Estimation du nombre de clusters à l'aide de l'algorithme de clustering spectral

Bruno Pelletier, Pierre Pudlo

Estimation du nombre de clusters à l'aide de l'algorithme de clustering spectral

Bruno Pelletier & Pierre Pudlo

Université Montpellier II
Institut de Mathématiques et Modélisation de Montpellier
Place Eugène Bataillon ; Case Courrier 051 ; 34095 Montpellier CEDEX
{pelletier,ppudlo}@math.univ-montp2.fr

Résumé. Le but de la classification non supervisée est de partitionner les données en classes ou clusters d'objets relativement similaires. Dans ce contexte, l'estimation du nombre de clusters est une question centrale. L'algorithme de clustering spectral est une méthode récente pour détecter de tels clusters. Il est basé sur la décomposition spectrale d'une certaine matrice de similarité. L'étude des valeurs propres de cette matrice de similarité fournit une méthode prometteuse pour estimer le nombre de clusters. Nous présenterons des illustrations numériques sur différentes données simulées. Nous utilisons la définition d'un cluster due à Hartigan (1975), qui repose sur les ensembles de niveau de la densité. Nous commençons donc par ôter les points de l'échantillon où la densité estimée est faible, et qui gènent la détection des clusters.

Abstract. Unsupervised classification aims at partionning the dataset into groups or clusters of similar objects. In this context, estimating the number of clusters is an essential issue. The spectral clustering algorithm is a recent method to detect clusters. It is based on the spectral decomposition of some similarity matrix. The study of the eigenvalues of this similarity matrix gives us a promissing way to estimate the number of clusters. We will provide detailed numerical analysis with different simulated datasets. We use Hartigan (1975)'s definition of a cluster, based on the level sets of the density. Thus, we begin by removing from the dataset the points where the estimated density is low and that distrub clusters' detection.

Mots-clés. Statistique mathématique, fouille de données, classification non supervisée, ensemble de niveaux.

Les données sont modélisées par $\{X_i\}_{i\geq 1}$, une suite de vecteurs aléatoires de $\mathbf{R}^d$, tirés suivant une loi de densité $\varphi : \mathbf{R}^d \to \mathbf{R}$. Hartigan (1975) a proposé une définition non paramétrique d'un cluster : les t-clusters sont les composantes connexes de l'ensemble de niveau $\mathcal{L}(t) := \{x : \varphi(x) \geq t\}$ de la densité. Dans toute la suite, t est fixé et omis dans les notations. Introduisons également un estimateur non paramétrique de la densité $\hat{\varphi}_n$, basé sur l'échantillon $X_1, \ldots, X_n$.

Ainsi, de cet échantillon de taille n, on extrait les points qui tombent dans l'ensemble de niveau t de la densité estimée :

$$J(n) = \{j \leq n : \hat{\varphi}_n(X_j) \geq t\}.$$

Cette étape préliminaire permet d'avoir des clusters bien séparés, et facile leur détection, même si l'estimation de la densité n'est pas de très bonne qualité.

1. Clustering spectral

MATRICES DE SIMILARITÉ. Soit $\{k(\cdot,\cdot;\varepsilon)\}_{\varepsilon>0}$ une famille de fonctions de similarité sur $\mathbf{R}^d$, invariante par rotation :

$$k(x,y;\varepsilon) = h(||x-y||^2/\varepsilon),$$

où $h : \mathbf{R}_+ \to \mathbf{R}_+$ est une fonction. Les deux matrices ci-dessous, indexées par $J(n)\times J(n)$, sont appelées respectivement matrice de similarité non normalisée, et matrice de similarité normalisée :

$$\begin{aligned} \mathbb{K}_{i,j}(n,\varepsilon) &= h(||X_i - X_j||^2/\varepsilon) \quad \text{et} \\ \mathbb{Q}(n,\varepsilon) &= \mathbb{D}(n,\varepsilon)^{-1}\mathbb{K}(n,\varepsilon) \end{aligned}$$

où $\mathbb{D}(n,\varepsilon)$ est la matrice diagonale dont les entrées sont

$$\mathbb{D}_{i,i}(n,\varepsilon) = (\text{Card } J(n))^{-1} \sum_{j\in J(n)} \mathbb{K}_{i,j}(n,\varepsilon).$$

On introduit les hypothèses suivantes.
(H1) La densité φ est de classe C^2 et son gradient ne s'annule pas sur l'ensemble $\{x : \varphi(x) = t\}$. La densité φ, son gradient $\nabla\varphi$ et sa matrice hessienne $\text{Hess}(\varphi)$ sont uniformément bornés sur $\mathbf{R}^d$.
(H2) La fonction h est de classe C^2, son support est compact et il existe un ouvert sur lequel elle est minorée par un réel strictement positif.

GRAPHE PONDÉRÉ ASSOCIÉ. Soit $\mathcal{G}_n$ le graphe complet dont les sommets sont les points extraits, i.e. $\{X_j : j \in J(n)\}$. La matrice $\mathbb{K}(n,\varepsilon)$ permet d'affecter des poids (aléatoires) aux arêtes du graphe $\mathcal{G}_n$: le poids de l'arête $\{X_i; X_j\}$ est donnée par $\mathbb{K}_{i,j}(n,\varepsilon)$, c'est-à-dire la similarité entre X_i et X_j. Alors, $\mathbb{Q}(n,\varepsilon)$ est la matrice de transition de la marche au hasard réversible sur ce graphe, dont les probabilités de transition sont proportionnelles aux poids des arêtes.

L'algorithme de clustering spectral repose sur les valeurs propres proches de 1 de la matrice $\mathbb{Q}(n,\varepsilon)$, ainsi que les vecteurs propres associés, voir par exemple von Luxburg (2007).

Par exemple, prenons pour h la fonction indicatrice de $[0;1]$, et regardons le graphe pondéré défini ci-dessus. En ôtant les arêtes $\{X_i; X_j\}$ dont le poids $\mathbb{K}_{i,j}(n,\varepsilon)$ est nul, on obtient le graphe de Biau et al. (2007). L'estimateur du nombre de t-clusters proposé dans cet article est le nombre de composantes connexes du graphe $\mathcal{G}_n$. C'est exactement l'ordre de multiplicité de la valeur propre 1 pour la matrice $\mathbb{Q}(n,\varepsilon)$. En particulier, les

auteurs de l'article montrent un résultat de consistance quand $n \to \infty$, lorsque $\varepsilon = \varepsilon(n)$ tend vers 0.

Opérateurs bornés sur un espace fonctionnel. Pour comprendre les propriétés des matrices $\mathbb{Q}(n, \varepsilon)$ quand $n \to \infty$, il est utile d'introduire des opérateurs sur un espace de fonctions. Si $f : \mathcal{L} \to \mathbf{R}$ est une fonction C^1, on note $\|f\|_W = \|f\|_\infty + \|\nabla f\|_\infty$. Et on s'intéresse à l'espace de Banach $W(\mathcal{L})$ des fonctions $f : \mathcal{L} \to \mathbf{R}$ de classe C^1, de norme $\|f\|_W$ finie. Notons P_n la mesure empirique de l'échantillon extrait, i.e.

$$P_n = (\text{Card } J(n))^{-1} \sum_{j \in J(n)} \delta_{X_j}.$$

On introduit les opérateurs linéaires empiriques

$$\begin{aligned} Q_n(\varepsilon) : W(\mathcal{L}) &\to W(\mathcal{L}) \\ f(\cdot) &\mapsto \int f(y) q_n(\cdot, \mathrm{d}y; \varepsilon). \end{aligned}$$

où

$$d_n(x; \varepsilon) = \int k(x, y; \varepsilon) P_n(\mathrm{d}y), \quad \text{et} \quad q_n(x, \mathrm{d}y; \varepsilon) = \frac{k(x, y; \varepsilon)}{d_n(x; \varepsilon)} P_n(\mathrm{d}y).$$

Ces opérateurs bornés sont reliés aux matrices précédemment définies via la fonction d'évaluation $\pi_n : f \in W(\mathcal{L}) \mapsto \{f(X_j)\}_{j \in J(n)}$ par

$$\pi_n \circ Q_n(\varepsilon) = \mathbb{Q}(n, \varepsilon) \circ \pi_n. \tag{1}$$

2. Consistance

Opérateur limite. À partir de la densité φ, de l'ensemble de niveau $\mathcal{L}$, et de la famille de fonctions de similarité, nous pouvons définir une famille de noyaux de transition

$$q(x, \mathrm{d}y; \varepsilon) = \frac{k(x, y; \varepsilon)}{d(x; \varepsilon)} \varphi(y) \mathrm{d}y.$$

où $d(x; \varepsilon) = \varphi(\mathcal{L})^{-1} \int_{\mathcal{L}} k(x, y; \varepsilon) \varphi(y) dy$ et $\varphi(\mathcal{L})$ est la probabilité que X_1 soit dans $\mathcal{L}$, i.e. $\int_{\mathcal{L}} \varphi(x) \mathrm{d}x$. Ces noyaux de transition définissent des opérateurs bornés

$$\begin{aligned} Q(\varepsilon) : W(\mathcal{L}) &\to W(\mathcal{L}) \\ f(\cdot) &\mapsto \int_{\mathcal{L}} f(y) q(\cdot, \mathrm{d}y; \varepsilon). \end{aligned}$$

Résultats. À $\varepsilon > 0$ fixé, quand $n \to \infty$, von Luxburg et al. (2008) montrent que $Q_n(\varepsilon)$ convergent presque sûrement vers $Q(\varepsilon)$ au sens de la convergence collective compacte. Pelletier et Pudlo (2009) montrent la convergence ci-dessous.

Théorème 1. *Quand $n \to \infty$, $Q_n(\varepsilon)$ converge presque sûrement vers $Q(\varepsilon)$ en norme opérateur.*

On en déduit le résultat suivant.

Proposition 2. *(i) $Q(\varepsilon)$ est un opérateur compact sur $W(\mathcal{L})$. Son spectre est composé de 0, et de valeurs propres isolées $\lambda \neq 0$. (ii) Les valeurs propres de $Q(\varepsilon)$ sont réelles, et de module plus petit que 1. (iii) Le spectre de $Q_n(\varepsilon)$ est composé de 0 et des valeurs propres de $\mathbb{Q}(n, \varepsilon)$.*

Démonstration. On voit facilement que $Q_n(\varepsilon)$ est un opérateur de rang fini. Cette remarque, ainsi que le théorème 1, montre le résultat *(i)*. Le point *(ii)* est dû au fait que $Q(\varepsilon)$ définit une marche au hasard réversible sur $\mathcal{L}$. Et *(iii)* est essentiellement dû à (1), voir par exemple von Luxburg et al. (2008).□

Convergence des valeurs propres. Soit $1 = \lambda_{n,1} \geq \lambda_{n,2} \geq \ldots \geq -1$ les valeurs propres de $\mathbb{Q}(n, \varepsilon)$, où les valeurs propres sont répétées autant de fois que leur ordre de multiplicité. On peut retrouver le nombre de t-clusters en cherchant le nombre de valeurs propres proches de 1. En effet, Pelletier et Pudlo (2009) montre le résultat ci-dessous

Théorème 3. *Soit $1 - \eta$ la plus grande valeur propre de $Q(\varepsilon)$ qui soit différente de 1 et κ le nombre de composantes connexes de $\mathcal{L}$, i.e. le nombre de t-clusters. Si ε est suffisamment petit, alors*
(i) les premières valeurs propres $\lambda_{n,1}, \ldots, \lambda_{n,\kappa}$ convergent presque sûrement vers 1 quand $n \to \infty$;
(ii) pour tout $\ell > \kappa$, il existe un rang n_0 tel que, pour tout $n \geq n_0$, $1 - \lambda_{n,\ell} > \eta/2$.

La démonstration de ce résultat repose sur la convergence en norme oprateur : tout sous-ensenble fini de valeurs propres isolées dépend continument de l'opérateur.

Pour estimer κ, ce théorème suggère donc d'étudier la suite $\{1-\lambda_{n,\ell}\}_{\ell \geq 1}$ et de regarder pour quelle valeur de ℓ cette suite décolle de zéro. C'est ce que nous faisons dans la section suivante, sur des données simulées.

Extension du résultat lorsque (H2) n'est pas vérifiée. L'opérateur limite $Q(\epsilon)$ et les opérateurs empiriques $Q_n(\epsilon)$ dépendent de la fonction h qui sert à construire les fonctions de similarité, même si cette dépendance n'est pas explicite dans les notations utilisées. On peut vérifier que ces opérateurs dépendent continûment de cette fonction h, avec comme topologie celle induite par la norme $\|h\| = \|h\|_\infty + \|\nabla h\|_\infty$. Ainsi, lorsque la fonction h est suffisamment proche d'une fonction qui vérifie les hypothèses (H2), le théorème 3 s'adapte alors ainsi : pour n suffisamment grand, les κ premières valeurs propres de $Q_n(\varepsilon)$ sont encore proches de 1, alors que les suivantes sont à distance au moins $\eta/2$ de 1.

Dans la pratique, il est classique d'utiliser la fonction $h : t \mapsto \exp(-t^2/2)$, qui n'est pas une fonction à support compact. Mais on peut approcher h, aussi près que l'on veut,

par des fonctions qui vérifient (H2). Ainsi, l'argument précédent justifie que l'on obtienne le même type de comportements que sous l'hypothèse (H2).

3. Estimation du nombre de cluters : exemples numériques

Nous proposerons plusieurs études numérique. Ainsi, nous analyserons empiriquement différents types de situations, en particulier lorsque les clusters ont des formes non-convexes et ne se séparent pas par des hyperplans.

Nous nous concentrons sur l'étude de deux poins spécifiques : (1) comment la qualité de l'estimateur $\hat{\varphi}_n$ de la densité influe sur l'algorithme, (2) dans le cas où il y $\kappa = 1$ ou 2 clusters, nous produirons une étude numérique du comportement des deux premières valeurs propres $\lambda_{n,1}$ et $\lambda_{n,2}$ via des méthodes de Monte-Carlo.

Bibliographie

[1] Biau, G., Cadre, B. et Pelletier, B. (2007) A graph-based estimator of the number of clusters. *ESAIM: Probability and Statistics*, 11, 272–280.
[2] Hartigan, J. A. (1975) *Clustering algorithms*, Editions John Wiley & Sons.
[3] von Luxburg, U. (2007) A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4), 395–416.
[4] von Luxburg, U., Belkin, M. et Bousquet, O. (2008) Consistency of spectral clustering. *The Annals of Statistics*, 36, 555–586.
[5] Pelletier, B. et Pudlo, P. (2009) *Strong consistency of spectral clustering on level sets*, soumis.