Hessian Eigenmaps:
new locally linear embedding techniques for high-dimensional data

David L. Donoho[*]   and   Carrie Grimes[†]

Abstract

We describe a method to recover the underlying parametrization of scattered data (m_i) lying on a manifold M embedded in high-dimensional Euclidean space. The method, *Hessian-based Locally Linear Embedding* (HLLE), derives from a conceptual framework of *Local Isometry* in which the manifold M, viewed as a Riemannian submanifold of the ambient Euclidean space $\mathbb{R}^n$, is locally isometric to an open, connected subset Θ of Euclidean space $\mathbb{R}^d$. Since Θ does not have to be convex, this framework is able to handle a significantly wider class of situations than the original Isomap algorithm.

The theoretical framework revolves around a quadratic form $\mathcal{H}(f) = \int_M \|H_f(m)\|_F^2 dm$ defined on functions $f : M \mapsto \mathbb{R}$. Here Hf denotes the Hessian of f, and $\mathcal{H}(f)$ averages the Frobenius norm of the Hessian over M. To define the Hessian, we use orthogonal coordinates on the tangent planes of M.

The key observation is that, if M truly is locally isometric to an open connected subset of $\mathbb{R}^d$, then $\mathcal{H}(f)$ has a $(d+1)$-dimensional nullspace, consisting of the constant functions and a d-dimensional space of functions spanned by the original isometric coordinates. Hence, the isometric coordinates can be recovered up to a linear isometry.

Our method may be viewed as a modification of the Locally Linear Embedding and our theoretical framework as a modification of the Laplacian Eigenmaps framework, where we substitute a quadratic form based on the Hessian in place of one based on the Laplacian.

Keywords: Manifold Learning; Locally Linear Embedding; Isomap; Tangent Coordinates; Isometry; Laplacian Eigenmaps; Hessian Eigenmaps.

[*]Corresponding author: Department of Statistics, Room 128, Sequoia Hall, Stanford University, Stanford, CA 94305, USA, Email: donoho@stat.stanford.edu Voice: (650) 723-2620 Fax: (650) 725-8977.

[†]Department of Statistics, Stanford University, Stanford, CA 94305-9025, USA.

1 Introduction

A recent article in *Science* [1] proposed to recover a low-dimensional parametrization of high-dimensional data by assuming that the data lie on a manifold M which, viewed as a Riemannian submanifold of the ambient Euclidean space, is globally isometric to a convex subset of a low-dimensional Euclidean space. This bold assumption has been surprisingly fruitful, although the extent to which it holds is not fully understood.

It is now known [4, 5] that there exist high-dimensional libraries of articulated images for which the corresponding data manifold is indeed locally isometric to a subset of a Euclidean space; however it is easy to see that in general, the assumption that the subset will be convex is unduly restrictive. Convexity can fail in the setting of image libraries due to (i) exclusion phenomena [4, 5], where certain regions of the parameter space would correspond to collisions of objects in the image, or (ii) unsystematic data sampling, which investigates only a haphazardly chosen region of the parameter space.

In this paper, we describe a method which works to recover a parametrization for data lying on a manifold which is locally isometric to an open, connected subset Θ of Euclidean space $\mathbb{R}^d$. Since this subset need not be convex, while the original method proposed in [1] demands convexity, the new proposal significantly expands of the class of cases which can be solved by isometry principles.

Justification of our method follows from properties of a quadratic form $\mathcal{H}(f) = \int_M ||H_f(m)||_F^2 dm$ defined on functions $f : M \mapsto \mathbb{R}$. $\mathcal{H}(f)$ measures the average, over the data manifold M, of the Frobenius norm of the Hessian of f. To define the Hessian, we use orthogonal coordinates on the tangent planes of M.

The key observation is that, if M is locally isometric to an open connected subset of $\mathbb{R}^d$, then $\mathcal{H}(f)$ has a $d+1$ dimensional nullspace, consisting of the constant function and a d-dimensional space of functions spanned by the original isometric coordinates. Hence, the isometric coordinates can be recovered, up to a rigid motion, from the null space of $\mathcal{H}(f)$.

Our method may be viewed as variant of the Local Linear Embedding method of Roweis and Saul [2]; and our conceptual framework may be viewed as a modification of the Laplacian Eigenmaps framework of Belkin and Niyogi [3]. In these modifications, we substitute a quadratic form based on the Hessian in place of the original ones based on the Laplacian. The point of interest is that a linear function has everywhere vanishing Laplacian, but not every function with everywhere vanishing Laplacian is linear, while a function is linear if and only if it has everywhere vanishing Hessian. By substituting the Hessian for the Laplacian, we find a global embedding which is nearly linear in every set of local tangent coordinates.

We describe an implementation of this procedure on sampled data and demonstrate that it performs consistently with the theoretical predictions on a variant of the 'Swiss Roll' example, where the data are not sampled from a convex region in parameter space.

2   Notation and Motivation

Suppose we have a parameter space $\Theta \subset \mathbb{R}^d$ and a smooth mapping $\psi : \Theta \mapsto \mathbb{R}^n$, where the embedding space $\mathbb{R}^n$ obeys $d < n$. We speak of the image $M = \psi(\Theta)$ as the manifold, although of course from the viewpoint of manifold theory it is actually the very special case of a single coordinate patch.

The vector θ can be thought of as some control parameters underlying a measuring device, and the manifold as the enumeration $m = \psi(\theta)$ of all possible measurements as the parameters vary. Thus the mapping ψ associates parameters to measurements.

In such a setting, we are interested in obtaining data examples $m_i, i = 1, ..., N$ showing (we assume) the results of measurements with many different choices of control parameters (θ_i, $i = 1, ..., N$.) We will speak of M as the data manifold, i.e. the manifold on which our data m_i must lie. In this paper, we consider only the situation where all data points m_i lie exactly in the manifold M.

There are several concrete situations related to image analysis and acoustics where this abstract model may apply:

- *Scene Variation* – Pose variations, facial gesturing.

- *Imaging Variations* – Changes in the position of lighting sources and changes in the spectral composition of lighting color.

- *Acoustic Articulations* – Changes in distance from source to receiver. Changes in position of the speaker, or in the direction of the speaker's mouth.

In all such situations, there is an underlying parameter controlling articulation of the scene; here are two examples.

- *Facial Expressions* – The tonus of several facial muscles control facial expression; conceptually, a parameter vector θ records the contraction of each of those muscles.

- *Pose Variations* – Several joint angles (shoulder, elbow, wrist, ...) control the combined pose of the elbow-wrist-finger system in combination.

We also speak of M as the *articulation manifold*.

In the above settings, and presumably many others, we can make measurements (m_i), but without having access to the corresponding articulation parameters (θ_i).

It would be interesting to be able to recover the underlying parameters θ_i from the observed points m_i on the articulation manifold. Thus we have the

Parametrization Recovery Problem: *Given a collection of data points (m_i) on an articulation manifold M, recover the mapping ψ and the parameter points θ_i.*

As stated, this is of course ill-posed, since if ψ is one solution, and $\phi : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a morphing of the Euclidean space $\mathbb{R}^d$, the combined mapping $\psi \circ \phi$ is another solution. For this reason, several extra assumptions must be made in order to uniquely determine solutions.

3 Isomap

In an insightful article, Tenenbaum, de Silva, and Langford [1] proposed a method that, under certain assumptions, could indeed recover the underlying parametrization of a data manifold. The assumptions were:

(ISO1) *Isometry.* The mapping ψ preserves geodesic distances. That is, define a distance between two points m and m' on the manifold according to the distance travelled by a bug walking along the manifold M according to the shortest path between m and m'. Then the isometry assumption says that

$$G(m, m') = |\theta - \theta'|, \qquad \forall m \leftrightarrow \theta, m' \leftrightarrow \theta',$$

where $|\cdot|$ denotes Euclidean distance in $\mathbb{R}^d$.

(ISO2) *Convexity.* The parameter space Θ is a convex subset of $\mathbb{R}^d$. That is, if θ, θ' is a pair of points in Θ, then the entire line segment $\{(1-t)\theta + t\theta' : t \in (0,1)\}$ lies in Θ.

Tenenbaum et al. [1] introduced a procedure, Isomap, which, under these assumptions, recovered Θ up to rigid motion. That is, up to a choice of origin and a rotation and possible mirror imaging about that origin, Isomap recovered Θ. In their paper, they gave an example showing successful recovery of articulation parameters from an image database that showed many views of a wrist rotating and a hand opening at various combinations of rotation/opening.

The stated assumptions lead to two associated questions:

(Q1) Whether interesting articulation manifolds have isometric structure; and

(Q2) Whether interesting parameter spaces are truly convex.

Donoho and Grimes [4, 5] studied these questions in the case of image libraries. Namely, they modeled images m as continuous functions $m(x, y)$ defined on the plane $(x, y) \in \mathbb{R}^2$, and focused attention on images in special articulation families defined by certain mathematical models. As one example, they considered images of a ball on a white background, where the underlying articulation parameter is the position of the ball's center. In this model, let B_θ denotes the ball of radius 1 centered at $\theta \in \mathbb{R}^2$, and define

$$m_\theta(x, y) = 1_{B_\theta}(x, y).$$

This establishes a correspondence between $\theta \in \mathbb{R}^2$ and m_θ in $L^2(\mathbb{R}^2)$. After dealing with technicalities associated with having $L^2(\mathbb{R}^2)$ as the ambient space in which M is embedded, they derived expressions for the metric structure induced from $L^2(\mathbb{R}^2)$ and showed that indeed, if Θ is a convex subset of $\mathbb{R}^2$, then isometry holds:

$$G(\theta, \theta') = |\theta - \theta'|, \qquad \forall \theta, \theta' \in \Theta$$

They found that isometry held for a dozen examples of interesting image articulation families, including cartoon faces with articulated eyes, lips, and brows. Hence (Q1) admits of positive answers in a number of interesting cases.

4

On the other hand, in their studies of image articulation families, Donoho and Grimes [4, 5] noted that (Q2) can easily have a negative answer. A simple example occurs with images showing *two* balls which articulate by translation, as in the single-ball case mentioned above, but where the ball centers obey *exclusion:* the two balls never overlap. In this case, the parameter space $\Theta \subset \mathbb{R}^4$ becomes nonconvex; writing $\theta = (\theta^1, \theta^2)$ as a concatenation of the parameters of the two ball centers, we see that it is missing a tube where $|\theta^1 - \theta^2| \leq 1$.

The case of two balls moving independently and subject to exclusion is merely one in a series of examples where the articulation manifold fails to obey (ISO1) and (ISO2), but instead obeys something weaker:

(LocISO1) *Local Isometry.* In a small enough neighborhood of each point m, geodesic distances to nearby points m' in M are identical to Euclidean distances between the corresponding parameter points θ and θ'.

(LocISO2) *Connectedness.* The parameter space Θ is a open connected subset of $\mathbb{R}^d$.

In such settings, the original assumptions of Isomap are violated, and, as shown in [4, 6], the method itself fails to recover the parameter space up to a linear mapping. Donoho and Grimes [6] pointed out the possibility of recovering non-convex Θ by applying Isomap to a suitable decomposition of M into overlapping geodesically convex pieces. However, a fully automatic procedure based on a general principle would be preferable in solving this problem. In this paper we propose such a procedure.

4 The $\mathcal{H}$-Functional

We now set up notation to define the quadratic form $\mathcal{H}(f)$ referred to in the abstract and introduction.

We suppose that $M \subset \mathbb{R}^n$ is a smooth manifold, and so the tangent space $T_m(M)$ is well-defined at each point $m \in M$. Thinking of the tangent space as a subspace of $\mathbb{R}^n$, we can associate to each such tangent space $T_m(M) \subset \mathbb{R}^n$ an orthonormal coordinate system using the inner product inherited from $\mathbb{R}^n$. (It will not matter in the least how this choice of coordinate system varies from point to point in M).

Think momentarily of $T_m(M)$ as an affine subspace of $\mathbb{R}^n$ which is tangent to M at m, with the origin $0 \in T_m(M)$ identified with $m \in M$. There is a neighborhood N_m of m such that each point $m' \in N_m$ has a unique closest point $V' \in T_m(M)$, and such that the implied mapping $m' \mapsto V'$ is smooth. The point in $T_m(M)$ has coordinates given by our choice of orthonormal coordinates for $T_m(M)$. In this way, we obtain local coordinates for a neighborhood N_m of $m \in M$. call them $x_1^{(tan,m)}, \dots, x_d^{(tan,m)}$, where we retain tan, m in the notation to remind us that they depend on the way in which coordinates were defined on $T_m(M)$.

We now use the local coordinates to define the Hessian of a function $f : M \mapsto \mathbb{R}$ which is C^2 near m. Suppose that $m' \in N_m$ has local coordinates $x = x^{(tan,m)}$. Then the rule $g(x) = f(m')$ defines a function $g : U \mapsto \mathbb{R}$, where U is a neighborhood of 0 in $\mathbb{R}^d$. Since the mapping $m' \mapsto x$ is smooth, the function g is C^2 We define the *Hessian of f at m in tangent coordinates* as the ordinary Hessian of g:

$$(H_f^{tan}(m))_{i,j} = \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} g(x)\Big|_{x=0}.$$

In short, at each point m, we use the tangent coordinates and differentiate f in that coordinate system. We call this construction the *tangent Hessian* for short.

We remark that, unfortunately, there is some ambiguity in this definition of the Hessian, because the entries in the Hessian matrix H_f^{tan} depend on the choice of coordinate system on the underlying tangent space $T_m(M)$. More properly, if we consider another choice of orthonormal coordinates in $T_m(M)$, we get another system of local coordinates in N_m, and for that set of local coordinates, we get another Hessian matrix, which can be quite different.

Luckily, it is possible to extract invariant information about f. Comparing two different Hessians, H and H', say, which could be obtained as a result of two different choices in the local coordinate system, we have the relation

$$H' = UHU^T, \tag{1}$$

where U is the orthonormal matrix translating one set of coordinates into the other. Although this says that H and H' can differ a great deal in their entries, it turns out that their sizes must be similar. For a d by d matrix A, let $\|A\|_F = (\sum_{ij} A_{ij}^2)^{1/2}$ denote the usual Frobenius norm of matrices; then for two matrices H and H' obeying (1), we have $\|H'\|_F = \|H\|_F$. Consequently, provided we always restrict attention just to the Frobenius norm of $H_f^{(tan)}$, our recipe gives a well-defined quantity.

We now consider a quadratic form defined on C^2 functions by

$$\mathcal{H}(f) = \int_M \|H_f^{(tan)}(m)\|_F^2 dm,$$

where dm stands for a probability measure on M which has strictly positive density everywhere on the interior of M. $\mathcal{H}(f)$ measures the average 'curviness' of f over the manifold M.

Theorem 1 *Suppose $M = \psi(\Theta)$ where Θ is an open connected subset of $\mathbb{R}^d$, and ψ is a locally isometric embedding of Θ into $\mathbb{R}^n$. Then $\mathcal{H}(f)$ has a $d+1$ dimensional nullspace, consisting of the constant function and a d-dimensional space of functions spanned by the original isometric coordinates.*

We give the proof in Appendix A.

Corollary 2 *Under the same assumptions as Theorem 1, the original isometric coordinates θ can be recovered, up to a rigid motion, by identifying a suitable basis for the null space of $\mathcal{H}(f)$.*

We sketch the argument for the corollary. Fix a point m_0 in M. Recall the orthogonal coordinate system chosen for $T_{m_0}(M)$ gives us a local coordinate system – a set of d functions $x_1^{(tan,m_0)}, ..., x_d^{(tan,m_0)}$ – defined on a neighborhood $N_{m_0} \subset M$. The nullspace of $\mathcal{H}$ is $d+1$ dimensional; consider the subspace $V \subset nullspace(\mathcal{H})$ consisting of those functions f orthogonal to 1 in the $L^2(M, dm)$ inner product. Within V, we can find a linearly independent set of functions $\psi_1,..., \psi_d$ making a basis for V, and such that, if x refers to local coordinates $x^{(tan,m_0)}$ in the vicinity of m_0, then $\psi_j(m) = x_j(m) + o(dist(m, m_0)^2)$. The functions ψ_j provide the required basis of V. The vector function $\psi(m) = (\psi_j(m))_{j=1}^d$ gives an inverse to ϕ, up to a rigid motion, i.e. we have $\psi(\phi(\theta)) = U\theta + \tau$, where the orthonormal matrix U effects orthogonal rotation, and adding τ effects a location shift. Hence, this recovers the original coordinates up to rigid motion under an exact isometric embedding ψ. □

5 Hessian LLE

We now consider the setting where we have sampled data (m_i) lying on M, and we would like to recover the underlying parametrization ψ and underlying parameter settings θ_i, at least up to rigid motion. Theorem 1 and its Corollary suggest the following algorithm for attacking this problem. We model our algorithm structure on the original Locally Linear Embedding algorithm [2]

Hessian LLE Algorithm:

Input: $(m_i : i = 1, ..., N)$ a collection of N points in $\mathbb{R}^n$.

Parameters: d, the dimension of the parameter space; k, the size of the neighborhoods for fitting.

Constraints: $min(k, n) > d$.

Output: $(w_i : i = 1, ..., N)$ a collection of N points in $\mathbb{R}^d$, the recovered parametrization.

Procedure:

- *Identify Neighbors.* For each data point m_i, $i = 1, ...n$, identify the indices corresponding to the k-nearest neighbors in Euclidean distance. Let $\mathcal{N}_i$ denote the collection of those neighbors. For each neighborhood $\mathcal{N}_i$, $i = 1, \ldots, N$, form a $k \times n$ matrix M^i whose rows consist of the re-centered points $m_j - \bar{m}_i$, $j \in \mathcal{N}_i$, where $\bar{m}_i = Ave\{m_j : j \in \mathcal{N}_i\}$.

- *Obtain Tangent Coordinates.* Perform a singular value decomposition of M^i, getting matrices U, D, and V; U is k by $min(k, n)$. The first d columns of U give the tangent coordinates of points in $\mathcal{N}_i$.

- *Develop Hessian Estimator.* Develop the infrastructure for least-squares estimation of the Hessian. In essence, this is a matrix H^i with the property that if f is a smooth function $f : M \mapsto \mathbb{R}$, and $\mathbf{f}_j = (f(m_i))$, then the vector v^i whose entries are obtained from $\mathbf{f}$ by extracting those entries corresponding to points in the neighborhood $\mathcal{N}_i$, then, the matrix vector product $H^i v^i$ gives a $d(d+1)/2$ vector whose entries approximate the entries of the Hessian matrix, $(\frac{\partial f}{\partial U_i \partial U_j})$.

- *Develop Quadratic Form.* Build a symmetric matrix $\mathcal{H}_{ij}$ having, in coordinate pair ij, the entry

$$\mathcal{H}_{i,j} = \sum_l \sum_r ((H^l)_{r,i}(H^l)_{r,j}).$$

  Here by H^l we mean, again, the $d(d+1)/2$ by k matrix associated with estimating the Hessian over neighborhood $\mathcal{N}_l$, where rows r correspond to specific entries in the Hessian matrix and columns i correspond to specific points in the neighborhood.

- *Find Approximate Null Space.* Perform an eigenanalysis of $\mathcal{H}$, and identify the $d + 1$-dimensional subspace corresponding to the $d + 1$ smallest eigenvalues. There will be an eigenvalue 0 associated with the subspace of constant functions; and the next d eigenvalues will correspond to eigenvectors spanning a d-dimensional space $\hat{V}_d$ in which our embedding coordinates are to be found.

- *Find Basis for Null Space.* Select a basis for $\hat{V}_d$ which has the property that its restriction to a specific fixed neighborhood $\mathcal{N}_0$ (the neighborhood may be chosen arbitrarily from those used in the algorithm) provides an orthonormal basis. The given basis has basis vectors $w^1, \ldots w^d$; these are the embedding coordinates.

The algorithm is a straightforward implementation of the idea of estimating tangent coordinates, the tangent hessian and the empirical version of the operator $\mathcal{H}$.

Remarks.

- *Coding Requirements.* This can be implemented easily in MATLAB, Mathematica, S-Plus, R, or similar quantitative programming environment; our Matlab implementation is available at `www-stat.stanford.edu/~carrie`.

- *Storage Requirements.* This is a 'spectral method', and involves solving the eigenvalue problem for an N by N matrix. Although it would appear to require $O(N^2)$ storage, which can be prohibitive, the storage required is actually proportional to $n \cdot N$ – i.e. the storage of the data points; in fact this storage can be kept on disk; the remaining storage is basically proportional to Nk. Note that the matrix $\mathcal{H}$ is a sparse matrix, with about $O(Nk)$ nonzero entries.

- *Computational Complexity.* In effect, the computational cost difference between a sparse and a full matrix using the sparse eigenanalysis implementation in Matlab 6.1 (`eigs.m`: uses Arnoldi methods) depends on the cost of computing a matrix-vector product using the input matrix. For our sparse matrix, the cost of each product is about $2kN$, while for a full matrix the cost is something like $2N^2$, making the overall cost of the sparse version $O(kN^2)$.

- *Building the Hessian Estimator.* Consider first the case $d = 2$. Form a matrix X^i consisting of the following columns:

$$X^i = \begin{bmatrix} 1 & U_{\cdot,1} & U_{\cdot,2} & (U_{\cdot,1}^2) & (U_{\cdot,2}^2) & (U_{\cdot,1} \times U_{\cdot,2}) \end{bmatrix}. \tag{2}$$

In the general case $d > 2$, create a matrix with $1 + d + d(d + 1)/2$ columns; the first $d + 1$ of these consist of a vector of ones, and then the first d columns of U; the last $d(d + 1)/2$ consist of the various cross-products and squares of those d columns. Perform the usual Gram-Schmidt orthonormalization process on the matrix X^i, yielding a matrix $\tilde{X}^i$ with orthonormal columns; then define H^i by extracting the last $d(d + 1)/2$ columns and transposing:

$$(H^i)_{r,l} = (\tilde{X}^i)_{l,1+d+r}.$$

- *Finding the Basis for the Null Space.* Let V be the N by d matrix of eigenvectors built from the nonconstant eigenvectors associated to the $(d+1)$ smallest eigenvalues, and let $V_{l,r}$ denote the l-th entry in the r-th eigenvector of $\mathcal{H}$. Define the matrix $(R)_{rs} = \sum_{j \in N_1} V_{j,r} V_{j,s}$. The desired N by d matrix of embedding coordinates is obtained from:

$$W = V \cdot R^{-1/2}.$$

In Section 7, we apply this recipe to a canonical isometric example.

6   Comparison to LLE / Laplacian Eigenmaps

The algorithm we have described bears substantial resemblance to the LLE procedure proposed by Roweis and Saul [2]. The theoretical framework we have described also bears substantial resemblance to the Laplacian Eigenmap framework of Belkin and Niyogi [3] , only with the Hessian replacing the Laplacian. The Laplacian Eigenmap setup goes as follows: Define the Laplacian operator in tangent coordinates by $\Delta^{(tan)}(f) = \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}$, and define the functional $\mathcal{L}(f) = \int_M (\Delta^{(tan)}(f))^2 \, dm$. This functional computes the average of the Laplacian operator over the manifold; Laplacian Eigenmap methods propose to solve embedding problems by obtaining the $d+1$ lowest eigenvalues of $\mathcal{L}$, and using the corresponding eigenfunctions to embed the data in low-dimensional space. The LLE method is an empirical implementation of the same principle, defining a discrete Laplacian based on a hearest-neighbor graph and embedding scattered n-dimensional data by using the first d nonconstant eigenvectors of the graph Laplacian.

7   Data Example

In this example, we take a random sample (m_i) on the Swiss Roll surface [2] in three dimensions. The resulting surface is like a rolled-up sheet of paper, and so is exactly isometric to Euclidean space (i.e. to a rectangular segment of $\mathbb{R}^2$). Successful results of LLE and Isomap on such data have been previously published [1, 2]. However, here we consider a change in sampling procedure. Instead of sampling parameters in a full rectangle, we sample from a rectangle with a missing rectangular strip punched out of the center. The resulting Swiss Roll is then missing the corresponding strip, and so is not convex (while still remaining connected).

Using this model, and the code provided for Isomap and LLE in [1] and [2], respectively, we test the performance of all three algorithms on a random sample of 600 points in three dimensions. The points were generated using the same code published by Roweis and Saul [2]. The results, as seen in Figure 1, show the dramatic effect that non-convexity can have on the resulting embeddings. Although the data manifold is still locally isometric to Euclidean space, the effect of the missing sampling region is to, in the case of Locally Linear Embedding, make the resulting embedding functions asymmetric and nonlinear with respect to the original parametrization. In the case of Isomap, the non-convexity causes a strong dilation of the missing region, warping the rest of the embedding. Hessian LLE, on the other hand, embeds the result almost perfectly into two-dimensional space.

The computational demands of Locally Linear Embedding algorithms are very different than those of the Isomap distance-processing step. LLE and HLLE are both capable of handling
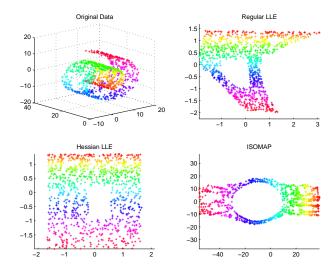
Figure 1: Top left: original data; Top right: LLE embedding (Roweis and Saul code, k=12); Bottom Left: Hessian Eigenmaps (Donoho and Grimes code, k=12); Bottom right: Isomap (Tenenbaum, et al. code, k=7). The underlying correct parameter space that generated the data is a square with a central square removed, similar to what is obtained by the Hessian approach, bottom left.

large N problems, since initial computations are performed only on smaller neighborhoods, while Isomap has to compute a full matrix of graph distances for the initial distance processing step. However, both LLE and HLLE are more sensitive to the dimensionality of the data space, n, because they must estimate a local tangent space at each point. While we introduce an orthogonalization step in HLLE that makes the local fits more robust to pathological neighborhoods than LLE, HLLE still requires effectively a numerical second differencing at each point that can be very noisy at low sampling density.

8   Discussion

We have derived a new Local Linear Embedding algorithm from a conceptual framework that provably solves the problem of recovering a locally isometric parametrization of a manifold M when such a parametrization is possible. The existing Isomap method can solve the same problem in the special case where M admits a globally isometric parametrization. This special case requires that M be geodesically convex, or equivalently that Θ be convex.

Note that in dealing with data points (m_i) sampled from a naturally-occurring manifold M, we can see no reason that the probability measure underlying the sampling must have geodesically convex support. Hence our local isometry assumption seems much more likely to hold in practice than the more restrictive global isometry assumption in Isomap.

Hessian LLE requires the solution of N separate k-by-k eigenproblems and, similar to Roweis and Saul's original LLE algorithm, a single N by N sparse eigenproblem. The sparsity of this eigenproblem can confer a substantial advantage over the general nonsparse eigenproblem. This is an important factor distinguishing LLE techniques from the Isomap technique, which poses a completely dense N by N matrix for eigenanalysis. In our experience, if we budget an

equal programming effort in the two implemntations, the implementation of Hessian LLE can solve much larger-scale data analysis problems (much larger N) than Isomap. (We are aware that there are modifications of the Isomap principle under development which attempt to take advantage of sparse graph structure (i.e. Landmark Isomap); however, the original Isomap principle discussed here is not posable as a sparse eigenproblem).

It is also interesting to contrast Hessian LLE with Roweis and Saul's LLE and Belkin and Niyogi's Laplacian Eigenmaps. In our opinion, Theorem 1 establishes the fundamental correctness of Hessian Eigenmaps in the setting of local isometry; it seems that no similar result can be true for Laplacian Eigenmaps, nor has an alternate conceptual setting been proposed where fundamental correctness has been proven. (I.e. we are not aware of a conceptual framework in which we can agree that there is a 'right answer' – a correct underlying parametrization – and in which it can be proven that Laplacian eigenmaps accurately recover that 'right answer' modulo some linear transformations.) In other respects, the methodologies are similar, and in fact our implementation owes its basic structure to the original Roweis and Saul implementation of LLE [2]. Also as in LLE, the Hessian-based method can perform poorly if bad results are obtained from a few of the tangent space coordinate fits.

A drawback of Hessian LLE versus the other methods just discussed is the Hessian approach requires estimation of second derivatives, and this is known to be numerically noisy or difficult in very high-dimensional data samples.

Our understanding of the phenomena involved in learning parametrizations is certainly not complete, and we still have much to learn about the implementation of spectral methods for particular types of data problems.

A   Appendix: Proof of Theorem 1

Since the mapping ψ is a locally isometric embedding, the inverse mapping $\phi = \psi^{-1} : M \mapsto \mathbb{R}^d$ provides a locally isometric coordinate system on M. Let $\theta_1, \ldots, \theta_d$ denote the isometric coordinates.

Let m be a fixed point in the interior of M and let f be a C^2 function on M. Define the pullback of f to Θ by $g(\theta) = f(\psi(\theta))$. Define the *Hessian in isometric coordinates* of f at m by

$$(H_f^{iso})_{i,j}(m) = \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} g(\theta)\Big|_{\theta = \phi(m)}.$$

This definition postulates knowledge of the underlying isometry ϕ, so the result is definitely not something we expect to be 'learnable' from knowledge of M alone. Nevertheless it provides an important benchmark for comparison. Now define the quadratic form

$$\mathcal{H}^{iso}(f) = \int_M \|H_f^{iso}\|_F^2 dm.$$

This is similar to the quadratic form $\mathcal{H}$, except that it is based on the Hessian in isometric coordinates rather than the Hessian in tangent coordinates. We will explore the nullspace of this quadratic form, and relate it to the nullspace of $\mathcal{H}$.

By its very definition, $\mathcal{H}^{iso}$ has a natural pullback from $L^2(M, dm)$ to $L^2(\Theta, d\theta)$. Indeed, letting $g : \Theta \mapsto \mathbb{R}$ be a function on $\Theta \subset \mathbb{R}^d$ with open interior, letting θ be an interior point, and letting $H_g^{euc}(\theta)$ denote the ordinary Hessian in Euclidean coordinates at θ, we have actually

defined H^{iso} *by* $H_f^{iso} \equiv H_{f^*}^{euc}$, where $f^*(\theta) = f(\psi(\theta))$ is the pullback of f to $\Theta \subset \mathbb{R}^d$. Hence, defining, for functions $g : \Theta \mapsto \mathbb{R}$ the quadratic form

$$\mathcal{H}^{euc}(g) = \int_\Theta \|H_g^{euc}\|_F^2 d\theta$$

(where $d\theta$ and dm are densities in $1-1$ correspondence under the 1-1 correspondence $\theta \leftrightarrow m$), we have

$$\mathcal{H}^{iso}(f) = \mathcal{H}^{euc}(f^*) \ \ \forall f \in C^\infty(M).$$

It follows that the nullspace of $\mathcal{H}^{iso}$ is in 1-1 correspondence with the nullspace of $\mathcal{H}^{euc}$ under the pullback by ψ. Consider then the nullspace of $\mathcal{H}^{euc}$.

Lemma 3 *Let $\Theta \subset \mathbb{R}^d$ be connected with open interior, and let $d\theta$ denote a strictly positive density on the interior of Θ. Let $W_2^2(\Theta)$ denote the usual Sobolev space of functions on Θ which have finite L^2 norm and whose first two distributional derivatives exist and belong to L^2. Then $\mathcal{H}^{euc}$, viewed as a functional on the linear space $W_2^2(\Theta)$, has a $d+1$-dimensional nullspace consisting of the span of the constant function together with the d coordinate functions θ_1, ..., θ_d.*

Proof of Lemma 3. It is of course obvious that the nullspace contains the span of the constant function and all the coordinate functions, since this span is simply all linear functions and since linear functions have everywhere vanishing Hessians. In the other direction, we show that the nullspace contains only these functions. Consider any function g in $C^\infty(\Theta)$ which is not exactly linear. Then there must be some second-order mixed derivative $\frac{\partial g}{\partial \theta_{i_1} \partial \theta_{i_2}}$ and which is non-vanishing on some ball: $\|\frac{\partial g}{\partial \theta_{i_1} \partial \theta_{i_2}}\|_{L^2(\Theta, d\theta)} > 0$. But

$$
\begin{aligned}
\mathcal{H}^{euc}(g) &= \int_\Theta \|H_g\|_F^2 d\theta \\
&= \int_\Theta \sum_{i,j} (\frac{\partial g}{\partial \theta_{i_1} \partial \theta_{i_2}})^2 d\theta = \sum_{i,j} \int_\Theta (\frac{\partial g}{\partial \theta_{i_1} \partial \theta_{i_2}})^2 d\theta \\
&= \sum_{i,j} \|\frac{\partial g}{\partial \theta_{i_1} \partial \theta_{i_2}}\|_{L^2(\Theta, d\theta)}^2 \geq \|\frac{\partial g}{\partial \theta_{i_1} \partial \theta_{i_2}}\|_{L^2(\theta, \theta_0)}^2 \quad > 0.
\end{aligned}
$$

Hence no smooth nonlinear function can belong to the nullspace of $\mathcal{H}^{euc}$. The openness of the interior of Θ implies that $C^\infty(\Theta)$ is dense in $W_2^2(\Theta)$, and so we can reach the same conclusion for all g in $W_2^2(\Theta)$. $\qquad \square$

By pullback, Lemma 3 immediately implies:

Corollary 4 *Viewed as a functional on $W_2^2(M)$, $\mathcal{H}^{iso}$ has a $d+1$-dimensional nullspace, consisting of the span of the constant functions and the d isometric coordinates $\theta_i(m)$ on M.*

We now show the same for the object of our original interest: $\mathcal{H}$. This follows immediately from the following lemma:

Lemma 5 *Let f be a function in $C^\infty(M)$ and let ψ be a local isometry between Θ and M. Then at every $m \in interior(M)$,*

$$\|H_f^{tan}(m)\|_F = \|H_f^{iso}(m)\|_F.$$

The lemma implies that

$$\mathcal{H}(f) = \mathcal{H}^{iso}(f) \quad \forall f \in C^\infty(M)$$

and allows us to see that $\mathcal{H}$ and $\mathcal{H}^{iso}$ have the same nullspace. Theorem 1 follows.

Proof of Lemma 5. Recall that $\|H_f^{tan}(m)\|_F$ is unchanged by variations in the choice of orthonormal basis for the tangent plane $T_m(M)$. Recall that $\phi = \psi^{-1}$ is a local isometry, and gives a coordinate system $\theta_1, \ldots, \theta_d$ on M and therefore induces a choice of coordinate system on $T_m(M)$ that is orthonormal because ϕ is a local isometry. Therefore we may assume that our choice of orthonormal basis for $T_m(M)$ is exactly the same as the choice induced by ϕ. Once this choice has been made, Lemma 5 follows if we can show that:

$$H_f^{tan}(m) = H_f^{iso}(m). \tag{A1}$$

This will follow if we show that for every vector $v \in T_m(M)$,

$$v' H_f^{tan}(m) v = v' H_f^{iso}(m) v. \tag{A2}$$

Given a vector $v \in T_m(M)$, let $\gamma_v : [0, \epsilon) \mapsto M$ denote the unit-speed geodesic in M that starts at $m = \gamma_v(0)$ and that has v for its tangent $\frac{d}{dt}\gamma_v|_{t=0} = v$. Consider the induced function $g_{geo,v}(t) = f(\gamma_v(t))$. Notice that by definition of isometric coordinates

$$g''_{geo,v}(0) = v' H_f^{iso}(m) v.$$

On the other hand, the tangent space $T_m(M)$ provides another local coordinate system for M. Let $\tau_m : T_m(M) \mapsto M$ denote the inverse, that maps from local coordinates back to M; this is the inverse of the mapping $m' \mapsto (x_i^{tan,m}(m'))$. Consider the path $\delta_v : [0, \epsilon) \mapsto T_m(M)$ defined by $d_v(t) = tv$; this corresponds to a path in M defined by $\delta_v(t) = \tau_m(d_v(t))$, i.e. projecting the path in $T_m(M)$ onto a neighborhood of m in M. Consider the induced function $g_{tan,v} = f(\delta_v(t))$. Notice that by definition of Tangent coordinates

$$g''_{tan,v}(0) = v' H_f^{tan}(m) v.$$

Hence, we have to show that

$$g''_{tan,v}(0) = g''_{geo,v}(0) \tag{A3}$$

for all $v \in T_m(M)$. This implies (A2) and hence (A1).

The key observation is that

$$|\delta_v(t) - \gamma_v(t)| = o(t^2), \qquad t \to 0. \tag{A4}$$

It follows that for every Lipschitz f that

$$|g_{tan,v}(t) - g_{geo,v}(t)| = |f(\delta_v(t)) - f(\gamma_v(t))| = o(t^2), \qquad t \to 0,$$

which proves (A3). The key relation (A4) follows by combining two basic facts.

Lemma 6 *Consider a geodesic $\gamma : [0, \epsilon) \mapsto M$ of a Riemannian submanifold M of $\mathbb{R}^n$, and view it as a space curve in $\mathbb{R}^n$. View $T_m(M)$ as a subspace of $\mathbb{R}^n$. The acceleration vector $\frac{d^2}{dt^2}\gamma$ of this space curve at the point $m = \gamma(0)$ is normal to $T_m(M)$.*

13

This lemma is a classic textbook fact about isometric embeddings; See Chapter 6 of [8]. Less well-known is that the same fact is true of $\delta_v(t)$:

Lemma 7 *Viewing $t \mapsto \delta_v(t)$ as a space curve in $\mathbb{R}^n$, its acceleration vector at $t = 0$ is normal to $T_m(M)$.*

In short, the acceleration components of both δ_v and γ_v at 0 reflect merely the extrinsic curvature of M as a curved submanifold of $\mathbb{R}^n$; neither curve has an acceleration component within $T_m(M)$. As both curves by construction have the same tangent at 0, namely v, the key relation (A4) holds, and Lemma 5 is proved. Lemma 7 does not seem to be so well-known as Lemma 6, so we prove it here.

Proof of Lemma 7. Let $c = n - d$ denote the codimension of M in $\mathbb{R}^n$. We can model M in the vicinity of m as the solution of a system of equations:

$$\eta_1(x) = \eta_2(x) = \cdots = \eta_c(x) = 0.$$

Let $t \mapsto x(t)$ be a path in $\mathbb{R}^n$ that starts in M at $t = 0$ and stays in M. Without loss of generality choose coordinates so $x(0) = 0 \in \mathbb{R}^n$. Then

$$\langle \nabla \eta_i, x(t) \rangle + \frac{1}{2}x(t)' H_{\eta_i} x(t) = o(t^2), \quad i = 1, \ldots, c,$$

where $\nabla \eta_i$ denotes the gradient of η_i and H_{η_i} denotes the Hessian of η_i, both evaluated at $x = 0$. Writing $x(t) = tv + (t^2/2)u + o(t^2)$, we conclude that

$$\langle \nabla \eta_i, v \rangle = 0, \quad i = 1, \ldots, c,$$

while

$$t^2 \langle \nabla \eta_i, u \rangle + t^2 v' H_{\eta_i} v = o(t^2), \quad i = 1, \ldots, c.$$

In short,

$$\langle \nabla \eta_i, u \rangle = -v' H_{\eta_i} v \quad i = 1, \ldots, c. \tag{A5}$$

As the $\nabla \eta_i$ span the normal space to M at m, it follows that u must have the specified coordinates in the normal space.

Now, by definition, $\delta_v(t)$ is the closest point in M to the point in the tangent plane represented by tv. But then, viewing $v \in T_m(M)$ as a vector in $\mathbb{R}^n$ we must have that $u \in \mathbb{R}^n$ satisfies

$$\min_u |tv - (tv + (t^2/2)u)| \text{ subject to (A5)}.$$

In short, u solves

$$\min_u |u| \text{ subject to (A5)}.$$

This minimum-norm problem has a unique solution, with u in the span of the vectors $\nabla \eta_i$, i.e. in the normal space perpendicular to $T_m(M)$. Hence the space curve $x(t)$ has no component of acceleration in $T_m(M)$. $\qquad \square$

References

[1] Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000) A global geometric framework for nonlinear dimensionality reduction, *Science*, **290**, pp. 2319–2323.

[2] Roweis, Sam T. & Saul, Lawrence K. (2000) Nonlinear dimensionality reduction by locally linear embedding, *Science*, **290**, pp. 2323–2326.

[3] Belkin, M. & Niyogi, P., Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, *Neural Computation, to appear.*

[4] Donoho, David L. & Grimes, Carrie (2002) When does Isomap recover the natural parametrization of families of articulated images? *Technical Report 2002-27*, Department of Statistics, Stanford University.

[5] Donoho, David L. & Grimes, Carrie (2002) When does geodesic distance recover the true hidden parametrization of families of articulated images? *Proc. European Symposium on Artificial Neural Networks,* **10**, (d-side publications), pp. 199-204.

[6] Donoho, David L. & Grimes, Carrie (2002) Local Isomap perfectly recovers the underlying parametrization for familites of occluded/lacunary images. *Submitted*

[7] Golub, G. & van Loan, C. (1993) *Matrix Computations, 2nd edition*, (Johns Hopkins University Press, Baltimore), pp. 499–507.

[8] Do Carmo, Manfredo Perdigão (1992) in *Riemannian Geometry*, eds. Kadison, Richard V. & Singer, Isidore M. (Birkhäuser, Boston), pp. 124-143.