Nonlinear Dimensionality Reduction on Graphs

Yanning Shen, Panagiotis A. Traganitis and Georgios B. Giannakis

Dept. of ECE and DTC, University of Minnesota, Minneapolis, USA

Emails: {shenx513, traga003, georgios}@umn.edu

Abstract—In this era of data deluge, many signal processing and machine learning tasks are faced with high-dimensional datasets, including images, videos, as well as time series generated from social, commercial and brain network interactions. Their efficient processing calls for dimensionality reduction techniques capable of properly compressing the data while preserving task-related characteristics, going beyond pairwise data correlations. The present paper puts forth a nonlinear dimensionality reduction framework that accounts for data lying on known graphs. The novel framework turns out to encompass most of the existing dimensionality reduction methods as special cases, and it is capable of capturing and preserving possibly nonlinear correlations that are ignored by linear methods, as well as taking into account information from multiple graphs. An efficient algorithm admitting closed-form solution is developed and tested on synthetic datasets to corroborate its effectiveness.

Index Terms—Dimensionality reduction, nonlinear modeling, graph signal processing

I. INTRODUCTION

The massive development of connected devices and highly precise instruments has introduced the world to vast volumes of high-dimensional data. Traditional data analytics cannot cope with these massive amounts, which motivates well investigating dimensionality reduction schemes capable of gleaning out efficiently low-dimensional information from large-scale datasets. Dimensionality reduction is a vital first step to render tractable critical learning tasks, such as large-scale regression, classification, and clustering, and allows for processing of datasets that might otherwise not be tractable.

Dimensionality reduction methods have been extensively studied by the signal processing and machine learning communities [2], [10], [16], [17]. Principal component analysis (PCA) [10] is the 'workhorse' method yielding low-dimensional representations that preserve most of the high-dimensional data variance. Multi-dimensional scaling (MDS) [12] on the other hand, maintains pairwise distances between data when going from high- to low-dimensional spaces, while local linear embedding (LLE) [16] only preserves relationships between neighboring data. Information from non-neighboring data is lost in LLE's low-dimensional representation, which may in turn influence the performance of ensuing tasks such as classification or clustering [6], [21]. It is also worth stressing that all aforementioned approaches capture and preserve linear relationships between data. However, for data residing on highly nonlinear manifolds using only linear relations might produce low-dimensional representations that are not accurate. Generalizing PCA, Kernel PCA [9] can capture nonlinear relationships between data, for a preselected

kernel function. In addition, Laplacian eigenmaps [2] preserve nonlinear similarities between neighboring data.

While all the aforementioned approaches have been successful in reducing the dimensionality of various types of data, they do not consider additional information during the dimensionality reduction process. This prior information may be task specific, e.g. provided by some "expert" or the physics of the problem, or it could be inferred from alternate views of the data, and can provide additional insights for the desired properties of the low-dimensional representations. In fMRI signals for instance, in addition to time series collected at different brain regions, one may also have access to the connectivity patterns among these regions. As shown in [8], [9], [18], [19] for PCA, this additional information can be encoded in a graph, and incorporated into the dimensionality reduction process through graph-aware regularization.

The present paper puts forth a novel framework for dimensionality reduction that can capture nonlinear relations between data, and also exploit additional information through graph regularization. This framework encompasses all aforementioned approaches, and markedly broadens their scope.

II. PRELIMINARIES AND PROBLEM STATEMENT

Consider a dataset with N vectors of dimension D collected as columns of the matrix $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_N]$. Without loss of generality, it will be assumed that the sample mean $N^{-1} \sum_{n=1}^{N} \mathbf{x}_n$ has been removed from each $\mathbf{x}_i$. Dimensionality reduction looks for a set of d-dimensional vectors $\{\mathbf{y}_i\}_{i=1}^{N}$, with $d < D$, that preserve certain properties of $\{\mathbf{x}_i\}$. MDS for instance aims to preserve pairwise distances among $\{\mathbf{x}_i\}$ when obtaining the corresponding low-dimensional representations $\{\mathbf{y}_i\}$, while LLE attempts to preserve neighborhoods. As will be shown in the ensuing subsections, all these dimensionality reduction schemes are special cases of kernel-based PCA.

A. Principal component analysis

Given $\mathbf{X}$, PCA finds a linear subspace of dimension d so that hopefully all the data lie on or close to it (in the least-squares sense). Specifically, PCA solves

$$\min_{\mathbf{U}_d, \{\mathbf{y}_i\}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{U}_d \mathbf{y}_i\|_2^2 \quad \text{s. to} \quad \mathbf{U}_d^\top \mathbf{U}_d = \mathbf{I} \quad (1)$$

where $\mathbf{U}_d \in \mathbb{R}^{D \times d}$ is an orthonormal matrix. The optimal solution of (1) is $\mathbf{y}_i = \mathbf{U}_d^\top \mathbf{x}_i$, where $\mathbf{U}_d$ is formed by the eigenvectors of $\mathbf{X}\mathbf{X}^\top = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$ corresponding to the d largest eigenvalues [7]. For future use, consider the singular value decomposition (SVD) $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$. Given $\{\mathbf{y}_i\}$, the original vectors can be recovered as $\mathbf{x}_i = \mathbf{U}_d \mathbf{y}_i$. PCA

thrives when data lie close to a d-dimensional hyperplane. Its complexity is that of eigendecomposing $\mathbf{XX}^\top$, i.e., $\mathcal{O}(ND^2)$, which means PCA is more affordable when $D \ll N$. In contrast, dimensionality reduction of small sets of high-dimensional vectors ($D \gg N$) becomes more tractable with dual PCA.

B. Dual PCA and Kernel PCA

The SVD of $\mathbf{Y}$ implies that $\mathbf{U}_d = \mathbf{YV\Sigma}^{-1}$, which in turn yields the low-dimensional vectors as $\boldsymbol{\psi}_i = \mathbf{U}_d^\top \mathbf{y}_i = \boldsymbol{\Sigma}_d^{-1} \mathbf{V}_d^\top \mathbf{Y}^\top \mathbf{y}_i$. Collecting $\boldsymbol{\Psi} := [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N]$, we have

$$\mathbf{Y} = \mathbf{U}_d^\top \mathbf{X} = \boldsymbol{\Sigma}_d \mathbf{V}_d^\top \qquad (2)$$

where $\boldsymbol{\Sigma}_d \in \mathbb{R}^{d \times d}$ is a diagonal matrix containing the d leading eigenvalues of $\mathbf{X}^\top \mathbf{X}$, and $\mathbf{V}_d \in \mathbb{R}^{N \times d}$ is the submatrix of $\mathbf{V}$ collecting the corresponding eigenvectors of $\mathbf{X}^\top \mathbf{X}$. The complexity of dual PCA is $\mathcal{O}(DN^2)$; therefore, it is preferable when $D \gg N$. Moreover, it can be readily verified that besides (1), $\mathbf{Y}$ in (2) is also the optimal solution to the following optimization problem

$$\min_{\mathbf{Y}} \|\mathbf{K}_x - \mathbf{Y}^\top \mathbf{Y}\|_F^2 \quad \text{s. to } \mathbf{YY}^\top = \boldsymbol{\Lambda}_d \qquad (3)$$

where $\mathbf{K}_x := \mathbf{X}^\top \mathbf{X}$, and $\boldsymbol{\Lambda}_d$ denotes a $d \times d$ diagonal matrix containing the d largest eigenvalues of $\mathbf{K}_x$. Compared to PCA, dual PCA needs only the inner products $\{\mathbf{x}_i^\top \mathbf{x}_j\}$ in order to obtain the low-dimensional representations, but not $\mathbf{X}$ itself. Hence, dual PCA can yield low-dimensional vectors $\{\mathbf{y}_i\}$ of general (non-metric) objects that are not necessarily expressed using vectors $\{\mathbf{x}_i\}$, so long as inner products (a.k.a correlations) of the latter are known. Furthermore, expanding the cost in (3), we can equivalently express it as

$$\min_{\mathbf{Y}:\mathbf{YY}^\top = \boldsymbol{\Lambda}_d} -\text{tr}(\mathbf{YK}_x\mathbf{Y}^\top) \Leftrightarrow \min_{\mathbf{Y}:\mathbf{YY}^\top = \boldsymbol{\Lambda}_d} \text{tr}(\mathbf{YK}_x^{-1}\mathbf{Y}^\top). \quad (4)$$

While PCA performs well for data that lie close to a hyperplane, this property might not be true for many datasets [9]. In such cases one may resort to kernel PCA. Kernel PCA "lifts" $\{\mathbf{x}_i\}$ using a nonlinear function $\boldsymbol{\phi}$, onto a higher (possibly infinite) dimensional space, where the data may lie on or near a linear hyperplane, and then finds low-dimensional representations $\{\mathbf{y}_i\}$. Kernel PCA is obtained by solving (3) with $[\mathbf{K}_x]_{i,j} = \kappa(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}^\top(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j)$, where $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ denotes a prescribed kernel function [5].

C. Local linear embedding

Another popular method that deals with data that cannot be presumed close to a hyperplane is LLE. It assumes that $\{\mathbf{x}_i\}$ lie on a smooth manifold, which can be locally approximated by tangential hyperplanes. Specifically, LLE assumes that each datum can be expressed as a linear combination of its neighbors; that is, $\mathbf{x}_i = \sum_{j \in \mathcal{N}_i} w_{ij}\mathbf{x}_j + \mathbf{e}_i$, where $\mathcal{N}_i$ is a set containing the indices of the nearest neighbors of $\mathbf{x}_i$, in the Euclidean distance sense. In order to solve for $\{w_{ij}\}$, the following optimization problem is considered

$$\mathbf{W} = \arg\min_{\check{\mathbf{W}}} \|\mathbf{X} - \mathbf{X}\check{\mathbf{W}}\|_F^2$$
$$\text{s. to } \check{w}_{ij} = 0, \quad \forall i \notin \mathcal{N}_j, \quad \sum_i \check{w}_{ij} = 1 \qquad (5)$$

where $\check{w}_{ij}$ denotes the (i, j)-th entry of $\check{\mathbf{W}}$. Upon obtaining $\mathbf{W}$ as the constrained least-squares solution in (5), LLE finds $\{\mathbf{y}_i\}$ that best preserve the neighborhood relationships encoded in $\mathbf{W}$, by solving

$$\min_{\mathbf{Y}:\mathbf{YY}^\top = \boldsymbol{\Lambda}_d} \|\mathbf{Y} - \mathbf{YW}\|_F^2$$
$$\Leftrightarrow \min_{\mathbf{Y}:\mathbf{YY}^\top = \boldsymbol{\Lambda}_d} \text{tr}[\mathbf{Y}(\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^\top \mathbf{Y}^\top] \qquad (6)$$

where $\boldsymbol{\Lambda}_d$ is a diagonal matrix. Conventional LLE adopts $\boldsymbol{\Lambda}_d = \mathbf{I}$, which is subsumed by the constraint in (6). Nonetheless, the difference is just a scaling of $\{\mathbf{y}_i\}$ when $\boldsymbol{\Lambda}_d \neq \mathbf{I}$. If the diagonal of $\boldsymbol{\Lambda}_d$ collects the d smallest eigenvalues of matrix $(\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^\top$, then (6) is a special case of kernel PCA with [cf. (4)]

$$\mathbf{K}_x = [(\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^\top]^\dagger \qquad (7)$$

where $\dagger$ denotes pseudo inverse. Similarly, other popular dimensionality reduction methods such as MDS and Laplacian eigenmaps can also be viewed as special cases of kernel PCA, by appropriately selecting $\mathbf{K}_x$ [4]. Thus, (4) can be viewed as an encompassing framework for nonlinear dimensionality reduction. This framework is the foundation of the general graph-aware methods we develop in the ensuing section.

III. GRAPH-AWARE NONLINEAR DIMENSIONALITY REDUCTION

Matrices $\mathbf{K}_x$ for dual PCA, kernel PCA and LLE depend only on $\mathbf{X}$. However, as mentioned in Sec. I, additional structural information potentially useful for the dimensionality reduction task may be available. This knowledge can be encoded in a graph and embodied in $\mathbf{Y}$ via graph regularization. Specifically, suppose there exists a graph $\mathcal{G}$ over which the data is smooth; that is, vectors $\{\mathbf{x}_i\}$ on nodes of $\mathcal{G}$ are also close to each other in Euclidean distance. With $\mathbf{A}$ denoting the adjacency matrix of $\mathcal{G}$, we have $[\mathbf{A}]_{ij} = a_{ij} \neq 0$ if node i is connected with node j. The Laplacian of $\mathcal{G}$ is defined as $\mathbf{L}_\mathcal{G} := \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is a diagonal matrix with entries $[\mathbf{D}]_{ii} = d_{ii} = \sum_j a_{ij}$. Now consider

$$\text{tr}(\mathbf{YL}_\mathcal{G}\mathbf{Y}^\top) = \sum_{i=1}^N \sum_{j \neq i}^N a_{ij}(\mathbf{y}_i - \mathbf{y}_j)^2 \qquad (8)$$

which is a weighted sum of the distances of adjacent $\mathbf{y}_i$'s on the graph. By minimizing (8) over $\mathbf{Y}$, the low-dimensional representations corresponding to adjacent nodes with large edge weights a_{ij} will be close to each other.

A. Kernel PCA on graphs

Introducing (8) as a regularization term to the original kernel PCA problem in (4), we arrive at

$$\min_{\mathbf{Y}:\mathbf{YY}^\top = \boldsymbol{\Lambda}_d} \text{tr}(\mathbf{YK}_x^{-1}\mathbf{Y}^\top) + \gamma \text{tr}(\mathbf{YL}_\mathcal{G}\mathbf{Y}^\top) \qquad (9)$$

where γ is a positive scalar, and $\boldsymbol{\Lambda}_d$ collects the d smallest eigenvalues of $\mathbf{K}_x^{-1} + \gamma\mathbf{L}_\mathcal{G} = \bar{\mathbf{V}}\boldsymbol{\Lambda}\bar{\mathbf{V}}^\top$. Combining the Laplacian regularization with the kernel PCA formulation, (9) is capable of finding $\{\mathbf{y}_i\}$ that preserve the "lifted" covariance captured by $\mathbf{K}_x$, while at the same time, promoting the

---

Algorithm 1 Local nonlinear embedding over graphs

Input: $\mathbf{X}$, γ
S1. Estimate $\mathbf{W}$ from $\mathbf{X}$.
S2. Obtain kernel matrix $\mathbf{K}_x$ via (7).
S3. Find low-dimensional representations $\mathbf{Y}$ using (10).

---

smoothness of the low-dimensional representations over the graph $\mathcal{G}$. Problem (9) admits a closed-form solution as

$$\mathbf{Y} = \mathbf{\Lambda}_d^{1/2}\bar{\mathbf{V}}_d^{\top} \tag{10}$$

where $\bar{\mathbf{V}}_d$ denotes the sub-matrix of $\bar{\mathbf{V}}$ containing columns corresponding to eigenvalues in $\mathbf{\Lambda}_d$. When the γ is set to 0, one obtains the solution of kernel PCA [cf. (2)].

Remark 1: When the kernel function needed to form $\mathbf{K}_x$ is not known a priori, one can use $\mathbf{K}_x = \sum_{q=1}^{Q} \theta_q \mathbf{K}_x^{(q)}$ in (9), where $\{\mathbf{K}_x^{(q)}\}$ is a known dictionary of kernel matrices, and $\{\theta_q\}$ are estimated as variables in (9) [1]. In addition, instead of directly using $\mathbf{L}_{\mathcal{G}}$, a family of graph kernels $r(\mathbf{L}_{\mathcal{G}}) := \mathbf{U}_{\mathcal{G}} r(\mathbf{\Lambda}) \mathbf{U}_{\mathcal{G}}^{\top}$ can be employed, where $r(.)$ is a scalar function of the eigenvectors of $\mathbf{L}_{\mathcal{G}}$. By appropriately selecting $r(.)$, different graph properties can be accounted for. As an example, when r sets eigenvalues above a certain threshold to 0, it acts as a sort of "low pass" filter over the graph. A detailed discussion of how to choose graph kernels can be found in [14], [15]. Furthermore, instead of prescribing an $r(\mathbf{L}_{\mathcal{G}})$, a data-driven dictionary-based approach could also be employed to learn the proper graph kernel for the task at hand [15].

Remark 2: Even though only a single graph regularizer is introduced in (9), our method is flexible to include multiple graph regularizers based on different graphs. Therefore, the proposed method offers a powerful tool for dimensionality reduction on so-called multi-layer graphs, which encode the relationships among data across multiple graphs [11], [22].

B. Local nonlinear embedding on graphs

Broadening the premise of LLE, we pursue here a more general dimensionality reduction framework that captures nonlinear correlations between neighboring data, in addition to the structure induced by the graph $\mathcal{G}$. To this end, suppose that each datum can be represented by its neighbors as

$$[\mathbf{x}_i]_m = \sum_{j \in \mathcal{N}_i} h_{ij}([\mathbf{x}_j]_m) + [\mathbf{e}_i]_m, \quad m = 1, \dots, D \tag{11}$$

where $\{h_{ij}(\cdot)\}_{i,j=1}^{N}$ are prescribed scalar nonlinear functions admitting a Pth-order expansion

$$h_{ij}(z) = \sum_{p=1}^{P} w_{ij}[p] z^p \tag{12}$$

where coefficients $\{w_{ij}[p]\}$ are to be determined. Taylor's expansion asserts that for P sufficiently large, (12) offers an accurate approximation for all memoryless differentiable nonlinear functions. Such a nonlinear model has been used for gragh topology identification [20]. With the estimated $\mathbf{W}$ at hand, the low-dimensional representations can be obtained via (10), with $\mathbf{K}_x$ as in (7); see also Algorithm 1.
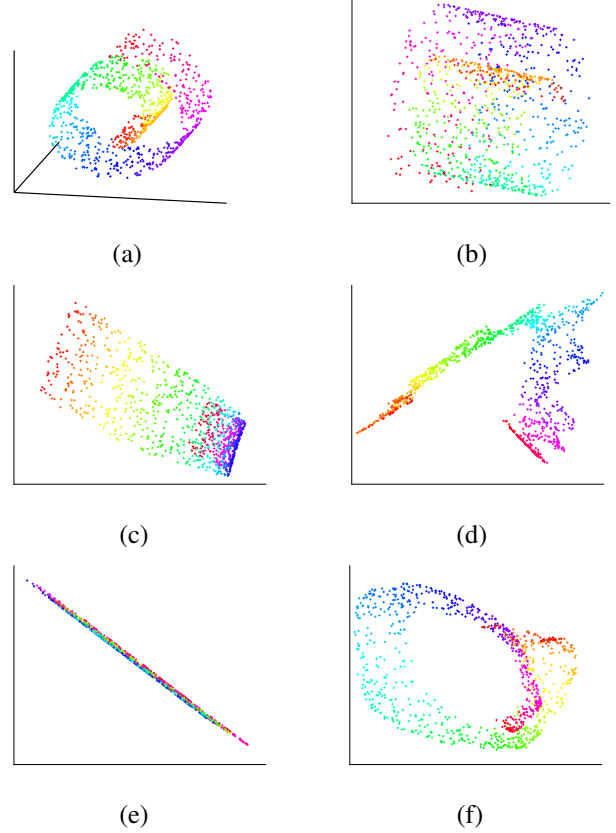

(a) (b)

(c) (d)

(e) (f)

Fig. 1: Embedding results of two manifolds: linear hyperplane and trefoil (a) visualization of $\{\mathbf{x}_i\}_{i=1}^{N}$; and $\{\mathbf{y}_i\}_{i=1}^{N}$ obtained from (b) PCA; (c) LLE with $K = 20$; (d) LNEG with $K = 20$; (e) LLE with $K = 40$; (f) LNEG with $K = 40$.

IV. NUMERICAL TESTS

The performance of the generalized version of LLE was assessed using tests on synthetic data. Alg. 1 is tested using $\mathbf{K}_x$ as in (7) for the locally nonlinear embedding (LNE) both without and with graph regularization (the latter abbreviated as LNEG), and is compared with LLE and PCA [10]. For all experiments, the graph $\mathcal{G}$ is constructed with adjacecy matrix $\mathbf{A}$s with (i, j)th entry $a_{ij} = \mathbf{x}_i^{\top}\mathbf{x}_j/\|\mathbf{x}_i\|\|\mathbf{x}_j\|$. Two types of tests are carried out in order to: a) evaluate embedding performance for a single manifold; and b) assess how informative the low-dimensional embeddings are for distinguishing different manifolds.

Embedding experiment. In the first experiment, the embedding performance of the proposed method is assessed. A 3-dimensional Swiss roll manifold is generated, and $1,000$ data are randomly sampled from the manifold as shown in Fig. 1 (a). Fig. 1 (b) showcases the 2-dimensional embeddings obtained from PCA, while Figs. 1 (c) and (d) illustrate the resulting embeddings from LLE and LNEG respectively, where neighborhoods of $K = 20$ data are considered. Figs. 1 (e) and (f) illustrate embeddings obtained by LLE and LNEG with $K = 40$. The regularization parameter of LNEG is set to $\gamma = 0.1$, and the polynomial order is set to $P = 2$. Clearly, by exploiting the nonlinear relationships
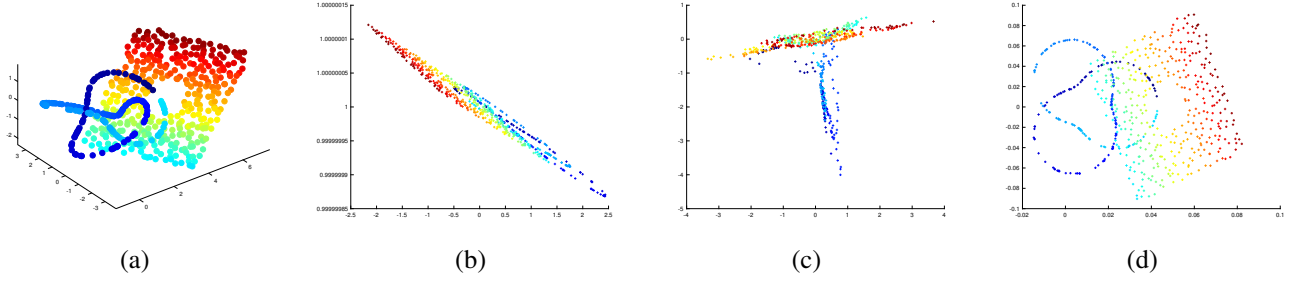
| (a) | (b) | (c) | (d) |

Fig. 2: Embedding results of two manifolds: a linear hyperplane with hole and a trefoil (a) visualization of two manifolds; and $\{\mathbf{y}_i\}_{i=1}^N$ obtained from (b) LLE with $K = 40$; and, (c) LNEG with $K = 40$ and $P = 2$; (d) PCA.
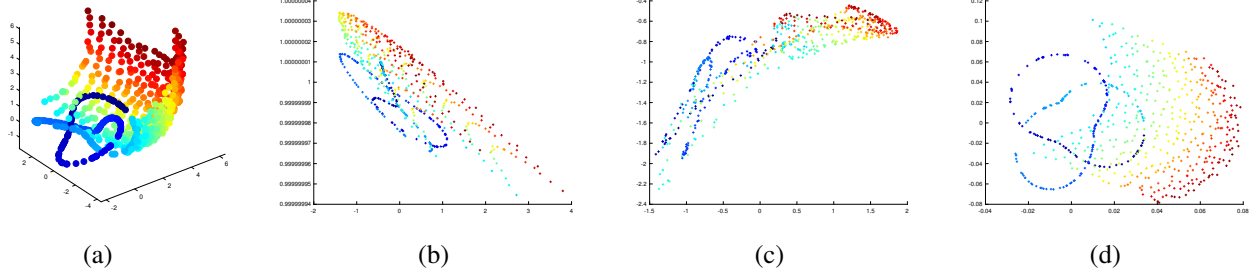

| (a) | (b) | (c) | (d) |

Fig. 3: Embedding results of two manifolds: a nonlinear sphere and a trefoil (a) visualization of two manifolds; and $\{\mathbf{y}_i\}_{i=1}^N$ obtained from (b) LLE with $K = 40$; (c) LNEG with $K = 40$ and $P = 3$; and (d) PCA.

	Plane-hole-trefoil			Sphere-trefoil		
K	LLE	LNE	LNEG	LLE	LNE	LNEG
5	0.25	0.18	0.18	0.29	0.21	0.17
10	0.44	0.21	0.18	0.39	0.27	0.16
20	0.15	0.13	0.17	0.48	0.26	0.14
30	0.21	0.20	0.17	0.46	0.28	0.19
40	0.36	0.20	0.17	0.39	0.21	0.20
PCA	0.49			0.43		

TABLE I: Clustering error rate on low-dimensional representations obtained from: LLE, LNE, LNEG and PCA.

between data, the resulting low-dimensional representations are capable of better preserving the structure of the manifold, thus allowing for more accurate visualization.

Clustering experiment. In this experiment, the ability of Alg. 1 to provide meaningful embeddings for clustering of different manifolds is assessed. Two 3-dimensional manifolds, a linear hyperplane with a hole around the origin and a trefoil are generated on the same ambient space as per [3], and 200 and 400 data are sampled from them, respectively. Here each manifold corresponds to a different cluster. Fig. 2(a) illustrates the sampled points from the generated manifolds. $\mathbf{Z}_1 \in \mathbb{R}^{3 \times 200}$ and $\mathbf{Z}_2 \in \mathbb{R}^{3 \times 400}$ contain the data generated from the linear hyperplane and the trefoil, respectively. Both manifolds are then linearly embedded in $\mathbb{R}^{100}$, that is $\mathbf{X}_i = \mathbf{P}\mathbf{Z}_i + \mathbf{E}_i$, where $\mathbf{P} \in \mathbb{R}^{100 \times 3}$ is an orthonormal matrix, and $\mathbf{E}$ is a noise matrix with entries sampled from a zero mean Gaussian distribution with variance 0.01. Afterwards, the 100-dimensional data in $\mathbf{X} := [\mathbf{X}_1 \ \mathbf{X}_2]$ are embedded into 2-dimensional representations $\mathbf{Y} \in \mathbb{R}^{2 \times 600}$ using LLE,

LNEG and PCA. Figures. 2(b), (c), and (d) depict the 2-dimensional embeddings $\mathbf{Y}$ provided by LLE, LNEG, and PCA, respectively. Similarly, Fig. 3 illustrates the resulting embeddings when $\mathbf{Z}_2$ is sampled from a nonlinear sphere. In both cases, the nonlinear methods result in embeddings that separate the two manifolds. To further assess the performance, K-means is carried out on the resulting $\mathbf{Y}$ [13]. Table I shows the clustering error when running K-means on the low-dimensional embeddings given by PCA, LLE, LNE and LNEG, across different values of K. The proposed approaches provide embeddings that enhance separability of the two manifolds, resulting in lower clustering error compared to LLE and PCA. In addition, greater performance gain is observed when both manifolds are nonlinear, as in the case of Fig. 3.

V. CONCLUSIONS

This paper introduced a general framework for nonlinear dimensionality reduction over graphs. By leveraging nonlinear relationships between data, low-dimensional representations were obtained to preserve these nonlinear correlations. Graph regularization was employed to account for additional prior knowledge when seeking the low-dimensional representations. An efficient algorithm that admits closed-form solution was developed, and several tests were conducted on simulated data to demonstrate its effectiveness. To broaden the scope of this study, several intriguing directions open up: a) extensive numerical tests on real data; b) development of data-dependent schemes that are capable of selecting appropriate kernels; c) online implementations that can handle streaming data; and d) generalizations to cope with large-scale graphs and high-dimensional datasets.

REFERENCES

[1] F. R. Bach, G. R. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proc.intl. Conf. on Machine learning*, New York, USA, 2004, pp. 6–13.

[2] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, Mar. 2003.

[3] E. Elhamifar and R. Vidal, "Sparse manifold clustering and embedding," in *Advances in Neural Information Processing Systems*, Granada, Spain, 2011, pp. 55–63.

[4] A. Ghodsi, "Dimensionality reduction -A short tutorial," *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada*, vol. 37, p. 38, 2006.

[5] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds," in *Proc. Intl. Conf. on Machine Learning*. Alberta, Canada: ACM, Jul. 2004, p. 47.

[6] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Journal of the Royal Statistical Society*, vol. 28, no. 1, pp. 100–108, Jan. 1979.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[8] B. Jiang, C. Ding, and J. Tang, "Graph-Laplacian PCA: Closed-form solution and robustness," in *Proc. Conf. on Computer Vision and Pattern Recognition*, Portland, Oregon, Jun. 2013, pp. 3492–3498.

[9] T. Jin, J. Yu, J. You, K. Zeng, C. Li, and Z. Yu, "Low-rank matrix factorization with multiple hypergraph regularizer," *Pattern Recognition*, vol. 48, no. 3, pp. 1011–1022, Mar. 2015.

[10] I. Jolliffe, *Principal Component Analysis*. Wiley Online Library, 2002.

[11] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of Complex Networks*, vol. 2, no. 3, pp. 203–271, 2014.

[12] J. B. Kruskal and M. Wish, *Multidimensional Scaling*. Sage, 1978, vol. 11.

[13] S. Lloyd, "Least-squares quantization in PCM," *IEEE Trans. Info. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[14] D. Romero, V. N. Ioannidis, and G. B. Giannakis, "Kernel-based Reconstruction and Kalman Filtering of Space-time Functions on Dynamic Graphs," *IEEE Journal on Special Topics in Signal Processing*, vol. 11, no. 6, 2017.

[15] D. Romero, M. Ma, and G. B. Giannakis, "Kernel-based reconstruction of graph signals," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 764–778, Feb. 2017.

[16] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[17] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Intl. Conf. on Artificial Neural Networks*, Lausanne, Switzerland, Oct. 1997, pp. 583–588.

[18] N. Shahid, N. Perraudin, V. Kalofolias, G. Puy, and P. Vandergheynst, "Fast robust PCA on graphs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 740–756, Feb. 2016.

[19] F. Shang, L. Jiao, and F. Wang, "Graph dual regularization non-negative matrix factorization for co-clustering," *Pattern Recognition*, vol. 45, no. 6, pp. 2237–2250, Jun. 2012.

[20] Y. Shen, B. Baingana, and G. B. Giannakis, "Kernel-based structural equation models for topology identification of directed networks," *IEEE Trans. Sig. Proc.*, vol. 65, no. 10, pp. 2503–2516, May 2017.

[21] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[22] P. A. Traganitis, Y. Shen, and G. B. Giannakis, "Topology inference of multilayer networks," in *Intl. Workshop on Network Science for Comms.*, Atlanta, GA, May 2017.