A unified probabilistic framework for robust manifold learning and embedding

Qi Mao¹ · Li Wang² · Ivor W. Tsang³

Received: 16 April 2016 / Accepted: 18 October 2016 / Published online: 23 December 2016
© The Author(s) 2016

Abstract This paper focuses on learning a smooth skeleton structure from noisy data—an emerging topic in the fields of computer vision and computational biology. Many dimensionality reduction methods have been proposed, but none are specially designed for this purpose. To achieve this goal, we propose a unified probabilistic framework that directly models the posterior distribution of data points in an embedding space so as to suppress data noise and reveal the smooth skeleton structure. Within the proposed framework, a sparse positive similarity matrix is obtained by solving a box-constrained convex optimization problem, in which the sparsity of the matrix represents the learned neighborhood graph and the positive weights stand for the new similarity. Embedded data points are then obtained by applying the maximum a posteriori estimation to the posterior distribution expressed by the learned similarity matrix. The embedding process naturally provides a probabilistic interpretation of Laplacian eigenmap and maximum variance unfolding. Extensive experiments on various datasets demonstrate that our proposed method obtains the embedded points that accurately uncover inherent smooth skeleton structures in terms of data visualization, and the method yields superior clustering performance compared to various baselines.

Editors: Bob Durrant, Kee-Eung Kim, Geoff Holmes, Stephen Marsland, Zhi-Hua Zhou, and Masashi Sugiyama.

✉ Li Wang
liwang8@uic.edu

Qi Mao
maoq1984@gmail.com

Ivor W. Tsang
Ivor.Tsang@uts.edu.au

¹ HERE North America, Chicago, IL, USA

² Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, IL, USA

³ Centre for Artificial Intelligence, University of Technology Sydney, Sydney, NSW, Australia

Keywords Dimensionality reduction · Probabilistic model · Manifold embedding · Structure learning

1 Introduction

Real world data, such as images and molecular profiles, is usually high-dimensional. In order to adequately handle such high-dimensional data, dimensionality reduction is usually used to transform the data from a high dimensional space into a reduced representation, which ideally corresponds to the intrinsic dimensionality of the data (Fukunaga 2013). It also facilitates data visualization and the classification of high-dimensional data by mitigating the curse of dimensionality and other undesirable properties (Jimenez and Landgrebe 1998).

During the last two decades, a plethora of dimensionality reduction methods have been proposed (van der Maaten et al. 2009; Burges 2009). Most of them aim to preserve certain information within the data. Principal component analysis (PCA; Jolliffe 1986), a classic method for this purpose, learns a subspace linearly spanned over some orthonormal bases by minimizing the reconstruction error (Burges 2009). However, the complex structure of data may be misrepresented by the linear manifold constructed using PCA. To overcome this issue, Kernel PCA (KPCA; Schölkopf et al. 1999) first maps the original space to a reproducing kernel Hilbert space (RKHS) using a kernel function, and then performs PCA in the RKHS. Hence, KPCA is a nonlinear generalization of traditional PCA, but its performance critically depends on the choice of the kernel function.

Several works have been proposed for model selection of the kernel function. Gaussian process latent variable model (GPLVM; Lawrence 2005) achieves a nonlinear generalization of probabilistic PCA (PPCA; Tipping and Bishop 1999) and learns a kernel function defined on a set of variables in a low-dimensional latent space by maximizing the log-likelihood of observed data in terms of covariance matrix in Gaussian process (Rasmussen 2006). However, the objective function of GPLVM is highly nonconvex, so it is easily trapped at the local optima. Maximum variance unfolding (MVU; Weinberger et al. 2004) is proposed to bypass the challenging issue of choosing a kernel function by learning a non-parametric kernel matrix directly, which retains pairwise distances encoded in a neighborhood graph constructed from the input data. Several variants of MVU have also been proposed, such as relaxation by the inequality constraints (MVUineq; Weinberger and Saul 2006) or ℓ_2 -norm over slacks of distance differences (MVU2; Weinberger and Saul 2006), and landmark MVU (ℓ MVU; Weinberger et al. 2005).

Another family of dimensionality reduction methods is sparse spectral manifold learning, which aims to find a manifold that is close to the intrinsic structure of data. Similar to MVU, a neighborhood graph has to be provided in advance, but only the pairwise similarities over the edges of the given graph are considered to approximate the manifold. Given a neighborhood graph and a kernel function, Laplacian eigenmap (LE; Belkin and Niyogi 2001) is proposed to find a mapping where the distances between a data point and its neighbors are minimized. Authors provide the theoretical analysis of the Laplacian matrix using graph spectral theory. However, LE also faces the difficulty of selecting kernel functions. Locally linear embedding (LLE; Saul and Roweis 2003) is proposed to preserve local geometry by learning a sparse similarity matrix based on the assumption that local patches over k -nearest neighbors are nearly linear and overlap with another one to form a manifold.

Recently, datasets with smooth skeleton structures have been emerging from the fields of computer vision (Weinberger and Saul 2006; Song et al. 2007) and computational biol-

ogy (Curtis et al. 2012; Sun et al. 2014). For example, human cancer is a dynamic disease that develops over an extended time period. Once initiated from a normal cell, the advance to malignancy can, to some extent, be considered to have a complex branching structure (Greaves and Maley 2012). It becomes very important to unveil this dynamic process using massive molecular profile data, which is generally in high-dimensional space corresponding to tens of thousands of genes (Curtis et al. 2012). More importantly, the data is generally noisy because not all genes provide useful expressions related to the cancer. Thus, the dynamic progression structure is generally assumed to reside in a low-dimensional space (Sun et al. 2014). Moreover, the visualization of cancer progression structure is also valuable for downstream analysis, e.g., providing critical insights into the process of disease, and informing the development of diagnostics, prognostics and targeted therapeutics. Datasets with this type of structure have also been widely studied in principal curve learning, but only for curve structures. (Hastie and Stuetzle 1989; Kégl et al. 2000). However, a curve is not an appropriate representation of complex structures such as **loops, bifurcations, and multiple disconnected components**.

Although the above methods work well under certain conditions, they lack a unified probabilistic framework to robustly learn a smooth skeleton structure from noisy data. Probabilistic models such as PPCA and GPLVM can deal with noisy data, but they find it difficult to model the neighborhood manifold. On the other hand, methods based on neighborhood manifold, such as MVUs, LE and LLE, either find it hard to learn the manifold structure of a smooth skeleton, or cannot be interpreted as probabilistic model for model selection and missing data, simultaneously. Thus, a specially designed dimensionality reduction method for uncovering skeleton structures in complex forms from noisy high-dimensional datasets is desperately needed.

In this paper, our goal is to recover a skeleton structure in an embedding space from noisy observed data. To achieve this goal, we propose a novel probabilistic model that directly learns a posterior distribution of the embedded points in the embedding space by directly incorporating noise of data points through a prior distribution and adopting pairwise distance constraints for modeling a smooth skeleton structure, simultaneously. Once the posterior distribution has been obtained, the embedded points are equivalent to the maximum a posteriori (MAP) estimation. The main contributions of this paper are summarized as follows:

- We propose a novel probabilistic dimensionality reduction framework that can simultaneously suppress data noise and uncover a smooth skeleton structure, in which a prior distribution is used to model the noise and the expected inequality constraints over pairwise distances of embedding points impose smooth skeleton structures.
- Under the proposed framework, the posterior distribution has an analytic expression with a sparse positive similarity matrix for representing a weighted neighborhood graph, so that both the skeleton structure and the similarity function are automatically adapted from the data.
- The embedding points are represented by the MAP estimation of the learned posterior distribution. Solving the MAP estimation is consistent with the optimization problem of LE, and gives a natural explanation for the use of KPCA for embedding.
- The resulting optimization problem for learning a sparse positive similarity matrix is convex with a box constraint, and the objective function is continuous and differentiable. Thus, there exists a global optimal solution, and the problem can be efficiently solved by fast and large-scale off-the-shelf optimization tools.
- Extensive experiments have been conducted from the perspectives of data visualization and clustering. Results on five synthetic datasets and nine real world datasets demonstrate

that the proposed method can correctly recover smooth skeleton structures from noisy data for data visualization and also achieve better clustering performance than various existing methods.

2 Related work

Our proposed method is a probabilistic model that can automatically learn a smooth skeleton structure from a given observed dataset with noise. Although many dimensionality reduction methods have been proposed in the literature, most of them are not suitable for our purposes. In this section, we will mainly discuss in details two methods, which are closely related to our proposed method—MVU (Weinberger et al. 2004) and maximum entropy unfolding (MEU; Lawrence 2012).

MVU (Weinberger et al. 2004) is a deterministic method that learns a kernel matrix $\mathbf{K}$ from a given dataset $\{\mathbf{y}_i\}_{i=1}^N$ as a set of observed data points in $\mathbb{R}^D$ and a neighborhood graph with $\mathcal{N}_i$ as the neighbors of $\mathbf{y}_i$. The formulation of MVU is defined as the following optimization problem given by

$$\begin{aligned} \max_{\mathbf{K} \succeq 0} \quad & \text{Tr}(\mathbf{K}) \\ \text{s.t.} \quad & K_{i,i} + K_{j,j} - 2K_{i,j} = \phi_{i,j}, \forall i, j \in \mathcal{N}_i \\ & \sum_{i,j} K_{i,j} = 0, \end{aligned}$$

where $\text{Tr}(\mathbf{K})$ represents the trace of $\mathbf{K}$, $K_{i,j}$ is the (i, j) th element of $\mathbf{K}$, and $\phi_{i,j}$ is the square distance of $\mathbf{y}_i$ and $\mathbf{y}_j$. The embedded data is then obtained by applying KPCA on $\mathbf{K}$. MVU is well suited to unfolding a manifold if the true manifold is given. MVU is formulated as a semidefinite programming (SDP) problem (Vandenberghe and Boyd 1996), but is impractical when scaled up to thousands of data points. Landmark MVU (ℓ MVU; Weinberger et al. 2005) alleviates the high computational complexity of MVU by introducing landmarks and linear transformation for kernel matrix factorization. However, this causes the additional difficulties of choosing landmarks, the nearest neighbors for linear transformation, and determining a neighborhood graph for retaining pairwise distances. MVU also faces challenges when handling data with noise. Some variants of MVUs including MVUineq and MVU2 (Weinberger and Saul 2006) can alleviate the noise in pairwise distances by relaxing the equality constraints, but they cannot model the noise of the data points. If the data is noisy, a precomputed neighborhood graph to approximate the manifold of data is not reliable any more. Moreover, the k -nearest neighbor graph used in MVUs is no longer appropriate to data where local regions have different densities (Elhamifar and Vidal 2011).

MEU (Lawrence 2012) was proposed to directly model the density of observed data $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]$ by minimizing the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951) between a base density $m(\mathbf{Y})$ and the density $p(\mathbf{Y})$ as

$$\min_{p(\mathbf{Y})} \int p(\mathbf{Y}) \log \frac{p(\mathbf{Y})}{m(\mathbf{Y})},$$

under the constraints on the expected squared inter-point distances $\phi_{i,j}$ of any two samples, $\mathbf{y}_i$ and $\mathbf{y}_j$. Let $m(\mathbf{Y})$ be a very broad, spherical Gaussian density with covariance $\lambda^{-1}\mathbf{I}$, where $\lambda > 0$ is a scaled parameter. The density function is then constructed as

$$p(\mathbf{Y}) \propto \exp\left(-\frac{1}{2}\text{Tr}(\lambda \mathbf{Y} \mathbf{Y}^T)\right) \exp\left(-\frac{1}{2} \sum_i \sum_{j \in \mathcal{N}_i} w_{i,j} \phi_{i,j}\right),$$

even though the explicit form of these constraints is not given. The kernel matrix $\mathbf{K} = (\mathbf{L} + \lambda \mathbf{I})^{-1}$ is obtained after the optimal $\mathbf{L}$ is achieved by maximizing the logarithmic function of $p(\mathbf{Y})$, and the embedded points are obtained for dimensionality reduction by applying KPCA, where $w_{i,j}$ is the (i, j) th element of a similarity matrix $\mathbf{W}$ to form the graph Laplacian $\mathbf{L} = \text{diag}(\mathbf{W} \mathbf{1}) - \mathbf{W}$. MEU models the density $p(\mathbf{Y})$ of observed data, which has to assume that the data features are i.i.d. given the model parameters. This assumption hardly hold if feature correlation exists. In order to obtain the kernel, MEU takes a pseudolikelihood approximation to learn $\mathbf{L}$ so as to avoid the positive semidefinite constraint on $\mathbf{K}$. However, pseudolikelihood is motivated by computational reasons, and it sacrifices the accuracy of the estimated kernel (Besag 1975).

3 Maximum posterior manifold embedding

3.1 Motivation

We are interested in learning a smooth skeleton from noisy data. Informally, a smooth skeleton is a special skeleton structure of manifolds that is embedded in a low-dimensional space of the observed data. Noisy data can be the data with noises added into the original feature space, or inserted as irrelevant dimensions. These special structures have been widely studied in principal curve learning for curve structures (Hastie and Stuetzle 1989; Kégl et al. 2000). However, curves are not appropriate representations of complex structures such as loops, bifurcations, and multiple disconnected components. In this paper, we further generalize skeleton structures to these complex forms. In nonlinear dimensionality reduction methods, the structure of embedded data might be different from the structure of original data (in the case of 2-D or 3-D datasets which can be visualized naturally), but the smooth skeleton structure should be retained, and the noise is removed or alleviated.

Figure 5 shows five synthetic datasets with clear, smooth skeleton structures. These are treated as the ground truth for evaluating the visualization performance of different methods. For example, the structure of “DistortedSShape” data is generated by adding noise into data that originally forms an S-shape structure. A method that finds a smooth bell-shape and retains the correct curve structure is preferred, even though the curves are demonstrated in different shapes. Also on the “2moons” dataset, the structure of two non-intersecting smooth curves could be identified as two disconnected skeleton structures for our interests. The detailed descriptions of these synthetic datasets and their corresponding skeleton structures are discussed in Sect. 4.3.1.

Figure 1 shows two real world datasets. A collection of teapot images is viewed from different angles (Weinberger et al. 2005). Each image contains 76×101 RGB pixels, so the pixel space has a dimension of 23,028, but the intrinsic structure has only one degree of freedom: the angle of rotation, as shown in Fig. 1a. Human cancer is a dynamic disease that develops over an extended time period. Once initiated from a normal cell, the advance to malignancy can to some extent be considered a Darwinian process—a multi-step evolutionary process—that responds to selective pressure (Greaves and Maley 2012). The disease progresses through a series of clonal expansions that result in tumor persistence and growth, and ultimately the ability to invade surrounding tissue and metastasize to distant organs. As shown in Fig. 1b, the evolution trajectories inherent in cancer progression are complex

Fig. 1 The real datasets with smooth skeleton structures. **a** A collect of teapot images form a smooth skeleton structure of a circle, **b** molecular profiles of cancer tissues have a complex and branching progression. **a** Teapot images, **b** cancer progress path

and branching (Greaves and Maley 2012). It has become critically important to uncover the progression path from massive molecular profile data (Sun et al. 2014).

In order to learn a set of embedded points that form a smooth skeleton from observed data, the manifold assumption over the embedded points is quite appropriate. In general, a sparse neighborhood graph over observed data points, e.g., a k -nearest neighbor graph, is manually crafted, and is used to approximate the manifold of the data. However, it is challenging to build a good neighborhood graph with a prefixed k if the local density of each observed data point is very different. Moreover, a smooth manifold cannot tolerate noise in observed data if we preserve deterministic distances as strictly as we do in MVU since the distances themselves are not reliable. To overcome the above issues, we should resort to a probabilistic model that represents the distance in a flexible fashion so as to learn a smooth skeleton structure to best approximate a true smooth skeleton.

3.2 Probabilistic modeling assumptions

We define $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^N$ as a set of observed data points in $\mathbb{R}^D$. The distance of any two points $\mathbf{y}_i$ and $\mathbf{y}_j$ can be computed in either the Euclidean space as $\phi_{i,j} = \|\mathbf{y}_i - \mathbf{y}_j\|_2^2$ or the RKHS $\mathcal{H}$ as $\phi_{i,j} = \|\varphi(\mathbf{y}_i) - \varphi(\mathbf{y}_j)\|_{\mathcal{H}}^2 = K_{i,i} + K_{j,j} - 2K_{i,j}$, where $K(\mathbf{y}_i, \mathbf{y}_j) = \langle \varphi(\mathbf{y}_i), \varphi(\mathbf{y}_j) \rangle_{\mathcal{H}}$ is a kernel function. Let the corresponding embedded data $\mathbf{x}$ of $\mathbf{y}$ be a set of points $\{\mathbf{x}_i\}_{i=1}^N$ in an embedding space $\mathbb{R}^d$ with $\mathbf{x}_i = [x_i^1, \dots, x_i^d]^T$.

In line with existing probabilistic dimensionality reduction models such as PPCA (Tipping and Bishop 1999), GPLVM (Lawrence 2005), and MEU (Lawrence 2012), we assume that $\mathbf{f}_k, k = 1, \dots, d$, are independent and identically distributed (i.i.d.) given by

$$p(\mathbb{F}|\mathbf{y}) = \prod_{k=1}^d p(\mathbf{f}_k|\mathbf{y}),$$

where $\mathbb{F} = \{\mathbf{f}_k\}_{k=1}^d$ and $\mathbf{f}_k = [x_1^k, \dots, x_N^k]^T \in \mathbb{R}^N$. This imposes non-correlation among features in the embedding space, and is widely employed by existing methods such as PPCA and MEU. In addition, the prior distribution of embedded points with the mean as zero and the covariance as either the identity matrix or scaled by a positive parameter has also been used in PPCA and MEU. Here, we take the scaled identity matrix as the covariance of the

prior distribution similar to MEU, that is, $p_0(\mathbf{f}_k) \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$. The reason for taking this prior is that λ provides the enough flexibility in our models to fit the noise of a given dataset.

Unlike the aforementioned probabilistic models, we propose to directly model the posterior distribution $p(\mathbb{F}|\mathbb{Y})$ and obtain the optimal embedded points $\mathbf{x}$ by applying the maximum a posteriori estimation on $p(\mathbb{F}|\mathbb{Y})$, which can seamlessly incorporate distance information and prior distribution. To achieve a flexible distance representation for smooth skeleton structures and obtain a sparse neighborhood graph for alleviating the effect of noises in the observed data, we take two critically important innovations. First, the expected Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ in terms of the posterior distribution $p(\mathbb{F}|\mathbb{Y})$ of embedding points is used as the representation of the pairwise distance between two embedded points, i.e.,

$$\int \|\mathbf{x}_i - \mathbf{x}_j\|^2 p(\mathbb{F}|\mathbb{Y}) d\mathbb{F} = \sum_{k=1}^d \int (x_i^k - x_j^k)^2 p(\mathbf{f}_k|\mathbb{Y}) d\mathbf{f}_k.$$

Second, we introduce an error tolerance variable $\xi_{i,j}$ so that $\phi_{i,j}$ and the expected distance of corresponding two embedded points $\mathbf{x}_i$ and $\mathbf{x}_j$ do not have to be strictly preserved. This is given by

$$\int \|\mathbf{x}_i - \mathbf{x}_j\|^2 p(\mathbb{F}|\mathbb{Y}) d\mathbb{F} \leq \phi_{i,j} + \xi_{i,j}, \xi_{i,j} \geq 0, \forall i, j. \quad (1)$$

It is worth noting that the neighborhood graph is unknown and our goal is to learn it from data. This is one of the key differences between our method and both MVU and MEU, which assume that the neighborhood graph is given as a prior.

The expected distance of these points with respect to the posterior distribution is much more flexible than the deterministic distances used in MVU since both $\mathbf{x}$ and $p(\mathbb{F}|\mathbb{Y})$ are variables to optimize. Moreover, this flexibility is further strengthened by the inequality constraints when these distances might not be strictly preserved. These constraints not only tolerate noises, but also result in a sparse neighborhood graph (see Sect. 3.3). More importantly, the smooth skeleton we seek can be found from the high flexibility of embedded points since two embedded points can move closely if the distance between two observed data is small according to (1). By contrast, deterministic constraints used in MVU are strictly preserved, so MVU cannot achieve the same effect. As a result, the not-restrictively-preserved constraints allow the embedded points to move flexibly to form a smooth manifold and provide the feasibility of the dimensionality reduction methods for automatically adapting the neighborhood manifold from data. We provide a detailed discussion from a duality perspective in Sect. 3.3.

3.3 Problem formulation via probabilistic modeling

Based on the assumptions discussed in Sect. 3.2, we propose to directly estimate a posterior distribution of embedded data points $\mathbf{x}$ in a low-dimensional space $\mathbb{R}^d$ by minimizing the KL divergence between the posterior distribution and a prior distribution with a set of constraints in terms of expected pairwise distances. This modeling technique has been widely used to learn a posterior distribution from data for problems such as classification (Jebara 2001), structured output prediction (Zhu and Xing 2009) and multiple kernel learning (Mao et al. 2015).

The joint posterior distribution over $\mathbb{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_d\}$ can be achieved by assuming that distributions $\{p(\mathbf{f}_k|\mathbb{Y})\}_{k=1}^d$ are independent, i.e., $p(\mathbb{F}|\mathbb{Y}) = \prod_{k=1}^d p(\mathbf{f}_k|\mathbb{Y})$ and solving the KL divergence with respect to the joint distribution given by

$$\begin{aligned}
& \min_{p(\mathbf{f}_k|\mathbf{y}) \in \mathcal{P}_d, \{\xi_{i,j}\}} \sum_{k=1}^d \int p(\mathbf{f}_k|\mathbf{y}) \log \frac{p(\mathbf{f}_k|\mathbf{y})}{p_0(\mathbf{f}_k)} d\mathbf{f}_k + C \sum_{i,j} \xi_{i,j} \\
& \text{s.t.} \sum_{k=1}^d \int (x_i^k - x_j^k)^2 p(\mathbf{f}_k|\mathbf{y}) d\mathbf{f}_k \leq \phi_{i,j} + \xi_{i,j}, \xi_{i,j} \geq 0, \forall i, j,
\end{aligned} \quad (2)$$

where $\mathcal{P}_d = \times_{k=1}^d \mathcal{P}_k$ is a Cartesian product of d i.i.d. probability spaces and $\mathcal{P}_k = \{p(\mathbf{f}_k|\cdot) | \int p(\mathbf{f}_k|\mathbf{y}) d\mathbf{f}_k = 1, p(\mathbf{f}_k|\mathbf{y}) \geq 0\}$, $\forall k$.

By introducing dual variables $w_{i,j} \geq 0$, $\beta_{i,j} \geq 0$ and τ_k the Lagrangian function of problem (2) can be formulated as

$$\begin{aligned}
L(\{p(\mathbf{f}_k|\mathbf{y})\}, \{w_{i,j}\}, \{\beta_{i,j}\}, \tau_k) &= \sum_{k=1}^d \int p(\mathbf{f}_k|\mathbf{y}) \log \frac{p(\mathbf{f}_k|\mathbf{y})}{p_0(\mathbf{f}_k)} d\mathbf{f}_k + C \sum_{i,j} \xi_{i,j} \\
&+ \sum_{i,j} \frac{w_{i,j}}{4} \left[\sum_{k=1}^d \int (x_i^k - x_j^k)^2 p(\mathbf{f}_k|\mathbf{y}) d\mathbf{f}_k - \phi_{i,j} - \xi_{i,j} \right] \\
&- \sum_{i,j} \beta_{i,j} \xi_{i,j} + \sum_{k=1}^d \tau_k \left(\int p(\mathbf{f}_k|\cdot) d\mathbf{f}_k - 1 \right)
\end{aligned}$$

where the dual variable $\frac{w_{i,j}}{4}$ is introduced to scale up dual variable $w_{i,j}$ for ease of representation, while still following the Lagrangian duality theorem (Boyd and Vandenberghe 2004). Moreover, the property of KL divergence automatically imposes the optimal $p(\mathbf{f}_k|\mathbf{y})$ as positive, so the positive constraints over $p(\mathbf{f}_k|\mathbf{y})$ are negligible in the Lagrangian function. According to the first order optimality condition (Boyd and Vandenberghe 2004), we have the following KKT conditions

$$1 + \log p(\mathbf{f}_k|\mathbf{y}) - \log p_0(\mathbf{f}_k) + \sum_{i,j} \frac{w_{i,j}}{4} (x_i^k - x_j^k)^2 + \tau_k = 0, k = 1, \dots, d \quad (3)$$

$$\int p(\mathbf{f}_k|\mathbf{y}) d\mathbf{f}_k = 1, k = 1, \dots, d \quad (4)$$

$$C - \frac{w_{i,j}}{4} - \beta_{i,j} = 0, \forall i, j \quad (5)$$

$$w_{i,j} \left[\sum_{k=1}^d \int (x_i^k - x_j^k)^2 p(\mathbf{f}_k|\mathbf{y}) d\mathbf{f}_k - \phi_{i,j} - \xi_{i,j} \right] = 0, \forall i, j, \quad (6)$$

$$\beta_{i,j} \xi_{i,j} = 0, \forall i, j. \quad (7)$$

Suppose that $p_0(\mathbf{f}_k) = \mathcal{N}(0, \lambda^{-1} \mathbf{I})$ is multivariate normal distribution with mean zero and covariance matrix $\lambda^{-1} \mathbf{I}$ where $\mathbf{I}$ is the identity matrix and $\lambda > 0$ is a parameter. Equations (3) and (4) lead to an analytic form of posterior distribution, for all $k = 1, \dots, d$, given by

$$\begin{aligned}
p(\mathbf{f}_k|\mathbf{y}) &\propto \exp \left(- \sum_{i,j} \frac{w_{i,j}}{4} (x_i^k - x_j^k)^2 - \frac{\lambda}{2} \sum_{i=1}^N x_i^2 \right) \\
&\sim \mathcal{N}(\mathbf{f}_k|0, (\mathbf{L} + \lambda \mathbf{I})^{-1}),
\end{aligned} \quad (8)$$

where $w_{i,j}$ is the (i, j) th element of matrix $\mathbf{W}$, $\mathbf{L} = \mathbf{D} - \mathbf{W}$, $\mathbf{D} = \text{diag}(\mathbf{W}\mathbf{1})$, and $\mathbf{1}$ is a vector of all ones.

By substituting (8) into the Lagrangian function, we obtain the dual problem

$$\begin{aligned} \max_{\mathbf{W}} \quad & -\log Z(\mathbf{W}) - \sum_{i,j} \frac{w_{i,j}}{4} \phi_{i,j} \\ \text{s.t.} \quad & 0 \leq w_{i,j} \leq 4C, w_{i,i} = 0, w_{i,j} = w_{j,i}, \forall i, j, \end{aligned} \quad (9)$$

where the partition function is defined as

$$Z(\mathbf{W}) = \sum_{k=1}^d \int p_0(\mathbf{f}_k) \exp \left(- \sum_{i,j} \frac{w_{i,j}}{4} (x_i^k - x_j^k)^2 \right) d\mathbf{f}_k.$$

By incorporating the prior distribution, the dual problem can be simplified to

$$\begin{aligned} \max_{\mathbf{W}} \quad & \frac{d}{2} \log \det(\mathbf{L} + \lambda \mathbf{I}) - \frac{1}{4} \langle \mathbf{W}, \Phi \rangle \\ \text{s.t.} \quad & \mathbf{L} = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}, \\ & 0 \leq w_{i,j} \leq 4C, w_{i,i} = 0, w_{i,j} = w_{j,i}, \forall i, j, \end{aligned} \quad (10)$$

where $\Phi = \{\phi_{i,j}\} \in \mathbb{R}^{N \times N}$ is the distance matrix of data.

As a result, the joint posterior density function can be further derived as a simple form given by

$$p(\mathbb{F}|\mathbb{Y}) \propto \exp \left(- \frac{1}{2} \text{Tr}(\mathbf{X}(\mathbf{L} + \lambda \mathbf{I})\mathbf{X}^T) \right), \quad (11)$$

where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] = [\mathbf{f}_1, \dots, \mathbf{f}_d]^T \in \mathbb{R}^{d \times N}$. In order to learn $p(\mathbb{F}|\mathbb{Y})$ from data, we must solve problem (10), which is discussed in Sect. 3.4.

Although matrix $\mathbf{W}$ is introduced as a dual variable for solving problem (2), it has several interesting properties. (i) According to (6), the expected distance is strictly preserved if $w_{i,j} > 0$ with a tolerance $\xi_{i,j}$. (ii) According to (10), we know that $w_{i,j}$ is small if $\phi_{i,j}$ is large. Thus, the dual variable $w_{i,j}$ can be interpreted as a similarity of the embedded points $\mathbf{x}_i$ and $\mathbf{x}_j$. (iii) Optimizing $w_{i,j}$ leads to a sparse positive similarity matrix, where the optimal solution $\mathbf{W}$ is sparser if C is larger. This property can be explained by the KKT conditions at the optimum of (2). Specifically, according to condition (5), if $w_{i,j} = 0$, we have $\beta_{i,j} = C$. And then, condition (7) implies that $\xi_{i,j} = 0$. According to (6), we have $\sum_{k=1}^d \int (x_i^k - x_j^k)^2 p(\mathbf{f}_k|\mathbb{Y}) d\mathbf{f}_k < \phi_{i,j}$, which means that the distance of embedded points $\mathbf{x}_i$ and $\mathbf{x}_j$ must be smaller than $\phi_{i,j}$. That is, $w_{i,j} = 0$ leads to a shrinkage of distance. C is the regularization parameter of $\sum_{i,j} \xi_{i,j}$ in (2). By minimizing ℓ_1 norm over $\xi_{i,j}$, the $\xi_{i,j}$ will result in many zeros if C increases. As a result, the larger the C is, the sparser the $\mathbf{W}$ is more likely to be.

According to the above properties, we are now ready to explain why the proposed model can achieve a skeleton structure from noisy data through the duality perspective of problem (2). First, our model is a probabilistic model represented by a posterior density function, which incorporates the prior distribution with precision λ to capture the noise of latent embedding points. Second, the expected pairwise distances in terms of the posterior density function are more robust than deterministic pairwise distances, so they tolerate noisy data points for preserving distance information. Third, the penalty term and inequality distance constraints imposes the sparsity of $\mathbf{W}$ so that many distances are not preserved, instead they shrink. The degree of shrinkage depends on the distance of original data points. If the distance between two original points is large, the shrinkage is also large. By combining these three factors, a large pairwise distance caused by noises tends to shrink so as to approximate the inherent

distance. If the data has an inherent skeleton structure, our model can correctly uncover the skeleton structure using shrinkage effect on noisy distances.

3.4 Sparse positive similarity matrix learning

Given Φ , we obtain $\mathbf{W}$ by solving problem (10). The objective function with respect to $\mathbf{W}$ is concave since the log determinant term is concave and the second term is linear (Boyd and Vandenberghe 2004). The general approach is to reformulate (10) as a SDP given by

$$\begin{aligned} \max_{\mathbf{W}, \mathbf{L} \succeq 0} \quad & \log \det(\mathbf{L} + \lambda \mathbf{I}) - \frac{1}{2d} \langle \mathbf{W}, \Phi \rangle \\ \text{s.t.} \quad & \mathbf{L} = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}, \\ & 0 \leq w_{i,j} \leq 4C, w_{i,i} = 0, w_{i,j} = w_{j,i}, \forall i, j, \end{aligned} \quad (12)$$

which could be solved with an existing SDP solver such as SDPT3 (Tutuncu et al. 2003). However, due to the high complexity of general SDP solvers, this is impractical for thousands of data points.

According to the properties of the Laplacian matrix, we have $\mathbf{L} = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W} \succeq 0$ if $\mathbf{W} \succeq 0$. In other words, $\mathbf{L}$ is guaranteed to be positive semidefinite for any non-negative $\mathbf{W}$. Thus, problem (10) can be treated as a box-constrained convex optimization problem, and can be reformulated as

$$\begin{aligned} \max_{w_{i,j}} \quad & \log \det(\text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W} + \lambda \mathbf{I}) - \frac{1}{d} \sum_{i,j < i} w_{i,j} \phi_{i,j} \\ \text{s.t.} \quad & 0 \leq w_{i,j} \leq 4C, \forall i, j < i, \end{aligned} \quad (13)$$

where only the variables corresponding to the lower triangle of $\mathbf{W}$ are optimized due to the symmetric property of pairwise distances. This reformulation significantly reduces the number of optimized variables and removes the challenging positive semidefinite constraint.

Problem (13) is convex with box constraints, so it can be solved efficiently even for large-scale problems with existing methods such as the L-BFGS-B solver (Byrd et al. 1995). Specifically, we obtain the derivative of the log determinant term with respect to $w_{i,j}$ as

$$\frac{\partial \log \det(\text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W} + \lambda \mathbf{I})}{\partial w_{i,j}} = \text{Tr}(\mathbf{Q}^{-1} \mathbf{A}_{i,j}), \forall i, j < i,$$

where $\mathbf{Q} = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W} + \lambda \mathbf{I}$, and the matrix $\mathbf{A}_{i,j}$ can be represented by

$$[\mathbf{A}_{i,j}](m, n) = \begin{cases} 1, & m = n = i \text{ or } m = n = j, \\ -1, & m = i \wedge n = j \text{ or } m = j \wedge n = i, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the partial gradient of problem (13) with respect to $w_{i,j}$ is written as

$$\partial_{w_{i,j}} = \text{Tr}(\mathbf{Q}^{-1} \mathbf{A}_{i,j}) - \frac{1}{d} \phi_{i,j}, \forall i, j. \quad (14)$$

The convergence analysis of L-BFGS-B (Byrd et al. 1995) is automatically applied to the proposed method.

Algorithm 1 MPME (Maximum Posterior Manifold Embedding)

-
- 1: **Input:** the observed data $\mathbf{y}$, parameters λ and C
  - 2: Compute distance matrix Φ from $\mathbf{y}$
  - 3: Learn $\mathbf{W}$ by solving (13) using the L-BFGS-B method
  - 4: Obtain $\mathbf{X}$ by solving (16) or applying KPCA on $(\mathbf{L} + \lambda\mathbf{I})^{-1}$
  - 5: **Output:** the embedded data points $\mathbf{x}$
-

3.5 Embedding via MAP estimation

Given the posterior distribution (11), we obtain the embedded points $\mathbf{X}$ by maximizing the logarithm of (11) as the maximum a posteriori estimation given by

$$\max_{\mathbf{X}} \log p(\mathbb{F}|\mathbf{y}) \Leftrightarrow \min_{\mathbf{X}} \frac{1}{2} \text{Tr}(\mathbf{X}(\mathbf{L} + \lambda\mathbf{I})\mathbf{X}^T). \quad (15)$$

Since the posterior distribution is derived in terms of pairwise distances, it is insensitive to the translations of the embeddings. There is a trivial solution for the above objective, i.e., $\mathbf{X} = \mathbf{0}$. To overcome this issue, we present two strategies from our observations. The first observation is that objective function (15) is similar to that of LE (Belkin and Niyogi 2001). Let $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_d]$ and $\mathbf{D}_\lambda = \text{diag}((\mathbf{W} + \lambda\mathbf{I})\mathbf{1})$. The constraint $\mathbf{F}^T \mathbf{D}_\lambda \mathbf{F} = \mathbf{I}$ is added to prevent the above issue. With the given constraint, solving problem (15) with respect to $\mathbf{X}$ is equivalent to solving the following generalized eigenvalue problem

$$(\mathbf{L} + \lambda\mathbf{I})\mathbf{F} = \mathbf{D}_\lambda \mathbf{F} \mathbf{A}, \quad (16)$$

where $\mathbf{A}$ is a diagonal matrix with the (k, k) th element as the k th eigenvalue and $\mathbf{f}_k$ is its associated eigenvector. If $\lambda = 0$, the objective function becomes LE. Hence, LE is a special case of formulation (15) with the above constraint. Another observation is that the posterior distribution is a matrix normal distribution (Gupta and Nagar 1999) given by

$$p(\mathbf{F}|\mathbf{y}) \sim \mathcal{MN}_{N,d}(\mathbf{0}, \mathbf{U}, \mathbf{I}), \quad (17)$$

where $\mathbf{U} = (\mathbf{L} + \lambda\mathbf{I})^{-1}$ is the sample-based covariance matrix and can also be interpreted as a regularized Laplacian kernel with regularization parameter $\lambda > 0$ (Smola and Kondor 2003). As a result, we can apply KPCA to $\mathbf{U}$ to achieve the embedded data points.

Our embedding process is similar to either LE or KPCA, but the proposed embedding framework provides a novel way to automatically learn a sparse positive similarity matrix $\mathbf{W}$ from a set of pairwise distances, and the sparse positive similarity matrix is purposely designed for learning the embedded points. This also provides a probabilistic interpretation why MVU takes KPCA as the embedding method after learning a kernel matrix. The pseudo-code of our proposed maximum posterior manifold embedding (MPME) is given in Algorithm 1. Solving problem (13) takes approximately $O(N^{2.37})$ for computing logdet and an inversion of matrix $\mathbf{Q}$ at each iteration of the L-BFGS-B solver. Solving (17) takes $O(N^3)$ of KPCA. Thus, the time complexity of Algorithm 1 takes the order of $O(N^3)$, which is the same as that of most spectral based methods, but is much faster than the SDP used in MVU.

3.6 Discussion

Our proposed MPME method takes several key components from existing dimensionality reduction methods such as preserving distances from MVU and the probabilistic modeling of embedded points from PPCA and MEU. However, our method is very different. In the

following, several key differences are discussed in detail by comparing our method to MVU and MEU.

3.6.1 Comparison with MVU

First, the variables to be optimized are different. Our model learns a sparse similarity matrix, while MVU learns a dense kernel with a positive semidefinite constraint. Second, the objective functions are different. Our model maximizes the posterior distribution of latent variables in a Bayesian way, while MVU maximizes the variance of latent data points in a deterministic way. Third, the constraints are different. Our model defines expected inequality constraints (1) and (2) with error tolerance, while MVU imposes strictly equality constraints. As discussed in [Zhu and Xing \(2009\)](#), the model with the expected inequality constraints can robustly tolerate inaccurate pairwise distances. It is worth noting that ℓ MVU takes inequality constraints, but they are treated as a relaxation from the optimization perspective. And, MVUineq and MVU also take a different relaxation of the equality constraints, but the incorporation of prior distribution and expected pairwise distance in our proposed method is extremely helpful for learning a smooth skeleton structure from noisy data.

3.6.2 Comparison with MEU

One of the key differences is that our framework directly models the posterior distribution $p(\mathbf{X}|\mathbf{y})$ of latent data, while MEU models the density $p(\mathbf{Y})$ of observed data. As a result, MEU has to assume that the data features are i.i.d.. However, this assumption is hardly satisfied if feature correlation exists. By contrast, our model assumes that the reduced features in the latent space are i.i.d., which is more reasonable than that used in MEU since the latent space is generally assumed to be formed by a set of orthogonal bases, such as PCA and KPCA. This difference also leads to a clear interpretation of methods such as LE and KPCA based on the learned matrix $\mathbf{L}$ as discussed in Sect. 3.5, while MEU does not give the reason why applying KPCA on $\mathbf{K}$ can work well for dimensionality reduction. Another key difference is the expected inequality constraints used in our model, which can be explicitly represented, so the posterior distribution is well defined instead of heuristically constructed in MEU. The third difference is that MEU takes pseudolikelihood approximation to learn $\mathbf{L}$ so as to avoid the positive semidefinite constraint on $\mathbf{K}$. However, pseudolikelihood is motivated by computational reasons, and it sacrifices the accuracy of the estimated kernel ([Besag 1975](#)). By contrast, our model does not have this issue since we directly obtain the posterior distribution. Moreover, the optimization problem is a box-constrained convex optimization problem, which can be solved globally and efficiently using existing optimization tools.

4 Experiments

We conducted experiments to evaluate the proposed MPME method by comparing it with various baselines in two different experiments. One experiment justifies the embedded points by visualizing them in a low-dimensional space (e.g., 2-D or 3-D), while the other numerically evaluates the embedded points using clustering performance. Before describing the two sets of experiments, we first introduce the general experimental setting and present the sensitivity analysis of the proposed method in terms of its parameters.

Table 1 Datasets used in the experiments

Task	Dataset	N	D	d	# of classes
Visualization	Circle	100	2	2	–
	DistortedSShape	100	2	2	–
	2moons	400	2	2	–
	Helix	600	3	2	–
	Twinpeaks	600	3	2	–
	Cancer	2133	367	3	–
	Teapot	400	23028	2	–
Clustering	Letter	5000	16	12	26
	Pendigits	3498	16	9	10
	Satimage	4435	36	6	6
	USPS	2007	256	32	10
	Isolet	3119	617	165	2
	Coil20	1440	1024	84	20
	YALE-B	2414	1024	116	38

The data with no class label is denoted by ‘–’. For the data with class labels, the d is equal to the dimension that can retain 95% energy after PCA is applied

4.1 Experimental setting

We compared our proposed MPME method to various existing dimensionality reduction methods, including PCA (Jolliffe 1986), KPCA (Schölkopf et al. 1999), LE (Belkin and Niyogi 2001), LLE (Saul and Roweis 2003), MVU (Weinberger et al. 2004), three variants of MVU, i.e., MVUineq (Weinberger and Saul 2006), MVU2 (Weinberger and Saul 2006), and ℓ MVU (Weinberger et al. 2005), GPLVM (Lawrence 2005), MEU (Lawrence 2012), and tSNE (van der Maaten and Hinton 2008) on the 14 datasets shown in Table 1. Most of the methods are implemented in the drtoolbox,¹ except for MEU,² and the variants of MVUs.³ These methods have several parameters for dimensionality reduction that must be set before applying them to the datasets. For LLE and LE, we employed the normal neighborhood selection strategy implemented in the drtoolbox to construct the neighborhood graphs. For methods using a kernel function, such as KPCA, LE, GPLVM, we employed a Gaussian kernel. For MPME, Φ was set to the Euclidean distances of the input data, and its parameters λ and C are discussed in Sect. 4.2. In the same way as the MVUs and MEU, MPME uses KPCA as the embedding step to obtain embedded points. The default settings of various other parameters were used as suggested in the drtoolbox and corresponding toolbox unless noted otherwise.

The datasets used in the experiments are summarized in Table 1. The visualization results of the embedded points learned by compared methods are shown in 2-D or 3-D for comparisons (see Sect. 4.3). The clustering results are reported with Kmeans on the learned embedding that was obtained by each compared method (see Sect. 4.4). Two popular evaluation criteria, accuracy and normalized mutual information (NMI), are used for comparing

¹ <https://lvdmaaten.github.io/drtoolbox/>.

² <https://github.com/SheffieldML/meu>.

³ <http://www.cs.cornell.edu/~kilian/code/files/mvu2012.zip>.

Fig. 2 The sensitivity analysis of the proposed MPME method on *DistortedSShape* by varying C in the grid $[0.1, 1, 10, 100, \infty]$ and fixing $\lambda = 10$. The *first row* shows the embedded points in 2-D space and the *second row* shows the adjacency matrix of $\mathbf{W}$. nz is the number of nonzeros, and the percentage sparsity is also given

Fig. 3 The sensitivity analysis of the proposed MPME method on *DistortedSShape* by varying λ in the grid $[0.1, 0.5, 1, 5, 10]$ and fixing $C = \infty$. The *first row* shows the embedded points in 2-D space and the *second row* shows the adjacency matrix of $\mathbf{W}$. nz is the number of nonzeros, and its percentage sparsity is also given

clustering methods (Nie et al. 2009). For fair comparisons, we set d to be the dimension that retains 95% of its energy after applying PCA. All methods use the number of true clusters as the number of clusters for Kmeans and the same reduced dimensionality for each dataset.

4.2 Parameter sensitivity analysis

Our proposed MPME method has three parameters: the low dimension d , the regularization parameter C , and the precision λ of prior distribution. For dimensionality reduction, we generally fix d to be 2 or 3 for data visualization, or preset a proper value for clustering as we do in this paper by retaining enough information from the original data. Thus, we study the sensitivity of C and λ by investigating one parameter and fixing others in terms of embedding visualization and its corresponding clustering performance.

We first conducted the sensitivity analysis for data visualization. Figures 2 and 3 demonstrate the skeleton structures of embedded points learned by the proposed method on *DistortedSShape* by varying $C \in [0.1, 1, 10, 100, \infty]$ with $\lambda = 10$ and $\lambda \in [0.1, 0.5, 1, 5, 10]$ with $C = \infty$, respectively. We can clearly see that the skeleton becomes smoother and noise is gradually removed when C increases. Moreover, the graph represented by the similarity

Fig. 4 The sensitivity analysis of the proposed MPME method in terms of clustering performance evaluation criteria such as accuracy and NMI on datasets **a** USPS and **b** YALE-B by varying λ in the grid $[0.1, 10]$ and C in $[0.1, \infty]$

matrix $\mathbf{W}$ becomes sparser when C increases. These empirical observations are consistent with our theoretical analysis in Sect. 3.3 that C is a parameter for controlling the sparsity of the learned similarity matrix. Moreover, parameter λ is also important, which can change the location of embedding points as shown in Fig. 3. The importance of λ becomes clear for clustering problems, which is discussed below.

We also investigated the clustering performance of the proposed method by varying λ and C . The clustering results on datasets *USPS* and *YALE-B* with respect to λ and C are reported in Fig. 4. These results imply that the clustering performance largely depends on the given datasets. We observed that the clustering performance is very sensitive to λ . A small λ is preferred on *USPS*, while a large λ is preferred on *YALE-B*. This is reasonable because data points in *USPS* form clear clustering structure, while the data points in *YALE-B* have a manifold structure (Nie et al. 2010).

4.3 Data visualization

For the data visualization in a 2-D/3-D Euclidean space, we performed experiments on five synthetic datasets and two real datasets.

4.3.1 Synthetic datasets

As discussed in Sect. 3.1, we are particularly interested in the manifold structures with a clear and smooth skeleton, on which the data is generated with noises. Datasets, *Circle* and *DistortedSShape*,⁴ were used in the principal curve learning (Kégl et al. 2000), where the underlying structures are a circle and an S-shaped curve, respectively. The two datasets are also noisy. The dataset *2moons*⁵ was used in manifold-based semi-supervised learning (Belkin et al. 2006) to distinguish the two classes, so two smooth curves exist that reflect the true skeleton structure. Following the work of van der Maaten et al. (2009), we generated two 3-D datasets, *helix* and *twinpeaks*, by using *drtoolbox*, where the data point $\mathbf{y}_i$ of *helix* is computed by $\mathbf{y}_i = [(2 + \cos(8p_i)) \cos(p_i), (2 + \cos(8p_i)) \sin(p_i), \sin(8p_i)]$, and the data point $\mathbf{y}_i$ of *twinpeaks* is computed by $\mathbf{y}_i = [1 - 2p_i, \sin(\pi - 2\pi p_i) \tanh(3 - 6q_i)]$ where p_i and q_i are two random numbers that were sampled from a uniform distribution with support $[0, 1]$. From the generative process, we know that *helix* demonstrates a circle structure and *twinpeaks* has two intersecting curves. Since most of existing dimensionality reduction methods are nonlinear, the structure of embedded data might be different from the structure of original data, but the smooth skeleton should be retained and noises should be removed or weakened.

Results for data visualization in 2-D embedding space on five synthetic datasets using nine different methods are shown in Fig. 5. The first three datasets consist of noisy data points scattered around the intrinsic skeleton structure. For example, the structure of *DistortedSShape* data is transformed from a distorted S-shape to smooth bell-shape using the proposed method and MEU, but the other methods failed to obtain a smooth structure by simply scaling the data. Also on the data *2moons* dataset, our proposed method and tSNE identified two smooth curves of the two-moon shape, while other methods failed. On the five tested datasets, only the proposed method correctly captured all the smooth skeleton structures. These observations are in line with the goal of learning a smooth skeleton structure from various datasets with different complexities of skeleton structures.

4.3.2 Cancer data

We interrogated a large-scale, publicly available breast cancer dataset (Curtis et al. 2012) for cancer progression modeling. The dataset contains the expression levels of over 25,000 gene transcripts obtained from 144 normal breast tissue samples and 1989 tumor tissue samples. Using a nonlinear regression method, a total of 359 genes were identified that may play a role in cancer development (Sun et al. 2014). We set MPME with $\lambda = 1.2$ and $C = \infty$, and visualize the embedded points in 3-D space, where $C = \infty$ is for the skeleton structure in terms of the sparsity solution of $\mathbf{W}$. For the purpose of understanding this in details, we also report the kernel matrices learned by variants of the MVUs.

Figure 6 shows the embedded points and the learned similarity matrix in 3-D space from MPME and the MVUs. Figure 7 shows the embedded points of the remaining methods. Each tumor sample is colored with its corresponding PAM50 subtype label, a molecular approximation that uses a 50-gene signature to group breast tumors into five subtypes including normal-like, luminal A, luminal B, HER2+ and basal (Parker et al. 2009). The basal and HER2+ subtypes are known to be the most aggressive types of breast tumor. The skeleton structure learned by MPME in the 3-D space suggests a linear bifurcating progression path,

⁴ <https://www.lri.fr/~kegl/researchUdeM/research/pcurves/implementations/Samples/>.

⁵ manifold.cs.uchicago.edu/manifold_regularization/2moons.mat.

Fig. 5 Five synthetic datasets and their corresponding visualization results of embedded data points in 2-D space using nine dimensionality reduction methods including the proposed MPME method, KPCA, LE, MVUs (MVU, MVUineq, and MVU2), GPLVM, MEU, and tSNE

starting from normal tissue samples, to normal-like samples, and through to luminal A before continuing to luminal B, and then diverging to either the basal or HER2+ subtypes. A conceptual linear evolution progression model was proposed that posits that basal tumors are

Fig. 6 The 3-D visualization results (*better viewing in color*) of breast cancer data in terms of embedded points and adjacency matrix of W (sorted by subtypes) learned by our proposed MPME method, or kernel matrices by MVU, MVUineq and MVU2, respectively. For ease of illustrating the embedding result obtained by MPME, we annotate each subtype of embedded samples using *eclipses* with *dashed lines* and the progression path using *curves*

Fig. 7 Results of breast cancer data including the 3-D visualization results of embedded points by KPCA, LE, GPLVM, MEU, tSNE and PCA, respectively

derived from luminal tumors, seeing Figure 6 in [Creighton \(2012\)](#). The revealed skeleton structure is consistent with the proposed branching architecture of cancer progression ([Sun et al. 2014](#)). For ease of understanding, we annotate each subtype of embedded samples obtained by MPME using ellipses with dashed lines and the progression path using curves. LE and MEU demonstrate similar trends, but they are more noisy and less smooth than MPME. tSNE can obtain a clustering structure, but the interconnections between the different subtypes are missing. MVUs do not have a clear skeleton structure for describing cancer progression, and also the clustering structures of subtypes are not as well formed as that of

MPME by comparing the learned kernels of MVUs and the similarity matrix of MPME. The remaining methods do not demonstrate any skeleton structures. These observations imply that our proposed MPME method can unveil a cancer progression path better than baseline methods from high-dimensional noisy breast cancer data.

4.3.3 Teapot data with added noises

The dataset *Teapot*⁶ is a collection of 400 teapot images that are taken successively as a teapot is rotated 360°. Each image consists of 76×101 RGB pixels (Weinberger et al. 2005), i.e., each image lies in 23,028 dimensional space. Each pixel value is divided by 255 so as to normalize each pixel to $[0, 1]$. As shown in Weinberger et al. (2004, 2005); Weinberger and Saul (2006), MVUs can correctly recover the circle structure from teapot data. Thus, unlike the previous experimental settings, we intentionally add noise to the data so that the circular embeddings were not as clear as in the original data. Specifically, we add the noise sampled from the uniform distribution in $[0, \rho]$ where $\rho \in [0, 0.1, 0.2, 0.4, 0.5]$, to each dimension of teapot images. For $\rho = 0$, the data is the same as the original teapot data. The larger the ρ becomes, the more noise the newly generated data has. Since the normalized pixel is in $[0, 1]$, the generated data with a noise rate $\rho \leq 0.5$ are expected to have enough information to uncover the circular structure.

The visualization results obtained by MPME and MVUs in 2-D space on five datasets with different noise rates are shown in Fig. 8. We made the following observations. (i) All methods can correctly uncover the circular structure of original teapot data in 2-D space ($\rho = 0$); (ii) When the added noise is small ($\rho \in [0.1, 0.2, 0.4]$), all methods still succeeded, but the circular structure obtained by MPME was smoother than the MVUs. (iii) When $\rho = 0.5$, all four methods failed due to the large error rate, but the embeddings of MPME demonstrate a smooth skeleton structure, while the MVUs do not. For datasets with a large proportion of noise ($\rho = 0.5$), we expected to see the circular structure in a high-dimensional space. Based on this conjecture, we conducted another set of experiments to learn the embeddings of noisy datasets with $\rho \in [0.4, 0.5]$ in 3-D space. The visualization results are shown in Fig. 9. It is clear that MPME can correctly recover the smooth skeleton structure—a circle, but the MVUs failed in most cases except for MVU2 where $\rho = 0.4$. These results imply that our proposed MPME method can robustly recover smooth skeleton structures from noisy data.

4.4 Clustering with dimensionality reduction

To evaluate the obtained embedded points for clustering problems, we conducted an experiment on seven benchmark datasets, including one object dataset, *Coil20* (Nene et al. 1996); one face dataset, *YALE-B* (Nie et al. 2009); one spoken letter recognition dataset, *Isolet* (Belkin et al. 2006); and four other datasets.⁷ In the experiments, we tuned λ in a range $[0.1, 10]$ and C in $[0.1, 1, 10, 100, \infty]$. We ran Kmeans with 20 random initializations and report the result of the best objective so as to alleviate the issue of local solution.

Table 2 reports the results of ten methods on seven datasets including their accuracy and normalized mutual information (NMI). We made the following observations. Kmeans on the embedded points after applying PCA provides a marginal improvement over Kmeans on the original space. That is, the reduced dimension by retaining 95% of its energy of data can capture the most information from the data. Except for MPME and tSNE, the other methods

⁶ <http://www.cc.gatech.edu/~lsong/data/teapotdata.zip>.

⁷ <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>.

Fig. 8 Results of teapot data in 2-D space using the proposed MPME method and variants of MVUs by varying the noisy ratio added to the original data in the range $[0, 0.1, 0.2, 0.4, 0.5]$. The larger the noise ratio is, the more noises are added to the original data. The visualization of embedded points and its adjacency matrix of learned similarity matrix by MPME are shown

Fig. 9 Results of teapot data in 3-D space using the proposed MPME method and variants of MVUs by varying the noisy ratio added to the original data in the range $[0.4, 0.5]$

Table 2 Clustering results of seven datasets using 10 methods in terms of accuracy and NMI

Method	Letter	Pendigits	Satimage	USPS	Coil20	YALE-B	Isolet
Accuracy							
Kmeans	0.2632	0.6544	0.6685	0.6153	0.6674	0.1081	0.5633
PCA	0.2634	0.6527	0.6681	0.6208	0.6674	0.1110	0.5633
KPCA	0.2596	0.7387	0.7145	0.1933	0.2243	0.0766	0.5008
LLE	0.3084	0.1215	0.6510	0.1624	0.3493	0.0667	0.5021
LE	0.0674	0.7742	0.6692	0.4828	0.3889	0.1475	0.5361
ℓ MVU	0.1484	0.5692	0.6886	0.3896	0.5146	0.0684	0.5989
GPLVM	0.2634	0.6527	0.6681	0.4709	0.6674	0.0671	0.5630
MEU	0.2518	0.7138	0.7398	0.6129	0.3590	0.3094	0.5476
tSNE	0.3856	0.7504	0.6257	0.7603	0.7319	0.3592	0.5976
MPME	0.3432	0.8276	0.7411	0.7344	0.7590	0.4010	0.6935
NMI							
Kmeans	0.3621	0.6669	0.6097	0.5657	0.7845	0.1694	0.0117
PCA	0.3591	0.6627	0.6090	0.5664	0.7845	0.1781	0.0117
KPCA	0.3528	0.6849	0.6028	0.1085	0.3104	0.0911	0.0006
LLE	0.3992	0.0059	0.5080	0.0106	0.3848	0.0810	0.0004
LE	0.0211	0.7397	0.5812	0.4522	0.5077	0.2534	0.0369
ℓ MVU	0.1379	0.6422	0.5818	0.3858	0.6595	0.0870	0.0292
GPLVM	0.3591	0.6627	0.6090	0.3824	0.7845	0.0716	0.0116
MEU	0.3341	0.7635	0.6826	0.5743	0.5067	0.4174	0.0151
tSNE	0.5059	0.7846	0.6266	0.7868	0.8743	0.5925	0.0288
MPME	0.4616	0.8139	0.6594	0.7611	0.8862	0.5301	0.1109

The best results are in bold

hardly outperform PCA, which is also observed in [van der Maaten et al. \(2009\)](#). Although MEU shares some properties with MPME, MPME still outperforms MEU as shown in Table 2 due to the key differences discussed in Sect. 3.6.2. These observations further confirm that both the inequality constraints for preserving expected distances and the probability model can contribute to a robust dimensionality reduction model by learning a proper similarity matrix for clustering problems.

5 Conclusion

We propose a novel probabilistic dimensionality reduction framework for learning a smooth skeleton structure from noisy data by directly modeling the posterior distribution of latent points with a set of constraints over the expected pairwise distances. Our model not only tolerates noisy data and uncovers a smooth skeleton structure by learning a sparse positive similarity matrix, but also gives a natural interpretation of LE and MVU from the probabilistic point of view. Extensive experiments demonstrated that our proposed method can achieve better visualization results on seven datasets, successfully unveil the circular structure of teapot data with added noises at different noise ratios. The method also yields superior clustering performance compared to various baseline methods on seven real datasets.

Acknowledgements This project is partially supported by the ARC Future Fellowship FT130100746 and ARC grant LP150100671.

References

- Belkin, M., & Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *NIPS* (Vol. 14, pp. 585–591). Cambridge: MIT Press.
- Belkin, M., Niyogi, P., & Sindhvani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7, 2399–2434.
- Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24(3), 179–195.
- Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.
- Burges, C. J. C. (2009). Dimension reduction: A guided tour. *FTML*, 2(4), 275–365.
- Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5), 1190–1208.
- Creighton, C. (2012). The molecular profile of luminal b breast cancer. *Biologics*, 15, 440.
- Curtis, C., Shah, S. P., Chin, S., et al. (2012). The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature*, 486(7403), 346–352.
- Elhamifar, E., & Vidal, R. (2011). Sparse manifold clustering and embedding. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *NIPS* (pp. 55–63). Granada, Spain: Granada Congress and Exhibition Centre.
- Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. New York: Academic press.
- Greaves, M., & Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381), 306–313.
- Gupta, A. K., & Nagar, D. K. (1999). *Matrix variate distributions* (Vol. 104). Boca Raton: CRC Press.
- Hastie, T., & Stuetzle, W. (1989). Principal curves. *JASA*, 84, 502–516.
- Jebara, T. (2001). *Discriminative, generative and imitative learning*. Ph.D. thesis, Massachusetts Institute of Technology.
- Jimenez, L. O., & Landgrebe, D. (1998). Supervised classification in high-dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data. *TSMC*, 28(1), 39–54.
- Jolliffe, J. T. (1986). *Principal component analysis*. Berlin: Springer.
- Kégl, B., Krzyzak, A., Linder, T., & Zeger, K. (2000). Learning and design of principal curves. *IEEE TPAMI*, 22(3), 281–297.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86.
- 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.
- Mao, Q., Tsang, I. W., Gao, S., & Wang, L. (2015). Generalized multiple kernel learning with data-dependent priors. *IEEE TNNLS*, 24(2), 248–261.
- Nene, S. A., Nayar, S. K., & Murase, H. (1996). *Columbia object image library (coil-20)*. Technical Report CUCS-005-96.
- Nie, F., Xu, D., Tsang, I. W., & Zhang, C. (2009). Spectral embedded clustering. In C. Boutilier (Ed.), *IJCAI* (pp. 1181–1186). Menlo Park, California: AAAI Press.
- Nie, F., Xu, D., Tsang, I. W., & Zhang, C. (2010). Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *TIP*, 19(7), 1921–1932.
- Parker, J., Mullins, M., Cheang, M., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8), 1160–1167.
- Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. Cambridge: The MIT Press.
- Saul, L. K., & Roweis, S. T. (2003). Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *JMLR*, 4, 119–155.
- Schölkopf, B., Smola, A., & Muller, K. (1999). Kernel principal component analysis. In B. Schölkopf, A. J. Smola, & C. J. C. Burges (Eds.), *Advances in Kernel methods–Support vector learning* (pp. 327–352). Cambridge: MIT Press.
- Smola, A. J., & Kondor, R. (2003). Kernels and regularization on graphs. In B. Schölkopf & M. K. Warmuth (Eds.), *ICML* (pp. 144–158). New York: Springer.
- Song, L., Smola, A., Gretton, A., & Borgwardt, K. (2007). A dependence maximization view of clustering. In Z. Ghahramani (Ed.), *ICML* (pp. 815–822). New York: ACM.

- Sun, Y., Yao, J., Nowak, N., & Goodison, S. (2014). Cancer progression modeling using static sample data. *Genome Biology*, 15(8), 440.
- Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 61(3), 611–622.
- Tutuncu, R., Toh, K., & Todd, M. (2003). Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, 95, 189–217.
- Vandenberghe, L., & Boyd, S. (1996). Semidefinite programming. *SIAM Review*, 38(1), 49–95.
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *JMLR*, 9(2579–2605), 85.
- van der Maaten, L., Postma, E. O., & van den Herik, H. J. (2009). *Dimensionality reduction: A comparative review*. Tilburg University Technical Report, TiCC-TR 2009-005.
- Weinberger, K. Q., Packer, B. D., & Saul, L. K. (2005). Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In R. G. Cowell & Z. Ghahramani (Eds.) *Proceedings of the 10th international workshop on artificial intelligence and statistics* (pp. 381–388).
- Weinberger, K. Q., & Saul, L. K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *IJCV*, 70(1), 77–90.
- Weinberger, K. Q., Sha, F., & Saul, L. K. (2004). Learning a kernel matrix for nonlinear dimensionality reduction. In C. E. Brodley (Ed.), *ICML* (p. 106). New York: ACM.
- Zhu, J., & Xing, E. P. (2009). Maximum entropy discrimination markov networks. *JMLR*, 10, 2531–2569.