A Duality View of Spectral Methods for Dimensionality Reduction

Lin Xiao LXIAO@CALTECH.EDU
Center for the Mathematics of Information, California Institute of Technology, Pasadena, CA 91125, USA

Jun Sun SUNJUN@STANFORD.EDU
Stephen Boyd BOYD@STANFORD.EDU
Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA

Abstract

We present a unified duality view of several recently emerged spectral methods for nonlinear dimensionality reduction, including Isomap, locally linear embedding, Laplacian eigenmaps, and maximum variance unfolding. We discuss the duality theory for the maximum variance unfolding problem, and show that other methods are directly related to either its primal formulation or its dual formulation, or can be interpreted from the optimality conditions. This duality framework reveals close connections between these seemingly quite different algorithms. In particular, it resolves the myth about these methods in using either the top eigenvectors of a dense matrix, or the bottom eigenvectors of a sparse matrix — these two eigenspaces are exactly aligned at primal-dual optimality.

1. Introduction

In many areas of information processing, such as machine learning and data mining, one is often confronted with the problem of dimensionality reduction, i.e., how to extract low dimensional structure from high dimensional data. In a concise mathematical framework, we are given a set of high dimensional data $x_1, \ldots, x_n$ in $\mathbf{R}^d$ (the inputs), and need to compute their "faithful" representations $y_1, \ldots, y_n$ in $\mathbf{R}^r$ (the outputs), with r much smaller than d. Here "faithful" roughly means that nearby inputs are mapped to nearby outputs, while faraway inputs are mapped to faraway outputs (Saul et al., 2005). It is usually assumed that the inputs were sampled from a low dimensional manifold embedded in $\mathbf{R}^d$. An ideal algorithm should be able to estimate the manifold's intrinsic dimension r, as well as to compute the low dimensional representations.

If the sampled data are mainly confined to a linear subspace, then this problem can be well handled by classical techniques such as principle component analysis (PCA) (Jolliffe, 1986) and metric multidimensional scaling (MDS) (Cox & Cox, 1994). Both of them are spectral methods, i.e., methods based on eigenvalue decomposition of either the covariance matrix (for PCA) or the Gram matrix (for MDS) of the input data. For data sampled from general nonlinear manifolds, however, these linear methods do not give satisfactory answers.

Recently, several new spectral methods have been devised to address nonlinear dimensionality reduction: Isomap (Tenenbaum et al., 2000), locally linear embedding (LLE) (Roweis & Saul, 2000), Laplacian eigenmaps (Belkin & Niyogi, 2003), Hessian LLE (Donoho & Grimes, 2003), maximum variance unfolding (MVU) (Weinberger & Saul, 2004; Sun et al., 2005), local tangent space alignment (Zhang & Zha, 2004), and geodesic nullspace analysis (Brand, 2004). Excellent overviews of these methods can be found in Saul et al. (2005) and Burges (2005).

As summarized in Saul et al. (2005), although these new methods share a similar computational structure, they are based on rather different geometric intuitions and intermediate computations. For example, Isomap tries to preserve the global pairwise distances of the input data as measured along the low dimensional manifold (geodesic distances); LLE and Laplacian eigenmaps try to preserve certain local geometric relationships of the data; MVU, on the other hand, preserves local distances but maximize a global objective — the total variance. Computationally, Isomap and MVU construct a dense matrix and use its top eigenvectors (eigenvectors associated with the largest eigenval-

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

ues) in producing the low dimensional representations, while LLE, Laplacian eigenmaps, and Hessian LLE construct a sparse matrix and use its bottom eigenvectors (eigenvectors associated with the smallest eigenvalues). In addition, methods using dense matrices (Gram matrix) can often detect the intrinsic dimension by a tellable gap between a few top eigenvalues and rest of the spectra, but methods using sparse matrices (e.g., Laplacian) do not yield such an estimate since their bottom eigenvalues are usually closely located. In the latter case, an additional step of estimating the intrinsic dimensionality is needed beforehand; see, e.g., Costa and Hero (2004) and references therein.

Each of these spectral methods for dimensionality reduction has its own advantages and disadvantages (Saul et al., 2005), and each can be favorable for different classes of data sets. Nevertheless, these seemingly very different methods are capable of producing quite similar results, at least for some pedagogical examples. In an effort of trying to better understand the connections between these methods, Ham et al. (2004) gave a kernel view of these algorithms, interpreting each of them as an instance of kernel PCA (Schölkopf et al., 1998) on specially constructed kernel matrices.

Our main contribution in this paper is to provide a unified duality view of different spectral methods for nonlinear dimensionality reduction. After a brief review of PCA and MDS in §2, we discuss in §3 the duality theory for the MVU problem (Sun et al., 2005), deriving two equivalent forms of its dual problem and discussing the implications of the optimality conditions. Next we explain how Isomap, LLE and Laplacian eigenmaps fit in the duality framework in §4, §5 and §6, respectively. We follow Saul et al. (2005) for basic descriptions of these algorithms. We show that Isomap is directly related to constructing an approximate optimal solution for the primal MVU problem, Laplacian eigenmaps simply use feasible solutions for the dual MVU problem, and the motivation behind LLE can find interpretation from the primal-dual optimality conditions for the MVU problem. We conclude the paper in §7 with further remarks.

2. PCA and MDS

In this section we briefly review PCA and MDS, as they are building blocks of other spectral methods. We emphasize their geometric intuitions that will be reminiscent in other methods. For convenience, we assume the inputs are centered at the origin, i.e., $\sum_i x_i = 0$.

PCA projects the inputs x_i onto a r-dimensional subspace that minimizes the approximation error. In other words, we need to find a projection matrix P of rank $r < d$ that solves the least-square problem

$$\text{minimize} \quad \textstyle\sum_{i=1}^n \|x_i - Px_i\|^2. \tag{1}$$

The optimal projection matrix can be factorized as $P = UU^T$ where $U \in \mathbf{R}^{d\times r}$ has orthonormal columns. The r-dimensional representations are given as

$$y_i = U^T x_i, \qquad i = 1, \ldots, n. \tag{2}$$

It is straightforward to show that the problem (1) is equivalent to

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n \|y_i\|^2 = \frac{1}{2n}\sum_{i,j}\|y_i - y_j\|^2 \\ \text{subject to} & U^TU = I, \quad y_i = U^T x_i \end{array} \tag{3}$$

where I denotes the identity matrix. Thus PCA computes the low dimensional projections that have maximum variance, or equivalently, maximum total pairwise distances.

The solution to PCA is obtained from the eigenvalue decomposition of the covariance matrix $C = \sum_{i=1}^n x_i x_i^T$. Suppose $C = \sum_{i=1}^d \lambda_i u_i u_i^T$, where λ_i is the i-th largest eigenvalue of C and u_i is the associated unit eigenvector. Then the optimal low dimensional representations can be computed using the equation (2) with $U = [u_1, \ldots, u_r]$.

MDS computes the low dimensional representations that most faithfully preserve the inner products between the high dimensional data points. That is, it finds $y_1, \ldots, y_n \in \mathbf{R}^r$ to solve the problem

$$\text{minimize} \quad \textstyle\sum_{i,j}(x_i^T x_j - y_i^T y_j)^2 = \|G - K\|_F^2$$

where G and K are the Gram matrices of the inputs and outputs, with $G_{ij} = x_i^T x_j$ and $K_{ij} = y_i^T y_j$, respectively; and $\|\cdot\|_F$ denotes the matrix Frobenius norm. Thus, MDS tries to best approximate the Gram matrix. In fact MDS is often motivated by preserving the pairwise distances. Let $D_{ij} = \|x_i - x_j\|^2$ and D be the matrix of squared pairwise Euclidean distances. It can be shown that

$$G = -\tfrac{1}{2}\left(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T\right) D \left(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T\right) \tag{4}$$

where $\mathbf{1}$ denotes the vector of all ones.

The solution to MDS is obtained from the eigenvalue decomposition of the Gram matrix G. Suppose $G = \sum_{k=1}^n \lambda_k v_k v_k^T$, where λ_k is the k-the largest eigenvalue of G and v_k is the corresponding unit eigenvector. The outputs of MDS are given by

$$y_i = \left[\sqrt{\lambda_1}(v_1)_i \ \ldots \ \sqrt{\lambda_r}(v_r)_i\right]^T, \quad i = 1, \ldots, n. \tag{5}$$

It turns out that MDS and PCA produce the same outputs. Note that we can write $C = XX^T$ and $G = X^TX$ with $X = [x_1 \ldots x_n]$, and the equivalence of their outputs can be easily established using the singular value decomposition of X. In both cases, a large gap between the r-th and the $(r+1)$-th eigenvalues indicates that the inputs can be well approximated by outputs in a subspace of dimension r.

3. Maximum variance unfolding

MVU is also known as semidefinite embedding (SDE) as it was first proposed in Weinberger and Saul (2004). This algorithm attempts to "unfold" the manifold by pulling the data points apart as far as possible, while faithfully preserving the local distances and angles between nearby input data.

The first step of the algorithm is to construct a undirected graph by connecting each input x_i with its k-nearest neighbors, where k is a small integer. Call this graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with node set $\mathcal{V} = \{1, \ldots, n\}$ representing the set of inputs, and $\{i,j\} \in \mathcal{E}$ if x_i is connected to x_j. We assume the graph is connected.

MVU attempts to find low dimensional representations $y_1, \ldots, y_n \in \mathbf{R}^r$ that have the maximum possible total variance, while preserving the local distances over each edge of the graph. This can be formulated as the quadratic programming problem

$$\begin{array}{ll} \text{maximize} & \sum_i \|y_i\|^2 = \frac{1}{2n}\sum_{i,j}\|y_i - y_j\|^2 \\ \text{subject to} & \sum_i y_i = 0 \\ & \|y_i - y_j\|^2 = D_{ij}, \quad \{i,j\} \in \mathcal{E}. \end{array} \tag{6}$$

Here the optimization variables are the y_i's, and the problem data are the D_{ij}'s and $\mathcal{E}$. (Recall that $D_{ij} = \|x_i - x_j\|^2$ are computed from the input data.) The constraint $\sum_i y_i = 0$ eliminates the translational degree of freedom. It is obvious that the objective of maximizing the total variance has root in PCA, cf. the formulation (3). It is also closely related MDS since it can also be interpreted as maximizing the total pairwise distances.

The quadratic program (6) is not convex, but it can be reformulated as one, in particular, a semidefinite program (SDP) (Vandenberghe & Boyd, 1996). Let K denote the Gram matrix of the outputs, with components $K_{ij} = y_i^T y_j$. Then SDP formulation is

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}\, K \\ \text{subject to} & K = K^T \succeq 0, \quad \mathbf{1}^T K \mathbf{1} = 0 \\ & K_{ii} + K_{jj} - 2K_{ij} = D_{ij}, \quad \{i,j\} \in \mathcal{E} \end{array} \tag{7}$$

where $K \succeq 0$ means that the matrix K is positive semidefinite (i.e., has only nonnegative eigenvalues).

The reformulation into SDP not only allows global and efficient solution of the MVU problem, but also gives the extra capability of estimating the intrinsic dimension. Note that in the quadratic program (6), we have to first choose the output dimension r before solving it, not to mention the hardness to find the global optimum. By solving the SDP (7), we obtain an optimal Gram matrix $K^\star$ without specifying the output dimension r. Then we can apply MDS on $K^\star$ to estimate r from the number of significant eigenvalues, and construct the low dimensional representations y_i from the associated eigenvectors as done in (5).

3.1. The dual MVU problem

Examining the dual of an optimization problem often gives further insight of the problem and offers theoretical and computational advantages (Boyd & Vandenberghe, 2004). The MVU problem is no exception.

We call the problem (7) the primal MVU problem. In forming the Lagrangian, we associate the dual variable $Z = Z^T \succeq 0$ with the constraint $K = K^T \succeq 0$, the dual variable $\nu \in \mathbf{R}$ with the constraint $\mathbf{1}^T K \mathbf{1} = 0$, and the dual variables W_{ij} with the constraints $K_{ii} + K_{jj} - 2K_{ij} = D_{ij}$ for $\{i,j\} \in \mathcal{E}$. For convenience, we write the last set of equality constraints as

$$\mathbf{Tr}\, K E^{\{i,j\}} = D_{ij}, \qquad \{i,j\} \in \mathcal{E}$$

where the $n \times n$ matrix $E^{\{i,j\}}$ has only four nonzero elements: $E^{\{i,j\}}_{ii} = E^{\{i,j\}}_{jj} = 1$, $E^{\{i,j\}}_{ij} = E^{\{i,j\}}_{ji} = -1$. We consider the dual variables W_{ij} as elements of a $n \times n$ matrix W with $W_{ij} = 0$ if $\{i,j\} \notin \mathcal{E}$. Thus we have the Lagrangian

$$\begin{aligned} L(K, Z, \nu, W) &= \mathbf{Tr}\, K + \mathbf{Tr}\, KZ - \nu \mathbf{1}^T K \mathbf{1} \\ &\quad - \sum_{\{i,j\} \in \mathcal{E}} W_{ij} \left(\mathbf{Tr}\, K E^{\{i,j\}} - D_{ij} \right) \\ &= \mathbf{Tr}\, K \left(I + Z - \nu \mathbf{1}\mathbf{1}^T - \sum_{\{i,j\} \in \mathcal{E}} W_{ij} E^{\{i,j\}} \right) \\ &\quad + \sum_{\{i,j\} \in \mathcal{E}} D_{ij} W_{ij}. \end{aligned}$$

The dual function is obtained as

$$\begin{aligned} g(Z, \nu, W) &= \sup_{K = K^T} L(K, Z, \nu, W) \\ &= \begin{cases} \sum_{\{i,j\} \in \mathcal{E}} D_{ij} W_{ij} & \text{if } I + Z - \nu \mathbf{1}\mathbf{1}^T - \sum_{\{i,j\} \in \mathcal{E}} W_{ij} E^{\{i,j\}} = 0 \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Eliminating Z from the equality, the feasibility condition in the above equation becomes

$$I - \nu \mathbf{1}\mathbf{1}^T - L \preceq 0, \qquad L = \textstyle\sum_{\{i,j\} \in \mathcal{E}} W_{ij} E^{\{i,j\}}.$$

Note that L is a weighted Laplacian of the graph $\mathcal{G}$. The above linear matrix inequality is equivalent to

$$\nu \geq 1/n, \qquad \lambda_{n-1}(L) \geq 1$$

where λ_{n-1} denotes the second smallest eigenvalue of a symmetric matrix. Here $\lambda_n(L) = 0$ with associated eigenvector $\mathbf{1}$. Thus the dual MVU problem is

$$\begin{array}{ll} \text{minimize} & \sum_{\{i,j\}\in\mathcal{E}} D_{ij}W_{ij} \\ \text{subject to} & \lambda_{n-1}(L) \geq 1 \\ & L = \sum_{\{i,j\}\in\mathcal{E}} W_{ij}E^{\{i,j\}}. \end{array} \tag{8}$$

This is a convex optimization problem because the function $\lambda_{n-1}(L)$ is concave under the implicit constraint $\lambda_n(L) = 0$ (Sun et al., 2005). Note that the dual variable ν does not appear in the problem.

Since both the objective $\sum D_{ij}W_{ij}$ and the constraint function $\lambda_{n-1}(L)$ in problem (8) are positive homogeneous in W, we can just as well maximize $\lambda_{n-1}(L)$ subject to a constraint on $\sum D_{ij}W_{ij}$. This leads to an alternative formulation of the dual MVU problem

$$\begin{array}{ll} \text{maximize} & \lambda_{n-1}(L) \\ \text{subject to} & \sum_{\{i,j\}\in\mathcal{E}} D_{ij}W_{ij} = c \\ & L = \sum_{\{i,j\}\in\mathcal{E}} W_{ij}E^{\{i,j\}} \end{array} \tag{9}$$

where the constant $c > 0$ can be chosen arbitrarily. This again is a convex optimization problem (e.g., can be formulated as as SDP). The two formulations of the dual MVU problem are equivalent in the following sense: If $W^\star$ is an optimal solution to problem (8) and let $c^\star$ denotes its optimal value, then $(c/c^\star)W^\star$ is an optimal solution to problem (9) with optimal value $\lambda^\star_{n-1} = c/c^\star$. A similar relationship holds backward.

The formulation (9) is closely related to the *absolute algebraic connectivity* problem (Fiedler, 1989), in which $c = |\mathcal{E}|$ and the weights W_{ij} are constrained to be nonnegative. The same formulation and its duality with MVU were studied by Sun et al. (2005) in the context of finding the fastest mixing continuous-time Markov chain on a graph.

3.2. Duality and optimality conditions

The following duality results hold for the primal MVU problem (7) and the dual MVU problem (8).

- *Weak duality.* For any primal feasible K and any dual feasible W, we have

  $$\mathbf{Tr}\, K \leq \textstyle\sum_{i,j} D_{ij}W_{ij}.$$

  (Note that $W_{ij} = 0$ if $\{i,j\} \notin \mathcal{E}$.) Thus, any dual feasible W gives an upper bound on the optimal value of the primal MVU problem. This can be seen by checking the duality gap:

  $$\begin{aligned} &\textstyle\sum_{i,j} D_{ij}W_{ij} - \mathbf{Tr}\, K \\ &= \textstyle\sum_{i,j} D_{ij}W_{ij} - \mathbf{Tr}\, LK + \mathbf{Tr}\, LK - \mathbf{Tr}\, K \\ &= \textstyle\sum_{i,j} \Big(D_{ij} - (K_{ii} + K_{jj} - 2K_{ij})\Big)W_{ij} \\ &\quad + \mathbf{Tr}(L - I)K - (1/n)\mathbf{1}^T K\mathbf{1} \\ &= \mathbf{Tr}\left(L - \left(I - (1/n)\mathbf{1}\mathbf{1}^T\right)\right)K \geq 0. \end{aligned} \tag{10}$$

  The last inequality holds because $\lambda_{n-1}(L) \geq 1$ implies that $L - \left(I - (1/n)\mathbf{1}\mathbf{1}^T\right)$ is positive semidefinite, and the trace of the product of two positive semidefinite matrices is nonnegative. If this gap is zero, then K is optimal for the primal, and W is optimal for the dual. In other words, zero gap is sufficient for optimality.

- *Strong duality.* There exist a primal-dual feasible pair $(K^\star, W^*)$ with zero duality gap, i.e.,

  $$\mathbf{Tr}\, K^\star = \textstyle\sum_{i,j} D_{ij}W^\star_{ij}.$$

  This means that optimal values of the primal and dual problems are the same. Strong duality follows from Slater's condition for constraint qualification (Boyd & Vandenberghe, 2004).

A pair $(K^\star, W^\star)$ is primal-dual optimal if and only if they satisfy the following Karush-Kuhn-Tucker (KKT) optimality conditions:

- primal feasibility

  $$\begin{aligned} &K^\star = K^{\star T} \succeq 0, \qquad \mathbf{1}^T K^\star \mathbf{1} = 0 \\ &K^\star_{ii} + K^\star_{jj} - 2K^\star_{ij} = D_{ij}, \quad \{i,j\} \in \mathcal{E} \end{aligned}$$

- dual feasibility

  $$L^\star = \textstyle\sum_{\{i,j\}\in\mathcal{E}} W^\star_{ij}E^{\{i,j\}}, \qquad \lambda_{n-1}(L^\star) \geq 1$$

- complementary slackness

  $$L^\star K^\star = K^\star \tag{11}$$

  This is the result of enforcing equality in (10).

Note that we always have $\lambda_{n-1}(L^\star) = 1$. Thus the complementary slackness condition (11) means that the range of $K^\star$ lies in the eigenspace (e.s.) of $L^\star$ associated with λ_{n-1}. Since $K^\star$ is a dense Gram matrix while $L^\star$ is a sparse weighted Laplacian, equation (11) means precisely

$$\text{top e.s. of dense } K^\star \subseteq \text{bottom e.s. of sparse } L^\star \tag{12}$$

Here "bottom e.s." means the eigenspace associated with λ_{n-1}. (We discard the eigenvector $\mathbf{1}$ of $L^\star$ associated with the smallest eigenvalue $\lambda_n = 0$.) Another direct consequence of (11) is

$$r \;\le\; \mathbf{Rank}\, K^\star \;\le\; \text{multiplicity of } \lambda_{n-1}(L^\star) \qquad (13)$$

where r is the dimension of the low dimensional representations obtained by performing MDS on $K^\star$. We have $r < \mathbf{Rank}\, K^\star$ if there is a significant gap in the nonzero eigenvalues of $K^\star$.

With the inequality (13), Sun et al. (2005) showed that the maximum-variance embeddings of a path must be one-dimensional, and for a ring it must be two-dimensional. It can also be show that the maximum-variance embedding of a tree can always be two-dimensional. Göring et al. (2005) studied similar graph embedding problems using duality theory for the absolute algebraic connectivity problem (9).

In the rest of the paper, we will show how various spectral methods for nonlinear dimensionality reduction are connected by the MVU duality theory.

4. Isomap

Isomap computes low dimensional representations of the high dimensional data that best preserve pairwise distances as measured along the submanifold from which they were sampled. It can be understood as a variant of MDS in which we use estimates of pairwise geodesic distances on the submanifold, instead of the standard Euclidean distances.

The algorithm has three steps. First it constructs the k-nearest neighbor graph, and assigns each edge a length that equals the Euclidean distance between the two nodes connected. The second step is to compute the pairwise distance Δ_{ij}, for all pairs of nodes i and j, as the length of the shortest paths connecting them on the graph (e.g., using Djikstra's algorithm). In the third step, it uses the pairwise distances Δ_{ij} as inputs to MDS as described in §2. More specifically, it computes a matrix G using (4) with D substituted by Δ, estimates the dimension r by the number of significant eigenvalues of G, and constructs the low-dimensional representations using (5). Note that in this case G may not be positive semidefinite.

4.1. Connection to MVU

Isomap can be interpreted as directly constructing an *approximate* solution for the primal MVU problem. We argue as follows. Consider the Riemannian structure on a manifold induced from the standard Euclidean metric on $\mathbf{R}^d$. The Euclidean distance between any two points on the manifold is always smaller than their geodesic distance. Thus the total pairwise Euclidean distances of the data points is upper bounded by their total pairwise geodesic distances. In addition, we see in (6) that maximizing the variance is equivalent to maximizing the total pairwise Euclidean distances. So in this sense, Isomap attempts to maximize the variance by directly using the geodesic distances.

This interpretation becomes accurate in the limit, with increasing sampling density ($n \to \infty$), if the submanifold is isometric to a convex subset of the Euclidean space. In particular, this condition guarantees the asymptotic convergence of the Isomap algorithm (Bernstein et al., 2000; Donoho & Grimes, 2002). In this case, the pairwise geodesic distances become feasible to the MVU problem, and the solution to MVU approaches its upper bound obtained by Isomap. Thus MVU converges to the same limit as Isomap.

If the above condition is not satisfied, then Isomap and MVU could behave quite differently (Weinberger & Saul, 2004). More general conditions for the asymptotic convergence of MVU is still an open question.

5. Locally linear embedding

LLE computes low dimensional representations of the high dimensional data that most faithfully preserve the local linear structure. The algorithm and Laplacian eigenmaps (see next section) differ from Isomap and MVU in that they use the bottom eigenvectors of a sparse matrix, as opposed to the top eigenvectors of a dense Gram matrix.

LLE has three steps. First, as other methods, it construct a k-nearest neighbor graph. However, this is a *directed* graph whose edges indicate nearest neighbor relations, which may or may not be symmetric. In this case, the set of edges $\mathcal{E}$ consists of *ordered* pairs (i, j) meaning that j is a neighbor of i. We let $\mathcal{N}_i = \{j | (i, j) \in \mathcal{E}\}$ to denote the set of neighbors of i. In the second step, LLE assigns a weight W_{ij} to each edge $(i, j) \in \mathcal{E}$ by solving the least-squares problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \left\| x_i - \sum_{j\in\mathcal{N}_i} W_{ij} x_j \right\|^2 \\ \text{subject to} & \sum_{j\in\mathcal{N}_i} W_{ij} = 1, \quad i = 1, \ldots, n. \end{array} \qquad (14)$$

(A regularization term may be added to the objective to obtain unique solution.) In the third step, LLE computes $y \in \mathbf{R}^r$ by solving another least-square problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \left\| y_i - \sum_{j\in\mathcal{N}_i} W_{ij} y_j \right\|^2 \\ \text{subject to} & \sum_i y_i = 0, \quad (1/n) \sum_i y_i y_i^T = I \end{array} \qquad (15)$$

It turns out that the solution to (15) can be obtained by computing the bottom $r+1$ eigenvectors of the matrix $(I-W)^T(I-W)$. Let these normalized eigenvectors be $v_n, v_{n-1}, \ldots, v_{n-r}$, associated with the bottom eigenvalues $0 = \lambda_n < \lambda_{n-1} \leq \cdots \leq \lambda_{n-r}$. We discard $v_n = (1/\sqrt{n})\mathbf{1}$ associated with $\lambda_n = 0$, and use the next r eigenvectors to form the outputs

$$y_i = [(v_{n-1})_i \ \ldots \ (v_{n-r})_i]^T, \qquad i = 1, \ldots, n. \quad (16)$$

5.1. Connection to MVU

The key idea behind LLE is that every point on the submanifold can be approximately reconstructed by a linear combination of its neighbors, i.e.,

$$x_i \approx \textstyle\sum_{j \in \mathcal{N}_i} W_{ij} x_j, \qquad i = 1, \ldots, n. \quad (17)$$

(Locally the manifold can be well approximated by its tangent space.) The sparse matrix W obtained by (14) encodes such local geometric properties of the inputs. We shall show that such local linear properties are hidden in the optimality conditions of the MVU problem, in particular, the complementarity condition (11).

Let $\widetilde{Y} = [\tilde{y}_1 \ \ldots \ \tilde{y}_n]$ be the outputs of MVU. Then we can write $K^\star = \widetilde{Y}^T\widetilde{Y}$. Now (11) implies $L^\star\widetilde{Y}^T = \widetilde{Y}^T$, which in turn can be written as

$$\tilde{y}_i = \textstyle\sum_{j \in \mathcal{N}_i} W^\star_{ij}(\tilde{y}_i - \tilde{y}_j), \qquad i = 1, \ldots, n$$

where $W^\star_{ij}$ are the optimal solutions to the dual MVU problem (8). This equation describes a local linear relationship of the data. In fact it can be converted to

$$(L^\star_{ii} - 1)\tilde{y}_i = \textstyle\sum_{j \in \mathcal{N}_i} W^\star_{ij}\tilde{y}_j, \qquad i = 1, \ldots, n \quad (18)$$

where $L^\star_{ii} = \sum_j W_{ij}$. We see that the equations (17) and (18) describe very similar linear relationships, except for a scaling factor and the fact that $W^\star$ in (18) is symmetric while W in (17) is nonsymmetric. Therefore the motivation behind LLE has an interpretation from the primal-dual optimality conditions for the MVU problem.

6. Laplacian eigenmaps

Laplacian eigenmaps compute low dimensional representations of the high dimensional data that most faithfully preserve proximity relations, mapping nearby inputs into nearby outputs.

First, the algorithm construct a undirected, k-nearest neighbor graph as in MVU and Isomap. Then it assigns positive weights W_{ij} to every edge of the graph; for example, let $W_{ij} = 1$ for all $\{i,j\} \in \mathcal{E}$, or let $W_{ij} = \exp(-\|x_i - x_j\|^2/\sigma^2)$ where σ^2 is a scalar parameter. In the last step, for a given dimension r, it finds outputs $y_i \in \mathbf{R}^r$ by solving the problem

$$\begin{array}{ll} \text{minimize} & \sum_{\{i,j\} \in \mathcal{E}} W_{ij}\|y_i - y_j\|^2 \\ \text{subject to} & \sum_i L_{ii} y_i y_i^T = I \end{array} \quad (19)$$

where $L_{ii} = \sum_j W_{ij}$ are the diagonal elements of the weighted Laplacian L. The cost function encourages nearby inputs to be mapped into nearby outputs.

The solution to (19) is obtained by computing the bottom $r+1$ unit eigenvectors of the generalized eigenvalue problem

$$L\,v_j = \lambda_j\,D_L\,v_j, \quad j = n, n-1, \ldots, n-r$$

where D_L denotes the diagonal matrix formed by taking the diagonals of L. This is equivalent to compute the bottom eigenvectors of the normalized Laplacian $D_L^{-1/2} L D_L^{-1/2}$ and then scale them by the diagonal matrix $D_L^{-1/2}$. The outputs y_i are given by (15) as in LLE. We can also use a variation of Laplacian eigenmaps where the constraint in (19) is changed to $\sum_i y_i y_i^T = I$. In this case, we simply use the bottom eigenvectors of L.

6.1. Connection to MVU

There is a great deal of freedom in choosing the edge weights W_{ij} (these are symmetric). We relate Laplacian eigenmaps to MVU by considering these weights as feasible solutions to the dual MVU problem (8). Note that the constraint $\lambda_{n-1}(L) \geq 1$ in (8) can always be satisfied by scaling up the weights, which does not change the eigenvectors. With this in mind, we can interpret the dual MVU problem as a particular way to choose the weights, with the objective

$$\text{minimize} \quad \textstyle\sum_{\{i,j\} \in \mathcal{E}} W_{ij}\|x_i - x_j\|^2. \quad (20)$$

(Note $D_{ij} = \|x_i - x_j\|^2$.) This objective has the similar form as (19), with outputs y_i substituted by inputs x_i.

Thus we can solve the dual MVU problem (8) first, finding the weights $W^\star$ that minimize the objective (20) subject to $\lambda_{n-1}(L) \geq 1$, then use $W^\star$ in (19) to compute the outputs. This two-step procedure is precisely like the one use in LLE, cf. (14) and (15). Moreover, with such a pre-optimization of the weights, Laplacian eigenmaps compute the bottom eigenvectors of $L^\star$, solution to the dual MVU problem. By the MVU duality theory, in particular (12), we know that they coincide with the top eigenvectors of the primal solution $K^\star$, given that they use the same dimension r.

Solving the dual MVU problem (8) to obtain $W^\star$ for Laplacian eigenmaps can be very costly, if we convert

this problem into an SDP and solve it by interior-point methods (Boyd & Vandenberghe, 2004). Solving SDPs is limited to problem size up to $n \approx 2000$. However, the alternative formulation (9) can be solved by subgradient-type algorithms, for problems with n up to $100,000$; see a similar problem in Boyd et al. (2004).

Unlike Isomap and MVU, the bottom eigenvalues of L in Laplacian eigenmaps do not have a tellable gap that allow us to estimate the dimensionality of the underlying manifold (LLE is similar). This can also be understood from the MVU duality theory — the bottom eigenvectors correspond to closely located eigenvalues, actually the same eigenvalue λ_{n-1} when using $L^\star$. The next smaller eigenvalue may be very close to λ_{n-1}, but its associated eigenvector(s) could have little contribution in building a faithful representation. In practice, we cannot expect to tell multiplicities of eigenvalues from numerical results, thus it is difficult to estimate the intrinsic dimension of the underlying manifold.

6.2. Extensions

Although producing roughly the same eigenspace for embedding, methods based on sparse matrices lose the scaling factors given by eigenvalues as done in methods based on dense matrices; cf. (16) and (5). Such scaling factors can be essential in obtaining isometric embeddings. An improvement in this direction can be achieved by adding a post-processing step using MVU.

Let V be a $n \times r$ matrix whose columns are the r bottom eigenvectors obtained from Laplacian eigenmaps or LLE (after discarding the constant vector associated with zero eigenvalue). We can approximate the Gram matirx in MVU by $K = VQV^T$, where Q is $r \times r$ and positive semidefinite. Then we form the SDP

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}\, VQV^T \\ \text{subject to} & Q = Q^T \succeq 0, \quad K = VQV^T \\ & K_{ii} + K_{jj} - 2K_{ij} \le D_{ij}, \quad \{i,j\} \in \mathcal{E} \end{array} \tag{21}$$

Comparing with (7), here the constraint $\mathbf{1}^T K \mathbf{1} = 0$ is automatically satisfied, but we have to relax the pairwise distance constraints to inequalities to preserve feasibility. Solving the SDP (21) costs much less computationally than solving (7) because the variable Q has size $r \times r$ instead of $n \times n$. In addition, we can recover the scaling factors using the eigenvalues of Q. In general, we can use more than r bottom eigenvectors from Laplacian eigenmaps to form V. This gives us the additional capability of estimating r from the gap in the eigenvalue spectra of Q.

A very similar approach has been explored by Weinberger et al. (2005). They choose a set of landmarks $z_1, \ldots, z_m \in \mathbf{R}^d$ ($m \ll n$) of the inputs and find a matrix $V \in \mathbf{R}^{n \times m}$ that best approximates all the inputs as $x_i \approx \sum_j V_{ij} z_j$. The matrix V is constructed from LLE. Then an SDP similar to (21) is solved to get the optimal landmark kernel Q. From Q they find low dimensional representations for the m landmarks $\tilde{z}_j \in \mathbf{R}^r$ and generate outputs $y_i = \sum_j V_{ij} \tilde{z}_j$. We note that the number of landmarks m, though much smaller than n, could still be much larger than r. Sha and Saul (2005) studied other extensions, e.g., conformal eigenmaps, that use SDP to post-process eigenvectors obtained from Laplacian eigenmaps or LLE.

7. Conclusions

We have shown that MVU duality theory reveals close connections between several spectral methods for nonlinear dimensionality reduction. In particular, Isomap can be considered as directly constructing an approximate optimal solution for the primal MVU problem. With increasing sampling density, these two methods converge to the same solution in the limit if the underlying submanifold is isometric to a convex subset of Euclidean space. The locally linear structure embraced by LLE can be interpreted from the optimality conditions of MVU. Laplacian eigenmaps use edge weights that are feasible to the dual MVU problem. Using the optimal weights for the dual MVU problem corresponds to a two-step procedure similar as in LLE. This duality framework also explains why using top eigenvectors of dense Gram-like matrices and using bottom eigenvectors of sparse Laplacian-like matrices can produce similar results — these two eigenspaces coincide at primal-dual optimality.

By capturing the simple yet key feature of maximizing variance, exactly or approximately, MVU duality theory offers a unified view of several spectral methods for nonlinear dimensionality reduction. Nevertheless, MVU is certainly not the best universal solution, and different methods may perform well on different class of problems. Currently we are experimenting with new variants and extensions suggested by the duality framework, and working on empirical results to illustrate the theoretical connections developed in this paper.

Acknowledgments

The authors are grateful to Lawrence Saul and Kilian Weinberger for insightful discussions. Part of this work was done when Lin Xiao was on a supported visit at the Institute for Mathematical Sciences, National University of Singapore.

References

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*, 1373–1396.

Bernstein, M., de Silva, V., Langford, J. C., & Tenenbaum, J. B. (2000). *Graph approximations to geodesics on embedded manifolds* (Technical Report). Stanford University.

Boyd, S., Diaconis, P., & Xiao, L. (2004). Fastest mixing Markov chain on a graph. *SIAM Review, problems and techniques section*, *46*, 667–689.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Brand, M. (2004). *From subspaces to submanifolds* (Technical Report 2004-134). Mitsubishi Electric Research Laboratories.

Burges, C. J. C. (2005). Geometric methods for feature extraction and dimensional reduction. In L. Rokach and O. Maimon (Eds.), *Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers*. Kluwer Academic Publishers.

Costa, J., & Hero, A. O. (2004). Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. on Signal Process.*, *52*, 2210–2221.

Cox, T., & Cox, M. (1994). *Multidimensional scaling*. London: Chapman & Hall.

Donoho, D. L., & Grimes, C. E. (2002). *When does Isomap recover the natural parameterization of families of articulated images?* (Technical Report 2002-27). Stanford University, Department of Statistics.

Donoho, D. L., & Grimes, C. E. (2003). Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, *100*, 5591–5596.

Fiedler, M. (1989). Laplacian of graphs and algebraic connectivity. In *Combinatorics and graph theory*, vol. 25 of *Banach Center Publications*, 57–70. Warsaw: PWN-Polish Scientific Publishers.

Göring, F., Helmberg, C., & Wappler, M. (2005). *Embedded in the shadow of the separator* (Preprint 2005-12). Fakultät für Mathematik, Technische Universität Chemnitz, Chemnitz, Germany.

Ham, J., Lee, D. D., Mika, S., & Schölkopf, B. (2004). A kernel view of the dimensionality reduction of manifolds. *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)* (pp. 369–376). Banff, Alberta, Canada.

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

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

Saul, L. K., & Roweis, S. T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, *4*, 119–155.

Saul, L. K., Weinberger, K. Q., Sha, F., Ham, J., & Lee, D. D. (2005). Spectral methods for dimensionality reduction. In B. Schöelkopf, O. Chapelle and A. Zien (Eds.), *Semisupervised learning*. MIT Press.

Schölkopf, B., Smola, A. J., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*, 1299–1319.

Sha, F., & Saul, L. K. (2005). Analysis and extension of spectral methods for nonlinear dimensionality reduction. *Proc. of 22nd International Conference on Machine Learning* (pp. 785–792). Bonn, Germany.

Sun, J., Boyd, S., Xiao, L., & Diaconis, P. (2005). The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Review*. Accepted for publication.

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

Vandenberghe, L., & Boyd, S. (1996). Semidefinite programming. *SIAM Review*, *38*, 49–95.

Weinberger, K. Q., Packer, B. D., & Saul, L. K. (2005). Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics* (pp. 381–388). Barbados, West Indies.

Weinberger, K. Q., & Saul, L. K. (2004). Unsupervised learning of image manifolds by semidefinite programming. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)* (pp. 988–995). Washington D. C.

Zhang, Z., & Zha, H. (2004). Principle manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, *26*, 313–338.