Published in Proc. ICANN'97, Springer LNCS 1327, pp. 595 - 600

Topology Representing Networks 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 we compare two methods for intrinsic dimensionality (ID) estimation based on optimally topology preserving maps (OTPMs). The first one is a direct approach, where the intrinsic dimensionality is estimated directly from the OTPM. We argue that this approach suffers from both practical and theoretical pitfalls. The second is a new approach which combines OTPMs with an efficient local principal component analysis (PCA). Exploiting the OTPM, local PCA can be shown to have only *linear time complexity* w.r.t. the dimensionality of the input space (in contrast to the prohibitive *cubic complexity* of the conventional approach), and hence the method becomes applicable even for very high dimensional input spaces as frequently encountered in computer vision. A local ID estimate is then obtained as the local number of significant eigenvalues. In addition to ID estimation the local subspaces as revealed by our local PCA can be directly used for further data processing tasks including classification and regression.
The workability of the new approach for ID estimation and subspace auto-association is demonstrated on a sequence of 64×64 pixel images (4096-dimensional input space).

1 Introduction

The intrinsic, or topological, dimensionality (ID) of N patterns in an n-dimensional space refers to the minimum number of "free" parameters needed to generate the patterns. It essentially determines whether the n-dimensional patterns can be described adequately in a subspace (submanifold) of dimensionality $d < n$, [5]. As pointed out in [3], knowledge of the ID is important in order to determine the number of features necessary to represent the data, to decide whether a reasonable 2d or 3d representation exists or to estimate the effectiveness of algorithms depending on the ID, as e.g. methods for constructing classifiers or training neural networks. It can be greatly helpful in problems like pattern recognition, industrial or medical diagnosis and data compression.

In this article, we will concentrate on two local approaches to ID-estimation based on optimally topology preserving maps (OTPMs) (see e.g. [5] for alternative approaches). The first one, [3], tries to directly estimate the local ID from the number of neighbors of a node in an OTPM. The second one, [1], uses an OTPM for efficient local PCA and estimates the ID as the number of significant eigenvalues. It is conceptually similar to that of Fukunaga and Olsen, [4], 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. In contrast to the direct approach, the local subspaces revealed by local PCA can be further used for data modeling.

In the remainder of this article we will first review OTPMs in section 2. We will then discuss the approach of Frisone et al. in section 3. Our own approach is summarized in section 4, and in section 5 we show how our local subspaces can be utilized for data modeling. A demonstration is given in section 6, and we close with a brief summary and outlook in section 7.

2 Optimally Topology Preserving Maps

Optimally Topology Preserving Maps (OTPMs) are closely related to Martinetz' Perfectly Topology Preserving Maps (PTPMs) [7] and are constructed in just the same way. The only reason to introduce them separately is that in order to form a PTPM the centers must be "dense" in the manifold M. Without prior knowledge this assumption cannot be checked, and in practice it will rarely be valid. OTPMs emerge if just the construction method for PTPMs is applied without checking for the density condition. Only in favorable cases one will obtain a PTPM (probably without noticing). OTPMs are nevertheless optimal in the sense of the topographic function introduced by Villmann in [11]: In order to measure the degree of topology preservation of a graph G with an associated set of centers 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 centers 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 $OTPM_T(S)$ is constructive: Simply pick $x \in T$ according $p_T(x)$, calculate the best and second best matching centers, c_{bmu} and c_{smu}, and connect bmu with smu. This procedure is just the essence of Martinetz' Hebbian learning rule for topology representing networks. Obviously, for a finite training set T the $OTPM_T(S)$ can be constructed in time $O(|T|)$. For a training set defined via a pdf $p_T(x)$, G will converge to $OTPM_T(S)$ with probability one.

For our purposes, $OTPM_T(S)$ has two important properties. First, it does 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.

3 Direct ID estimation with OTPMS

Frisone et al. have been the first ones trying to exploit the benevolent properties of OTPMs for ID estimation. They tried to directly infer the ID from the number of direct neighbors of nodes in an OTPM by relating this number to the maximum kissing number in sphere packings (Kiss-SPP). The problem here is to find a packing of d-dimensional spheres of equal size so that the number τ of spheres touching (kissing) each other is maximal [2]. Kiss-SPP has only been solved for $d = 1, 2, 3, 8, 24$ and there exist optimal solutions for lattices of spheres for $d = 4, 5, 6, 7$, [2].

Analyzing the hypothetical analogy between the number of neighbors and the maximum kissing number one realizes that it rests on three assumptions: First, that the centers have been optimally distributed in the manifold (in the sense of the lowest quantization error), second, that the optimal distribution is realized by a lattice quantizer and third, that the problem of finding the best lattice quantizer is equivalent to finding the lattice with highest kissing number.

While there is some evidence that the last two assumptions hold at least for small d, they are in fact open questions, [2]. Anyway, lattices and other regular (optimal) center distributions can only emerge for very large number of centers (infinitely many) and, of course, an even larger numbers of training samples. Finally, a vector quantization algorithm generating the optimal distribution for this large number of centers (by annealing?) in finite time does not exist.

This requirement for a huge number of training data, long training times and lack of theoretical foundation appears to exclude the direct approach from practical applications.

4 Efficient ID estimation based on local PCA of OTPMs

Similar to the direct approach of Frisone, our ID estimation procedure rests on the fact that the number of neighbors of a node in an OTPM only depends on the intrinsic dimensionality d and is independent of the input dimensionality n.

It proceeds in four stages (batch-variant). First, generate a set of N centers $S = \{c_1, \ldots, c_N\}$ as the output of a vector quantization algorithm working on the training set T. Second, calculate the graph G as the optimally topology preserving map, $OTPM_T(S)$, of S w.r.t. T. Third, for each node $i \in G$ perform a principal component analysis of its correlation matrix $\frac{1}{m_i} A^T A$, $A^T = [c_{1_i} - c_i, \ldots, c_{m_i} - c_i]$, with $(c_{j_i} - c_i)$ the difference vectors between c_i and c_{j_i}, the center of its j-th direct topological neighbor in G. Finally, exclude eigenvectors corresponding to very small eigenvalues.

As a result of the vector quantization stage the centers are placed within the manifold M and noise orthogonal to M is filtered out. $OTPM_T(S)$ is constructed by simply connecting nodes corresponding to best and second best matching centers on presentation of T.

The central "trick" is to use the difference vectors $(c_{j_i} - c_i)$ for PCA of each local subspace and not the data in a local region itself, as e.g. in [4] or

[6]: First, the difference vectors have very low noise component orthogonal to M (due to the noise reduction property of the vector quantizing stage), and second, the number of neighbors m_i of a node in an OTPM does only depend on the intrinsic dimensionality d and is small for small d. Straightforward PCA of the correlation matrix $\frac{1}{m_i}A^TA$ nevertheless would take time $O(n^3)$,[9], yet the m_i eigenvectors and m_i eigenvalues can be obtained by PCA of AA^T as well, cf. [8], taking only time $O(m_i^3)$. Since AA^T clearly can be computed in time $O(m_i^2 n)$, and the number of neighbors m of a node in an OTPM does not depend on n but the intrinsic dimensionality d, local PCA of the correlation matrix takes only time $O(m(d)^2 n + m(d)^3)$ and hence scales only linearly (optimally) with the input dimensionality.

Deciding, what size an eigenvalue as obtained by each local PCA must have to indicate an associated intra-manifold eigenvector, amounts to determining a threshold. We adopted the Dα criterion from Fukunaga et. al., [4], that regards an eigenvalue μ_i as significant if $\frac{\mu_i}{\max_j \mu_j} > \alpha\%$. If no prior knowledge concerning the distribution of the noise is available, different values of α have to be tested.

5 Local subspaces for data modeling

Local subspace analysis as described in section 4 supplies us with a set of (orthonormal) eigenvectors $e_1^i, \ldots, e_{l_i}^i$, $l_i \leq m_i$, spanning the local subspace for each center $c_i \in S$. These subspaces can be used straightforwardly to improve existing local approximation schemes including RBF networks and Local Linear Maps (LLMs),[10], by first projecting stimuli to the relevant subspaces. Here we demonstrate, how local subspaces can be used for compact coding by locally linear data modeling, [6]. In this approach, new data is modeled as

$$\hat{x} = c_{bmu} + \sum_{i=1}^{l_{bmu}} ((x - c_{bmu})^T e_i^{bmu}) e_i^{bmu}, \tag{1}$$

i.e. as the center of the best matching unit (Euclidean distance) and the projection to the subspace of the bmu, respectively. Using speech and (pre-processed) image data with typically low intrinsic dimensionality, Kambhatla and Leen demonstrated that this method compares well to (and even outperforms) standard bottle-neck Backpropagation networks. They, however, used conventional PCA on the data in the Voronoi cells. With help of local PCA based on OTPMs, local linear modeling now scales up linearly for high dimensional input spaces.

6 Experimental Results

In this demonstration we want to investigate an image sequence generated by taking 180 snapshots (every 2°) with a resolution of 64×64 pixels (4096-dimensional input space) of a robot rotating a cylindric grey ramp around its z-axis (from 0° to 360°). Since the background remains constant, the images lie on a closed 1-dimensional trajectory in image space with ID $d = 1$.

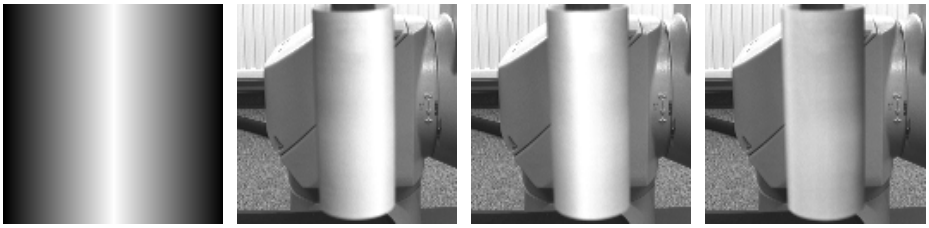

Fig. 1. Grey ramp under different rotations. From left to right: Original (symmetric) grey ramp, grey ramp wrapped around a bottle with part of the robot arm in the background under 0°, 45°, 90° rotation

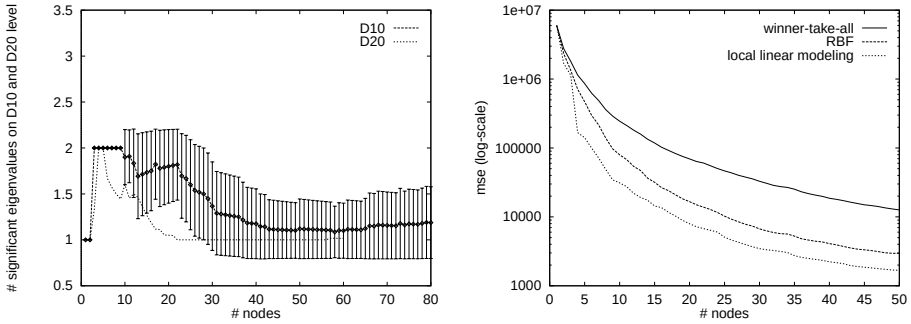

Fig. 2. Left: ID plots for rotating grey ramp on D10 and D20 level with error bars for the D10 level. Right: Reconstruction error (mse) for locally linear modeling, RBF and winner-take-all network on the D20 level.

Figure 2, left, shows the ID estimates obtained as the mean number of significant local eigenvalues by our procedure for different numbers of centers on the D10 and D20 level. The standard deviations of the estimates are included as error bars. The plots clearly indicate an ID of one or at least two. In figure 2, right, the reconstruction error (mse) for local linear modeling with the subspaces constructed on the D20 level is depicted (averaged over 180 test images). For comparison, the reconstruction error obtained with an RBF network and a simple winner take all scheme (same center distributions[1]) are also included. For a given number of centers, local linear modeling is clearly superior.

7 Summary

We have investigated two algorithms for ID estimation based on OTPMs. While the first approach pioneers an interesting idea, it generally turns out to be of

[1] As a vector quantizer for generating the center distributions we used an incremental version of the LBG algorithm, cf. [1]. Adding the $(N+1)$.th center where quantization is worst and keeping the old distribution of the remaining N centers, the LBG algorithm only needs to adjust centers in the near surrounding of the new one.

little practical value. The second approach combines OTPMs with local PCA and thereby directly extends and improves the classical algorithm of Fukunaga and Olsen, [4], in terms of computational complexity and noise sensitivity. Scaling only linearly with the dimensionality of the input space, it turns out to be ideally suited for local subspace approximation of low dimensional manifolds in high dimensional input spaces in general. These subspaces can then be utilized by local subspace methods, as e.g. local linear data modeling.

Going beyond this article, we currently apply our local subspace construction method for constructing Hyper Basis Function networks and improved LLMs for visual learning. First results concerning appearance based robot grasping and pose recognition are very encouraging.

References

1. 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. (submitted to IEEE PAMI).
2. J.H. Conway and N.J.A. Sloane. *Sphere Packings, Lattices and Groups*. Grundlehren der mathematischen Wissenschaften 290. Springer Verlag NY, 1988.
3. F.Frisone, P.Morasso, F.Firenze, and L.Ricciardiello. Application of topology-representing networks to the estimation of the intrinsic dimensionality of data. In *Proc. of the International Conference on Artificial Neural Networks*, volume 1, pages 323–327, 1995.
4. K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 20(2):176–183, 1971.
5. A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentince Hall, 1988.
6. N. Kambhatla and T.K. Leen. Fast non-linear dimension reduction. In *Advances in Neural Information Processing Systems, NIPS 6*, pages 152–159, 1994.
7. T. Martinetz and K. Schulten. Topology representing networks. In *Neural Networks*, volume 7, pages 505–522, 1994.
8. H. Murase and S. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.
9. 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.
10. H. Ritter, T. Martinetz, and K. Schulten. *Neuronale Netze*. Addison-Wesley, 1991.
11. T. Villmann, R. Der, and T. Martinetz. A novel aproach to measure the topology preservation of feature maps. *ICANN*, pages 289–301, 1994.

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