Comparison of kernel density estimators with assumption on number of modes

Raphaël Coudret\textsuperscript{(1)}, Gilles Durrieu\textsuperscript{(2)}, Jérôme Saracco\textsuperscript{(1)}

\textsuperscript{(1)}CQFD team, INRIA Bordeaux - Sud-Ouest and IMB, University of Bordeaux
\textsuperscript{(2)}LMBA, University of South Brittany

COMPSTAT 2012, August 29th, 2012, Limassol
Summary

Taking the number of modes into account
  Kernel density estimators
  The critical bandwidth ...

Other probability density estimators
  A plug-in method
  Level sets of a density

Simulations
  A beta mixture
  The asymmetric claw density

Concluding remarks and prospects
Taking the number of modes into account
Kernel density estimators

Purpose of this talk:
Estimate a probability density function assuming its number of modes is known.

▶ Non-parametric framework besides this assumption.
▶ Insight about the presence of modes (mixture models).
▶ Approaches related to multimodality tests.
▶ Use of kernel density estimators.
Kernel density estimators

Let \( \hat{f}_{K,h} \) be the kernel density estimator of a density \( f \). It is defined for all \( t \in \mathbb{R} \) by:

\[
\hat{f}_{K,h}(t) := \frac{1}{n h} \sum_{i=1}^{n} K \left( \frac{t - X_i}{h} \right),
\]

for

- independent and identically distributed random variables \( X_1, \ldots, X_n \) generated from \( f \),
- a kernel \( K \), which is most of the time a probability density function (uniform distribution, normal distribution, ...)
- a positive real \( h \): the bandwidth.
The critical bandwidth ...

... for the Gaussian kernel

We choose the Gaussian kernel defined by $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$.
Let $N(f)$ be the number of modes of $f$.

The critical bandwidth is defined for $k \geq 1$ as

$$h_{\text{crit},k} := \min \limits_{N(\hat{f}_K,h) \leq k} (h).$$

Proposition (Silverman, 1981).

If $K$ is the Gaussian kernel, the function $h \rightarrow N(\hat{f}_K,h)$ is right-continuous, decreasing and piecewise constant.

This property ensures that $h_{\text{crit},k}$ can be computed by a binary search algorithm.
The critical bandwidth ...

... for the Gaussian kernel

Required assumptions on \( f \):

\[(\mathcal{H}1) \quad f \text{ is uniformly continuous on } \mathbb{R},\]
\[(\mathcal{H}2) \quad f \text{ has a bounded support } [r, s],\]
\[(\mathcal{H}3) \quad f \in C^2([r, s]),\]
\[(\mathcal{H}4) \quad \lim_{t \downarrow r} f'(t) > 0 \text{ and } \lim_{t \uparrow r} f'(t) < 0,\]
\[(\mathcal{H}5) \quad N(f) \leq k \text{ and if } z \text{ is the location of a mode or an antimode, } f''(z) \neq 0,\]
\[(\mathcal{H}6) \quad \text{the points of } ]r, s[ \text{ where } f' \text{ is equal to 0 are positions of either a mode or an antimode.}\]
The critical bandwidth ...

... for the Gaussian kernel

**Theorem.**
If assumptions \((\mathcal{H}1) - (\mathcal{H}6)\) are verified then, for every \(k \geq N(f)\), for every \(\varepsilon > 0\), when \(n \to \infty\),

\[
\mathbb{P} \left( \sup_{t \in \mathbb{R}} \left| \hat{f}_{K,h_{crit,k}}(t) - f(t) \right| > \varepsilon \right) \to 0,
\]

and,

\[
\mathbb{P} \left( \int_{\mathbb{R}} \left| \hat{f}_{K,h_{crit,k}}(t) - f(t) \right| dt > \varepsilon \right) \to 0.
\]

**Proof.**
Using properties on \(h_{crit,k}\) given by Mammen, Marron and Fisher (1991) and those in Devroye (1987) and in Devroye and Wagner (1980) concerning random bandwidths.
The critical bandwidth ...

... for the uniform kernel

When $K(t) = \mathbb{I}_{[-\frac{1}{2}, \frac{1}{2}]}(t)$, $h_{crit,k}$ can be computed for each $k \geq 1$.

BUT

Theorem.

When $K$ is the uniform kernel, for every probability density $f$ from which $X_1, \ldots, X_n$ are generated, $h_{crit,k}$ increases with $n$.

AND

Devroye (1987) and Devroye and Wagner (1980) require that $h_{crit,k} \rightarrow 0$ when $n \rightarrow \infty$ to obtain satisfying asymptotic results.
Other probability density estimators
A plug-in method

Choice of the bandwidth (Sheather and Jones, 1991):

\[ h_{SJ} := \arg\min_h (\widehat{AMISE}(h)), \]

where

- \( \widehat{AMISE}(h) \) is a plug-in estimator of \( AMISE(h) \),
- \( AMISE(h) \) is an asymptotic expansion of \( MISE(h) \),
- \( MISE(h) := \mathbb{E} \left[ \int_{\mathbb{R}} \left( \hat{f}_{K,h}(t) - f(t) \right)^2 dt \right] \)

The bandwidth \( h_{SJ} \) is asymptotically close to

\[ h_{opt} := \arg\min_h (MISE(h)). \]
Level sets of a density

For all $\lambda > 0$, a level set $\Gamma(\lambda)$ of a density $f$ is defined as:

$$\Gamma(\lambda) := \{x : f(x) \geq \lambda\}.$$

**Properties:**

- Let $\mathcal{U}$ be the set of the unions of at most $N(f)$ disjoint intervals of $\mathbb{R}$. Then, for all $\lambda \geq 0$, $\Gamma(\lambda) \in \mathcal{U}$.

- $\forall t \in \mathbb{R}, \ f(t) = \int_{0}^{\infty} \mathbb{I}_{\Gamma(\lambda)}(t) d\lambda$. 

**Figure:** An example of a level set.
Level sets of a density

Polonik (1995) provides an estimator $\hat{\Gamma}_{n,U}(\lambda)$ of $\Gamma(\lambda)$ for every $\lambda \geq 0$, such that

$$\hat{\Gamma}_{n,U}(\lambda) \in U.$$ 

Then, we can estimate $f$ with $\hat{f}_P$, defined by:

$$\hat{f}_P(t) := \int_0^\infty \mathbb{I}_{\hat{\Gamma}_{n,U}(\lambda)}(t) d\lambda.$$ 

Polonik showed the following property:

$$\mathbb{P} \left( \int_\mathbb{R} \left| \hat{f}_P(t) - f(t) \right| dt > \varepsilon \right) \to 0,$$

when $n \to \infty$. 
Simulations
A beta mixture

\[ X \sim \begin{cases} \mathcal{B}(\alpha_1, \beta_1) & \text{with probability } p_1, \\ \mathcal{B}(\alpha_2, \beta_2) & \text{with probability } p_2 = 1 - p_1, \end{cases} \]

with \( \alpha_1 = \beta_2 = 2, \beta_1 = 5, \alpha_2 = 10 \) and \( p_1 = \frac{2}{3} \).

Quality criteria of an estimate \( \hat{f} \) of \( f \):

\[ \text{IAE}(\hat{f}) := \int_{\mathbb{R}} |\hat{f}(t) - f(t)| \, dt, \]

and

\[ \eta(\hat{f}) := \arg\min_{\hat{z} \in \hat{Z}} \hat{f}(\hat{z}) - \arg\min_{z \in Z} f(z), \]

where \( Z \) and \( \hat{Z} \) are respectively the sets of the locations of the antimodes of \( f \) and \( \hat{f} \).
A beta mixture

Figure: An example of density estimates with: \( \hat{f}_{K,h_{\text{crit}},2} \) and the Gaussian kernel, \( \hat{f}_{K,h_{\text{crit}},2} \) and the uniform kernel, \( \hat{f}_{K,h_{\text{SJ}}} \) and the Gaussian kernel, \( \hat{f}_P \).

Points: Locations of \( \text{argmin}_{\hat{z} \in \hat{Z}} \hat{f}(\hat{z}) \) for each estimate.
A beta mixture

**Figure:** Boxplots of 100 $\text{IAE}(\hat{f})$ measures for various density estimators $\hat{f}$.

- **Top, left:** $\hat{f}_{K,h_{\text{crit}},2}$ where $K$ is the Gaussian kernel.
- **Top, right:** $\hat{f}_{K,h_{\text{crit}},2}$ where $K$ is the uniform kernel.
- **Bottom, left:** $\hat{f}_{K,h_{\text{SJ}}}$ where $K$ is the Gaussian kernel.
- **Bottom, right:** $\hat{f}_{P}$.

More graphics ...
The asymmetric claw density

\[ X \sim \begin{cases} 
\mathcal{N}(0, 1) & \text{with probability } \frac{1}{2}, \\
\mathcal{N}\left(l + \frac{1}{2}, \left(\frac{2^{-l}}{10}\right)^2\right) & \text{with probability } \frac{2^{1-l}}{31},
\end{cases} \text{ for } l \in \{-2, -1, 0, 1, 2\}. \]

We have \( N(f) = 5 \) where \( f \) is the probability density of \( X \). Some modes of \( f \) are very sharp. Some other modes are smooth.

- \( f \) has special local features.
- The bandwidth is a global parameter.

Behavior of \( h_{\text{crit},k} \) and \( h_{\text{SJ}} \) ?
The asymmetric claw density

Figure: An example of density estimates with \( \hat{f}_{K,h_{\text{crit},5}} \) and \( \hat{f}_{K,h_{\text{SJ}}} \) when \( K \) is the Gaussian kernel. Points: locations of \( \arg\min_{\tilde{z} \in \tilde{Z}} \hat{f}(\tilde{z}) \) for each estimator.
The asymmetric claw density

Figure: Groups of 100 $\eta(\hat{f})$ measures for the estimators (left) $\hat{f}_{K,h_{\text{crit}},5}$ and (right) $\hat{f}_{K,h_{\text{SJ}}}$, where $K$ is the Gaussian kernel.
Concluding remarks and prospects
Concluding remarks and prospects

- We can choose $N(\hat{f}_K, h_{crit,k})$. It is not possible for $\hat{f}_K, h_{SJ}$ and for $\hat{f}_P$.

- Alike numerical convergences of the IAE for a simple case (beta mixture).

- Good results for $h_{crit,k}$ when estimating $\text{argmin}_{z \in Z} f(z)$.

- Do not use $h_{crit,k}$ with the uniform kernel.

- How to choose $k$ in $h_{crit,k}$?

- Application: biological data.
Taking the number of modes into account
Other probability density estimators
Simulations
Concluding remarks and prospects

References

R. Coudret, G. Durrieu, and J. Saracco.
Comparison of kernel density estimators with assumption on number of modes.
Submitted paper.

L. P. Devroye.
A Course in Density Estimation.

L. P. Devroye and P. J. Wagner.
The strong uniform consistency of kernel density estimates.

E. Parzen.
On estimation of probability density function and mode.

W. Polonik.
Density estimation under qualitative assumptions in higher dimensions.

S. J. Sheather and M. C. Jones.
A reliable data-based bandwidth selection method for kernel density estimation.

B. W. Silverman.
Using kernel density estimates to investigate multimodality.
A beta mixture

Figure: Groups of 100 $\eta(\hat{f})$ measures for various density estimators.

Top, left: $\hat{f}_{K,h\text{crit}}$ where $K$ is the Gaussian kernel.

Top, right: $\hat{f}_{K,h\text{crit}}$ where $K$ is the uniform kernel.

Bottom, left: $\hat{f}_{K,h\text{SJ}}$ where $K$ is the Gaussian kernel.

Bottom, right: $\hat{f}_P$. 
The asymmetric claw density

Figure: Boxplot of 100 $IAE(\hat{f})$ measures for the estimators (left) $\hat{f}_{K,h_{crit,5}}$ and (right) $\hat{f}_{K,h_{SJ}}$, where $K$ is the Gaussian kernel.