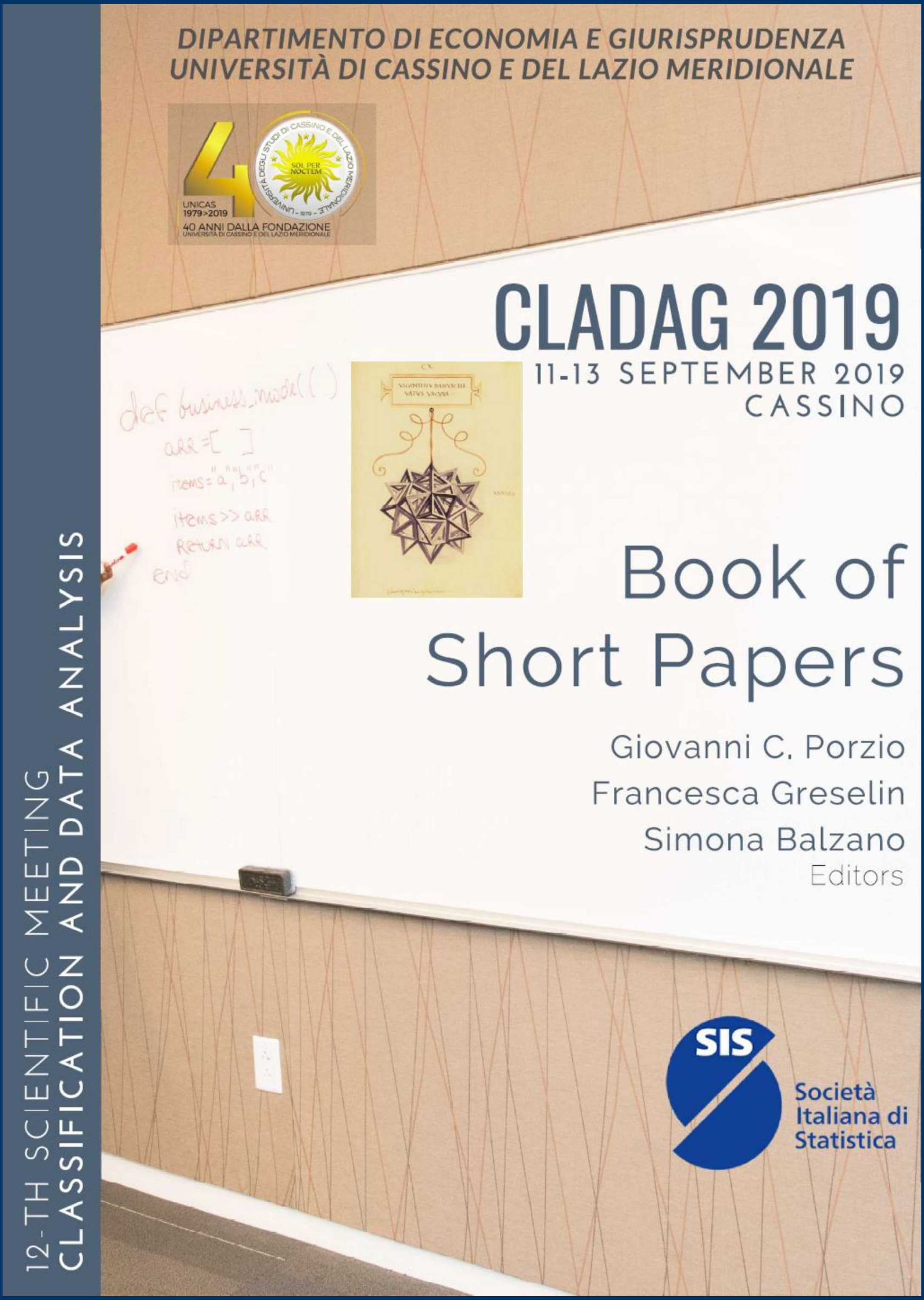
UNIVERSITÀ degli STUDI di CASSINO e del LAZIO MERIDIONALE

COLLANA SCIENTIFICA



DIPARTIMENTO DI ECONOMIA E GIURISPRUDENZA
UNIVERSITÀ DI CASSINO E DEL LAZIO MERIDIONALE

UNICAS 1979>2019
40 ANNI DALLA FONDAZIONE

CLADAG 2019

11-13 SEPTEMBER 2019
CASSINO

Book of Short Papers

Giovanni C. Porzio
Francesca Greselin
Simona Balzano
Editors

12-TH SCIENTIFIC MEETING
CLASSIFICATION AND DATA ANALYSIS

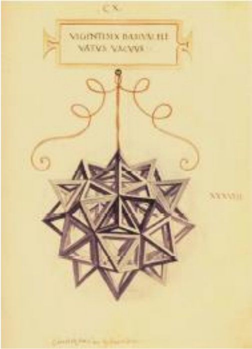
SIS
Società Italiana di Statistica

EUC
EDIZIONI UNIVERSITA' DI CASSINO

2019

Università di Cassino e del Lazio Meridionale
Centro Editoriale di Ateneo
Palazzo degli Studi Località Folcara, Cassino (FR), Italia

ISBN 978-88-8317-108-6



CLADAG 2019

Book of Short Papers

Giovanni C. Porzio

Francesca Greselin

Simona Balzano

Editors

2019

ASYMPTOTICS FOR BANDWIDTH SELECTION IN NONPARAMETRIC CLUSTERING

Alessandro Casa[1], José E. Chacón[2] and Giovanna Menardi[1]

[1] Dipartimento di Scienze Statistiche, Università degli studi di Padova,
(e-mail: casa@stat.unipd.it, menardi@stat.unipd.it)

[2] Departamento de Matemáticas, Universidad de Extremadura,
(e-mail: jechacon@unex.es)

ABSTRACT: In the framework of nonparametric clustering, clusters are defined as the domains of attraction of the modes of the density function assumed to underlie the data. To identify clusters, an estimate of the density is then needed, with kernel density estimator taking the lion's share. When resorting to these methods a fine tuning of the amount of smoothing, governing the modal structure of the density, is required. While thoroughly analyzed in the context of density estimation, this issue has been scarcely studied for clustering purposes. In this work the problem is addressed from an asymptotic perspective. A sensible distance among groupings is introduced and its asymptotic expression is derived and exploited in order to obtain a bandwidth selection procedure specifically tailored for nonparametric clustering.

KEYWORDS: modal clustering, kernel estimator, gradient bandwidth, mean shift.

1 Introduction

Density-based clustering pursues the aim of providing a statistical formalization to the widespread, yet ill-posed, problem of finding groups in a set of data. According to the nonparametric - or *modal* - formulation, clusters are seen as the domains of attraction of the modes of the density assumed to underlie the data, usually estimated by nonparametric methods. Linking the notion of cluster to the features of the underlying density frames the problem into a standard inferential context. As a consequence the concept of induced clustering, the partition implied by the characteristics of the density itself, is defined with the ideal population clustering being the one induced by the true density.

Regardless of the specific nonparametric density estimator adopted, the selection of a smoothing parameter is required. This choice represents a relevant issue since under- or over-smoothed estimates may lead to deceiving indications about the modal structure of the density, and hence about the number of groups.

The selection of the amount of smoothing is usually addressed via the minimization of some measure of distance which quantifies the discrepancy between the estimate and the target density. Standard references are the *Integrated Squared Error* and its expected value (MISE), or its asymptotic counterpart. While for the explicit task of density estimation, the distance criterion is usually selected to provide good estimates in a global sense, the same may be suboptimal in a clustering framework, where a focus on the local characteristics of the density would be more adequate to identify the modal regions.

The aim of this work is to address the problem of nonparametric density estimation for the final purpose of modal clustering. Density estimation is performed via the minimization of an appropriate metric relying on the comparison between the partitions induced by the estimated distribution and the true one, i.e. the ideal population clustering. A manageable asymptotic approximation of the considered metric is provided, which allows to define the optimal amount of smoothing for nonparametric clustering when a kernel estimator is adopted.

2 Optimal bandwidth for the asymptotic distance in measure

Let us assume that the observed data $\mathcal{X} = \{x_i\}_{i=1,\dots,n}$, are sampled from a random variable $\mathbb{X}$ with unknown density f. For mathematical tractability, in the following we restrict our attention to the univariate case, i.e. $x_i \in \mathbb{R}$.

A standard choice to estimate f is to resort to the kernel estimator

$$\hat{f}_h(x) = (1/nh)\sum_{i=1}^{n} K[(x-x_i)/h]$$

where K is a kernel function and $h > 0$ is the bandwidth which controls for the amount of smoothing and, then, the modal structure.

To tailor the choice of h for clustering purposes, we consider the *distance in measure* (Chacón, 2015) between $\hat{\mathcal{C}}_h = \{\hat{C}_1,\dots,\hat{C}_r\}$, the clustering induced by $\hat{f}_h$, and $\mathcal{C}_0 = \{C_{0,1},\dots,C_{0,s}\}$, the ideal population one, induced by the true f:

$$d(\hat{\mathcal{C}}_h, \mathcal{C}_0) = \frac{1}{2}\min_{\sigma\in\mathcal{P}_s}\left\{\sum_{i=1}^{r}\mathbb{P}(\hat{C}_i\Delta C_{0,\sigma(i)}) + \sum_{i=r+1}^{s}\mathbb{P}(C_{0,\sigma(i)})\right\}, \quad (1)$$

where $\mathcal{P}_s$ is the set of permutations of $\{1,\dots,s\}$, $C\Delta C_0 = (C\cap C_0^c)\cup(C^c\cap C_0)$ and with possibly $r \le s$. This distance can be seen as the minimal probability mass that needs to be moved to transform one clustering into the other. Being sample-specific, the distance in measure is subject to a random variability.

Hence, the *Expected Distance in Measure* $\mathrm{EDM(h)} = \mathbb{E}[d(\hat{\mathcal{C}}_h, \mathcal{C}_0)]$ is alternatively considered as a non-stochastic error distance. The optimal bandwidth is then defined as $h_{EDM} = \mathrm{argmin}_{\mathrm{h}>0}\mathrm{EDM(h)}$.

Under some regularity assumptions, it can be proved (Casa *et al.*, 2019) that $\mathrm{EDM(h)}$ is asymptotically equivalent to

$$\mathrm{AEDM}(h) = \sum_{j=1}^{r-1} \frac{f(m_j)}{f^{(2)}(m_j)} \psi\Big(\frac{1}{2}\mu_2(K) f^{(3)}(m_j) h^2, R(K^{(1)}) f(m_j)(nh^3)^{-1}\Big) \quad (2)$$

where $\psi(\mu, \sigma^2) = (2/\pi)^{1/2}\Big\{\sigma e^{-\mu^2/(2\sigma^2)} + |\mu| \int_0^{|\mu|/\sigma} e^{-z^2/2} dz\Big\}$, m_j is the j^{th} local minimum of f, $g^{(l)}$ denotes the l^{th} derivative of a function g, $\mu_2(K) = \int_{-\infty}^{\infty} x^2 K(x) dx$, and $R(K^{(1)}) = \int_{-\infty}^{\infty} K^{(1)}(x)^2 dx$.

Since neither the $\mathrm{EDM(h)}$ nor the $\mathrm{AEDM(h)}$ admit an explicit representation of their minima, the idea is to rely on a tight upper bound. The study of the behaviour of $\psi(\cdot,\cdot)$ allows us to introduce two different upper bounds, whose minimizers can be computed explicitly. It follows that

$$\begin{aligned}
h_{AB1} &= \left(\frac{9R(K^{(1)}) \left(\sum_{j=1}^{r-1} f(m_j)^{3/2} / f^{(2)}(m_j) \right)^2}{2\pi \mu_2(K)^2 \left(\sum_{j=1}^{r-1} f(m_j) |f^{(3)}(m_j)| / f^{(2)}(m_j) \right)^2} \right)^{1/7} n^{-1/7} \\
h_{AB2} &= \left(\frac{24R(K^{(1)}) \sum_{j=1}^{r-1} f(m_j)^{3/2} / f^{(2)}(m_j)}{11\mu_2(K)^2 \sum_{j=1}^{r-1} f(m_j)^{1/2} f^{(3)}(m_j)^2 / f^{(2)}(m_j)} \right)^{1/7} n^{-1/7} .
\end{aligned}$$

Note that, since the derived bandwidths are depending on some unknown quantities, from an operational point of view we need to resort to plug-in strategies.

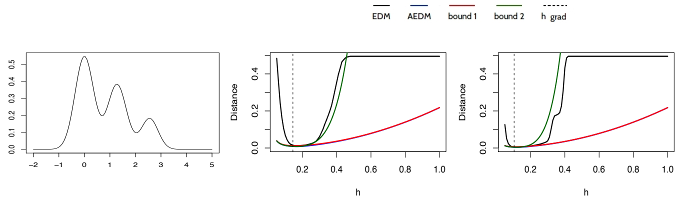
3 Some results and conclusions

In this section we present an excerpt of the numerical results obtained in one-dimensional setting in order to evaluate the performances of the proposed selectors as well as the quality of the introduced asymptotic approximations.

The top panel of Table 1 shows the quality of the derived approximations to the EDM, as a function of the bandwidth, when all the involved quantities are known. The approximations improve as the sample size increases and they appear to behave satisfactorily especially around the value of h minimizing the EDM. In the bottom panel we can see the results, in terms of EDM, of the data-based bandwidth selectors over $B = 1000$ synthetic samples, along with

Table 1. *Top panel: true density (left); EDM, AEDM and the bounds vs h for $n = 1000, 10000$ (middle and right panels). The vertical dashed line is associated to the gradient bandwidth. Bottom panel: EDM estimates (and standard errors) at the optimum h according to the AEDM, the two bounds, and the gradient bandwidth.*

	n=1000	n=10000
Density	EDM approximation	



	DM estimate	
$\hat{h}_{AEDM}$	0.015 (0.018)	0.005 (0.003)
$\hat{h}_{AB1}$	0.013 (0.010)	0.005 (0.003)
$\hat{h}_{AB2}$	0.014 (0.011)	0.005 (0.003)
$\hat{h}_{GRAD}$	0.013 (0.009)	0.005 (0.003)

the performances of the gradient bandwidth, representing a sensible competitor in this framework, obtained via MISE minimization. The proposed selectors $\hat{h}_{AB1}$ and $\hat{h}_{AB2}$ led to more accurate clusterings than h_{AEDM}, with a slight preference for the former. The gradient-based bandwidth, in turn, not only produces competitive results, but its Monte Carlo average distance in measure appears lower than the one produced by the asymptotic EDM minimizers. In fact, a deeper insight into the standard errors of the obtained distances shows that $\hat{h}_{AEDM}$, as well as $\hat{h}_{AB1}$ and $\hat{h}_{AB2}$, produce more variable results, due to a higher sensitivity of the minimizers to the plugged in pilot estimates.

For a complete exposition of the results, alongside with a multivariate generalization, see Casa *et al.*, 2019.

References

Casa, A., Chacón, J.E., & Menardi, G. 2019. Modal clustering asymptotics with applications to bandwidth selection. *arXiv preprint arXiv:1901.07300*.

Chacón, J.E. 2015. A population background for nonparametric density-based clustering. *Statistical Science*, **30**(4), 518–532.