Published in Proc. CAIP'97, Springer LNCS 1296, pp. 9-16

An Algorithm for Intrinsic Dimensionality Estimation

J. Bruske, G. Sommer

Computer Science Institute
Christian-Albrechts University zu Kiel, Germany
email:jbr@informatik.uni-kiel.de

Abstract. In this paper a new method for analyzing the *intrinsic dimensionality* (ID) of low dimensional manifolds in high dimensional feature spaces is presented. The basic idea is to first extract a low-dimensional representation that captures the *intrinsic topological structure* of the input data and then to analyze this representation, i.e. to estimate the intrinsic dimensionality. Compared to previous approaches based on local PCA the method has a number of important advantages: First, it can be shown to have only *linear time complexity* w.r.t. the dimensionality of the input space (in contrast to the cubic complexity of the conventional approach) and hence becomes applicable even for very high dimensional input spaces. Second, it is *less sensitive to noise* than former approaches, and, finally, the extracted representation can be directly used for further data processing tasks including auto-association and classification.
The presented method for ID estimation is illustrated on a synthetic data set. It has also been successfully applied to ID estimation of full scale image sequences, see [BS97].

1 Introduction

Adopting the classification in [JD88], there are two primary approaches for estimating the intrinsic dimensionality[1]. The first one is the *global approach* in which the swarm of patterns is unfolded or flattened in the d-dimensional space. Benett's algorithm [Ben69] and its successors as well as variants of MDSCAL [Kru64] for intrinsic dimensionality estimation belong to this category. The second approach is a *local* one and tries to estimate the intrinsic dimensionality directly from information in the neighborhood of patterns without generating configurations of points or projecting the pattterns to a lower dimensional space. Pettis' [PBJD79], Fukunaga and Olsen's [FO71] as well as Trunk's [Tru76] and Verveer and Duin's method [VD95] belong to this category.

Our approach belongs to the second category as well and is based on optimally topology preserving maps (OTPMs) and local principal component analysis (PCA) using a number of evenly distributed pointers in the manifold. It is

[1] The intrinsic, or topological, dimensionality of N patterns in an n-dimensional space refers to the minimum number of "free" parameters needed to generate the patterns [JD88]. It essentially determines whether the n-dimensional patterns can be described adequately in a subspace (submanifold) of dimensionality $m < n$.

conceptually similar to that of Fukunaga and Olsen, [FO71], using local PCA as well, but by utilizing OTPMs can be shown to better scale with high dimensional input spaces (linear instead of cubic) and to be more robust against noise. The local subspaces as revealed by local our PCA can be directly used for data modeling, as e.g. in [KL94].

2 Foundations

In this section we want to make the reader familiar with the basic ingredients of our algorithm for ID estimation to be presented in the next section. We fist introduce $OTPM$s, the underlying representation, and then turn to efficient PCA for $m < n$ points, the underlying method used for analyzing the $OTPM$, and finally comment on the problem of estimating the ID by local PCA, the general approach of our algorithm.

2.1 Constructing Optimally Topollogy Preserving Maps

Optimally Topology Preserving Maps ($OTPM$s) are closely related to Martinetz' Perfectly Topology Preserving Maps (PTPMs) [MS94] and are constructed in just the same way. The only reason to introduce them separatly is that in order to form a PTFM the pointers must be "dense" in the manifold M. Without prior knowledge this assumption cannot be checked, and in practice it will rarely be valid. $OTPM$s emerge if just the construction method for PTFMs is applied without checking for the density condition. Only in favourable cases one will obtain a PTFM (probably without noticing). $OTPM$s are nevertheless optimal in the sense of the topographic function introduced by Villmann in [VDM94]: In order to measure the degree of topology preservation of a graph G with an associated set of pointers S, Villmann effectively constructs the $OTPM$ of S and compares G with the $OTPM$. By construction, the topographic function just indicates the highest (optimal) degree of topology preservation if G is an $OTPM$.

Definition1 OTPM. Let $p(x)$ be a probability distribution on the input space R^n, $M = \{x \in R^n | p(x) \neq 0\}$ a manifold of feature vectors, $T \subseteq M$ a training set of feature vectors and $S = \{c_i \in M | i = 1, \ldots, N\}$ a set of pointers in M.

We call the undirected graph $G = (V, E)$, $|V| = N$, an *optimally topology preserving map of S given the training set T*, $OTPM_T(S)$, if

$$(i,j) \in E \;\Leftrightarrow\; \exists x \in T\, \forall k \in V \backslash \{i,j\} : \max\{\| c_i - x \|, \| c_j - x \|\} \leq \| c_k - x \|$$

Corolary 1 *If $T = M$ and if S is dense in M then $OTPM_T(S)$ is a PTPM.*

Note that the definition of the $OTPM$ is constructive: For calculating the $OTPM_T(S)$ simply pick $x \in T$ according $p_T(x)$, calculate the best and second best matching pointers, c_{bmu} and c_{smu}, and connect bmu with smu. If repeated

infinitely often, G will converge to $OTPM_T(S)$ with probability one. This procedure is just the essence of Martinetz' Hebbian learning rule.

For use in intrinsic dimensionality estimation and elsewhere, $OTPM_T(S)$ has two important properties. First, it does indeed only depend on the intrinsic dimensionality of T, i.e. it is independent of the dimensionality of the input space. Embedding T into some higher dimensional space does not alter the graph. Second, it is invariant against scaling and rigid transformations (translations and rotations). Just by definition it is the representation that optimally reflects the intrinsic (topological) structure of the data.

2.2 Efficient PCA for fewer points than dimensions

With $S = \{c_i \in R^n | i = 1, \ldots, N\}$ and $A^T = [c_1 - \bar{c}, \ldots, c_N - \bar{c}]$ the basic trick from linear algebra for $N < n$ is to calculate the PCA of $\hat{\Sigma} = AA^T$ instead of a PCA of the original covariance matrix $\Sigma = A^T A$. The eigenvalues of Σ, $\mu_1, \ldots, \mu_N$, are then identical to the eigenvalues $\nu_1, \ldots, \nu_N$ of $\hat{\Sigma}$ and the eigenvectors of Σ, $u_1, \ldots, u_N$, can be calculated from the eigenvectors $v_1, \ldots, v_N$ of $\hat{\Sigma}$ by setting $u_i = A^T v_i$. This can be simply checked by

$$\hat{\Sigma} v_i = \nu_i v_i \Leftrightarrow AA^T v_i = \nu_i v_i \Leftrightarrow A^T AA^T v_i = \nu_i A^T v_i \Leftrightarrow \Sigma(A^T v_i) = \nu_i A^T v_i$$

Since the PCA of the $N \times N$ matrix $\hat{\Sigma}$ can be calculated in $O(N^3)$, [PTVF88], and $\hat{\Sigma} = AA^T$ clearly can be computed in time $O(N^2 n)$, it takes only time $O(N^2 n + N^3)$ instead of $O(n^3)$ to calculate the PCA of the covariance matrix of S.

2.3 On the problem of ID estimation with local PCA

We assume the data points $x \in T$ to be noisy samples of a vector valued function $f : R^r \to R^n$

$$x = f(k) + \eta \tag{1}$$

where $k = [k_1, \ldots, k_r]$ is an r-dimensional parameter vector and η denotes the noise. The function f can be imagined to describe an r-dimensional hypersurface S in n-dimensional space. The effect of noise is to render the surface not infinitely thin (see [VD95]). Within a small region a linear approximation of the data set (such as provided by the eigenvectors of local PCAs) is only valid if the largest variance in directions n_j perpendicular to S is much smaller than the smallest variance in directions s_i of S, i.e.

$$\frac{\min_i Var(s_i)}{\max_j Var(n_j)} \gg 1. \tag{2}$$

Here, $Var(s_i)$, the intra-surface variance, depends on the size of the local region and $Var(n_j)$ depends on the variance caused by the noise *and* the fact that S cannot be exactly represented as a linear surface. This leads to a basic dilemma for any ID estimation algorithm based on local PCA: If the region is

too large, $Var(n_j)$ might be high due to the non-linear nature of S. If, on the other hand, the region is too small, the noise is still there and will eventually dominate $Var(s_i)$. The solution is to search for the region size that gives the best compromise.

Closely related to the problem of noise is the problem of having available only a limited set of data. In order to make local PCA approaches work, the data set has to be large enough to represent the non-linearities and to allow for filtering out the noise.

3 Dimensionality Analysis with *OTPM*s

The basic procedure for intrinsic dimensionality analysis with $OTPM$s works as follows: To find a set S of N pointers which reflects the distribution of T we first employs a clustering algorithm for T whose output are N cluster centers. Then we calculate the graph G as the optimal topology preserving map of S given T. The final step is to perform for each node v_i a principal component analysis of the correlation matrix of the difference vectors $c_j - c_i$ of the pointers c_j associated with the nodes v_j adjacent to v_i in G. The result of this analysis, i.e. eigenvalues and vectors for each node, is the output of the procedure and subjected to further analysis. Provided the complexity of the clustering algorithm is independent of the intrinsic dimensionality d, the serial time complexity is $O(n + m(d, T, S)^3)$, where $m(d, T, S)$ is the maximum number of direct neighbors of a node in the $OTPM$ as depending on the intrinsic dimensionality, the training set T and the set of pointers S. Bounds on $m(d, T, S)$ or even a functional form are hard to derive, yet m stays constant for constant ID, is independent of the input dimension n, and experiments confirm that it is indeed small for small IDs.

In the rest of this section we will first comment on the use of clustering algorithms and then extend the procedure to derive our actual ID estimation method.

3.1 Clustering in TPCA

The reason for clustering the data prior to construction of the $OTPM$ and not just drawing N pointers randomly from T is twofold: First the distribution of the pointers should reflect the underlying distribution $p_T(x)$ as accurately as possible and second we would like to eliminate noise on the data. Any vector quantization algorithm which aims at minimizing the (normalized) quantization error

$$J = \frac{1}{n} \sum_{i=1}^{N} \int_{V_i} \| x - c_i \|^2 \, p(x) dx, \tag{3}$$

where V_i denotes the Voronoi cell of c_i, is a good choice since by minimizing the total variance it will preferably place the pointers within the manifold M and filter out orthogonal noise. This holds because as long as criterion (2) is fulfilled placing pointers within the surface and hence reducing the intra-surface

variance causes the largest decrease in J. From information theory it is known that it also produces a distribution of pointers which reflects the probability density, e.g. [Zad82].

3.2 The ID estimation procedure

Eventuallywe must decide how many dominant eigenvalues exist in each local region, i.e. what size an eigenvalue as obtained by each local PCA must exceed to indicate an associated intra-surface eigenvector. This amounts to determining a threshold. We adopted the Dα criterion from Fukunaga et. al. [FO71] which regards an eigenvalue μ_i as significant if

$$\frac{\mu_i}{\max_j \mu_j} > \alpha\%. \tag{4}$$

If no prior knowledge is available, different values of α have to be tested. Otherwise, knowledge of the largest noise component can be used to calculate α.

A second problem is that due to the noise/non-linearity dilemma mentioned in section 2.3 we do not know the optimal local region sizes in advance and, in particular, do not know the optimal number of pointers N. Monitoring the development of the local eigenvalues for a growing number of pointers ($N = 1, \ldots$) and searching for characteristic transitions is the most natural way to proceed. In this case, one does not want to cluster all the $N+1$ pointers from scratch but rather would like to incrementally build on the existing N clusters, i.e. just add one new cluster and modify the existing ones if necessary. Using the LBG vector quantization algorithm, [LBG80], we start with $N = 1$ and add a new pointer by first searching the cluster with highest intra cluster variance. In this cluster we then search for the training sample x with the highest quantization error, add a new pointer at x, take this configuration of $N+1$ pointers as the new starting configuration for the LBG algorithm and run *tpca* for the $N+1th$ round. This procedure of first searching for the worst quantized cluster helps to alleviate problems with outliers which could lead to multiple insertions at the same point if only the worst quantized example was considered.

Finally, if we have reason to believe that the data set has constant intrinsic dimensionality (i.e. has been generated by one function and not by a mixture of functions) our estimate of the intrinsic dimensionality will be the average of all local ID estimates together with its standard deviation. The ID estimate and its standard deviation is then plotted versus the number of pointers N, with different plots resulting from different choices of α. In the next section we will demonstrate that these plots actually do give very fine and characteristic hints on the ID of the data set. Our estimation procedure is interactive because the user has to choose a set of thresholds α and the final decision on the ID depends on his inspection of the ID plots. Yet for reasons already indicated and further illustrated in the next section, without prior knowledge a fully automated procedure based on local PCA which outputs the ID estimate given the data set does not make sense.

4 Experimental Results

In order to provide an impression of the characteristics of our ID estimation procedure we here apply it to a mixture of noisy data sets of different intrinsic structure and dimensionality. The data set is descibed and illustrated in figure 1.

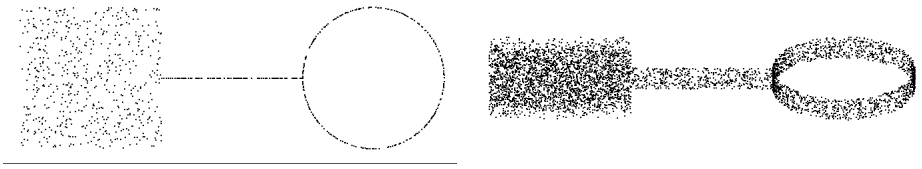

Fig. 1. Two views of the Square-Line-Circle data set. The 3d data set consists of 5000 random dots within a circle, a line and a square in the xy-plane with uniform noise in the z-direction. The noise has a variance of 1/12. The data density is approximately uniform over the data set. Left: View on the xy-plane, Right: Rotation of 60° around x-axis

Figure 2 shows the ID estimation procedure in progress for a growing number of pointers on the D10 level. From top to bottom, left to right with 5, 10, 20, 35, 45, and 70 nodes in the $OTPM$. Dark circles indicate a local ID estimate of one, medium dark circles an estimate of two and light circles of three (D10 criterion). For only five nodes the $OTPM$ indicates a one dimensional connection structure for the circle and the line and a two dimensional one for the square, identical to the ID estimates (by local PCA of the $OTPM$). For 10 nodes the $OTPM$ has already grasped the intrinsic structure of the data set. For 20 nodes we also get the correct local ID estimates for the line-data and the square but the ID estimate of the circle data is still two instead of one. This is due to the curvature (non-linearity) of the circle. From 35 to 45 nodes even the true ID of the circle is revealed because the number of pointers has now become large enough for a linear approximation of the circle on the D10 level. For even higher numbers of pointers the distribution of pointers as obtained by the LBG algorithm will eventually approximate the noise, i.e. leave the surface. From now on (see figure 2 for 70 nodes) the ID will be overestimated.

The mean squared quantization error for the Square-Line-Circle data set

$$mse = \frac{1}{|T|} \sum_{i=1}^{N} \sum_{x \in V_i} \| x - c_i \|^2 \tag{5}$$

for e.g. $N = 45$ nodes is 0.29 which is only about three times the variance of the noise. Subtracting the noise variance, only two times the noise variance remains for the average local intra-surface variance. Clearly, a simple local PCA approach as e.g. that of Fukunaga et al. (taking the unfiltered data as input to

the local PCA) would not yield the correct local ID estimates on a D10 level for that local region size but would detect the noise variance as a second or third most significant eigenvalue on any level. This is what we refer to as the increased robustness against noise and the increased discrimination ability of our procedure.

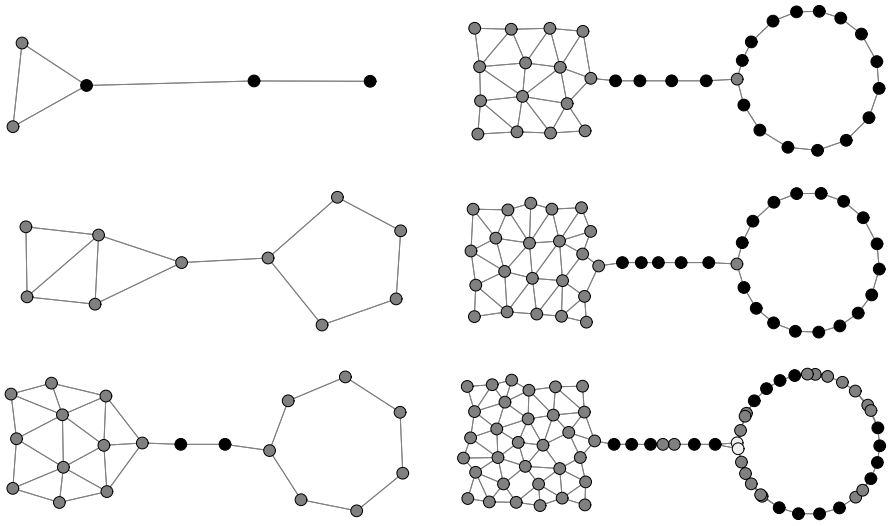

Fig. 2. Local ID estimation for the Square-Line-Circle data set for a growing number of pointers (nodes in the $OTPM$) on the D10 level. From top to bottom, left to right: 5, 10, 20, 35, 45, 70 nodes. Dark circles indicate a local ID estimate of one, medium dark circles an estimate of two and light circles of three dimensions.

Further applications of our ID estimation technique, including ID estimation of a sequence of full scale images, can be found in [BS97]. Due to limited space they had to be omitted here.

5 Conclusion

We have presented an algorithm for estimating the intrinsic dimensionality of low dimensional submanifolds embedded in high dimensional feature spaces. The algorithm belongs to the category of local ID-estimation procedures, is based on local PCA and directly extends and improves its predecessor, the algorithm of Fukunaga and Olsen, [FO71], in terms of computational complexity and noise sensitivity. The main ideas are first to cluster the data, second to construct an $OTPM$ and third to use the $OTPM$ and not the data itself for local PCA.

Besides tests on an illustrative artificial data set (this article) the procedure has been successfully applied to ID-estimation of image sequences with image

resolutions of up to 256×256 pixels, [BS97]. Such application is out of reach for conventional ID-estimation procedures based on local PCA and to the best of our knowledge has not been tackled before.

*OTPM*s together with eigenvectors and eigenvalues returned by local PCA are not only useful for ID estimation but can be used for linear approximation of the data and construction of auto-associators in quite an obvious way. Such associators will work by projecting new data to the local subspaces spanned by the eigenvectors, i.e. by projecting to a linear approximation of the manifold.

References

[Ben69] R. S. Bennett. The intrinsic dimensionality of signal collections. *IEEE Transactions on Information Theory*, 15:517–525, 1969.

[BS97] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. Technical Report 9703, Inst. f. Inf. u. Prakt. Math. Christian-Albrechts-Universitaet zu Kiel, 1997.

[FO71] K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 20(2):176–183, 1971.

[JD88] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[KL94] N. Kambhatla and T.K. Leen. Fast non-linear dimension reduction. In *Advances in Neural Information Processing Systems, NIPS 6*, pages 152–159, 1994.

[Kru64] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[LBG80] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transaction on Communications*, 28(1):84–95, 1980.

[MS94] T. Martinetz and K. Schulten. Topology representing networks. In *Neural Networks*, volume 7, pages 505–522, 1994.

[PBJD79] K. Pettis, T. Bailey, T. Jain, and R. Dubes. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 1:25–37, 1979.

[PTVF88] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C - The Art of Scientific Computing*. Cambridge University Press, 1988.

[Tru76] G. V. Trunk. Statistical estimation of the intrinsic dimensionality of a noisy signal collection. *IEEE Transactions on Computers*, 25:165–171, 1976.

[VD95] P. J. Verveer and R. P.W. Duin. An evaluation of intrinsic dimensionality estimators. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 17(1):81–86, 1995.

[VDM94] T. Villmann, R. Der, and T. Martinetz. A novel approach to measure the topology preservation of feature maps. *ICANN*, pages 289–301, 1994.

[Zad82] P. L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, 1982.

This article was processed using the LaTeX macro package with LLNCS style