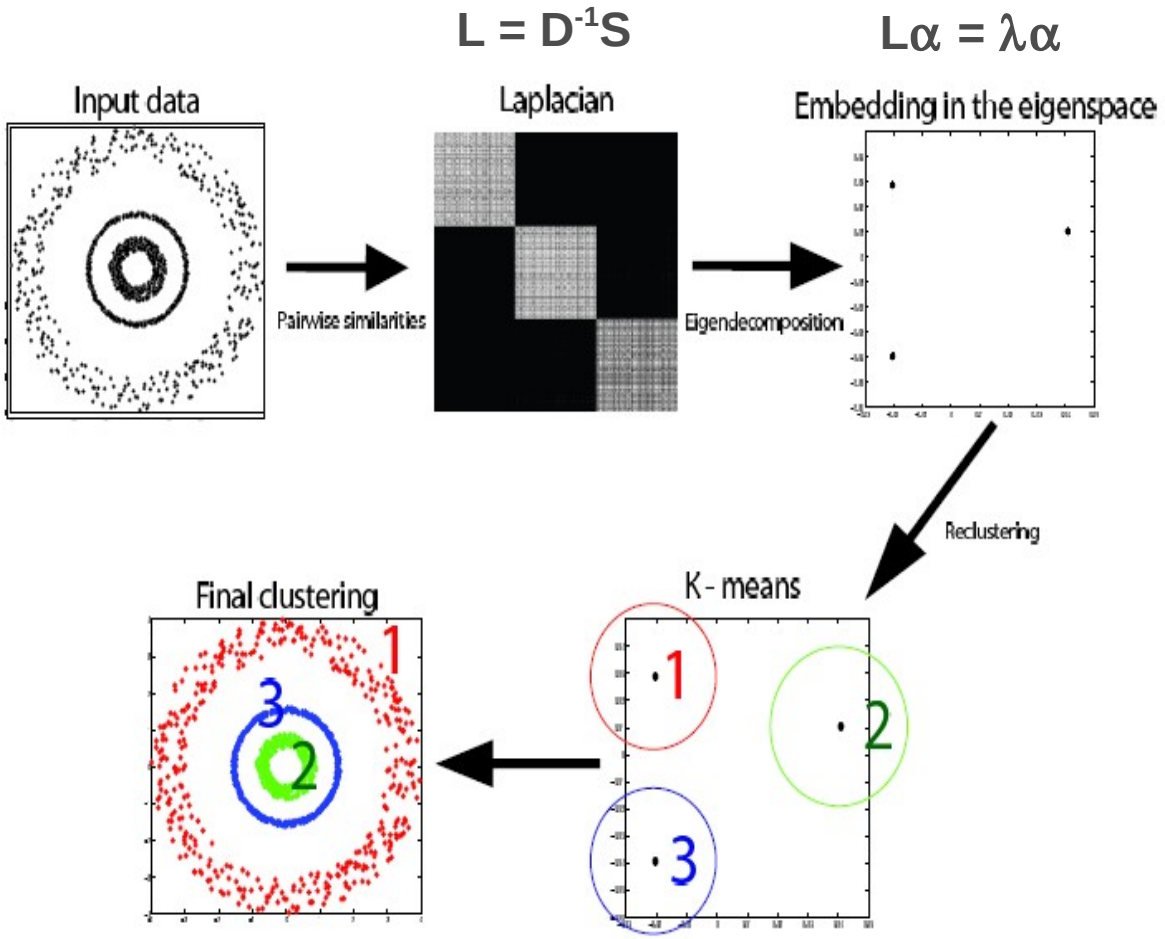
Soft Kernel spectral clustering (SKSC)

Rocco Langone[*], Raghvendra Mall[*] and Johan A. K. Suykens[*]

[*]Department of Electrical Engineering (ESAT), SCD / I Minds Future Health Department, KU Leuven, B-3001 Leuven Belgium

IJCNN 2013
Dallas - 07/08/2013

Classical Spectral Clustering [1]



Advantages:

- can discover complex clustering boundaries
- outperforms many other methods in practice

Drawbacks:

- No model selection
- No out-of-sample extension

Kernel Spectral Clustering (KSC, [2])

Primal problem:

Degree matrix

$$\min_{w^{(l)},e^{(l)},b_l} \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)^T} w^{(l)} - \frac{1}{2N}\sum_{l=1}^{k-1} \gamma_l e^{(l)^T} D^{-1} e^{(l)}$$

$$\text{such that } e^{(l)} = \Phi w^{(l)} + b_l 1_N$$

→ Clustering model

Dual problem:

Centering matrix — Training kernel matrix

$$D^{-1} M_D \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}$$

Out-of-sample extension:

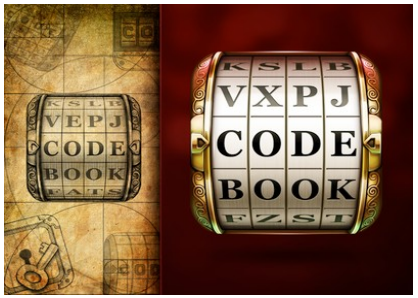
Score variable (projection) — RBF kernel parameter

$$e^{(l)}_{\text{test}} = \Omega_{\text{test}} \alpha^{(l)} + b_l 1_{\text{Ntest}}$$

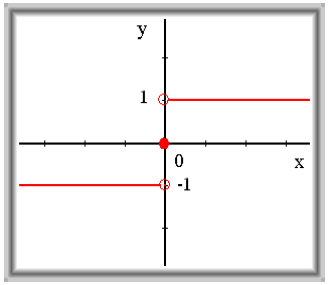
$$\Omega_{\text{test,ri}} = K(x^{\text{test}}_r, x_i), \quad K(x_i, x_j) = \exp(-||x_i - x_j||^2_2/\sigma^2)$$

Cluster assignment:

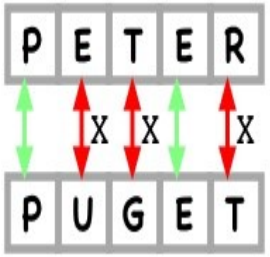
1) Form Training codebook



2) Binarize projections



3) ECOC decoding



BLF

Balanced Linefit (BLF):

User-defined parameter

Validation set

Number of clusters

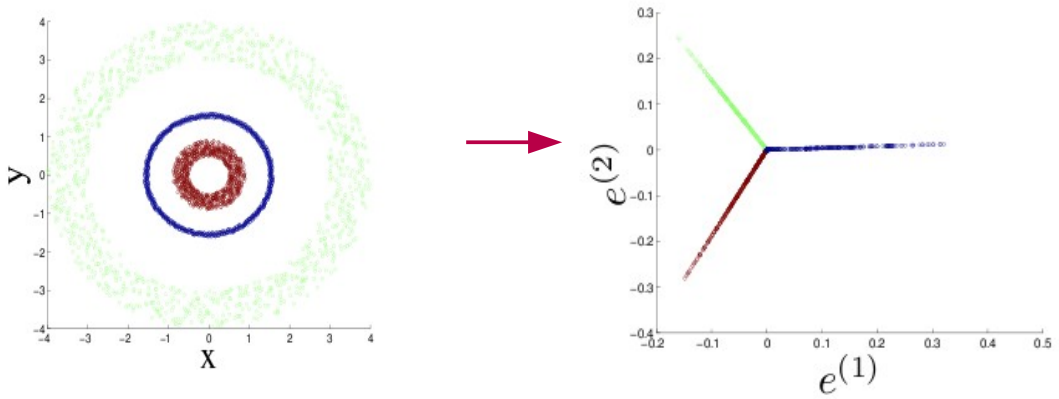
$$\mathrm{BLF}(\mathcal{D}^V, k) = \eta \mathrm{linefit}(\mathcal{D}^V, k) + (1-\eta)\mathrm{balance}(\mathcal{D}^V, k)$$

Linefit: equals **0** when the score variables are distributed spherically and **1** when they are collinear

Balance: equals **1** when the clusters have the same number of elements and **0** in extremely unbalanced cases

Input space

Projection space



Drawbacks:

- free parameter (η) not clear how to tune
- tends to be biased toward k=2 clusters
- assumes few overlap between the clusters.

Soft Kernel Spectral Clustering (SKSC, [3])

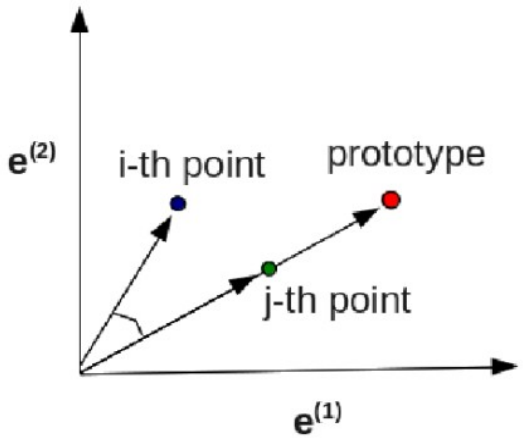
Motivations:

• Hard clustering works fine when the clusters are compact and well separated, while the performance can decrease dramatically when they overlap
• Soft clustering improves the interpretability of the results.

Idea:

• Use KSC as initialization step to find a first division into groups
• Refine this grouping by re-calculating the prototypes in the score variable space
• New cluster assignments are performed by means of the cosine distance between each point and the prototypes
• Subjective probability of belonging to each cluster (soft assignments) is calculated based on these distances [4].

Cosine distance



Soft membership

$$cm_i^{(q)} = \frac{\prod_{j \neq q} d_{ij}^{\cos}}{\sum_{p=1}^{k} \prod_{j \neq p} d_{ij}^{\cos}}$$

New model selection scheme

Average membership strength (AMS):

- calculate a kind of mean membership per cluster indicating the average degree of belonging of the points to a cluster.
- do the same for every cluster take the mean.

$$\mathrm{AMS} = \frac{1}{k}\sum_{p=1}^{k}\frac{1}{n_p}\sum_{i=1}^{n_p} cm_i^{(p)}$$

Advantages:

- does not have any parameter to tune
- it can be used in an effective way also in case of overlapping clusters.

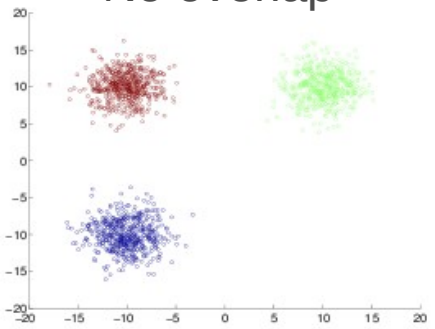
Clustering results: Toy data I

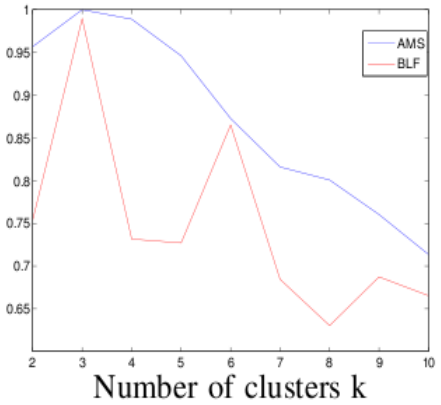
Three Gaussian clouds

Model selection

Cluster quality evaluation

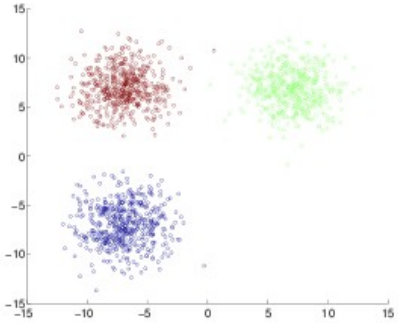
No overlap

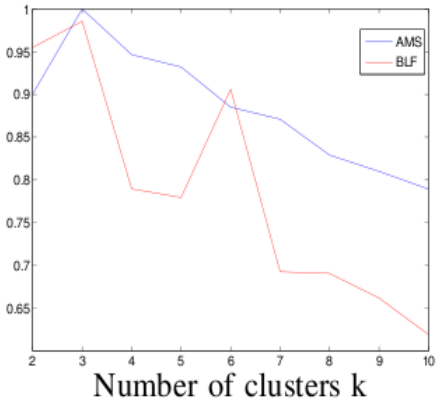




	KSC	SKSC
Silhouette	0.96	0.96
DBI	0.24	0.24
ARI	1	1

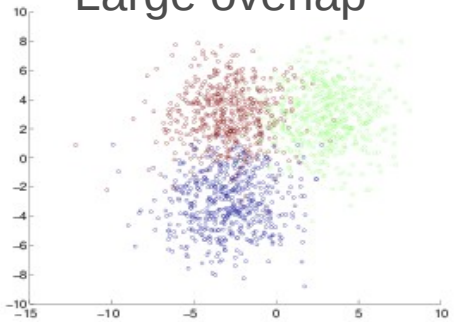
Small overlap

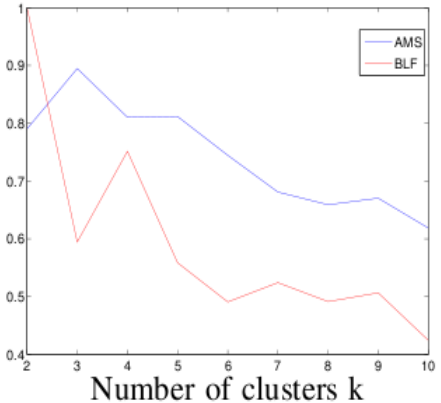




	KSC	SKSC
Silhouette	0.90	**0.91**
DBI	0.34	**0.33**
ARI	0.96	**0.99**

Large overlap

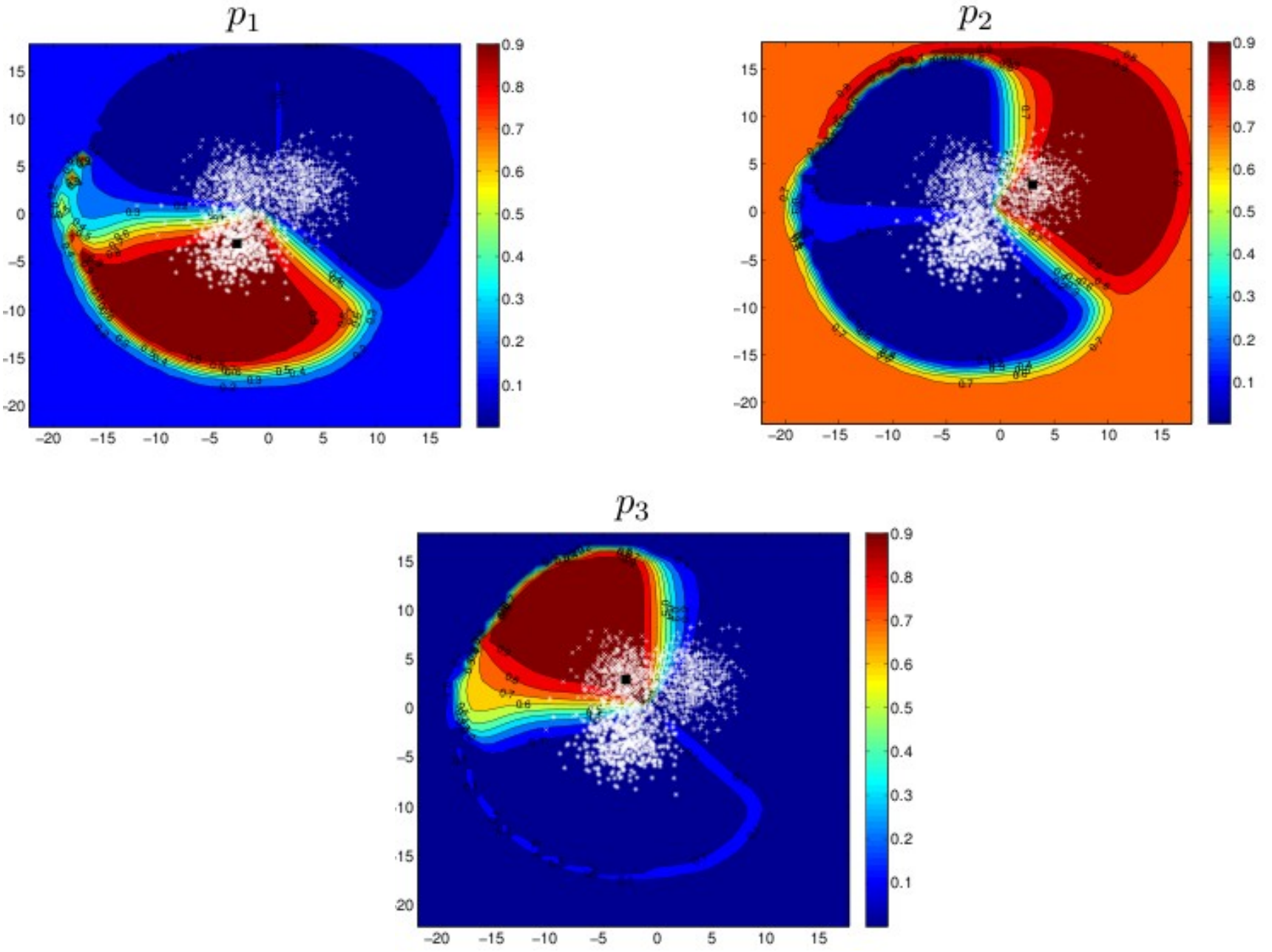




	KSC	SKSC
Silhouette	0.59	**0.64**
DBI	0.79	**0.76**
ARI	0.65	**0.74**

Example SKSC results

Three Gaussians with large overlap:



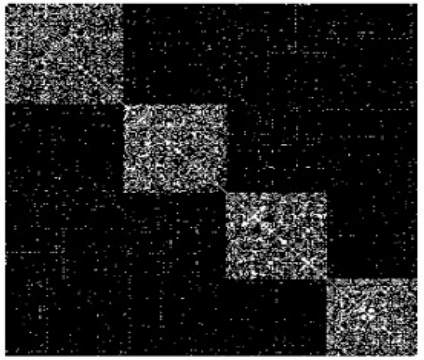
Clustering results: Toy data II

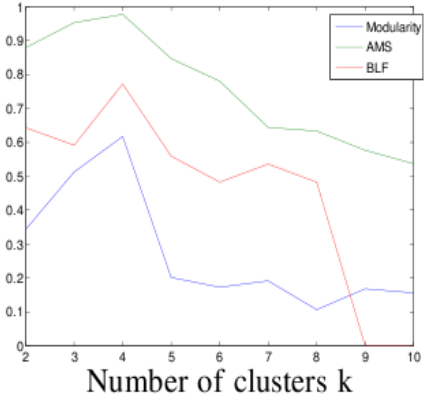
Synthetic network

Model selection

Cluster quality evaluation

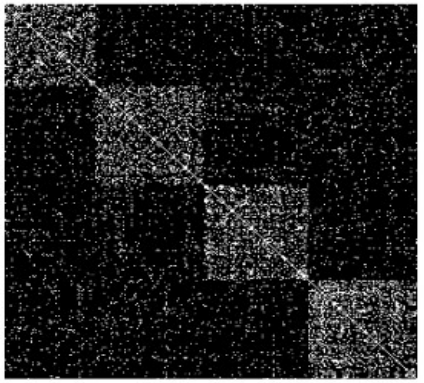
Small overlap

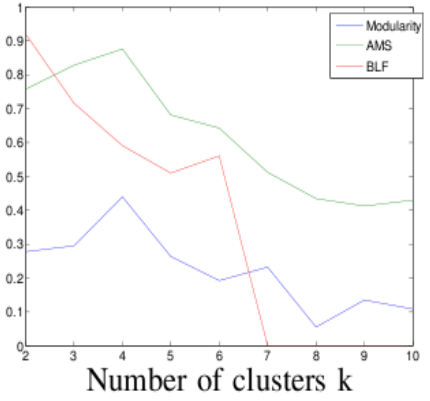




	KSC	SKSC
Modularity	0.68	**0.69**
Conductance	0.0040	0.0040
ARI	0.98	**0.99**

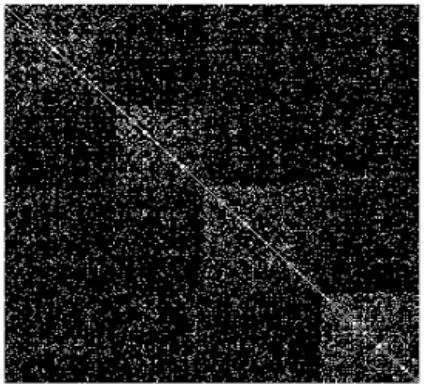
Medium overlap

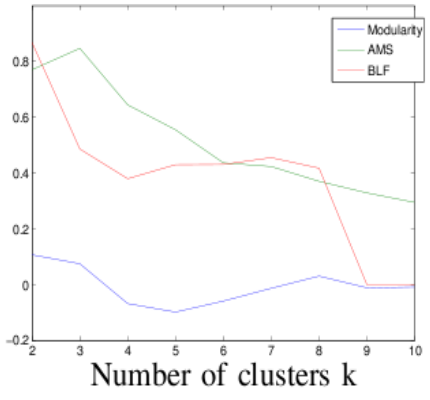




	KSC	SKSC
Modularity	0.53	**0.54**
Conductance	0.0042	0.0042
ARI	0.96	**0.97**

Large overlap





	KSC	SKSC
Modularity	0.34	**0.36**
Conductance	0.0043	0.0043
ARI	0.65	**0.66**

Clustering results: Image Segmentation

Dataset:

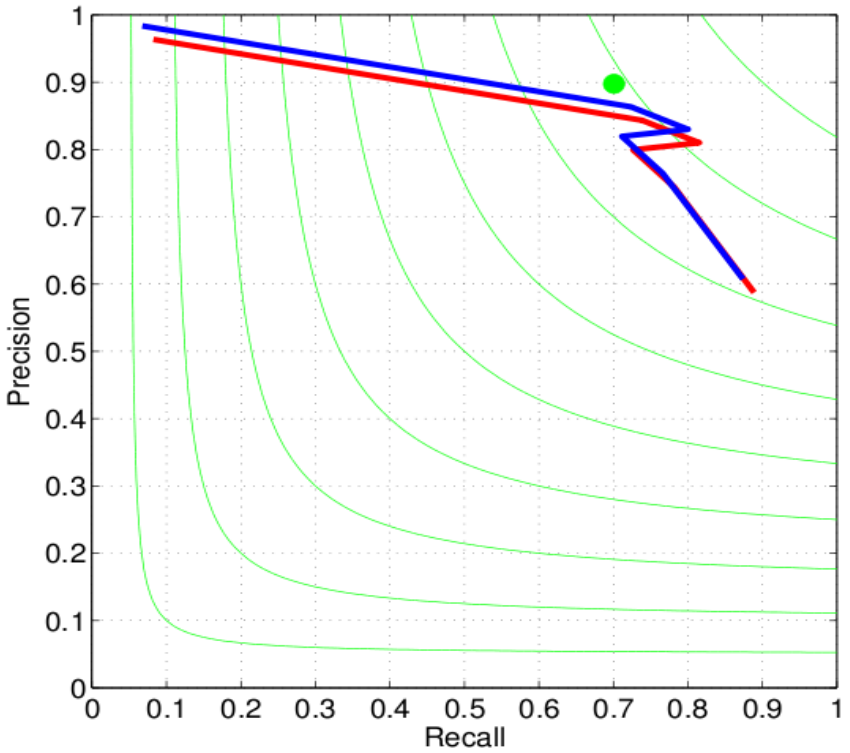
We utilized 10 color images from the Berkeley image dataset].

Segmentation quality:

F-measure (2*Precision*Recall)/(Precision + Recall) with respect to human ground-truth boundaries

Model comparison:

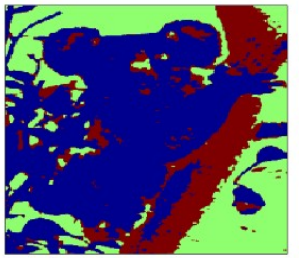
SKSC
KSC



SKSC segmentation example:

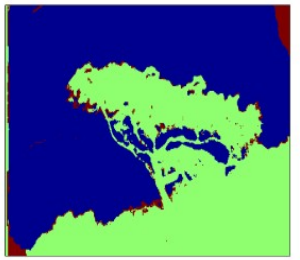
ID 69015





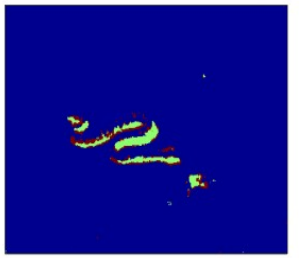
ID 147091





ID 196073





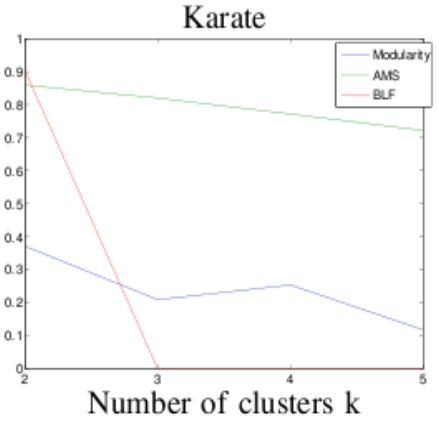
Clustering results: Community detection

Datasets:

Model selection:

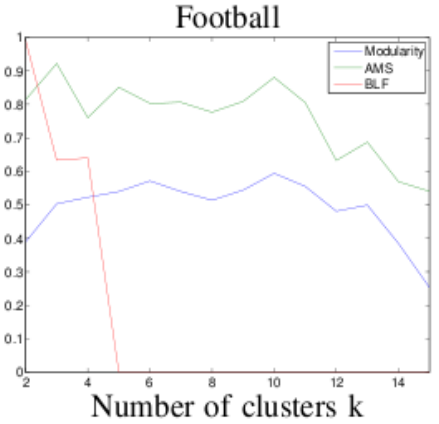
Clustering results:

Karate: 34 member nodes of the Zachary's karate club network, formed by **2** communities



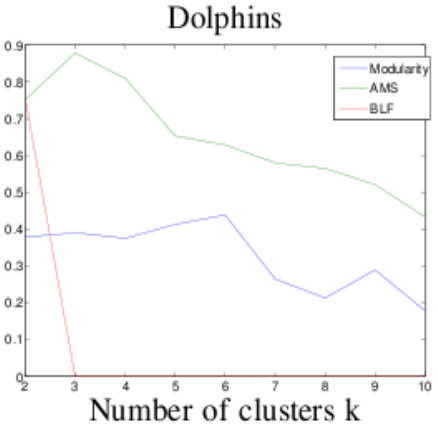
	KSC	SKSC
Modularity	0.371	0.371
Conductance	0.083	0.083
ARI	1	1

Football: American college football Games network formed by 115 nodes (the teams) and 616 edges (the games). It can be divided into **12** communities according to athletic conferences.



	KSC	SKSC
Modularity	0.594	**0.597**
Conductance	0.772	**0.766**
ARI	0.818	0.818

Dolphins: this network depicts the associations between 62 dolphins. It can be naturally split in 3 main Groups: female-female, male-male and mixed sex associations.



	KSC	SKSC
Modularity	**0.378**	0.370
Conductance	**0.146**	0.148
ARI	**0.935**	0.872

Conclusions

- We proposed an algorithm, SKSC, to provide moderated outputs for Kernel Spectral Clustering
- A new model selection scheme, AMS, has also been introduced
- On some toy data and real life examples we showed that SKSC and AMS can outperform respectively KSC and BLF, mostly when the clusters present some degree of overlap.

References

[1] U. von Luxburg, “A tutorial on spectral clustering,” Statistics and Computing, vol. 17, no. 4, pp. 395–416, 2007.
[2] 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, vol. 32, no. 2, pp. 335–347, February 2010.
[3] Langone R., Mall R., Suykens J.A.K., "Soft Kernel Spectral Clustering", IJCNN 2013.
[4] Ben-Israel and C. Iyigun, “Probabilistic d-clustering,” J. Classif., vol. 25, no. 1, pp. 5–26, Jun. 2008.