Unsupervised Learning of Image Manifolds by Semidefinite Programming

Kilian Q. Weinberger and Lawrence K. Saul
Department of Computer and
Information Science
University of Pennsylvania
Email: {kilianw, lsaul}@cis.upenn.edu

***Abstract*— Can we detect low dimensional structure in high dimensional data sets of images and video? The problem of dimensionality reduction arises often in computer vision and pattern recognition. In this paper, we propose a new solution to this problem based on semidefinite programming. Our algorithm can be used to analyze high dimensional data that lies on or near a low dimensional manifold. It overcomes certain limitations of previous work in manifold learning, such as Isomap and locally linear embedding. It also bridges two recent developments in machine learning: semidefinite programming for learning kernel matrices and spectral methods for nonlinear dimensionality reduction. We illustrate the algorithm on easily visualized examples of curves and surfaces, as well as on actual images of faces, handwritten digits, and solid objects.**

I. Introduction

Many data sets of images and video are characterized by far fewer degrees of freedom than the actual number of pixels per image. The problem of dimensionality reduction is to understand and analyze these images in terms of their basic modes of variability—for example, the pose and expression of a human face, or the rotation and scaling of a solid object. Ultimately, this is a problem that must be solved by robust systems for computer vision and pattern recognition [1], [2], [3]. It is also of great interest to researchers in biological vision and computational neuroscience [4].

Mathematically, we can view an image as a point in a high dimensional vector space whose dimensionality is equal to the number of pixels in the image [5], [6]. If the images in a data set are effectively parameterized by a small number of continuous variables, then they will lie on or near a low dimensional *manifold* in this high dimensional space [7]. Though one can imagine other types of hidden structure in ensembles of images (such as clusters [8] or parts [9]), in this paper, we shall focus solely on continuous modes of variability and the unsupervised learning of image manifolds.

Beyond its applications in computer vision, manifold learning is best described as a problem at the intersection of statistics, geometry, and computation. The problem is illustrated in Fig. 1. Given high dimensional data sampled from a low dimensional manifold, how can we efficiently compute a faithful (nonlinear) embedding? In the last few years, researchers have uncovered a large family of algorithms for computing such embeddings from the top or bottom eigenvectors of an appropriately constructed matrix. These algorithms—including Isomap [10], locally linear embedding (LLE) [11], [12], hessian LLE [13], Laplacian eigenmaps [14], and others [15]—can reveal low dimensional manifolds that are not detected by classical linear methods, such as principal component analysis (PCA) [16].

Our main contribution in this paper is a new algorithm for manifold learning based on semidefinite programming. Like Isomap and LLE, it relies on efficient and tractable optimizations that are not plagued by spurious local minima. Interestingly, though, our algorithm is based on a completely different geometric intuition (and optimization), and it overcomes certain well-known limitations of previous work. Our algorithm also reveals an interesting and unexpected connection to recent work on kernel methods in pattern recognition [17].

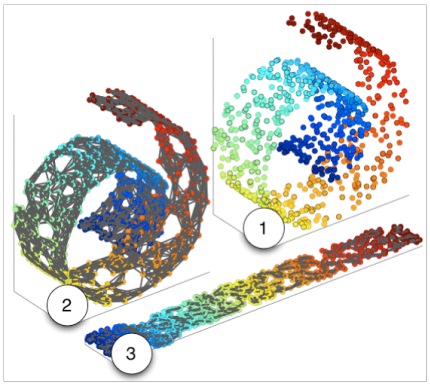

Fig. 1. The problem of manifold learning, as illustrated for $N = 800$ data points sampled from a "Swiss roll". (1). A discretized manifold is revealed by forming the graph that pairwise connects each data point and its $k = 6$ nearest neighbors (2). An unsupervised algorithm must learn the faithful two dimensional embedding that unfolds the Swiss roll while preserving the local geometry of nearby data points (3).

The organization of this paper is as follows. In section II, we review classical methods for linear dimensionality reduction, then introduce the particular class of nonlinear transformations that we consider for unsupervised learning of image manifolds. In section III, we show how to formulate manifold learning as a highly tractable problem in semidefinite programming; this leads to a simple algorithm, called *semidefinite embedding* (SDE), for analyzing high dimensional data that lies on or near a low dimensional manifold. In section IV, we present

results of the algorithm on several data sets, including easily visualized examples of curves and surfaces, as well as images of faces, handwritten digits, and solid objects. In section V, we contrast our algorithm with previous approaches in manifold learning and nonlinear dimensionality reduction. Finally, in section VI, we conclude by describing several directions for future work.

II. Dimensionality Reduction

We study dimensionality reduction as a problem in unsupervised learning. Given N high dimensional inputs $\vec{X}_i \in \mathcal{R}^D$ (where $i = 1, 2, \ldots, N$), the problem is to compute outputs $\vec{Y}_i \in \mathcal{R}^d$ in one-to-one correspondence with the inputs that provide a faithful embedding in $d < D$ dimensions. By "faithful", we mean that nearby points remain nearby and that distant points remain distant; we shall make this intuition more precise in what follows. Ideally, an unsupervised learning algorithm should also estimate the intrinsic dimensionality d of the manifold sampled by the inputs $\vec{X}_i$.

Our algorithm for manifold learning builds on classical methods for dimensionality reduction. We therefore begin by briefly reviewing the linear methods of principal component analysis (PCA) [16] and metric multidimensional scaling (MDS) [18]. The generalization from subspaces to manifolds is then made by introducing the idea of local isometry.

A. Linear Methods

PCA and MDS are based on simple geometric intuitions. In PCA, the inputs are projected into the lower dimensional subspace that maximizes the projected variance; the basis vectors of this subspace are given by the top eigenvectors of the $D \times D$ covariance matrix, $C = \frac{1}{N} \sum_i \vec{X}_i \vec{X}_i^T$. (Here and in what follows, we assume without loss of generality that the inputs are centered on the origin: $\sum_i \vec{X}_i = \vec{0}$.)

In MDS with classical scaling, the inputs are projected into the subspace that best preserves their pairwise squared distances $|\vec{X}_i - \vec{X}_j|^2$ or, more precisely, their dot products $\vec{X}_i \cdot \vec{X}_j$. The outputs of MDS are computed from the top eigenvectors of the $N \times N$ Gram matrix of dot products, with elements $G_{ij} = \vec{X}_i \cdot \vec{X}_j$. Note that a set of vectors is determined up to rotation by its Gram matrix of dot products.

Though based on somewhat different geometric intuitions, PCA and MDS yield the same results—essentially a rotation of the inputs followed by a projection into the subspace with the highest variance. The correlation matrix of PCA and the Gram matrix of MDS have the same rank and eigenvalues up to a constant factor. Both matrices are semipositive definite, and gaps in their eigenvalue spectra indicate that the high dimensional inputs $X_i \in \mathcal{R}^D$ lie to a good approximation in a lower dimensional subspace of dimensionality d, where d is the number of appreciably positive eigenvalues. These linear methods for dimensionality reduction generate faithful embeddings when the inputs are mainly confined to a low dimensional subspace; in this case, their eigenvalues also reveal the correct underlying dimensionality. They do not generally succeed, however, in the case that the inputs lie on a low dimensional manifold.

B. From Subspaces to Manifolds

We will refer to any method that computes a low dimensional embedding from the eigenvectors of an appropriately constructed matrix as a method in *spectral embedding*. If PCA and MDS are linear methods in spectral embedding, what are their nonlinear counterparts? In fact, there are several, most of them differing in the geometric intuition they take as starting points and in the generalizations of linear transformations that they attempt to discover.

The nonlinear method we propose in this paper is based fundamentally on the notion of *isometry*. (For the sake of exposition, we defer a discussion of competing nonlinear methods based on isometries [10], [13] to section V.) Formally, two Riemannian manifolds are said to be isometric if there is a diffeomorphism such that the metric on one pulls back to the metric on the other. Informally, an isometry is a smooth invertible mapping that looks locally like a rotation plus translation, thus preserving distances along the manifold. Intuitively, for two dimensional surfaces, the class of isometries includes whatever physical transformations one can perform on a sheet of paper without introducing holes, tears, or self-intersections. Many interesting image manifolds are isometric to connected subsets of Euclidean space [19].

Isometry is a relation between manifolds, but we can extend the notion in a natural way to data sets. Consider two data sets $X = \{\vec{X}_i\}_{i=1}^N$ and $Y = \{\vec{Y}_i\}_{i=1}^N$ that are in one-to-one correspondence. Let the $N \times N$ binary matrix η indicate a neighborhood relation on X and Y, such that we regard $\vec{X}_j$ as a neighbor of $\vec{X}_i$ if and only if $\eta_{ij} = 1$ (and similarly, for $\vec{Y}_j$ and $\vec{Y}_i$). We will say that *the data sets X and Y are locally isometric under the neighborhood relation η if for every point $\vec{X}_i$, there exists a rotation and translation that maps $\vec{X}_i$ and its neighbors precisely onto $\vec{Y}_i$ and its neighbors.*

We can translate the above definition into various sets of equality constraints on X and Y. To begin, note that the local mapping between neighborhoods will exist if and only if the distances and angles between points and their neighbors are preserved. Thus, whenever both $\vec{X}_j$ and $\vec{X}_k$ are neighbors of $\vec{X}_i$ (that is, $\eta_{ij}\eta_{ik} = 1$), for local isometry we must have that:

$$\left(\vec{Y}_i - \vec{Y}_j\right) \cdot \left(\vec{Y}_i - \vec{Y}_k\right) \;=\; \left(\vec{X}_i - \vec{X}_j\right) \cdot \left(\vec{X}_i - \vec{X}_k\right). \qquad (1)$$

Eq. (1) is sufficient for local isometry because the triangle formed by any point and its neighbors is determined up to rotation and translation by specifying the lengths of two sides and the angle between them. In fact, such a triangle is similarly determined by specifying the lengths of all its sides. Thus, we can also say that X and Y are locally isometric under η if whenever $\vec{X}_i$ and $\vec{X}_j$ are themselves neighbors (that is, $\eta_{ij} = 1$) or are common neighbors of another point in the data set (that is, $[\eta^T \eta]_{ij} > 0$), we have:

$$\left|\vec{Y}_i - \vec{Y}_j\right|^2 \;=\; \left|\vec{X}_i - \vec{X}_j\right|^2. \qquad (2)$$

This is an equivalent characterization of local isometry as eq. (1), but expressed only in terms of pairwise distances. Finally, we can express these constraints purely in terms of dot products. Let $G_{ij} = \vec{X}_i \cdot \vec{X}_j$ and $K_{ij} = \vec{Y}_i \cdot \vec{Y}_j$ denote the Gram matrices of the inputs and outputs, respectively. We can rewrite eq. (2) as:

$$K_{ii} + K_{jj} - K_{ij} - K_{ji} \;=\; G_{ii} + G_{jj} - G_{ij} - G_{ji}. \qquad (3)$$

Eq. (3) expresses the conditions for local isometry purely in terms of Gram matrices; it is in fact this formulation that will form the basis of our algorithm for manifold learning.

III. Semidefinite Embedding

We can now formulate the problem of manifold learning more precisely, taking as a starting point the notion of local isometry. In particular, given N inputs $\vec{X}_i \in \mathcal{R}^D$ and a prescription for identifying "neighboring" inputs, can we find N outputs $\vec{Y}_i \in \mathcal{R}^d$, where $d < D$, such that the inputs and outputs are locally isometric, or at least approximately so? Alternatively, we can state the problem in terms of Gram matrices: can we find a Gram matrix K_{ij} that satisfies the constraints in eq. (3), and for which the vectors $\vec{Y}_i$ (which are determined up to a rotation by the elements of the Gram matrix) lie in a subspace of dimensionality $d < D$, or at least approximately lie in such a subspace? In this section, we show how this can be done by a constrained optimization over the cone of semidefinite matrices.

Like PCA and MDS, the algorithm we propose for manifold learning is based on a simple geometric intuition. Imagine each input $\vec{X}_i$ as a steel ball that is connected to its k nearest neighbors by rigid rods. The effect of the rigid rods is to fix the distances and angles between nearest neighbors, no matter what other forces are applied to the inputs. Now imagine that the inputs are pulled apart, maximizing their total variance subject to the constraints imposed by the rigid rods. Fig. 1 shows the unraveling effect of this transformation on inputs sampled from the Swiss roll. The goal of this section is to formalize the steps of this transformation—in particular, the constraints that must be satisfied by the final solution, and the nature of the optimization that must be performed.

A. Constraints

The constraints that we need to impose for local isometry are naturally represented by a graph with N nodes, one for each input. Consider the graph formed by connecting each input to its k nearest neighbors, where k is a free parameter of the algorithm. For simplicity, we assume that the graph formed in this way is connected; if not, then each connected component should be analyzed separately. The constraints for local isometry under this neighborhood relation are simply to preserve the lengths of the edges in this graph, as well as the angles between edges at the same node. In practice, it is easier to deal only with constraints on distances, as opposed to angles. To this end, let us further connect the graph by adding edges between the neighbors of each node (if they do not already exist). Now by preserving the distances of all edges in this new graph (see Fig. 2), we preserve both the distances of edges and the angles between edges in the original graph—because if all sides of a triangle are preserved, so are its angles.

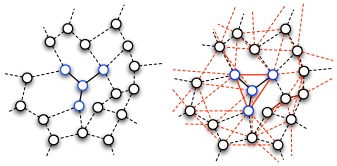

Fig. 2. In the left graph, each node is connected to its nearest neighbors; in the right graph, direct connections are also added between the neighbors. (The original and added edges involving just the middle node are shown in bold.) Preserving the distances of edges in the right graph is equivalent to preserving the distances of edges and the angles between edges in the left graph.

In addition to imposing the constraints represented by the "neighborhood graph", we also constrain the outputs $\vec{Y}_i$ to be centered on the origin:

$$\sum_i \vec{Y}_i = \vec{0}. \qquad (4)$$

Eq. (4) simply removes a translational degree of freedom from the final solution. The centering constraint can be expressed in terms of the Gram matrix K_{ij} as follows:

$$0 = \left|\sum\nolimits_i \vec{Y}_i\right|^2 = \sum_{ij} \vec{Y}_i \cdot \vec{Y}_j = \sum_{ij} K_{ij}. \qquad (5)$$

Note that eq. (5) is a linear equality constraint on the elements of the output Gram matrix, just like eq. (3).

Because the geometric constraints on the outputs $\vec{Y}_i$ are so naturally expressed in terms of the Gram matrix K_{ij} (and because the outputs are determined up to rotation by their Gram matrix), we may view manifold learning as an optimization over Gram matrices K_{ij} rather than vectors $\vec{Y}_i$. Not all matrices, however, can be interpreted as Gram matrices: only symmetric matrices with nonnegative eigenvalues can be interpreted in this way. Thus, we must further constrain the optimization to the cone of semidefinite matrices [20].

In sum, there are three types of constraints on the Gram matrix K_{ij}, arising from local isometry, centering, and semidefiniteness. The first two involve linear equality constraints; the last one is not linear, but importantly it is *convex*. We will exploit this property in what follows. Note that there are $O(Nk^2)$ constraints on $O(N^2)$ matrix elements, and that the constraints are not incompatible, since at the very least they are satisfied by the input Gram matrix G_{ij} (assuming, as before, that the inputs $\vec{X}_i$ are centered on the origin).

B. Optimization

What function of the Gram matrix can we optimize to "unfold" a manifold, as in Fig. (1)? As motivation, consider the ends of a piece of string, or the corners of a flag. Any slack in the string serves to decrease the (Euclidean) distance between its two ends; likewise, any furling of the flag serves to bring its corners closer together. More generally, we observe that any "fold" between two points on a manifold serves to decrease

the Euclidean distance between the points. This suggests an optimization that we can perform to compute the outputs $\vec{Y}_i$ that unfold a manifold sampled by inputs $\vec{X}_i$. In particular, we propose to maximize the sum of pairwise squared distances between outputs:

$$\mathcal{T}(Y) = \frac{1}{2N} \sum_{ij} \left| \vec{Y}_i - \vec{Y}_j \right|^2 . \tag{6}$$

By maximizing eq. (6), we pull the outputs as far apart as possible, *subject to the constraints in the previous section*.

Before expressing this objective function in terms of the Gram matrix K_{ij}, let us verify that it is indeed bounded, meaning that we cannot pull the outputs infinitely far apart. Intuitively, the constraints to preserve local distances (and the assumption that the graph is connected) prevent such a divergence. More formally, let $\eta_{ij}=1$ if $\vec{X}_j$ is one of the k nearest neighbors of $\vec{X}_i$, and zero otherwise, and let τ be the maximal distance between any two such neighbors:

$$\tau = \max_{ij} \left[\eta_{ij} \left| \vec{X}_i - \vec{X}_j \right|^2 \right] . \tag{7}$$

Assuming the graph is connected, then the longest path through the graph has a distance of at most $N\tau$. We observe furthermore that given two nodes, the distance of the path through the graph provides an upper bound on their Euclidean distance. Thus, for all outputs $\vec{Y}_i$ and $\vec{Y}_j$, we must have $|\vec{Y}_i - \vec{Y}_j| < N\tau$. Using this to provide an upper bound on the objective function in eq. (6), we obtain:

$$\mathcal{T}(Y) \;\leq\; \frac{1}{2N} \sum_{ij} (N\tau)^2 \;=\; \frac{N^3\tau^2}{2}. \tag{8}$$

Thus, the objective function cannot increase without bound if we enforce the constraints to preserve local distances.

We can express the objective function in eq. (6) directly in terms of the the Gram matrix K_{ij} of the outputs $\vec{Y}_i$. Expanding the terms on the right hand side, and enforcing the constraint that the outputs are centered on the origin, we obtain:

$$\begin{aligned}
\mathcal{T}(Y) &= \frac{1}{2N} \sum_{ij} \left(\left| \vec{Y}_i \right|^2 + \left| \vec{Y}_j \right|^2 + 2\vec{Y}_i \cdot \vec{Y}_j \right), && (9)\\
&= \sum_i \left| \vec{Y}_i \right|^2 , && (10)\\
&= \sum_i K_{ii}, && (11)\\
&= \mathrm{Tr}(K). && (12)
\end{aligned}$$

Thus, we can interpret the objective function for the outputs in several ways: as a sum over pairwise distances in eq. (6), as a measure of variance in eq. (10), or as the trace of their Gram matrix in eq. (12). The second interpretation is reminiscent of PCA, but whereas in PCA we compute the linear projection that maximizes variance, here we compute the locally isometric embedding. Put another way, the objective function for maximizing variance remains the same; we have merely changed the allowed form of the dimensionality reduction. We also emphasize that in eq. (12), we are maximizing the trace, not minimizing it. While a standard relaxation to minimizing the rank [21] of a semidefinite matrix is to minimize its trace, the intuition here is just the opposite: we will obtain a low dimensional embedding by maximizing the trace of the Gram matrix.

Let us now collect the costs and constraints of this optimization. The problem is to maximize the variance of the outputs $\{\vec{Y}_i\}_{i=1}^N$ subject to the constraints that they are centered on the origin and locally isometric to the inputs $\{\vec{X}_i\}_{i=1}^N$. In terms of the input Gram matrix $G_{ij} = \vec{X}_i \cdot \vec{X}_j$, the output Gram matrix $K_{ij} = \vec{Y}_i \cdot \vec{Y}_j$ and the adjacency matrix η_{ij} indicating nearest neighbors, the optimization can be written as:

Maximize $\mathrm{Tr}(K)$ **subject to** $K \succeq 0$, $\sum_{ij} K_{ij} = 0$,
and $\forall ij$ **such that** $\eta_{ij}=1$ **or** $[\eta^T \eta]_{ij}=1$,
$K_{ii}+K_{jj}-K_{ij}-K_{ji} \;=\; G_{ii}+G_{jj}-G_{ij}-G_{ji}.$

This problem is an instance of semidefinite programming (SDP) [20]: the domain is the cone of semidefinite matrices intersected with hyperplanes (represented by equality constraints), and the objective function is linear in the matrix elements. The optimization is bounded above by eq. (8); it is also convex, thus eliminating the possibility of spurious local maxima. There exists a large literature on efficiently solving SDPs, as well as a number of general-purpose toolboxes. The results in this paper were obtained using the SeDuMi toolbox [22] in MATLAB.

C. Spectral Embedding

From the Gram matrix learned by semidefinite programming, we can recover the outputs $\vec{Y}_i$ by matrix diagonalization. Let $V_{\alpha i}$ denote the i^{th} element of the α^{th} eigenvector, with eigenvalue λ_α. Then the Gram matrix can be written as:

$$K_{ij} = \sum_{\alpha=1}^{N} \lambda_\alpha V_{\alpha i} V_{\alpha j}. \tag{13}$$

An N-dimensional embedding that is locally isometric to the inputs $\vec{X}_i$ is obtained by identifying the α^{th} element of the output $\vec{Y}_i$ as:

$$Y_{\alpha i} = \sqrt{\lambda_\alpha} V_{\alpha i}. \tag{14}$$

The eigenvalues of K are guaranteed to be nonnegative. Thus, from eq. (14), a large gap in the eigenvalue spectrum between the d^{th} and $(d+1)^{\text{th}}$ eigenvalues indicates that the inputs lie on or near a manifold of dimensionality d. In this case, a low dimensional embedding that is *approximately* locally isometric is given by truncating the elements of $\vec{Y}_i$. This amounts to projecting the outputs into the subspace of maximal variance, assuming the eigenvalues are sorted from largest to smallest. The quality of the approximation is determined by the size of the truncated eigenvalues; there is no approximation error for zero eigenvalues. The situation is analogous to PCA and

MDS, but here the eigenvalue spectrum reflects the underlying dimensionality of a manifold, as opposed to merely a subspace.

The three steps of the algorithm, which we call Semidefinite Embedding (SDE), are summarized in Table I. In its simplest formulation, the only free parameter of the algorithm is the number of nearest neighbors in the first step (though one can imagine more elaborate schemes). The second step of the algorithm, involving semidefinite programming, is the most computationally intensive. The first and third steps of SDE resemble those of other algorithms for manifold learning, discussed in section V; SDE has rather different properties, however, due to the particular nature of its second step.

(I) Nearest Neighbors	Compute the k nearest neighbors of each input. Form the graph that connects each input to its neighbors, as well as each neighbor to other neighbors of the same input.
(II) Semidefinite Programming	Compute the Gram matrix of the maximum variance embedding that is centered on the origin and preserves the distances of all edges in the neighborhood graph.
(III) Spectral Embedding	Extract a low dimensional embedding from the dominant eigenvectors of the Gram matrix learned by semidefinite programming.

TABLE I

THE THREE STEPS OF SEMIDEFINITE EMBEDDING (SDE).

IV. RESULTS

We used several data sets of curves, surfaces, and images to evaluate the algorithm in Table I for low dimensional embedding of high dimensional inputs.

Fig. 1 shows $N = 800$ inputs sampled off a "Swiss roll" [10]. The inputs to the algorithm had $D = 8$ dimensions, consisting of the three dimensions shown in the figure, plus five extra dimensions[1] filled with low variance Gaussian noise. The bottom plot of the figure shows the unfolded Swiss roll extracted from the Gram matrix learned by semidefinite programming. The top three eigenvectors are plotted, but the variance in the third dimension (shown to scale) is negligible. The eigenvalue spectrum in Fig. (8) reveals two dominant eigenvalues—a major eigenvalue, representing the unwrapped length of the Swiss roll, and a minor eigenvalue, representing its width. (The unwrapped Swiss roll is much longer than it is wide.) The other eigenvalues are nearly zero, indicating that SDE has discovered the correct underlying dimensionality ($d = 2$) of these inputs.

Fig. 3 shows another easily visualized example. The left plot shows $N = 539$ inputs sampled from a trefoil knot in $D = 3$ dimensions; the right plot shows the $d = 2$ embedding discovered by SDE using $k = 4$ nearest neighbors. The color coding reveals that local neighborhoods have been preserved. In this case, the underlying manifold is a one-dimensional curve, but due to the cycle, it can only be represented in Euclidean space by a circle. The eigenvalue spectrum in Fig. (8) reveals two dominant eigenvalues; the rest are essentially zero, indicating the underlying (global) dimensionality ($d = 2$) of the circle.

[1] For $K = 6$ nearest neighbors, the noise in extra dimensions helps to prevent the manifold from "locking up" when it is unfolded subject to the equality constraints in eqs. (1–3).

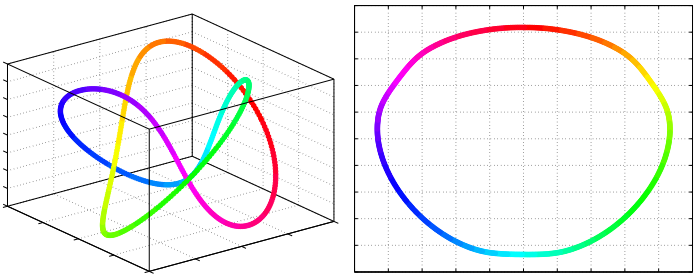

Fig. 3. Left: $N = 539$ inputs sampled along a trefoil knot in $D = 3$ dimensions. Right: $d = 2$ embedding computed by SDE using $k = 4$ nearest neighbors. The color coding shows that the embedding preserves local neighborhoods.

Fig. 4 shows the results of SDE applied to color images of a three dimensional solid object. The images were created by viewing a teapot from different angles in the plane. The images have 76×101 pixels, with three byte color depth, giving rise to inputs of $D = 23028$ dimensions. Though very high dimensional, the images in this data set are effectively parameterized by one degree of freedom—the angle of rotation. SDE was applied to $N = 400$ images spanning 360 degrees of rotation, with $k = 4$ nearest neighbors used to generate a connected graph. The two dimensional embedding discovered by SDE represents the rotating object as a circle—an intuitive result analogous to the embedding discovered for the trefoil knot. The eigenvalue spectrum of the Gram matrix learned by semidefinite programming is shown in Fig. (8); all but the first two eigenvalues are practically zero, indicating the underlying (global) dimensionality ($d = 2$) of the circle.

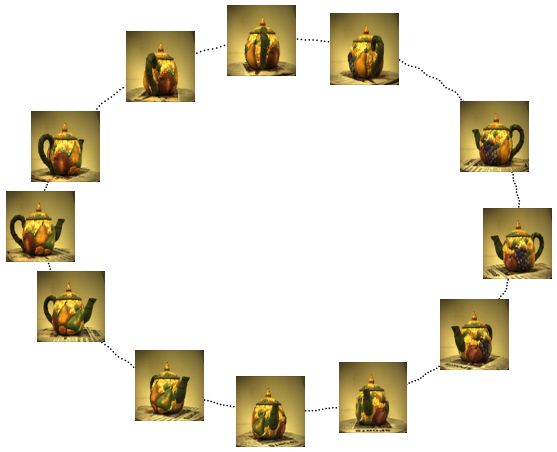

Fig. 4. Two dimensional embedding of $N = 400$ images of a rotating teapot, obtained by SDE using $k = 4$ nearest neighbors. For this experiment, the teapot was rotated 360 degrees; the low dimensional embedding is a full circle. A representative sample of images are superimposed on top of the embedding.

Fig. 5 was generated from the same data set of images; however, for this experiment, only $N = 200$ images were

used, sampled over 180 degrees of rotation. In this case, the eigenvalue spectrum from SDE detects that the images lie on a one dimensional curve (see Fig. 8), and the $d=1$ embedding in Fig. 5 orders the images by their angle of rotation.

Fig. 6 shows the results of SDE on a data set of $N=1000$ images of faces. The images contain different views and expressions of the same face. The images have 28×20 grayscale pixels, giving rise to inputs with $D=560$ dimensions. The plot in Fig. 6 shows the first two dimensions of the embedding discovered by SDE, using $k=4$ nearest neighbors. Interestingly, the eigenvalue spectrum in Fig. 8 indicates that most of the variance of the spectral embedding is contained in the first three dimensions.

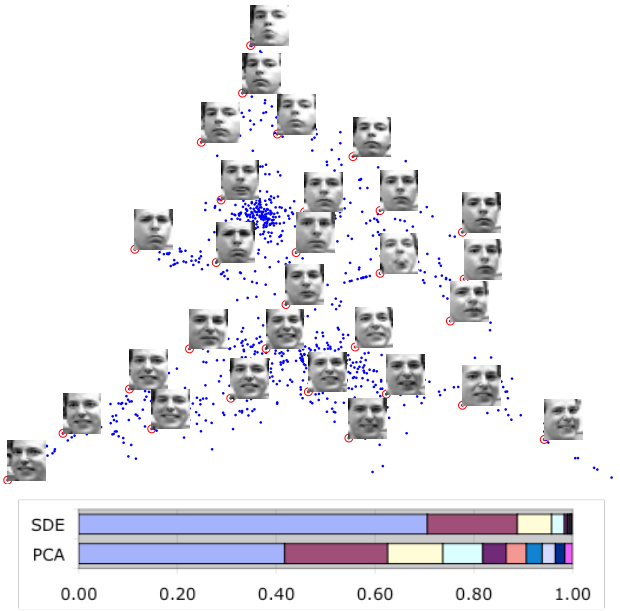

Fig. 6. *Top:* two dimensional embedding of $N=1000$ images of faces, obtained by SDE using $k=4$ nearest neighbors. Representative faces are shown next to circled points. *Bottom:* eigenvalues of SDE and PCA on this data set, indicating their estimates of the underlying dimensionality. The eigenvalues are shown as a percentage of the trace of the output Gram matrix for SDE and the trace of the input Gram matrix for PCA. The eigenvalue spectra show that most of the variance of the nonlinear embedding is confined to many fewer dimensions than the variance of the linear embedding.

Fig. 7 shows the results of SDE applied to another data set of images. In this experiment, the images were a subset of $N=953$ handwritten TWOS from the USPS data set of handwritten digits [23]. The images have 16×16 grayscale pixels, giving rise to inputs with $D=256$ dimensions. Intuitively, one would expect these images to lie on a low dimensional manifold parameterized by such features as size, slant, and line thickness. Fig. 7 shows the first two dimensions of the embedding obtained from SDE, with $k=4$ nearest neighbors. The eigenvalue spectrum in Fig. 8 indicates a latent dimensionality significantly larger than two, but still much smaller than the actual number of pixels.

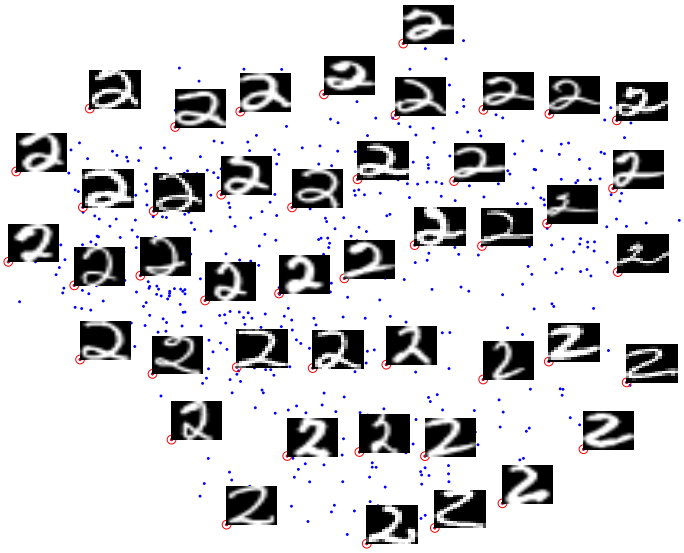

Fig. 7. Results of SDE using $k=4$ nearest neighbors on $N=953$ images of handwritten TWOS. Representative images are shown next to circled points.

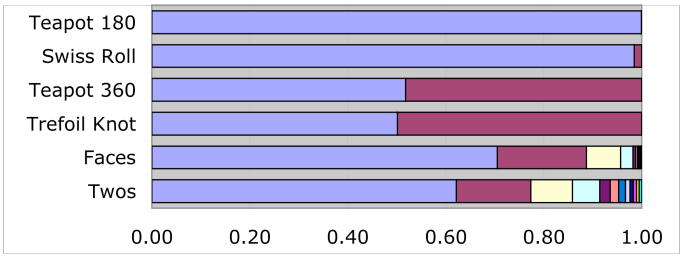

Fig. 8. Eigenvalue spectra from SDE on the data sets in this paper. The eigenvalues are shown as a percentage of the trace of the Gram matrix learned by semidefinite programming. SDE unambiguously identifies the correct underlying dimensionality of the Swiss roll, trefoil knot, and teapot data sets. The images of faces and handwritten digits give rise to many fewer non-zero eigenvalues than the actual number of pixels.

V. Related Work

Our work bridges two recent developments in machine learning: spectral methods for nonlinear dimensionality reduction and semidefinite programming for learning kernel matrices. We discuss each in turn.

A. Manifold Learning

The last few years have witnessed a number of developments in spectral methods for manifold learning. Recently proposed algorithms include Isomap [10], locally linear embedding (LLE) [11], [12], hessian LLE (hLLE) [13], and Laplacian eigenmaps [14]; there are also related algorithms for clustering [24], [25]. All these algorithms share the same basic structure as SDE, consisting of three steps: (i) computing neighborhoods in the input space, (ii) constructing a square matrix with as many rows as inputs, and (iii) spectral embedding via the top or bottom eigenvectors of this matrix. SDE is based on a rather different geometric intuition, however, and as a result, it has different properties.

Fig. 5. One dimensional embedding of $N = 200$ images of a rotating teapot, obtained by SDE using $k = 4$ nearest neighbors. For this experiment, the teapot was only rotated 180 degrees. Representative images are shown ordered by their location in the embedding.

Table II compares these algorithms. Each algorithm attempts to estimate and preserve a different geometric signature of the manifold sampled by the inputs. Isomap estimates geodesic distances between inputs; LLE estimates the coefficients of local linear reconstructions; hLLE and Laplacian eigenmaps estimate the Hessian and Laplacian on the manifold, respectively; SDE estimates local angles and distances. Of these algorithms, only Isomap, hLLE, and SDE attempt to learn isometric embeddings; they are therefore the easiest to compare (since they seek the same solution, up to rotation and scaling). The results on the data set in Fig. 9 reveal some salient differences between these algorithms. While SDE and hLLE reproduce the original inputs up to isometry, Isomap fails in this example because the sampled manifold is not isometric to a *convex* subset of Euclidean space. (This is a key assumption of Isomap, one that is not satisfied by many image manifolds [19].) Moreover, comparing the eigenvalue spectra of the algorithms, only SDE detects the correct underlying dimensionality of the inputs; Isomap is foiled by non-convexity, while the eigenvalue spectra of LLE and hLLE do not reveal this type of information [12], [13].

Overall, the different algorithms for manifold learning in Table II should be viewed as complementary; each has its own advantages and disadvantages. LLE, hLLE, and Laplacian eigenmaps construct sparse matrices, and as a result, they are easier to scale to large data sets. On the other hand, their eigenvalue spectra do not reliably reveal the underlying dimensionality of sampled manfolds, as do Isomap and SDE. There exist rigorous proofs of asymptotic convergence for Isomap [19], [26] and hLLE [13], but not for the other algorithms. On the other hand, SDE by its very nature provides finite-size guarantees that its constraints will lead to locally isometric embeddings. We are not aware of any finite-size guarantees provided by the other algorithms, and indeed, the Hessian estimation in hLLE relies on numerical differencing, which can be problematic for small sample sizes. Finally, while the different algorithms have different computational bottlenecks, it is fair to say that SDE is the most computationally demanding. Scaling SDE to larger data sets will likely require special-purpose solvers for its instance of semidefinite programming. This is an important direction for future work.

B. Kernel Methods

Along with the growing interest in manifold learning, the last few years have also witnessed an explosion of interest in kernel methods for pattern recognition [17]. Kernel methods rely on an implicit mapping of inputs to a higher (and potentially infinite) dimensional feature space. The kernel

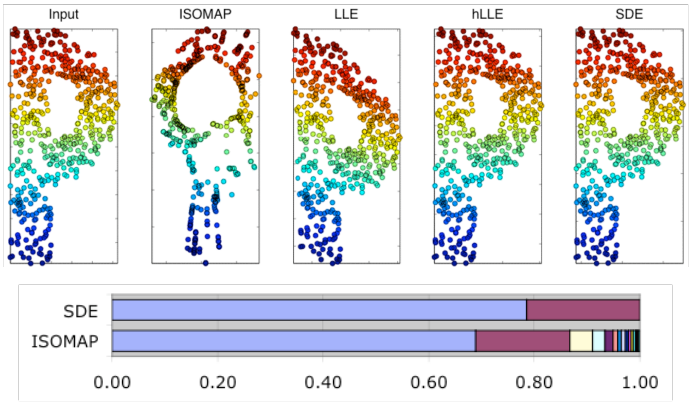

Fig. 9. *Top:* embedding of a non-convex two dimensional data set ($N = 500$) by different algorithms for manifold learning. Isomap, LLE, and hLLE were run with $k = 10$ nearest neighbors; SDE, with $k = 5$ nearest neighbors. Only hLLE and SDE reproduce the original inputs up to isometry. *Bottom:* only SDE has an eigenvalue spectrum that indicates the correct underlying dimensionality ($d = 2$).

Algorithm	Matrix	Mapping	Signature
Isomap	dense	isometric	geodesic distances
SDE	dense	isometric	local distances
LLE	sparse	conformal	local angles
hLLE	sparse	isometric	Hessian
Laplacian eigenmaps	sparse	proximity preserving	discrete Laplacian

TABLE II

COMPARISON OF MANIFOLD LEARNING ALGORITHMS IN TERMS OF THE MATRICES THEY COMPUTE, THE MAPPINGS THEY LEARN, AND THE GEOMETRIC SIGNATURES THEY EXPLOIT.

function specifies the dot product between the feature vectors formed in this way from the original inputs. The "kernel trick" is to replace the dot products $\vec{X} \cdot \vec{X}'$ that appear in linear algorithms for pattern recognition by the kernel function $K(\vec{X}, \vec{X}')$. Support vector machines [27] for classification and kernel PCA [28] for nonlinear dimensionality reduction are examples of algorithms that were conceived in this way, with a kernel matrix that stores the pairwise dot products between inputs in feature space. In most kernel machines, the kernel function is simply specified a priori; the most popular choices involve polynomial and Gaussian kernels. Recent work in supervised learning, however, has investigated the possibility of learning kernel matrices by semidefinite programming [29].

The Gram matrix K_{ij} learned by SDE can be viewed as a kernel matrix between inputs. While there have been attempts to interpret the matrices constructed by Isomap and LLE as

kernels [17], [30], [31], the interpretation for SDE is arguably the most straightforward. The kernel matrix learned by SDE is interesting in several respects. First, it is based on variance maximization, as opposed to margin maximization [17], [29]; the former applies to unsupervised learning, whereas the latter requires (at least some) labeled examples. Second, whereas most kernel functions are chosen to map the inputs into a higher dimensional feature space, the kernel matrix learned by SDE does just the opposite, typically mapping the inputs into a lower dimensional space. Finally, SDE may be viewed as a special version of kernel PCA [28]—ideally suited for manifold discovery—in which the kernel matrix itself is learned (in a completely unsupervised manner) from unlabeled examples.

VI. Discussion

Our initial results for SDE seem promising. SDE has different properties than algorithms such as Isomap and LLE, and many of these properties can be construed as advantages. Like Isomap (and unlike LLE), its eigenvalue spectrum reveals the underlying dimensionality of sampled manifolds; unlike Isomap, however, it does not assume that the inputs are isometric to a convex subset of Euclidean space. To our knowledge, SDE is also the first algorithm for manifold learning based on semidefinite programming.

There are many important directions for future work. Perhaps the most urgent is the investigation of faster methods for solving the semidefinite program in SDE. This study used a generic solver that did not exploit the special structure of the constraints. A specialized solver should allow us to scale SDE up to larger data sets and larger neighborhood sizes. Also, we can relax the constraints in eqs. (1–3) without altering the basic structure of the semidefinite program. Introducing slack variables to relax these constraints may improve the robustness of the algorithm on noisy data sets.

As has been done for Isomap [10], [19], [26] and hLLE [13], it would be desirable to formulate SDE in the continuum limit and to construct rigorous proofs of asymptotic convergence. Such theoretical results would almost certainly provide additional insight into the behavior of the algorithm.

Other directions for future work include the use of SDE kernels in support vector machines [27], the investigation of image manifolds with different topologies [32] (such as those isometric to low dimensional spheres or torii), and the extrapolation of SDE kernels to out-of-sample inputs [30]. Indeed, to the extent that SDE provides a new connection between work in manifold learning and kernel methods, we hope it will lead to further advances in both areas.

References

[1] J. K.-C. L. and Ming-Hsuan Yangand D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, vol. 1, 2003, pp. 313–320.

[2] R. Pless and I. Simon, "Using thousands of images of an object," in *Proceedings of the Sixth Joint Conference on Information Science*, 2002, pp. 684–687.

[3] R. Pless, "Image spaces and video trajectories: Using Isomap to explore video sequences," in *Proceedings of the Ninth International Conference on Computer Vision (ICCV 2003)*, 2003, pp. 1433–1440.

[4] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, pp. 2268–2269, 2000.

[5] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3(1), pp. 71–86, 1991.

[6] D. Beymer and T. Poggio, "Image representation for visual learning," *Science*, vol. 272, p. 1905, 1996.

[7] H. Lu, Y. Fainman, and R. Hecht-Nielsen, "Image manifolds," in *Applications of Artificial Neural Networks in Image Processing III, Proceedings of SPIE*, N. M. Nasrabadi and A. K. Katsaggelos, Eds., vol. 3307. Bellingham, WA: SPIE, 1998, pp. 52–63.

[8] S. Gordon, J. Goldberger, and H. Greenspan, "Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations," in *Proceedings of the Ninth International Conference on Computer Vision (ICCV 2003)*, 2003, pp. 370–377.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[10] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[11] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[12] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[13] D. L. Donoho and C. E. Grimes, "Hessian eigenmaps: locally linear embedding techniques for high-dimensional data," *Proceedings of the National Academy of Arts and Sciences*, vol. 100, pp. 5591–5596, 2003.

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15(6), pp. 1373–1396, 2003.

[15] M. Brand, "Charting a manifold," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003.

[16] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[17] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002.

[18] T. Cox and M. Cox, *Multidimensional Scaling*. London: Chapman & Hall, 1994.

[19] D. L. Donoho and C. E. Grimes, "When does Isomap recover the natural parameterization of families of articulated images?" Department of Statistics, Stanford University, Tech. Rep. 2002-27, August 2002.

[20] L. Vandenberghe and S. P. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38(1), pp. 49–95, March 1996.

[21] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the American Control Conference*, vol. 6, June 2001, pp. 4734–4739.

[22] J. F. Sturm, "Using SeDuMu 1.02, a MATLAB toolbox for optimization overy symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, 1999.

[23] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 16(5), pp. 550–554, May 1994.

[24] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pp. 888–905, August 2000.

[25] A. Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, pp. 849–856.

[26] H. Zha and Z. Zhang, "Isometric embedding and continuum Isomap," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, 2003, pp. 864–871.

[27] V. Vapnik, *Statistical Learning Theory*. N.Y.: Wiley, 1998.

[28] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[29] G. R. G. Lanckriet, N. Christianini, P. L. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semi-definite programming," in *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*, 2002, pp. 323–330.

[30] Y.Bengio, J.-F. Paiement, and P.Vincent, “Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering,” Départmente d’Informatique et Recherche Opérationnelle, Université de Montréal, Tech. Rep., 2003.
[31] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, “A kernel view of the dimensionality reduction of manifolds,” Max-Planck-Institut für Biologische Kybernetik, Tübingen, Tech. Rep. TR-110, July 2003.
[32] R. Pless and I. Simon, “Embedding images in non-flat spaces,” Washington University, Tech. Rep. WU-CS-01-43, December 2001.