On the choice of an appropriate bandwidth for modal clustering

Scelta di un appropriato parametro di lisciamento per il clustering modale

Alessandro Casa, José E. Chacón and Giovanna Menardi

Abstract In *modal* clustering framework groups are regarded as the domains of attraction of the modes of probability density function underlying the data. Operationally, to obtain a partition, a nonparametric density estimate is required and kernel density estimator is commonly considered. When resorting to these methods a relevant issue regards the selection of the smoothing parameter governing the shape of the density and hence possibly the modal structure. In this work we propose a criterion to choose the bandwidth, specifically tailored for the clustering problem since based on the minimization of the distance between a partition of the data induced by the kernel estimator and the whole-space partition induced by the true density.
Abstract Nell'ambito del clustering, l'approccio modale associa i gruppi ai domini di attrazione delle mode della funzione di densità sottostante i dati. L'individuazione dei gruppi richiede una stima non parametrica della densità, spesso basata su metodi kernel. Un problema rilevante, a tale scopo, riguarda la selezione del parametro di lisciamento che governa la forma della densità e, di conseguenza, la struttura modale. In questo lavoro si propone un criterio per la selezione del parametro di lisciamento, specificamente orientato al problema del clustering non parametrico e basato sulla minimizzazione di una misura di distanza tra la partizione dei dati indotta da uno stimatore kernel e la partizione dello spazio indotta dalla vera funzione di densità.

Key words: modal clustering, distance in measure, bandwidth selection, kernel density estimator

Alessandro Casa, Giovanna Menardi
Dipartimento di Scienze Statistiche, Università degli Studi di Padova
via C. Battisti 241, 35121, Padova; e-mail: casa@stat.unipd.it, menardi@stat.unipd.it
José E. Chacón
Departamento de Matemáticas, Universidad de Extremadura,
E-06006 Badajoz, Spain; e-mail: jechacon@unex.es

1 Introduction

Distance-based clustering is probably the most common approach to the unsupervised problem of obtaining a partition of a set of data into a number of groups. In spite of an intuitive interpretation and conceptual simplicity, this approach lacks of a 'ground truth', thus preventing the possibility to resort to formal statistical procedures. The density-based approach to cluster analysis overcomes such drawback by providing a formal definition of cluster, based on some specific features of the probability density function assumed to underlie the data. This approach has been developed following two distinct directions. The parametric one hinges on modelling the density function by means of a mixture distribution, where clusters are associated to the mixture components. Readers can refer to [3] for a recent review. This work focuses on the nonparametric - or *modal* - formulation, which is built on the concept of clusters as the "domains of attraction" of the modes of the density underlying the data [7]. The local maxima of the density are regarded to as the archetypes of the clusters, which are represented by the sorrounding regions (see Figure 1 for an illustration). These concepts have been translated into a formal definition of cluster by [1], resorting to notions and tools borrowed from Morse theory (see [2] for an introduction). Operationally, modal clustering has been pursued by two different strands of methods, both based on a preliminary nonparametric estimate of the density. The first strand looks directly for the modes of the estimated density and associates each cluster to the set of points along the steepest ascent path towards a mode, while the second one associates the clusters to the estimated density level sets of the sample space. For a detailed review see [4].

Modal clustering is appealing for several reasons. The outlined notion of cluster is close to the intuition of groups as dense regions; consistently, clusters are not constrained to have some particular pre-determined shape and resorting to nonparametric tools allow to mantain this flexibility. Also, since clusters are the domains of attraction of the density modes, the number of clusters is an intrinsic property of the data generator mechanism and its determination is itself an integral part of the estimation procedure. Furthermore, the existence of a formalized notion of cluster, based on the features of the density, allows to define an ideal population clustering goal, and frames the clustering problem into a standard inferential context.

Despite enjoying these relevant strenghts, when resorting to the nonparametric formulation, some criticalities have to be faced, mostly related to the estimation of the density underlying the data. Firstly obtaining nonparametric density estimates is usually computationally burdensome. This issue gets worse when working in high-dimensional spaces where nonparametric estimators suffer of the "curse of dimensionality". A relevant issue is that, regardless of the specific choice of the nonparametric density estimator, the selection of a smoothing parameter is required. Choosing this parameter turns out to be crucial since an inaccurate choice could lead to a misleading resulting estimate: too large values may lead to cover interesting structures, while too small values may lead to the appearance of spurious modes. If a kernel density estimator, the most common choice in the considered framework, is employed, the selection of the smoothing parameter is based on some reference

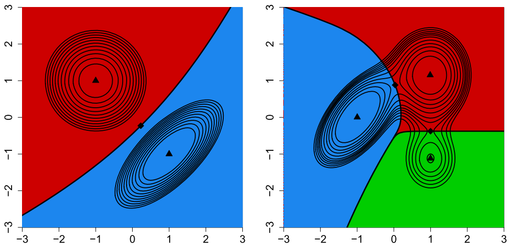

Fig. 1 Partitions induced by the modes of the density function in two examples of mixtures of bivariate normal densities.

rule or on criteria attempting to estimate properly the underlying density function. Even if these criteria have proved to produce appropriate clustering results in different situations, we believe that the clustering problem, being of a different nature with respect to the estimation of the density, would require a different rationale. In this work, a possible way to choose the optimal amount of smoothing, hinging on the specific clustering aim, is discussed. After formally defining a convenient loss function to measure the distance between data and population clustering, in the following we obtain its asymptotic expansion which, through a minimization, allows a focused selection of the smoothing parameter. Implications of this selection are finally discussed.

2 Kernel density estimation

According to the nonparametric formulation of density-based clustering the observed data $\mathscr{X} = \{x_i\}_{i=1,\dots,n}, x_i \in \mathbb{R}$, are supposed to be sampled from a random variable $\mathbb{X}$ with unknown density f. Note that, initially, we restrict our attention to the univariate case to allow a more rigorous treatment of the problem, with the intention to generalize the results to higher dimensional situations. To obtain a partition of the data adopting a nonparametric clustering perspective, regardless of its operational formulation (level set-based or mode seeking-based), an estimate $\hat{f}$ of the true density f is needed. In the rest of the paper we focus on the kernel density estimator

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) , \tag{1}$$

where K is the kernel, usually a symmetric density function, and $h > 0$ is the bandwidth, controlling the smoothness of the resulting estimate. A large value for h will tend to oversmooth the density, possibly covering some revelant features, while a small value will lead to an undersmoothed estimate where spurious modes (i.e. clusters) could arise.

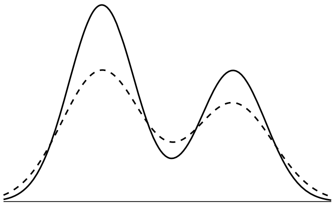

Fig. 2 Two density functions that are not close but induce exactly the same clustering.

The usual approach to select the bandwidth consists in minimizing some specific optimality criterion: the most used one is the *Mean Integrated Squared Error* (MISE)

$$MISE(h) = \mathbb{E}\int_{\mathbb{R}} \{\hat{f}_h(x) - f(x)\}^2 dx \tag{2}$$

which employs the $L2$-distance to assess the performance of the estimator. However, this expression does not have a tractable closed form and its asymptotic approximation -AMISE- is usually minimized, instead. Since both the MISE and the AMISE depend on the unknown f, different approaches to estimate them have been pursued, such as the ones based on *least square cross validation*, *biased cross validation* or *plug-in bandwidth selectors*. For a more comprehensive review and comparison see, e.g., [8].

3 The proposed selector

The selectors introduced in Section 2 are designed to choose the bandwidth so that it induces an appropriate estimate of the density. Nonetheless density estimation and clustering are two different problems with different requirements. It has been shown (e.g. [1]) that, even if two density functions are not really close, they can produce exactly the same partition of the data; see Figure 2 for an example. Furthermore, modal clustering strongly depends on some specific characteristics of the density function (the gradient for the mode seeking-based formulation and the high-density regions for the level set-based one) while, minimizing criteria such as the AMISE, an appropriate estimate is required in a global sense.

Our contribution hence finds its motivation in the lack of bandwidth selectors specifically conceived for nonparametric cluster analysis. In a similar fashion, even without specifically referring to clustering, [6] develop a plug-in type bandwidth selector that is appropriate for estimation of the highest density regions thus focusing on the modal regions, particularly relevant in the clustering formulation outlined above.

When density estimate is employed to subsequently partition the data, a more specifically tailored and appropriate performance measure than the $L2$-distance should be considered to select the amount of smoothing. Recalling [1], a natural way to quantify the performance of a data-based clustering is to consider the distance in measure between sets, where the measure has to be intended as the probability distribution with density f.

Formally, let $\mathscr{C} = \{C_1, \dots, C_r\}$ and $\mathscr{D} = \{D_1, \dots, D_r\}$ be two clusterings with the same number of groups r, their distance in measure can be measured by

$$d_1(\mathscr{C}, \mathscr{D}) = \underset{\nu \in \mathscr{P}}{min} \sum_{i=1}^{r} \mathbb{P}(C_i \Delta D_{\nu(i)}) \ , \tag{3}$$

where $\mathscr{P}$ is the set of the permutation of $\{1, \dots, r\}$, and $C\Delta D = (C \cap D^c) \cup (C^c \cap D)$. In the following we will actually consider

$$d_P(\mathscr{C}, \mathscr{D}) = \frac{1}{2} \underset{\nu \in \mathscr{P}}{min} \left\{ \sum_{i=1}^{r} \mathbb{P}(C_i \Delta D_{\nu(i)}) + \sum_{i=r+1}^{s} \mathbb{P}(D_{\nu(i)}) \right\} \ , \tag{4}$$

accounting for the intrinsic redundancy in (3) and for the possibility of having clustering with different number of groups. This distance can be seen as the minimal probability mass that needs to be moved to transform one clustering into the other.

Consider $\mathscr{C}_0$ as the ideal population clustering induced by the true density f and $\hat{\mathscr{C}}_n$ a data-based partition obtained from the sample $\mathscr{X}$. The idea is to quantify the quality of $\hat{\mathscr{C}}_n$ by measuring its distance in measure from $\mathscr{C}_0$. For large n, since the estimated number converges to the true number of clusters, it can be shown [1, Theorem 4.1] that (4) could be written as

$$d_P(\hat{\mathscr{C}}_n, \mathscr{C}_0) = \sum_{j=1}^{r-1} |F(\hat{m}_j) - F(m_j)| \ , \tag{5}$$

where F is the distribution function associated with f while $m_1, \dots, m_{r-1}$ and $\hat{m}_1, \dots, \hat{m}_{r-1}$ denote respectively the local minima (i.e. cluster boundaries in the univariate setting) of f and $\hat{f}$. Through two Taylor expansions, under some regularitiy conditions [1, Theorem 4.1], we obtain

$$|F(\hat{m}_j) - F(m_j)| \simeq \frac{f(m_j)}{f^{(2)}(m_j)} |\hat{f}^{(1)}(m_j)| \ , \tag{6}$$

where $f^{(j)}$ is the $j-th$ derivative of f. To obtain an asymptotic expression for (6) we have to study further the limit behavior of $\hat{f}^{(1)}(m_j)$. Considering that, if $h \to 0$ and $nh^{2r+1} \to \infty$, it is known that

$$(nh^{2r+1})^{1/2} \{\hat{f}^{(r)}(x) - K_h * f^{(r)}(x)\} \sim \mathscr{N}(0, R(K^{(r)})f(x)) \ , \tag{7}$$

where $R(K^{(r)}) = \int_{\mathbb{R}} (K^{(r)}(x))^2 dx$ and $(h * g)(x) = \int h(x-y)g(y)dy$. For a detailed treatment of the behaviour of kernel estimators at the critical points of a density, see [5]. Studying appropriately the bias term in (7), considering $r = 1$ and focusing on the local minima (i.e. $x = m_j$) we end up obtaining the limit distribution for the quantity of interest

$$n^{2/7}\hat{f}^{(1)}(m_j) \sim \mathscr{N}\left(\frac{\beta^2 f^{(3)}(m_j)\mu_2(K)}{2}, \frac{R(K^{(1)})f(m_j)}{\beta^3}\right) ,$$

where $\mu_2(K) = \int_{\mathbb{R}} x^2 K(x)dx$ and $\beta = n^{1/7}h$.
Thus, considering the property of a *folded normal distribution* and after some algebra, the asymptotic *expected distance in measure* (EDM) between a data clustering $\hat{\mathscr{C}}_n$ and the ideal population clustering $\mathscr{C}_0$ is given by

$$\begin{aligned}\mathbb{E}(d_P(\hat{\mathscr{C}}_n,\mathscr{C}_0)) &= \sum_{j=1}^{r-1}\frac{f(m_j)}{f^{(2)}(m_j)}\mathbb{E}(|\hat{f}^{(1)}(m_j)|)\\ &\simeq \sum_{j=1}^{r-1}\frac{f(m_j)}{f^{(2)}(m_j)}n^{-2/7}\{2\sigma^2\phi_\sigma(\mu)+\mu[1-2\Phi_\sigma(-\mu)]\} , \qquad (8)\end{aligned}$$

where ϕ_σ and Φ_σ denote respectively the density and the distribution function of a $\mathscr{N}(0,\sigma^2)$ random variable, $\mu = \beta^2 f^{(3)}(m_j)\mu_2(K)/2$ and $\sigma^2 = R(K^{(1)})f(m_j)/\beta^3$. The optimal bandwidth h_{d_P} for modal clustering purposes can be then obtained as $h_{d_P} = argmin_h \mathbb{E}(d_P(\hat{\mathscr{C}}_n,\mathscr{C}_0))$ by means of numerical optimization, after obtaining a suitable estimate of the unknown quantities $f(\cdot)$ and $f^{(2)}(\cdot)$ in the guise of the MISE/AMISE minimization. Another viable solution would be to work with a more manageable upper bound of (8) in order to obtain an explicit formula for the minimizer.

Further work is required to evaluate the performance of the proposed bandwidth selector as well as its comparison with some alternatives. There is much room for proceeding, and a multivariate extension of the discussed selector is needed to provide it with a concrete usability in more realistic settings.

References

1. Chacón, J.E.: A population background for nonparametric density-based clustering. Stat Sci, 30(4): 518-532 (2015).
2. Matsumoto, Y.: An introduction to Morse Theory. Amer. Math. Soc. (2002)
3. McNicholas, P.D.: Model-based clustering. J Classif, 33(3): 331-373 (2016).
4. Menardi, G.: A review on modal clustering. Int Stat Rev, 84(3): 413-433 (2016).
5. Romano, J.P.: On weak convergence and optimality of kernel density estimates of the mode. Ann Stat, 16(2):629-647 (1988).
6. Samworth, R.J & Wand, M.P.: Asymptotics and optimal bandwidth selection for highest density region estimation. Ann Stat, 38(3): 1767-1792 (2010).
7. Stuetzle, W.: Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. J Classif, 20(1): 25-47 (2003).
8. Wand, M.P. & Jones, M.C.: Kernel smoothing. Chapman & Hall (1994)