Highly Sparse Reductions to Kernel Spectral Clustering

Raghvendra Mall, Rocco Langone, and Johan A.K. Suykens

Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven,
Kasteelpark Arenberg, 10 B-3001 Leuven, Belgium
{raghvendra.mall,rocco.langone,johan.suykens}@esat.kuleuven.be

Abstract. Kernel spectral clustering is a model-based spectral clustering method formulated in a primal-dual framework. It has a powerful out-of-sample extension property and a model selection procedure based on the balanced line fit criterion. This paper is an improvement of a previous work which sparsified the kernel spectral clustering method using the line structure of the data projections in the eigenspace. However, the previous method works only in the case of well formed and well separated clusters as in other cases the line structure is lost. In this paper, we propose two highly sparse extensions of kernel spectral clustering that can overcome these limitations. For the selection of the reduced set we use the concept of angles between the data projections in the eigenspace. We show the effectiveness and the amount of sparsity obtained by the proposed methods for several synthetic and real world datasets.

1 Introduction

Clustering algorithms are widely used tools in fields like data mining, machine learning, graph compression and many other tasks. The aim of clustering is to divide data into natural groups present in a given dataset. Clusters are defined such that the data present within the group are more similar to each other in comparison to the data between clusters. Spectral clustering methods [1,2,3] are generally better than the traditional k-means techniques. A new spectral clustering algorithm based on weighted kernel principal component analysis (PCA) formulation was proposed in [4]. The method is based on a model built in a primal-dual optimization framework. The model has a powerful out-of-sample extension property which allows to infer cluster affiliation for unseen data.

The data points are projected to the eigenspace and the projections are expressed in terms of non-sparse kernel expansions. In [5] sparsification of this clustering model was done by exploiting the line structure of the projections when the clusters are well formed and well separated. However, *the method fails when the clusters are overlapping and for real world datasets where the projections in the eigenspace do not follow a line structure as mentioned in [8]. In this paper, we propose methods to handle these issues.* We locate the mean of each cluster in the eigenspace. The mean of the cluster in the eigenspace is located on the least squares linear regressor for all the points in that cluster. We use angular

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 163–169, 2013.

distance to locate the projections close to the mean in the eigenspace and then select these projections based on their euclidean distance from the origin in the eigenspace. The main advantage of these sparse reductions is that it results in much simpler and faster predictive models.

2 Kernel Spectral Clustering

We provide a brief description of the kernel spectral clustering methodology. Given a dataset $\mathcal{D} = \{x_i\}_{i=1}^{N_{tr}}, x_i \in \mathbb{R}^d$, the training points are selected by maximizing the quadratic Rènyi criterion as depicted in [8,9,10]. Here x_i represents the i^{th} training point and the training set is represented by X_{tr}. The number of data points in the training set is N_{tr}. Given $\mathcal{D}$ and the number of clusters k, the primal problem of the spectral clustering via weighted kernel PCA is formulated as follows [4]:

$$\begin{aligned} \min_{w^{(l)},e^{(l)},b_l} \quad & \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)\intercal} w^{(l)} - \frac{1}{2N_{tr}}\sum_{l=1}^{k-1}\gamma_l e^{(l)\intercal} D_\Omega^{-1} e^{(l)} \\ \text{such that} \quad & e^{(l)} = \Phi w^{(l)} + b_l 1_{N_{tr}}, l = 1,\ldots,k-1, \end{aligned} \tag{1}$$

where $e^{(l)} = [e_1^{(l)},\ldots,e_{N_{tr}}^{(l)}]^\intercal$ are the projections onto the eigenspace, $l = 1,\ldots,k-1$ indicates the number of score variables required to encode the k clusters, $D_\Omega^{-1} \in \mathbb{R}^{N_{tr}\times N_{tr}}$ is the inverse of the degree matrix associated to the kernel matrix Ω. Φ is the $N_{tr} \times n_h$ feature matrix, $\Phi = [\phi(x_1)^\intercal;\ldots;\phi(x_{N_{tr}})^\intercal]$ and $\gamma_l \in \mathbb{R}^+$ are the regularization constants. We note that $N_{tr} < N$ i.e. the number of points in the training set is less than the total number of points in the dataset. The kernel matrix Ω is obtained by calculating the similarity between each pair of data point in the training set. Each element of Ω, denoted as $\Omega_{ij} = K(x_i, x_j) = \phi(x_i)^\intercal\phi(x_j)$ is obtained for example by using the radial basis function (RBF) kernel. The clustering model is then represented by:

$$e_i^{(l)} = w^{(l)\intercal}\phi(x_i) + b_l, i = 1,\ldots,N_{tr}, \tag{2}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the mapping to a high-dimensional feature space n_h, b_l are the bias terms, $l = 1,\ldots,k-1$. The projections $e_i^{(l)}$ represent the latent variables of a set of $k-1$ binary cluster indicators given by $\text{sign}(e_i^{(l)})$ which can be combined with the final groups using an encoding/decoding scheme. The decoding consists of comparing the binarized projections w.r.t. codewords in the codebook and assigning cluster membership based on minimal Hamming distance. The dual problem corresponding to this primal formulation is:

$$D_\Omega^{-1} M_D \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}, \tag{3}$$

where M_D is the centering matrix which is defined as $M_D = \mathrm{I}_{N_{tr}} - (\frac{(1_{N_{tr}} 1_{N_{tr}}^\intercal D_\Omega^{-1})}{1_{N_{tr}}^\intercal D_\Omega^{-1} 1_{N_{tr}}})$. The $\alpha^{(l)}$ are the dual variables and the positive definite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ plays the role of similarity function. The corresponding predictive model

is $\hat{e}^{(l)}(x) = \sum_{i=1}^{N_{tr}} \alpha_i^{(l)} K(x, x_i) + b_l$ which provides clustering inference for unseen data points x. Thus the model has an out-of-sample extension property. For selection of the hyper-parameters of the model i.e. k and σ for RBF kernel we use the Balanced Angular Fit (BAF) criterion proposed in [6,7].

3 Sparse Reductions to the KSC Model

3.1 Related Work

The projections of the data points in the eigenspace are expressed in terms of non-sparse kernel expansions. The primal vectors $w^{(l)} = \sum_{i=1}^{N_{tr}} \alpha_i^{(l)} \phi(x_i)$ can be approximated by a reduced set. The objective is to approximate $w^{(l)}$ by a new weight vector $\tilde{w}^{(l)} = \sum_{i=1}^{R} \beta_i^{(l)} \phi(\tilde{x}_i)$ minimizing $||w^{(l)} - \tilde{w}^{(l)}||_2^2$ where $\tilde{x}_i$ is the i^{th} point in the reduced set $\mathcal{RS}$ whose cardinality is R. In [5], it was shown that if the reduced set $\mathcal{RS}$ is known then the $\beta^{(l)}$ co-efficients can be obtained by solving the linear system:

$$\Omega^{\psi\psi} \beta^{(l)} = \Omega^{\psi\phi} \alpha^{(l)}, \tag{4}$$

where $\Omega_{mn}^{\psi\psi} = K(\tilde{x}_m, \tilde{x}_n)$, $\Omega_{mi}^{\psi\phi} = K(\tilde{x}_m, x_i)$, $m, n = 1, \ldots, R, i = 1, \ldots, N_{tr}$ and $l = 1, \ldots, k-1$.

This reduced set can be built by selecting points whose projections in the eigenspace occupy certain positions or by using an elastic net penalization. Two methods to attain the same was proposed in [5,8]. The amount of sparsity attained by introducing penalization is not as much as that obtained by selection of points based on their position. The method based on selecting points from certain positions as proposed in [5] works when the clusters are well formed and separated. This is because in that condition we obtain a line structure corresponding to the projections of the data points in the eigenspace.

3.2 Proposed Methods

We propose two methods for selection of points to form the reduced set $\mathcal{RS}$. In case when the clusters are not well formed and overlapping, the projections of the corresponding data points in the eigenspace loses the line structure. We estimate the mean of all the projections for a particular cluster. According to the properties of least squares linear regressor [11], the linear regressor fit for all the projections belonging to that cluster in the eigenspace passes through the mean of the projections. However, it might so happen that in the input space there is no actual data point corresponding to that mean. So, in order to select points existing in the input space, we use the concept of angular similarity.

We select all the projections from a cluster whose cosine distance w.r.t. mean for that cluster (e_{μ_i}) is less than threshold t i.e. $1 - \cos(e_j, e_{\mu_i}) < t$. Here $i = 1, \ldots, k-1$ and $j = 1, \ldots, N_{C_i}$, e_j is a projection and N_{C_i} is the number of points in the i^{th} cluster. We initially set $t = 0$ and increase it using an iterative procedure. During each iteration, we increase the value of the t by 0.1 until we have non-empty selection set corresponding to that cluster. We observe in our experiments that one iteration is enough for most of the datasets.

First Proposed Method - Once the selection set is obtained, we calculate the euclidean distance of these projections from the origin. For the first method of selection, we select the projection which is the farthest and at median distance from the origin. The projection which is the farthest from the origin generally corresponds to a point which is close to the center of the cluster in the input space. This is because the projection value say $e_c^{(l)} = \sum_{i=1}^{N_{tr}} \alpha_i^{(l)} K(x_i, x_c) + b_l$ is dependent on the $\alpha_i^{(l)}$ and the kernel evaluation of $K(x_i, x_c)$. According to the nearly piecewise structure of the eigenvectors, all the points belonging to that cluster have nearly similar value of $\alpha_i^{(l)}$. The cluster center in the input space has maximum similarity (≈ 1) to points in that cluster and minimum similarity (≈ 0) to the points in other clusters. Thus, since the effect of the $\alpha_i^{(l)}$ is nearly constant for all the points belonging to a cluster, it can be concluded that the point corresponding to the cluster center in the input space is the projection whose euclidean distance is farthest from the origin in the eigenspace . The median point is selected to provide more stability to the reduced set $\mathcal{RS}$. Thus, if there are k clusters in a dataset, the number of points required to build the reduced set is $2k$ and the sparsity is given as:

$$Sparsity = 1 - \frac{2k}{N_{tr}} \tag{5}$$

Second Proposed Method - For the second method, we select 10% of the projections from the selection set obtained as a result of cosine distance for each cluster. We keep the vector containing the angular distance between these projections and the corresponding mean in a sorted order. Let the size of this vector be S_i. Then the size of the reduced set for each cluster is $r_i = \lceil \frac{S_i}{10} \rceil$, $i = 1, \ldots, k$. We divide this vector into $\lfloor \frac{S_i}{r_i} \rfloor$ bins for each cluster and select a projection from each bin. The minimum value that sparsity can take for this method is $1 - \frac{N_{tr}}{10 \times N_{tr}}$. This is when all the projections corresponding to each cluster maintain a line structure. This is because the cosine similarity of all the projections w.r.t. mean is maximum. However, in other cases when the projections deviate from the line structure then the amount of sparsity introduced is:

$$Sparsity = 1 - \sum_{i=1}^{k} \frac{1}{N_{tr}} \frac{S_i}{r_i}. \tag{6}$$

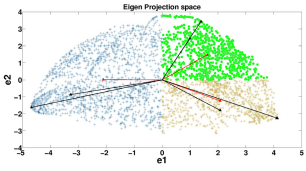
4 Experimental Results

Table 1 provides information about internal quality metrics namely silhouette criterion (*sil*), davies-bouldin index (*db*) and *Sparsity* (*S*) in percentage for the 2 proposed methods. The benchmark datasets are obtained from `http://cs.joensuu.fi/sipu/datasets/`. We compare with method proposed in [5] and the $L_2 + L_1$ penalization based method in [8]. Higher values of *sil* are better and lower values of *db* represents better clustering quality. We highlight the best results in Table 1 for the 9 real world datasets.

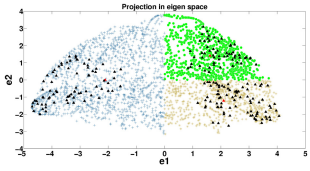
Table 1. Experimental Results on various benchmark datasets

Dataset	1^{st} Highly Sparse Method			2^{nd} Highly Sparse Method			Method in [5]			$L_2 + L_1$ Penalization		
	sil	*db*	*S* (%)	*sil*	*db*	*S* (%)	*sil*	*db*	*S* (%)	*sil*	*db*	*S* (%)
Breast	**0.665**	**0.928**	**98.0**	0.645	0.955	90	0.612	0.975	97.00	0.639	0.933	96.65
Bridge	0.2787	2.05	**99.5**	**0.28**	2.009	91.7	0.265	2.15	99.26	0.275	**1.72**	98.57
Glass	**0.433**	1.85	78	0.35	**1.71**	90	0.32	1.99	69.15	0.41	1.81	**96.88**
Iris	**0.3975**	**0.87**	86.7	0.343	1.12	**91**	0.31	1.25	80.0	0.309	1.306	86.77
MLF	0.74	**1.07**	**99.7**	0.701	1.127	96	0.70	1.15	99.5	**0.795**	1.158	99.55
MLJ	0.823	**0.448**	**99.5**	0.82	0.453	94.8	0.801	0.67	99.2	**0.881**	0.67	95.51
Thyroid	0.499	1.198	**93.75**	**0.512**	**1.183**	90	0.492	1.73	91.0	0.51	1.24	93.75
Wdbc	**0.565**	**1.28**	**97.6**	0.5535	1.28	90	0.54	1.30	95.12	0.56	1.303	**97.6**
Wine	0.282	1.86	88.7	**0.3**	**1.86**	**92.5**	0.273	1.92	82.5	0.29	1.91	86.8

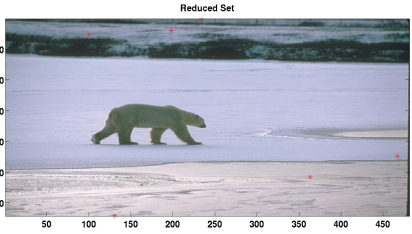
(a) Polar Bear Image



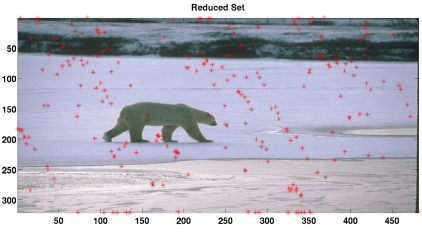
(b) Eigenspace for 1^{st} Highly Sparse Method.



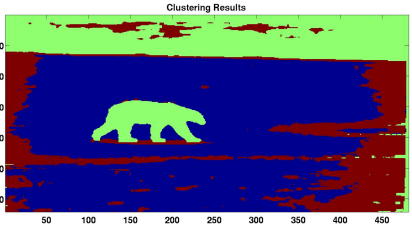
(c) Eigenspace for 2^{nd} Highly Sparse Method.



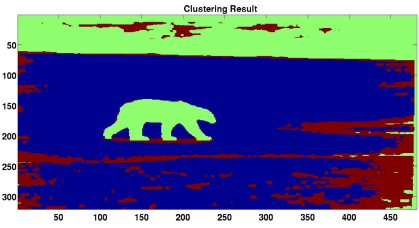
(d) Reduced Set for 1^{st} Highly Sparse Method.



(e) Reduced Set for 2^{nd} Highly Sparse Method.



(f) Clustering Results for 1^{st} Highly Sparse Method.



(g) Clustering Results for 2^{nd} Highly Sparse Method.

Fig. 1. Steps involved for the 2 proposed highly sparse KSC models for an image dataset

An image segmentation experiment using the χ^2 kernel is shown in Figure 1. The total number of pixels is $154,401$ (321×481). The training set consists of $N_{tr} = 7,500$ pixels and the validation set consists of $10,000$ pixels. Both these set are selected by maximizing the quadratic Rènyi entropy. After validation we

obtain $k = 3$ for kernel parameter $\sigma_\chi = 2.807$. The 1^{st} clustering model uses just 6 pixels out of $7,500$ pixels while the second method uses 225 pixels out of $7,500$ training pixels. Since original cluster memberships for this image is not known beforehand, we use 2 internal quality metrics - the silhouette criterion (sil) and the davies-bouldin index (db) as described in [12]. For the 1^{st} highly sparse KSC model the sil value is 0.39 and the db is 1.35 and for the 2^{nd} proposed method these values are 0.32 and 1.15 respectively.

Figure 1a represents the image to be segmented. Figures 1b and 1c showcase the eigenspace. In Figure 1b the red lines represent the cluster means and the black lines represent the farthest and the median projection for that cluster. Similarly, in Figure 1c the red triangles represent the cluster means and the black triangles correspond to 10% of the projections. Figures 1d and 1e highlight the pixels selected from the image as the reduced set $\mathcal{RS}$. These pixels are marked by red-colored '*' reference. Figures 1f and 1g depict the clustering results for the highly sparse KSC models.

5 Conclusion

We proposed 2 highly sparse reductions to KSC model based on a reduced set method. This was achieved by selection of those projections from the eigenspace which occupy certain positions w.r.t. mean projection for each cluster. The clustering model only depended on this reduced set $\mathcal{RS}$ obtained from the training points. This made the clustering model simpler and the predictive model faster as less number of computations were required for out-of-sample extensions. The simulations showed the applicability of the proposed sparse method on various overlapping and real world datasets.

Acknowledgments. This work was supported by Research Council KUL, ERC AdG A-DATADRIVE-B, GOA/10/09MaNet, CoE EF/05/006, FWO G.0588.09, G.0377.12, SBO POM, IUAP P6/04 DYSCO, COST intelliCIS.

References

1. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Proceedings of the Advances in Neural Information Processing Systems, pp. 849–856. MIT Press, Cambridge (2002)
2. Luxburg, U.: A tutorial on Spectral clustering. Statistics and Computing 17(4), 395–416
3. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Intelligence 22(8), 888–905 (2000)
4. Alzate, C., Suykens, J.A.K.: Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(2), 335–347 (2010)
5. Alzate, C., Suykens, J.A.K.: Highly Sparse Kernel Spectral Clustering with Predictive Out-of-sample extensions. In: ESANN, pp. 235–240 (2010)

6. Mall, R., Langone, R., Suykens, J.A.K.: Kernel Spectral Clustering for Big Data Networks. Entropy 15(5), 1567–1586 (2013)
7. Langone, R., Mall, R., Suykens, J.A.K.: Soft Kernel Spectral Clustering. IJCNN (2013)
8. Alzate, C., Suykens, J.A.K.: Sparse kernel spectral clustering models for large-scale data analysis. Neurocomputing 74(9), 1382–1390 (2011)
9. Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
10. Girolami, M.: Orthogonal series density estimation and the kernel eigenvalue problem. Neural Computation 14(3), 1000–1017 (2002)
11. Kenney, J.F., Keeping, E.S.: Linear Regression and Correlation. Mathematics of Statistics 3(1), ch. 15, 252–285
12. Rabbany, R., Takaffoli, M., Fagnan, J., Zaiane, O.R., Campello, R.J.G.B.: Relative Validity Criteria for Community Mining Algorithms. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 258–265 (2012)