

SOME RECENT RESULTS ON THE PERFORMANCE AND IMPLEMENTATION OF MANIFOLD LEARNING ALGORITHMS

Xiaoming Huo

Georgia Institute of Technology, School of Industrial and Systems Engineering
765 Ferst Dr., Atlanta, GA 30332-0205

ABSTRACT

Manifold learning is becoming an important research topic in both statistics and machine learning. Many works have appeared. A series of newly proposed algorithms have been proven to be effective in a wide range of applications. In many cases, the theoretical properties are not completely known. It is interesting to see that many algorithms depends on specific linear invariant subspaces. We recently derived performance bounds on these algorithms [1, 2]. Moreover, detailed analysis on these algorithms, especially the considerations of locally utilized dimensions and the property of the associated algorithms, demonstrates that there are common principles in designing an efficient manifold learning algorithm. We reveal these principles and utilize them to analyze existing manifold learning algorithms. Our products include interpretation of some reported numerical examples, as well as predicted advantages and disadvantages of existing methods.

This paper is a condensed presentation of some results from recent works [3, 1, 2]. A purpose of making this documentation is to satisfy the requirement of an upcoming workshop, which will be held right before the Annual Meeting of INFORMS in 2006 at Pittsburgh PA, USA. See "<http://dm.section.informs.org/WorkshopCFP.doc>" for workshop details.

Many results reported here are joint works with Andrew K. Smith, a co-author of [1, 2].

1. INTRODUCTION

Manifold-based learning is an emerging and promising approach in nonparametric dimension reduction. In [3], authors review the state-of-the-art mathematical developments, as well as some interesting applications. The following provides a brief summary.

A manifold is a topological space that is locally Euclidean (i.e., around every point, there is a neighborhood that is topologically the same as the open unit ball in the Euclidean space). A good example of a manifold is the Earth. Locally, at each point on the surface of the Earth, we have a 3-D coordinate system: two for location and the last one for

the altitude. Globally, it is a 2-D sphere embedded in a 3-D space.

Manifolds offer a powerful framework of dimension reduction. The key idea of dimension reduction is to find the most succinct low dimensional structure that is embedded in a higher dimensional space. Historically, Occam's razor has been used to justify dimension reduction. The key idea of Occam's razor is to choose the simplest model from a set of equivalent models to explain a given phenomenon. It is easy to see that a manifold structure gives a dimension reduction. Moreover, if the data are indeed generated according to a manifold, then a manifold-based learning is, in some sense, optimal.

A detailed description on some manifold-based dimension reduction algorithms is given in [3]. The next section (Section 2) offers a quick summary. The main idea is to give an overview. Recent works [1, 2] have established some theoretical performance bounds on manifold-learning algorithms. This paper should be considered an advertisement for the contents of [1, 2].

2. A SURVEY

In [3], five categories of methods/algorithms are presented:

1. Classical methods, including principal component analysis (PCA). Other methods are related, such as factor analysis and techniques in multivariate analysis.

Principal Component Analysis (PCA) is one of the most classical methods in dimensional reduction. PCA is also known as the Karhunen-Loève transform, or singular value decomposition (SVD). The key idea of PCA is to find the low-dimensional linear subspace which captures the maximum proportion of the variation within the data. A key assumption of PCA is that the underlying structure is a linear subspace.

2. Semi-classical methods, including multidimensional scaling (MDS) [4, 5].

MDS is the name of a group of methods that have found a wide range of applications. The key idea is

to find a mapping from a high-dimensional space to a low-dimensional space, such that the pairwise distances between the observed points are preserved the best. An intuitive example is to recover the relative positions of cities from the inter-city distances. Imagine that the exact locations (coordinates) of N cities are lost. However, we have the driving distances between pairs of them. These distances form an matrix. Based on this matrix, MDS can recover a 2-D coordinate system that includes the locations of these cities, subject to a rigid motion (a combination of rotation, shifting, and reflection), such that the distances among the points on this 2-D plane are close to the driving distances among those cities.

The above in fact gives an example of metric MDS [6, 7], which is relative to nonmetric MDS [4, 8] that are reviewed in [3].

3. Manifold searching methods, including generative topographic mapping (GTM) [9], local linear embedding (LLE) [10], and ISOMAP [11]. These methods inspired a lot of contemporary works. New and better methods have been introduced, as described in the following. Due to their historical contribution, these papers are highly cited.
4. Methods rooted in continuum spectral theory, including the Laplacian eigenmaps [12] and Hessian eigenmaps [13], which are based on elegant theory in spectral analysis, and then discretize the results in the continuum to generate numerical approaches. We will analyze the performance of hessian eigenmaps later. These methods overcome some shortcomings of earlier methods (e.g., LLE and ISOMAP). For example, Hessian eigenmaps do not require the domain of the manifold to be convex.
5. Advanced manifold methods, including charting [14] and local tangent space alignment (LTSA) [15]. These methods are based on global alignment. The key insight in these methods is the realization that the global alignment can be achieved via an eigenvalue computation. In our numerical experiments, most of the time, LTSA is the most desirable.

All the above methods have the spirit of finding the embedded geometric structure, i.e., a manifold. Different methods are based on different ideas. In order to determine their performance, it may seem like methods should be analyzed separately. As a matter of fact, many of them eventually become null-space searching algorithms. (Recall that null-spaces are spanned by the solutions of a system of linear equations corresponding to a predetermined matrix.) Hence, if we can characterize the behavior of null-spaces under uncertainty, we can provide a unified analysis of some of the

mentioned methods. In [3], it is reviewed that LLE, Hessian eigenmaps, and LTSA are null space-based methods. The results that will be mentioned in the next two sections rely on analysis of null-spaces too.

3. ANALYSIS OF LTSA

In this section, we review the main result in [1]. As mentioned before, manifold learning (ML) algorithms are novel and model-free dimension reduction (DR) approaches. Researchers in *manifold learning* have invented many efficient algorithms. We would like to emphasize the differences of these algorithms from traditional statistical DR methods:

1. There is *no* parametric model assumed for the observed data; although we assume a mapping between the observations and a set of intrinsic (low-dimensional) vectors.
2. The sampling density is high enough to ensure the recovery of the underlying structure on its support.

We adopt a statistical model: $y_i = f(x_i) + \varepsilon_i, i = 1, 2, \dots, n$, where y_i denotes a high-dimensional noisy observation, $f(\cdot)$ is a mapping satisfying some local regularity conditions, x_i is a low-dimensional intrinsic parameter, ε_i is a random error, and n is the sample size. The objective of DR is to find the set $\{x_i\}$, without any parametric model assumption on f except local smoothness. Our performance analysis will be different from manifold learning, which in general considers *noiseless* observations: $y_i = f(x_i), i = 1, 2, \dots, n$; i.e., ignoring additive random errors. In numerical simulations, all of these algorithms are observed to be robust against errors.

The main contribution of [1] is to establish a performance property of a manifold learning algorithm under the presence of errors. The key idea in our analysis is to treat the solutions of manifold learning algorithms as invariant subspaces, and then carry out a matrix perturbation analysis. It has been reported by many (e.g., [10, 12, 13, 14, 15]) that solutions of their manifold learning algorithms correspond to invariant subspaces which are spanned by the eigenvectors associated with the 2nd through $(d+1)$ st smallest eigenvalues of certain matrices. The form of such a matrix depends on the details of the algorithm. These subspaces are clearly invariant, because they are spanned by eigenvectors [16, Section I.3.4].

LTSA is chosen because it is representative. First of all, in numerical simulation (e.g., using the tools offered by [17]), we find empirically that LTSA performs among the best of the available algorithms. Second, the solution to each step of the LTSA algorithm is an invariant subspace, which allows us to analyze its performance by applying matrix perturbation theory. Third, the similarity between many

manifold learning algorithms (e.g., LLE, Laplacian eigenmaps and Hessian eigenmaps) that their solutions can all be interpreted as invariant subspaces indicates that results for LTSA can be generalized to other algorithms.

The theoretical result in [1] gives a worst case bound on the performance of LTSA. To be more specific, let $x_i, i = 1, 2, \dots, n$, denote a set of low-dimensional vectors. For reasons which will become evident later, we call this set the *true parametrization*. Let $y_i, i = 1, 2, \dots, n$, denote the observed high-dimensional vectors that are generated according to $y_i = f(x_i) + \varepsilon_i$, and assume that f is locally regular. Let $\{\tilde{x}_i, 1 \leq i \leq n\}$ denote the estimated parameter set. Let $\mathcal{R}(\tilde{X})$ (resp., $\mathcal{R}(X)$) denote the invariant subspace that is associated with the set $\{\tilde{x}_i, 1 \leq i \leq n\}$ (resp., $\{x_i, 1 \leq i \leq n\}$). (Details regarding invariant subspaces can be found in [1].) We prove the following regarding the distance between the two invariant subspaces, which consequently gives the worst case analysis on the performance of LTSA:

$$\begin{aligned} & \|\tan(\mathcal{R}(\tilde{X}), \mathcal{R}(X))\|_2 \\ & \leq 4 \cdot \frac{C_3(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau) \cdot \|\sum_{i=1}^n S_i\|_\infty}{\ell_{\min}}, \end{aligned}$$

where C_3 is a constant that depends on the dimension of the observations, the regularity of the function f and the value of an algorithmic parameter that is used in LTSA, σ is an upper bound on the absolute values of the random errors, τ denotes the size of the neighborhoods, within which f is assumed to be well-behaved, $\|\sum_{i=1}^n S_i\|_\infty$ is equal to the maximum number of times that a single observation appears in nearest-neighbor sets, and ℓ_{\min} is a constant determined by the global structure of the mapping f . The above inequality is established under the conditions that $\tau \rightarrow 0$, $\frac{\sigma}{\tau} \rightarrow 0$, and $\ell_{\min} \rightarrow 0$ at a rate slower than the first two so that the right hand side of the inequality goes to 0; more specifically,

$$\frac{C_3(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau) \cdot \|\sum_{i=1}^n S_i\|_\infty}{\ell_{\min}} \rightarrow 0.$$

Performance analysis will provide theoretical foundation for the application of ML algorithms. To the best of our knowledge, [1] is the first attempt of this kind.

The aforementioned result differs from most existing classical DR methodologies. Due to the amount of literature in DR and the space limitation, we just discuss two of the most popular branches in DR, while mentioning that there are many more.

- Principal component analysis (PCA) and multidimensional scaling (MDS), together with many extensions, are widely known in statistics. In contrast to PCA, a manifold learning algorithm does *not* require the underlying structure to be a linear subspace. Unlike

MDS, a manifold learning algorithm does not impose the same pairwise distances in the data (or observation) space. For example, MDS will fail in the pedagogical numerical example that is presented in [1]. That example gives a case in which it is not necessary to keep the pairwise distances between observations.

- A recent branch of research in DR involves the idea of a *central subspace*. The most recent work that we are aware of in this area is Cook and Ni [18]. A central subspace is defined only when a response is present - hence it is a supervised learning problem, compared to the unsupervised learning problem that is discussed here. A central subspace is still a globally linear subspace, however, while manifold learning makes no such assumption.

There are still open questions to be addressed (Section 3.1). In addition to a discussion on the relation of LTSA to existing DR methodologies, we will also address relation with known results as well (Section 3.2). The following is quoted from [1]. We refer to the original paper for more detailed information.

3.1. Open Questions

The rate of convergence of ℓ_{\min} is determined by the topological structure of f . It is important to estimate this rate of convergence, but this issue has not been addressed here.

We assume that $\tau \rightarrow 0$. One can imagine that it is true when the error bound (σ) goes to 0 and when the x_i 's are sampled with a sufficient density in the support of f . An open problem is how to derive the rate of convergence of $\tau \rightarrow 0$ as a function of the topology of f and the sampling scheme. After doing so, we may be able to decide where our theorem is applicable.

We assume that the covering P_i is given, such that $\tau \rightarrow 0$ holds. Given a covering scheme, such as choosing the k -nearest neighbors, a verification of $\tau \rightarrow 0$ and a derivation of its corresponding rate is an open question too. The answer to this will depend on the topology of f , which is not covered in this paper, and the sampling scheme.

3.2. Relation with Existing Works

The error analysis in the original paper about LTSA is the closest to our result. However, Zhang and Zha [15] do not interpret their solutions as invariant subspaces, and hence their analysis does not yield a worst case bound as we have derived here.

Reviewing the original papers on LLE [10], Laplacian eigenmaps [12], and Hessian eigenmaps [13] reveals that their solutions are subspaces spanned by a specific set of eigenvectors. This naturally suggests that results analogous

to ours may be derivable as well for these algorithms. A recent book chapter [3] stresses this point. After deriving corresponding upper bounds, we can establish different proofs of consistency than those presented in these papers.

ISOMAP, another popular manifold learning algorithm, is an exception. Its solution cannot immediately be rendered as an invariant subspace. However, ISOMAP calls for MDS, which can be associated with an invariant subspace; one may derive an analytical result through this route.

The DR problem considered here is an unsupervised learning problem. There are supervised learning problems that are of similar flavor, e.g., contour regression and inverse regression [19]. As mentioned before, the rich literature in the supervised counterpart (e.g., the concepts of *exhaustiveness*, *sufficiency*, *central subspaces*, etc) gives motivation to derive corresponding results in the supervised framework.

4. PERFORMANCE ANALYSIS ON HESSIAN LLE

The following content is mainly from [2].

In [13], the authors present a new nonlinear dimensionality reduction algorithm — Hessian locally linear embedding (HLLE). Along with it, they present an intriguing Theorem which, intuitively, suggests that their algorithm is consistent — that is, with a sufficiently large sample, the algorithm can recover the underlying parameters up to an isometry. However, the Theorem does not actually establish this property. It is a statement about a functional in the continuum, which involves unknown quantities, while the algorithm forms a discrete estimate of this functional based on the sample data points. Thus, in order to establish rigorously the consistency of this method, several issues of convergence need to be investigated. In [2], it is shown that the estimated quantities used in the algorithm converge to their counterparts in the continuous manifold.

Paper [2] makes several contributions. First, the results give new understanding of the asymptotic properties of the HLLE algorithm. To our knowledge, this is the first time that the consistency of the algorithm has been proven. The proof also yields insight into the factors that affect the performance of the algorithm, and the implications of various geometric properties of the underlying manifold on the ability of HLLE and similar algorithms to recover the manifold structure. Second, the authors propose a modified estimator of the Hessian matrix and demonstrate that it results in a small improvement in performance in terms of Procrustes error [20]. If this small improvement in performance is viewed as significant, then obviously the contribution is important. If the improvement is judged to be insignificant, then they have provided stronger theoretical support for the existing methodology — i.e., in this case, HLLE performs almost as well as a modified estimator, which has many optimality properties due to the fact that it is a least-squares

estimate. Finally, authors provide new insight into the connections between HLLE and LTSA [15].

5. COMPARISON OF LTSA AND HLLE

The following is from [2] and is mainly due to the first author (Andrew K. Smith) of that paper.

Since we have now established the consistency of both LTSA and HLLE as long as certain regularity conditions are imposed on the underlying manifold, it seems natural to wonder if the two algorithms are, in some sense, equivalent. In this Section, we investigate this question and show that there is, in fact, a strong similarity between the two, despite the considerable differences in their actual implementations.

Due to space, we omit a review of the notations and assume that readers are familiar with the notations that are used in [2]. Consider again the Taylor expansion of a C^2 function $f : \mathcal{M} \rightarrow \mathbb{R}$ at a sample point:

$$\begin{aligned} f(y_{i_j}) &= f(\bar{y}_i) + J_f(\bar{y}_i)(y_{i_j} - \bar{y}_i) \\ &\quad + \frac{1}{2}(y_{i_j} - \bar{y}_i)^T H_f(\bar{y}_i)(y_{i_j} - \bar{y}_i) \\ &\quad + O(\|y_{i_j} - \bar{y}_i\|^3) \end{aligned}$$

Now, recall that LTSA, in its second step, finds the null space of the matrix (see [15] and [1] for details)

$$\widehat{S}\widehat{P}_k \begin{pmatrix} I - \widehat{U}_1\widehat{U}_1^T & & & 0 \\ & I - \widehat{U}_2\widehat{U}_2^T & & \\ & 0 & \ddots & \\ & & & I - \widehat{U}_N\widehat{U}_N^T \end{pmatrix} \widehat{P}_k \widehat{S}^T$$

while HLLE finds the null space of

$$\widehat{S} \begin{pmatrix} \widehat{Q}_1\widehat{Q}_1^T & & & 0 \\ & \widehat{Q}_2\widehat{Q}_2^T & & \\ & 0 & \ddots & \\ & & & \widehat{Q}_N\widehat{Q}_N^T \end{pmatrix} \widehat{S}^T.$$

The connection between the two may be viewed as follows: First, recall the Theorem proven in [13]: The functional

$$\mathcal{H}(f) = \int_{\mathcal{M}} \|H_f(m)\|_F^2 dm \quad (1)$$

has a $d+1$ -dimensional nullspace, consisting of the constant function and the d isometric coordinate functions. LTSA and HLLE are both ways of finding functions which are well-approximated by their estimated derivatives given by the optimal d -dimensional approximation over each neighborhood — the difference is only in the way they define “well-approximated.”

In view of assumptions in [2], we may assume that

$$U_i U_i^T \approx J_g(\bar{y}_i) J_g^T(\bar{y}_i) \quad (2)$$

by noting that these two matrices are projections, and therefore are functions only of the column space of the two matrices, and are invariant with respect to the bases chosen. To see this, suppose that the columns of two matrices, say A and $B \in \mathbb{R}^{m \times n}$, form orthonormal bases of the same subspace. Then we have $A = BV$ for some orthogonal $V \in \mathbb{R}^{n \times n}$. Then, by definition, the projection onto the column space of A is given by $AA^T = BV(BV)^T = BVV^TB^T = BB^T$, which is the projector onto the column space of B .

Suppose that $f_k : \mathcal{M} \rightarrow \mathbb{R}, k = 1, 2, \dots, d$ are the global coordinate functions of the data points, and let $F = (f_1, f_2, \dots, f_d)^T$. Then, because the f_k are functions of θ , the underlying parameters, it is easy to see that $\mathcal{R}(J_F^T(\bar{y}_i)) \subset \mathcal{R}(J_g(\bar{y}_i))$. On the other hand, $J_F^T(\bar{y}_i)$ clearly has rank d , as does $J_g(\bar{y}_i)$. Thus, $\mathcal{R}(J_F^T(\bar{y}_i)) = \mathcal{R}(J_g(\bar{y}_i))$. Further, F must be a locally *linear* function of θ . Therefore, we have

$$\begin{aligned} & \|(I - J_F^T(\bar{y}_i)J_F(\bar{y}_i))g(F(Y_i))\| \\ &= \|(I - J_g(\bar{y}_i)J_g^T(\bar{y}_i))g(F(Y_i))\| \\ &\approx \|(I - \widehat{U}_i\widehat{U}_i^T)g(F(Y_i))\| \\ &= O(\tau^2) \end{aligned} \quad (3)$$

where $f(Y_i) \stackrel{\text{def}}{=} (f(y_{i_1}), f(y_{i_2}), \dots, f(y_{i_n}))^T$. Thus, $(I - \widehat{U}_i\widehat{U}_i^T)f(Y_i)$ may be viewed as the approximate error of the first-order Taylor expansion of f , using \widehat{U}_i as an approximation to $J_g^T(\bar{y}_i)$.

Meanwhile, HLLC minimizes $\widehat{Q}_i\widehat{Q}_i^T$, which, as we have seen, is a (somewhat crude) estimate of $\|H_f(\bar{y}_i)\|_F^2$. To see the connection with (1), notice that, if we set $\tau_{min} \stackrel{\text{def}}{=} \min_{i,j} \{\|y_{i_j} - \bar{y}_i\|\}$, we have

$$\begin{aligned} \tau_{min}^2 \|H_f(\bar{y}_i)\| &\leq \|(y_{i_j} - \bar{y}_i)^T H_f(\bar{y}_i)(y_{i_j} - \bar{y}_i)\|_F \\ &\leq \tau^2 \|H_f(\bar{y}_i)\|. \end{aligned}$$

So HLLC seeks functions that minimize the second term in (1). In this sense, LTSA may be interpreted as seeking d orthogonal scalar functions which minimize

$$\sum_{l=1}^d \sum_{i=1}^N \sum_{j=1}^k \|f_l(y_{i_j}) - (f_l(\bar{y}_i) + J_{f_l}(\bar{y}_i)(y_{i_j} - \bar{y}_i))\|_2^2$$

while HLLC seeks d orthogonal scalar functions which minimize

$$\sum_{l=1}^d \sum_{i=1}^N \sum_{j=1}^k \|(y_{i_j} - \bar{y}_i)^T H_{f_l}(\bar{y}_i)(y_{i_j} - \bar{y}_i)\|_F^2.$$

(See [15] for more details on the interpretation of LTSA as an optimization problem.) Essentially, then, recalling (1) and (4), we see that the two are just different ways of

exploiting a Taylor expansion by assuming that the observations are smooth functions of the underlying parameters. The difference is simply that LTSA seeks functions for which the first-order Taylor approximation is most accurate, while HLLC seeks to minimize the second term in the Taylor expansion. Asymptotically, of course, these are equivalent since the second term dominates the remainder as $\tau \rightarrow 0$. We expect, therefore, that the difference between the results of the two algorithms, after allowing for a possible rotation and reflection, is of $O(\tau^3)$. Notice that this is consistent with some experiments in [2], in which we observe that the Procrustes errors for both LTSA and HLLC converge to zero, and seem to do so at nearly the same rate, since the error curves coincide almost exactly for larger values of N . However, the above explanation leads us to expect that the two might differ more substantially if the underlying manifold has large third- and higher-order derivatives at least at some points.

What, then, should we make of the differences between the two algorithms? From a computational perspective, LTSA is the clear winner. It only requires the computation of the pseudo-inverse of the left-singular vectors of each local singular value decomposition, and leads to a sparse eigenvalue problem. HLLC, on the other hand, requires the comparatively difficult computation of both the second-order matrix of cross products and its QR -factorization at every neighborhood. In practice, LTSA is far faster (in our particular simulations, about an order of magnitude faster). From a purely statistical perspective, however, there is no clear winner. The importance of the higher-order terms in the Taylor expansion seems to be specific to each particular application. If we have reason to suspect that higher-order derivatives may be large, then LTSA may offer a significant improvement. However, if we anticipate that the data may be explained by a simple curve (in particular, if we suspect *a priori* that the underlying manifold may be represented as a function with only first- and second-order terms in the parameters), then LTSA may be more sensitive to noise in the data, while HLLC would be relatively more stable. We suspect, therefore, that neither algorithm will strictly dominate the other in terms of performance — the choice of which algorithm is preferable will depend on the particular problem under consideration.

6. FINAL COMMENT

Readers are encouraged to read [1, 2] for more specific information.

7. REFERENCES

- [1] X. Huo and A. K. Smith, “Performance analysis of a manifold learning algorithm in di-

- mension reduction,” Tech. Rep., Georgia Institute of Technology, Atlanta, GA, March 2006, <http://www2.isye.gatech.edu/statistics/papers/06-06.pdf>.
- [2] A. K. Smith and X. Huo, “Perturbation analysis of a manifold-learning algorithm – hessian locally linear embedding (hllc),” Tech. Rep., Georgia Institute of Technology, Atlanta, GA, September 2006.
- [3] X. Huo, X. Ni, and A. K. Smith, “A survey of manifold-based learning methods,” in *Mining of Enterprise Data*, T. W. Liao and E. Triantaphyllou, Eds. Springer, New York, 2005, Submitted. Also available at <http://www2.isye.gatech.edu/statistics/papers/06-10.pdf>.
- [4] J. B. Kruskal, “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis,” *Psychometrika*, vol. 29, pp. 1–27, 1964.
- [5] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer-Verlag, New York, 1997.
- [6] W. S. Torgerson, “Multidimensional scaling: I. theory and method,” *Psychometrika*, vol. 17, pp. 401–419, 1952.
- [7] G. Young and A. S. Householder, “Discussion of a set of points in terms of their mutual distances,” *Psychometrika*, vol. 3, pp. 19–22, 1938.
- [8] R. N. Shepard, “The analysis of proximities: multidimensional scaling with an unknown distance function: I & II,” *Psychometrika*, vol. 27, pp. 125–140 & 219–246, 1962.
- [9] C. M. Bishop, M. Svensen, and C. K. I. Williams, “GTM: The generative topographic mapping,” *Neural Computation*, vol. 10, no. 1, pp. 215–234, 1998.
- [10] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, pp. 2323–2326, 2000.
- [11] J. B. Tenenbaum, V. de Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, pp. 2319–2323, 2000.
- [12] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.
- [13] D. L. Donoho and C. E. Grimes, “Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data,” *Proceedings of the National Academy of Arts and Sciences*, vol. 100, pp. 5591–5596, 2003.
- [14] M. Brand, “Charting a manifold,” in *Neural Information Processing Systems*. Mitsubishi Electric Research Labs, March 2003, vol. 15, MIT Press.
- [15] Z. Zhang and H. Zha, “Principal manifolds and nonlinear dimension reduction via local tangent space alignment,” *SIAM Journal of Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.
- [16] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, MA, 1990.
- [17] T. Wittman, “MANifold learning Matlab demo,” <http://www.math.umn.edu/~wittman/mani/index.html>, April 2005.
- [18] R. D. Cook and L. Ni, “Sufficient dimension reduction via inverse regression: a minimum discrepancy approach,” *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 410–428, June 2005.
- [19] B. Li, H. Zha, and F. Chiaromonte, “Contour regression: a general approach to dimension reduction,” *Annals of Statistics*, vol. 33, no. 4, pp. 1580–1616, August 2005.
- [20] R. Sibson, “Studies in the robustness of multidimensional scaling: procrustes statistics,” *J. Roy. Statist. Soc. Ser. B*, vol. 40, no. 2, pp. 234–238, 1978.