Serdica
Mathematical Journal
Сердика
Математическо списание

For further information on
Serdica Mathematical Journal
which is the new series of
Serdica Bulgaricae Mathematicae Publicationes
visit the website of the journal http://www.math.bas.bg/~serdica
or contact: Editorial Office
Serdica Mathematical Journal
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: serdica@math.bas.bg

Serdica Math. J. **35** (2009), 109–116

Serdica
Mathematical Journal

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

DENOISING MANIFOLDS FOR DIMENSION REDUCTION

Arvind K. Jammalamadaka

Communicated by S. T. Rachev

Abstract. Locally Linear Embedding (LLE) has gained prominence as a tool in unsupervised non-linear dimensional reduction. While the algorithm aims to preserve certain proximity relations between the observed points, this may not always be desirable if the shape in higher dimensions that we are trying to capture is observed with noise. This note suggests that a desirable first step is to remove or at least reduce the noise in the observations before applying the LLE algorithm. While careful denoising involves knowledge of (i) the level of noise (ii) the local sampling density and (iii) the local curvature at the point in question, in most practical situations such information is not easily available. Under the model we discuss, a simple averaging of the neighboring points does reduce the noise and is easy to implement. We consider the Swiss roll example to illustrate how well this procedure works. Finally we apply these ideas on biological data and perform clustering after such a 2-step procedure of denoising and dimension reduction.

2000 *Mathematics Subject Classification:* 68T01, 62H30, 32C09.

Key words: Nonlinear dimension reduction, locally linear embedding, noise reduction, smoothing, nearest neighbors, clustering.

1. Introduction. There is considerable literature on trying to find low dimensional structure in high-dimensional data. This includes such classical linear approaches as principal components analysis, independent components analysis, singular value decomposition, etc. (see for instance [3]), and nonlinear methods which include locally linear embedding [6], Laplacian eigenmaps [1] and Isomap [9]. A recent paper by Sha and Saul [7] addresses the question of preserving local features such as the angles in the embedding created by LLE. Such an exercise may not be meaningful if the observed data comes with errors, because the angles then may be more an artifact of the errors in the observations, rather than the intrinsic shape of the high-dimensional data. One obvious remedy might be to first smooth such data in order to reduce the noise, which is what we propose here. Whenever the data is noisy as in the model discussed here, the use of LLE or indeed any other dimension-reduction technique should be preceded by an initial smoothing step as a way to reduce the noise. What we propose for smoothing, is to replace any observed point by the average of its k nearest neighbors. In Section 3 we study the effectiveness of this procedure, using the Swiss roll example and consider the choice of k. In the final section, we consider yeast cell-cycle data due to Spellman et al [8] as containing a certain lower-dimensional manifold with errors around it, and we see how the smoothing affects the Kullback-Liebler divergence between 5 biologically classified groups in this data in the reduced LLE space.

2. A model with errors. Consider the problem of nonlinear dimension reduction where we have a high-dimensional data set, $X = [x_1, \ldots, x_N]$, $x_i \in R^m$ which is to be reduced to lower dimensional data $Y = [y_1, \ldots, y_N]$, $y_i \in R^d$ where $d < m$. This is the typical problem except that here we assume that the observations y_i come with errors around some "true" submanifold in the high-dimensional space. We consider the following noise model (see also [10]):

$$x_i = f(t_i) + \epsilon_i, \qquad i = 1, \ldots, N,$$

where $x_i \in R^m$ is observed around the manifold $f(\cdot)$ with noise ϵ_i.

For instance, in the Swiss roll example, $m = 3$, and the manifold can be parameterized by t in three dimensions as follows:

$$f(t) = [(3\pi/2 + 3\pi t)\cos(3\pi/2 + 3\pi t), \quad \alpha t, \quad (3\pi/2 + 3\pi t)\sin(3\pi/2 + 3\pi t)], \quad 0 < t < 1, \tag{1}$$

where α is an arbitrary height coefficient.

Instead of observing this manifold deterministically at randomly sampled points $\{t_i\}$, we assume each coordinate comes with error, that is

$$x_i = f(t_i) + \sigma \mathbf{randn}(m, 1),$$

where $m = 3$ is the dimension of the input space and **randn** is the standard normal distribution. This corresponds to independent and identically distributed errors on each of the coordinates, though this can be generalized. Figure 1 shows the actual Swiss roll manifold, the sampled version when the observations are made without error, and the same Swiss roll when each coordinate has a Gaussian error with $\sigma = 1$.

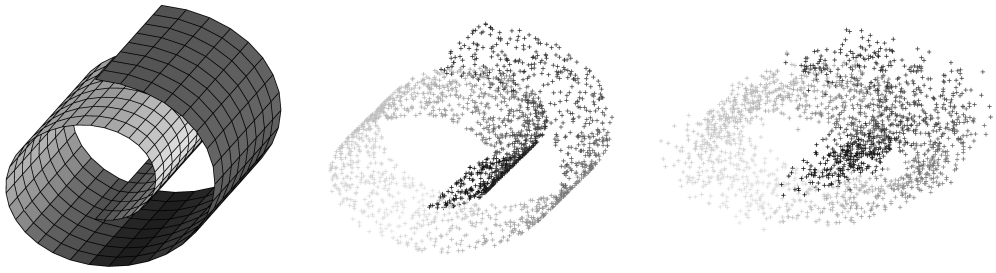

Fig. 1. Swiss roll, sampled without and with noise

Clearly the noise makes for "fuzz" around the actual manifold. Our proposal to reduce it is to replace each point by the average value of its k nearest neighbors i.e.,

Find the k nearest neighbors $\{x_{ij}\}$ of x_i, $j = 1, \ldots, k$ and replace each x_i by

$$\bar{x}_i = \frac{x_i + \sum x_{ij}}{(k+1)}.$$

This has the effect of removing the outliers and in general bringing the points closer the manifold, on the average. In fact, writing t_{ij} as the point that corresponds to the j^{th} neighbor x_{ij} on the manifold, that is, $Ex_{ij} = f(t_{ij})$, the error

after smoothing, can be expressed as

$$
\begin{aligned}
& E\left[\bar{x}_i - f(t_i)\right]\left[\bar{x}_i - f(t_i)\right]' \\
& \quad = E\left[\frac{x_i + \sum x_{ij}}{(k+1)} - f(t_i)\right]\left[\frac{x_i + \sum x_{ij}}{(k+1)} - f(t_i)\right]' \\
& \quad = \frac{E\left[(x_i - f(t_i)) + \sum (x_{ij} - f(t_i))\right]\left[(x_i - f(t_i)) + \sum (x_{ij} - f(t_i))\right]'}{(k+1)^2} \\
& \quad = \frac{E[\epsilon\epsilon']}{(k+1)} + \frac{\sum_j \sum_{j'} [f(t_{ij}) - f(t_i)][f(t_{ij'}) - f(t_i)]'}{(k+1)^2}. \qquad (2)
\end{aligned}
$$

In the equation above, the first term represents the fact we are averaging over $(k+1)$ points and will decrease with k, whereas the second term representing the geometry of the manifold around t_i tends to increase as one moves away from t_i. A further analysis of the second term using a Taylor expansion of $f(\cdot)$ around t_i which explores the tangent plane and curvature, is planned. Intuitively, larger the k, meaning farther one ventures away from t_i, the larger the second term. This is similar in spirit to the variance, bias trade-off in choosing the window-width in density estimation. We explore this empirically in the case of the Swiss roll, in the next section.

An alternative approach to reconstructing a curve out of noisy data is to use so-called "principal curves" as in Hastie and Stuetzle [2] and "principal manifolds" (see [10]). We plan to explore these in subsequent discussions.

3. Swiss roll illustration. In this section, we use the Swiss roll example to illustrate how smoothing reduces noise and discuss the choice of k. Figure 1 gives the actual Swiss roll described by Equation (1). It has been sampled at 2000 points, giving the middle picture and finally the picture on the right is the result of adding independent Gaussian noise with $\sigma = 1$ to each of the 3 coordinates. The goal is to recover as closely as possible. As remarked earlier, the "optimal" choice of k depends on the level of the noise σ, the sampling rate as well as the behavior of the curve in a neighborhood of the point. Given all these as they are in our case, we explore how the choice of k affects the "error", namely:

$$\text{Error}(k) = \sum_i \left[\bar{x}_i - f(t_i)\right]\left[\bar{x}_i - f(t_i)\right]'$$

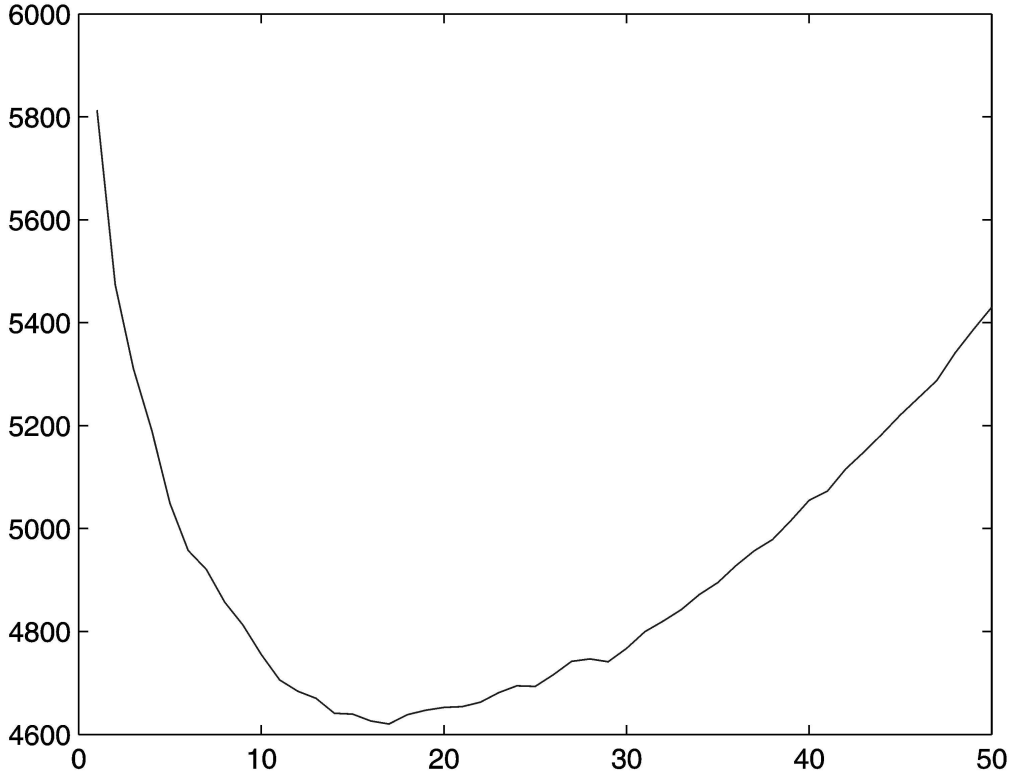

Fig. 2. Error versus k

which is plotted in the following graph.

As can be seen from the graph in Figure 2, the error is a minimum when $k = 17$. The effect of the smoothing can be seen in Figure 3.

4. Cluster distances in smoothed biological data based on LLEs. Spellman et al. [8] analyze time-series data for the yeast cell-cycle. We focus in particular on the time-series gene expression data at 18 time-points, corresponding to the α-factor experiments. The $N = 798$ genes in this experiment have been biologically classified into 5 groups. We use the LLEs as a means of dimension reduction taking 20 points for the local embedding. It is entirely unclear what this 18-dimensional manifold looks like, but since this is indeed noisy data, it suggests that it might benefit by the smoothing of the kind that we discuss here, which we do using an empirically chosen value of $k = 12$. We use

Fig. 3. Noisy swiss roll, before and after smoothing

		MG1	G1	S	S/G2	G2/M
MG1	before smoothing	0.0000	8.4503	50.5820	30.4196	18.5799
	after smoothing	0.0000	10.6744	68.7780	86.4581	17.6147
	improvement	**0.0000**	**2.2241**	**18.1960**	**56.0385**	**-0.9652**
G1	before smoothing			14.2859	6.7652	7.7739
	after smoothing			22.2337	18.6195	9.9770
	improvement			**7.9477**	**11.8543**	**2.2030**
S	before smoothing				6.8534	22.5413
	after smoothing				12.2103	25.6964
	improvement				**5.3569**	**3.1550**
S/G2	before smoothing					5.3671
	after smoothing					6.9041
	improvement					**1.5370**

Fig. 4. Distances between biologically defined clusters, before and after smoothing

the resulting lower dimensional space to do model based clustering which leads to results more in conformity with the biological classes, but more dramatically when we find the Kullback-Liebler Divergence as a measure of distance between the 5 biological classes there is a much stronger separation when LLE is used

after smoothing than before smoothing. The K-L divergence we compute is:

$$D(f,g) = \int f \cdot \log\left(\frac{f}{g}\right) + \int g \cdot \log\left(\frac{g}{f}\right)$$
$$= (1/2)\text{trace}\{(\mu_f - \mu_g)'(\Sigma_f^{-1} + \Sigma_g^{-1})(\mu_f - \mu_g) + \Sigma_f\Sigma_g^{-1} + \Sigma_g\Sigma_f^{-1} - 2I\}$$

The table in Figure 4 gives these distances in the lower dimensional space for the 5 biologically meaningful classes (1) for the original data and (2) when the data is smoothed to remove some noise. The conclusion is therefore that smoothing seems to help improve the discrimination between these various classes, and substantially so in some cases.

The author would like to thank Prof. Tommi Jaakkola of MIT for introduction to this topic.

REFERENCES

[1] M. Belkin, P. Niyogi. Laplacian eigenmaps for dimension reduction and data representation. Technical Report, Dept. of Statistics, Univ. of Chicago, 2001.

[2] T. Hastie, W. Stuetzle. Principal curves. *J. Amer. Statist. Assoc.* , **84** (1988), 502-516.

[3] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. Springer, New York, 2001.

[4] S. Kullback. Information Theory and Statistics. Dover Publications, 1997.

[5] T. A. Myrvoll, F. K. Soong. Optimal Clustering of Multivariate Normal Distributions using Divergence and its applications to HMM Adaptation. IEEE Trasactions on ICASSP, 2003, 552–555.

[6] S. T. Roweis, L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** (2000), 2323–2326.

[7] F. Sha, L. K. Saul. Analysis and Extension of Spectral Methods for Nonlinear Dimensionality Reduction. Proc. of the 22nd Int. Conference on Machine Learning, Bonn, Germany, 2005.

[8] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, D. Botstein, B. Futcher. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cervisiae by Microarray Hybridization. *Molecular Biology of the Cell* **9**, (1998), 3273–3297.

[9] J. Tenenbaum, V. De Silva, J. Langford. A global geometric framework for nonlinear dimension reduction. *Science* **290** (2000), 2319–2323.

[10] Z. Zhang, H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM J. Sci. Comput.* **26** (2004), 313–338.

Arvind K. Jammalamadaka
MIT EECS
77 Massachusetts Ave.
Cambridge, MA 02139 USA
*e-mail:*ajamma@mit.edu

Received October 28, 2008
Revised November 24, 2008