Latent Smooth Skeleton Embedding

Li Wang
Department of Mathematics,
Statistics and Computer Science
University of Illinois at Chicago
liwang8@uic.edu

Qi Mao
HERE Company
qimao.here@gmail.com

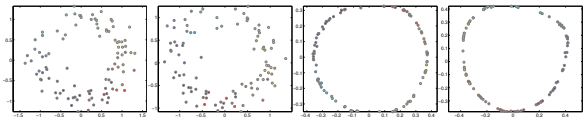
Ivor W. Tsang
Centre for Artificial Intelligence
University of Technology Sydney
Ivor.Tsang@uts.edu.au

Abstract

Learning a smooth skeleton in a low-dimensional space from noisy data becomes important in computer vision and computational biology. Existing methods assume that the manifold constructed from the data is smooth, but they lack the ability to model skeleton structures from noisy data. To overcome this issue, we propose a novel probabilistic structured learning model to learn the density of latent embedding given high-dimensional data and its neighborhood graph. The embedded points that form a smooth skeleton structure are obtained by maximum a posteriori (MAP) estimation. Our analysis shows that the resulting similarity matrix is sparse and unique, and its associated kernel has eigenvalues that follow a power law distribution, which leads to the embeddings of a smooth skeleton. The model is extended to learn a sparse similarity matrix when the graph structure is unknown. Extensive experiments demonstrate the effectiveness of the proposed methods on various datasets by comparing them with existing methods.

In many fields of science and engineering, one is often confronted with the problem of dimensionality reduction (Burges 2009; Van der Maaten, Postma, and van den Herik 2009). The problem aims to extract low-dimensional structures from high-dimensional datasets, which are generally characterized by much fewer degrees of freedom than actual number of features.

In this paper, we are particularly interested in unveiling a smooth skeleton structure in a latent space from data with noise. Figure 1 illustrates an intuitive example in which synthetic data points are drawn from a smooth circle with noises in two-dimensional space. It is challenging to recover the circle (Figures 1(c) and 1(d)) from the noisy data without any prior knowledge of the structure. Datasets with a smooth skeleton structure have become widely accessible in computer vision (Weinberger and Saul 2006) and computational biology (Curtis et al. 2012). In the study of human cancer, a widely accepted hypothesis is that human cancer is a dynamic disease developed over an extended period with the accumulation of genetic alterations (Greaves and Maley 2012). The evolution trajectories of tumor persistence, growth, and ultimately metastasis, are complex and branching (Greaves and Maley 2012). Massive molecular profile data from excised tumor tissue makes it feasible to uncover the branching architecture of cancer evolution (Mao et al. 2015). However, learning a smooth branching structure embedded in a low-dimensional space from high-dimensional noisy datasets poses a great challenge.



(a) Original data (b) MVU (c) PSL-ℓ_2 (d) PSL-ℓ_1

Figure 1: A synthetic example illustrating the motivation for unveiling the smooth skeleton structure from noisy data. The data is drawn from a circle with added noise. Each point is colored for the purpose of illustration. Our two proposed methods PSL-ℓ_2 and PSL-ℓ_1 are compared with MVU.

Existing methods mostly rely on distances (or similarities) to model the intrinsic structure of data. They either provide a similarity matrix as a prior (Belkin and Niyogi 2001; Schölkopf, Smola, and Muller 1999), or learn a similarity measurement based on a subset of distances in a local region (Elhamifar and Vidal 2011; Saul and Roweis 2003), or directly learn a kernel matrix from data (Weinberger, Packer, and Saul 2005; Xiao, Sun, and Boyd 2006; Mao and Tsang 2010). These distances become unreliable if the data is noisy. Moreover, they lack the ability to model a smooth skeleton from noisy data. As shown in Figure 1, the strict distance preservation in maximum variance unfolding (MVU) (Weinberger, Sha, and Saul 2004) fails to capture the smooth circle from the data (see Figure 1 (b)).

We aim to learn a smooth skeleton from noisy data. To achieve this goal, we first present expected distances of embedded data points following an unknown density. We then propose a novel probabilistic structured learning model to learn the density of latent embedding variables given high-dimensional data. The main contributions of this paper are:

1) By directly modeling the unknown density of embedding latent variables, the proposed model can be considered as a probabilistic version of MVU. The embedding data is obtained by MAP estimation.

2) Our duality analysis shows that the unknown density has an analytic solution in the form of a sparse similarity

matrix or a regularized Laplacian kernel (Smola and Kondor 2003), and the eigenvalues of the kernel learned by the proposed model follows a power law distribution, which leads to a smooth skeleton of embedded points, while the duality view of MVU cannot achieve this.

3) The proposed model possesses a variety of advantageous properties from probabilistic and discriminative viewpoints, including the robust representation of expected distances, easy extension for error tolerance, model selection of neighborhood structures, and global optimum of the resulting convex optimization problems.

4) We further extend the proposed model for two settings: a neighborhood graph is given as a priori but distances are noisy, and the graph is unknown. An efficient alternating direction of multiplier method (ADMM) is proposed to handle an optimization problem that generalizes both cases.

Related Work

Let $\mathbb{Y} = \{\mathbf{y}_i\}_{i=1}^N$ be a set of N data points where $\mathbf{y}_i \in \mathbb{R}^D$. MVU aims to find a set of embedded data points $\mathbb{X} = \{\mathbf{x}_i\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $d < D$ such that the variance of the embedded points is maximized subject to constraints such that distances between nearby data points are preserved (Weinberger, Sha, and Saul 2004).

MVU consists of three steps. The first step is to compute the k-nearest neighbors $\mathcal{N}_i$ of data point $\mathbf{y}_i, \forall i$. Let $\phi_{i,j} = ||\mathbf{y}_i - \mathbf{y}_j||^2$ and $D_{i,j} = ||\mathbf{x}_i - \mathbf{x}_j||^2$. The second step is to solve the following optimization problem given by

$$\max_{\mathbb{X}} \sum_{i=1}^N ||\mathbf{x}_i||^2 : \sum_{i=1}^N \mathbf{x}_i = 0, D_{i,j} = \phi_{i,j}, \forall i, j \in \mathcal{N}_i, \quad (1)$$

where the first constraint eliminates the translational degree of freedom on the embedded data points by constraining them to be centered at the origin; the remaining constraints preserve distances between k-nearest neighbors. Instead of optimizing $\mathbb{X}$, MVU reformulates (1) as a semidefinite programming by learning a kernel matrix $\mathbf{K}$ with the (i,j)th element denoted by $K_{i,j} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ with a semidefinite constraint $\mathbf{K} \succeq 0$ for a valid kernel (Scholkopf and Smola 2001). We have $D_{i,j} = ||\mathbf{x}_i - \mathbf{x}_j||^2 = K_{i,i} + K_{j,j} - 2K_{i,j}$. The resulting problem is given by

$$\max_{\mathbf{K}} \mathrm{Tr}(\mathbf{K}) : \sum_{i,j} K_{i,j} = 0, \mathbf{K} \succeq 0, D_{i,j} = \phi_{i,j}, \forall i, j \in \mathcal{N}_i,$$

where $\sum_{i,j} K_{i,j} = \left\langle \sum_{i=1}^N \mathbf{x}_i, \sum_{j=1}^N \mathbf{x}_j \right\rangle = 0$ is a relaxed constraint for ease of kernelization. The final step is to obtain the embedding $\mathbb{X}$ by applying KPCA (Schölkopf, Smola, and Muller 1999) on the optimal $\mathbf{K}$.

A duality view of the MVU problem has been studied in (Xiao, Sun, and Boyd 2006). Define $\mathbf{E}^{i,j}$ as an $N \times N$ matrix consisting of only four nonzero elements: $\mathbf{E}^{i,j}[i,i] = \mathbf{E}^{i,j}[j,j] = 1, \mathbf{E}^{i,j}[i,j] = \mathbf{E}^{i,j}[j,i] = -1$. The preserving constraints can be rewritten as $\mathrm{Tr}(\mathbf{K}\mathbf{E}^{i,j}) = \phi_{i,j}, \forall i, j \in \mathcal{N}_i$. Thus, the dual MVU problem is

$$\min \sum_{i,j \in \mathcal{N}_i} w_{i,j}\phi_{i,j} : \lambda_{N-1}(\mathbf{L}) \geq 1, \mathbf{L} = \sum_{i,j \in \mathcal{N}_i} w_{i,j}\mathbf{E}^{i,j}, \quad (2)$$

where $w_{i,j}$ is the dual variable subject to the preserving constraint associated with edge (i,j), and λ_{N-1} denotes the second smallest eigenvalue of a symmetric matrix.

Expected Distance Preserving

Even though MVU has been successfully applied to a number of datasets, two challenging problems exist. 1) The distances over noisy data lack the reliability for modeling embedded points, and the manifold constructed over these distances may not directly reflect a smooth skeleton structure. Thus, it is challenging to learn a smooth skeleton structure from noisy data in high dimension. 2) The k-nearest neighbor graph might be not adequate for modeling points with large variances in different regions (Elhamifar and Vidal 2011). It is important to automatically learn a neighborhood graph so as to better approximate the true structure. We resolve these issues by proposing a novel probabilistic model, which can be viewed as a probabilistic version of MVU.

Proposed Probabilistic Model

As the embedding $\mathbb{X}$ is the variable of interest, we treat it as a random variable. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N \times d}$ be a matrix of embedded points. Define $p(\mathbf{X})$ as the density of embedded points, and the bases of the embedding space are assumed to be independent so that $p(\mathbf{X}) = \prod_{k=1}^d p(\mathbf{f}_k)$ where $[\mathbf{f}_1, \ldots, \mathbf{f}_d] = \mathbf{X}^T$. This assumption is commonly used in spectral methods. We can now reformulate the distance of embedded points $\mathbf{x}_i$ and $\mathbf{x}_j$ as the expected distance with respect to density $p(\mathbf{X})$ given by $\mathbb{E}[||\mathbf{x}_i - \mathbf{x}_j||^2] = \sum_{k=1}^d \int (f_{i,k} - f_{j,k})^2 p(\mathbf{f}_k) \mathrm{d}\mathbf{f}_k$, where the equality holds due to $||\mathbf{x}_i - \mathbf{x}_j||^2 = \sum_{k=1}^d (f_{i,k} - f_{j,k})^2$ and the independence over features. The centralized constraint can be reformulated for each reduced dimension as $\mathbb{E}[\sum_{i=1}^N f_{i,k}] = \int \mathbf{1}^T \mathbf{f}_k p(\mathbf{f}_k) \mathrm{d}\mathbf{f}_k = 0, \forall k$. In addition, we assume that the prior distribution $p_0(\mathbf{X}) = \prod_{k=1}^d p_0(\mathbf{f}_k)$ where $p_0(\mathbf{f}_k)$ is a multivariate normal distribution with zero mean and covariance matrix $\gamma^{-1}\mathbf{I}$, i.e. $\mathbf{f}_k \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$. In order to learn $p(\mathbf{X})$ from data $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{D \times N}$ given a set of expected constraints, the principle of maximum entropy is used. Given a prior distribution and a neighborhood graph, we minimize the following optimization problem

$$\begin{aligned} \min_{\{p(\mathbf{f}_k) \in \mathcal{P}_k\}_{k=1}^d} & \sum_{k=1}^d \int p(\mathbf{f}_k) \log \frac{p(\mathbf{f}_k)}{p_0(\mathbf{f}_k)} \mathrm{d}\mathbf{f}_k \quad (3) \\ \text{s.t. } & \mathbb{E}[||\mathbf{x}_i - \mathbf{x}_j||^2] = \phi_{i,j}, \forall i, j \in \mathcal{N}_i, \\ & \mathbb{E}\left[\sum_{i=1}^N f_{i,k}\right] = 0, \forall k, \end{aligned}$$

where $\mathcal{P} = \times_{k=1}^d \mathcal{P}_k$ is a Cartesian product of d i.i.d. probability spaces and $\mathcal{P}_k = \{\int p(\mathbf{f}_k)\mathrm{d}\mathbf{f}_k = 1, p(\mathbf{f}_k) \geq 0\}$ is a feasible set of density functions.

Proposition 1. *Problem (3) has an analytic solution*

$$p(\mathbf{f}_k) = \mathcal{N}\big(0, (\mathbf{L} + \gamma\mathbf{I})^{-1}\big), \forall k, \quad (4)$$

where Laplacian matrix $\mathbf{L} = diag(\mathbf{W}\mathbf{1}) - \mathbf{W}$, the (i,j)th element $w_{i,j}$ of $\mathbf{W}$ is the dual variable of its corresponding distance preserving constraint, and the dual optimization problem is convex given by

$$\begin{aligned} \max_{\mathbf{W}} & \frac{d}{2} \log\det(\mathbf{L} + \gamma\mathbf{I}) - \frac{1}{4}\langle \mathbf{W}, \mathbf{\Phi} \rangle \quad (5) \\ s.t. & \mathbf{L} + \gamma\mathbf{I} \succ 0, w_{i,j} = 0, \forall i, j \notin \mathcal{N}_i, \end{aligned}$$

where $\mathbf{\Phi}$ is a distance matrix with the (i,j)th element as $\phi_{i,j}$ if for any $i,j \in \mathcal{N}_i$ and ∞ otherwise, and $\langle \cdot, \cdot \rangle$ is the standard trace inner product.

According to Proposition 1, there are several interesting properties. First, the zero mean constraint holds automatically in (3). This can be verified by the posterior distribution with zero mean. Second, our model can obtain smooth skeleton structure of embedding, which will be discussed in details below.

Analyses from Spectrum and Optimization

The objective function contains log-determinant of $\mathbf{L} + \gamma\mathbf{I}$, which can be equivalently formulated as $\log\det(\mathbf{L} + \gamma\mathbf{I}) = \sum_{i=1}^{N} \log(\lambda_i(\mathbf{L}) + \gamma)$, where λ_i denotes the ith largest eigenvalue of a symmetric matrix. Thus, the log determinant can be related to the negative log-likelihood of a power law distribution of λ_i as $p(\lambda_i) \propto \lambda_i^{-\theta}$ where θ is called the power law exponent, and γ is a positive term used to make $\lambda_i + \gamma > 0$. The power law distribution imposes large values on a small set of eigenvalues, while the remaining eigenvalues have small values. In the proposed model, $\theta = \frac{d}{2}$. This is critically different from the dual MVU problem where the second smallest eigenvalue is maximized (Xiao, Sun, and Boyd 2006). The positive term γ represents the sensitivity threshold for the parameters and is used to smooth out the scale-free property (Liu and Ihler 2011). If $\gamma \gg \lambda_i$, we have $\log(\lambda_i + \gamma) \approx \frac{1}{\gamma}\lambda_i + \log(\gamma)$. This generalizes the ℓ_1 regularizer over the eigenvalues. According to the above spectral analysis, the proposed model puts more emphasis on the whole spectrum of the Laplacian matrix following a power law distribution. The difference between MVU and our model will be illustrated in the experiments.

From an optimization perspective, the proposed unfolding framework provides a novel approach to learn a sparse similarity matrix $\mathbf{W}$ automatically from a set of pairwise distances, and the similarity matrix is intentionally designed for learning the embedded points that achieve a smooth skeleton structure because: (i) The proposed formulation (3) learns a posterior distribution of embedded points since the expected distance of these points with respect to the posterior distribution has much more flexibility than the deterministic distances used in MVU. (ii) The smoothness of the manifold structure is achieved by the expected constraints over an infinite number of possible candidates of embedded points, where distances can be varied flexibly so that these distances may not be strictly preserved. In contrast, deterministic constraints used in MVU are strictly preserved. (iii) The not-restrictively-preserved constraints allow the embedded points to move flexibly to form a smooth skeleton in terms of $\langle \mathbf{W}, \Phi \rangle$ in (5). In other words, by minimizing $\langle \mathbf{W}, \Phi \rangle$, if two points are close on the given graph, their corresponding embedded points are also close.

Embedding via MAP Estimation

Once $\mathbf{W}$ has been obtained, the posterior distribution of embedding is explicitly represented as $p(\mathbf{X}) = \prod_{k=1}^{d} p(\mathbf{f}_k)$, which is the same as the matrix normal distribution (Gupta and Nagar 1999) given by $p(\mathbf{F}|\mathbb{Y}) \sim \mathcal{MN}_{N,d}(\mathbf{0}, \mathbf{U}, \mathbf{I})$, where $\mathbf{U} = (\mathbf{L} + \gamma\mathbf{I})^{-1}$ is the sample-based covariance matrix and can also be interpreted as a regularized Laplacian kernel with regularization parameter γ (Smola and Kondor 2003). Given the posterior distribution, we can obtain the point estimate of $\mathbf{X}$ by using MAP estimation. Thus, from probabilistic point of view, we can apply KPCA on $\mathbf{U}$ to achieve the embedded data points similar to MVU for embedding. For reference, we name the proposed model as Probabilistic Structured Learning (PSL).

Latent Smooth Skeleton Embedding

In addition to imposing constraints for strictly preserving distances, we also consider variants of the difference between two pairwise distances. One is to tolerate the errors of distances on the edges of a given neighborhood graph, the other is to learn the neighborhood graph from data. Both relaxations can be formulated according to the maximum entropy density estimation with generalized regularization (Dudík, Phillips, and Schapire 2007). Next, we propose the two variants of PSL to learn smooth skeleton embedding on noisy high-dimensional data.

Known Neighborhood Structure

Suppose that a neighborhood graph $\mathcal{G}$ is known in advance and can reliably capture the underlying structure of the data. However, the distance $\phi_{i,j}$ is not reliable due to noisy samples or features. In this case, we aim to obtain a set of embedded points that can preserve the distances with a penalty on the violated pair of points corresponding to an edge in $\{(i,j) : \forall i, j \in \mathcal{N}_i\}$. By introducing variable $\xi_{i,j}$ for the distance violation on edge (i,j) and parameter $C > 0$, PSL with ℓ_2 regularization (PSL-ℓ_2) can be formulated as

$$\min_{\{p(\mathbf{f}_k)\in\mathcal{P}_k\}_{k=1}^{d}} \sum_{k=1}^{d} \int p(\mathbf{f}_k) \log \frac{p(\mathbf{f}_k)}{p_0(\mathbf{f}_k)} \mathrm{d}\mathbf{f}_k + \frac{C}{2} \sum_{i,j\in\mathcal{N}_i} \xi_{i,j}^2 \quad (6)$$
$$\text{s.t. } \mathbb{E}[||\mathbf{x}_i - \mathbf{x}_j||^2] - \phi_{i,j} = \xi_{i,j}, \forall i, j \in \mathcal{N}_i.$$

Proposition 2. *Problem (6) has an analytic solution (4) where $\mathbf{W}$ can be obtained by solving*

$$\max_{\mathbf{W}} \frac{d}{2} \log\det(\mathbf{L} + \gamma\mathbf{I}) - \frac{1}{4}\langle \mathbf{W}, \mathbf{\Phi} \rangle - \frac{1}{2C}||\mathbf{W}||_F^2 \quad (7)$$
$$\text{s.t. } \mathbf{L} + \gamma\mathbf{I} \succ 0, w_{i,j} = 0, \forall i, j \notin \mathcal{N}_i,$$

where $||\mathbf{W}||_F$ is the Frobenius norm of $\mathbf{W}$, $\mathbf{\Phi}$ is a matrix with the (i,j)th element as $\phi_{i,j}$ if for any $i,j \in \mathcal{N}_i$ and ∞ otherwise. The violation can be computed as $\xi_{i,j} = \frac{w_{i,j}}{C}$.

According to Proposition 2, a large C imposes a small violation of pairwise distances. As $C \to \infty$, problem (6) is equivalent to (3). Note that MVU was extended for data without noise by introducing a tolerance term similar to (6) (Weinberger and Saul 2006), whose motivation was only to overcome the "lock up" effect of embedding data if strictly preserving distances is impossible for data without noise.

Unknown Neighborhood Structure

If the data does not provide a reliable neighborhood structure as a prior, we propose to automatically learn a sparse graph by imposing sparsity on $\mathbf{W}$. To achieve this goal, we modify the constraints in (6) such that the absolute difference

of distances between $\mathbb{E}[||\mathbf{x}_i - \mathbf{x}_j||^2]$ and $\phi_{i,j}$ is restricted in a range between $-\beta\phi_{i,j}$ and $\beta\phi_{i,j}$, which is the scaled distance by β. In other words, we can interpret the parameter β as a tolerance for the deviation of an embedding distance from its associated observed distance, which we want to preserve. This variant of the MAP unfolding with the defined constraints called PSL with ℓ_1 regularization (PSL-ℓ_1) is formulated as

$$\min_{\{p(\mathbf{f}_k)\in\mathcal{P}_k\}_{k=1}^d} \sum_{k=1}^{d} \int p(\mathbf{f}_k) \log \frac{p(\mathbf{f}_k)}{p_0(\mathbf{f}_k)} \mathrm{d}\mathbf{f}_k \tag{8}$$
$$\text{s.t. } \left|\mathbb{E}[||\mathbf{x}_i - \mathbf{x}_j||^2] - \phi_{i,j}\right| \leq \beta\phi_{i,j}, \forall i, j.$$

Proposition 3. *Problem (8) has an analytic solution (4) where* $\mathbf{W}$ *can be obtained by solving*

$$\max_{\mathbf{W}} \frac{d}{2} \log\det(\mathbf{L} + \gamma\mathbf{I}) - \frac{1}{4}\langle \mathbf{W}, \mathbf{\Phi}\rangle - \beta||\mathbf{\Phi} \odot \mathbf{W}||_1 \tag{9}$$
$$\text{s.t. } \mathbf{L} + \gamma\mathbf{I} \succ 0$$

where $||\mathbf{\Phi} \odot \mathbf{W}||_1 = \sum_{i=1}^N \sum_{j=1}^N \phi_{i,j}|w_{i,j}|$ *and* $\mathbf{\Phi}$ *is a matrix with the* (i,j)*th element as* $\phi_{i,j}$*.*

Proposition 3 shows that the objective function of (9) leads to a sparse similarity matrix due to the ℓ_1 regularization. That is, $w_{i,j}$ approaches to 0 if $\phi_{i,j}$ is large. This is consistent with the intuition that two data points are dissimilar if they are distant. Problem (9) is similar to the model proposed for sparse structure learning (Lake and Tenenbaum 2010). The key difference is that the latter model has a nonnegative constraint on $\mathbf{W}$, i.e. $\mathbf{W} \geq 0$, while our model does not have this constraint. In addition, MEU (Lawrence 2012) for estimating graph structure by imposing ℓ_1 regularizer over the kernel matrix based on the assumption that original features are identically independent. However, our model takes weighted ℓ_1 regularizer over $\mathbf{W}$ and is derived from the independence of latent variables, which is commonly used in spectral methods. A very recent work (Mao, Wang, and Tsang 2016) can learn a sparse similarity matrix for shrinkage effect, but our method effectively controls the deviation of distances using the absolute difference.

Optimization Algorithm

We propose to solve the following optimization problem

$$\min_{\mathbf{W},\mathbf{G}} \langle \mathbf{W}, \mathbf{A}\rangle + \Omega(\mathbf{W}) - \log\det(\mathbf{G}) \tag{10}$$
$$\text{s.t. } \mathbf{G} = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W} + \gamma\mathbf{I}, \mathbf{G} \succ \mathbf{0}, \mathbf{W} \in \mathcal{S},$$

where $\mathcal{S}$ is a set of symmetric matrices. It is easy to see that problem (10) is a generic problem of (5), (7), and (9).

ADMM (Boyd et al. 2011) is employed to solve (10). By introducing $\mathbf{Z} \in \mathbb{R}^{N\times N}$ as the multiplier of the linear matrix equation and $\tau > 0$ as the penalty parameter for the violation of the linear matrix constraint, we have an augmented Lagrangian function of problem (10) as $L_\tau(\mathbf{G}, \mathbf{W}, \mathbf{Z}, \tau) = \langle \mathbf{W}, \mathbf{A}\rangle + \Omega(\mathbf{W}) - \log\det(\mathbf{G}) - \langle \mathbf{Z}, \mathbf{G} - g_\gamma(\mathbf{W})\rangle + \frac{\tau}{2}||\mathbf{G} - g_\gamma(\mathbf{W})||_F^2$, where $g_\gamma(\mathbf{W}) = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W} + \gamma\mathbf{I}$. ADMM iteratively solves following subproblems until convergence:

$$\mathbf{G}^{(t+1)} = \arg\min_{\mathbf{G}\succ 0} \frac{\tau}{2}||\mathbf{G} - \mathbf{Q}^{(t)}||_F^2 - \log\det(\mathbf{G}), \tag{11}$$

$$\mathbf{W}^{(t+1)} = \arg\min_{\mathbf{W}\in\mathcal{S}} \langle \mathbf{W}, \mathbf{A}\rangle + \Omega(\mathbf{W}) + \frac{\tau}{2}||g_\gamma(\mathbf{W}) - \mathbf{P}^{(t)}||_F^2, \tag{12}$$

$$\mathbf{Z}^{(t+1)} = \mathbf{Z}^{(t)} - \tau(\mathbf{G}^{(t+1)} - g_\gamma(\mathbf{W}^{(t+1)})), \tag{13}$$

where $\mathbf{Q}^{(t)} = g_\gamma(\mathbf{W}^{(t)}) + \frac{1}{\tau}\mathbf{Z}^{(t)}$ and $\mathbf{P}^{(t)} = \mathbf{G}^{(t+1)} - \frac{1}{\tau}\mathbf{Z}^{(t)}$. Next, we discuss detailed methods for solving the above subproblems separately.

First, problem (11) has an analytic solution, which is shown in the following proposition.

Proposition 4. *Let the eigen-decomposition of* $\mathbf{Q}^{(t)}$ *be* $\mathbf{Q}^{(t)} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ *where* $\mathbf{V}$ *is an orthogonal matrix whose column vectors are eigenvectors of* $\mathbf{Q}^{(t)}$ *and* $\mathbf{\Lambda} = diag([\lambda_i])$ *is a diagonal matrix with corresponding eigenvalues. The optimal solution of (11) is* $\mathbf{G}^{(t+1)} = \mathbf{V}diag([\widetilde{\lambda}_i])\mathbf{V}^T$ *where*

$$\widetilde{\lambda}_i = \frac{\lambda_i}{2} + \sqrt{\frac{\lambda_i^2}{4} + \frac{1}{\tau}}, \forall i. \tag{14}$$

Next, we solve subproblem (12). Before detailing the optimization method, we present the following lemma, which reduces the number of variables to be optimized.

Lemma 1. *Given* $\mathbf{A}$*, problem (12) with* Ω *defined above has* $w_{i,i} = 0, \forall i$ *at the optimum.*

According to Lemma 1 and the symmetry property of $\mathbf{W}$, the optimization problem (12) can be rewritten as a problem with respect to $\mathcal{W} = \{w_{i,j} : \forall (i,j) \in \mathcal{E}_S\}$ where $\mathcal{E}_S = \{(i,j) : \forall i, j < i \wedge j \in \mathcal{N}_i\}$. Let $\mathbf{w}$ be a vectorized representation of optimized variables in $\mathcal{W}$ by applying a mapping function $\varphi(i,j) = \ell$ that maps indexes $(i,j) \in \mathcal{E}_S$ to the ℓth element of $\mathbf{w}$. As a result, problem (12) can be written as $\min_{\mathbf{w}} h^{(t)}(\mathbf{w}) + \Omega(\mathbf{w})$, where

$$h^{(t)}(\mathbf{W}) = \langle \mathbf{W}, \mathbf{A}\rangle + \frac{\tau}{2}||g_\gamma(\mathbf{W}) - \mathbf{P}^{(t)}||_F^2. \tag{15}$$

The following proposition shows that the objective function (15) is a quadratic function with respect to $\mathbf{w}$.

Proposition 5. *Let* $\mathbf{w}$ *be the vectorized representation of variables in* $\mathcal{W}$ *with the mapping function* φ *defined above. Define* ϕ *and* $\mathbf{p}$ *as the vectorizations of* $\mathbf{\Phi}$ *and* $\mathbf{P}^{(t)}$ *with the same mapping function, respectively. We have that minimizing the objective function (15) with respect to* $\mathbf{W}$ *is equivalent to minimizing the following quadratic function with respect to* $\mathbf{w}$ *given by*

$$h^{(t)}(\mathbf{w}) = \frac{\tau}{2}\mathbf{w}^T\mathbf{B}\mathbf{w} + \tau\mathbf{w}^T\mathbf{c}$$

where $\mathbf{B} = \sum_i \mathbf{b}_i\mathbf{b}_i^T + 2\mathbf{I}$*,* $\mathbf{b}_i$ *is a binary vector with the* ℓ*th element, and for* $j \neq i$*, if* $j < i$*,* $\mathbf{b}_i[\varphi(i,j)] = 1$*, and if* $j > i$*,* $\mathbf{b}_i[\varphi(j,i)] = 1$*, and the remaining elements of* $\mathbf{b}_i$ *are* 0*s, and* $\mathbf{c} = 2\mathbf{p} + \sum_i(\gamma - p_{i,i})\mathbf{b}_i + \frac{1}{d\tau}\boldsymbol{\phi}$ *.*

According to Proposition 5, we can equivalently transform the constrained optimization problem (12) into an unconstrained quadratic optimization problem with a smaller number of variables to be optimized. To take full advantage of structure $\mathbf{B}$, we seek optimization methods (Byrd et al. 1995; Schmidt 2010) to solve problem (12) by exploring the first order information since we can compute the objective $h^{(t)}(\mathbf{w}) = \tau(\frac{1}{2}\sum_{i=1}^N(\mathbf{b}_i^T\mathbf{w})^2 + ||\mathbf{w}||^2 + \mathbf{w}^T\mathbf{c})$ and its gradient $\nabla_{\mathbf{w}}h^{(t)}(\mathbf{w}) = \tau(\sum_{i=1}^N \mathbf{b}_i[\mathbf{b}_i^T\mathbf{w}] + 2\mathbf{w} + \mathbf{c})$ very efficiently as they only involve the inner product of two vectors, and either $\mathbf{b}_i$ or $\mathbf{w}$ are sparse in general.

According to Propositions 4 and 5, the reformulated problem follows the standard form of ADMM since the positive

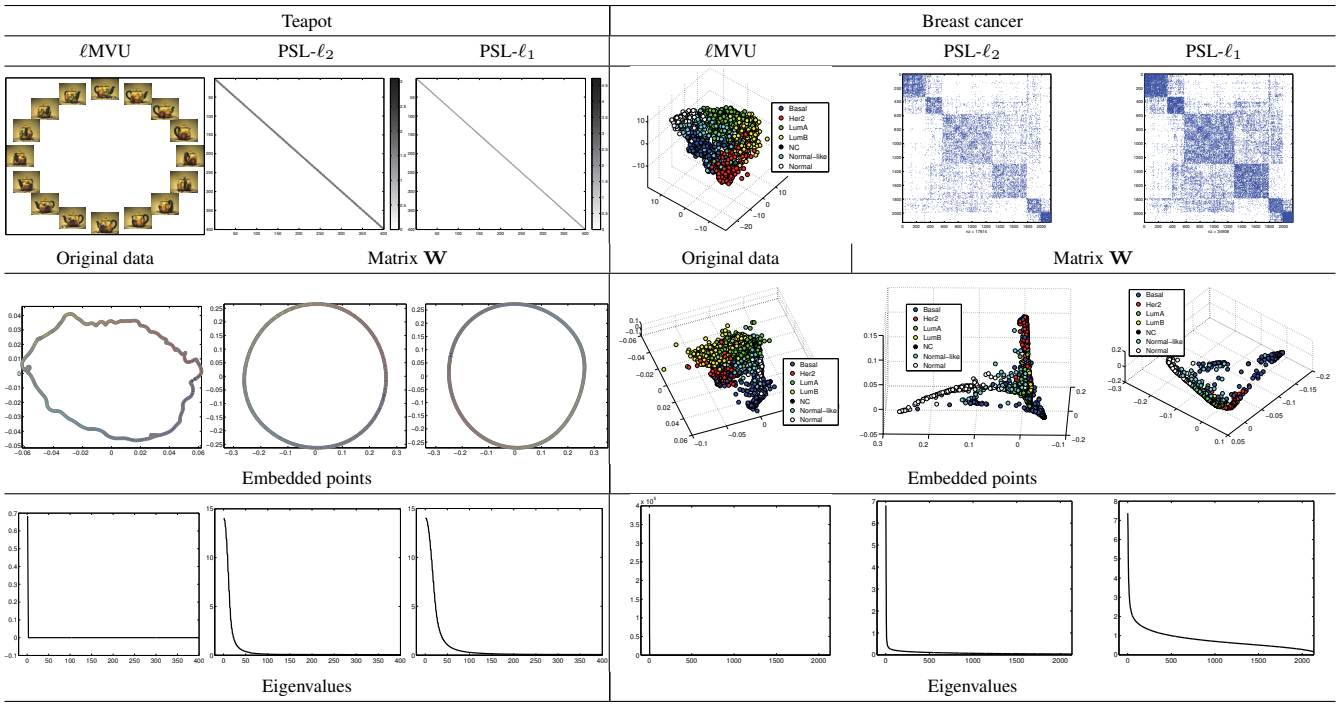


Figure 2: Results of ℓMVU, PSL-ℓ_2, and PSL-ℓ_1 on Teapot and Breast cancer, including the embedded points in 2-D or 3-D space, the eigenvalues of the learned kernel matrix, and the sparse similarity matrix learned by PSL-ℓ_2 and PSL-ℓ_1. Eigenvalues are plotted in descending order and the similarity matrix is rearranged in terms of encoded colors in ascending order.

definite constraint on $\mathbf{G}$ automatically holds due to the log determinant function and the vectorized $\mathbf{w}$ is unconstrained. As a result, the stopping criterion discussed in Section 3.3.1 of work (Boyd et al. 2011) is used and the penalty parameter τ is varied adaptively according to the primal and dual residues as given in Section 3.4.1 of the same work. The convergence property of ADMM proved in (Boyd et al. 2011) is thus readily adapted to the proposed algorithm. Our method has computational complexity $O(N^3)$, which is the same as that of most of spectral based methods, but is much faster than semi-definite programming used in MVU.

Experiments

Experiments were conducted on various datasets to evaluate the proposed methods. The first experiment is to verify the embedded points by visualizing them in 2-D or 3-D space, while the second experiment is to evaluate clustering performance on the embedded points.

Nonlinear Dimensionality Reduction

We evaluated the proposed methods on datasets in which the high-dimensional points were sampled from a low-dimensional skeleton structure. Two datasets were used in this experiment. One is teapot data (Weinberger, Sha, and Saul 2004) which consists of 400 color images of a teapot viewed from different angles in the plane (rotated $360°$), and each image consists of 76×101 RGB pixels in a $23,028$ dimensional vector space (Weinberger and Saul 2006). The other is breast cancer data, which contains the expression levels of over $25,000$ gene transcripts obtained from 144 normal breast tissue samples and $1,989$ tumor tissue samples (Mao et al. 2015). All data points are encoded with colors for checking the distribution of embedded points.

Landmark MVU (ℓMVU) (Weinberger, Packer, and Saul 2005) was evaluated for computational consideration, and the number of landmarks was set to 40. As KPCA and MVU have been compared thoroughly in (Weinberger, Sha, and Saul 2004), KPCA is not reported. As the ℓ_2 regularized model is a generalized version of PSL in the case of C approaching infinity, we set $C = 10^3$. A very broad, spherical Gaussian density with $\gamma = 10^{-5}$ is used as a surrogate for the uninformative prior. For PSL-ℓ_1, β controls the sparsity of the similarity matrix, so our model does not have a preset number of nearest neighbors. We set $\beta = 10^3$ to promote the sparsity of $\mathbf{W}$.

Figure 2 shows the results of ℓMVU and our proposed methods on Teapot and Breast cancer. First, we observe that all three methods can correctly recover the circle structure of the teapot images, but our proposed methods obtain a skeleton of embedded points that is much smoother than that obtained by ℓMVU. The smoothness of the skeleton structure becomes much clearer on Breast cancer data since the data tends to be very noisy. The skeleton structures of our proposed methods suggest a linear bifurcating progression path, starting from the normal tissue samples, and diverging to either luminal A or basal subtypes. The linear trajectory through luminal A continues to luminal B and to the HER2+ subtype. This is consistent with the branching ar-

Table 1: Clustering results of 11 methods on seven datasets in terms of accuracy and NMI. The best results are in bold.

Dataset	COIL20	Isolet	Pendigits	Satimage	USPS	YALE-B	Letter
(N, c)	(1440, 20)	(3119, 2)	(3498, 10)	(4435, 6)	(2007, 10)	(2414, 38)	(5000, 26)
(D, d)	(1024, 84)	(617, 165)	(16, 9)	(36, 6)	(256, 32)	(1024, 116)	(16, 12)
Accuracy							
Kmeans	0.6674	0.5633	0.6544	0.6685	0.6153	0.1081	0.2632
PCA	0.6674	0.5633	0.6527	0.6681	0.6208	0.1110	0.2634
KPCA	0.6694	0.5643	0.7384	0.6728	0.6313	0.1135	0.2752
LLE	0.3493	0.5021	0.1215	0.6510	0.1624	0.0667	0.3084
LE	0.2035	0.5005	0.7607	0.6870	0.3767	0.0597	0.0634
ℓMVU	0.5042	0.5989	0.5692	0.6886	0.3896	0.0684	0.1484
GPLVM	0.6674	0.5630	0.6527	0.6681	0.4709	0.0671	0.2634
MEU	0.3590	0.5476	0.7138	0.7398	0.6129	0.3094	0.2518
SMCE	0.3813	0.5162	0.8333	0.7143	0.6009	0.2933	0.2824
PSL-ℓ_2	0.7181	**0.6794**	0.8208	0.7132	0.6487	**0.4138**	0.2816
PSL-ℓ_1	**0.7319**	0.5973	**0.8468**	**0.7454**	**0.7309**	0.3521	**0.3320**
NMI							
Kmeans	0.7845	0.0117	0.6669	0.6097	0.5657	0.1694	0.3621
PCA	0.7845	0.0117	0.6627	0.6090	0.5664	0.1781	0.3591
KPCA	0.7893	0.0121	0.6846	0.6110	0.5769	0.1742	0.3664
LLE	0.3848	0.0004	0.0059	0.5080	0.0106	0.0810	0.3992
LE	0.2357	0.0003	0.7554	0.6038	0.2943	0.0515	0.0213
ℓMVU	0.6494	0.0292	0.6422	0.5818	0.3858	0.0870	0.1379
GPLVM	0.7845	0.0116	0.6627	0.6090	0.3824	0.0716	0.3591
MEU	0.5067	0.0151	0.7635	0.6826	0.5743	0.4174	0.3341
SMCE	0.4920	0.0165	0.8071	0.6515	0.6176	0.4042	0.3890
PSL-ℓ_2	0.8412	**0.0952**	0.7999	0.6526	0.7130	**0.6094**	0.3940
PSL-ℓ_1	**0.8517**	0.0392	**0.8230**	**0.6953**	**0.7641**	0.4796	**0.4482**

chitecture of cancer progression (Greaves and Maley 2012). However, ℓMVU does not obtain a clear structure. Moreover, the skeleton structure learned by PSL-ℓ_1 is slightly better than the structure by PSL-ℓ_2. This is partially because the varied neighborhood on Breast cancer data is better than fixed neighborhood. We also demonstrate the distribution of eigenvalues. ℓMVU obtained a kernel with few non-zero eigenvalues, while our methods learned a kernel with eigenvalues that follow a power law distribution. Our methods obtain a sparse similarity matrix, while ℓMVU does not. These observations are in line with our theoretical analysis, and imply that our proposed methods can learn smooth skeleton structures of embedded points from noisy data.

Clustering with Dimensionality Reduction

We evaluated clustering performance on embedded points obtained by both proposed methods and existing dimensionality reduction methods. The datasets used in the experiments are listed in Table 1, where the reduced dimensions are set so that retain 95% energy of each dataset after applying PCA. We compared the proposed methods PSL-ℓ_2 and PSL-ℓ_1 with PCA, KPCA (Schölkopf, Smola, and Muller 1999), LLE (Belkin and Niyogi 2001), LE (Saul and Roweis 2003), ℓMVU (Weinberger, Packer, and Saul 2005), GPLVM (Lawrence 2005), MEU (Lawrence 2012) and SMCE (Elhamifar and Vidal 2011). For methods with a k-nearest neighborhood graph as the input, we either obtained k using the normal neighborhood selection strategy (Van der Maaten, Postma, and van den Herik 2009) or tuned k in a range $\{5, 10, 15, 20, 30, 50, 100\}$ for the best performance. For methods that use Gaussian kernel, we set the bandwidth parameter to the estimated standard deviation within neighborhoods (Weinberger, Sha, and Saul 2004). The parameter β in PSL-ℓ_1 and regularization parameter in SMCE were tuned in a large range $\{0.1, 1, 10, 10^2, 10^3\}$. Other parameters use default setting suggested in drtoolbox (*lvdmaaten.github.io/drtoolbox/*), MEU and SMCE. Kmeans on the original data was used as the baseline. To alleviate the non-convex issue of Kmeans, we ran Kmeans with 20 random initialization, and the clustering with the best objective value was evaluated in terms of accuracy and normalized mutual information (NMI) (Nie et al. 2009). All methods used the same reduced dimension and conducted Kmeans on embedded points with the number of true clusters.

Table 1 shows the clustering results of 11 methods on seven datasets in terms of accuracy and NMI. We make several observations from the results. First, Kmeans on the embedded data points obtained by half the existing dimensionality reduction methods may not be better than PCA, which was also observed in (Van der Maaten, Postma, and van den Herik 2009). Second, the proposed methods can consistently obtain robust results on most of the datasets in terms of both accuracy and NMI. Third, methods with learning a sparse graph structure can obtain better results than methods with a fixed neighborhood graph as input. Although PSL-ℓ_2 does not learn a graph, its robust distance modeling alleviate this issue, so the clustering performance is comparable with PSL-ℓ_1 and even better than SMCE. Lastly, the proposed methods perform much better than probabilistic models such as GPLVM and MEU due to the strategy of directly modeling the posterior distribution of embedded points. This is also verified by the superior performance of PSL to MVU where a smooth skeleton might be better for clustering problems than only unfolding the data by preserving distances.

Conclusion

In this paper, we proposed a novel probabilistic model for nonlinear dimensionality reduction. Unlike MVU, we model the posterior distribution of embedded points by preserving the expected pairwise distances encoded in a given neighborhood graph. The duality of this model can be interpreted as maximizing the log-likelihood of a power law distribution over the eigenvalues of a sparse dual matrix, which leads to a smooth skeleton. Two variants of the model are also discussed for noise samples and graph structure learning. The formulated problems are convex and can be efficiently solved by ADMM. Extensive experiments demonstrate that the proposed model achieves a smooth skeleton of embedded points and outperforms various existing methods on seven datasets in terms of clustering performance.

Acknowledgments

Ivor W. Tsang is supported by the ARC Future Fellowship FT130100746 and ARC grant LP150100671.

References

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, 585–591.

Boyd, S.; Parikh, N. amd Eric, C.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *FTML* 3(1):1–122.

Burges, C. J. C. 2009. Dimension reduction: a guided tour. *FTML* 2(4):275–365.

Byrd, R. H.; Lu, P.; Nocedal, J.; and Zhu, C. 1995. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* 16(5):1190–1208.

Curtis, C.; Shah, S. P.; Chin, S.; et al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486(7403):346–352.

Dudík, M.; Phillips, S. J.; and Schapire, R. E. 2007. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *JMLR* 8(6).

Elhamifar, E., and Vidal, R. 2011. Sparse manifold clustering and embedding. In *NIPS*, 55–63.

Greaves, M., and Maley, C. C. 2012. Clonal evolution in cancer. *Nature* 481(7381):306–313.

Gupta, A. K., and Nagar, D. K. 1999. *Matrix variate distributions*, volume 104. CRC Press.

Lake, B., and Tenenbaum, J. 2010. Discovering structure by learning sparse graph. In *Proceedings of the 33rd Annual Cognitive Science Conference*.

Lawrence, N. D. 2005. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *JMLR* 6:1783–1816.

Lawrence, N. D. 2012. A unifying probabilistic perspective for spectral dimensionality reduction: Insights and new models. *JMLR* 13(1):1609–1638.

Liu, Q., and Ihler, A. T. 2011. Learning scale free networks by reweighted l1 regularization. In *AISTATS*, 40–48.

Mao, Q., and Tsang, I. W. 2010. Parameter-free spectral kernel learning. *UAI*.

Mao, Q.; Yang, L.; Wang, L.; Goodison, S.; and Sun, Y. 2015. SimplePPT: A simple principal tree algorithm. In *SDM*.

Mao, Q.; Wang, L.; and Tsang, I. W. 2016. A unified probabilistic framework for robust manifold learning and embedding. *Machine Learning*.

Nie, F.; Xu, D.; Tsang, I. W.; and Zhang, C. 2009. Spectral embedded clustering. In *IJCAI*, 1181–1186.

Saul, L. K., and Roweis, S. T. 2003. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *JMLR* 4:119–155.

Schmidt, M. 2010. *Graphical model structure learning with l1-regularization*. Ph.D. Dissertation, University of British Columbia, Vancouver.

Scholkopf, B., and Smola, A. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schölkopf, B.; Smola, A.; and Muller, K. 1999. Kernel principal component analysis. *Advances in Kernel Methods - Support Vector Learning* 327–352.

Smola, A. J., and Kondor, R. 2003. Kernels and regularization on graphs. In *COLT*. 144–158.

Van der Maaten, L.; Postma, E. O.; and van den Herik, H. J. 2009. Dimensionality reduction: A comparative review. *Tilburg University Technical Report, TiCC-TR 2009-005*.

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

Weinberger, K.; Packer, B.; and Saul, L. 2005. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *AISTATS*, 381–388.

Weinberger, K.; Sha, F.; and Saul, L. 2004. Learning a kernel matrix for nonlinear dimensionality reduction. In *ICML*, 106.

Xiao, L.; Sun, J.; and Boyd, S. 2006. A duality view of spectral methods for dimensionality reduction. In *ICML*, 1041–1048. ACM.