Estimating the Intrinsic Dimensionality of Hyperspectral Images

Jörg Bruske *, Erzsébet Merényi†

Abstract. Estimating the intrinsic dimensionality (ID) of an intrinsically low (d-) dimensional data set embedded in a high (n-) dimensional input space by conventional Principal Component Analysis (PCA) is computationally hard because PCA scales cubic ($O(n^3)$) with the input dimension [11]. Besides this computational drawback, global PCA will overestimate the ID if the data manifold is curved. In this paper we apply ID_OTPM [1], a new algorithm for ID estimation based on Optimally Topology Preserving Maps [7] to image sequences. In particular, we utilize ID_OTPM for ID estimation of an AVIRIS data set, a hyperspectral remote sensing image cube, with input dimension of the individual image planes $n = 257880$.

Most interestingly, our experiments suggest that the inter-band dimension d_b of the AVIRIS data set is between one and two, whereas the spectral dimension d_s is about four. These results provide important clues for compression, visualization and classification of the the AVIRIS data set.

1. Introduction

An example of hyperspectral imagery is from the sensor AVIRIS, which takes 194 spatially co-registered images at 194 different wavelengths (see [8], this volume). It is like a stack of 194 images of the exact same spatial region. The 194-dimensional spectrum associated with each spatial pixel identifies the surface material within the respective pixel. Yet building classifiers who automatically determine the surface material at a given pixel location from the 194-d reflectance spectrum for this pixel has turned out far from trivial (again see [8] for a brief overview), and hence a more detailed analysis of the data set seems appropriate. One question we might ask is how many "clusters" we can find in the data set, hoping that it corresponds to the number of surface classes. A more basic question is what the intrinsic dimensionality of the data set is like.

The intrinsic, or topological, dimensionality of N patterns in an n-dimensional space determines whether the n-dimensional patterns can be described

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

†Lunar and Planetary Laboratory, The University of Arizona, email: erzsebet@mars.lpl.arizona.edu

adequately in a subspace (submanifold) of dimensionality $m < n$ [4]. 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. For example, if the ID of a data set is 2 or 3, the data can be mapped to a 2 or 3 dimensional map [6] and visualized for monitoring or diagnosis purposes without distortions. And in classifier and regressor design, particular within the neural network approach, the complexity of classifiers (number of basis functions, hidden units) with best generalization properties is well known to depend on the ID [2].

Our approach to ID estimation (ID_OTPM) is based on optimally topology preserving maps (OTPMs) and local principal component analysis (PCA). It is conceptually similar to that of Fukunaga and Olsen [3] 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

For the convenience of the reader, we will now briefly review some basic properties of Optimally Topology Preserving Maps and provide a condensed description of our ID estimator. More details as well as an extended discussion can be found in [1].

2.1. Optimally Topology Preserving Maps

Optimally Topology Preserving Maps (OTPMs) are closely related to Martinetz' Perfectly Topology Preserving Maps (PTPMs) [7] and emerge if just the construction method for PTPMs is applied without checking for Martinetz' density condition[1]. 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 [12]: 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.

Definition 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

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

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

[1]This density condition could only be checked if the data manifold was known

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. Finally, if $T = M$ and if S is dense in M then $OTPM_T(S)$ will become a PTPM.

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.

2.2. Efficient ID estimation based on local PCA of OTPMs

Central to our ID estimation procedure is 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.

ID_OTPM 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 $\frac{1}{m_i}A^TA$, $A^T = [c_{1_i} - c_i, \ldots, c_{m_i} - c_i]$, 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 main "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 [3] or [5]: 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 $O(n^3)$ [11], yet the m_i eigenvectors and m_i eigenvalues can be obtained by PCA of AA^T as well, cf. [9], 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., [3], 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.

3. Experimental results

Given the 194 bands of the AVIRIS data set with 512×614 pixels each, the intrinsic dimensionality of this data set can be defined in two ways. The first is the inter-band dimensionality that is obtained if we regard each of the 194 bands (images) as a point in 512×614 dimensional image space. Hence we have the problem of estimating the ID for 194 257880-d points. The other way of looking at the data is to focus on the 194-d spectrum at each pixel and determine the the ID of 257880 194-d points. The latter is referred to as the spectral dimension.

Figure 1 depicts the results of applying ID_OTPM to the estimation of the inter-band dimensionality, working with 194 257880-d points. It shows the ID estimates obtained as the mean number of significant local eigenvalues by ID_OTPM for different numbers of centers on the 1% (D1) and 10% (D10) level. The standard deviations of the estimates are included as error bars on the D10 level. The plots clearly indicate an ID between one or two.

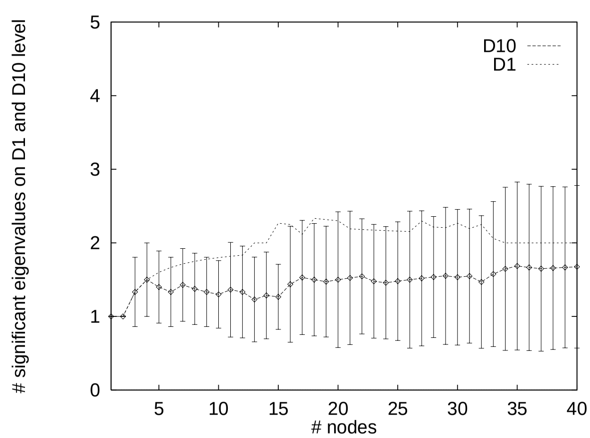

Figure 1: ID plots for estimating the inter-band dimension on D1 and D10 level with errorbars on D10 level.

On the other hand, figure 2 shows the ID-estimate for the spectral dimension, working with 257880 194-d points. ID estimation on the 10% (D10) level suggests that the spectral dimension is about 4. The plot for the 1% (D1) level confirms a low intrinsic dimensionality, yet taking into account more noise and curvature as on the 10% level returns higher estimates (whether the additional small eigenvalues bear important information or not can only be decided in classification experiments). Again, the plots were obtained as the mean number of significant local eigenvalues by ID_OTPM for different numbers of centers, standard deviations included as error bars on the D10 level.

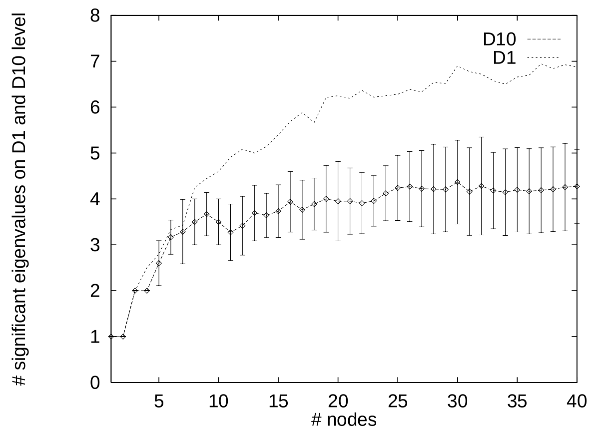

Figure 2: ID plots for estimating the spectral dimension on D1 and D10 level with errorbars on D10 level.

4. Conclusion and Outlook

Applying ID_OTPM we were able to estimate the inter-band dimensionality of the AVIRIS data set between one and two and the spectral dimensionality as about four. But what is the benefit of this? First of all, the results are encouraging in that they indicate that in both cases the ID is quite low and hence local approximation schemes (as e.g. mixture models, RBF networks or extended SOMs) are indeed promising candidates for *classification* of the AVIRIS data. Second, the low intrinsic dimensionality of the data allows *compression* of the data. If e.g. local linear modeling is applied [5], instead of transmitting 257880 194-dimensional vectors it suffices to initially transmit a small number of codebook vectors and to code each vector as the index of the best matching code-vector and the 4 projection coefficients to the local subspace (5-tuple). Third, in order not to work with 194-d inputs for a classifier one may try to *reduce the input dimension* to 4. Here, a an autoassociative bottleneck network with 4 hidden nodes lends itself for mapping the 194-d data to a 4 dimensional coordinate system [10]. Also, the low intrinsic dimensionality of the data justifies the use of 2-d SOMs for *visualization* or *clustering* of AVIRIS data, since the effective dimensionality reduction is only from 4 to 2 dimensions. Finally, the difference between inter-band and spectral dimensionality gives insight into the physical process: Tuning the wavelength (1 parameter) causes a smooth 1-dimensional transition between the different bands, yet reveals more information (4-d spreading) in the spectrum. This explains the enhanced discriminative power of multiband remote sensing

We want to point out that the approach presented in this paper does not only return the local ID estimates but also the sets of orthonormal vectors spanning the local subspaces which can be directly used for subspace modeling of the data.

References

[1] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE PAMI*, 20(5):572–575, 1998.

[2] R. Duin. Superlearning capabilities of neural networks? In *Proc. 8th Scandinavian Conference on Image Analysis*, pages 547–554, Tromso, Norway, 1993.

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

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

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

[6] T. Kohonen. *Self-Organizing Maps.* Springer, 1995.

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

[8] Erzsebet Merenyi. Self-organizing anns for planetary surface composition research. In *Proc. of the ESANN*, 1999.

[9] H. Murase and S. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[10] E. Oja. Data compression, feature extraction, and autoassociation in feedforward neural networks. In *Artificial Neural Networks*, pages 737–745. Elsevier Science Publishers, 1991.

[11] 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.

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