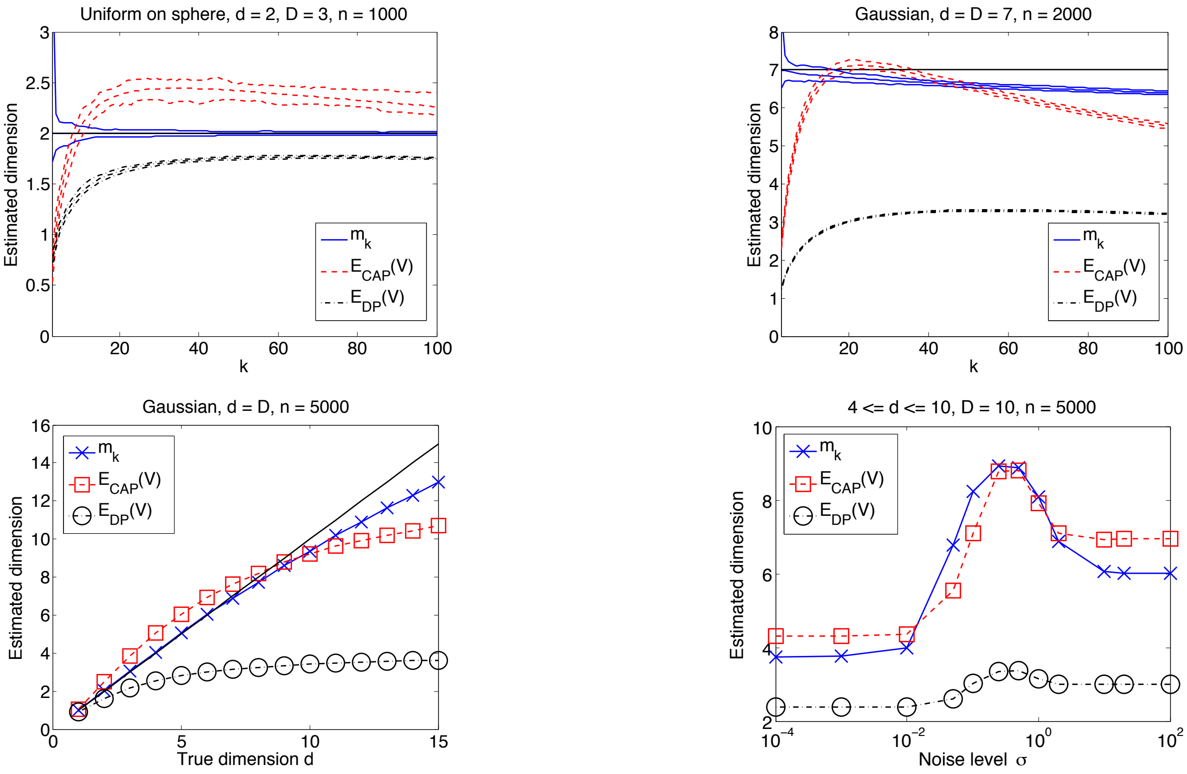
	$\boldsymbol{E}_{\mathrm{CAP}}(\boldsymbol{R})$	$\boldsymbol{E}_{\mathrm{DP}}(\boldsymbol{R})$
$n = 5 \cdot 10^4$, $k = 500$	6.77 (±0.19)	4.36 (±0.01)
$n = 5 \cdot 10^5$, $k = 1000$	7.58 (±0.12)	5.01 (±0.01)
$n = 5 \cdot 10^5$, $k = 2500$	6.99 (±0.13)	4.90 (±0.02)
$n = 5 \cdot 10^6$, $k = 3000$	7.77 (±0.11)	5.48 (±0.01)
$n = 5 \cdot 10^6$, $k = 8000$	7.44 (±0.14)	5.41 (±0.01)
$n = 5 \cdot 10^7$, $k = 5000$	7.95 (±0.20)	5.84 (±0.02)

Dependence on k. The top row of Figure 4 shows the results of the estimators $\hat{m}_k$, $E_{\mathrm{DP}}(V)$ and $E_{\mathrm{CAP}}(V)$ as a function of k. In the left figure the data consists of 1000 points sampled from a uniform distribution on the hypersphere $\mathbb{S}^2 \subseteq \mathbb{R}^3$. We can see that $\hat{m}_k$ performs best and yields a perfect estimate for all values of k in the range of consideration. However, this estimator explicitly uses distance values. Although using much less information, our estimators perform well and yield a correct result after rounding for a broad range of k too. The right figure deals with 2000 points from a 7-dim Gaussian $N_7(0, I)$. In this higher-dimensional case the situation is different: while $E_{\mathrm{CAP}}(V)$ still performs reasonable and yields a correct result, at least for a not too small range of k, $E_{\mathrm{DP}}(V)$ constantly underestimates the dimension. This confirms our findings of Section 2.3.

Dependence on the sample size n. As we have proved in Section 3, both E_{CAP} and E_{DP} converge to the true dimension if $n \to \infty$ and k is chosen appropriately. In Table 2 we show the results for increasing sample size n in the case of a 7-dim Gaussian $N_7(0, I)$. We can see that for E_{DP} the convergence is painfully slow and even with $n = 5 \cdot 10^7$ points it still underestimates the dimension. E_{CAP} needs a lot fewer sample points in order to give a valuable result — compare with the previous paragraph. However, here it has a tendency to slightly "overshoot" (which is a consequence of a suboptimal choice of k).

Bias with respect to the true dimension. The left figure in the bottom row of Figure 4 shows the behavior of the estimators with respect to the true dimension d. We can see that as the true dimension d increases, the property of underestimating the dimension of E_{DP} is shared by E_{CAP} and even by $\hat{m}_k$ (although to a much slighter extent).

Noisy data. In the right figure in the bottom row of Figure 4 we plot the estimated dimensions as a function of the noise level σ for 5000 points drawn from $U(0, 1)^4 \times N_6(0, \sigma I)$ and k set to 20. Here $U(0, 1)$



Uniform on sphere, d = 2, D = 3, n = 1000

Gaussian, d = D = 7, n = 2000

Gaussian, d = D, n = 5000

4 <= d <= 10, D = 10, n = 5000

Figure 4: Top row: The estimators $\hat{m}_k$, $E_{\text{DP}}(V)$ and $E_{\text{CAP}}(V)$ as a function of k (average over 100 datasets together with min/max over the 100 datasets) for 1000 points from a uniform distribution on the hypersphere $\mathbb{S}^2 \subseteq \mathbb{R}^3$ (left) and 2000 points from a 7-dim Gaussian $N_7(0, I)$ (right). The solid black lines show the true dimension. Bottom row: Left: Estimated dimensions as a function of the true dimension d (solid black line). 5000 points from a d-dim Gaussian $N_d(0, I)$, $k = 20$. Right: Estimates as a function of the noise level σ. 5000 points from $U(0, 1)^4 \times N_6(0, \sigma I)$, $k = 20$.

denotes a uniform distribution on the unit interval. When σ is small, the last six components of the data can be considered as noise and the dimension of the data should be four. As σ increases, the noise level gets so high that the data actually should be considered as 10-dimensional. Finally, the role of the "true" data and the noise gets inverted, the first four components are dominated by the six last ones and considered as noise, hence the dimension should be 6. This behavior is reflected by all three estimators under consideration. However, again the performance of $E_{\text{DP}}(V)$ is very poor, completely failing to correctly determine either of the various dimensions.

5 DISCUSSION

As opposed to all existing dimension estimators in the literature, we consider a setting where we only have access to very restricted information about the data. Instead of actual distance measurements all we get to know is *who* are the k nearest neighbors of each data point (but we do *not* know distances to the neighbors). In the light of the findings of Kleindessner and von Luxburg (2014) and Terada and von Luxburg (2014), who show in different scenarios that such ordinal information uniquely determines the geometry of the data, it is not too surprising that dimensionality estimation in this setting is possible. However, the question is how to do it successfully in practice.

The main message of our paper is twofold. First, we show that the "obvious" estimator E_{DP}, which is based on the doubling property of the Lebesgue measure, albeit being consistent in the large sample limit, performs only poorly in practice. Second, we provide an alternative, not so obvious estimator E_{CAP} which appears to be more well-behaved. Our experiments demonstrate that E_{CAP} achieves results that are comparable to those of distance-based estimators from the literature or are only slightly worse.

Acknowledgements

U.v.L. acknowledges funding of the German Research Foundation (grant LU1718/1-1 and Research Unit 1735).

References

N. Ailon. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *JMLR*, 13:137–164, 2012.

R. Bennett. The intrinsic dimensionality of signal collections. *IEEE Trans. Information Theory*, 15(5): 517–525, 1969.

F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(10):1404–1407, 2002.

J. Costa and A. Hero. Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets. In *European Signal Processing Conference (EUSIPCO)*, 2004.

J. Costa, A. Girotra, and A. Hero. Estimating local intrinsic dimension with k-nearest neighbor graphs. In *Statistical Signal Processing Workshop (SSP)*, 2005.

B. Eriksson and M. Crovella. Estimating intrinsic dimension via clustering. In *Statistical Signal Processing Workshop (SSP)*, 2012.

A. Farahmand, C. Szepesvári, and J.-Y. Audibert. Manifold-adaptive dimension estimation. In *International Conference on Machine Learning (ICML)*, 2007.

K. Fukunaga and D. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. Computers*, 20(2):176–183, 1971.

P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica*, 9:189–208, 1983.

M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in Rˆd. In *International Conference on Machine Learning (ICML)*, 2005.

B. Kégl. Intrinsic dimension estimation using packing numbers. In *Neural Information Processing Systems (NIPS)*, 2002.

M. Kleindessner and U. von Luxburg. Uniqueness of ordinal embedding. In *Conference on Learning Theory (COLT)*, 2014.

E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Neural Information Processing Systems (NIPS)*, 2005.

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian J. Math. Stat.*, 4(1): 66–70, 2011.

B. McFee and G. Lanckriet. Partial order embedding with multiple kernels. In *International Conference on Machine Learning (ICML)*, 2009.

B. McFee and G. Lanckriet. Learning multi-modal similarity. *JMLR*, 12:491–523, 2011.

K. Pettis, T. Bailey, A. Jain, and R. Dubes. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1(1):25–37, 1979.

R. Rosales and G. Fung. Learning sparse metrics via linear programming. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

B. Shaw and T. Jebara. Structure preserving embedding. In *International Conference on Machine Learning (ICML)*, 2009.

K. Sricharan, R. Raich, and A. Hero. Optimized intrinsic dimension estimator using nearest neighbor graphs. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. Kalai. Adaptively learning the crowd kernel. In *International Conference on Machine Learning (ICML)*, 2011.

J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

Y. Terada and U. von Luxburg. Local ordinal embedding. In *International Conference on Machine Learning (ICML)*, 2014.

G. Trunk. Statistical estimation of the intrinsic dimensionality of data collections. *Information and Control*, 12:508–525, 1968.

U. von Luxburg and M. Alamgir. Density estimation from unweighted k-nearest neighbor graphs: a roadmap. In *Neural Information Processing Systems (NIPS)*, 2013.

U. von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large random neighborhood graphs. *JMLR*, 15:1751–1798, 2014.