Asymptotics for bandwidth selection in nonparametric clustering

12th Scientific Meeting - Classification and Data Analysis Group

Alessandro Casa[1]

José E. Chacòn[2] and Giovanna Menardi[1]

Università degli Studi di Padova[1]
Universidad de Extremadura[2]

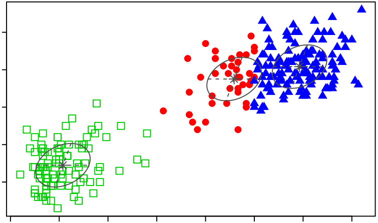
casa@stat.unipd.it

Cassino, 12th September 2019

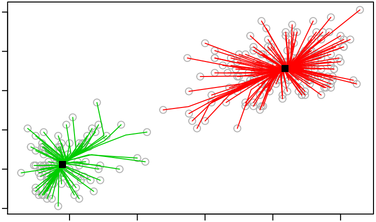
Density-based clustering

- **Framework**: density-based clustering
  - Statistical formalization of the clustering problem;
  - Identifies a partition of the sample space and not only of the data points;
  - Definition of cluster based on features of the density underlying the data.
- Developed according to two different paradigms:

Model-based

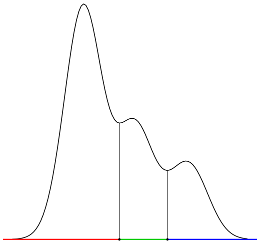


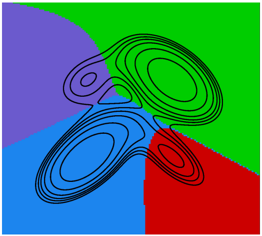
Modal



Modal clustering

- **Nonparametric formulation**: clusters correspond to the domains of attraction of the modes of the density;
- Operationally required:
  - A density estimate;
  - A method to locate modal regions of the density.

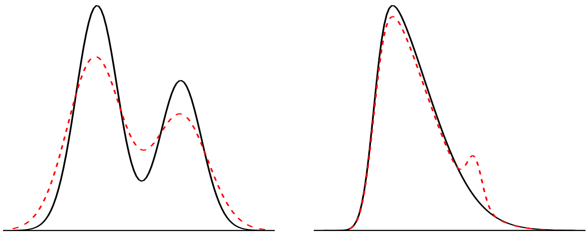




Modal clustering

- Kernel density estimator (KDE) is usually considered to obtain a nonparametric density estimate;
- Let $\mathcal{X} = \{x_i\}_{i=1,\ldots,n}$, $x_i \in \mathbb{R}$ be a sample from a r.v. X with density f. Define KDE as
$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$
  - $K(\cdot)$ is the kernel function;
  - h is the bandwidth.
- Bandwidth h controls the amount of smoothing:
$\longrightarrow$ if wrongly chosen could lead to cover interesting features or highlight spurious ones.

Aim and contribution



- Usual bandwidth selectors aim to obtain appropriate estimates of the density from a global perspective
  BUT $\longrightarrow$ Density estimation and clustering are different problems with different requirements;
- **Contribution**: propose an asymptotically optimal modal clustering-oriented bandwidth by minimizing a measure of distance.

Distance between clusterings

- **Distance in measure** (DM) between clusterings $\mathcal{C}_0 = \{C_1, \ldots, C_r\}$ and $\hat{\mathcal{C}}_n = \{\hat{C}_r, \ldots, \hat{C}_r\}$

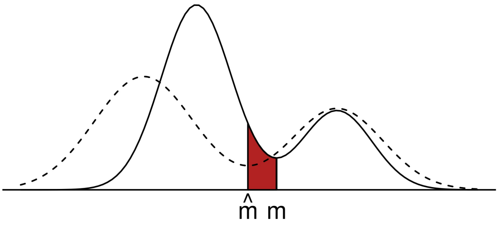
$$d(\mathcal{C}_0, \hat{\mathcal{C}}_n) = \frac{1}{2} \min_{\sigma \in \mathcal{P}_r} \left\{ \sum_{j=1}^{r} \mathbb{P}(C_j \Delta \hat{C}_{\sigma(j)}) \right\}$$

  - $\mathcal{P}_r$: set of permutations of $\{1, \ldots, r\}$;
  - $C \Delta \hat{C} = (C \cap \hat{C}^c) \cup (C^c \cap \hat{C})$;

- Easily generalizable for clusterings with different number of groups;
- Optimal bandwidth:

$$h_{opt} = \operatorname*{argmin}_h \mathbb{E}[d(\mathcal{C}_0, \hat{\mathcal{C}}_n)]$$

Distance between clusterings



- **Interpretation**: minimal probability mass that has to be moved to transform one clustering to the other.
- For n large DM is

$$d(\mathcal{C}_0, \hat{\mathcal{C}}_n) = \sum_{j=1}^{r-1} |F(\hat{m}_j) - F(m_j)| \ ,$$

  - F cdf of f;
  - $\{m_j\}$ and $\{\hat{m}_j\} = \{\hat{m}_j(h)\}$ minima of f and $\hat{f}_h$, $j = 1, \ldots, r$.

Distance between clusterings

- **Asymptotic Expected Distance in Measure** (AEDM)

$$\begin{aligned}\mathbb{E}(d_P(C_0,\hat{C}_n)) &\simeq \sum_{j=1}^{r-1}\frac{f(m_j)}{f^{(2)}(m_j)}n^{-2/7}\psi(\mu,\sigma)\\ &\simeq \sum_{j=1}^{r-1}\frac{f(m_j)}{f^{(2)}(m_j)}n^{-2/7}\{2\sigma^2\phi_\sigma(\mu)+\mu[1-2\Phi_\sigma(-\mu)]\}\end{aligned}$$

  - $\mu=\mu(h)=n^{2/7}h^2f^{(3)}(m_j)\mu_2(K)/2$;
  - $\sigma^2=\sigma^2(h)=R(K^{(1)})f(m_j)/(n^{3/7}h^3)$;
  - $\phi_\sigma(\cdot)$ and $\Phi_\sigma(\cdot)$ density and cdf of $\mathcal{N}(0,\sigma^2)$;
  - $R(K^{(r)})=\int_{\mathbb{R}}(K^{(r)}(x))^2dx$ and $\mu_2(K)=\int_{\mathbb{R}}x^2K(x)dx$;
- Two summands behaving as *Integrated Squared Bias* and *Integrated Variance* for the MISE;
- Problems:
  - Dependence on f, $f^{(2)}$ and $f^{(3)}$;
  - Not available a closed form for h_{opt}.

Optimal bandwidth

- Introduce two different upper bounds:

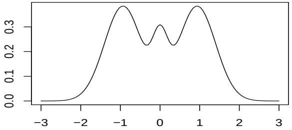
$$\begin{aligned}\psi(\mu,\sigma) &\leq \sigma\sqrt{2/\pi} + |\mu| \\ \psi(\mu,\sigma) &\leq \sigma\sqrt{2/\pi} + \mu^2/(\sigma\sqrt{2\pi}))\end{aligned}$$

- Minimization of the bounded versions of the AEDM leads to

$$h_{opt,b_1} = \left(\frac{3R(K')^{1/2} \sum_{j=1}^{r-1} \frac{f(m_j)^{3/2}}{f^{(2)}(m_j)}}{\mu_2(K)\sqrt{2\pi} \sum_{j=1}^{r-1} \frac{f(m_j)|f^{(3)}(m_j)|}{f^{(2)}(m_j)}} \right)^{2/7} n^{-1/7}$$

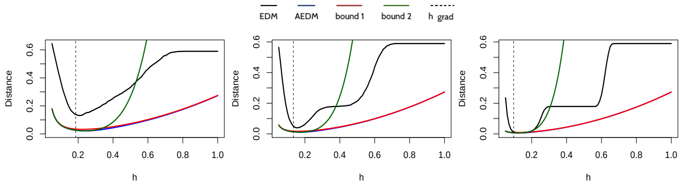
$$h_{opt,b_2} = \left(\frac{24R(K') \sum_{j=1}^{r-1} \frac{f(m_j)^{3/2}}{f^{(2)}(m_j)}}{11\mu_2^2(K) \sum_{j=1}^{r-1} \frac{f(m_j)^{1/2} f^{(3)}(m_j)^2}{f^{(2)}(m_j)}} \right)^{1/7} n^{-1/7} .$$

Numerical results



	n =100	n=1000	n=10000

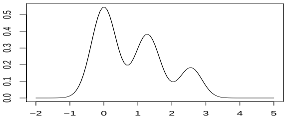
EDM approximation



DM estimate

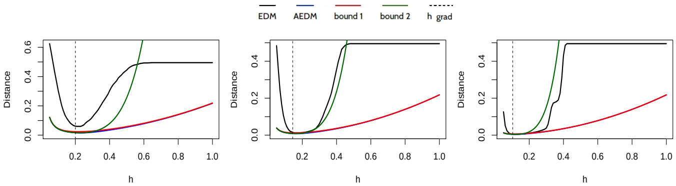
h_{aedm}	0.324 (0.200)	0.061 (0.070)	0.010 (0.016)
h_{bound1}	0.301 (0.195)	0.053 (0.066)	0.011 (0.017)
h_{bound2}	0.318 (0.199)	0.058 (0.069)	0.010 (0.016)
h_{grad}	0.256 (0.159)	0.092 (0.076)	0.008 (0.005)
$\%\hat{r} = r$	2.8	58.0	100.0

Numerical results



	n =100	n=1000	n=10000

EDM approximation



DM estimate			
h_{aedm}	0.160 (0.175)	0.017 (0.034)	0.006 (0.007)
h_{bound1}	0.144 (0.169)	0.017 (0.030)	0.006 (0.007)
h_{bound2}	0.157 (0.174)	0.017 (0.034)	0.006 (0.007)
h_{grad}	0.179 (0.158)	0.013 (0.009)	0.005 (0.003)
$\%\hat{r} = r$	42.7	99.7	100.0

Concluding remarks and future work

- We have obtained a good asymptotic approximation to the EDM
  $\longrightarrow$ problems in data-driven procedures due to plug-in strategies;
- Gradient bandwidth seems to be an appropriate choice when resorting to modal clustering;
- Directions of **future work**:
  - Study bandwidth selection problem for mode estimation and the estimation of the number of modes
    $\longrightarrow$ possibly beneficial in the plug-in steps.

Relevant references

1. Chacón, J.E. (2015). *A population background for nonparametric density-based clustering*. Statistical Science, 30(4).
2. Devroye, L. & Györfi, L. (1985). *Nonparametric density estimation: the L_1 view*. Wiley & Sons.
3. Grund, B. and Hall, P. (1995). *On the minimisation of the ℓ^p error in mode estimation.* The Annals of Statistics, **23**(6), 2264–2284.
4. Menardi, G. (2016). *A review on modal clustering*. International Statistical Review, 84(3).
5. Romano, J.P. (1988). *On weak convergence and optimality of kernel density estimates of the mode*. The Annals of Statistics, 16(2).
6. Samworth, R.J. & Wand, M.P. (2010). *Asymptotics and optimal bandwidth selection for highest density region estimation estimation*. The Annals of Statistics, 38(3).
7. Wand, M.P. & Jones, M.C. (1994). *Kernel smoothing*. Chapman & Hall.