Highly Sparse Kernel Spectral Clustering with Predictive Out-of-Sample Extensions

Carlos Alzate and Johan A.K. Suykens

K.U.Leuven - Dept. of Electrical Engineering ESAT - SCD
Kasteelpark Arenberg 10, B-3001 Leuven - Belgium

Abstract. Kernel spectral clustering has been formulated as a primal - dual optimization setting allowing natural extensions to out-of-sample data together with model selection in a learning framework which is important for obtaining a good generalization performance. In this paper, we propose a new sparse method for kernel spectral clustering. The approach exploits the structure of the eigenvectors and the corresponding projections of the data when the clusters are well formed. Experimental results with toy data and images show highly sparse clustering models with predictive capabilities.

1 Introduction

Classical spectral clustering algorithms [1, 2] have been recently proven to be very successful compared to traditional techniques such as k-means. These formulations have roots in graph theory and are solved via eigenvalue problems where certain eigenvectors contain information about the groups present on the data [3]. One issue with classical spectral clustering is that the clusters cannot be easily extended to out-of-sample data. A new spectral clustering algorithm based on a weighted version of kernel PCA was introduced in [4]. This approach is formulated in a primal and dual optimization framework allowing to extend the clustering model to out-of-sample points via projections onto the eigenvectors. The projections are expressed in terms of non-sparse kernel expansions. In this paper, we propose a method to sparsify the clustering model by exploiting the structure of the projections when the clusters are well formed. The proposed approach is based on a reduced set method for approximating the projections. The reduced set points are chosen such that they follow a special structure on the projections. This paper is organized as follows. Section 2 summarizes the kernel spectral clustering method. Section 3 describes the new sparse method. Section 4 contains the experiments and in Section 5, conclusions are given.

2 Predictive Kernel Spectral Clustering

2.1 Primal - Dual Formulation

Given training data $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ and the number of desired clusters k, the following clustering model can be assumed:

$$e^{(l)} = \Phi w^{(l)} + b_l 1_N, l = 1, \ldots, n_e$$

where $e^{(l)} = [e_1^{(l)}; \ldots; e_N^{(l)}]$ is the compact form of $e_i^{(l)} = {w^{(l)}}^T \varphi(x_i) + b_l$, $\Phi = [\varphi(x_1)^T; \ldots; \varphi(x_N)^T]$ is the $N \times d_h$ feature matrix, $\varphi : \mathbb{R}^d \to \mathbb{R}^{d_h}$ is the mapping to a high-dimensional feature space of dimension d_h, b_l are bias terms and $i = 1, \ldots, N, l = 1, \ldots, n_e$. The projections $e^{(l)}$ represent the latent variables of a set of n_e binary cluster indicators obtained by $\text{sign}(e^{(l)})$ which can be encoded to obtain the final k groups. Consider the following constrained optimization problem in the primal space [4]:

$$\begin{aligned} \min_{w^{(l)}, e^{(l)}, b_l} \frac{1}{2N} \sum_{l=1}^{n_e} \gamma_l {e^{(l)}}^T V e^{(l)} - \frac{1}{2} \sum_{l=1}^{n_e} {w^{(l)}}^T w^{(l)} \\ \text{such that } e^{(l)} = \Phi w^{(l)} + b_l 1_N, l = 1, \ldots, n_e \end{aligned} \tag{1}$$

where γ_l are regularization parameters and V is a positive definite weight matrix typically chosen to be diagonal. The KKT optimality conditions of the Lagrangian of (1) are: $\frac{\partial \mathcal{L}}{\partial w^{(l)}} = 0 \to w^{(l)} = \Phi^T \alpha^{(l)}$, $\frac{\partial \mathcal{L}}{\partial e^{(l)}} = 0 \to \alpha^{(l)} = \gamma_l V e^{(l)}$, $\frac{\partial \mathcal{L}}{\partial b_l} = 0 \to 1_N^T \alpha^{(l)} = 0$, $\frac{\partial \mathcal{L}}{\partial \alpha^{(l)}} = 0 \to e^{(l)} = \Phi w^{(l)} + b_l 1_N$, $l = 1, \ldots, n_e$. Eliminating the primal variables $w^{(l)}, e^{(l)}, b_l$ leads to the dual eigenvalue problem [4]:

$$V M \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)} \tag{2}$$

where $\lambda_l = N/\gamma_l, l = 1, \ldots, n_e$, $M = I_N - 1/(1_N^T V 1_N) 1_N 1_N^T V$ and $\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$. The projections written in terms of the dual variables become

$$e_i^{(l)} = \sum_{j=1}^{N} \alpha_j^{(l)} K(x_j, x_i) + b_l$$

and the bias terms: $b_l = -1/(1_N^T V 1_N) 1_N^T V \Omega \alpha^{(l)}, l = 1, \ldots, n_e$. In the case when the weight matrix $V = I$, then (2) becomes kernel PCA. In the same way, if $V = D^{-1} = \text{diag}(1/d_1; \ldots; 1/d_N)$ where $d_i = \sum_{j=1}^{N} \Omega_{ij}$ then (2) is related to the random walks model for spectral clustering [5, 6]. The relationship can be interpreted as first applying a weighted centering to the kernel matrix[1] and then applying the random walks method. This special centering consists of removing the weighted mean of the data points in the feature space and is induced by the bias terms in the primal model. The centering weights are given by $D^{-1} 1_N$.

2.2 Piecewise Constant Eigenvectors and Encoding

If the kernel matrix Ω represents the similarity matrix of a graph with k connected components and $V = D^{-1}$, then the eigenvectors $\alpha^{(l)}$ associated to the $k-1$ largest eigenvalues of $D^{-1} M \Omega$ are piecewise constant and indicators of the corresponding connected parts of the graph. Thus, we set $n_e = k - 1$. A key difference with classical spectral clustering algorithms is the fact that 1_N

[1] Since $M^T \alpha^{(l)} = \alpha^{(l)}$ and V is invertible, the eigenvalue problem (2) is equivalent to $M \Omega M^T \alpha^{(l)} = \lambda V^{-1} \alpha^{(l)}$ where the weighted mean is removed from the rows and columns of the kernel matrix.

is not an eigenvector of $D^{-1}M\Omega$. This can be seen from the KKT optimality condition $1_N^T\alpha^{(l)} = 0$ which imposes that the eigenvectors should have zero mean. Each cluster is now represented as a single point in the $\mathbb{R}^{k-1}$ eigenspace. Moreover, these single points are always in different orthants due also to the KKT optimality conditions. One way to encode the eigenvectors is to consider that two points are in the same cluster if they are in the same orthant in the corresponding eigenspace. A codebook can be obtained from the rows of the matrix containing the $k-1$ binarized leading eigenvectors in the columns.

2.3 Out-of-Sample Extensions and Decoding

The projections $e^{(l)}$ define the cluster indicators for training data. In the case of an out-of-sample data point x, the projections become:

$$\hat{z}^{(l)}(x) = \sum_{i=1}^{N} \alpha_i^{(l)} K(x, x_i) + b_l.$$

The possibility of extending the clustering model to out-of-sample data in a natural way corresponds to one of the main advantages of this formulation. In classical spectral clustering, extensions to out-of-sample data are not clear and should rely on approximations such as the Nyström method [7]. Out-of-sample extensions also allow performing spectral clustering in a learning framework with training, validation and test stages which becomes important for generalization. Decoding consists of comparing the binarized projections with respect to the codewords in the codebook and assigning cluster membership based on minimal Hamming distance.

2.4 Collinearity and Model Selection

If the eigenvectors are piecewise constant, then data points in the same cluster are collinear in the projections space. This is due to the fact that the projections of test points $\{x_t\}_{t=1}^{N_t}$ can then be rewritten as $\hat{z}_t^{(l)} = c_p^{(l)} \sum_{j \in \mathcal{A}_p} K(x_t, x_j) + \sum_{u \notin \mathcal{A}_p} \alpha_u^{(l)} K(x_t, x_u) + b_l$ where $c_p^{(l)}$ is the constant value corresponding to the p-th cluster $\mathcal{A}_p$ in the l-th eigenvector. Assuming that $K(x_i, x_j) \to 0$ when x_i and x_j are in different clusters then the term $\sum_{u \notin \mathcal{A}_p} \alpha_u^{(l)} K(x_t, x_u)$ vanishes and the clusters are visualized as lines in the projections space (which holds for the common RBF kernel and the χ^2 kernel). Since the cluster membership depends on the orthant in which the projected variables are located, an intuitive membership *certainty* measure is the distance of a given data point in the projection space from the origin (assuming zero mean projected variables). The larger the distance, the more certain the point belongs to the corresponding cluster. Thus, the tips of the lines can serve as cluster prototypes. The Balanced Line Fit (BLF) criterion introduced in [4] is an average measure of collinearity and balance of the obtained clusters on validation data and can be used for obtaining the number of clusters k and the kernel parameters.

3 Sparse Model

Since typically the number of required eigenvectors is much less than N, specialized techniques such as the Lanczos method [8] can be used in order to solve (2) efficiently. However, storing the full $D^{-1}M\Omega$ matrix would still be required. A way to overcome this issue is to build a clustering model on a subset of the dataset and infer the cluster membership of the remaining data points using the out-of-sample extension. Obtaining a representative subset of the available data can be done in a greedy manner by adding points to the training pool such that the quadratic Renyi entropy is maximized [9, 10]. Also note that the projections are expressed in terms of non-sparse kernel expansions. The primal vectors $w^{(l)} = \sum_{i=1}^{N} \alpha_i^{(l)} \varphi(x_i)$ can be approximated by a reduced set method. The objective is then to approximate $w^{(l)}$ by $\tilde{w}^{(l)} = \sum_{j=1}^{R} \beta_j^{(l)} \varphi(\tilde{x}_j)$ where $\tilde{x}_j$ corresponds to the reduced set of points and $R \leq N$. If the reduced set is known, one approach to obtain the reduced set coefficients $\beta^{(l)}$ is given by $\min_{\beta^{(l)}} ||w^{(l)} - \tilde{w}^{(l)}||_2^2$ which leads to solving the linear system $\Omega^{\Psi\Psi}\beta^{(l)} = \Omega^{\Psi\Phi}\alpha^{(l)}$, where $\Omega_{mn}^{\Psi\Psi} = K(\tilde{x}_m, \tilde{x}_n), \Omega_{mi}^{\Psi\Phi} = K(\tilde{x}_m, x_i)$, $m, n = 1, \ldots, R, i = 1, \ldots, N$ and $l = 1, \ldots, k-1$ [11, 12]. In this way, after finding $\beta^{(l)}$, the projections of a data point x can be approximated by $z^{(l)}(x) \approx \sum_{j=1}^{R} \beta_j^{(l)} K(\tilde{x}_j, x)$. A reduced set can be built by considering input data points that correspond to certain positions sampled from the lines. We propose to use both ends of the lines (endpoints and points closest to the origin) together with the median point. Thus, the clustering model only depends of $R = 3k$ data points and typically $k \ll N$.

4 Experimental Results

A toy experiment using the RBF kernel can be seen in Figure 1. The dataset consists of 3 Gaussian clouds in 2D for a total number of $6,000$ data points. The training set was selected using the quadratic Renyi entropy as described in [9] and $N = 600$ points were selected. The validation set contained $1,200$ randomly selected points and the BLF was used on validation data to find $k = 3$ and the RBF kernel parameter $\sigma = 0.56$. Using these tuning parameters, the clustering model is trained and the reduced set is obtained from the projections of training data. The cluster indicators of the remaining data points are then inferred via the approximated projections computed using the reduced set. An image segmentation experiment using the χ^2 kernel is shown in Figure 2. The total number of pixels is $154,401$ (321×481). The training set consisted of $N = 1,000$ pixels selected using the entropy criterion. The validation set contained $20,000$ randomly selected pixels for model selection using the BLF with obtained $k = 3$ and $\sigma_\chi = 0.04$. Note the strong line structures present on the validation set projections. Computation time for eigendecomposition and prediction was 34 seconds in a 2.4 GHz Core 2 Duo, 4 GB RAM and MATLAB 7.9. The clustering model depends only on 9 out of $1,000$ input data points which corresponds to a sparseness of 99.1%.

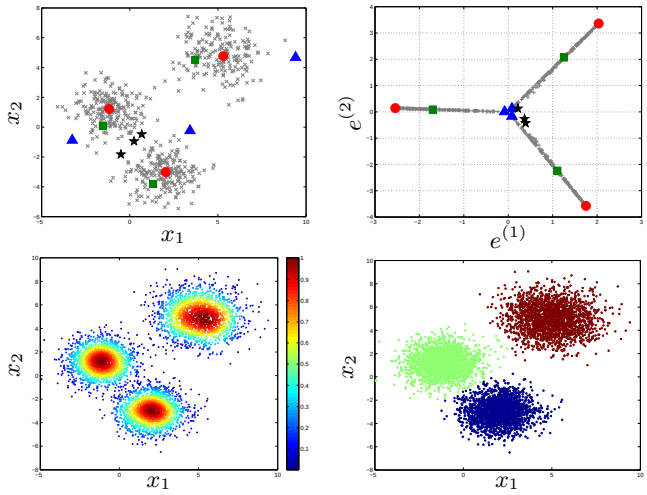


Fig. 1: **Top left:** Training set and reduced set points. **Top right:** Training set projections used to obtain the reduced set points ($R = 9$). The stars are outside the lines and correspond to points in zones of overlap. **Bottom left:** Cluster membership certainty of the full dataset. **Bottom right:** Inferred clustering of the full dataset using the out-of-sample extension and the proposed method.

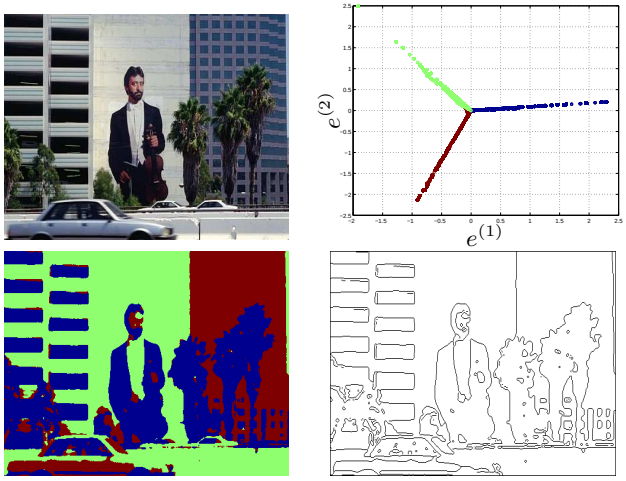


Fig. 2: **Top left:** Original image. **Top right:** Validation set projections for model selection with the BLF. Different colors indicate different clusters. **Bottom left:** Segment-label image. **Bottom right:** Inferred segmentation. The clusters were inferred using only $R = 9$ out of $1,000$ training data points.

5 Conclusions

A new highly sparse approach based on a reduced set method is proposed for kernel spectral clustering. The methodology allows predicting the cluster indicators of out-of-sample points. The clustering model only depends on a reduced set of training points which are selected by exploiting the structure of the projections. Due to the predictive capability of the algorithm, model selection can be done by selecting parameters such that the projections on validation data are as collinear as possible. The simulations show the applicability of the proposed sparse method.

Acknowledgements: This work was supported by grants and projects for the Research Council K.U.Leuven (GOA-Mefisto 666, GOA-Ambiorics, several PhD / Postdocs & fellow grants), the Flemish Government FWO: PhD / Postdocs grants, projects G.0240.99, G.0211.05, G.0407.02, G.0197.02, G.0080.01, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0226.06, G.0302.07, ICCoS, ANMMM; AWI;IWT:PhD grants, GBOU (McKnow) Soft4s, the Belgian Federal Government (Belgian Federal Science Policy Office: IUAP V-22; PODO-II (CP/01/40), the EU(FP5-Quprodis, ERNSI, Eureka 2063-Impact;Eureka 2419-FLiTE) and Contracts Research/Agreements (ISMC / IPCOS, Data4s, TML,Elia, LMS, IPCOS, Mastercard). Johan Suykens is a professor at the K.U.Leuven, Belgium. The scientific responsibility is assumed by its authors.

References

[1] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[2] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 14*, pages 849–856, MIT Press, 2002.

[3] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[4] C. Alzate and J. A. K. Suykens. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):335–347, February 2010.

[5] M. Meila and J. Shi. A random walks view of spectral segmentation. In *Artificial Intelligence and Statistics AISTATS*, 2001.

[6] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[7] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, February 2004.

[8] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1996.

[9] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[10] M. Girolami. Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 14(3):669–688, 2002.

[11] B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K. R. Müller, M. Scholz, G. Rätsch, and A. J. Smola. Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10(5):1000–1017, September 1999.

[12] C. Alzate and J. A. K. Suykens. Sparse kernel models for spectral clustering using the incomplete Cholesky decomposition. In *Proc. of the 2008 International Joint Conference on Neural Networks (IJCNN 2008)*, pages 3555–3562, 2008.