On the choice of an appropriate bandwidth for modal clustering

49th Scientific Meeting of the Italian Statistical Society

Alessandro Casa[1]

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

Università degli Studi di Padova[1]

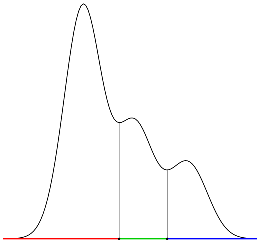
Universidad de Extremadura[2]

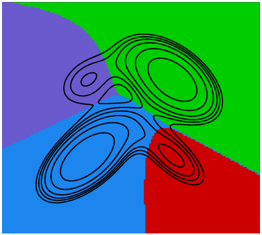
casa@stat.unipd.it

21st June 2018

Density-based clustering

- **Framework**: density-based clustering $\longrightarrow$ definition of cluster based on features of the density underlying the data.
- **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.





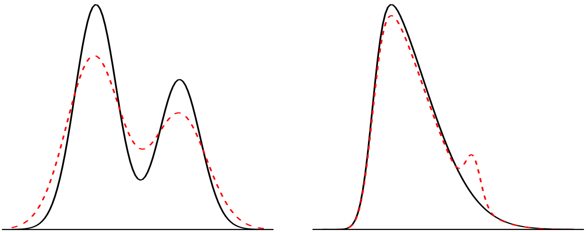
Modal clustering

- Kernel density estimator (KDE) is usually considered to obtain a nonparametric density estimate;
- Let $\mathcal{X} = \{x_i\}_{i=1,\dots,n}$, $x_i \in \mathbb{R}$ be a sample from a r.v. X with density f. Define KDE as

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

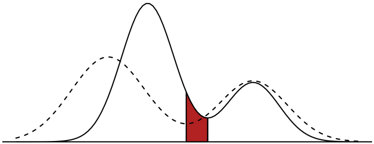
  - $K(\cdot)$ is the kernel function;
  - h is the bandwidth.
- Bandwidth h controls the amount of smoothing: $\rightarrow$ if wrongly chosen could lead to cover interesting features or highlight spurious ones.

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 $\mathcal{C}_0$ and $\hat{\mathcal{C}}_n$ for large n

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

  with F cdf of f, $\{m_j\}$ and $\{\hat{m}_j\} = \{\hat{m}_j(h)\}$ minima of f and $\hat{f}_h$.

- **Interpretation**: minimal probability mass that has to be moved to transform one clustering to the other;
- Optimal bandwidth:

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

.

Distance between clusterings

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

$$\mathbb{E}(d_P(\mathcal{C}_0, \hat{\mathcal{C}}_n)) \simeq \sum_{j=1}^{r-1} \frac{f(m_j)}{f^{(2)}(m_j)} n^{-2/7} \psi(\mu, \sigma)$$

$$\simeq \sum_{j=1}^{r-1} \frac{f(m_j)}{f^{(2)}(m_j)} n^{-2/7} \{2\sigma^2 \phi_\sigma(\mu) + \mu[1 - 2\Phi_\sigma(-\mu)]\}$$

  - $\mu = \mu(h) = n^{2/7} h^2 f^{(3)}(m_j) \mu_2(K)/2$;
  - $\sigma^2 = \sigma^2(h) = R(K^{(1)}) f(m_j)/(n^{3/7} h^3)$;
  - $\phi_\sigma(\cdot)$ and $\Phi_\sigma(\cdot)$ density and cdf of $\mathcal{N}(0, \sigma^2)$;
  - $R(K^{(r)}) = \int_{\mathbb{R}} (K^{(r)}(x))^2 dx$ and $\mu_2(K) = \int_{\mathbb{R}} x^2 K(x) dx$;
- 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

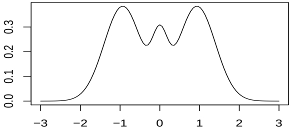
- Introduce two different upper bounds:

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

- Minimization of the bounded versions of the AEDM leads to

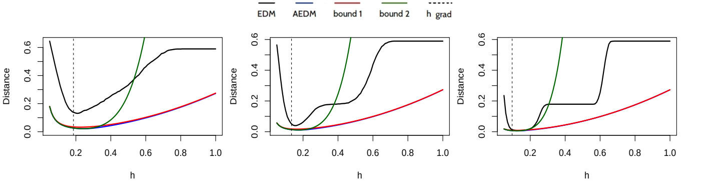
$$\begin{aligned}h_{opt,b_1} &= \left(\frac{3R(K')^{1/2}\sum_{j=1}^{r-1}\frac{f(m_j)^{3/2}}{f^{(2)}(m_j)}}{\mu_2(K)\sqrt{2\pi}\sum_{j=1}^{r-1}\frac{f(m_j)|f^{(3)}(m_j)|}{f^{(2)}(m_j)}}\right)^{2/7} n^{-1/7} \\ h_{opt,b_2} &= \left(\frac{24R(K')\sum_{j=1}^{r-1}\frac{f(m_j)^{3/2}}{f^{(2)}(m_j)}}{11\mu_2^2(K)\sum_{j=1}^{r-1}\frac{f(m_j)^{1/2}f^{(3)}(m_j)^2}{f^{(2)}(m_j)}}\right)^{1/7} n^{-1/7}\ .\end{aligned}$$

Numerical results



	n =100	n=1000	n=10000

EDM approximation



DM estimate

	n =100	n=1000	n=10000
h_{aedm}	0.265 (0.173)	0.060 (0.070)	0.010 (0.016)
h_{bound1}	0.237 (0.155)	0.053 (0.066)	0.011 (0.017)
h_{bound2}	0.257 (0.168)	0.058 (0.069)	0.010 (0.016)
h_{grad}	0.183 (0.031)	0.092 (0.075)	0.008 (0.005)

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**:
    - Find suitable nonparametric alternatives to better locate the minima and to estimate the features of the density at those points $\longrightarrow$ local bandwidth ?
    - Extend the results in multivariate situations.

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. Menardi G. (2016). *A review on modal clustering*. International Statistical Review, 84(3).
4. Romano, J.P. (1988). *On weak convergence and optimality of kernel density estimates of the mode*. The Annals of Statistics, 16(2).
5. Samworth, R.J. & Wand, M.P. (2010). *Asymptotics and optimal bandwidth selection for highest density region estimation estimation*. The Annals of Statistics, 38(3).
6. Wand, M.P. & Jones, M.C. (1994). *Kernel smoothing*. Chapman & Hall.