Nonlinear Optimization

"A Duality View of Spectral Methods for Dimensionality Reduction" [6]

Telmo Cunha (73487)

Wednesday 6th July, 2022

Abstract

This project is based on the paper titled "A Duality View of Spectral Methods for Dimensionality Reduction" written by Lin Xiao, Jun Sun and Stephen Boyd. The goal is to understand how different methods for nonlinear dimensionality reduction are connected, in particular manifold learning methods. This is achieved by formulating the dual problem of a relaxed version of the maximum variance unfolding optimization problem (MVU) and then analyzing the primal-dual optimality results. The authors arrive at a connection between different spectral methods (top eigenvectors of dense matrices versus bottom eigenvectors of sparse matrices) used in the solutions of distinct manifold learning methods, Isomap, Locally Linear Embeddings and Laplacian Eigenmaps.

Introduction to Manifold Learning

Manifold learning consists of finding a lower dimensional embedding of data, i.e. given a set of datapoints $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, we want to find $\{y_i\}_{i=1}^n$, $y_i \in \mathbb{R}^r$, where ideally $r << d$, under the assumption that this data lies approximately within a submanifold of this higher dimensional space. This typically involves a graph formulation which reflects a discrete approximation of this underlying submanifold (done using either kNN or ϵ-neighbors). The usefulness of dimensionality reduction comes from a reduction in complexity allowing for better interpretability of the data and to understand how different features are related. The particular case of non-linear dimensionality reduction allows to capture non-linear relationships between the data which is lost under linear methods such as PCA and MDS.

To obtain these embeddings, for the methods discussed in the paper, one relies on spectral decomposition, either considering the top eigenvectors of a dense matrix or the bottom eigenvectors of a sparse matrix. These methods are distinct in their geometric motivation and are a priori unrelated so the connection between them via the duality of MVU is particularly interesting.

Before introducing MVU we take a brief look at two linear methods, Principal Component Analysis (PCA) and Metric Multidimensional Scaling (MDS) since they serve as motivation for the remaining methods. In fact, Isomap relies directly on MDS with a different input.

Principal Component Analysis (PCA)

Given input data $\{x_i\}_{i=1}^n$ (w.l.o.g. assuming $\sum_i x_i = 0$), $x_i \in \mathbb{R}^d$, we want to find $\{y_i\}_{i=1}^n$, $y_i \in \mathbb{R}^r$, where ideally $r << d$, under the assumption that the data lies close to a subspace of the higher dimensional feature space. One possible formulation of PCA is the following:

$$\min_P \quad \sum_{i=1}^n \|x_i - Px_i\|^2 \tag{1}$$

where $P_{n\times n}$ is an orthogonal projection ($P^2 = P$ and $P = P^T$) matrix of rank $r < d$ (non-convex constraint) which can be factorized as $P = UU^T$, where $U_{n\times r}$ has orthonormal columns.

- Why can we factorize $P = UU^T$?
  Suppose P projects on a subspace S with orthonormal basis $\{u_1, ..., u_r\}$, $u_i \in \mathbb{R}^n$. We can write P as $P = \sum_{i=1}^r \langle u_i, \cdot \rangle u_i$. So, given any x, we have $Px = \sum_{i=1}^r u_i^T x u_i$. Defining $U = [u_1\ u_2\ ...\ u_r]$, we have:

$$UU^T x = \begin{bmatrix} | & & | \\ u_1 & \dots & u_r \\ | & & | \end{bmatrix} \begin{bmatrix} \langle u_1, x \rangle \\ ... \\ \langle u_r, x \rangle \end{bmatrix} = \sum_{i=1}^r \langle u_i, x \rangle u_i = Px \tag{2}$$

The embeddings (projected datapoints) into the $r-$dimensional subspace are given by $y_i = U^T x_i$ for $i = 1, ..., n$.

We can reformulate this problem in a different way. For a single x_i we have:

$$\begin{aligned}\|x_i - Px_i\|^2 = \|x_i - UU^T x_i\|^2 = (x_i - UU^T x_i)^T (x_i - UU^T x_i) = \|x_i\|^2 - 2x_i^T UU^T x_i + (UU^T x_i)^T (UU^T x_i) = \\ = \|x_i\|^2 - 2x_i^T UU^T x_i + x_i^T UU^T UU^T x_i = \|x_i\|^2 - x_i^T UU^T x_i = \|x_i\|^2 - (U^T x_i)^T U^T x_i = \|x_i\|^2 - \|y_i\|^2\end{aligned} \tag{3}$$

since $y_i = U^T x_i$ and $U^T U = I$. Thus, we have an equivalence of the following problems:

$$\begin{aligned} \min_{y_i} \quad & \|x_i\|^2 - \|y_i\|^2 \\ \text{s.t.} \quad & U^T U = I \\ & y_i = U^T x_i \end{aligned} \qquad\qquad \begin{aligned} \max_{y_i} \quad & \|y_i\|^2 = \frac{1}{2n} \sum_{i,j} \|y_i - y_j\|^2 \\ \text{s.t.} \quad & U^T U = I \\ & y_i = U^T x_i \end{aligned}$$

The equivalence on the right can be seen from the following computation. First, we have $\sum_i y_i = 0$ since $\sum_i U^T x_i = U^T \sum_i x_i = U^T 0 = 0$. Then:

$$\begin{aligned} \frac{1}{2n} \sum_{i,j} \|y_i - y_j\|^2 &= \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n (y_i - y_j)^T (y_i - y_j) = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n y_i^T y_i + y_j^T y_j - 2 y_i^T y_j = \\ &= \frac{1}{2n} \left((n y_1^T y_1 + \sum_{j=1}^n y_j^T y_j) + ... + (n y_n^T y_n + \sum_{j=1}^n y_j^T y_j) \right) - 2 \sum_{i=1}^n \sum_{j=1}^n y_i^T y_j = \\ &= \frac{1}{2n} \left(n \sum_{i=1}^n y_i^T y_i + n \sum_{j=1}^n y_j^T y_j \right) - 2 \sum_{i=1}^n \sum_{j=1}^n y_i^T y_j = \sum_{i=1}^n \|y_i\|^2 - 2 \sum_{i=1}^n \sum_{j=1}^n y_i^T y_j = \\ &= \sum_{i=1}^n \|y_i\|^2 - 2 \left(y_1^T (\sum_{j=1}^n y_j) + ... + y_n^T (\sum_{j=1}^n y_j) \right) = \sum_{i=1}^n \|y_i\|^2 \end{aligned} \tag{4}$$

The solution to PCA is obtained from the eigenvalue decomposition of the covariance matrix $\Sigma = \sum_i x_i x_i^T$, which is a symmetric positive semidefinite matrix. Considering its eigenvalue decomposition $\Sigma = U \Lambda U^T$ we have $\Sigma = \sum_{i=1}^d \lambda_i u_i u_i^T$, where we consider $\lambda_1 \geq ... \geq \lambda_d \geq 0$. Since EVD is equivalent to SVD for symmetric matrices retaining the first r terms of the sum gives the best rank r approximation to Σ (Eckart–Young–Mirsky theorem). The first $y_i = U^T x_i$, $i = 1, ..., r$ are called the principal components, where U is as defined above.

To actually derive the solution to PCA one can use the method of Lagrange multipliers. Considering the problem for the first principal component as:

$$\begin{aligned} \max_{\alpha_i} \quad & \alpha_i^T \Sigma \alpha_i = \text{Var}(\alpha_i^T \text{x}) = \text{Var}(y_i) \\ \text{s.t.} \quad & \alpha_i^T \alpha_i = 1 \end{aligned} \tag{5}$$

The method of Lagrange multipliers exploits the fact that the maximum occurs at points of tangency between level curves. Setting $f(\alpha_i) = \alpha_i^T \Sigma \alpha_i$ and $g(\alpha_i) = \alpha_i^T \alpha_i$ we want to solve the following system of $(d+1)$ equations:

$$\begin{cases} \nabla f(\alpha_*) = \lambda \nabla g(\alpha_*) \\ \alpha_*^T \alpha_* = 1 \end{cases} \tag{6}$$

From the first equation we obtain $2\Sigma\alpha_* = 2\lambda\alpha_*$, i.e. α_* is an eigenvector of Σ with eigenvalue λ. To actually maximize $\alpha_i^T \Sigma \alpha_i$ we compute $\alpha_*^T \Sigma \alpha_* = \alpha_*^T \lambda \alpha_* = \lambda \alpha_*^T \alpha_* = \lambda$. Thus, we should pick the eigenvector corresponding to the largest eigenvalue to obtain maximum variance. Letting $\alpha_* = \alpha_1$, this gives our first principal component $\text{y}_1 = \alpha_1^T \text{x}$. To determine the next principal component we want to solve a similar problem with the added constraint that $\text{Cov}(y_1, y_i) = \alpha_1^T \Sigma \alpha_i = \lambda_i \alpha_1^T \alpha_i = 0$, we have then:

$$\begin{aligned} \max_{\alpha_i} \quad & \alpha_i^T \Sigma \alpha_i \\ \text{s.t.} \quad & \alpha_i^T \alpha_i = 1 \\ & \alpha_1^T \alpha_i = 0 \end{aligned} \tag{7}$$

This can again be solved using the Lagrange multipliers method which results in α_2 as an eigenvector of Σ with second largest eigenvalue. One can proceed similarly to obtain the remaining principal components, see [4].

Metric Multidimensional Scaling (MDS)

The goal here is the same as PCA, we consider the input data $\{x_i\}_{i=1}^n$ (assuming $\sum_i x_i = 0$), $x_i \in \mathbb{R}^d$, and we want to find $\{y_i\}_{i=1}^n$, $y_i \in \mathbb{R}^r$, where ideally $r << d$, under the assumption that the data lies close to a subspace of the higher dimensional feature space.

The idea of MDS is to find a lower dimensional representation of the data which closely approximates the inner products, i.e. we want to solve the following optimization problem:

$$\min_K \quad \sum_{i,j} (x_i^T x_j - y_i^T y_j)^2 := \|G - K\|_F^2 \tag{8}$$

where G and K are known as the Gram matrices of the inputs and outputs, respectively, where $G_{ij} = x_i^T x_j$ and $K_{ij} = y_i^T y_j$. Under this Gram matrix formulation the embedding dimension r is hidden in K. Since $K = Y^T Y$ this matrix will be at most of rank r.

The matrix G is symmetric, since the inner-product is symmetric, and positive semidefinite, which is easily seen by bilinearity of the inner-product:

$$y^T G y = \sum_i \sum_j y_i G_{ij} y_j = \sum_i \sum_j y_i \langle x_i, x_j \rangle y_j = \sum_i \sum_j \langle y_i x_i, y_j x_j \rangle = \left\langle \sum_i y_i x_i, \sum_j y_j x_j \right\rangle = \| \sum_i y_i x_i \|^2 \geq 0 \tag{9}$$

Another possible formulation of MDS is via pairwise distances $D_{ij} = \|x_i - x_j\|^2$, where the Gram matrix G is obtained from D by the following formula:

$$G = -\frac{1}{2}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) D (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) \tag{10}$$

Since the matrix G is symmetric it admits an eigenvalue decomposition (EVD) which coincides with its singular value decomposition (SVD). By the Eckart–Young–Mirsky theorem the best rank r approximation to G (under this norm) is given by its top r eigenvectors, i.e. for any rank r, $n \times n$ matrix A we have:

$$\|G - A\|_F \geq \|G - \sum_{k=1}^r \lambda_k v_k v_k^T\|_F \tag{11}$$

Thus, the solution to MDS is obtained from the eigenvalue decomposition of $G = \sum_{k=1}^n \lambda_k v_k v_k^T$, we have $K = \sum_{k=1}^r \lambda_k v_k v_k^T$ and $K = Y^T Y$, and the output is given by the top r eigenvectors:

$$y_i = \left(\sqrt{\lambda_1}(v_1)_i, \sqrt{\lambda_2}(v_2)_i, ..., \sqrt{\lambda_r}(v_r)_i \right) \tag{12}$$

for $i = 1, ..., n$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$ are the eigenvalues of G with corresponding eigenvectors v_i. The exact computation can be seen in the derivation of equation (23).

It is possible to show that MDS and PCA produce the same outputs, considering we can write $\Sigma = X^T X$ and $G = X^T X$, via singular value decomposition.

A large gap between eigenvalues λ_r and λ_{r+1} shows that the inputs are well approximated by the outputs in dimensions r for both PCA and MDS. This again follows from the fact that SVD is the best rank r approximation and coincides with the EVD. Typically in PCA one retains eigenvectors corresponding to the top eigenvalues which explain $\sim 90\%$ of the variance.

Maximum Variance Unfolding (MVU)

As seen in class, the idea of maximum variance unfolding is to approximate the underlying submanifold via a graph $G = (V, E)$ (figure 1-1) where the set of edges E is defined via kNN or $\epsilon-$neighbors (figure 1-2) and then, find an embedding where we maximize the variance, or equivalently, the norm of the embedded points (figure 1-3).

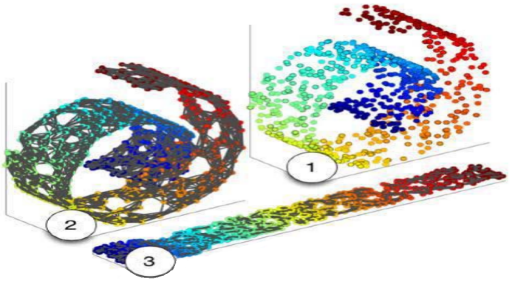

Figure 1: MVU on a "Swiss Roll". Adapted from: [7]

Considering the input data $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, we want to find an embedding $\{y_i\}_{i=1}^n$, $y_i \in \mathbb{R}^r$, where ideally $r << d$, but now the assumption is that the data lies close to a submanifold of the higher dimensional feature space, as opposed to a subspace, as was the case for PCA and MDS.

Suppose we constructed a connected graph $G = (V, E)$ (if it is not connected one considers each connected component as an independent problem) where the edges are obtained via kNN. The goal of the embedding is to maximize total variance under local constraints, where we want to keep edge lengths and angles between edges on the same node. The problem formulation is simpler when considering only constraints on distances. In order to do this we further connect the graph, i.e. given a vertex x_i we connect all neighbors x_j of x_i with each other. Then, the distance constraints over this new graph is equivalent to the distance and angle constraints of the initial graph and can be formulated as:

$$\text{Whenever } (i,j) \in E \implies \|y_i - y_j\|^2 = \|x_i - x_j\|^2 \tag{13}$$

We further consider the embedded data to have zero mean:

$$\sum_i y_i = 0, \tag{14}$$

which removes a translational degree of freedom from the embedding. Considering all of this, the optimization problem is given by:

$$\begin{aligned} \max_{y_i} \quad & \sum_{i=1}^n \|y_i\|^2 = \frac{1}{2n} \sum_{i,j=1}^n \|y_i - y_j\|^2 \\ \text{s.t.} \quad & \sum_{i=1}^n y_i = 0 \\ & \|y_i - y_j\|^2 = \|x_i - x_j\|^2, \text{ if } (i,j) \in E \end{aligned} \tag{15}$$

This is a non-convex problem since we are maximizing a convex quadratic and the second constraint is quadratic in the variables (and thus not an affine constraint). The equivalence of the objective function was seen in the PCA section, equation (4).

We now reformulate this problem, as seen in class, by expanding the terms in equation (15). We rewrite the problem in the following way:

$$\begin{aligned} \max_{y_i} \quad & \sum_{i=1}^n y_i^T y_i \\ \text{s.t.} \quad & 0 = \sum_{i=1}^n y_i = \|\sum_{i=1}^n y_i\|^2 = \left\langle \sum_{i=1}^n y_i, \sum_{j=1}^n y_j \right\rangle = \sum_{i=1}^n \sum_{j=1}^n y_i^T y_j \\ & y_i^T y_i + y_j^T y_j - 2 y_i^T y_j = \|x_i - x_j\|^2, \text{ if } (i,j) \in E \end{aligned} \tag{16}$$

Defining the matrix variable $Y_{r\times n} = [y_1\, y_2 \ldots y_n]$ we rewrite the problem as:

$$\begin{aligned} \max_{Y} \quad & \mathbf{Tr}(Y^TY) \\ \text{s.t.} \quad & \mathbf{1}^T(Y^TY)\mathbf{1} = 0 \\ & e_i^T(Y^TY)e_i + e_j^T(Y^TY)e_j - 2e_i^T(Y^TY)e_j = \|x_i - x_j\|^2, \text{ if } (i,j) \in E \end{aligned} \tag{17}$$

where the e_i, e_j are canonical basis vectors in $\mathbb{R}^n$. Introducing another matrix variable $K_{n\times n} = Y^TY$ we obtain:

$$\begin{aligned} \max_{K,Y} \quad & \mathbf{Tr}(K) \\ \text{s.t.} \quad & \mathbf{1}^TK\mathbf{1} = 0 \\ & K_{ii} + K_{jj} - 2K_{ij} = \|x_i - x_j\|^2, \text{ if } (i,j) \in E \\ & K = Y^TY \end{aligned} \tag{18}$$

The objective function is now linear and the constraints are linear and affine except for the last one, which has the non-convex part of the problem.

Since Y does not appear in the objective function its role is only to constrain the variable K. We now prove that we have the following equivalence:

$$K = Y^TY,\ Y_{r\times n} \Leftrightarrow K = K^T \succeq 0, \text{ and } \operatorname{rank}(K) \leq r \tag{19}$$

Proof.

($\Rightarrow$) $K = K^T$ since $K = Y^TY$ and $K^T = (Y^TY)^T = Y^TY = K$, or equivalently from $K_{ij} = \langle y_i, y_j\rangle = \langle y_j, y_i\rangle = K_{ji}$, by symmetry of the inner-product. We have $K \succeq 0$ since, given any $x \in \mathbb{R}^n$:

$$x^TKx = x^TY^TYx = (Yx)^T(Yx) = \|Yx\|^2 \geq 0 \tag{20}$$

Under the assumption that $Y_{r\times n}$ is an horizontal matrix, i.e. $r << n$, we have:

$$Kx = Y^TYx = Y^T(Yx) \tag{21}$$

Therefore, for any $y = Yx \in \mathbb{R}^r$ the column space of K, its range, is a linear combination of the columns of Y^T, therefore its dimension is at most r which is equivalent to $\operatorname{rank}(K) \leq r$.

($\Leftarrow$) Since K is symmetric, by the finite spectral theorem we have $K = Q\Lambda Q^T$ where Q is an orthogonal matrix, i.e. $QQ^T = Q^TQ = I$. Furthermore, since $\operatorname{rank}(K) \leq r$ we have at most r non-zero eigenvalues on the diagonal Λ. We can write:

$$K = Q\Lambda Q^T = \left(\begin{array}{c|c} Q'_{n\times r} & \end{array}\right)\left(\begin{array}{c|c} \Lambda'_{r\times r} & 0_{r\times(n-r)} \\ \hline 0_{(n-r)\times r} & 0_{(n-r)\times(n-r)} \end{array}\right)\left(\begin{array}{c} Q'^T_{r\times n} \\ \hline \end{array}\right) = Q'\Lambda'Q'^T \tag{22}$$

By the positive semidefinite property the eigenvalues in Λ' are non-negative. Thus, we can write $\Lambda' = \Lambda'^{1/2}\Lambda'^{1/2}$ to obtain:

$$Q'\Lambda'Q'^T = Q'(\Lambda'^{1/2})^T\Lambda'^{1/2}Q'^T = \underbrace{(\Lambda'^{1/2}Q'^T)^T}_{Y^T}\underbrace{\Lambda'^{1/2}Q'^T}_{Y} = Y^TY \tag{23}$$

□

The optimization problem finally becomes:

$$\begin{aligned} \max_{K} \quad & \mathbf{Tr}(K) \\ \text{s.t.} \quad & \mathbf{1}^TK\mathbf{1} = 0 \\ & K_{ii} + K_{jj} - 2K_{ij} = \|x_i - x_j\|^2, \text{ if } (i,j) \in E \\ & K = K^T \succeq 0 \\ & \operatorname{rank}(K) \leq r \end{aligned} \tag{24}$$

So, except for the last constraint on the rank which makes the problem non-convex, the remaining problem is an SDP. To move forward one relaxes the problem by simply dropping the rank constraint on K which means that we are enlarging the feasible set. The relaxed formulation (SDP) is then:

$$\begin{aligned} \max_{K} \quad & \mathbf{Tr}(K) \\ \text{s.t.} \quad & \mathbf{1}^T K \mathbf{1} = 0 \\ & K_{ii} + K_{jj} - 2K_{ij} = \|x_i - x_j\|^2, \text{ if } (i,j) \in E \\ & K = K^T \succeq 0 \end{aligned} \tag{25}$$

The solution of this problem gives us a matrix $K^\star$ where one of two possibilities occurs:

- If we actually have $\text{rank}(K) \leq r$ we recover Y through equation (23), $Y = \Lambda'^{1/2} Q'^T$. The solution comes from the top eigenvectors of K.

  Note that, for any rotation matrix R (which must satisfy $R^T R = I$), we have $K = Q'(\Lambda'^{1/2})^T R^T R \Lambda'^{1/2} Q'^T \implies Y = R\Lambda'^{1/2} Q'^T$, so the solution Y is unique up to rotations.

- If we are not in the above case, which implies that $\text{rank}(K) > r$, we just proceed as before and consider only the first r eigenvectors where $\lambda_{r+1} \neq 0$, with the possibility that the remaining ones are also non-zero. By reconstructing K through the top r eigenvectors we are projecting on the space of matrices with rank smaller or equal to r, however, by doing this we might be violating the remaining constraints.

  Just like in PCA, a large gap between the eigenvalues indicates that the manifold is well captured by only the first of these. This fact comes again from singular value decomposition since it provides the best rank r approximation to the matrix K.

Before proceeding let's rewrite the constraint on the edges in the following way (as the authors do):

$$K_{ii} + K_{jj} - 2K_{ij} = \|x_i - x_j\|^2 \Leftrightarrow \mathbf{Tr}(KE^{\{i,j\}}) = D_{ij}, \quad (i,j) \in E \tag{26}$$

where $E^{\{i,j\}}$ is a $n \times n$ matrix with the only non-zero entries being $E^{\{i,j\}}_{ii} = E^{\{i,j\}}_{jj} = 1$, $E^{\{i,j\}}_{ij} = E^{\{i,j\}}_{ji} = -1$ and $D_{ij} = \|x_i - x_j\|^2$. This can be seen via the following computation:

$$KE^{\{i,j\}} = K_{n\times n} \begin{array}{c} \\ 1 \\ \dots \\ i \\ \dots \\ j \\ \dots \\ n \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & \dots & i & \dots & j & \dots & n \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} = \begin{bmatrix} | & & K_{1i} - K_{1j} & | & & -K_{1i} + K_{1j} & | & & | \\ 0 & \dots & \vdots & 0 & \dots & \vdots & 0 & \dots & 0 \\ | & & K_{ii} - K_{ij} & | & & -K_{ii} + K_{ij} & | & & | \\ 0 & \dots & \vdots & 0 & \dots & \vdots & 0 & \dots & 0 \\ | & & K_{ji} - K_{jj} & | & & -K_{ji} + K_{jj} & | & & | \\ 0 & \dots & \vdots & 0 & \dots & \vdots & 0 & \dots & 0 \\ | & & K_{ni} - K_{nj} & | & & -K_{ni} + K_{nj} & | & & | \end{bmatrix} \tag{27}$$

and we see that the trace is precisely $K_{ii} - K_{ij} - K_{ji} + K_{jj} = K_{ii} + K_{jj} - 2K_{ij}$, since K is symmetric.

Duality on MVU

Recalling the SDP that we obtained (by relaxing the non-convex rank condition) and rewriting the edge constraints, we obtain the (Primal) MVU problem which is given by:

$$\begin{aligned} \max_{K} \quad & \mathbf{Tr}(K) \\ \text{s.t.} \quad & \mathbf{1}^T K \mathbf{1} = 0 \\ & \mathbf{Tr}(KE^{\{i,j\}}) = D_{ij}, \text{ if } (i,j) \in E \\ & K^T = K \succeq 0 \end{aligned} \tag{28}$$

Our goal now is to dualize this problem so we introduce first the dual variables for each constraint:

- $\mathbf{1}^T K \mathbf{1}$: We consider the dual variable $\nu \in \mathbb{R}$.
- $\mathbf{Tr}(KE^{\{i,j\}}) = D_{ij}$: For each $(i,j) \in E$ we consider the dual variable $W_{ij} \in \mathbb{R}$. We can write this as a matrix W where whenever $(i,j) \notin E$, $W_{ij} = 0$. Note that $W_{ij} = W_{ji}$, since it is associated to the same constraint.
- $K^T = K \succeq 0$: We introduce the dual variable $Z^T = Z \succeq 0$ which goes int the objective function as $+\mathbf{Tr}(KZ)$ since we want an upper bound since it is a maximization problem. Because $KZ \succeq 0$, the trace, which is equal to the sum of the eigenvalues, is positive. This comes from a generalization of the inner-product for matrices where $\langle B, A \rangle = \mathbf{Tr}(B^T A)$.

The Lagrangian function $\mathcal{L}(K, W, \nu, Z)$ is given by:

$$\begin{aligned}\mathcal{L}(K, W, \nu, Z) =& \mathbf{Tr}(K) + \mathbf{Tr}(KZ) - \nu \mathbf{1}^T K \mathbf{1} - \sum_{(i,j)\in E} W_{ij}(\mathbf{Tr}(KE^{\{i,j\}}) - D_{ij}) = \\ =& \mathbf{Tr}\left(K(I + Z - \nu \mathbf{1}\mathbf{1}^T - \sum_{(i,j)\in E} W_{ij} E^{\{i,j\}})\right) + \sum_{(i,j)\in E} D_{ij} W_{ij}\end{aligned} \tag{29}$$

where the last equality comes from the linearity of the trace and the computation:

$$\mathbf{Tr}(K\mathbf{1}\mathbf{1}^T) = \mathbf{Tr}\left(K \begin{bmatrix} | & & | \\ 1 & \dots & 1 \\ | & & | \end{bmatrix}\right) = \mathbf{Tr}\left(\begin{bmatrix} \sum_j K_{1j} & \dots & \sum_j K_{1j} \\ \vdots & \dots & \vdots \\ \sum_j K_{nj} & \dots & \sum_j K_{1j} \end{bmatrix}\right) = \sum_{i=1}^n \sum_{j=1}^n K_{ij} = \mathbf{1}^T K \mathbf{1} \tag{30}$$

The dual objective function is then given by:

$$g(W, \nu, Z) = \sup_K \mathcal{L}(K, W, \nu, Z) \tag{31}$$

The Lagrangian in eq. (29) is an affine function, analogous to $y = kx + d$ in 1 dimension, and thus the supremum is always $+\infty$ unless $(I + Z - \nu \mathbf{1}\mathbf{1}^T - \sum_{(i,j)\in E} W_{ij} E^{\{i,j\}}) = 0$. This can be seen by considering for example $Z = W = 0$ and $\nu = 0$ (since these are feasible) and then the first diagonal term in the product is K_{11} which we can take to $+\infty$. We have then:

$$g(W, \nu, Z) = \sup_K \mathcal{L}(K, W, \nu, Z) = \begin{cases} \sum_{(i,j)\in E} D_{ij} W_{ij}, \text{ if } I + Z - \nu \mathbf{1}\mathbf{1}^T - \sum_{(i,j)\in E} W_{ij} E^{\{i,j\}} = 0 \\ +\infty, \text{ otherwise} \end{cases} \tag{32}$$

And the (Dual) MVU problem is given by:

$$\begin{aligned} \min_{W,\nu,Z} \quad & \sum_{(i,j)\in E} D_{ij} W_{ij} \\ \text{s.t.} \quad & Z^T = Z \succeq 0 \\ & I + Z - \nu \mathbf{1}\mathbf{1}^T - L = 0, \qquad L = \sum_{(i,j)\in E} W_{ij} E^{\{i,j\}} \end{aligned} \tag{33}$$

Since Z is not part of the objective function its influence is to constraint the other variables, in particular we can write:

$$I - \nu \mathbf{1}\mathbf{1}^T - L = -Z \preceq 0 \tag{34}$$

where the term on the left is already symmetric by construction, so we do not lose any information by discarding Z from the problem. The only constraint on the dual formulation is then the LMI given by:

$$I - \nu \mathbf{1}\mathbf{1}^T - L \preceq 0 \tag{35}$$

We now show that this LMI is equivalent to the two following relations:

$$\nu \geq \frac{1}{n}, \qquad \lambda_{n-1}(L) \geq 1 \tag{36}$$

where λ_{n-1} represents the second smallest eigenvalue of L.

First recall that L is a symmetric matrix, since it is a weighted sum of the symmetric matrices $E^{\{i,j\}}$. Furthermore, by construction, it has an eigenvector $\mathbf{1} = [1 \dots 1]^T$ associated to the eigenvalue $\lambda_n = 0$. This can be seen from the following computation:

$$L\mathbf{1} = \begin{bmatrix} \sum_{j=1}^n L_{1j} \\ \vdots \\ \sum_{j=1}^n L_{nj} \end{bmatrix} = \begin{bmatrix} \sum_{(1,j)\in E} W_{1j} - W_{1j} \\ \vdots \\ \sum_{(n,j)\in E} W_{nj} - W_{nj} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{37}$$

Since L is a symmetric matrix we have an eigenvalue decomposition for L, by the finite spectral theorem, where $L = Q\Lambda Q^T$ with Q an orthogonal matrix, i.e. $QQ^T = Q^TQ = I$.

- $\nu \geq \frac{1}{n}$: Multiplying by $\mathbf{1}^T$ on the left and $\mathbf{1}$ on the right (which is equivalent to summing all entries of the matrix) we get:

$$\mathbf{1}^T \left(I - \nu \mathbf{1}\mathbf{1}^T - L\right) \mathbf{1} \preceq 0 \Leftrightarrow n - \nu n^2 - \underbrace{\mathbf{1}^T L \mathbf{1}}_{=0} \leq 0 \Leftrightarrow \nu \geq \frac{1}{n} \tag{38}$$

- $\lambda_{n-1}(L) \geq 1$: From the above, we can write $L = Q\Lambda Q^T$ and we know that $\mathbf{1}$ is an eigenvector of L with eigenvalue 0. Then, we can write:

$$L = \begin{bmatrix} | & & | \\ q_1 & \dots & q_n \\ | & & | \end{bmatrix} \left[\begin{array}{c|c} 0 & \\ \hline & B \end{array}\right] \begin{bmatrix} — & q_1^T & — \\ & \vdots & \\ — & q_n^T & — \end{bmatrix} \tag{39}$$

where the unit eigenvector q_1 is colinear with $\mathbf{1}$. Let $V_{n\times(n-1)}$ be the matrix with orthogonal columns $\{q_2, ..., q_n\}$ associated with the eigenvalues in the diagonal matrix B. We have then $V^TLV = B_{(n-1)\times(n-1)}$. From $L \succeq I - \nu\mathbf{1}\mathbf{1}^T$, eq. (35), we obtain:

$$V^TLV \succeq V^T(I - \nu\mathbf{1}\mathbf{1}^T)V \Leftrightarrow B \succeq \underbrace{V^TIV}_{=I} - \nu V^T\mathbf{1}\mathbf{1}^TV \Leftrightarrow B \succeq I \tag{40}$$

where the last term on the right vanishes since $\mathbf{1}$ is colinear with q_1 and therefore orthogonal to all rows of V^T. Since B is diagonal we conclude that all its eigenvalues are larger or equal to 1 which is equivalent to stating that the second smallest eigenvalue of L is larger or equal to 1, i.e. $\lambda_{n-1}(L) \geq 1$.

Note that we would still have to prove the backwards relation, i.e. that $\lambda_{n-1}(K) \geq 1$, $\nu \geq \frac{1}{n}$ and the fact that L is symmetric with an eigenvector $\mathbf{1}$ corresponding to eigenvalue 0 gives back the LMI in (35). This was not confirmed, but we believe that the same idea backwards would lead to the desired result.

From these computations, the dual problem can be formulated as:

$$\begin{aligned} \min_W \quad & \sum_{(i,j)\in E} D_{ij}W_{ij} \\ \text{s.t.} \quad & \lambda_{n-1}(L) \geq 1, \qquad L = \sum_{(i,j)\in E} W_{ij}E^{\{i,j\}} \end{aligned} \tag{41}$$

where ν is removed from the problem since it is not part of the objective function and does not constrain the other variables.

This problem is convex because the function $\lambda_{n-1}(L)$ is concave under the constraint $\lambda_n(L) = 0$, since we can always write it as the convex constraint $-\lambda_{n-1}(L) + 1 \leq 0$. We now show why this function is concave (following the argument in [8]), we have:

$$\lambda_{n+1}(L(W)) = \inf_{\|u\|=1, u^T\mathbf{1}=0} u^TL(W)u = \inf_{\|u\|=1, u^T\mathbf{1}=0} \sum_{(i,j)\in E} W_{ij}(u_i - u_j)^2 \tag{42}$$

and, being a pointwise infimum of a family of linear functions of W it is concave, see section 3.2 of [2]. This is the same idea used in PCA to obtain the second principal component where we ask it to have unit norm and be orthogonal to the first principal component, with $L(W)$ instead of the covariance matrix Σ.

Since both the objective function and the constraint are positive homogeneous the problem can be recast in the following form:

$$\begin{aligned} \max_W \quad & \lambda_{n-1}(L) \\ \text{s.t.} \quad & \sum_{(i,j)\in E} D_{ij}W_{ij} = c, \qquad L = \sum_{(i,j)\in E} W_{ij}E^{\{i,j\}} \end{aligned} \tag{43}$$

for arbitrary $c > 0$. The equivalence here is in the following sense: If $W^\star$ is the optimal solution of (41) with optimal value $c^\star$, then $\lambda^\star_{n-1} = \frac{c}{c^\star}$ is the optimal value of (43) with optimal solution $(\frac{c}{c^\star})W^\star$. This problem is seen to be convex since we are maximizing a concave function, seen in (42), under affine constraints. We now justify this equivalence.

Definition. Positive homogeneous function
A function $f : V \to W$, between vectors spaces over $\mathbb{R}$, is positive homogeneous if, for any $\alpha > 0$ $\exists k \in \mathbb{Z}$ such that, $f(\alpha x) = \alpha^k x$, for every $x \in V$.

The objective function in (41) is positive homogeneous of degree 1 since:

$$\sum_{(i,j)\in E} D_{ij}(\alpha W_{ij}) = \alpha \sum_{(i,j)\in E} D_{ij}W_{ij} \tag{44}$$

The constraint in (41) is positive homogeneous of degree 1, from the eigenvalue equation $Lx = \lambda x$ we have:

$$(\alpha L)x = \alpha Lx = \alpha\lambda x \tag{45}$$

Thus, any eigenvalue λ of L becomes $\alpha\lambda$, i.e. $\lambda_{n-1}(\alpha L) = \alpha\lambda_{n-1}(L)$.

We show this result over a general setting. Consider two positively homogeneous functions of degree 1, $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ where g is also continuous. We show how to convert the following problem on the left into the problem on the right:

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & g(x) \geq 1 \end{aligned} \qquad\qquad \begin{aligned} \max_x \quad & g(x) \\ \text{s.t.} \quad & f(x) = 1 \end{aligned}$$

To show this we further assume that the optimal value of the problem on the left, call it $p^\star$, satisfies $p^\star > 0$ and that the set of minimizers is non-empty, $x^\star \neq \emptyset$. By continuity of g at the optimal value we must have $g(x) = 1$. To see this, suppose that this was not the case, i.e. at $x^\star$ we have $g(x^\star) > 1$. Then, by continuity of g, there exists $\epsilon > 0$ and a ball of radius ϵ centered at $x^\star$ such that $g(x) > 1$ on that ball. Then, by reducing x towards the origin, the constraint will still be satisfied but, because f is positive homogeneous and always non-negative we are necessarily reducing its value, which contradicts the fact that $x^\star$ was the optimal value. This shows that the problem on the left is equivalent to the following problem:

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & g(x) = 1 \end{aligned} \tag{46}$$

Now we have equivalence of the following problems:

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & g(x) = 1 \end{aligned} \Leftrightarrow \begin{aligned} \min_x \quad & f(x)/g(x) \\ \text{s.t.} \quad & g(x) = 1 \end{aligned} \Leftrightarrow \begin{aligned} \min_x \quad & f(x)/g(x) \\ \text{s.t.} \quad & g(x) > 0 \end{aligned}$$

The first is clear since $g(x) = 1$. The second one comes from the fact that enlarging the feasible set in this way does not change the minimizers of the objective function. To see this note the following, the function $f(x)/g(x)$ will not change over a positive scaling since both f and g are positive homogeneous of the same degree. When defining $g(x) > 0$ we are allowing the function to be valued at any of these rays but these rays do not add any extra possible values for the ratio $f(x)/g(x)$. We now have the following equivalence:

$$\begin{aligned} \min_x \quad & f(x)/g(x) \\ \text{s.t.} \quad & g(x) > 0 \end{aligned} \Leftrightarrow \begin{aligned} \min_x \quad & f(x)/g(x) \\ \text{s.t.} \quad & g(x) > 0 \\ & f(x) > 0 \end{aligned}$$

This is equivalent because the set $\{x | g(x) > 0\} \subset \{x \,|\, f(x) > 0\}$ which implies that $\{x \,|\, g(x) > 0\} \cap \{x | f(x) > 0\} = \{x \,|\, g(x) > 0\}$ and thus the feasible set remains unchanged. To see why this is the case let $\bar{x}$ be a point such that $g(\bar{x}) > 0$ then:

$$g\left(\frac{\bar{x}}{g(\bar{x})}\right) = 1$$

by the positive homogeneous property of g. Therefore the point $\bar{x}/g(\bar{x})$ is feasible to the problem (46) which implies that $f(\frac{\bar{x}}{g(\bar{x})}) > 0 \implies f(\bar{x}) > 0$, by positive homogeneity of f. We have still the further equivalences:

$$\begin{array}{lll} \min\limits_{x} & f(x)/g(x) & \\ \text{s.t.} & g(x) > 0 & \\ & f(x) > 0 & \end{array} \Leftrightarrow \begin{array}{lll} \max\limits_{x} & g(x)/f(x) & \\ \text{s.t.} & g(x) > 0 & \\ & f(x) > 0 & \end{array} \Leftrightarrow \begin{array}{lll} \max\limits_{x} & g(x)/f(x) & \\ \text{s.t.} & f(x) > 0 & \end{array}$$

The first equivalence comes the fact that minimizing a ratio is the same as maximizing its reciprocal. On the second equivalence we are dropping a constraint which enlarges the feasible set. Contained within $f > 0$ we have the possibility that $g(x) > 0$ or $g(x) \leq 0$. However, since we are maximizing the ratio $\frac{g(x)}{f(x)}$ with $f(x) > 0$ the solution will always come from the part $g(x) > 0$. Thus, enlarging the set in this way does not alter the optimal solution of the problem. Now we finally have:

$$\begin{array}{lll} \max\limits_{x} & g(x)/f(x) & \\ \text{s.t.} & f(x) > 0 & \end{array} \Leftrightarrow \begin{array}{lll} \max\limits_{x} & g(x)/f(x) & \\ \text{s.t.} & f(x) = 1 & \end{array} \Leftrightarrow \begin{array}{lll} \max\limits_{x} & g(x) & \\ \text{s.t.} & f(x) = 1 & \end{array}$$

This equivalences now follow the same arguments we started with.

Now, this holds for our case because the primal MVU problem has an always non-negative function thus, since the dual function is always an upper bound by weak duality, the objective function of the dual must also be non-negative, i.e. the optimal value of (28), call it $c^\star$ satisfies $c^\star \geq 0$. Furthermore, the eigenvalue being given by a polynomial equation is also a continuous function. The relation between the optimal values and solutions of the problem can be seen via the following computations:

At the optimal solution of (41), we have:

$$\begin{aligned} &\sum_{(i,j)\in E} D_{ij} W^\star_{ij} = c^\star \\ &\lambda_{n-1}(L^\star) = 1 \\ &L^\star = \sum_{(i,j)\in E} W^\star_{ij} E^{\{i,j\}} \end{aligned} \tag{47}$$

By positive homogeneity, for any $c > 0$ the first equation is equivalent to:

$$\sum_{(i,j)\in E} D_{ij} W^\star_{ij} = c^\star \Leftrightarrow \sum_{(i,j)\in E} D_{ij} \left(\frac{c}{c^\star} W^\star_{ij}\right) = c \tag{48}$$

So, considering this equation as a constraint on a new variable $W' = \frac{c}{c^\star} W$, when it is satisfied we must have $\lambda_{n-1}(W'^\star) = \lambda_{n-1}(\frac{c}{c^\star} W^\star) = \frac{c}{c^\star} \lambda_{n-1}(W^\star) = \frac{c}{c^\star}$.

Primal-Dual Optimality conditions

The Primal problem (28) is a convex problem (SDP) thus, if Slater's condition holds, we have strong duality. The authors state that this is the case but we were not able to explicitly construct a strictly feasible K, i.e. $K \succ 0$ satisfying the first two constraints in problem (28). An alternative here would be to search for a Slater points via the dual formulation.

Under the assumption that Slater's condition holds this means that, at the optimal solutions $(K^\star, W^\star)$, we have zero duality gap:

$$\mathbf{Tr}(K^\star) = \sum_{(i,j)\in E} D_{ij} W^\star_{ij} \tag{49}$$

We consider the following statement from [6] as an assumption (since we are not familiar with the KKT conditions), "A pair $(K^\star, W^\star)$ is primal-dual optimal if and only if they satisfy the following KKT optimality conditions". We have:

- **Primal feasibility**: The pair $(K^\star, W^\star)$ must satisfy the primal problem constraints:

$$\begin{aligned} &K^\star = K^{\star T} \succeq 0 \\ &\mathbf{1}^T K^\star \mathbf{1} = 0 \\ &\mathbf{Tr}(K^\star E^{\{i,j\}}) = D_{ij}, \text{ if } (i,j) \in E \end{aligned} \tag{50}$$

- **Dual feasibility**: The pair $(K^\star, W^\star)$ must satisfy the dual problem constraints:

$$\begin{aligned} &\lambda_{n-1}(L^\star) \leq 1 \\ &L^\star = \sum_{(i,j)\in E} W^\star_{ij} E^{\{i,j\}} \end{aligned} \tag{51}$$

- **Complementary slackness**: It further satisfies:

$$L^\star K^\star = K^\star \tag{52}$$

However, equation (52) can also be seen from zero duality gap, see equation (10) in [6], when the equality holds.

We have seen that $\lambda_{n-1}(L^\star) = 1$ thus, equation (52) means the following: given any $y = K^\star x$, i.e. y is in the column space or range of $K^\star$, then we have:

$$L^\star K^\star x = K^\star x \Leftrightarrow L^\star y = y \tag{53}$$

Which means that any vector in the range of $K^\star$ is an eigenvector of $L^\star$ with eigenvalue $\lambda_{n-1}(L^\star) = 1$, in particular any eigenvector of $K^\star$ is an eigenvector of $L^\star$ with eigenvalue $\lambda_{n-1}(L^\star) = 1$. Since $L^\star$ is a sparse matrix (it only has non-zero elements whenever $(i,j) \in E$) and $K^\star$ is a dense matrix (each entry is an inner-product between datapoints) we can write this as:

$$\text{top e.s. of dense } K^\star \subseteq \text{bottom e.s. of sparse } L^\star \tag{54}$$

where bottom eigenspace here means the eigenspace associated with $\lambda_{n-1}(L^\star) = 1$ ($\lambda_n = 0$ associated with eigenvector $\mathbf{1}$ is discarded). Another consequence of (52) is the following:

$$r \leq \text{rank}(K^\star) \leq \text{(geometric) multiplicity of } \lambda_{n-1}(L^\star) \tag{55}$$

The inequality on the right is clear since any vector in the range of $K^\star$ must be in the eigenspace of $\lambda_{n-1}(L^\star) = 1$ and thus it is spanned by the independent eigenvectors associated with $\lambda_{n-1}(L^\star) = 1$ (geometric multiplicity).

The inequality on the left comes from considering r to be the relevant dimension coming from MDS applied on $K^\star$, i.e. they correspond to the top eigenvectors of $K^\star$ after which their magnitude drops but is not necessarily zero.

Isomap (Isometric mapping) [3]

We now present a very brief outline of the idea behind Isomap:

- **Step 1:** Given datapoints $\{x_i\}_{i=1}^n$ we first construct a graph $G = (V, E, w)$ wheres nodes are connected by an edge either by considering kNN or $\epsilon-$neighbourhoods. The (symmetric) weight function w will have entries $w_{ij} = w_{ji} = \|x_i - x_j\|_2$, i.e. whenever $(i,j) \in E$, w_{ij} holds the Euclidean distance between the pairs of points, which can be stored in a matrix W where $W_{ij} = 0$ whenever $(i,j) \notin E$.
- **Step 2:** For all pairs of points $(i,j) \notin E$ we estimate their geodesic distance (i.e. the shortest distance within the underlying submanifold) by considering the shortest distance between these points through the edges of the graph. The distances are added to the zero entries in W above.
- **Step 3:** From W, which is a matrix holding these pairwise distances, and equation (10) we compute a matrix G which is then used in the MDS optimization problem (8). The embedding dimension r is estimated by the number of significant eigenvalues of G and the outputs are given by (12).

Connection with MVU

Isomap can be seen as an approximation to the primal MVU problem. Given any two points x_i and x_j in the dataset, defining their geodesic distance along the underlying submanifold as $\text{gd}(x_i, x_j)$ we certainly have:

$$\text{gd}(x_i, x_j) \geq \|x_i - x_j\|_2 \tag{56}$$

Therefore, the total pairwise Euclidean distances is upper bounded by the total pairwise geodesic distances, i.e.:

$$\sum_{i,j} \text{gd}(x_i, x_j) \geq \sum_{i,j} \|x_i - x_j\|_2 \tag{57}$$

We have seen that the objective function of the primal MVU problem (28) is equivalent to maximizing pairwise distances. We can view Isomap as maximizing this variance by considering approximations to the geodesic distances, under its equivalence with PCA. In particular, in the limit where $n \to \infty$ (under a technical condition which guarantees convergence of the isomap algorithm, that the submanifold must be isometric to a convex subset of Euclidean space) MVU approaches the upper bound given by Isomap. Intuitively, when we are stretching the manifold the final Euclidean distance converges to the geodesic distance.

Locally Linear Embedding [5]

Locally Linear Embedding exploits the geometric notion that the neighbourhood of each point is closely approximated by its tangent space, provided that the submanifold is well approximated by the sample. Figure (2) shows the idea of approaching sections of the sphere by its tangent space.

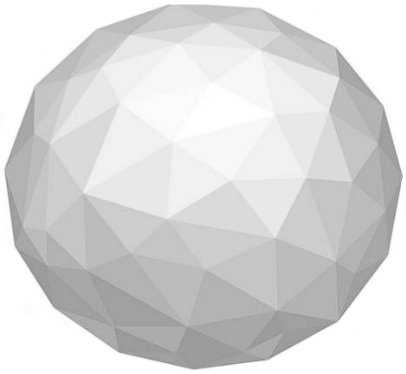

Figure 2: Local linear structure on a sphere. Source: (istockphoto.com)

The steps of the algorithm are the following:

- **Step 1:** Given datapoints $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, we first construct a graph $G = (V, E, w)$ wheres nodes are connected by an edge considering kNN. Contrary to the previous cases, of Isomap and MVU, the graph here is a directed graph. Thus $(i,j) \in E \not\Longrightarrow (j,i) \in E$, though it might be the case that both are in E. Define the set $\mathcal{N}_i = \{x_j | (i,j) \in E\}$ as the set of neighbours of point x_i.

- **Step 2:** LLE first determines the weights W_{ij} of each edge $(i,j) \in E$ by solving the following problem:

$$\begin{aligned} \min_{W_{ij}} \quad & \sum_{i=1}^{n} \|x_i - \sum_{x_j \in \mathcal{N}_i} W_{ij} x_j\|^2 \\ \text{s.t.} \quad & \sum_{x_j \in \mathcal{N}_i} W_{ij} = 1, \quad i = 1, ..., n \end{aligned} \tag{58}$$

  We are reconstructing each datapoint x_i via its neighbours where the constraint forces a convex combination of the neighbour points, i.e. we are approximating x_i via the affine space generated by the neighbours. According to [5], for any particular point the optimal weights are invariant under rotations, translations and scaling of the data and thus they manifest an intrinsic property of the submanifold.

- **Step 3:** The last step finds a lower dimensional embedding of the data $\{y_i\}_{i=1}^n$, $y_i \in \mathbb{R}^r$, $r << d$, by solving the following problem with the weights W_{ij} from the previous problem:

$$\begin{aligned} \min_{y_i} \quad & \sum_{i=1}^{n} \|y_i - \sum_{x_j \in \mathcal{N}_i} W_{ij} y_j\|^2 \\ \text{s.t.} \quad & \sum_i y_i = 0 \\ & \frac{1}{n} \sum_i y_i y_i^T = I \end{aligned} \tag{59}$$

The constraints, besides asking that the embeddings have zero mean, force a non-trivial solution to the problem. The solution to (59) is given by the bottom $r+1$ eigenvectors of the matrix $A := (I-W)^T(I-W)$. Suppose the normalized eigenvectors of A are $\{v_n, ..., v_{n-r}\}$ associated with the smallest eigenvalues $0 = \lambda_n < \lambda_{n-1} \leq ... \leq \lambda_{n-r}$. The first eigenvector $v_n = \frac{1}{\sqrt{n}}\mathbf{1}$ associated to $\lambda_n = 0$ is discarded and the solution is given by:

$$y_i = [(v_{n-1})_i \dots (v_{n-r})_i]^T, \quad i = 1, ..., n \tag{60}$$

The first eigenvector is discarded because it is constant, it assigns the same coordinate for all embedded points which adds an irrelevant extra dimension.

Connection with MVU

The main idea in LLE is that any point x_i can be closely approximated by its k neighbours, that is:

$$x_i \approx \sum_{x_j \in \mathcal{N}_i} W_{ij} x_j, \quad i = 1, ..., n \tag{61}$$

Letting $\tilde{Y} = [\tilde{y}_1 \dots \tilde{y}_n]$ be the outputs of MVU we have $K^\star = \tilde{Y}^T\tilde{Y}$. From (52) we obtain $L^\star\tilde{Y}^T = \tilde{Y}^T$, which can be written as:

$$\tilde{y}_i = \sum_{x_j \in \mathcal{N}_i} W^\star_{ij}(\tilde{y}_i - \tilde{y}_j) \tag{62}$$

We can check this from the following, we have $(L^\star\tilde{Y}^T)^T = \tilde{Y}(L^\star)^T = \tilde{Y}$ which is given by:

$$\begin{bmatrix} | & & | \\ \tilde{y}_1 & \dots & \tilde{y}_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ \tilde{y}_1 & \dots & \tilde{y}_n \\ | & & | \end{bmatrix} \begin{bmatrix} L^\star_{11} & \dots & L^\star_{n1} \\ L^\star_{12} & \dots & L^\star_{n2} \\ \vdots & & \vdots \\ L^\star_{1n} & \dots & L^\star_{nn} \end{bmatrix} \tag{63}$$

Therefore we have:

$$\tilde{y}_i = \sum_{j=1}^{n} L^\star_{ij}\tilde{y}_j = \sum_{x_j \in \mathcal{N}_i} W^\star_{ij}\tilde{y}_i - W^\star_{ij}\tilde{y}_j = \sum_{x_j \in \mathcal{N}_i} W^\star_{ij}(\tilde{y}_i - \tilde{y}_j) \tag{64}$$

Which can be rewritten as:

$$(L^\star_{ii} - 1)\tilde{y}_i = \sum_{x_j \in \mathcal{N}_i} W^\star_{ij}\tilde{y}_j, \quad i = 1, ..., n \tag{65}$$

this follows immediately from the first equality in (64) and the negative coefficient on the $\tilde{y}_j$.

We see then that equations (61) and (65) reflect the same notion of a local linear relationship up to a scaling factor $(L^\star_{ii} - 1)$ and the fact that $W^\star$ is symmetric while, in general, W will not be. This shows that the geometric idea behind LLE is embedded in MVU via the primal-dual optimality conditions.

Laplacian Eigenmaps [1]

Under the same framework as the previous problems the idea behind Laplacian Eigenmaps is the following:

- **Step 1:** As before one starts by constructing an undirected graph $G = (V, E, w)$ where nodes (datapoints) are connected either by kNN or $\epsilon-$neighbours, this is called the adjacency graph.
- **Step 2:** Next one chooses the weights where, according to [1], one typically considers two possibilities. Either one considers the heat kernel $W_{ij} = exp(-\|x_i - x_j\|^2/\sigma^2)$ or a simple formulation where the entries of connected nodes is $W_{ij} = 1$ and 0 otherwise.
- **Step 3:** The embeddings $y_i \in \mathbb{R}^r$ are found by solving the following optimization problem:

$$\begin{aligned} \min_{y_i} \quad & \sum_{(i,j)\in E} W_{ij}\|y_i - y_j\|^2 \\ \text{s.t.} \quad & \sum_i L_{ii} y_i y_i^T = I \end{aligned} \tag{66}$$

  where $L_{ii} = \sum_j W_{ij}$ is the diagonal element of the weighted Laplacian L (which is an extension of the usual Laplacian operator to the graph setting) and is equal to the degree of each vertex, i.e. the number of edges which connect to that vertex ($\deg(v_i)$). The idea behind this method is to keep points which were close still close after the embedding, where the constraint removes an arbitrary scaling factor.

The solution to (66) is obtained from the bottom $r+1$ unit eigenvectors of the generalized eigenvalue problem:

$$Lv_j = \lambda_j D v_j, \quad j = n, n-1, ..., n-r \tag{67}$$

where D is a diagonal matrix with $D_{ii} = \deg(v_i)$.

The solution to the generalized eigenvalue problem can be shown to be equivalent (not verified) to finding the bottom eigenvectors of $D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ and then scale them by $D^{-\frac{1}{2}}$. The embeddings are given as in LLE by:

$$y_i = [(v_{n-1})_i \dots (v_{n-r})_i]^T, \quad i = 1, ..., n \tag{68}$$

where again the first eigenvector is discarded by the same reasons as before.

Connection with MVU

Rewriting the dual MVU problem (41):

$$\begin{aligned} \min_{W} \quad & \sum_{(i,j)\in E} W_{ij} \|x_i - x_j\|^2 \\ \text{s.t.} \quad & \lambda_{n-1}(L) \geq 1, \qquad L = \sum_{(i,j)\in E} W_{ij} E^{\{i,j\}} \end{aligned} \tag{69}$$

We see that the objective function here has the same form as the objective function in (66) but over the initial datapoints. To find the embeddings one could first solve the dual MVU problem to find $W^\star$, solving (41), and then input $W^\star$ in (66). This method is precisely the one used in LLE, where one first estimates the weights (58) and then computes the outputs (59). We should note that the optimal weights in (66) can always be made feasible to the dual MVU problem (41) since the constraint $\lambda_{n-1}(L) \geq 1$ can always be satisfied by scaling up the weights, since the function is positive homogeneous, and this does not change the eigenvectors of L, i.e. the embedding is the same.

Furthermore, considering the Laplacian Eigenmaps problem (66) with weights given by this dual solution $W^\star$, the embeddings are obtained by the bottom eigenvectors of $L^\star(W^\star)$ which are solutions to the transformed dual MVU problem (43) and coincide with the top eigenvectors of $K^\star$ via the complementary slackness condition (52).

Conclusion

Despite the existence of several different methods for manifold learning, by using duality theory on the MVU problem we see there are significant connections between these methods.

We saw Isomap can be considered as an approximation to the optimal solution of the primal MVU problem, with convergence under certain conditions. That the objective function of LLE, describing the local linear relationships, can be retrieved via the dual MVU problem. With Laplacian Eigenmaps we saw its solution via the bottom eigenvectors relates to the dual MVU problem and connects to the top eigenspace of the solution via the Primal MVU problem. Furthermore, by using the optimal weights $W^\star$ from the dual formulation of MVU in the Laplacian Eigenmaps problem we obtain a similar two-step procedure as in LLE.

To conclude, we confirm that duality is an incredible tool which allows to connect methods which were a priori unrelated.

References

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] Joshua B. Tenenbaum, Vin de Silva and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

[4] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[5] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000.

[6] L. Xiao, J. Sun and S. Boyd. A duality view of spectral methods for dimensionality reduction. *International Conference on Machine Learning (ICML)*, 2006.

[7] Kilian Q. Weinberger and Lawrence K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *IEEE - Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[8] Jun Sun, Stephen Boyd, Lin Xiao and Persi Diaconis. The fastest mixing markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM*, 2006.