Multiple Kernel Learning for Spectral Dimensionality Reduction

Diego Hernán Peluffo-Ordóñez[1], Andrés Eduardo Castro-Ospina[2(✉)], Juan Carlos Alvarado-Pérez[3,4], and Edgardo Javier Revelo-Fuelagán[5]

[1] Universidad Cooperativa de Colombia – Pasto, Pasto, Colombia
[2] Research Center of the Instituto Tecnologico Metropolitano, Medellin, Colombia
andrescastro@itm.edu.co
[3] Universidad de Salamanca, Salamanca, Spain
[4] Universidad Mariana, Pasto, Colombia
[5] Universidad de Nariño, Pasto, Colombia

Abstract. This work introduces a multiple kernel learning (MKL) approach for selecting and combining different spectral methods of dimensionality reduction (DR). From a predefined set of kernels representing conventional spectral DR methods, a generalized kernel is calculated by means of a linear combination of kernel matrices. Coefficients are estimated via a variable ranking aimed at quantifying how much each variable contributes to optimize a variance preservation criterion. All considered kernels are tested within a kernel PCA framework. The experiments are carried out over well-known real and artificial data sets. The performance of compared DR approaches is quantified by a scaled version of the average agreement rate between K-ary neighborhoods. Proposed MKL approach exploits the representation ability of every single method to reach a better embedded data for both getting more intelligible visualization and preserving the structure of data.

Keywords: Dimensionality reduction · Generalized kernel · Kernel PCA · Multiple kernel learning

1 Introduction

The aim of dimensionality reduction (DR) is to extract a lower dimensional, relevant information from high-dimensional data, being then a key stage within the design of pattern recognition and data mining systems. Indeed, when using adequate DR stages, the system performance can be enhanced as well as the data visualization can become more intelligible. The range of DR methods is diverse, including those classical approaches such as principal component analysis (PCA) and classical multidimensional scaling (CMDS), which are respectively based on variance and distance preservation criteria [1]. Recent methods of DR are focused

This work is supported by the Faculty of Engineering of Universidad Cooperativa de Colombia-Pasto, and the ESLINGA Research Group.

A. Pardo and J. Kittler (Eds.): CIARP 2015, LNCS 9423, pp. 626–634, 2015.
DOI: 10.1007/978-3-319-25751-8_75

on the data topology preservation. Mostly such a topology is driven by graph-based approaches where data are represented by a non-directed and weighted graph. In this connection, the weights of edge graphs are certain pairwise similarities between data points, the nodes are data points, and a non-negative similarity (also affinity) matrix holds the pairwise edge weights. Spectral methods such as Laplacian eigenmaps (LE) [2] and locally linear embedding (LLE) [3] were the pioneer ones to incorporate similarity-based formulations. Also, given the fact that the rows of the normalized similarity matrix can be seen as probability distributions, divergence-based methods have emerged (i.e., stochastic neighbor embedding (SNE) [4]). Spectral approaches for DR have been widely used in several applications such as relevance analysis [5,6], dynamic data analysis [7,8] and feature extraction [9,10]. Because of being graph-driven methods and involving then similarities, spectral approaches can be easily represented by kernels [11], which means that a wide range of methods can be set within a Kernel PCA framework [12]. At the moment to choose a method, aspects such as nature of data, complexity, aim to be reached and problem to be solved should be taken into consideration. In this regard, as mentioned above, there exists a variety of DR spectral methods making the selection of a method a nontrivial task. Also, some problems may require the combination of methods so that the properties of different methods are simultaneously taken into account to perform the DR process and the quality of resultant embedded space is improved.

The purpose of this work is to provide a multiple kernel learning (MKL) approach allowing for both selecting a DR method, and combining different methods to exploit the representation ability of every single method to reach a better embedded space than the one obtained when using only one method. This approach starts with kernel representations of conventional spectral methods as explained in [11]. Then, a generalized kernel is calculated by means of a linear combination of kernel matrices whose coefficients are estimated by an adapted variable relevance approach proposed in a previous work [6]. Similar approaches have been applied on dynamic data clustering [13] and image segmentation [14]. The experiments are carried out over well-known data sets, namely an artificial `Spherical shell`, a `Swiss roll` toy set, and `MNIST` image bank [15]. The DR performance is quantified by a scaled version of the average agreement rate between K-ary neighborhoods as described in [16].

The rest of this paper is organized as follows: Section 2 outlines the proposed MKL approach for dimensionality reduction. Section 3 describes the experimental setup as well as section 4 presents the results and discussion. Finally, some final remarks are drawn in section 5.

2 Multiple kernel Learning for Dimensionality Reduction

In mathematical terms, the goal of DR is to embed a high dimensional data matrix $\boldsymbol{Y} \in \mathbb{R}^{D \times N}$ into a low-dimensional, latent data matrix $\boldsymbol{X} \in \mathbb{R}^{d \times N}$, being $d < D$. Then, observed data and latent data matrices are formed by N data points, denoted respectively by $\boldsymbol{y}_i \in \mathbb{R}^D$ and $\boldsymbol{x}_i \in \mathbb{R}^d$, with $i \in \{1, \ldots, N\}$.

Kernel PCA, as PCA, maximizes a variance criterion, which can be seen as an inner product criterion when data matrix is centered. Let $\boldsymbol{\Phi} \in \mathbb{R}^{D_h \times N}$ be an unknown high dimensional representation space such that $D_h \ggg D$, and $\phi(\cdot)$ be a function that maps data from the original dimension to a higher one, such that $\phi(\cdot) : \mathbb{R}^D \to \mathbb{R}^{D_h}, \boldsymbol{y}_i \mapsto \phi(\boldsymbol{y}_i)$.

Given this, we can write the i-th column vector of matrix $\boldsymbol{\Phi}$ as $\boldsymbol{\Phi}_i = \phi(\boldsymbol{y}_i)$. Consequently, the inner product on the high-dimensional vector space is $\phi(\boldsymbol{y}_i)^\top \phi(\boldsymbol{y}_j) = k(\boldsymbol{y}_i, \boldsymbol{y}_j) = k_{ij}$, where $k(\cdot, \cdot)$, followed from Mercer's condition, is a kernel function. In matrix terms, we get that the kernel matrix is $\boldsymbol{K} = \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$.

Since Kernel PCA is developed under the condition that matrix $\boldsymbol{\Phi}$ has zero mean, we must ensure this condition by centering the kernel matrix as follows:

$$\begin{aligned} \boldsymbol{K} \leftarrow & \boldsymbol{K} - \frac{1}{N}\boldsymbol{K}\mathbf{1}_N\mathbf{1}_N^\top - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{K} + \frac{1}{N^2}\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{K}\mathbf{1}_N\mathbf{1}_N^\top \\ & = (\boldsymbol{I}_N - \mathbf{1}_N\mathbf{1}_N^\top)\boldsymbol{K}(\boldsymbol{I}_N - \mathbf{1}_N\mathbf{1}_N^\top), \end{aligned} \tag{1}$$

where $\mathbf{1}_N$ and $\boldsymbol{I}_N$ are N-dimensional all ones vector and identity matrix, respectively.

The aim of our MKL approach is to get a generalized kernel $\widetilde{\boldsymbol{K}} \in \mathbb{R}^{N \times N}$ from a linear combination of a set of kernels $\{\boldsymbol{K}^{(1)}, \ldots, \boldsymbol{K}^{(M)}\}$ to input a DR approach based on kernels. Ensuring linear independency, the generalized kernel can be written as:

$$\widetilde{\boldsymbol{K}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{K}^{(m)}. \tag{2}$$

Here, we propose to estimate the coefficients by using an adapted version of the variable ranking approach proposed in [6]. In [13], authors apply MKL based on a ranking vector to cluster time-varying data in a sequence of frames. A cumulative kernel is calculated to track the dynamic behavior, having each kernel a corresponding data matrix (one per frame). Unlike, in this approach we have a single data matrix, and then the ranking vector should be calculated using directly the kernel matrices. Define a matrix $\boldsymbol{\mathcal{K}} \in \mathbb{R}^{N^2 \times M}$ holding the vectorization of the kernel matrices. Likewise, suppose that a lower-rank representation $\widehat{\boldsymbol{\mathcal{K}}} \in \mathbb{R}^{N^2 \times M}$ of matrix $\widehat{\boldsymbol{\mathcal{K}}}$ is known. Regarding any othonormal matrix $\boldsymbol{U} = [\boldsymbol{u}^{(1)} \cdots \boldsymbol{u}^{(c)}] \in \mathbb{R}^{M \times c}$, we can write the lower-rank matrix as

$$\widehat{\boldsymbol{\mathcal{K}}} = \boldsymbol{\mathcal{K}}\boldsymbol{U}. \tag{3}$$

So, the full-rank matrix can be then estimated as $\boldsymbol{\mathcal{K}} = \widehat{\boldsymbol{\mathcal{K}}}\boldsymbol{U}^\top$. Similarly to the feature extraction problem stated in [5,9], here we propose to maximize the variance of $\widehat{\boldsymbol{\mathcal{K}}}$ by solving the following optimization problem:

$$\max_{\boldsymbol{U}} \ \operatorname{tr}(\boldsymbol{U}^\top \boldsymbol{\mathcal{K}}\boldsymbol{\mathcal{K}}\boldsymbol{U}) \tag{4a}$$

$$\text{s. t.} \quad \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_c. \tag{4b}$$

As demonstrated in [6], previous problem has a dual version that can be expressed as

$$\min_{\boldsymbol{U}} \|\boldsymbol{\mathcal{K}} - \widehat{\boldsymbol{\mathcal{K}}}\|_F^2 \tag{5a}$$

$$\text{s. t.} \quad \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_c, \tag{5b}$$

where $\|\cdot\|_F$ stands for Frobenius norm. Since this formulation is a quadratic problem subject to orthonormal constraints, a feasible solutions is selecting $\boldsymbol{U}$ as the eigenvectors related to the c largets eigenvalues of $\boldsymbol{\mathcal{K}}\boldsymbol{\mathcal{K}}$.

Finally, the coefficients α_m of the linear combination to calculate the generalized kernel are the ranking values quantifying how much each column of matrix $\boldsymbol{\mathcal{K}}$ (each kernel) contributes to minimizing the cost function given in (5a). Again, applying the variable relevance approach presented in [6], we can calculate the ranking vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]$ using:

$$\boldsymbol{\alpha} = \sum_{m=1}^{c} \lambda_m \boldsymbol{u}^{(m)} \circ \boldsymbol{\alpha}^{(m)}, \tag{6}$$

where $\circ$ denotes Hadamard (element-wise) product. Given the problem formulation, possitivenes of $\boldsymbol{\alpha}$ is guaranteed and then can be directly used to perform the linear combination.

3 Experimental Setup

Databases. Experiments are carried out over three conventional data sets. The first data set is an artificial spherical shell ($N = 1500$ data points and $D = 3$). The second data set is a randomly selected subset of the `MNIST` image bank [15], which is formed by 6000 gray-level images of each of the 10 digits ($N = 1500$ data points –150 instances for all 10 digits– and $D = 24^2$). The third data set is a toy set here called `Swiss roll` ($N = 3000$ data points and $D = 3$). Figure 1 depicts examples of the considered data sets.

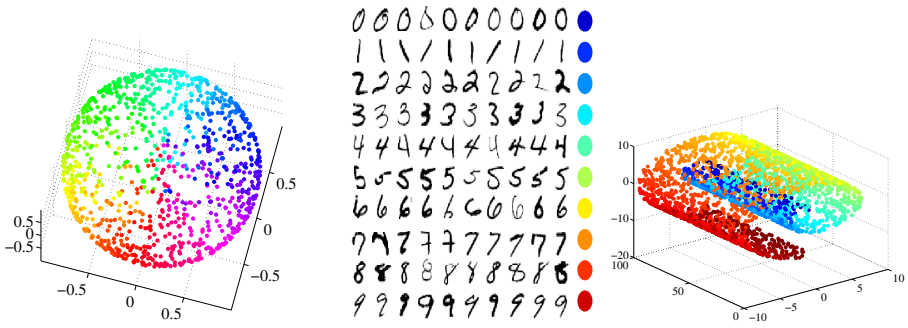
Kernels for DR. Three kernel approximations for spectral DR methods [11] are considered. Namely, classical multidimensional scalling (CMDS), locally linear embedding (LLE), and graph Laplacian eigenmaps (LE). CMDS kernel is the double centered distance matrix $\boldsymbol{D} \in \mathbb{R}^{N \times N}$ so

$$\boldsymbol{K}^{(1)} = \boldsymbol{K}_{CMDS} = -\frac{1}{2}(\boldsymbol{I}_N - \boldsymbol{1}_N \boldsymbol{1}_N^\top)\boldsymbol{D}(\boldsymbol{I}_N - \boldsymbol{1}_N \boldsymbol{1}_N^\top), \tag{7}$$

where the ij entry of $\boldsymbol{D}$ is given by $d_{ij} = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2$ and $||\cdot||_2$ stands for Euclidean norm.

A kernel for LLE can be approximated from a quadratic form in terms of the matrix $\boldsymbol{\mathcal{W}}$ holding linear coefficients that sum to 1 and optimally reconstruct observed data. Define a matrix $\boldsymbol{M} \in \mathbb{R}^{N \times N}$ as $\boldsymbol{M} = (\boldsymbol{I}_N - \boldsymbol{\mathcal{W}})(\boldsymbol{I}_N - \boldsymbol{\mathcal{W}}^\top)$ and λ_{max} as the largest eigenvalue of $\boldsymbol{M}$. Kernel matrix for LLE is in the form

$$\boldsymbol{K}^{(2)} = \boldsymbol{K}_{LLE} = \lambda_{max}\boldsymbol{I}_N - \boldsymbol{M}. \tag{8}$$



(a) *Sphere* (b) *MNIST* (c) *Swiss Roll*

Fig. 1. The three considered datasets.

Since kernel PCA is a maximization problem of the covariance of the the high-dimensional data represented by a kernel, LE can be expressed as the pseudo-inverse of the graph Laplacian $\boldsymbol{L}$:

$$\boldsymbol{K}^{(3)} = \boldsymbol{K}_{LE} = \boldsymbol{L}^{\dagger}, \tag{9}$$

where $\boldsymbol{L} = \boldsymbol{\mathcal{D}} - \boldsymbol{S}$, $\boldsymbol{S}$ is a similarity matrix and $\boldsymbol{\mathcal{D}} = \text{Diag}(\boldsymbol{S}\boldsymbol{1}_N)$ is the degree matrix. All previously mentioned kernels are widely described in [11]. The similarity matrix $\boldsymbol{S}$ is formed in such a way that the relative bandwidth parameter is estimated keeping the entropy over neighbor distribution as roughly $\log K$ where K is the given number of neighbors as explained in [17]. The number of neighbors is established as $K = 30$.

As well, a RBF kernel is also considered: $\boldsymbol{K}^{(4)} = \boldsymbol{K}_{RBF}$ whose ij entry are given by $\exp(-0.5||\boldsymbol{y}_i - \boldsymbol{y}_j||/\sigma^2)$ with $\sigma = 0.1$. For all methods, input data is embedded into a 2-dimensional space, then $d = 2$.

Accordingly, the MKL approach is performed considering $M = 4$ kernels. The generalized kernel provided $\widetilde{\boldsymbol{\mathcal{K}}}$ here as well as the individual kernels $\boldsymbol{K}^{(1)}, \ldots, \boldsymbol{K}^{(M)}$ are tested on kernel PCA as explained in [12].

Performance Measure: To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [16] is used, which is ranged within the interval $[0, 1]$. Since $R_{NX}(K)$ is calculated at each perplexity value from 2 to $N-1$, a numerical indicator of the overall performance can be obtained by calculating its area under the curve (AUC). The AUC assesses the dimension reduction quality at all scales, with the most appropriate weights.

Notwithstanding, it is important to note that kernels approximations are suboptimal and input parameters are not properly set, which means that under other settings, the quality measure and resultant embedding data might be significantly different. Here, just basic settings are considered in order to show the benefit of MKL rather than the individual methods.

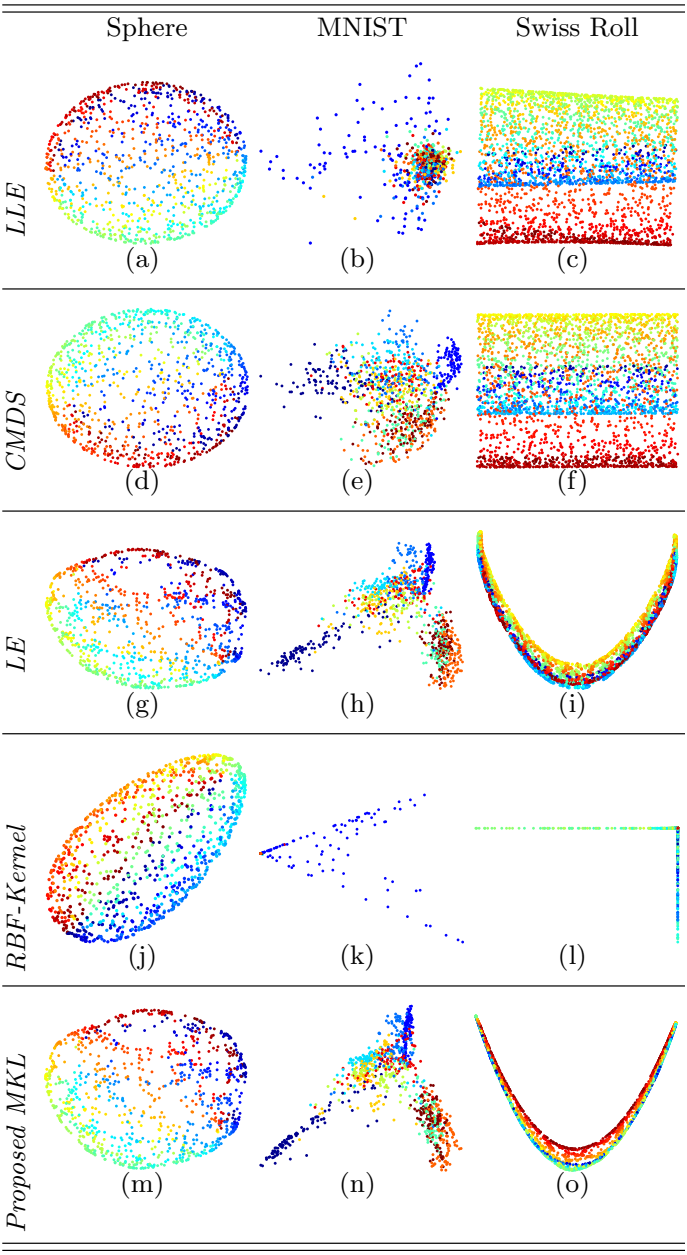
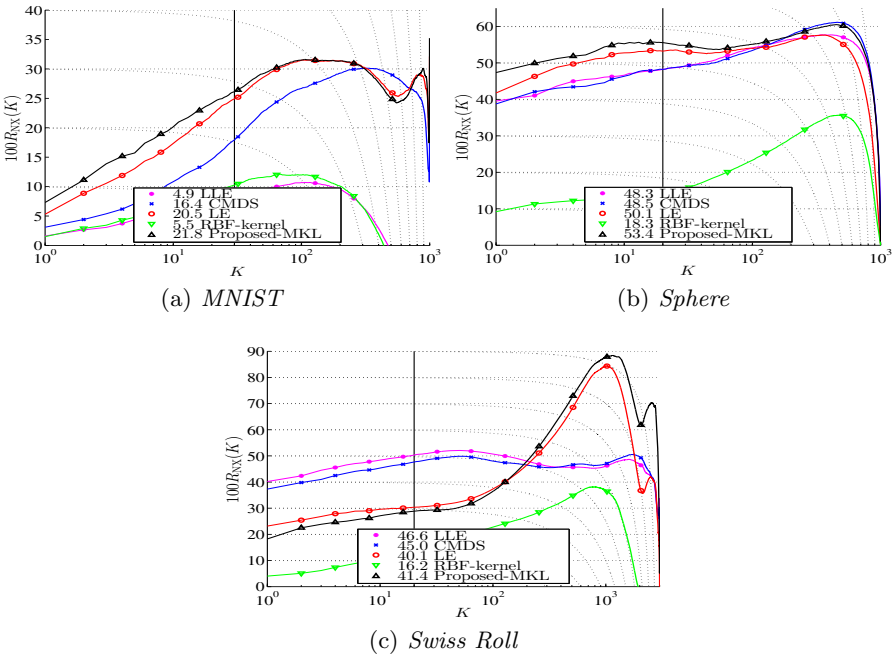


Fig. 2. 2D representations for selected methods over all data sets.



(a) *MNIST* (b) *Sphere*

(c) *Swiss Roll*

Fig. 3. Performances for the three considered datasets.

4 Results and Discussion

Figure 2 shows the resultant embedding data using the corresponding kernels of the studied methods, and the proposed generalized kernel for MKL. Comparing resultant embedding representations with the $R_{NX}(K)$ curves shown in Figure 3, we can appreciate that proposed MKL approach determines the best one among the considered methods, since embedding data reached by MKL resemble to the one of the best method. In this case, the best method is LE, which gets more intelligible representation since either underling clusters are better formed (see Figure 2(h)), or the manifold is better represented -resembling an object unfolding (see Figures 2(g) and 2(i)).

Additionally, the generalized kernel used in a kernel PCA may improve the quality of representation as can be appreciated from Figure 3. Indeed, the area under the curve reached by our MKL is the highest for two of the tested data sets. Particularly, for *Swiss roll* data set, our approach gets higher AUC than the baseline LE but is not the highest one. Nonetheless, differently the other considered methods, the R_{NX} curve of proposed MKL approach has a right-sided asymmetric plotting, which means that our approach is focused on specific structures of data -in this case, the global structure.

5 Conclusions and Future Work

In this work, a multiple kernel learning approach for dimensionality reduction tasks is presented. The core of this approach is the generalized kernel that is calculated by means of a linear combination of kernel matrices representing spectral dimensionality reduction methods, where the coefficients are obtained from a variable ranking based on a variance criterion. Proposed approach improves both data visualization and preservation by exploiting the representation ability of every single technique.

As future work, new multiple kernel learning approaches will be explored by combining kernel representations arising from other dimensionality reduction methods, aimed at reaching a good trade-off between preservation of data structure and intelligible data visualization.

References

1. Borg, I.: Modern multidimensional scaling: Theory and applications. Springer (2005)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15**(6), 1373–1396 (2003)
3. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
4. Hinton, G.E., Roweis, S.T.: Stochastic neighbor embedding. In: Advances in Neural Information Processing Systems, pp. 833–840 (2002)
5. Wolf, L., Bileschi, S.: Combining variable selection with dimensionality reduction. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 801–806, June 2005
6. Peluffo, D.H., Lee, J.A., Verleysen, M., Rodríguez-Sotelo, J.L., Castellanos-Domínguez, G.: Unsupervised relevance analysis for feature extraction and selection: a distance-based approach for feature relevance. In: International Conference on Pattern Recognition, Applications and Methods - ICPRAM 2014
7. Langone, R., Alzate, C., Suykens, J.A.: Kernel spectral clustering with memory effect. Statistical Mechanics and its Applications, Physica A (2013)
8. Maestri, M., Cassanello, M., Horowitz, G.: Kernel PCA performance in processes with multiple operation modes. Chemical Product and Process Modeling **4**(5), 7 (2009)
9. Wolf, L., Shashua, A.: Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. Journal of Machine Learning **6**, 1855–1887 (2005)
10. Unsupervised feature relevance analysis applied to improve ECG heartbeat clustering. Computer Methods and Programs in Biomedicine (2012)
11. Ham, J., Lee, D.D., Mika, S., Schölkopf, B.: A kernel view of the dimensionality reduction of manifolds. In: Proceedings of the Twenty-first International Conference on Machine learning, p. 47. ACM (2004)
12. Peluffo-Ordóñez, D., Lee, J., Verleysen, M.: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: IEEE Symposium Series on Computational Intelligence (2014)

13. Peluffo-Ordónez, D., Garcia-Vega, S., Langone, R., Suykens, J., Castellanos-Dominguez, G., et al.: Kernel spectral clustering for dynamic data using multiple kernel learning. In: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2013)
14. Molina-Giraldo, S., Álvarez-Meza, A.M., Peluffo-Ordoñez, D.H., Castellanos-Domínguez, G.: Image segmentation based on multi-kernel learning and feature relevance analysis. In: Pavón, J., Duque-Méndez, N.D., Fuentes-Fernández, R. (eds.) IBERAMIA 2012. LNCS, vol. 7637, pp. 501–510. Springer, Heidelberg (2012)
15. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
16. Lee, J.A., Renard, E., Bernard, G., Dupont, P., Verleysen, M.: Type 1 and 2 mixtures of Kullback-Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing (2013)
17. Cook, J., Sutskever, I., Mnih, A., Hinton, G.E.: Visualizing similarity data with a mixture of maps. In: International Conference on Artificial Intelligence and Statistics, pp. 67–74 (2007)