Spectral Clustering & Reproducing Kernels

Ilaria Giulini

INRIA Saclay

Joint work with Olivier Catoni

23 October 2015

RDMath IdF
Domaine d'Intérêt Majeur (DIM)
en Mathématiques

îledeFrance

Clustering: task of grouping objects into classes (clusters) according to their similarities.

Spectral clustering methods use data-dependent matrices (Laplacian matrix) to perform unsupervised clustering.

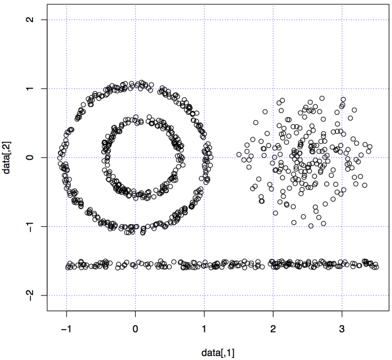
Setting: Spectral clustering in a Hilbert space
(where the points are i.i.d. according to an unknown distribution whose support is a union of compact connected components).

Our approach:

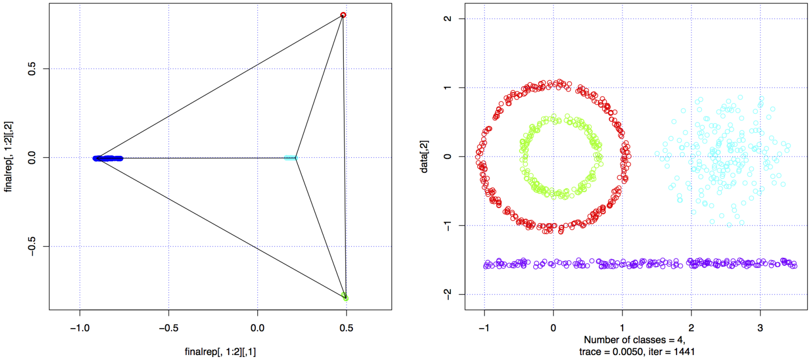
- View spectral clustering as a change of representation in a RKHS
- Modify the Ng, Jordan, Weiss algorithm
  (interpretation in terms of Markov chains with exp transitions)
- Estimate automatically the number of clusters.

EXAMPLE

Goal: Cluster $X_1, \ldots, X_n \in \mathbb{R}^2 \quad (n = 900)$



EXAMPLE



Note: clusters are at the vertices of a simplex

$\longrightarrow$ classification becomes trivial

Ng, Jordan, Weiss algorithm

Input:

- $X_1, \ldots, X_n$ the points to cluster
- c the number of clusters

1. Form $A_{ij} = \begin{cases} \exp(-\beta \|X_i - X_j\|^2) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$
2. Construct $L = D^{-1/2} A D^{-1/2}$ where $D_{ii} = \sum_j A_{ij}$
3. Compute c largest eigenvectors $v_1, \ldots, v_c$ of L and form $X = \begin{bmatrix} v_1 \ldots v_c \end{bmatrix}_{n \times c}$
4. Cluster each (renormalized) row of X into c clusters (e.g. via k-means)

Intuition

Let decompose

$$L = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\,U^\top$$

The Ng, Jordan, Weiss algorithm is based on

$$U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_c, 0, \ldots, 0)\,U^\top$$

Idea: Replace the projection with a smooth cut-off of the eigenvalues.

More precisely

$$U\,\mathrm{diag}(\lambda_1^m, \ldots, \lambda_n^m)\,U^\top$$

UNDERLYING INTEGRAL OPERATORS

Idea: View previous matrices as empirical versions of underlying integral operators.

Assume $X_1, \dots, X_n \in \mathcal{H} \sim \mathrm{P}$ (unknown).

$$
\begin{array}{ccc}
A \quad \text{(affinity matrix)} & \longleftrightarrow & K(x,y) = \exp\big(-\beta\|x-y\|^2\big) \\
L = D^{-1/2} A D^{-1/2} & \longleftrightarrow & \bar{K}(x,y) = \mu(x)^{-1/2} K(x,y) \mu(y)^{-1/2} \\
D_{ii} = \sum_j A_{ij} & \longleftrightarrow & \mu(x) = \int K(x,z)\, \mathrm{dP}(z) \\
\{\xi_i\}_{i=1}^n \mapsto \frac{1}{n} \sum_{j=1}^n L_{ij} \xi_j & \longleftrightarrow & L_{\bar{K}} : f \mapsto \int \bar{K}(x,z) f(z)\, \mathrm{dP}(z)
\end{array}
$$

Markov chain analysis of spectral clustering

The matrix A is used to form $M = D^{-1}A$ (Markov matrix with exp transitions).

Note:

$$\{\xi_i\}_{i=1}^n \mapsto \frac{1}{n}\sum_{j=1}^n L_{ij}\xi_j = \frac{1}{n}\sum_{j=1}^n D_{ii}^{1/2} M_{ij} D_{jj}^{-1/2}\xi_j$$

To determine clusters, use $M^{\exp(\beta T)}$

Hope for a similar behavior in the continuous case:

Define $M(x,y) = \mu(x)^{-1}K(x,y)$, so that

$$L_{\bar{K}} : f \mapsto \int \bar{K}(x,z) f(z)\, \mathrm{dP}(z) = \mu(x)^{1/2} \int M(x,z)\mu(z)^{-1/2} f(z)\, \mathrm{dP}(z)$$

Idea: Consider an iterate of $L_{\bar{K}}$.

Remark that

$$L_{\bar{K}}^{2m} f(x) = L_{\bar{K}_{2m}} f(x) = \int \bar{K}_{2m}(x, z)\, f(z)\, \mathrm{dP}(z),$$

where

$$\bar{K}_{2m}(x, y) = \int \bar{K}(y, z_1)\bar{K}(z_1, z_2) \ldots \bar{K}(z_{2m-1}, x)\, \mathrm{dP}^{\otimes(2m-1)}(z_1, \ldots, z_{2m-1})$$

whereas the kernel M defines a Markov chain $(Z_k)_{k\in\mathbb{N}}$ with transitions

$$M(x, y) = \frac{\mathrm{dP}_{Z_k \mid Z_{k-1}=x}}{\mathrm{dP}}(y)$$

and invariant measure Q defined by its density $\mathrm{dQ}/\mathrm{dP} = \mu$.

Proposition. For any $x, y \in \mathrm{supp}(\mathrm{P})$,

$$\left\langle \frac{\mathrm{dP}_{Z_m \mid Z_0 = x}}{\mathrm{dQ}}, \frac{\mathrm{dP}_{Z_m \mid Z_0 = y}}{\mathrm{dQ}} \right\rangle_{L^2_{\mathrm{Q}}} = \mu(x)^{-1/2} \bar{K}_{2m}(x, y) \mu(y)^{-1/2}$$

Introduce

$$K_m(x, y) = \bar{K}_{2m}(x, x)^{-1/2} \bar{K}_{2m}(x, y) \bar{K}_{2m}(y, y)^{-1/2}$$

In the new representation points are concentrated around ON vectors

Ideal algorithm in terms of kernels

Let $K(x, y) = \exp\big(-\beta\|x - y\|^2\big)$

1. Form (Laplacian operator)
$$\bar{K}(x, y) = \mu(x)^{-1/2} K(x, y) \mu(y)^{-1/2}$$
2. Construct
$$\bar{K}_{2m}(x, y) = \int \bar{K}(y, z_1)\bar{K}(z_1, z_2)\ldots\bar{K}(z_{2m-1}, x)\, \mathrm{dP}^{\otimes(2m-1)}(z_1, \ldots, z_{2m-1})$$
3. Renormalize to obtain
$$K_m(x, y) = \bar{K}_{2m}(x, x)^{-1/2}\bar{K}_{2m}(x, y)\bar{K}_{2m}(y, y)^{-1/2}$$
4. Cluster points according to the new representation defined by the symmetric kernel K_m.

Next step

- Construct an empirical algorithm

  by estimating the kernels

$$\bar{K}(x,y) = \mu(x)^{-1/2} K(x,y) \mu(y)^{-1/2}$$

  and

$$\bar{K}_{2m}(x,y) = \int \bar{K}(y,z_1)\bar{K}(z_1,z_2)\ldots\bar{K}(z_{2m-1},x)\,\mathrm{dP}^{\otimes(2m-1)}(z_1,\ldots,z_{2m-1})$$

- Provide convergence results

Toward an empirical algorithm: Gram operators

Idea: Link the previous kernels ($\bar{K}$ and $\bar{K}_{2m}$) with Gram operators

Note: the kernel $\bar{K}$ defines

- a RHKS $\mathcal{H}$ where

$$\bar{K}(x,y) = \langle \phi_{\bar{K}}(x), \phi_{\bar{K}}(y) \rangle_{\mathcal{H}}$$

- a Gram operator

$$\begin{aligned}
\mathcal{G}_{\bar{K}} \phi_{\bar{K}}(x) &= \int \langle \phi_{\bar{K}}(x), \phi_{\bar{K}}(z) \rangle_{\mathcal{H}} \, \phi_{\bar{K}}(z) \, \mathrm{dP}(z) \\
&= \int \bar{K}(x,z) \, \phi_{\bar{K}}(z) \, \mathrm{dP}(z)
\end{aligned}$$

An estimator of $\bar{K}$

Goal: Estimate $\bar{K}(x,y) = \mu(x)^{-1/2} K(x,y) \mu(y)^{-1/2}$ where

$$\mu(x) = \int K(x,z)\, \mathrm{dP}(z)$$

Note: The kernel $A(x,y) = K(x,y)^{1/2} = \exp\left(-\frac{\beta}{2}\|x-y\|^2\right)$ defines

- a RKHS $\mathcal{H}_A$ where $A(x,y) = \langle \phi_A(x), \phi_A(y) \rangle_{\mathcal{H}_A}$
- a Gram operator $\mathcal{G}_A v = \int \langle v, \phi_A(z) \rangle_{\mathcal{H}_A} \phi_A(z)\, \mathrm{dP}(z)$

so that

$$\mu(x) = \int \langle \phi_A(x), \phi_A(z) \rangle_{\mathcal{H}_A}^2\, \mathrm{dP}(z) = \langle \mathcal{G}_A \phi_A(x), \phi_A(x) \rangle_{\mathcal{H}_A}$$

An estimator of $\bar{K}$

Given any estimator of $\mathcal{G}_A$, we can estimate

$$\mu(x) = \langle \mathcal{G}_A \phi_A(x), \phi_A(x) \rangle_{\mathcal{H}_A} \quad \simeq \hat{\mu}(x)$$

and thus we estimate $\bar{K}(x,y) = \mu(x)^{-1/2} K(x,y) \mu(y)^{-1/2}$ with

$$\hat{K}(x,y) = \hat{\mu}(x)^{-1/2} K(x,y) \hat{\mu}(y)^{-1/2}$$

An estimator of $\bar{K}_{2m}$

Proposition. With the previous notation,

$$\bar{K}_{2m}(x, y) = \langle \mathcal{G}_{\bar{K}}^{2m-1} \phi_{\bar{K}}(x), \phi_{\bar{K}}(y) \rangle_{\mathcal{H}}$$

where $\mathcal{G}_{\bar{K}} \phi_{\bar{K}}(x) = \int \bar{K}(x, z)\, \phi_{\bar{K}}(z)\, \mathrm{dP}(z)$.

We need to estimate $\mathcal{G}_{\bar{K}}$ that depends on $\bar{K}$ and P. Thus

$$\bar{K}_{2m}(x, y) = \langle \mathcal{G}_{\bar{K}}^{2m-1} \phi_{\bar{K}}(x), \phi_{\bar{K}}(y) \rangle_{\mathcal{H}} \quad \simeq \langle \mathcal{G}_{\hat{K}}^{2m-1} \phi_{\hat{K}}(x), \phi_{\hat{K}}(y) \rangle_{\mathcal{H}}$$

where $\mathcal{G}_{\hat{K}} \phi_{\hat{K}}(x) = \int \hat{K}(x, z)\, \phi_{\hat{K}}(z)\, \mathrm{dP}(z)$ (still unknown!)

An estimator of $\bar{K}_{2m}$

Given $\hat{\mathcal{Q}}$ any estimator of $\mathcal{G}_{\hat{K}}$ we obtain

$$\begin{aligned}\bar{K}_{2m}(x,y) &\simeq \langle \mathcal{G}_{\hat{K}}^{2m-1}\phi_{\hat{K}}(x), \phi_{\hat{K}}(y)\rangle_{\mathcal{H}} \\ &\simeq \langle \hat{\mathcal{Q}}^{2m-1}\phi_{\hat{K}}(x), \phi_{\hat{K}}(y)\rangle_{\mathcal{H}} =: \hat{K}_{2m}(x,y)\end{aligned}$$

where $\phi_{\hat{K}}(x) = \chi(x)\phi_{\bar{K}}(x)$ and $\chi(x) = \Big(\mu(x)/\hat{\mu}(x)\Big)^{1/2}$.

Recall: $\hat{K}_{2m}(x,y) = \langle \hat{\mathcal{Q}}^{2m-1} \phi_{\hat{K}}(x), \phi_{\hat{K}}(y) \rangle_{\mathcal{H}}$

where $\phi_{\hat{K}}(x) = \chi(x)\phi_{\bar{K}}(x)$ and $\chi(x) = \Big(\mu(x)/\hat{\mu}(x)\Big)^{1/2}$.

PROPOSITION. For any $x, y \in \mathrm{supp}(\mathrm{P})$,

$$
\begin{aligned}
&|\hat{K}_{2m}(x,y) - \bar{K}_{2m}(x,y)| \\
&\qquad \leq \frac{\max\{1, \|\chi\|_\infty\}^2}{\mu(x)^{1/2}\mu(y)^{1/2}} \left(\|\hat{\mathcal{Q}}^{2m-1} - \mathcal{G}_{\bar{K}}{}^{2m-1}\|_\infty + 2\|\chi - 1\|_\infty \right)
\end{aligned}
$$

and

$$
\|\hat{\mathcal{Q}}^{2m-1} - \mathcal{G}_{\bar{K}}{}^{2m-1}\|_\infty \leq (2m-1)\|\hat{\mathcal{Q}} - \mathcal{G}_{\bar{K}}\|_\infty \left(1 + \|\hat{\mathcal{Q}} - \mathcal{G}_{\bar{K}}\|_\infty\right)^{2m-2}
$$

Choice of m

Notation:

- let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$ be the eigenvalues of $\hat{Q}$
- let p the maximal number of classes

The number of iterations m is the solution of

$$\left(\frac{\hat{\lambda}_p}{\hat{\lambda}_1}\right)^m \simeq \frac{1}{100}$$

Note: p can be overestimated

Estimate of a Gram operator

Recall: We have seen that $|\hat{K}_{2m}(x,y) - \bar{K}_{2m}(x,y)|$ depends on

- $\chi(x) = \Big(\mu(x)/\hat{\mu}(x)\Big)^{1/2}$
- $\|\hat{\mathcal{Q}} - \mathcal{G}_{\bar{K}}\|_\infty$

Last step: Provide some estimate of Gram operators

Notation:

- Let K be a symmetric kernel
- let $\mathcal{H}$ be the RKHS defined by K

Goal: Estimate

$$\mathcal{G}v = \int \langle v, z \rangle_{\mathcal{H}} \, z \, \mathrm{dP}(z), \qquad v \in \mathcal{H}$$

from an i.i.d. sample $X_1, \ldots, X_n \in \mathcal{H} \sim \mathrm{P}$

Assume that $\mathrm{tr}(\mathcal{G}) < +\infty$

The empirical estimator

The classical empirical estimator is defined by

$$\bar{\mathcal{G}}v = \frac{1}{n}\sum_{i=1}^{n}\langle v, X_i\rangle \, X_i$$

Let

- $R = \max_{i=1,\dots,n} \|X_i\|$
- $X \in \mathcal{H}$ be a r.v. of law P.

Assume that

$$\kappa = \sup_{\theta} \frac{\mathbb{E}[\langle \theta, X\rangle^4]}{\mathbb{E}[\langle \theta, X\rangle^2]^2} < +\infty$$

THEOREM. With probability $\geq 1 - 2\epsilon$,

$$\|\mathcal{G} - \bar{\mathcal{G}}\|_\infty \leq 4 \max\{\|\mathcal{G}\|_\infty, \sigma\} \Big[B_*(\|\mathcal{G}\|_\infty) + \tau_*(\|\mathcal{G}\|_\infty) \Big] + \sigma$$

where

$$B_*(\|\mathcal{G}\|_\infty) = \sqrt{\frac{2.032(\kappa - 1)}{n} \left(\frac{0.73 \, \mathrm{tr}(\mathcal{G})}{\max\{\|\mathcal{G}\|_\infty, \sigma\}} + b + \log(\epsilon^{-1}) \right)} + \sqrt{\frac{98.5\kappa \, \mathrm{tr}(\mathcal{G})}{n \max\{\|\mathcal{G}\|_\infty, \sigma\}}},$$

$$\tau_*(\|\mathcal{G}\|_\infty) = \frac{0.86 \, R^4}{n(\kappa - 1) \max\{\|\mathcal{G}\|_\infty, \sigma\}^2} \left[\frac{0.73 \, \mathrm{tr}(\mathcal{G})}{\max\{\|\mathcal{G}\|_\infty, \sigma\}} + b + \log(\epsilon^{-1}) \right]$$

and $b \simeq \log(\log(n)) \leq 4.35$ if $n \leq 10^{20}$.

A MORE ROBUST ESTIMATOR

It is possible to use a PAC-Bayesian approach to construct a more robust estimator $\hat{\mathcal{G}}$ such that

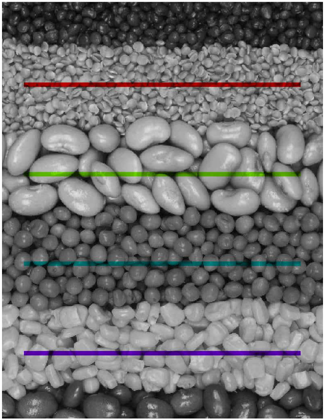
THEOREM. With probability $\geq 1 - 2\epsilon$,

$$\|\mathcal{G} - \hat{\mathcal{G}}\|_\infty \leq 4 \max\left\{\|\mathcal{G}\|_\infty, \sigma\right\} B_*(\|\mathcal{G}\|_\infty) + \sigma.$$

Note: In light tail situations, $\bar{\mathcal{G}}$ and $\hat{\mathcal{G}}$ behave in the same way

Work in progress: image classification

Test the algorithm in the setting of image classification



Work in progress: choice of β

Recall: we consider the Gaussian kernel

$$K(x,y) = K_\beta(x,y) = \exp\left(-\beta\|x-y\|^2\right)$$

The choice of β is based on the estimation of the trace of

$$L_\beta f(x) = \int K_\beta(x,z) f(z)\,\mathrm{dP}(z)$$

Note: Let $\lambda_1 \geq \lambda_2 \geq \ldots$ be the eigenvalues of L_β

$$\sum_i \lambda_i = \int K_\beta(x,x)\,\mathrm{dP}(x) = 1$$

$$\sum_i \lambda_i^2 = \int K_\beta(x,z)^2\,\mathrm{dP}(x)\mathrm{dP}(z) \leq 1$$

Work in progress: choice of β

Note:

$$F(\beta) = \int K_\beta(x,z)^2 \,\mathrm{dP}(x)\mathrm{dP}(z) \quad \begin{cases} \longrightarrow 1 & \text{if } \beta \to 0 \\ \longrightarrow 0 & \text{if } \beta \to \infty \end{cases}$$

Thus $F(\beta)$ controls the spread of the eigenvalues

$\longrightarrow$ we have to choose β sufficiently large

Goal: Find a way to calibrate β

Thank you