Combining local and global smoothing in non-parametric density estimation

Adelchi Azzalini
Dipartimento di Scienze Statistiche
Università di Padova, Italia

Department of Statistical Sciences, University of Cape Town
22nd July 2019

The broad context

Density estimation: *very* schematically

$$\text{density estimation} = \begin{cases} \text{parametric} \\ \quad \text{in-between: } \approx \emptyset \\ \text{non-parametric} \end{cases}$$

Our aim:

- move a step into the $\approx \emptyset$ space
- a bit more specifically, a step in the multivariate domain

Multivariate densities

- Context: non-parametric density estimation in $\mathbb{R}^d$
- problems grow with increasing d
- this is called ‘the curse of dimensionality’
- especially frustrating as it clashes with common perception: real data structures are not *really* that complex

About dimensionality

D. W. Scott (2015), Multivariate Density Estimation, 2nd edition, p.217

7.1 INTRODUCTION

The practical focus of most of this book is on density estimation in "several dimensions" rather than in very high dimensions. While this focus may seem misleading at first glance, it is indicative of a different point of view toward counting dimensions. Multivariate data in $\Re^d$ are almost never d-dimensional. That is, the *underlying structure* of data in $\Re^d$ is almost always of dimension lower than d. Thus, in general, the full space may be usefully partitioned into subspaces of signal and noise. Of course, this partition is not precise, but the goal is to eliminate a significant number of dimensions so as to encourage a parsimonious representation of the underlying structure.

The driving idea

The idea, in broad terms

- Target: alleviate the problem of dimensionality
- broad idea: adopt an intermediate formulation
  stay between parametric and non-parametric formulation
- need to insert a 'light' parametric structure

- This broad idea is open to various interpretations
- we explore one of them

Simplify the dependence structure

Simplify the dependence structure

- are all variables jointly related to all variables?
- perhaps, in some cases
- for other cases, we want to trim the dependence depth
- ...but, to reduce 'dependence depth', we must define it

Adopt a non-parametric density estimate

- Available sample: $(y_1, \ldots, y_n)$, where $y_i \in \mathbb{R}^d$
- Consider classical kernel density estimate (KDE):

$$\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{det(h)} K\left(h^{-1}(x - y_i)\right), \qquad x \in \mathbb{R}^d,$$

  having chosen
  - a kernel function K, e.g. density $\mathrm{N}_d(0, I_d)$,
  - a diagonal matrix h of positive smoothing parameters
- However, other estimates could be used.
- Aim: suitably modify classical KDE $\tilde{f}(x)$

Borrowing tools

- Borrow tools from log-linear models theory
- Consider d-dimensional frequency table
- Cell (log-)probabilities are expressed in a hierarchical structure
- For instance, if $d = 3$:

$$\begin{aligned}\log \pi_{rst} &= \lambda_0 + [\lambda_r^{(1)} + \lambda_s^{(2)} + \lambda_t^{(3)}] \\ &\quad + [\lambda_{rs}^{(12)} + \lambda_{rt}^{(13)} + \lambda_{st}^{(23)}] + [\lambda_{rst}^{(123)}]\end{aligned}$$

  with constraints among the λ's
- Simplify dependence structure by eliminating high-order terms
- Next, we plug this idea in the density estimation context

Adjust cells frequencies

- From sample $(y_1, \ldots, y_n)$, build d-dimensional frequency table
- Fit log-linear model to this table, with terms up to m-th order
- If j-th cell has frequency n_j, denote its fitted frequency by $\hat{n}_j$
- Proposal: sample point y_i belonging to j-th cell is given weight

$$w_{j(i)} = \hat{n}_j / n_j$$

so that the whole cell has weight $\hat{n}_j$

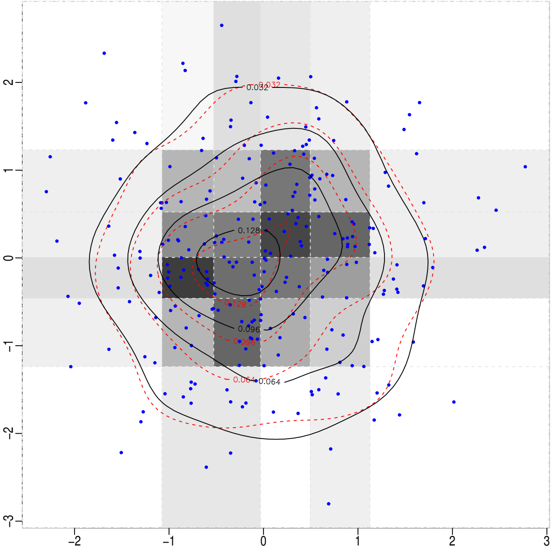
- Modified estimate:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{w_{j(i)}}{det(h)} K\left(h^{-1}(x - y_i)\right), \qquad x \in \mathbb{R}^d,$$

- local smoothing is combined with frequencies from log-linear model (the global smoother)

An ultra-simple illustration with $n = 250$ data from $\mathrm{N}_2(0, I_2)$

With $d = 2$, the only reduced log-linear model has $m = 1$ (independence)



Details

There are various practical details to handle:

- choice of cells/subdivisions for each axis?
- the smoothing parameter (as ever!)
- what to do when $n_j = 0$?
- choice of m?

Refer to the published paper for most of these points.
Here only examine choice of m.

Numerical work

Simulations set-up

Sample data with density

$$f(x) = \pi\, f_1(x) + (1-\pi)\, f_2(x)$$

considering

- either $0 < \pi < 1$ or $\pi = 1$
- f_1, f_2 either skew-normal or skew-t with $\nu = 2, 5, \infty$
- various correlation structures for f_1, f_2
- d from 3 to 5
- $m = 2$ or $m = 3$, provided $m < d$
- in most cases $n = 500$, sometimes $n = 250$ or 1000
- for each setting, $N = 2500$ samples

A full-factorial experiment is ‘impossible’, only a subset of cases.

Choose summary quantities

- Choice of summary quantities is not so obvious.
- Start from measure of error at a point x:

$$e_0(x) = \frac{|\tilde{f}(x) - f(x)|}{f(x)^{1/2}}, \qquad e(x) = \frac{|\hat{f}(x) - f(x)|}{f(x)^{1/2}}$$

  where

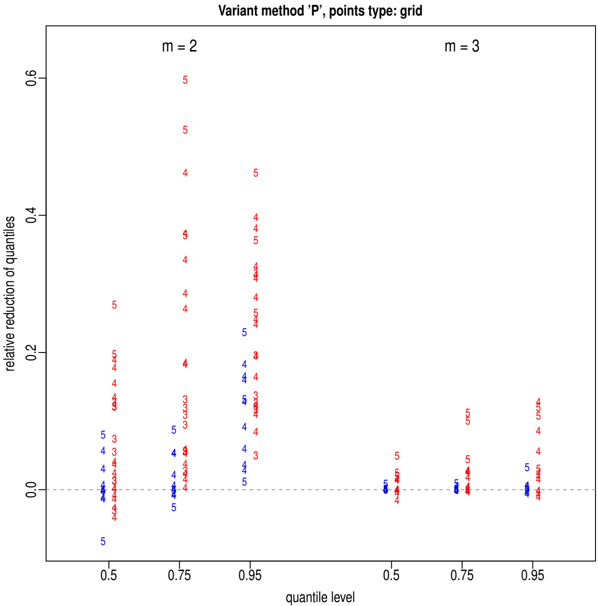
  $e_0(x)$ refers to classical KDE

  $e(x)$ refers to new proposal
- consider quantiles $Q_0(p), Q(p)$ at levels $p = (0.5, 0.75, 0.95)$
- final summary:

$$R(p) = \frac{Q_0(p) - Q(p)}{Q_0(p)}$$

- if $R(p) > 0$ there is an improvement over classical method

Summary outcome



Comments

- in most cases $R(p) > 0$, often by a good margin
- $m = 2$ superior to $m = 3$
- red points (unimodal) higher than blue points (mixtures)
- similar indications from other summaries
  (e.g. evaluation at the observed points, instead of a fixed grid)

Operational indication: just use $m = 2$

Closing

Application to density-based cluster analysis

- A natural application: density-based clustering methods
- clusters are associated to sets in $\mathbb{R}^d$ having high density
- specific exploration with R package `pdfCluster`
  using new estimate of the density

Clustering olive oil data

Clustering olive-oil data: true versus reconstructed groups
with R package `pdfCluster`

	classical KDE			new estimate		
	1	2	3	1	2	3
South	321	0	2	323	0	0
Sardinia	0	98	0	0	98	0
Centre-North	0	45	106	0	22	129
ARI		0.873			0.937	

Paper

Azzalini, A. (2016). Combining local and global smoothing in multivariate density estimation. *Stat*, 4.129.