Combining cross-validation and plug-in methods
for kernel density bandwidth selection

Carlos Tenreiro

CMUC and DMUC, University of Coimbra
The nonparametric density estimation problem

The Parzen-Rosenblatt kernel density estimator

Cross-validation and plug-in methods for bandwidth selection

Combining cross-validation and plug-in methods

(based on a recent joint work with J.E. Chacón, Universidad de Extremadura, Spain)

This is in order to obtain a data-based bandwidth selector that presents an overall good performance for a large set of underlying densities.
We have observations

\[ X_1, X_2, \ldots, X_n \]

independent and identically distributed real-valued random variables with unknown probability density function $f$:

\[ P(a < X < b) = \int_a^b f(x) \, dx \]

for all $-\infty < a < b < +\infty$.

We want to estimate $f$ based on the previous observations.

The goal in nonparametric density estimation is to estimate $f$ making only minimal assumptions about $f$. 
Nonparametric density estimation

- Exploring data is one of the goals of nonparametric density estimation.
- **Hidalgo Stamp Data**: thickness of 485 postage stamps that were printed over a long time in Mexico during the 19th century.

The idea is to gain insights into the number of different types of papers that were used to print the postage stamps.
In this talk, we will restrict our attention to another well known density estimator introduced by Rosenblatt (1956) and Parzen (1962): the kernel density estimator.

The motivation given by Rosenblatt (1956) for this density estimator is based on the fact

\[ f(x) = F'(x) \]

where the cumulative distribution function \( F \) can be estimated by the empirical distribution function given by

\[ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[\infty, x]}(X_i), \]

with

\[ I_A(y) = \begin{cases} 1 & \text{if } y \in A \\ 0 & \text{if } y \notin A. \end{cases} \]
If \( h \) is a small positive number we could expect that

\[
f(x) \approx \frac{F(x + h) - F(x - h)}{2h}
\]

\[
\approx \frac{F_n(x + h) - F_n(x - h)}{2h}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2h} I_{[x-h,x+h]}(X_i)
\]

\[
= \frac{1}{nh} \sum_{i=1}^{n} K_0 \left( \frac{x - X_i}{h} \right),
\]

where

\[
K_0(\cdot) = \frac{1}{2} I_{[-1,1]}(\cdot).
\]
The Parzen-Rosenblatt kernel estimator is obtained by replacing $K_0$ by a general symmetric density function $K$:

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

where:

- $h = h_n$, the bandwidth or smoothing parameter, is a sequence of strictly positive real numbers converging to zero as $n$ tends to infinity;
- $K$, the kernel, is a bounded and symmetric density function.

Contrary to the histogram estimator, the Parzen-Rosenblatt estimator gives regular estimates for $f$ if we take for $K$ a regular density function.
The choice of the bandwidth is a crucial issue for kernel density estimation.

Kernel density estimates for the Hidalgo Stamp Data (n=485):

\[ K(x) = (2\pi)^{-1/2}e^{-x^2/2} \]
The role played by $h$

- For the mean integrated square error

$$MISE(f; n, h) = \mathbb{E} \int \{ f_h(x) - f(x) \}^2 dx,$$

we have

$$MISE(f; n, h) = \int \text{Var} f_h(x) \, dx + \int \{ \mathbb{E} f_h(x) - f(x) \}^2 \, dx$$

$$\sim \frac{1}{nh} \int K^2(u) \, du + \frac{h^4}{4} \int u^2 K(u) \, du \int f''(x)^2 \, dx.$$

- If $h$ is too small we obtain an estimator with a small bias but with a large variability.

- If $h$ is too large we obtain an estimator with a large bias but with a small variability.
Choosing the bandwidth corresponds to balancing bias and variance.

$n = 1000$ and $h = 0.1$

undersmoothing
small $h$
small bias but large variability
Choosing the bandwidth corresponds to balancing bias and variance.

$n = 1000$ and $h = 0.5$

oversmoothing
large $h$
small variability but large bias
The role played by $h$:

- The main challenge in smoothing is to determine how much smoothing to do. With $n = 1000$ and $h = 0.36$, the kernel is almost right.

- The choice of the kernel is not so relevant to the estimator behaviour.
Data-based bandwidth selectors

- Under general conditions on the kernel and on the underlying density function, for each \( n \in \mathbb{N} \) there exists an optimal bandwidth \( h_{\text{MISE}} = h_{\text{MISE}}(n; f) \) in the sense that

\[
\text{MISE}(f; n, h_{\text{MISE}}) \leq \text{MISE}(f; n, h), \quad \text{for all } h > 0.
\]

- But \( h_{\text{MISE}} \) depends on the unknown density \( f \) ...

- We are interested in methods for choosing \( h \) that are based on the observations \( X_1, \ldots, X_n \),

\[
h = h(X_1, \ldots, X_n),
\]

and satisfy

\[
h(X_1, \ldots, X_n) \approx h_{\text{MISE}},
\]

for a large class of densities.
For each $h > 0$, we start by considering an unbiased estimator of $\text{MISE}(f; n, h) - R(f)$, given by

$$CV(h) = \frac{R(K)}{nh} + \frac{1}{n(n-1)} \sum_{i \neq j} \left( \frac{n-1}{n} K_h * K_h - 2K_h \right)(X_i - X_j),$$

and we take for $h$ the value $\hat{h}_{CV}$ which minimises $CV(h)$.

(Rudemo, 1982; Bowman, 1984)

Under some regularity conditions on $f$ and $K$ we have

$$\frac{\hat{h}_{CV}}{h_{MISE}} - 1 = O_p \left( n^{-1/10} \right).$$

(Hall, 1983; Hall & Marron, 1987)
The plug-in method is based on a simple idea that goes back to Woodroofe (1970).

We start with an asymptotic approximation $h_0$ for the optimal bandwidth $h_{\text{MISE}}$:

$$h_0 = c_K \psi^{-1/5} n^{-1/5}$$

where

$$c_K = R(K)^{-1/5} \left( \int u^2 K(u) du \right)^{-2/5}$$

and

$$\psi_r = \int f(r)(x)f(x)dx, \quad r = 0, 2, 4, \ldots$$

The plug-in bandwidth selector is obtained by replacing the unknown quantities in $h_0$ by consistent estimators:

$$\hat{h}_{\text{PI}} = c_K \hat{\psi}^{-1/5} n^{-1/5}$$
Estimating the functional $\psi_r$

- A class of kernel estimators of $\psi_r$ was introduced by Hall and Marron (1987a, 1991) and Jones and Sheather (1991):

$$\hat{\psi}_r(g) = \frac{1}{n^2} \sum_{i,j=1}^{n} U_g^{(r)}(X_i - X_j),$$

where $g$ is a new bandwidth and $U$ is a bounded, symmetric and $r$-times differentiable kernel.

- For $U = \phi$, the bandwidth that minimises the asymptotic mean square error of $\hat{\psi}_r(g)$ is given by

$$g_{0,r} = \left(2|\phi^{(r)}(0)||\psi_{r+2}|^{-1}\right)^{1/(r+3)} n^{-1/(r+3)}.$$

- In the case of the estimation of $\psi_4$, this bandwidth depends (again) on the unknown quantity $\psi_6$!
Multistage plug-in estimation of $\psi_r$

In order to estimate $\psi_4$ we have then the following schema:

<table>
<thead>
<tr>
<th>to estimate</th>
<th>we consider</th>
<th>we need</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_4$</td>
<td>$\hat{\psi}<em>4(\hat{g}</em>{0,4})$</td>
<td>$\psi_6$</td>
</tr>
<tr>
<td>$\psi_6$</td>
<td>$\hat{\psi}<em>6(\hat{g}</em>{0,6})$</td>
<td>$\psi_8$</td>
</tr>
<tr>
<td>$\psi_8$</td>
<td>$\hat{\psi}<em>8(\hat{g}</em>{0,8})$</td>
<td>$\psi_{10}$</td>
</tr>
<tr>
<td>$\psi_{4+2(\ell-1)}$</td>
<td>$\hat{\psi}<em>{4+2(\ell-1)}(\hat{g}</em>{0,4+2(\ell-1)})$</td>
<td>$\psi_{4+2\ell}$</td>
</tr>
</tbody>
</table>

where

$$\hat{g}_{0,r} = \left(2|\phi^{(r)}(0)||\hat{\psi}_{r+2}|^{-1}\right)^{1/(r+3)} n^{-1/(r+3)}.$$
The usual strategy to stop this cyclic process, is to use a parametric estimator of $\psi_{4+2\ell}$ based on some parametric reference distribution family.

The standard choice for the reference distribution family is the normal or Gaussian family:

$$f(x) = (2\pi \sigma)^{-1/2} e^{-x^2/(2\sigma^2)}.$$  

In this case, $\psi_{4+2\ell}$ is estimated by

$$\hat{\psi}_{4+2\ell}^{NR} = \phi^{(4+2\ell)}(0)(2\hat{\sigma}^2)^{-(5+2\ell)/2},$$

where $\hat{\sigma}$ denotes any scale estimate.
For a fixed $\ell \in \{1, 2, \ldots\}$ the $\ell$-stage estimator of $\psi_4$ is:

\[
\hat{\psi}^{\text{NR}}_{4+2\ell} \leadsto \hat{g}_{0,4+2(\ell-1)} \mapsto \hat{\psi}_{4+2(\ell-1)} \\
\hat{g}_{0,4+2(\ell-2)} \mapsto \hat{\psi}_{4+2(\ell-2)} \\
\hat{g}_{0,4+2(\ell-3)} \mapsto \hat{\psi}_{4+2(\ell-3)} \\
\vdots \\
\hat{g}_{0,4} \mapsto \hat{\psi}_4 \leadsto \hat{\psi}_{4,\ell}
\]
Multistage plug-in bandwidth selector

- Depending on the number \( \ell \in \{1, 2, \ldots\} \) of considered pilot stages of estimation we get different estimators \( \hat{\psi}_{4,\ell} \) of \( \psi_4 \).

- The associated \( \ell \)-stage plug-in bandwidth selector for the kernel density estimator is given by

\[
\hat{h}_{\Pi,\ell} = c_K \hat{\psi}_{4,\ell}^{-1/5} n^{-1/5}
\]

If \( f \) has bounded derivatives up to order \( 4 + 2\ell \) then

\[
\frac{\hat{h}_{\Pi,\ell}}{h_{\text{MISE}}} - 1 = O_p \left( n^{-\alpha} \right),
\]

with \( \alpha = 2/7 \) for \( \ell = 1 \) and \( \alpha = 5/14 \) for all \( \ell \geq 2 \).

(CT, 2003)
Two-stage plug-in bandwidth selector

- For the standard choice $\ell = 2$ we have:

\[
\hat{\psi}^{\text{NR}}_{8} \leadsto \hat{g}_{0,6} \leftrightarrow \hat{\psi}_{6} \\
\hat{g}_{0,4} \leftrightarrow \hat{\psi}_{4} \leadsto \hat{\psi}_{4,2}
\]

- The associated **two-stage plug-in bandwidth selector** is given by

\[
\hat{h}_{\text{PI},2} = c_{K} \hat{\psi}^{-1/5}_{4,2} n^{-1/5}
\]
Finite sample behaviour

- Nonparametric density estimation
- Kernel density estimator
- The role of $h$
- Data-based bandwidth selectors
- CV bandwidth
- PI bandwidth
- Estimating $\psi_r$
- Multistage PI
- Bandwidth
- Combining PI & CV
- Combining PI & CV
- References

Standard normal density:

Skewed unimodal density:

$n = 400$

ISE

$PI.2$ $CV$

$PI.2$ $CV$

$0.000$ $0.010$ $0.015$

$0.000$ $0.005$ $0.010$

$0.020$ $0.010$ $0.000$

$0.000$ $0.005$ $0.010$

PhD Program UC|UP

February 18, 2011 – 22
Finite sample behaviour

Nonparametric density estimation
Kernel density estimator
The role of $h$
Data-based bandwidth selectors
CV bandwidth
PI bandwidth
Estimating $\psi_r$
Multistage PI bandwidth
Combining PI & CV
Combining PI & CV
References

Strongly skewed density:

Asymmetric multimodal density:

$n = 400$

ISE

PI.2 CV

n = 400

ISE

PI.2 CV
Finite sample behaviour

Standard normal density:

Skewed unimodal density:

ISE

number of stages

References

PhD Program UC|UP  February 18, 2011 – 24
Finite sample behaviour

Strongly skewed density:

Asymmetric multimodal density:

References
From a finite-sample point of view the performance of $\hat{h}_{\text{PI},\ell}$ strongly depends on the considered number of stages.

For strongly skewed or asymmetric multimodal densities the standard choice $\ell = 2$ gives poor results.

The natural question that arises from the previous considerations is:

How can we choose the number of pilot stages $\ell$?

This is an old question posed by Park and Marron (1992).

In order to answer this question, the idea developed by Chacón and CT (2008) was to combine plug-in and cross-validation methods.
Combining plug-in and cross-validation procedures

- We started by fixing minimum and a maximum number of pilot stages
  \[ \underline{L} \quad \text{and} \quad \overline{L} \]
  and by choosing a stage \( \ell \) among the set of possible pilot stages
  \[ \mathcal{L} = \{\underline{L}, \underline{L} + 1, \ldots, \overline{L}\} \]

- This is equivalent to select one of the bandwidths
  \[ \hat{h}_{\text{PI},\ell} = c_K \hat{\psi}_{4,\ell}^{-1/5} n^{-1/5}, \ell \in \mathcal{L}. \]

- Recall that each one of these bandwidths has good asymptotic properties.
In order to select one of the previous multistage plug-in bandwidths we consider a weighted version of the cross-validation criterion function given by

$$CV_\gamma(h) = \frac{R(K)}{nh} + \frac{\gamma}{n(n-1)} \sum_{i \neq j} (\frac{n-1}{n} K_h * K_h - 2K_h)(X_i - X_j),$$

for some $0 < \gamma \leq 1$ that needs to be fixed by the user.

Finally, we take the bandwidth

$$\hat{h}_{PI,\ell} = c K \hat{\psi}_{4,\ell}^{-1/5} n^{-1/5},$$

where

$$\ell = \arg\min_{\ell \in L} CV_\gamma(\hat{h}_{PI,\ell}).$$
Asymptotic behaviour

If $f$ has bounded derivatives up to order $4 + 2\bar{L}$ and

$$|\psi_{4+2\ell}| \geq |\psi_{4+2\ell}^{NR}(\sigma_f)|, \quad \text{for all } \ell = 1, 2, \ldots, \bar{L},$$

(1)

then

$$\frac{\hat{h}_{PI,\ell}}{h_{MISE}} - 1 = O_p\left(n^{-\alpha}\right)$$

with $\alpha = 2/7$ for $\bar{L} = 1$ and $\alpha = 5/14$ for $\bar{L} \geq 2$.

(Chacón & CT, 2008)

□ Condition (1) is not very restrictive due to the smoothness of the normal distribution.

□ This result justifies the recommendation of using $\bar{L} = 2$. 
Asymptotic behaviour

Proof:

- From the definite-positivity property of the class of gaussian based kernels used in the multistage estimation process one can prove that

\[ P(\Omega_{L,L}) \to 1 \]

where

\[ \Omega_{L,L} = \left\{ \hat{h}_{PI,L} \leq \hat{h}_{PI,L-1} \leq \cdots \leq \hat{h}_{PI,L+1} \leq \hat{h}_{PI,L} \right\}. \]

- The conclusion follows easily from the asymptotic behaviour of \( \hat{h}_{PI,L} \) and \( \hat{h}_{PI,L} \), since for a sample in \( \Omega_{L,L} \) we have

\[
\frac{\hat{h}_{PI,L}}{h_{MISE}} - 1 \leq \frac{\hat{h}_{PI,\hat{\ell}(L)}}{h_{MISE}} - 1 \leq \frac{\hat{h}_{PI,L}}{h_{MISE}} - 1.
\]
**On the role played by $\bar{L}$ and $\gamma$**

Nonparametric density estimation
Kernel density estimator
The role of $h$
Data-based bandwidth selectors
CV bandwidth
PI bandwidth
Estimating $\psi_\gamma$
Multistage PI bandwidth
Combining PI & CV

Distribution of $\text{ISE}(\hat{h}_{\text{PI},\hat{\ell}})$ as a function of $\bar{L}$ and $\gamma$ ($n = 200$)
Choosing $\bar{L}$ and $\gamma$ in practice

- The boxplots show that a larger value for $\bar{L}$ is recommended especially for hard-to-estimate densities.

  The new bandwidth $\hat{h}_{\text{PI},\ell}$ is quite robust against the choice of $\bar{L}$ whenever a sufficiently large value is taken for $\bar{L}$.

  We decide to take $\bar{L} = 30$.

- Regarding the choice of $\gamma$, small values of $\gamma$ are more appropriate for easy-to-estimate densities, whereas large values of $\gamma$ are more appropriate for hard-to-estimate densities.

  In order to find a compromise between these two situations we decide to take $\gamma = 0.6$.

- We expect to obtain a new data-based bandwidth selector that presents a good overall performance for a wide range of density features.
Finite sample behaviour

Standard normal density:

Skewed unimodal density:

n = 400

ISE

PI.2 New CV

0.000 0.005 0.010 0.015

0.020 0.010 0.005

PI.2 New CV

References
Finite sample behaviour

Strongly skewed density:

Asymmetric multimodal density:

References
From this plot we can identify seven modes ... seven different types of paper were used (probably).
