Spectral clustering, reproducing kernels and Markov chains with exponential transitions

Olivier Catoni

CREST – EXCESS,
CNRS UMR 9194
Olivier.Catoni@ensae.fr

Séminaire Parisien de Statistiques,

Institut Henri Poincaré

October 12, 2015

Joint work with Ilaria Giulini and Xiayang Zhou

Clustering a probability measure

A Markov chain approach

Consider a separable Hilbert space $\mathscr{X}$, the family of kernels

$$A_\beta(x, y) = \exp\bigl(-\beta \|x - y\|^2\bigr), \qquad x, y \in \mathscr{X},$$

and a probability measure $\mathrm{P} \in \mathscr{M}_+^1(\mathscr{X})$, with compact support $\mathrm{supp}(\mathrm{P})$.
Let $\mu_\beta(x) = \int A_\beta(x, y) \, \mathrm{dP}(y)$, $M_\beta(x, y) = \mu_\beta(x)^{-1} A_\beta(x, y)$, consider the Markov chain Z_m, $m \in \mathbb{N}$ with transitions

$$\frac{\mathrm{d}}{\mathrm{dP}} \mathbb{P}_{Z_{m+1} \,|\, Z_m = x}(y) = M_\beta(x, y), \qquad m \in \mathbb{N},$$

and the invariant measure Q with density $\frac{\mathrm{dQ}}{\mathrm{dP}}(x) = \mu_\beta(x)$.
Define the representation

$$R(x) = \frac{\mathrm{d}}{\mathrm{dQ}} \mathbb{P}_{Z_m \,|\, Z_0 = x} \in \mathbb{L}^2(\mathrm{Q}), \qquad x \in \mathrm{supp}(\mathrm{P}).$$

and the kernel

$$K_m(x,y) = \langle R(x), R(y) \rangle_{\mathbb{L}^2(\mathrm{Q})}.$$

Remark that, since $\mu(y)M(y,z) = \mu(z)M(z,y)$,

$$\begin{aligned} K_m(x,y) &= \int \frac{\mathrm{d}}{\mathrm{dQ}} \mathbb{P}_{Z_m | Z_0 = x}(z) \frac{\mathrm{d}}{\mathrm{dQ}} \mathbb{P}_{Z_{2m} | Z_m = z}(y) \mathrm{dQ}(z) \\ &= \frac{\mathrm{d}}{\mathrm{dQ}} \mathbb{P}_{Z_{2m} | Z_0 = x}(y). \end{aligned}$$

Cycle decomposition

Let $\mathscr{G}_T = \{(x, y) \in \text{supp}(\mathrm{P})^2; \|y - x\| < T\}$ and let $\mathscr{C}_T$ be the connected components of $\mathscr{G}_T$.

Conjecture : $\lim\limits_{\beta \to \infty} K_{\exp(\beta T^2)}(x, y) = \sum\limits_{C \in \mathscr{C}_T} Q(C)^{-1} \mathbb{1}(\{x, y\} \subset C).$

(True when $\text{supp}(P)$ is finite.)
Consequence : putting
$H_m(x, y) = K_m(x, x)^{-1/2} K_m(x, y) K_m(y, y)^{-1/2}$,

$$\lim_{\beta \to \infty} H_{\exp(\beta T^2)}(x, y) = \sum_{C \in \mathscr{C}_T} \mathbb{1}(\{x, y\} \subset C).$$

For suitable values of β and m, the kernel H_m should define a RKHS in which the points of $\text{supp}(\mathrm{P})$ are clustered around the vertices of a simplex.

Link with Gram operators

Consider the symmetric Laplacian kernel $L(x,y) = \mu(x)^{-1/2} A(x,y) \mu(y)^{-1/2}$ and the representation ϕ_A in the RKSH $\mathscr{H}$ defined by the kernel A, so that

$$\langle \phi_A(x), \phi_A(y) \rangle_{\mathscr{H}} = A(x,y), \qquad x, y \in \mathrm{supp}(\mathrm{P}).$$

Define the representation $\phi_L : \mathrm{supp}(\mathrm{P}) \longrightarrow \mathscr{H}$ as $\phi_L(x) = \mu(x)^{-1/2} \phi_A(x)$. It satisfies

$$\langle \phi_L(x), \phi_L(y) \rangle_{\mathscr{H}} = L(x,y).$$

Introduce the Gram operator of $\mathrm{P} \circ \phi_L^{-1} \in \mathscr{M}_+^1(\mathscr{H})$ defined as

$$\begin{aligned} \mathscr{G} : & \mathscr{H} \to \mathscr{H} \\ & u \mapsto \mathscr{G}(u) = \int \langle u, \phi_L(y) \rangle_{\mathscr{H}} \phi_L(y) \mathrm{dP}(y) \end{aligned}$$

Link with Gram operators

Remark that

$$\begin{aligned}
K_m(x,y) &= \frac{\mathrm{d}}{\mathrm{dQ}}\mathbb{P}_{Z_{2m}|Z_0=x}(y) = \mu(y)^{-1}\frac{\mathrm{d}}{\mathrm{dP}}\mathbb{P}_{Z_{2m}|Z_0=x}(y) \\
&= \mu(y)^{-1}\int M(x,z_1)\cdots M(z_{2m-1},y)\,\mathrm{dP}(z_1)\dots\mathrm{dP}(z_{2m-1}) \\
&= \mu(x)^{-1/2}\mu(y)^{-1/2}\int L(x,z_1)\cdots L(z_{2m-1},y)\,\mathrm{dP}(z_1)\dots\mathrm{dP}(z_{2m-1}) \\
&= \mu(x)^{-1/2}\mu(y)^{-1/2}\big\langle \mathscr{G}^{2m-1}\big(\phi_L(x)\big),\phi_L(y)\big\rangle_{\mathscr{H}} \\
&= \mu(x)^{-1}\mu(y)^{-1}\big\langle \mathscr{G}^{2m-1}\big(\phi_A(x)\big),\phi_A(y)\big\rangle_{\mathscr{H}}
\end{aligned}$$

Therefore $H_m(x,y) = \dfrac{\langle \phi_S(x),\phi_S(y)\rangle_{\mathscr{H}}}{\|\phi_S(x)\|_{\mathscr{H}}\,\|\phi_S(y)\|_{\mathscr{H}}}$, $x,y\in\mathrm{supp}(\mathrm{P})$,

where $\phi_S(x) = \mathscr{G}^{(2m-1)/2}\big(\phi_A(x)\big)$.

Clustering a statistical sample

Let $X_1, \ldots, X_n$ be n independent copies of $X \sim \mathrm{P}$. Consider some estimator $\widehat{\mathscr{G}}$ of $\mathscr{G}$, and the clustering algorithm based on

$$\widehat{H}_m(x,y) = \frac{\langle \widehat{\phi}_S(x), \widehat{\phi}_S(y) \rangle_{\mathscr{H}}}{\|\widehat{\phi}_S(x)\|_{\mathscr{H}} \|\widehat{\phi}_S(y)\|_{\mathscr{H}}}$$

where

$$\widehat{\phi}_S(x) = \widehat{\mathscr{G}}^{(2m-1)/2}\big(\phi_A(x)\big).$$

$$\begin{aligned} \Big| \langle \widehat{\phi}_S(x), \widehat{\phi}_S(y) \rangle_{\mathscr{H}} - \langle \phi_S(x), \phi_S(y) \rangle \Big| &\leq \big\| \widehat{\mathscr{G}}^{2m-1} - \mathscr{G}^{2m-1} \big\|_{\infty} \\ &\leq (2m-1) \big\| \widehat{\mathscr{G}} - \mathscr{G} \big\|_{\infty} \Big(1 + \big\| \widehat{\mathscr{G}} - \mathscr{G} \big\|_{\infty} \Big)^{2m-2} \end{aligned}$$

Comparison with the algorithm of Ng, Jordan, and Weiss

Consider the plugging estimator $\widehat{\mathscr{G}}$ obtained by replacing P with the empirical measure $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$. We get

$$\widehat{\mathscr{G}}(u) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n}A(X_i,X_j)\right)^{-1}\langle u,\phi_A(X_i)\rangle_{\mathscr{H}}\phi_A(X_i),$$

therefore, considering the vector $\overline{D}_i = \sum_{i=1}^{n}A(X_i,X_j)$ and the $n\times n$ matrices $\overline{A}_{i,j} = A(X_i,X_j)$, and $\overline{L}_{i,j} = \overline{D}_i^{-1/2}\overline{A}_{i,j}\overline{D}_j^{-1/2}$,

Comparison with the algorithm of Ng, Jordan, and Weiss

we obtain

$$\langle \widehat{\mathscr{G}}^{2m-1} \phi_A(X_i), \phi_A(X_j) \rangle_{\mathscr{H}} = \overline{D}_i^{1/2} \overline{L}_{i,j}^{2m} \overline{D}_j^{1/2},$$

$$\widehat{H}_m(X_i, X_j) = \left(\overline{L}_{i,i}^{2m}\right)^{-1/2} \overline{L}_{i,j}^{2m} \left(\overline{L}_{j,j}^{2m}\right)^{-1/2},$$

whereas the Ng, Jordan and Weiss algorithm can be described as based on the scalar product

$$\widehat{\widehat{H}}(X_i, X_j) = \overline{\overline{L}}_{i,i}^{-1/2} \overline{\overline{L}}_{i,j} \overline{\overline{L}}_{j,j}^{-1/2},$$

where, if we decompose $\overline{L} = U \, \mathbf{diag}(\lambda_1, \dots, \lambda_n) \, U^\top$, and introduce the orthogonal projection Π_r on the r first coordinates of $\mathbb{R}^n$, $\overline{\overline{L}} = U \Pi_r U^\top$.

Comparison with the algorithm of Ng, Jordan, and Weiss

Therefore, to derive our algorithm from the N. J. & W. algorithm, we have to replace the hard cut-off Π_r by the smooth cut-off $\mathbf{diag}(\lambda_1^{2m}, \ldots, \lambda_n^{2m})$ that does not assume that the number of classes r is known in advance. (Another minor difference is that N. J. & W. take $\overline{A}_{i,i} = 0$.)

Convergence bounds

Introduce

$$\hat{\widehat{\mathscr{G}}}(u) = \frac{1}{n}\sum_{i=1}^{n}\mu(X_i)^{-1}\langle u, \phi_A(X_i)\rangle_{\mathscr{H}}\phi_A(X_i)$$

and $\chi(x) = \dfrac{\mu(x)}{\widehat{\mu}(x)} - 1$, where $\widehat{\mu}(x) = \dfrac{1}{n}\sum_{i=1}^{n} A(x, X_i)$. As $\|\widehat{\mathscr{G}} - \hat{\widehat{\mathscr{G}}}\|_\infty \le (1 + \|\hat{\widehat{\mathscr{G}}} - \mathscr{G}\|_\infty)\|\chi\|_\infty$,

$$\|\widehat{\mathscr{G}} - \mathscr{G}\|_\infty \le \|\hat{\widehat{\mathscr{G}}} - \mathscr{G}\|_\infty\Big(1 + \|\chi\|_\infty\Big) + \|\chi\|_\infty.$$

Convergence bounds

Let $\phi_{A^{1/2}} : \mathrm{supp}(\mathrm{P}) \longrightarrow \mathscr{H}_{1/2}$ be the feature map defined by the kernel $A(x,y)^{1/2}$. We see that

$$\begin{aligned}\mu(x) &= \int \langle \phi_{A^{1/2}}(x), \phi_{A^{1/2}}(y) \rangle^2_{\mathscr{H}_{1/2}} \, \mathrm{dP}(y) \\ &= \langle \mathscr{G}_{1/2}\big(\phi_{A^{1/2}}(x)\big), \phi_{A^{1/2}}(x) \rangle_{\mathscr{H}_{1/2}},\end{aligned}$$

where

$$\mathscr{G}_{1/2}(u) = \int \langle u, \phi_{A^{1/2}}(y) \rangle \phi_{A^{1/2}}(y) \, \mathrm{dP}(y),$$

so that the estimation of $\mu(x)$ can be deduced from the estimation of the Gram operator $\mathscr{G}_{1/2}$.

Convergence bounds

Let $\mathscr{H}$ be some separable Hilbert space, $Z \in \mathscr{H}$ some random variable, and $Z_1, \ldots, Z_n$ a sample made of n independent copies of Z.

Let $\sup\Big\{\mathbb{E}\big(\langle\theta, Z\rangle^4\big); \theta \in \mathscr{H}, \mathbb{E}\big(\langle\theta, Z\rangle^2\big) \leq 1\Big\} \leq \kappa < \infty$,

$$\sigma = \frac{100\kappa\mathbb{E}(\|Z\|^2)}{n/128 - 4.35 - \log(\epsilon^{-1})},$$

$$\tau(t) = \frac{0.86 \max\{\|Z_i\|^4\}}{n(\kappa-1)\max\{t,\sigma\}^2}\left(\frac{0.73\,\mathbb{E}(\|Z\|^2)}{t} + 4.35 + \log(\epsilon^{-1})\right),$$

$$\zeta(t) = \sqrt{2.04(\kappa-1)\left(\frac{0.73\mathbb{E}(\|Z\|^2)}{\max\{t,\sigma\}} + 4.35 + \log(\epsilon^{-1})\right)} + \sqrt{\frac{98.5\kappa\mathbb{E}(\|Z\|^2)}{\max\{t,\sigma\}}},$$

$$B(t) = \frac{n^{-1/2}\zeta(t)}{1 - 4n^{-1/2}\zeta(t)}$$

Convergence bounds

Let $\overline{\mathbb{E}}(\langle\theta,Z\rangle^2) = \dfrac{1}{n}\sum_{i=1}^{n}\langle\theta,Z_i\rangle^2$.

With probability at least $1-2\epsilon$, for any $\theta\in\mathscr{H}$ such that $\|\theta\|=1$,

$$\left|\frac{\max\{\sigma,\overline{\mathbb{E}}(\langle\theta,Z\rangle^2)\}}{\max\{\sigma,\mathbb{E}(\langle\theta,Z\rangle^2)\}}-1\right| \leq B\Big(\mathbb{E}(\langle\theta,X\rangle^2)\Big) + \frac{\tau\Big(\mathbb{E}(\langle\theta,Z\rangle^2)\Big)}{\Big[1-\tau\Big(\mathbb{E}(\langle\theta,Z\rangle^2)\Big)\Big]_+\Big[1-B\Big(\mathbb{E}(\langle\theta,X\rangle^2)\Big)\Big]_+}.$$

Convergence bounds

Let us consider the Gram operator $\mathscr{G}(u) = \mathbb{E}\big(\langle u, Z\rangle Z\big)$ and its empirical estimate $\widehat{\mathscr{G}}(u) = \frac{1}{n}\sum_{i=1}^{n}\langle u, Z_i\rangle Z_i$. With probability at least $1-2\epsilon$,

$$\|\widehat{\mathscr{G}} - \mathscr{G}\|_\infty \leq \|\mathscr{G}\|_\infty B(\|\mathscr{G}\|_\infty) \\ + \inf_{\sigma>0}\left[\frac{\sigma\,\tau(\sigma)}{\big[1-\tau(\sigma)\big]_+\big[1-B(\sigma)\big]_+} + \sigma\right].$$

Remarking that $\inf_{x\in\mathrm{supp}(\mathrm{P})}\mu(x) \geq \sigma$ for n large enough, that $\|\phi_S(x)\|_{\mathscr{H}} \geq \mu(x)$ and putting everything together gives, for any fixed values of β and m, a finite sample deviation bound in $n^{-1/3}$ for

$$\sup_{x,y\in\mathrm{supp}(\mathrm{P})}\left|\widehat{H}_m(x,y) - H_m(x,y)\right|.$$

Choice of the scale parameter β

We can choose β by fixing the value of

$$F(\beta) = \int A_\beta(x,y)^2 \, \mathrm{dP}(x)\mathrm{dP}(y) = \sum_{i=1}^{\infty} \lambda_i^2,$$

$$\text{estimated by } \overline{F}(\beta) = \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} A_\beta(X_i, X_j)^2.$$

where $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of the principal component analysis of $\phi_A(x)$, that is the eigenvalues of the Gram operator $u \mapsto \mathbb{E}\big[\langle u, \phi_A(X)\rangle \phi_A(X)\big]$. Remark that λ_i defines a probability measure on the eigenvectors, since $\sum_{i=1}^{\infty} \lambda_i = \mathbb{E}\big(A_\beta(x,x)\big) = 1$, so that $F(\beta)$ controls the spread of this distribution, that is the spread of the initial representation $\phi_A(X)$ other different directions of the Hilbert space $\mathscr{H}$.

Choice of the number of iterations m

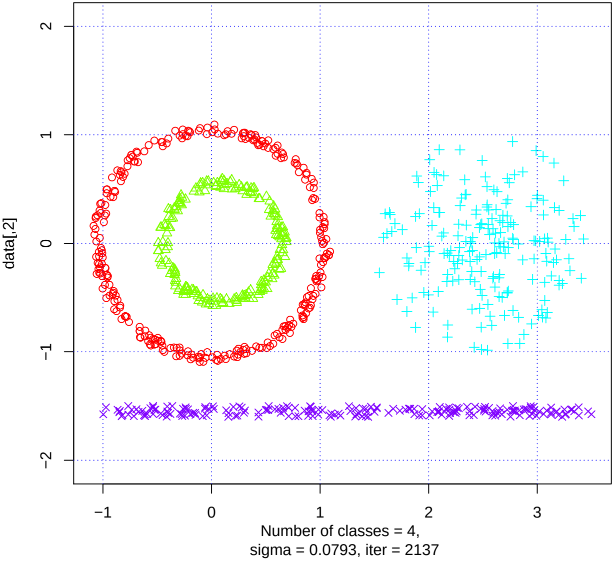
To choose the number of iterations m in practice, assuming that we know an upper bound r of the number of classes, we may fix the ratio

$$\rho = \left(\frac{\lambda_{r+1}}{\lambda_1} \right)^{2m}$$

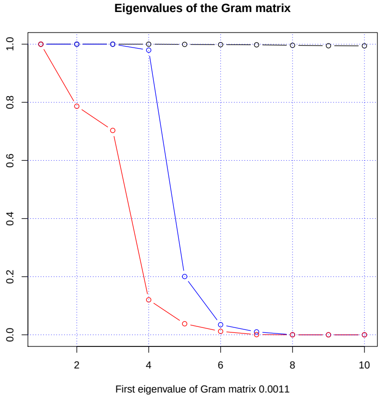
where this time, λ_i are the eigenvalues of the estimate of the Gram operator $u \mapsto \mathbb{E}\big[\langle u, \phi_L(X) \rangle \phi_L(X)\big]$. In the following simulations, we took $\rho = 1/100$, and the result does not seem to be very sensitive to the precise value of ρ, as long as it is small. We get

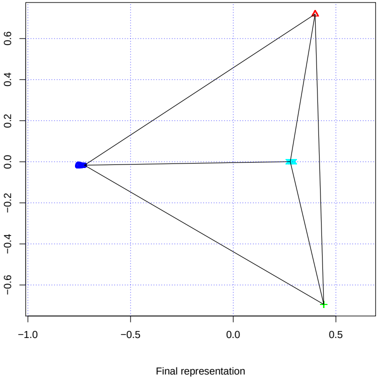
$$m = \left\lceil \frac{\log(\rho^{-1})}{2 \log(\lambda_1 / \lambda_{r+1})} \right\rceil.$$

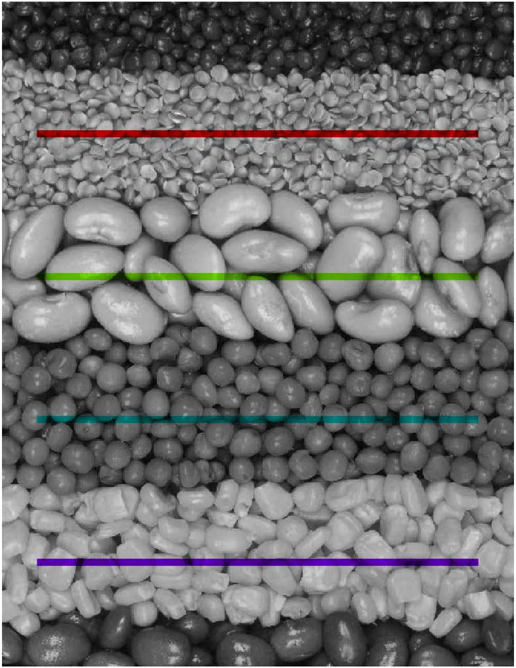
Examples of simulations



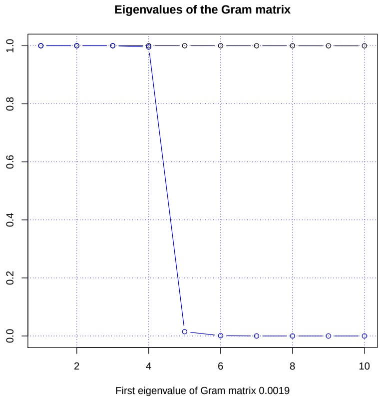
Examples of simulations

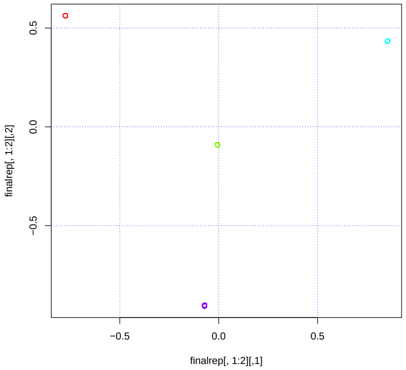






Examples of simulations





Examples of simulations