Parameterless Isomap with Adaptive Neighborhood Selection

Nathan Mekuz and John K. Tsotsos

Center for Vision Research (CVR) and
Department of Computer Science and Engineering,
York University, Toronto, Canada M3J 1P3
{mekuz, tsotsos}@cs.yorku.ca

Abstract. Isomap is a highly popular manifold learning and dimensionality reduction technique that effectively performs multidimensional scaling on estimates of geodesic distances. However, the resulting output is extremely sensitive to parameters that control the selection of neighbors at each point. To date, no principled way of setting these parameters has been proposed, and in practice they are often tuned ad hoc, sometimes empirically based on prior knowledge of the desired output. In this paper we propose a parameterless technique that adaptively defines the neighborhood at each input point based on intrinsic dimensionality and local tangent orientation. In addition to eliminating the guesswork associated with parameter configuration, the adaptive nature of this technique enables it to select optimal neighborhoods locally at each point, resulting in superior performance.

1 Introduction

Dimensionality reduction is a statistical tool commonly used to map data in high-dimensional space such as images, speech signals, etc. into lower dimensionality. The transformed data is often more suitable for regression analysis or classification than the original input data. Social sciences use dimensionality reduction extensively to uncover latent variables that explain observed phenomena. The underlying assumption is that observed high-dimensional samples lie on or near a lower-dimensional manifold embedded within the original high-dimensional space, and the purpose of the reduction is to project the high-dimensional data into a more compact representation while preserving certain properties of the data.

Traditional linear dimensionality reduction algorithms include Principal Component Analysis (PCA) [1] - a transformation that maximizes retained variance and Linear Discriminant Analysis (LDA) [2] - a projection that maximizes separation based on class labels. Nonlinear approaches include kernel PCA [3] - an application of linear PCA on data first transformed to typically higher dimensionality through some nonlinear kernel.

A recent surge in interest in locally linear manifold learning technique has resulted in the introduction of several new techniques, including, Isomap [4],

K. Franke et al. (Eds.): DAGM 2006, LNCS 4174, pp. 364–373, 2006.

Locally Linear Embedding (LLE) [5] and its derivatives Laplacian eigenmaps [6], Hessian eigenmaps [7] and others, for goals ranging from visualization problems to classification. These techniques view the manifold as a patchwork of connected linear surfaces, and attempt to preserve certain properties in the projection. If the manifold is continuous and sufficiently well sampled, then using Taylor's theorem, small patches can be approximated as linear. If parts of the manifold are linear, globally nonlinear methods may be overly complex, and difficult to train due to the large number of parameters. On the other hand, locally linear techniques may model the manifold effectively by fitting parts of it separately if they are able to decompose it into linear components. However, local modeling is sensitive to noise and the modeling of noise remains a challenge. Consequently, most locally linear techniques do not address the issue of noise.

Isomap [4] is a popular locally linear technique that works by assuming isometry of geodesic distances in the manifold. The geodesic distance is defined as the distance of the shortest path between two points that passes on the embedded manifold [8]. Isomap estimates geodesic distances by constructing a graph with Euclidean distances between neighboring points as edge weights and computing shortest paths in the graph. Finally, classical MDS is applied to compute an optimal embedding. A computationally efficient implementation that computes shortest paths to only a subset of landmark data points is presented in [9].

A central problem in Isomap and many other locally linear techniques (e.g., [5,6,7,10]) lies in the selection of neighbors that form local patches. The shape of the manifold is in most cases unknown but a common assumption is that in small patches the surface is smooth, and that close neighbors of a data point likely lie on the same part of the manifold and have a similar orientation. Therefore, properties of the locality at each data point are commonly estimated using its nearest neighbors. Two formulations are commonly used: a fixed number of neighbors (k nearest neighbors), or all neighbors within a fixed radius (ϵ hypersphere). The k nearest neighbors version is more common since the sparseness of the resulting structures is guaranteed. For example, the cost matrix used to compute an LLE embedding can have at most $4kN$ nonzero elements. Efficient versions exist of the Dijkstra algorithm (used in Isomap) that take advantage of the sparseness of the input graph. On the other hand, if an ϵ hypersphere is used, it is difficult to predict if a selected radius will include any neighbors at all at every point.

With either formulation, the choice of parameter typically has a dramatic effect on the transformation. If the neighborhoods are too small, disconnected clusters tend to form. Isomap maps the manifold in this case as a set of disjoint components, while LLE applies regularization on the cost matrix, but in both cases the global structure is lost. Since LLE performs a set of local optimizations, it is highly dependent on links created by sufficiently large neighborhoods to discern global structure. On the other hand, setting the neighborhood to a size that is too large creates links to parts of the manifold that are geodesically far. Isomap is especially sensitive to this problem since the shortest paths algorithm will tend to drain multiple paths through such shortcuts, affecting distance estimates globally. However, with small neighborhood sizes, the computed graph geodesic greatly overestimates the true geodesic distances in linear surfaces.

If the dimensionality reduction technique used assumes linear patches, then a good strategy for selecting these parameters needs to consider the (estimated) orientation of the manifold at each point. The selection should be data-driven and depend on such factors as curvature and density. But since curvature and density may vary over the manifold, one global setting may not work well for the entire manifold. In practice, examples such as the popular "Swiss roll" are presented where curvature and density are fairly constant everywhere. The parameters are often configured ad hoc, often by empirically evaluating the embeddings produced with different settings. However, if the Swiss roll is stretched (as in Figure 1), forming areas of varying curvature, then no global setting of k produces satisfying results.

In this paper, we describe a practical strategy for selecting a neighborhood size adaptively that does not require any parameters, based on estimates of intrinsic dimensionality and tangent orientation. We apply our technique to the Isomap algorithm and demonstrate simple manifolds where traditional neighborhood formulations fail, while our technique generates satisfactory mappings. The elimination of the parameter does not reduce the technique's flexibility, since there is no way to configure this parameter automatically without prior knowledge of the desired output.

The rest of the paper is organized as follows: Section 2 reviews the estimation of intrinsic dimensionality and motivates its use in estimating local tangent space orientation. Section 3 discusses our technique for estimating the orientation of local tangent space and compares it against previous work in the field. Section 4 outlines our proposed neighborhood selection technique. In Section 5 we present experimental results on datasets and finally Section 6 concludes with a discussion.

2 Intrinsic Dimensionality

The intrinsic dimensionality of a data set is commonly defined as the smallest number of dimensions that can be used to adequately explain the data. What constitutes an adequate explanation is subjective and depends on the user and the application. Nevertheless, intrinsic dimensionality is key in dimensionality reduction, since knowledge of the intrinsic dimensionality in every part of the manifold eliminates over- or underfitting. For a complete treatment on the subject of dimensionality see [11].

Wang et al. [12] propose an adaptive neighborhood selection heuristic based on estimates of local tangent orientation, and apply it to a variation of LLE they call Local Tangent Space Alignment (LTSA). Their proposed technique assumes fixed intrinsic dimensionality everywhere that is equal to the target dimensionality specified by the transformation, and uses user-specified parameters to threshold the projection of points in the neighborhood onto the complement and tangent spaces. However, in some applications, it is convenient to separate these two variables,

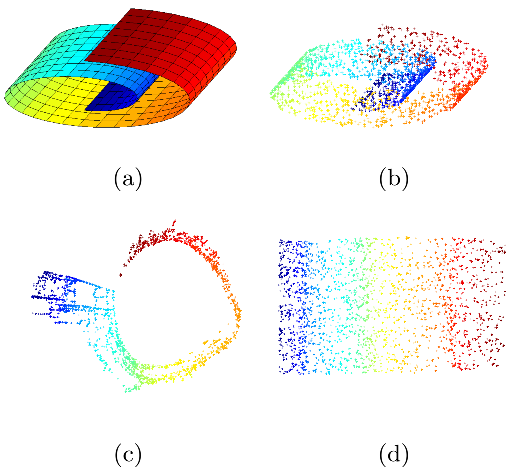


Fig. 1. (a). The classic 'Swiss roll' manifold stretched to aspect ratio 0.4. Unlike the more common circular Swiss roll, curvature varies along the manifold. (b). A uniform-density sample drawn from the manifold, $N = 2000$. (c). The best embedding computed by Isomap. A setting of $k = 4$ was used. Higher values result in more shortcuts created from data points to different parts of the manifold and more convoluted embeddings, while $k < 3$ results in several disconnected components, with no unifying global structure. (d). The projection obtained by Isomap using our adaptive neighborhood technique successfully unrolls the Swiss roll into a flat surface.

e.g., visualization, where target dimensionality is typically limited to 2 or 3. We therefore use target dimensionality in the MDS step to produce the final embedding, but intrinsic dimensionality when estimating the local geometry at a point.

Several techniques have been proposed to estimate intrinsic dimensionality from data in problems where it is unknown. Techniques that apply PCA (globally or locally) and threshold the resulting eigenvalues include [14]. Geometric methods include Costa et al. [15] - an estimator based on the length of minimal spanning trees on graph geodesics. We have found a maximum likelihood estimator recently proposed by Levina and Bickel [16] to work well on our data. The technique assumes constant density in small neighborhoods and approximates the number of samples in hyperspheres of growing radius as a Poisson process. Then the rate of the process $\lambda(t)$ at intrinsic dimensionality m can be expressed as,

$$\lambda(t) = \frac{f(x)\pi^{m/2} m t^{m-1}}{\Gamma(m/2+1)} \tag{1}$$

where $f(x)$ is the sampling density and $\Gamma(\cdot)$ is the Gamma function. A maximum likelihood estimate of the intrinsic dimensionality at point $\boldsymbol{x}_i$ given c neighboring observations is then (see [16] for complete details),

$$\hat{m}_c(\boldsymbol{x}_i) = \left[\frac{1}{c-1}\sum_{j=1}^{c-1} \log \frac{T_c(\boldsymbol{x}_i)}{T_j(\boldsymbol{x}_i)}\right]^{-1} \tag{2}$$

where $T_j(\boldsymbol{x}_i)$ represents the distance from $\boldsymbol{x}_i$ to its j'th nearest observation. The authors propose averaging over all points to obtain the global intrinsic dimensionality. An optimal c can be obtained by minimizing the estimator's standard deviation over different sample sizes drawn from the data. The resulting estimator is asymptotically unbiased (negatively biased otherwise) but enjoys remarkably low variance, making it perhaps possible to apply it semi-locally or in clusters.

3 Estimating the Local Tangent Space

In this section we outline our technique for estimating local tangent space. Using knowledge or an estimate of the intrinsic dimensionality m, we seek to estimate the orientation of the manifold at each data point $\boldsymbol{x}_i$ and compute an orthonormal basis $\boldsymbol{A}_i$ for it. While here we use a global value for m, local values can be used if a reliable estimator exists.

Hyperplanes (if linearity is assumed) through a neighborhood may be fitted by retaining the vectors corresponding to the highest singular values (up to the desired dimensionality) of the singular value decomposition (SVD) of $[\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k}] - \boldsymbol{x}_i \mathbf{1}^T$ where $\boldsymbol{x}_{i_j}$ are the neighbors of $\boldsymbol{x}_i$ and $\mathbf{1} = [1, \ldots, 1]^T$. Previous approaches that explicitly compute local tangent orientation include Medioni et al. [10], which estimates intrinsic dimensionality and orientation simultaneously at each point, by performing an eigen-decomposition at each data point, and retaining the largest eigenvectors up to the largest drop in the eigenvalues. Although a voting scheme is used to improve the estimator's variance, the resulting intrinsic dimensionality estimates are still too noisy for adaptive neighborhood selection. Wang et al. [12] compute a least squares fit about the mean of the neighborhood rather than $\boldsymbol{x}_i$, but the neighborhood size is indirectly controlled by several user-specified parameters.

Our technique also uses SVD to compute tangent orientation, but the neighbors used in the computation are selected based on estimated local sampling density, an approach inspired by [16]. Under ideal sampling conditions, $m + 1$ points define an m-dimensional hyperplane. However, in practice, degenerate configurations are often observed where the points are not in general position, and thus define a rank-deficient space (e.g., collinearity). This is a likely scenario if marginal densities vary along different axes. If the singular values λ_j resulting from the decomposition are ordered in non-increasing order, such that $\lambda_1 \geq \ldots \geq \lambda_m \ldots \geq \lambda_k$, in order to ensure a sound m-dimensional basis, λ_m must be sufficiently high and λ_{m+1} low. Specifically, singular value λ_m must be significant enough so that it represents an observation that lies in a direction orthogonal to the directions represented by singular values $\lambda_1 \ldots \lambda_{m-1}$, rather than leftovers from projections of other neighbors. Therefore, an appropriate threshold for λ_m is the expected radius to a neighboring point at $\boldsymbol{x}_i$, denoted $\tilde{T}_1(\boldsymbol{x}_i)$.

A coarse estimate of this radius may be obtained by taking the distance from $\boldsymbol{x}_i$ to its nearest neighbor. However, this estimate is unreliable as it is based on only one observation. A better strategy is to infer the radius from a further neighbor taking advantage of the robust properties of order statistics. If the volume of an m-dimensional hypersphere of radius r is $\pi^{m/2} r^m [\Gamma(\frac{m}{2}+1)]^{-1}$, then the expected number of observations $N(r)$ in the hypersphere follows,

$$E\big[N(r)\big] \propto r^m \tag{3}$$

and

$$T_{N(r)}(\boldsymbol{x}_i) \approx r \tag{4}$$

combining Eq. (3) and (4), we obtain,

$$\tilde{T}_1(\boldsymbol{x}_i) = (1/k)^{(1/m)} T_k(\boldsymbol{x}_i) = (1/k)^{(1/m)} |\boldsymbol{x}_{i_k} - \boldsymbol{x}_i|_2 \tag{5}$$

We iteratively increase k, generating a new estimate $\tilde{T}_1(\boldsymbol{x}_i)$ of the expected radius at $\boldsymbol{x}_i$ in each iteration according to Eq. (5), until $\lambda_m \geq \tilde{T}_1(\boldsymbol{x}_i)$. Upon termination, basis $\boldsymbol{A}_i$ is constructed from the vectors corresponding to the m highest singular values. This basis defines the estimated tangent space at $\boldsymbol{x}_i$ and is used in the next step of our algorithm to select neighbors that are consistent with it.

4 Selection of Neighbors

Using the basis $\boldsymbol{A}_i$ that defines the estimated tangent orientation at point $\boldsymbol{x}_i$, a neighborhood can be selected that includes nearby points that agree with the computed tangent. Wang et al. [12] examine the ratio of the matrix (Frobenius) norms of the projections of the points under consideration and compare it to a user-specified threshold η. While one global threshold can lead to adaptive neighborhood selection according to the local curvature and density at each point, it is unclear how η should be set. However, the optimal η probably depends on the intrinsic dimensionality since it determines the dimensionality of the projections (and hence their norms). Another anomaly is that the lowest possible ratio of norms (zero) is realized when exactly m neighbors are considered. To overcome this hitch, the neighborhood is initialized to a 'sufficiently large' K (another user-specified parameter) and iteratively shrunk until the specified η ratio is reached. If this step fails, a neighborhood is selected such that the ratio is minimized. Then an expansion step is performed where points that were discarded in the previous step are added back as neighbors if their projection norm ratio satisfies the user-specified threshold η, while 'skipping' nearer neighbors whose ratios do not. This is perhaps done to accommodate noise, but as Figure 2(a) demonstrates, it is a dangerous strategy in techniques like Isomap, since it can potentially result in invalid shortcuts to different parts of the manifold and these are likely to have adverse global effects. Another pitfall is that 'steps' in the manifold may be smoothed out, as depicted in Figure 2(b). If the projection ratio at the initial K falls below the user-specified η, the algorithm is trapped in a local minimum and fails to uncover the correct orientation.

In contrast, our strategy for selecting neighbors at a point $\boldsymbol{x}_i$ is a direct extension of our approach for estimating the tangent space, outlined in the previous

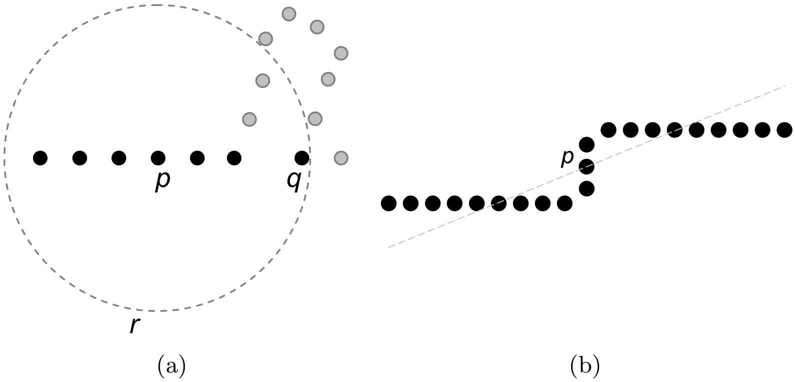


Fig. 2. (a). The neighbors (depicted as dark points) of point p defined according to [12]. Point q is considered a neighbor while closer points are excluded, resulting in an invalid shortcut from p to q. (b). Estimation of the tangent space at point p using [12]. Setting the initial K to a high value such that all the points in the figure are included results in a fit that appears good (high tangent space projection norm relative to complement space projection), but an incorrect tangent. The neighborhood contraction step in [12] is trapped in a local minimum and fails to uncover the correct vertical tangent.

section. We incrementally grow the neighborhood of $\boldsymbol{x}_i$ one point at a time, monitoring each new point's projection onto the complement space at $\boldsymbol{x}_i$, and testing the resulting norm against our estimate of radius to nearest neighbor $\tilde{T}_1(\boldsymbol{x}_i)$. New neighbors $\boldsymbol{x}_{i_j}$ are added iteratively until

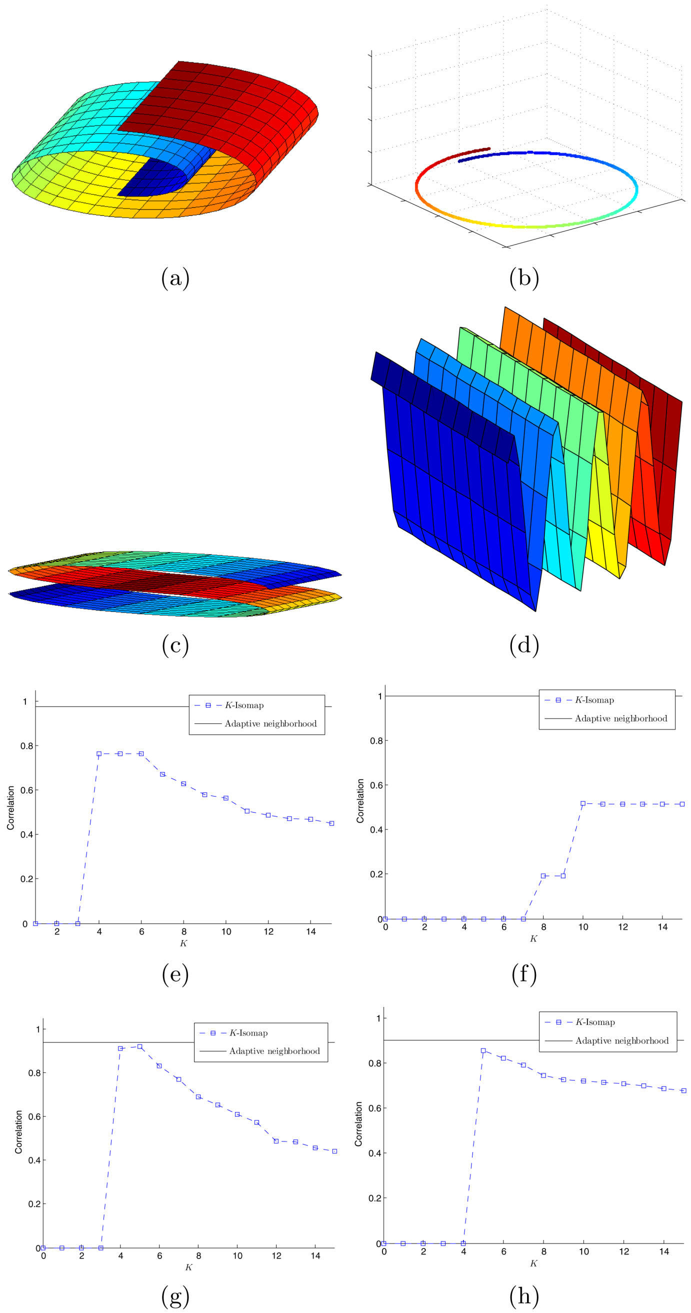
$$|(\boldsymbol{I} - \boldsymbol{A}_i\boldsymbol{A}_i^T)(\boldsymbol{x}_{i_j} - \boldsymbol{x}_i)|_2 < \tilde{T}_1(\boldsymbol{x}_i) \tag{6}$$

or equivalently,

$$\sqrt{|\boldsymbol{x}_{i_j} - \boldsymbol{x}_i|_2^2 - |\boldsymbol{A}_i(\boldsymbol{x}_{i_j} - \boldsymbol{x}_i)|_2^2} < \tilde{T}_1(\boldsymbol{x}_i) \tag{7}$$

is violated. To avoid improper shortcutting as illustrated in Figure 2(a), the iteration terminates when a neighboring point breaks the above condition. This process can be viewed as the inclusion of points within a hypercylinder of radius $\tilde{T}_1(\boldsymbol{x}_i)$ about the estimated tangent space defined by the basis $\boldsymbol{A}_i$. The estimate $\tilde{T}_1(\boldsymbol{x}_i)$ may be further refined at each iteration as new neighbors are added, according to the criterion in Eq. (5).

In contrast to [12], our technique may add an unlimited number of neighbors as long as the linear tangent space assumption is upheld. In linear sections of the manifold, all points are added as neighbors. This is a desirable property for Isomap, since in planar regions, geodesic distances are now correctly estimated as Euclidean distances (whereas normally the graph geodesic significantly overestimates distances). In fact, if the entire input manifold is linear, all geodesic distances are estimated as Euclidean distances, and Isomap degenerates into PCA.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 3. (a)-(d). Synthetic 2-D and 1-D manifolds embedded in 3-D. (e)-(h). Corresponding plots of the correlation between true geodesic distances and Isomap estimates using k-Isomap *(dashed line)* starting with the lowest value of k that yields a global mapping and our adaptive technique *(solid line)*.

5 Experimental Results

We have tested our technique with the Isomap algorithm on several datasets. Our algorithm's performance on manifolds with relatively constant curvature and uniform sampling (the Swiss roll and S-curve) matched that of Isomap with manually-selected optimal values of k. Figure 3 depicts the correlations between true and estimated geodesic distances on the stretched Swill roll and a 3-D spiral for different values of k and using our adaptive technique. For these structures, Isomap failed to compute satisfactory embeddings with any settings of the algorithm's parameters. On the other hand, our technique adaptively selected neighbors at each point resulting in superior interpolation in relatively flat surfaces while avoiding invalid shortcuts between different parts of the manifold. Here we report correlation to true distances as an objective qualitative measure. However, qualitatively, even small differences in correlation values translate into dramatic effects in terms of the resulting embedding. For example, the embedding produced by k-Isomap for the stretched Swiss roll in Figure 3(a) ($k = 4$, correlation=0.78) can be seen in Figure 1(c). As a sanity check, we also ran our algorithm on the Isomap face database (698 images of synthetic faces under varying illumination and pose). Our adaptive technique appears to produce a satisfactory embedding, but since the true manifold is unknown, quantitative analysis is not possible.

6 Summary

We have presented a parameterless adaptive technique for selecting a neighborhood at each point in nonlinear manifold learning, in particular Isomap. To date, nonlinear manifold learning techniques have relied on user-specified parameters that cannot be set in a principled way. Additionally, the use of one global setting results in suboptimal learning if curvature or density vary. Our technique eliminates the guesswork associated with tuning these parameters and enables modeling of manifolds that cannot be modeled effectively with one global setting. In addition to eliminating user-input parameters, our technique offers several advantages over previous work on adaptive selection.

We have demonstrated the effectiveness of the technique on several simulated and real datasets. The technique produces good results on our data and we are currently investigating several possible extensions. In its present form, the technique assumes that observations are sampled directly from the manifold with no noise. We are currently looking at ways to incorporate a noise model, as well as robust voting schemes to improve tangent space and neighborhood estimation at each point.

Acknowledgments. The authors thank Konstantinos Derpanis and Erich Leung for reviewing this manuscript and providing useful comments.

References

1. Jolliffe, I.T.: Principal Component Analysis. Springer-Verlag, New York (1986)
2. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley and Sons, Inc., New York (2000)
3. Schölkopf, B., Smola, A., Mller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation **10**(5) (1998) 1299–1319
4. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500) (2000) 2319–2323
5. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500) (2000) 2323–2326
6. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems. Number 14, Cambridge, MA, MIT Press (2002)
7. Donoho, D.L., Grimes, C.E.: Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Arts and Sciences **100** (2003) 5591–5596
8. Abraham, R., Marsden, J.E.: Foundations of Mechanics. Second edn. Addison-Wesley (1978)
9. de Silva, V., Tenenbaum, J.B.: Global versus local methods in nonlinear dimensionality reduction. In S. Becker, S.T., Obermayer, K., eds.: Advances in Neural Information Processing Systems 15. (MIT Press)
10. Medioni, G., Lee, M.S., Tang, C.K.: Computational Framework for Segmentation and Grouping. Elsevier Science Inc., New York, NY, USA (2000)
11. Falconer, K.: Fractal Geometry: Mathematical Foundations and Applications. John Wiley & Sons (1990)
12. Wang, J., Zhang, Z., Zha, H.: Adaptive manifold learning. In: NIPS. (2004)
13. Lorenz, E.: Deterministic nonperiodic flow. **20** (1963) 130–141
14. Fukunaga, K., Olsen, D.: An algorithm for finding intrinsic dimensionality of data. IEEE Transactions on Computer **20**(2) (1971) 176–183
15. Costa, J., Girotra, A., Hero, A.O.: Estimating local intrinsic dimension with k-nearest neighbor graphs. In: IEEE Workshop on Statistical Signal Processing (SSP), Bordeaux (2005)
16. Levina, E., Bickel, P.J.: Maximum likelihood estimation of intrinsic dimension. In Saul, L.K., Weiss, Y., Bottou, l., eds.: Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA (2005) 777–784