# Theoretical analysis of LLE based on its weighting step

Yair Goldberg and Ya'acov Ritov[*]

Department of Statistics and
The Center for the Study of Rationality
The Hebrew University

March 29, 2011

**Abstract**

The local linear embedding algorithm (LLE) is a widely used nonlinear dimension-reducing algorithm. However, its large sample properties are still not well understood. In this paper we present new theoretical results for LLE based on the way that LLE computes its weight vectors. We show that LLE's weight vectors are computed from the high-dimensional neighborhoods and are thus highly sensitive to noise. We also demonstrate that in some cases LLE's output converges to a linear projection of the high-dimensional input. We prove that for a version of LLE that uses the low-dimensional neighborhood representation (LDR-LLE), the weights are robust against noise. We also prove that for conformally embedded manifold, the pre-image of the input points achieves a low value of the LDR-LLE objective function, and that close-by points in the input are mapped to close-by points in the output. Finally, we prove that asymptotically LDR-LLE preserves the order of the points of a one-dimensional manifold. The Matlab code and and all data sets in the presented examples are available online.

*Keywords: Locally Linear Embedding (LLE), dimension reduction , manifold learning, LDR-LLE*

## 1 Introduction

The locally linear embedding algorithm (LLE) (Roweis and Saul 2000) belongs to a class of recently developed nonlinear dimension-reducing algorithms that include Isomap (Tenenbaum et al. 2000),

[*] This research was supported in part by Israeli Science Foundation grant 209/6. We are grateful to the anonymous reviewers of early versions of this manuscript for their helpful suggestions. Helpful discussions with Alon Zakai and Jacob Goldberger are gratefully acknowledged.

Laplacian Eigenmap (Belkin and Niyogi 2003), Hessian Eigenmap (Donoho and Grimes 2004), LTSA (Zhang and Zha 2004), and MVU (Weinberger and Saul 2006). The underlying assumption when using this group of algorithms is that the data is sitting on, or next to, an embedded manifold of low dimension within the original high-dimensional space. The goal of the algorithms is to find an embedding that maps the input points to the lower-dimensional space. Here a manifold is defined as a topological space that is locally equivalent to a Euclidean space. LLE was found to be useful in data visualization (Roweis and Saul 2000 Xu et al. 2008) and in image processing applications such as image denoising (Shi et al. 2005) and human face detection (Chen et al. 2007). It is also applied in different fields of science, such as chemistry (L'Heureux et al. 2004), biology (Wang et al. 2005), and astrophysics (Xu et al. 2006).

LLE attempts to recover the domain structure of the input data set in three steps. First, LLE assigns neighbors to each input point. Second, for each input point LLE computes weight vectors that best linearly reconstruct the input point from its neighbors. Finally, LLE finds a set of low-dimensional output points that minimize the sum of reconstruction errors, under some normalization constraints.

Saul and Roweis (2003, Section 5.4) suggested a modification of LLE that computes the weight vectors found in the second step of LLE by first finding the best low-dimensional representation for the neighborhood of each point, and then computing the weights with respect to these low-dimensional neighborhoods. We refer to this version of LLE as LLE with low-dimensional neighborhood representation (LDR-LLE). Numerical comparisons between these two versions of neighborhood representation were presented by Saul and Roweis (2003), and later in extended form by Goldberg and Ritov (2008).

In the following we present a theoretical analysis of LLE based on the way that the weight vectors are computed in the second step of LLE. The analysis is divided into two parts. First we study LLE with the usual weighting scheme. We show that LLE's neighborhood description captures the structure of the *high*-dimensional space, and not that of the *low*-dimensional domain. We show two main consequences of this observation. First, the weight vectors are highly sensitive to noise. This implies that a small perturbation of the input may yield an entirely different embedding. Second, we explain why this can cause LLE to converge to a linear projection of the high-dimensional input (see also Wu and Hu 2006). Numerical results that demonstrate our claims are provided.

We then move to analysis of LLE with low-dimensional neighborhood representation (LDR-LLE). We prove a number of theoretical results. We first prove that the weights computed by LDR-LLE are robust against noise. We then prove that when LDR-LLE is used on input points sampled from a manifold that is conformally embedded, the pre-image of the input points achieves a low value of the objective function. We also prove that embedding of LDR-LLE converges to a continuous map of the input, that is, LDR-LLE maps close-by points in the input to close-by points in the output. Finally, we prove that for a large enough sample, LDR-LLE preserves the order of the points of a one-dimensional manifold. Note that this is not true for LLE (see for example Fig. 1C and the explanation there).

The paper is organized as follows. The description of LLE is presented in Section 2. LDR-LLE is described in Section 3. The results for LLE appear in Section 4. Theoretical results regarding LDR-LLE appear in Section 5. Section 6 summarizes the main results. Detailed proofs appear in the Supplemental Materials.

## 2 Description of LLE

The input data $X = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^D$ for LLE is assumed to be sitting on or next to a $d$-dimensional manifold $\mathcal{M}$. We refer to $X$ as an $n \times D$ matrix, where each row stands for an input point. The goal of LLE is to recover the underlying $d$-dimensional structure of the input data $X$. LLE attempts to do so in three steps.

First, LLE assigns neighbors to each input point $x_i$. This can be done, for example, by choosing the input point's $K$-nearest neighbors based on the Euclidian distances in the high-dimensional space. Let the neighborhood matrix of $x_i$ be denoted by $X_i$, where $X_i$ is the $K \times D$ matrix with rows $\eta_j - x_i$ and $\eta_j$ is the $j$-th neighbor of $x_i$.

Second, LLE computes weights $w_i = (w_{ij})_j$ that best linearly reconstruct $x_i$ from its neighbors. These weights minimize the reconstruction error function

$$\varphi_i(w_i) = \|x_i - \sum_j w_{ij} x_j\|^2 \, , \tag{1}$$

where $w_{ij} = 0$ if $x_j$ is not a neighbor of $x_i$, and $\sum_j w_{ij} = 1$. With some abuse of notation, we will also refer to $w_i$ as a $K \times 1$ vector, where we omit the entries of $w_i$ for non-neighbor points. Using this notation, we may write $\varphi_i(w_i) = w_i' X_i X_i' w_i$.

Finally, given the weights found above, LLE finds a set of low-dimensional output points $Y = \{y_1, \ldots, y_n\} \in \mathbb{R}^d$ that minimize the sum of reconstruction errors

$$\Phi(Y) = \sum_{i=1}^{n} \|y_i - \sum_j w_{ij} y_j\|^2 , \tag{2}$$

under the normalization constraints $Y'\mathbf{1} = 0$ and $n^{-1}Y'Y = I$, where $\mathbf{1}$ is vector of ones. These constraints force a unique minimum of the function $\Phi$.

The function $\Phi(Y)$ can be minimized by finding the $d$-bottom non-zero eigenvectors of the sparse matrix $(I - W)'(I - W)$, where $W$ is the matrix of weights. Note that the $p$-th coordinate ($p = 1, \ldots, d$), found simultaneously for all output points $y_i$, is equal to the eigenvector with the $p$-smallest non-zero eigenvalue. This means that the first $p$ coordinates of the LLE solution in $q$ dimensions, $p < q$, are exactly the LLE solution in $p$ dimensions (Roweis and Saul 2000). Equivalently, if an LLE output of dimension $q$ exists, then a solution for dimension $p$, $p < q$, is merely a linear projection of the $q$-dimensional solution on the first $p$ dimensions.

When the number of neighbors $K$ is greater than the dimension of the input $D$, each data point can be reconstructed perfectly from its neighbors, and the local reconstruction weights are no longer uniquely defined. In this case, regularization is needed and one needs to minimize

$$\varphi_i^{\text{reg}}(w_i) = \|x_i - \sum_j w_{ij} x_j\|^2 + \delta \|w_i\|^2 , \tag{3}$$

where $\delta$ is a small constant. Saul and Roweis (2003) suggested $\delta = \frac{\Delta}{K}\text{trace}(X_i X_i')$ with $\Delta \ll 1$. Regularization can be problematic for the following reasons. When the regularization constant is not small enough, it was shown by Zhang and Wang (2007) that the correct weight vectors cannot be well approximated by the minimizer of $\varphi_i^{\text{reg}}(w_i)$. Moreover, when the regularization constant is relatively high, it produces weight vectors that tend towards the uniform vectors $w_i = (1/K, \ldots, 1/K)$. Consequently, the solution for LLE with a large regularization constant is close to that of the Laplacian Eigenmap algorithm (see Belkin and Niyogi 2003, Section 5). In addition, Lee and Verleysen (2007) demonstrated that the regularization parameter must be tuned carefully, since LLE can yield completely different embeddings for different values of this parameter.

# 3 Description of LDR-LLE

In this section we present the modification of LLE that computes the low-dimensional structure of the input points' neighborhoods suggested by Saul and Roweis (2003, Section 5.4). This is done by finding the best representation of rank $d$ (in the $l_2$ sense) for the neighborhood of each point, and then computing the weights with respect to these $d$-dimensional neighborhoods.

We begin by finding a rank-$d$ representation for each local neighborhood. Recall that $X_i$ is the $K \times D$ neighborhood matrix of $x_i$, whose $j$-th row is $\eta_j - x_i$, where $\eta_j$ is the $j$-th neighbor of $x_i$. We assume that the number of neighbors $K$ is greater than $d$, since otherwise $x_i$ cannot (in general) be reconstructed by its neighbors. We say that $X_i^P$ is the best rank-$d$ representation of $X_i$, if $X_i^P$ minimizes $\|X_i - Y\|_2$ over all the $K \times D$ matrices $Y$ of rank $d$. Let $ULV'$ be the SVD of $X_i$, where $U$ and $V$ are orthogonal matrices of size $K \times K$ and $D \times D$, respectively, and $L$ is a $K \times D$ matrix, where $L_{jj} = \lambda_j$ are the singular values of $X_i$ for $j = \min(K, D)$, ordered from the largest to the lowest, and $L_{ij} = 0$ for $i \neq j$. We denote

$$U = \left( \begin{array}{cc} U_1, & U_2 \end{array} \right) \; ; \; L = \left( \begin{array}{cc} L_1, & 0 \\ 0, & L_2 \end{array} \right) \; ; \; V = \left( \begin{array}{cc} V_1, & V_2 \end{array} \right) \tag{4}$$

where $U_1 = (u_1, \ldots, u_d)$ and $V_1 = (v_1, \ldots, v_d)$ are the first $d$ columns of $U$ and $V$, respectively, $U_2$ and $V_2$ are the last $K - d$ and $D - d$ columns of $U$ and $V$, respectively, and $L_1$ and $L_2$ are of dimension $d \times d$ and $(K-d) \times (D-d)$, respectively. Then by Corollary 2.3-3 of Golub and Loan (1983), $X_i^P$ can be written as $U_1 L_1 V_1'$.

For LLE, the weight vectors are found by minimizing (1). For $X_i^P$, the solution for this minimization problem is not unique, since by the construction all the vectors spanned by $u_{d+1}, \ldots, u_K$ zero this function. Thus, one can choose the weight vector in the span of $u_{d+1}, \ldots, u_K$ that has the smallest $l_2$ norm (Saul and Roweis 2003, Section 5). In other words, the weight vector can be found as

$$\underset{\substack{w_i \in \text{span}\{u_{d+1}, \ldots, u_K\} \\ w_i'\mathbf{1} = 1}}{\text{argmin}} \|w_i\|^2 . \tag{5}$$

Note that it is assumed that $\mathbf{1} \notin \text{span}\{u_1, \ldots, u_d\}$. This is true whenever the neighborhood points are in *general position*, i.e., no $d+1$ of them lie in a $(d-1)$-dimensional plane. To understand this, note that if $\mathbf{1} \in \text{span}\{u_1, \ldots, u_d\}$, then $(I - \frac{1}{K}\mathbf{1}\mathbf{1}')X_i^P = (I - \frac{1}{K}\mathbf{1}\mathbf{1}')U_1 L_1 V_1'$ is of rank $d - 1$. Since $(I - \frac{1}{K}\mathbf{1}\mathbf{1}')X_i^P$ is the projected neighborhood after centering, we obtained that the dimension of the centered projected neighborhood is of dimension $d-1$, and not $d$ as assumed, and therefore the points are not in general position. See also Assumption (A2) in Section 5 and the discussion that follows.

The following lemma shows how to compute the vector $w_i$ that minimizes (5).

**Lemma 3.1.** *Assume that the points of $X_i^P$ are in general position. Then the vector $w_i$ that minimizes* (5) *is given by*

$$w_i = \frac{U_2 {U_2}' \boldsymbol{1}}{\boldsymbol{1}' U_2 {U_2}' \boldsymbol{1}} \, . \tag{6}$$

Using Lemma 3.1, we can write LDR-LLE as follows. First, LDR-LLE assigns neighbors to each input point $x_i$, as in LLE, obtaining the matrix $X_i$. Second, the weight vectors are computed as follows. Write $X_i = ULV'$ and $U_2 = (u_{d+1} \ldots , u_K)$. The weights are given by

$$w_i = \frac{U_2 {U_2}' \mathbf{1}}{\mathbf{1}' U_2 {U_2}' \mathbf{1}} \, .$$

Finally, the $d$-dimensional embedding is found by minimizing $\Phi(Y)$ (see (2)), as in LLE.

Note that the difference between LDR-LLE and LLE is in the second step. LDR-LLE computes the *low*-dimensional neighborhood representation of each neighborhood and obtains its weight vector, while LLE computes the weight vector for the original *high*-dimensional neighborhoods. One consequence of this approach is that the weight vectors $w_i$ of LDR-LLE are less sensitive to perturbation, as shown in Theorem 5.1.

# 4 Preservation of high-dimensional neighborhood structure by LLE

In this section we focus on the computation of the weight vectors, which is performed in the second step of LLE. We first show that LLE characterizes the *high*-dimensional structure of the neighborhood. We explain how this can lead to the failure of LLE to find a meaningful embedding of the input. Two additional consequences of preservation of the high-dimensional neighborhood structure are discussed. First, LLE's weight vectors are sensitive to noise. Second, LLE's output may tends toward a linear projection of the input data when the number of input points tends to infinity. These claims are demonstrated using numerical examples.

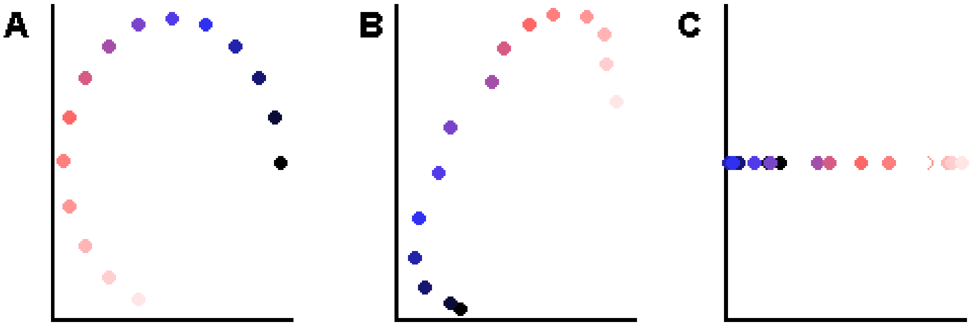

Figure 1: The input for LLE is the 16-point open ring that appears in (A). The two-dimensional output of LLE is given in (B). LLE finds and preserves the two-dimensional structure of each of the local neighborhoods. The one-dimensional output of LLE appears in (C). Note that LLE fails to unfold the ring (compare to Fig. 5). The computation was performed using 4-nearest-neighbors, and regularization constant $\Delta = 10^{-9}$.

We begin by showing that LLE preserves the high-dimensional neighborhood structure. We use the example that appears in Fig. 1. The input is a sample from an open ring which is a one-dimensional manifold embedded in $\mathbb{R}^2$. For each point on the ring, we define its neighborhood using its 4 nearest neighbors. Note that its *high*-dimensional ($D = 2$) neighborhood structure is curved, while the *low*-dimensional structure ($d = 1$) is a straight line. The two-dimensional output of LLE (see Fig. 1) is essentially a reconstruction of the input. In other words, LLE's weight vectors preserve the curved shape of each neighborhood.

The one-dimensional output of the open ring is presented in Fig. 1C. Recall that the one-dimensional solution is a linear projection of the two-dimensional solution, as explained in Section 2. In the open-ring example, LLE clearly fails to find an appropriate one-dimensional embedding, because it preserves the two-dimensional curved neighborhood structure. We shall now show that this holds true in some additional cases.

The swissroll output in Fig. 2B shows that the overall three-dimensional structure of the swissroll is preserved in the three-dimensional embedding. The two-dimensional output of LLE appears in Fig. 2C. It can be seen that LLE does not succeed in finding a meaningful embedding in this case. Fig. 3 presents the 'S' curve, with similar results.

We performed LLE, here and in all other examples, using the LLE Matlab code as it appears

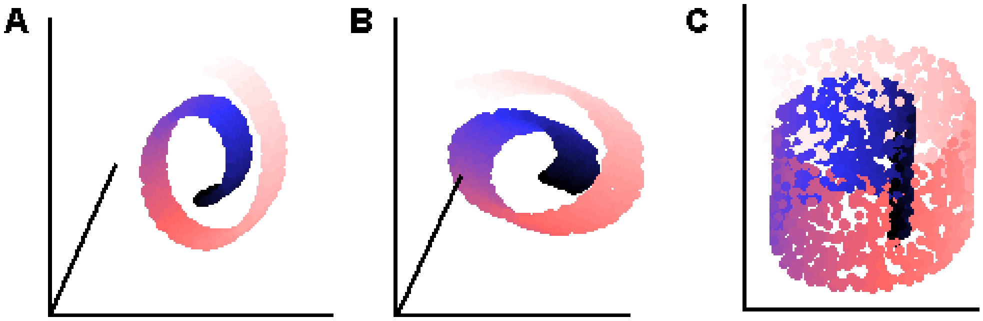

Figure 2: (A) LLE's input, a 2000-point swissroll. (B) The three-dimensional output of LLE. It can be seen that LLE finds the overall three-dimensional structure of the input. (C) The two-dimensional output of LLE. Note that LLE fails to unfold the swissroll.

on the LLE website [1]. The code that produced the input data for the swissroll (Fig. 2A) was also taken from the LLE website. We used the default values of 2000 sample points and $K = 12$ nearest neighbors, with $\Delta = 10^{-9}$ as the regularization constant. It should be noted that using a large regularization constant improved the results. However, weight vectors produced with large regularization constant do not approximate the neighborhood and tend toward a uniform vector.

The 'S' curve data (Fig 3A) was obtained by embedding the 2000-point sample produced using the code taken from the LLE website in $\mathbb{R}^D$, with $D = 15$. This embedding was obtained by adding a normal random vector with zero mean and $10^{-6}I$ variance matrix to each point. We used $K = 12$ in the computation. Note that since $K < D$, *no regularization is needed*. The failure to find the low-dimensional embedding is, therefore, inherent and is not due to the choice of regularization constant. It should be noted that roughly the same result was obtained when using the original three-dimensional 'S' curve with $\Delta = 10^{-9}$. The open ring, swissroll, and 'S' curve data sets can be found online (see Section A).

We now discuss the sensitivity of LLE's weight vectors $\{w_i\}$ to noise. Figure 4 shows that an arbitrarily small change in the neighborhood can cause a large change in the weight vectors. This result can be understood by noting how the vector $w_i$ is obtained. It can be shown (Saul and Roweis 2003) that $w_i$ equals $(X_i X_i')^{-1}\mathbf{1}$, up to normalization. Sensitivity to noise is therefore expected when the condition number of $X_i X_i'$ is large (see Golub and Loan 1983, Section 2).

[1] `http://www.cs.toronto.edu/~roweis/lle/`. The changes in the Matlab function *eigs* were taken into account.

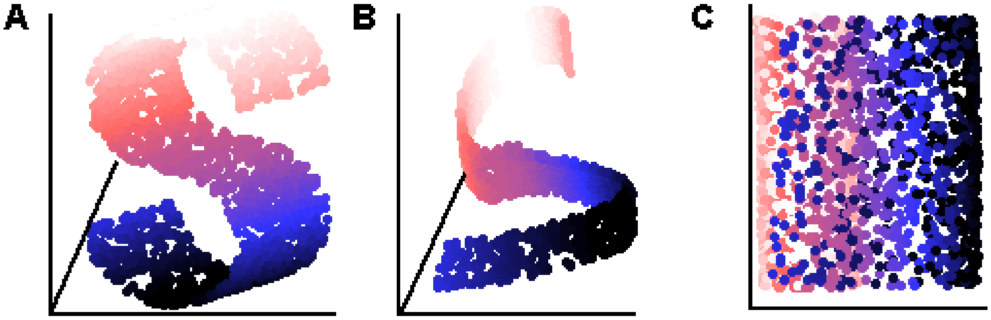

Figure 3: (A) The first three dimensions of LLE's input, a 2000-point 'S' curve embedded in $\mathbb{R}^{15}$. (B) The three-dimensional output of LLE. It can be seen that LLE finds the overall three-dimensional structure of the input. (C) The two-dimensional output of LLE. Note that LLE fails to unfold the 'S' curve.

One way to solve this problem is to enforce regularization, with its associated problems (see Section 2). We note that the sensitivity of LLE's weights to noise means that two similar inputs can result in widely varying outputs. This is clearly an undesirable property, since the parametric representation of two similar inputs is expected to be similar.

One more implication of the fact that LLE preserves the high-dimensional neighborhood structure is that LLE's output may tend to a linear projection of the input data. Wu and Hu (2006) proved for a finite data set that when the reconstruction errors are exactly zero for each of the neighborhoods, and under some dimensionality constraint, the output of LLE must be a linear projection of the input data. Here, we present a simple argument that explains why LLE's output tends to a linear projection when the number of input points tends to infinity, and show numerical examples that strengthen this claim. For simplicity, we assume that the input data is normalized.

Our argument is based on two claims. First, note that LLE's output for dimension $d$ is a linear projection of LLE's output for dimension $D$ (see Section 2). Second, note that by definition, the LLE output is a set of points $Y$ that minimizes the sum of reconstruction errors $\Phi(Y)$. For normalized input $X$ of dimension $D$, when the number of input points tends to infinity, each point is well reconstructed by its neighboring points. Therefore the reconstruction error $\varphi_i(w)$ tends to zero for each point $x_i$. This means that the input data $X$ tends to minimize the sum of reconstruction errors $\Phi(Y)$. Hence, the output points $Y$ of LLE for output of dimension $D$ tend to the input points (up to a rotation). The result of these two claims is that when the neighborhoods

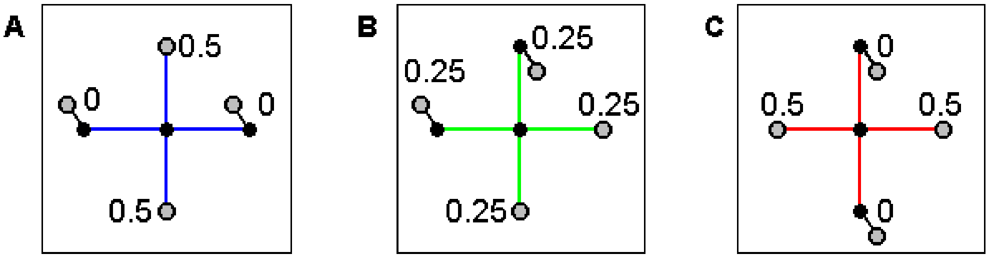

Figure 4: The effect of a small perturbation on the weight vector computed by LLE. All three panels show the same unperturbed neighborhood, consisting of a point and its four nearest-neighbors (black points), all sitting in the two-dimensional plane. Each panel shows a different small perturbation of the original neighborhood (gray points). All perturbations are in the direction orthogonal to the plane of the original neighborhood. (A) and (C): Both perturbations are in the same direction. (B) Perturbations are of equal size, in opposite directions. The unique weight vector for the center point is denoted for each case. These three different weight vectors vary widely, even though the different perturbations can be arbitrarily small.

are reconstructed well, any requested solution of dimension $d < D$ tends to a linear projection of the $D$-dimensional solution, i.e., a linear projection of the input data.

The result that LLE tends to a linear projection is of an asymptotical nature. However, numerical examples show that this phenomenon can occur even when the number of points is relatively small. This is indeed the case for the outputs of LLE shown in Figs. 1C, 2C, and 3C, for the open ring, the swissroll, and the 'S' curve, respectively.

Note that the linear projection does not occur to LDR-LLE. This is due to the fact that the $d$-dimensional output of LDR-LLE is not a projection of the embedding in dimension $q$, $q > d$. This is because the weight vectors $w_i$ are computed differently for different values of output dimension $d$. In particular, the input data no longer minimize $\Phi$ when $d < D$, and therefore the linear projection problem does not arise (see Fig 5 for numerical example).

# 5 Theoretical results for LDR-LLE

In this section we prove theoretical results regarding the computation of LDR-LLE. We first show that a small perturbation of the neighborhood has a small effect on the weight vector. Then

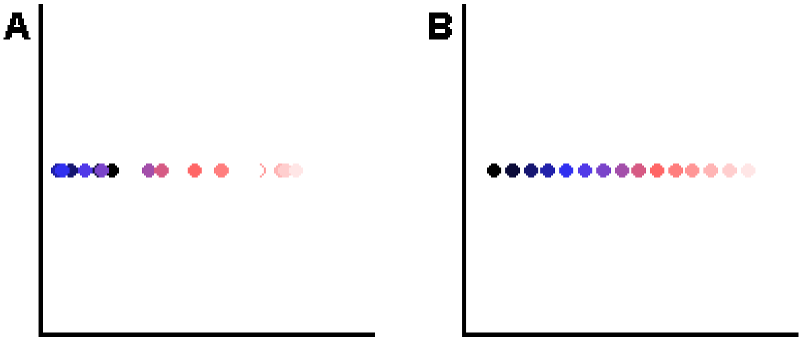

Figure 5: The outputs of LLE and LDR-LLE for the open ring (Fig. 1A) appear in (A) and (B) respectively

.

we show that the set of original points in the low-dimensional domain, that is the pre-image of the input points, achieves a low value of the objective function $\Phi$. We also show that that close-by points in the input are mapped to close-by points in the output. Finally, we prove that asymptotically LDR-LLE preserves the order of the points of a one-dimensional manifold.

We start with some definitions. Let $\Omega \subset \mathbb{R}^d$ be a compact set and let $f : \Omega \to \mathbb{R}^D$ be a smooth conformal mapping. This means that the inner products on the tangent bundle at each point are preserved up to a scalar $c$ that may change continuously from point to point. Note that the class of isometric embeddings is included in the class of conformal embeddings. Let $\mathcal{M}$ be the $d$-dimensional image of $\Omega$ in $\mathbb{R}^D$. Assume that the input $X = \{x_1, \ldots, x_n\}$ is a sample taken from $\mathcal{M}$. For each point $x_i$, define the neighborhood $X_i$ and its low-dimensional representation $X_i^P$ as in Section 3. Let $X_i = ULV'$ and $X_i^P = U_1 L_1 V_1{}'$ be the SVDs of the $i$-th neighborhood and its projection, respectively. Denote the singular values of $X_i$ by $\lambda_1^i \geq \ldots \geq \lambda_K^i$, where $\lambda_j^i = 0$ if $D < j \leq K$. Denote the mean vector of the projected $i$-th neighborhood by $\mu_i = \frac{1}{K}\mathbf{1}' X_i^P$.

For the proofs of the theorems we require that the local high-dimensional neighborhoods satisfy the following two assumptions:

(A1) For each $i$, $\lambda_{d+1}^i \ll \lambda_d^i$.

(A2) There is an $\alpha < 1$ such that for all $i$, $\frac{1}{K}\mathbf{1}' U_1 U_1{}' \mathbf{1} < \alpha$.

The first assumption states that for each $i$, the neighborhood $X_i$ is essentially $d$-dimensional. For our purposes it enough to demand $\lambda_{d+1}^i < \min\left\{(\lambda_d^i)^2, \frac{\lambda_d^i}{72}\right\}$. The second assumption is equivalent to the requirement that points in each projected neighborhood be in general position (see discussion in Section 4). We now show that this is equivalent to the requirement that the variance-covariance matrix of the projected neighborhood is not degenerate. Denote $S = \frac{1}{K} X_i^{P\prime} X_i^P = \frac{1}{K} V_1 L_1^2 V_1{}'$; then

$$\frac{1}{K}\mathbf{1}'U_1U_1{}'\mathbf{1} = \frac{1}{K}\mathbf{1}'(U_1L_1V_1{}')(V_1L_1^{-2}V_1{}')V_1L_1U_1{}'\mathbf{1} = \mu' S^{-1}\mu\,.$$

Note that since $S - \mu\mu'$ is positive definite, so is $I - S^{-1/2}\mu\mu' S^{-1/2}$. Since the only eigenvalues of $I - S^{-1/2}\mu\mu' S^{-1/2}$ are 1 and $1 - \mu' S^{-1}\mu$, we obtain that $\mu' S^{-1}\mu < 1$.

**Theorem 5.1.** *Let $E_i$ be a $K \times D$ matrix such that $\|E_i\|_F = 1$. Let $\widetilde{X}_i = X_i + \varepsilon E_i$ be a perturbation of the i-th neighborhood. Assume (A1) and (A2) and $\varepsilon < \min\left(\frac{(\lambda_d^i)^4}{72}, \frac{(\lambda_d^i)^2(1-\alpha)}{72}\right)$ and that $\lambda_1^i < 1$. Let $w_i$ and $\tilde{w}_i$ be the weight vectors computed by LDR-LLE for $X_i$ and $\widetilde{X}_i$, respectively, as defined by* (5)*. Then*

$$\|w_i - \tilde{w}_i\| < \frac{20\varepsilon}{(\lambda_d^i)^2(1-\alpha)}\,.$$

See proof in the Supplemental Materials.

Note that the assumption that $\lambda_1^i < 1$ can always be fulfilled by rescaling the matrix $X_i$ since rescaling the input matrix $X$ has no influence on the value of $w_i$.

Fig. 4 demonstrates why no bound similar to Theorem 5.1 exists for the weights computed by LLE. In the example we see a point on the grid with its 4-nearest neighbors, where some noise was added. Although $\lambda_1 \approx \lambda_2 \approx 1 - \alpha \approx 1$, and $\varepsilon$ is arbitrarily small, the distance between each pair of vectors is at least $\frac{1}{2}$. Conversely, the bound of Theorem 5.1 states that for $\varepsilon = 10^{-2}, 10^{-4}$, and $10^{-6}$ the upper bound on the distance when using the LDR-LLE is $20 \cdot 10^{-2}, 20 \cdot 10^{-4}$, and $20 \cdot 10^{-6}$, respectively. The empirical results shown in Fig. 6 are even lower.

For the second theoretical result we require some additional definitions.

The *minimum radius of curvature* $r_0 = r_0(\mathcal{M})$ is defined to be:

$$\frac{1}{r_0} = \max_{\gamma, t}\left\{\|\ddot{\gamma}(t)\|\right\}\,,$$

where $\gamma$ varies over all unit-speed geodesics in $\mathcal{M}$ and $t$ is in a domain of $\gamma$. The *minimum branch separation* $s_0 = s_0(\mathcal{M})$ is defined as the largest positive number for which $\|x - \tilde{x}\| < s_0$ implies that $d_{\mathcal{M}}(x, \tilde{x}) \leq \pi r_0$, where $x, \tilde{x} \in \mathcal{M}$, and $d_{\mathcal{M}}(x, \tilde{x})$ are the geodesic distance between $x$ and $\tilde{x}$ (for both definitions, see Bernstein et al. 2000).

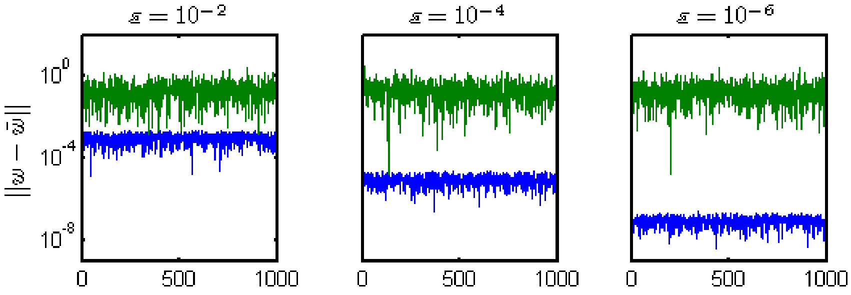

Figure 6: The effect of neighborhood perturbation on the weight vectors of LLE and of LDR-LLE. The original neighborhood consists of a point on the two-dimensional grid and its 4-nearest neighbors, as in Fig. 4. A 6-dimensional noise matrix $\varepsilon E$ where $\|E\|_F = 1$ was added to the neighborhood for $\varepsilon = 10^{-2}, 10^{-4}$, and $10^{-6}$, with 1000 repetitions for each value of $\varepsilon$. Note that no regularization is needed since $K < D$. The graphs show the distance between the vector $w = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ and the vectors computed by LLE (top graph, in green) and by LDR-LLE (bottom graph, in blue). Note the log scale in the $y$ axis.

Define the radius $r(i)$ of neighborhood $i$ to be

$$r(i) = \max_{j \in \{1,\ldots,K\}} \|\eta_j - x_i\|$$

where $\eta_j$ is the $j$-th neighbor of $x_i$. Finally, define $r_{\max}$ to be the maximum over $r(i)$ .

We say that the sample is *dense* with respect to the chosen neighborhoods if $r_{\max} < s_0$. Note that this condition depends on the manifold structure, the given sample, and the choice of neighborhoods. However, for a given compact manifold, if the distribution that produces the sample is supported throughout the entire manifold, then this condition is valid with probability increasing towards 1 as the size of the sample is increased and the radius of the neighborhoods is decreased.

**Theorem 5.2.** *Let $\Omega$ be a compact convex set in $\mathbb{R}^d$ which is equal to the closure of its interior. Let $f : \Omega \to \mathbb{R}^D$ be a smooth conformal mapping. Let $X$ be an $n$-point sample taken from $f(\Omega)$, and let $Z = f^{-1}(X)$, i.e., $z_i = f^{-1}(x_i)$. Assume that the sample $X$ is dense with respect to the choice of neighborhoods and that assumptions (A1)and (A2) hold. Then, if the weight vectors are chosen according to LDR-LLE,*

$$\frac{\Phi(Z)}{n} = \max_i \lambda^i_{d+1} \mathcal{O}\left(r^2_{\max}\right) . \tag{7}$$

See proof in the Supplemental Materials.

The theorem states that the pre-image $Z$ has a small value of $\Phi$ and thus is a reasonable embedding, although not necessarily the minimizer (see Goldberg et al. 2008). This observation is not trivial for two reasons. First, it is not known apriori that $\{f^{-1}(\eta_j)\}$, the pre-image of the neighbors of $x_i$, are also neighbors of $z_i = f^{-1}(x_i)$. When short circuits occur, this need not be true (see Balasubramanian et al. 2002). Second, the weight vectors $\{w_i\}$ characterize the projected neighborhood, which is only an approximation to the true neighborhood. Nevertheless, the theorem shows that $Z$ has a low $\Phi$ value.

The above discussion raises the question of assessing the quality of the embedding $Y$ obtained by the algorithm. While there is no single accepted measure for quality of embedding (see, for example,Venna and Kaski (2006);Chen (2006);Goldberg and Ritov (2009)), a reasonable demand is that close-by points in the pre-image $Z = f^{-1}(X)$ should be mapped to close-by points in the embedding $Y$. In the following we prove that when the number of points in the sample tends to infinity, close-by points in $Z$ are indeed mapped to close-by points in $Y$ with probability tends to one, at least for inner points.

Before we state this result we need to discuss the size of the neighborhoods. Let $X = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^D$ be the input data. Consider a neighborhood graph of the input matrix $X$ with $n$ vertices, and an edge between vertices $i$ and $j$ if $x_i$ is in the neighborhood of $x_j$ or vice versa. Note that if $K$ remains bounded as $n$ grows to infinity, the graph is likely to be unconnected when the observations are taken from a random sample from a positive density on the manifold. For example, for $d = 1$, the gaps are asymptotically exponentially distributed (locally i.i.d.). The number of cliques is approximately $n$ times the (positive) probability that an exponential random variable will be larger than two independent Gamma random variables with $K - 1$ degrees of freedom. But when the graph is not connected, the mapping of the cliques is arbitrary, and the resulting mapping will not resemble the original manifold. We need, therefore, to ensure enough overlapping between the adjacent neighborhoods.

In the following we take the radii of the neighborhoods to zero while ensuring that the number of observations within each neighborhood grows to infinity. More specifically we assume the following:

(A3) The input sample is taken from a continuous density on $f(\Omega)$ that is bounded away from zero and infinity. Moreover, there is an $\varepsilon_o > 0$ such that for all $x \in f(\Omega)$, and $0 < r \leq \varepsilon_o$, the probability of $B(x, r)$ is greater than $\delta r^d$, where $\delta > 0$ is some constant and $B(x, r) \subset \mathbb{R}^D$ is the ball of radius $r$ around $x$.

Note that since the radii tend to zero, we obtain that sample is dense for all $n$ large enough.

As before let $f : \Omega \to \mathbb{R}^D$ be a conformal mapping. For each $z \in \Omega$, let $J(z)$ be the Jacobian of $f$ at $z$. Since $f$ is conformal, there is a continuous function $c : \Omega \to \mathbb{R}_+$ such that $J(z)'J(z) = c(z)I$, where $I$ is the $d \times d$ identity matrix. Without loss of generality, we assume that $c(z) \geq \pi/2$, since we can always multiply $f$ by $1/\min(c(z))$. Let $\partial\Omega = \Omega/\Omega^o$ be the boundary of $\Omega$ and let $\text{dist}(z, \partial\Omega) = \min_{z' \in \partial\Omega} \|z - z'\|$ denote the distance of the point $z$ from the boundary.

We are now ready to state the theorem. To simplify the proof, the following theorem is expressed in terms of balls and not $K$-neighborhoods.

**Theorem 5.3.** *Let $f : \Omega \to \mathbb{R}^D$ as in Theorem 5.2 and let $X = X_n$ be an $n$-size sample taken from $f(\Omega)$. Assume that (A1)-(A3) hold. Let the neighborhoods be defined by balls with radius $r$, where $r = r_n \to 0$ while $nr^d \to \infty$. Let $\rho = \rho_n$ be such that $\rho/r \to 0$ but $n\rho^d \to \infty$. Then*

$$\frac{1}{n} \sum_{\{i:\mathrm{dist(z_i,\partial\Omega)>2r+\rho}\}} \max_{\{j:\|z_i - z_j\| < \rho\}} \|y_i - y_j\|^2 \leq \mathcal{O}_p(\rho/r). \tag{8}$$

*Furthermore, if $nr^{d(d+1+\eta)} \to \infty$ for some $\eta > 0$, then for any $\varepsilon > 0$ with probability converging to 1*

$$\max_{\{i:\mathrm{dist(z_i,\partial\Omega)>2r}\}} \max_{\{j:\|z_i - z_j\| < r^{d+1+\eta}\}} \|y_i - y_j\|^2 \leq \varepsilon. \tag{9}$$

The proof is given in the Supplemental Materials.

Theorem 5.3 proves that close-by points are mapped to close-by points. However, are far points mapped to far points? The following theorem argues that this is the case at least for $d = 1$. We show that for $d = 1$, $Y$ retains the order of $Z$, at least if the neighborhood size is selected with care.

**Theorem 5.4.** *Consider the setup of the previous theorems. Suppose $d = 1$ and the curve has a bounded curvature. Suppose the the points are taken from a bounded density on an interval which is bounded away from 0. Let the i-th neighborhood be the $K$ points proceeding and the $K$ points following $x_i$ on the curve. Suppose $K/n \to 0$, but $K^{9/7}/n \to \infty$. Suppose, with out loss of* *generality, that the pre-image is $z_1 < \cdots < z_n$. Then there are positive $\epsilon_n, \delta_n \to 0$, such that for all $i, j$, $\ell > 0$ such that $\epsilon_n n < i < j < j + \ell < (1 - \epsilon_n)n$ we have $y_{i+\ell} - y_i = y_{j+\ell} - y_j + o_p(\delta_n)$.*

In other words, the theorem states that all non-negligible differences between any two ordered inner points have the same sign, thus order is preserved. Note that when the sample size is large enough, with probability that tends to 1, one can choose the neighborhoods structure used in this theorem.

The proof is given in the Supplemental Materials.

# 6 Summary

In this work we study theoretical properties of the algorithm LLE. We demonstrated two limitations of LLE. First, we showed that the weight vectors computed by LLE are highly sensitive to noise. Second, we showed that LLE may converge to a linear projection of the high-dimensional input when the number of input points tends to infinity. We showed that this is a result of the fact that LLE captures the high-dimensional structure of the neighborhoods, and not the low-dimensional manifold structure.

As opposed to LLE, the LDR-LLE version of LLE finds the best low-dimensional representation for the neighborhood of each point. We proved that the weights computed by LDR-LLE are robust against noise. We also proved that when LDR-LLE is used on input points sampled from a conformally embedded manifold, the pre-image of the input points achieves a low value of the objective function. In addition we proved that close-by points in the input are mapped to close-by points in the output. Finally, we proved that asymptotically LDR-LLE preserves the order of the points of a one-dimensional manifold.

We believe that the results presented here are only the first step in the understanding of the theoretical properties of LLE. Many other questions are still open: The theoretical properties of LLE for isometrically embedded manifolds are not known, and it is not clear under which conditions LLE succeeds to find the underlying structure of the manifold. We hope that the theoretical tools derived in this paper will be of service in future studies of LLE and other manifold-learning techniques.

# A Supplemental Materials

**Supplementary Proofs** Detailed proofs for Lemma 3.1 and Theorems 5.1-5.4.

**Code and data sets** The archive file LDR_LLE.zip contains the MATLAB code and all data sets used in this work, as well as a readme.pdf file that describes all of the other files in the archive.

# References

M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva, and J. C. Langford. The isomap algorithm and topological stability. *Science*, 295(5552):7, 2002.

M. Belkin and P. Niyogi. Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Comp.*, 15(6):1373–1396, 2003.

M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University., 2000.

J. Chen, R. Wang, S. Yan, S. Shan, X. Chen, and W. Gao. Enhancing human face detection by resampling examples through manifolds. *IEEE Transactions on Systems, Man and Cybernetics, Part A.*, 37(6):1017–1028, 2007.

L. Chen. *Local multidimensional scaling for nonlinear dimension reduction, graph layout and proximity analysis.* PhD thesis, University of Pennsylvania, 2006.

D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. U.S.A.*, 100(10):5591–5596, 2004.

Y. Goldberg and Y. Ritov. LDR-LLE: LLE with low-dimensional neighborhood representation. In *Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II*, ISVC 08, pages 43–54, 2008.

Y. Goldberg and Y. Ritov. Local Procrustes for manifold embedding: a measure of embedding quality and embedding algorithms. *Machine Learning*, 77:1–25, 2009.

Y. Goldberg, A. Zakai, D. Kushnir, and Y. Ritov. Manifold learning: The price of normalization. *J. Mach. Learn. Res.*, 9:1909–1939, 2008.

G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, Maryland, 1983.

J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction.* Springer, Berlin, 2007.

P. J. L'Heureux, J. Carreau, Y. Bengio, O. Delalleau, and S. Y. Yue. Locally linear embedding for dimensionality reduction in qsar. *J. Comput. Aided Mol. Des.*, 18:475–482, 2004.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low-dimensional manifolds. *J. Mach. Learn. Res.*, 4:119–155, 2003. ISSN 1533-7928.

R. Shi, I. Shen, and W. Chen. Image denoising through locally linear embedding. In *CGIV '05: Proceedings of the International Conference on Computer Graphics, Imaging and Visualization*, pages 147–152. IEEE Computer Society, 2005.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

J. Venna and S. Kaski. Local multidimensional scaling. *Neural Netw.*, 19(6):889–899, 2006.

M. Wang, H. Yang, Z. H. Xu, and K. C. Chou. SLLE for predicting membrane protein types. *J. Theor. Biol.*, 232(1):7–15, 2005.

K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

F. C. Wu and Z. Y. Hu. The LLE and a linear mapping. *Pattern Recognition*, 39(9):1799–1804, 2006.

W. Xu, X. Lifang, Y. Dan, and H. Zhiyan. Speech visualization based on locally linear embedding (LLE) for the hearing impaired. In *BMEI (2)*, pages 502–505, 2008.

X. Xu, F. C. Wu, Z. Y. Hu, and A. L. Luo. A novel method for the determination of redshifts of normal galaxies by non-linear dimensionality reduction. *Spectroscopy and Spectral Analysis*, 26 (1):182–186, 2006.

Z. Zhang and J. Wang. MLLE: Modified locally linear embedding using multiple weights. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1593–1600. MIT Press, Cambridge, MA, 2007.

Z. Y. Zhang and H. Y. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comp*, 26(1):313–338, 2004.