Intrinsic Dimensionality Estimation With Optimally Topology Preserving Maps

J. Bruske and G. Sommer

Abstract—A new method for analyzing the *intrinsic dimensionality* (ID) of low-dimensional manifolds in high-dimensional feature spaces is presented. Compared to a previous approach by Fukunaga and Olsen, the method has only *linear instead of cubic time complexity* w.r.t. the dimensionality of the input space. Moreover, it is *less sensitive to noise* than the former approach. Experiments include ID estimation of synthetic data for comparison and illustration as well as ID estimation of an image sequence.

Index Terms—Intrinsic dimensionality estimation, topology preservation, principal component analysis, vector quantization.

1 Introduction

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 [3]. (It has long been noticed that this 19th-century notion of dimensionality is unprecise and fraught with problems, see, e.g., [1] for a short review. Yet there exists a precise definition of the *topological dimensionality*, given by Brouwer in 1913 [2], and it is this type of dimensionality we try to estimate, as opposed to the *fractal* or *Hausdorff* dimension.) It essentially determines whether the n-dimensional patterns can be described adequately in a subspace (submanifold) of dimensionality $m < n$. By providing a bound on the number of parameters needed to describe a data set, ID estimation is a valuable tool in system identification, classifier and regressor design as well as in data visualization. If the ID of a data set is two or three, the data can be mapped to a 2D or 3D map [4] and visualized for monitoring or diagnosis purposes without distortions. In classifier and regressor design, particularly within the neural network approach, the complexity of classifiers (number of basis functions, hidden units) with best generalization properties depends on the ID [5]. In, e.g., nonlinear dimension reduction with auto-associative five-layer bottle-neck networks [6], the number of hidden units in the encoding middle layer should directly correspond with the ID. Another example is local linear modeling of data [7], where the dimension of local subspaces should correspond with the local ID. The approach presented in this paper not only returns the local ID estimates but also the sets of orthonormal vectors spanning the local subspaces and hence can be directly used for local linear modeling.

Adopting the classification in [3], there are two primary approaches for estimating the intrinsic dimensionality. 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 [8] and its successors as well as variants of MDSCAL [9] 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 patterns to a lower-dimensional space. Pettis et al.'s [10], Fukunaga and Olsen's [11], as well as Trunk's [12] and Verveer and Duin's method [13] 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). It is conceptually similar to that of Fukunaga and Olsen (abbreviated ID_FO in the following) 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.

2 ID Estimation With OTPMs

The local approaches to ID estimation assume that the patterns $x \in T$, T the data set, are noisy samples of a differentiable vector valued function

$$F: R^d \to R^n, \quad x = f(k) + \eta \tag{1}$$

where $k = [k_1, \ldots, k_d]$ is a d-dimensional vector of d independent parameters and $\eta \in R^n$ denotes the noise. The function f can be imagined to describe a d-dimensional surface in n-dimensional space, which, due to the noise, is not infinitely thin. Assuming that in small local regions this surface can be linearly approximated by d-dimensional hyperplanes, the basic idea behind ID_FO is to perform local PCA in local regions of the data set. Ideally, each local PCA should return d significant eigenvalues corresponding to the spanning vectors $s_1, \ldots, s_d$ of the d-dimensional hyperplane (and hence indicating a local ID of d) and $n - d$ eigenvalues being either zero or insignificant (indicating noise) corresponding to eigenvectors $n_1, \ldots, n_{n-d}$ perpendicular to the surface. For this approach to work, within each local region the largest variance in direction perpendicular to the surface must be much smaller than the smallest variance in direction of the surface, i.e.,

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

Here, $Var(s_i)$, the intrasurface variance, depends on the size of the local region and the pattern density, and $Var(n_j)$ depends on the variance caused by the noise *and* the fact that the surface cannot be exactly represented as a linear surface. This leads to a *curvature/noise dilemma* for ID estimation algorithms based on local PCA: If the region is too large, $Var(n_j)$ might be high due to the curvature of the surface. If, on the other hand, the region is too small, the noise is still there and will eventually dominate $Var(s_i)$. A more basic problem is that ID estimation in the presence of noise actually becomes an ill-posed problem because without prior knowledge observing variance in some direction, it is impossible to tell whether it corresponds to intrasurface variance or noise. Imagine just a straight line in 3D space and some small uniform noise in the sampling process. Does the data describe a (1D) line or a filled (3D) cylinder? The answer depends on the scale (size of the local region) the data is looked at, and consequently an ID estimation algorithm should provide estimates on different scales, leaving the final decision to the (biased) user.

In ID_FO as well as our procedure, both the curvature/noise dilemma and the ill-posedness of the problem are dealt with by providing estimates for different local region sizes (i.e., on different scales). By tracking the development of ID estimates as a function of different local region sizes, the user must decide which estimate most closely fits his expectations. Of course, in order to make local PCA approaches work, the data set has to be large enough to represent the nonlinearities and to allow for filtering out the noise.

Our algorithm improves ID_FO with respect to computing time (linear instead of cubic scaling of PCA with the input dimension)

• *The authors are with the Computer Science Institute, Christian-Albrechts University Kiel, Preusserstr. 1-9, 24105 Kiel, Germany. E-mail: jbr@informatik.uni-kiel.de.*

Manuscript received 6 Feb. 1997; revised 2 Mar. 1998. Recommended for acceptance by V. Nalwa.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 106429.

0162-8828/98/$10.00 © 1998 IEEE

and noise sensitivity by not working on the data itself but on an intermediate representation as introduced in the following.

2.1 Optimally Topology Preserving Maps

The idea behind *topology preserving maps* is to represent a manifold $M \subseteq R^n$ by a topology preserving graph $G = (V, E)$, $|V| = N$, in which each node $i \in V$ is associated with a pointer $c_i \in M$ and a pattern $x \in M$ is mapped to

$$i = \arg\min_i \{\|c_i - x\|\}.$$

In order to be topology preserving, pointers c_i, c_j which are neighbored in M should be associated with nodes adjacent in G and vice versa. Defining pointers in M to be neighbored if their induced Voronoi cells $V_i^{(M)}$,

$$V_i^{(M)} = \left\{x \in M \middle| \|c_i - x\| \le \|c_j - x\|,\ 1 \le j \ne i \le N\right\},$$

have a common border, G is called a *perfectly topology preserving map* [14] if it corresponds to the *induced Delaunay triangulation* $D_M(S)$ of the set of pointers $S = \{c_1, \ldots, c_N\}$, i.e.,

$$(i, j) \in E \Leftrightarrow V_i^{(M)} \cap V_j^{(M)} \ne \emptyset.$$

Now, how can we construct a topology preserving map given the pattern set $T \subseteq R^n$ and a pointer set S without knowledge of M? The idea is to construct the graph $G = (V, E)$ by simply going through the pattern set and for each $x \in T$ to calculate the best and second best matching pointers, c_{bmu} and c_{smu}, and to connect *bmu* with *smu* in G. We call such a graph fulfilling

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

the optimally topology preserving map[1] $OTPM_T(S)$ of S given the training set T. As follows directly from Theorem 3 in [14], $OTPM_T(S)$ is perfectly topology preserving, $OTPM_T(S) = D_M(S)$, if $T = M$ and S is "dense" in M. In the more general case that $M \ne T \subset R^n$ and S is not dense, the condition

$$\forall x \in T \quad (c_{bmu(x)}, c_{smu(x)}) \in D_M(S), \tag{3}$$

ensures that $OTPM_T(S)$ contains no edges not also contained in $D_M(S)$. While (2) implicitly relates the local pattern density, noise level, and curvature for ID-estimation by local PCA to make sense, condition (3) relates curvature, pointer (sampling) density, and level of noise for perfect topology preservation of $OTPM_T(S)$. Additionally, in order to obtain all edges in $D_M(S)$, the pattern density must be high enough to induce them during OTPM construction. The number of pointers must increase at least linear with increasing ID (as the simplest d-dimensional geometric entity is the d-dimensional simplex with d + 1 pointers) and hence the number of patterns must increase at least quadratic (the d-dimensional simplex has $\frac{(d+1)d}{2}$ edges).

For our purposes, $OTPM_T(S)$ 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 The ID Estimator

The improved ID estimation procedure ID_OTPM based on OTPMs is summarized in Fig. 1. For a growing number of up to

1. "Optimal" here refers to the topographic function of T. Villmann et al. [15] qualify $OTPM_T(S)$ as the map with the highest degree of topology preservation.

```
input   training set T ⊂ R^n,
        maximal number of pointers N_max,
        significance level α.

S_1 = { arbitrary x ∈ T }
for N = 1 to N_max {
   S'_N = LBG(T, S_N)
   G = OTPM_T(S'_N)
   for_all_nodes (i ∈ G){
      Q_i = {(c_j − c_i) | c_i, c_j ∈ S'_N ∧ (i, j) ∈ E_G}
      ID_i = #_significant_eigenvalues(PCA(Q_i), α)
   }
   S_{N+1} = S'_N ∪ arg max_{x∈T} {min_{c_i ∈ S_N} ||x − c_i||}
}
```

Fig. 1. ID_OTPM: local ID estimation with OTPMs.

$N_{max} \le |T|$ pointers (corresponding to the shrinking local region sizes in ID_FO), it proceeds as follows:

- First, generate a set of N pointers $S'_N = \{c_1, \ldots, c_N\}$ as the output of a vector quantization algorithm working on the training set T. Here we use the LBG [16] algorithm LBG(T, S_N) with initial pointer set S_N.
- Second, calculate the graph G as the optimally topology-preserving map, $OTPM_T(S'_N)$, of S'_N for T.
- Third, for each node $i \in G$, perform a principal component analysis on the set $Q_i \subset R^n$ consisting of all the m_i difference vectors $(c_j - c_i)$ between the pointer of node i and all of its m_i direct neighbors in G. Estimate the local intrinsic dimensionality ID_i as the number of significant eigenvalues (see below) as returned by PCA. Finally, for the next round, extend the initial pointer set for the LBG-stage by including the pattern with the largest quantization error.

As a result of the vector quantization stage, the pointers are placed on the principal surfaces of M, and noise orthogonal to M is largely filtered out. $OTPM_T(S'_N)$ is constructed by simply connecting nodes corresponding to the best and second-best matching pointers on presentation of T.

As mentioned before, the central "trick" is to use the difference vectors $(c_j - c_i)$ for PCA of each local subspace and not the data in a local region itself: 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 depends only on the intrinsic dimensionality d and is small for small d. Straightforward PCA of Q_i nevertheless would take time $O(n^3)$, [17], yet the number of vectors in Q_i is m_i and hence the local PCAs can be performed in time $O\left(m_i^2 n + m_i^3\right)$. Since LBG() and OTPM-construction scale linear in n as well, ID_OTPM scales linearly (optimally) with the input dimensionality.

Following the discussion in Section 2, the problem of deciding whether an eigenvalue is significant is again ill-posed, because one does not know the intrasurface variance, the noise, and the curvature to concretize "$\gg 1$" in (2). We have adopted the same $D\alpha$ criterion as did Fukunaga and Olsen regarding an eigenvalue μ_i as significant if

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

TABLE 1
COMPUTING TIME WITH 32 LOCAL REGIONS AS A FUNCTION OF THE INPUT DIMENSION n FOR THE HELIX DATA SET ON A SPARC 4 WORKSTATION

n	ID_FO t_{est}[sec]	ID_OTPM t_{Lind}[sec]	ID_OTPM t_{est}[sec]
3	0.79	0.79	1.2
50	14.75	5.48	8.36
100	38.97	9.24	13.86
150	78.4	11.62	17.69
200	143.67	16.74	24.5

Typical values for α are five, 10, and 20. As in ID_FO, different values of α have to be tested, and again it is up to the user to prefer a certain interpretation. Yet in ID_OTPM, noise is largely reduced, and hence for the same amount of noise on the data, α can usually be smaller than in ID_FO.

3 EXAMPLES

Previous work on comparing different ID estimators including ID_FO on artificial and "natural" data sets [18], [13] concluded that despite the need for interpretation and interaction by the user (which we think is indispensable), ID_FO is one of the most reliable and easy-to-use methods, yet confirmed that it is sensitive to noise and suffers from quickly increasing computing time with increasing dimensions. Here we want to demonstrate that ID_OTPM overcomes these two problems using a helix data set and an image sequence. More experiments demonstrating the workability of ID_OTPM can be found in [19], including examples for ID up to five.

The helix data set consists of 1,000 noised samples generated by

$$(x_1, x_2, x_3) = \left(r\cos t, r\sin t, \frac{p}{2\pi}t\right) + (\eta_1, \eta_2, \eta_3),$$

$$\eta_i \text{ uniform from } \left[-a_\eta, +a_\eta\right],\ r = 2,\ p = 2,\ t \in [0, 4\pi]$$

Table 1 shows the *scaling of computing time*[2] with increasing input dimension n for the helix data set. Data was generated with $a_\eta = 0.5$, the additional $n-3$ dimensions being filled up with uniform noise with the same amplitude a_η. Using the cyclic Jacobi method as described in [17] for eigenvalue decomposition in both algorithms, ID_OTPM scales linearly with n and for 200 dimensions is already six times faster than the superlinearly (cubic) scaling ID_FO. Note, that more than half of the time of ID_OTPM is used for vector quantization. The helix is also used to compare the noise sensitivity of the two algorithms. Fig. 2 and Fig. 3 show the global[3] ID estimates obtained by each algorithm as a function of the number of nodes in ID_OTPM (respectively, local regions in ID_OF) for different noise amplitudes a_η on the D20 level. While ID_OTPM indicates the true ID for a_η up to 1.0, ID_OF has problems even for $a_\eta = 0.5$. For $a_\eta = 1.5$, ID_OF indicates the full input dimension, while ID_OTPM returns an ID estimate of about two, indicating the near cylindrical (2D) distribution of the data for that amount of noise.

In a second experiment, we investigate an image sequence generated by taking 180 snapshots (every 2°) with a resolution of 256 × 256 pixels (65,536-dimensional input space) of a robot rotating a cylindric gray ramp around its z-axis (from 0° to 360°), see Fig. 4. Since the background remains constant, the images lie on a closed 1D trajectory in image space with ID $d = 1$. The noise in the measurement process is approximately Gaussian with a standard deviation of 1.75 gray values per pixel. ID-estimation with ID_OTPM on the D05 level (Fig. 5) indicates that the ID is at most two.[4] Estimation on the D10 level indicates an ID between one and two, whereas estimation on the D20 level indicates an intrinsic dimensionality of one, the true ID. It is interesting to notice that in spite of the 65,536-dimensional input space, the ID-estimate never exceeds two on all three levels. The explanation, revealed by an analysis of the *OTPMs* for each number of nodes, is that the edges in the *OTPM* actually form a (1D) trajectory, i.e., the intrinsic structure (topology) is correctly represented by a 1D graph. Due to the nonlinearity of the trajectory, however, the local PCA taking the two difference vectors of a pointer and its two topological neighbors as input does not indicate a 1D local structure on each level.

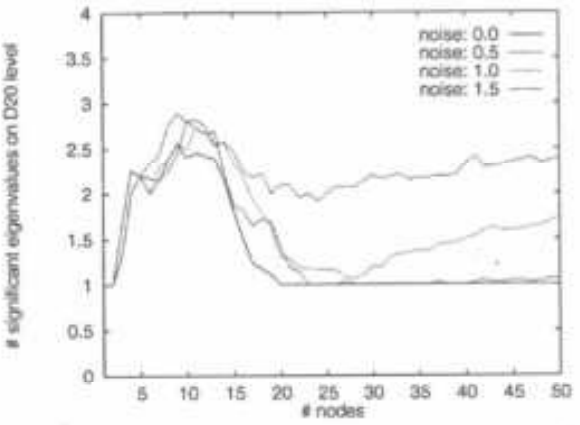

Fig. 2. ID estimates for ID_OTPM on D20 level for helix data set with different amounts of noise.

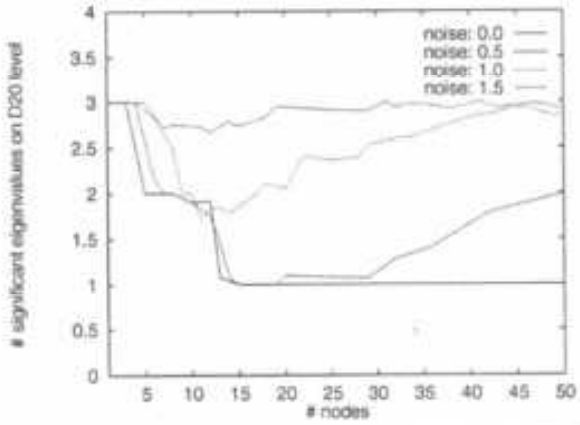

Fig. 3. ID estimates for ID_FO on D20 level for helix data set with different amounts of noise.

4 DISCUSSION

We have presented an algorithm for estimating the intrinsic dimensionality of low-dimensional submanifolds embedded in high

2. Implementation and compilation are not optimized.

3. Global ID estimates are obtained by averaging over all local ID estimates.

4. The reader should bear in mind that in this and the following experiment, we do *not* try to estimate any properties of the objects in the scene, e.g., the shape of the cylinder, but the number of free parameters that generated the image sequence. Each image is just treated as one point in 65,536-D image space.

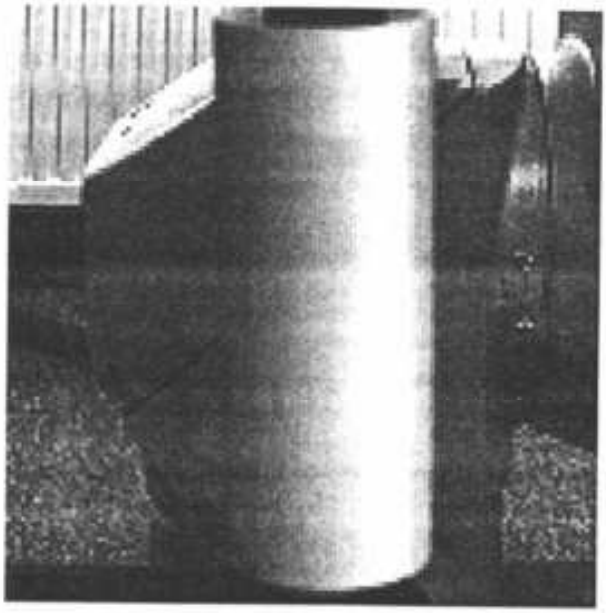

Fig. 4. Rotating gray ramp with part of the robot arm in the background.

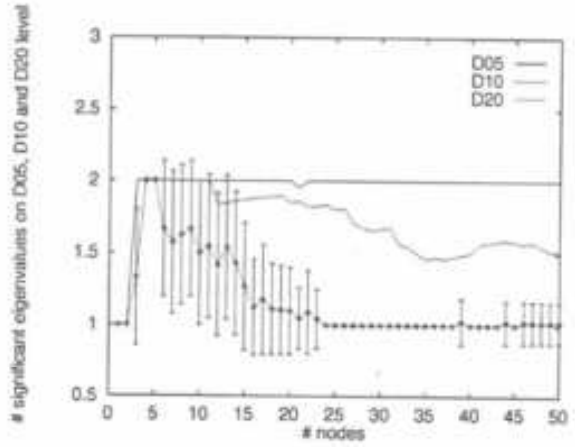

Fig. 5. ID plots on D05, D10, and D20 levels. Error bars indicate the standard deviation of the global ID estimate for the D20 level.

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 [11], 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.

Clustering is responsible for an even distribution of the cluster pointers and for noise reduction, i.e., placing the pointers in the manifold. The local PCA taking difference vectors of pointers as an input benefits from the noise reduction property of the clustering stage. Its output, the eigenvalues, gives a better hint at the local ID than those of straightforward local PCA on the data itself always including the full variance of the noise.

Constructing the *OTPM* for the cluster pointers provides a low-dimensional representation of the data which optimally reflects the intrinsic (topological) structure of the data. Independent of the dimension of the input space and invariant w.r.t. scaling and rigid transformations, it provides an ideal basis for ID estimation. Exploiting the *OTPM* for local PCA, our ID estimation procedure has only linear time complexity in the dimension of the input space, and the invariance properties directly transfer to the estimate.

*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 the linear approximation of the manifold.

References

[1] B. Mandelbrot, *Die fraktale Geometrie der Natur*. Basel: Birkhaeuser Verlag, 1991.
[2] A. Heyting and H. Freudenthal, *Collected Works of L.E.J. Brouwer*. New York: North Holland Elsevier, 1975.
[3] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, N.J.: Prentice Hall, 1988.
[4] T. Kohonen, *Self-Organizing Maps*. New York: Springer, 1995.
[5] R. Duin, "Superlearning Capabilities of Neural Networks?" *Proc. Eighth Scandinavian Conf. Image Analysis*, pp. 547-554, Tromso, Norway, 1993.
[6] E. Oja, "Data Compression, Feature Extraction, and Autoassociation in Feedforward Neural Networks," *Artificial Neural Networks*. New York: Elsevier Sciences, 1991, pp. 737-745.
[7] N. Kambhatla and T.K. Leen, "Fast Non-Linear Dimension Reduction," *Advances in Neural Information Processing Systems*, NIPS 6, pp. 152-159, 1994.
[8] R.S. Bennett, "The Intrinsic Dimensionality of Signal Collections," *IEEE Trans. Information Theory*, vol. 15, pp. 517-525, 1969.
[9] J.B. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, vol. 29, pp. 1-27, 1964.
[10] K. Pettis, T. Bailey, T. Jain, and R. Dubes, "An Intrinsic Dimensionality Estimator From Near-Neighbor Information," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 1, pp. 25-37, 1979.
[11] K. Fukunaga and D.R. Olsen, "An Algorithm for Finding Intrinsic Dimensionality of Data," *IEEE Trans. Computers*, vol. 20, no. 2, pp. 176-183, 1971.
[12] G.V. Trunk, "Statistical Estimation of the Intrinsic Dimensionality of a Noisy Signal Collection," *IEEE Trans. Computers*, vol. 25, pp. 165-171, 1976.
[13] P.J. Verveer and R.P. Duin, "An Evaluation of Intrinsic Dimensionality Estimators," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 1, pp. 81-86, Jan. 1995.
[14] T. Martinetz and K. Schulten, "Topology Representing Networks," *Neural Networks*, vol. 7, pp. 505-522, 1994.
[15] T. Villmann, R. Der, and T. Martinetz, "A Novel Approach to Measure the Topology Preservation of Feature Maps," *ICANN*, pp. 289-301, 1994.
[16] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84-95, 1980.
[17] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C—The Art of Scientific Computing*. Cambridge England: Cambridge Univ. Press, 1988.
[18] N. Wyse, R. Dubes, and A. Jain, "A Critical Evaluation of Intrinsic Dimensionality Algorithms," E. Gelsema and L. Kanal, eds., *Pattern Recognition in Practice*. New York: North-Holland Publishing Co., 1980, pp. 415-425.
[19] J. Bruske and G. Sommer, "Intrinsic Dimensionality Estimation With Optimally Topology Preserving Maps," Tech. Rep. 9,703, Inst. f. Inf. u. Prakt. Math., Christian-Albrechts-Universitaet zu Kiel, Kiel, Germany, 1997.