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

Carlos Tenreiro

CMUC and DMUC, University of Coimbra

Overview

- 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:

$$\mathrm{P}(a < X < b) = \int_a^b f(x)\, dx$$

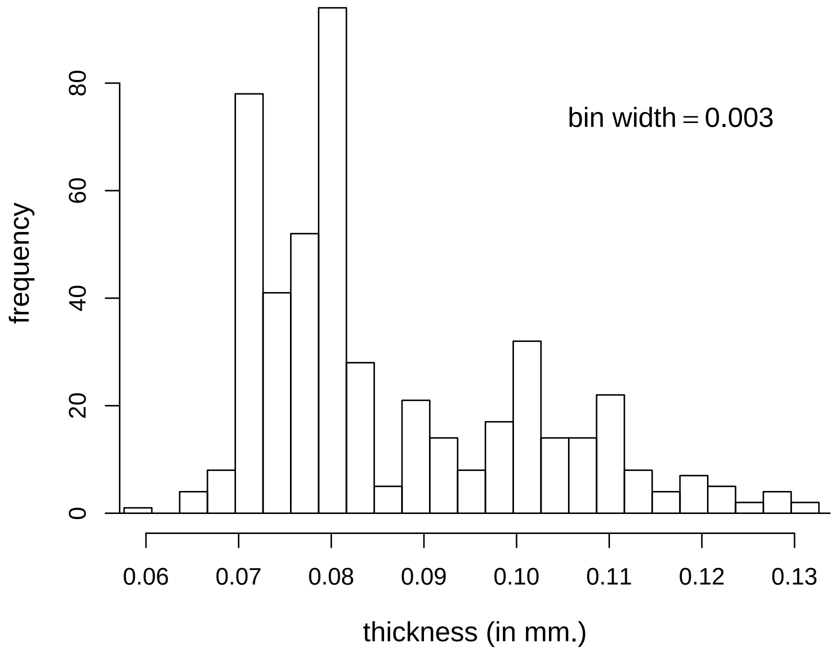
for all $-\infty < a < b < +\infty$.

- 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

$$\begin{aligned} f(x) &\approx \frac{F(x+h) - F(x-h)}{2h} \\ &\approx \frac{F_n(x+h) - F_n(x-h)}{2h} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} I_{]x-h,x+h]}(X_i) \\ &= \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\frac{x - X_i}{h}\right), \end{aligned}$$

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:

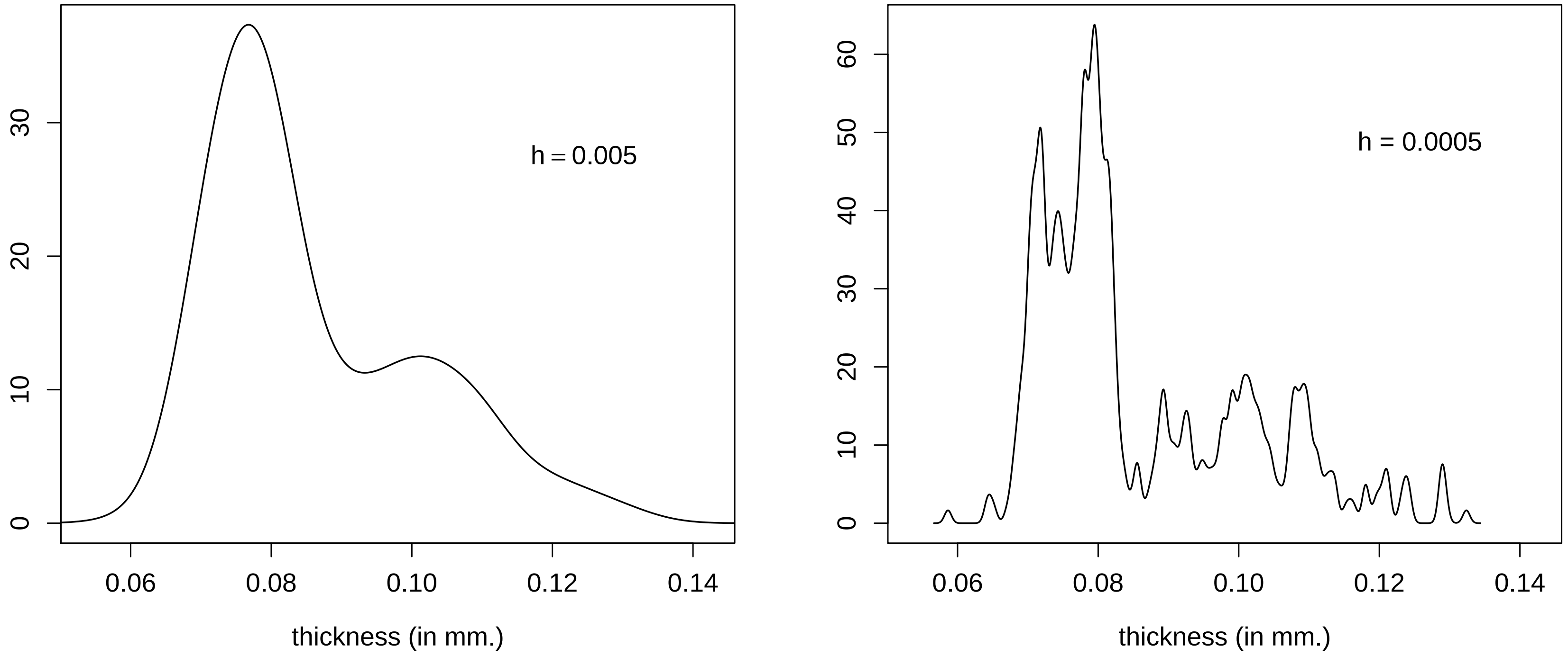
  – $h = h_n$, the bandwidth or smoothing parameter, is a sequence of strictly positive real numbers converging to zero as n tends to infinity;

  – 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

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

we have

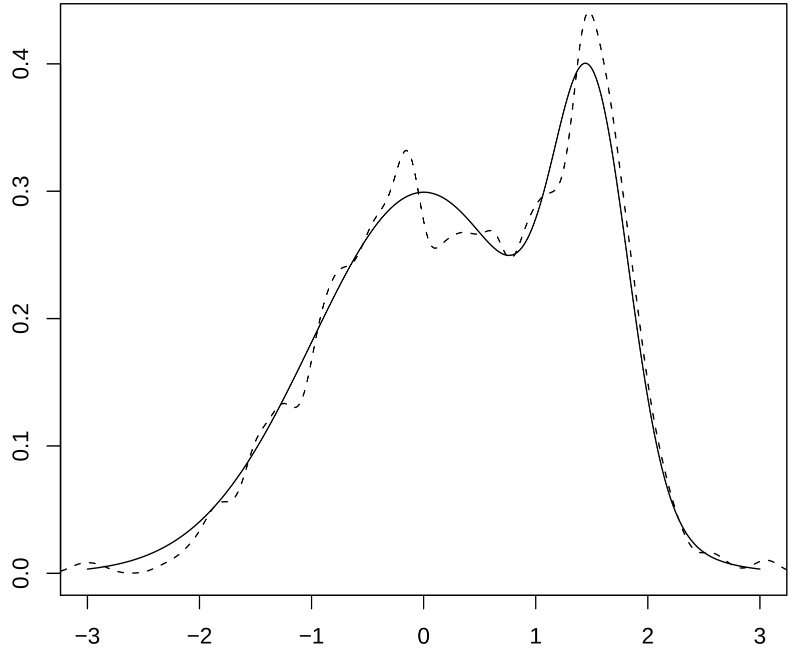
$$\begin{aligned}
&\text{MISE}(f; n, h) \\
&= \int \text{Var} f_h(x)\, dx + \int \{\text{E} f_h(x) - f(x)\}^2\, dx \\
&\sim \frac{1}{nh} \int K^2(u) du + \frac{h^4}{4} \int u^2 K(u) du \int f''(x)^2 dx.
\end{aligned}$$

□ 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 $h = 0.1$

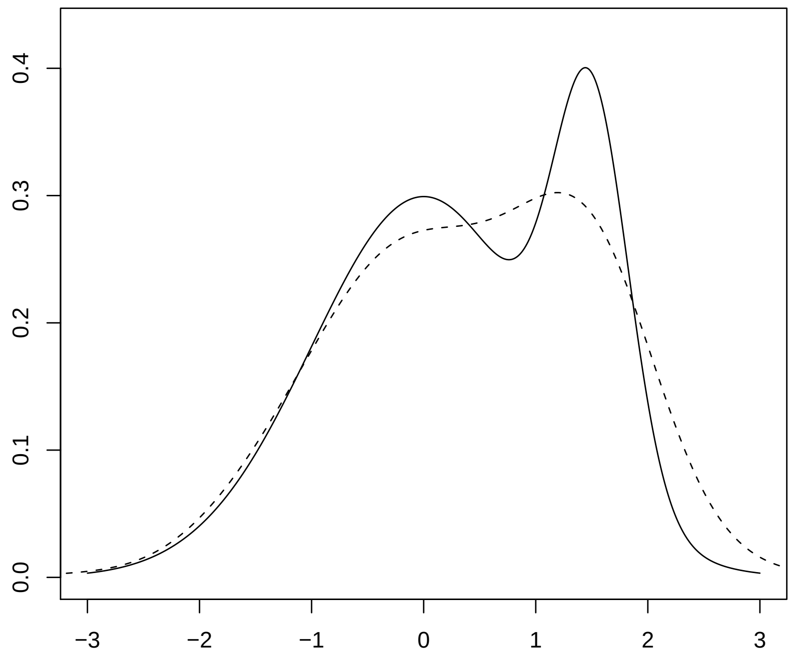


undersmoothing
small h
small bias but large variability

The role played by h

- Choosing the bandwidth corresponds to balancing bias and variance.

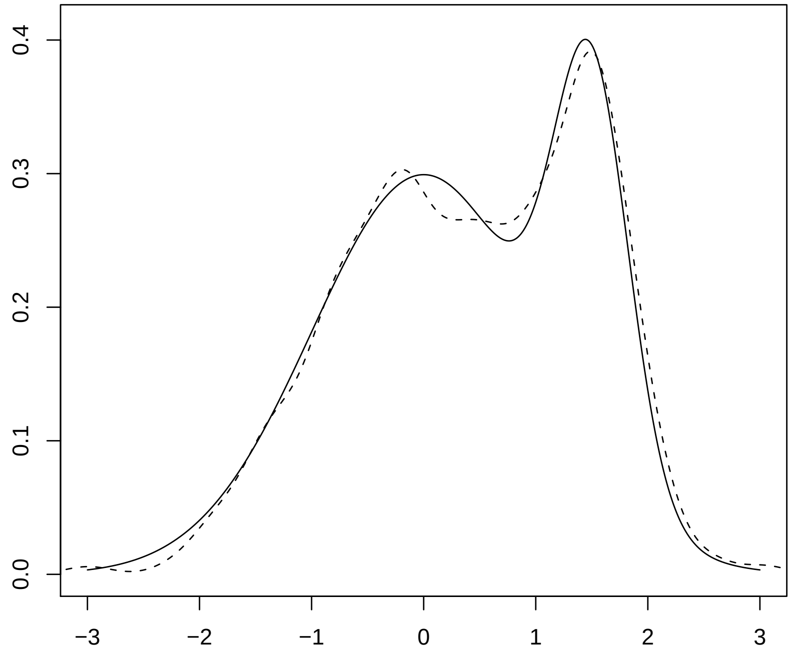
$n = 1000$ and $h = 0.5$



oversmoothing
large h
small variability but large bias

- The main challenge in smoothing is to determine how much smoothing to do.

$n = 1000$ and $h = 0.36$



almost right

- The choice of the kernel is not so relevant to the estimator behaviour.

- 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

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

- But h_{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 classe of densities.

Cross-validation bandwidth selection

□ For each $h > 0$, we start by considering an unbiased estimator of $\mathrm{MISE}(f; n, h) - R(f)$, given by

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

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

(Rudemo, 1982; Bowman, 1984)

Under some regularity conditions on f and K we have

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

(Hall, 1983; Hall & Marron, 1987)

Plug-in bandwidth selection

- 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_{MISE}:

$$h_0 = c_K \psi_4^{-1/5} n^{-1/5}$$

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:

$$\hat{h}_{\mathrm{PI}} = c_K \hat{\psi}_4^{-1/5} n^{-1/5}$$

- A class of kernel estimators of ψ_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 ψ_4, this bandwidth depends (again) on the unknown quantity ψ_6!

□ In order to estimate ψ_4 we have then the following schema:

to estimate		we consider		we need
ψ_4	$\longrightarrow$	$\hat{\psi}_4(\hat{g}_{0,4})$	$\longrightarrow$	ψ_6
ψ_6	$\longrightarrow$	$\hat{\psi}_6(\hat{g}_{0,6})$	$\longrightarrow$	ψ_8
ψ_8	$\longrightarrow$	$\hat{\psi}_8(\hat{g}_{0,8})$	$\longrightarrow$	ψ_{10}
		$\vdots$		
$\psi_{4+2(\ell-1)}$	$\longrightarrow$	$\hat{\psi}_{4+2(\ell-1)}(\hat{g}_{0,4+2(\ell-1)})$	$\longrightarrow$	$\psi_{4+2\ell}$

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

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

where $\hat{\sigma}$ denotes any scale estimate.

Multistage plug-in estimation of ψ_4

- For a fixed $\ell \in \{1, 2, \ldots\}$ the ℓ-stage estimator of ψ_4 is:

$$
\hat{\psi}^{\mathrm{NR}}_{4+2\ell} \rightsquigarrow
\boxed{
\begin{array}{ccc}
\hat{g}_{0,4+2(\ell-1)} & \mapsto & \hat{\psi}_{4+2(\ell-1)} \\
 & \swarrow & \\
\hat{g}_{0,4+2(\ell-2)} & \mapsto & \hat{\psi}_{4+2(\ell-2)} \\
 & \swarrow & \\
\hat{g}_{0,4+2(\ell-3)} & \mapsto & \hat{\psi}_{4+2(\ell-3)} \\
 & \swarrow & \\
 & \vdots & \\
 & \swarrow & \\
\hat{g}_{0,4} & \mapsto & \hat{\psi}_4
\end{array}
}
\rightsquigarrow \hat{\psi}_{4,\ell}
$$

- Depending on the number $\ell \in \{1, 2, \ldots\}$ of considered pilot stages of estimation we get different estimators $\hat{\psi}_{4,\ell}$ of ψ_4.

- The associated ℓ-stage plug-in bandwidth selector for the kernel density estimator is given by

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

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

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

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:

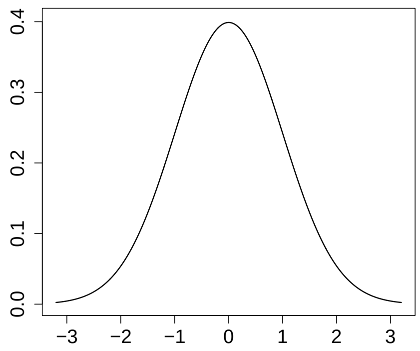
$$\hat{\psi}_8^{\mathrm{NR}} \rightsquigarrow \boxed{\begin{array}{ccc} \hat{g}_{0,6} & \mapsto & \hat{\psi}_6 \\ & \swarrow & \\ \hat{g}_{0,4} & \mapsto & \hat{\psi}_4 \end{array}} \rightsquigarrow \hat{\psi}_{4,2}$$

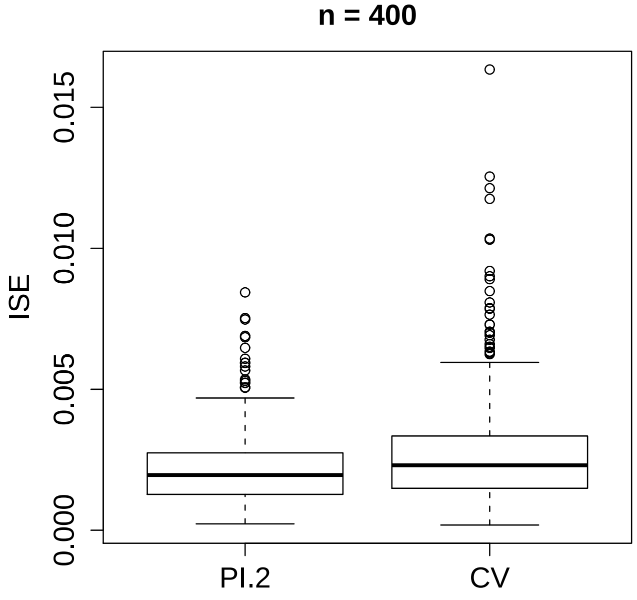
- The associated two-stage plug-in bandwidth selector is given by

$$\hat{h}_{\mathrm{PI},2} = c_K \hat{\psi}_{4,2}^{-1/5} n^{-1/5}$$

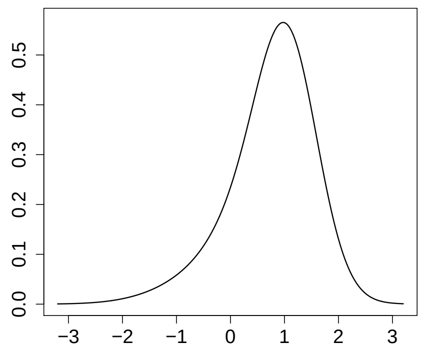
Finite sample behaviour

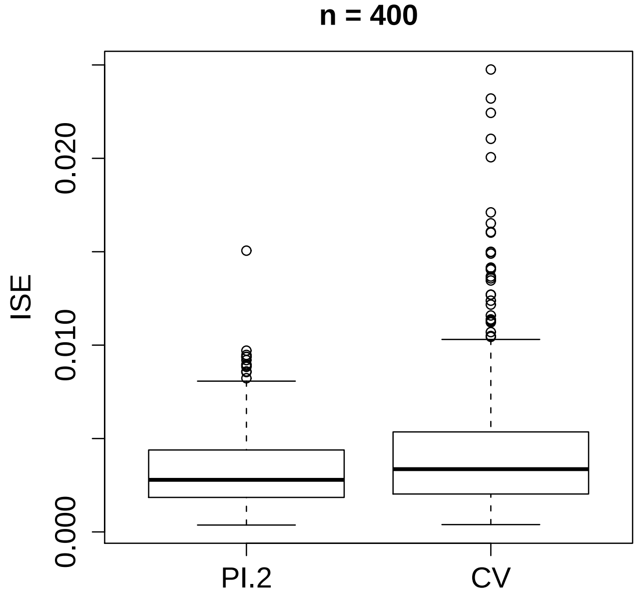
Standard normal density:





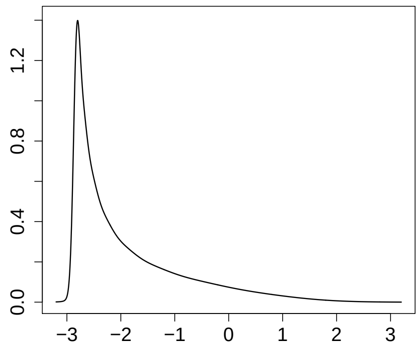
Skewed unimodal density:

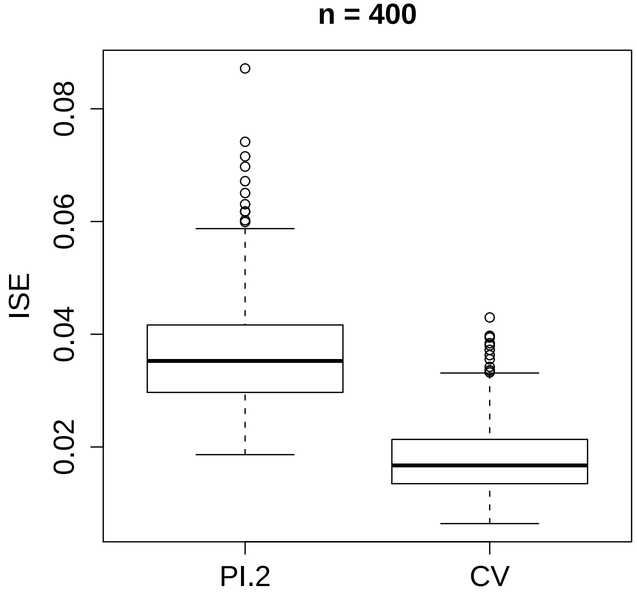




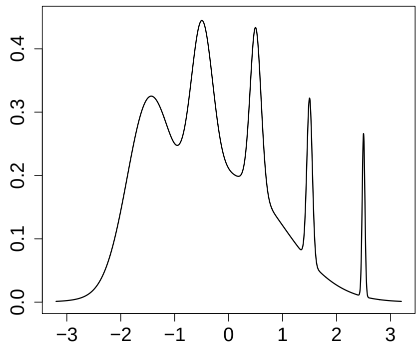
Finite sample behaviour

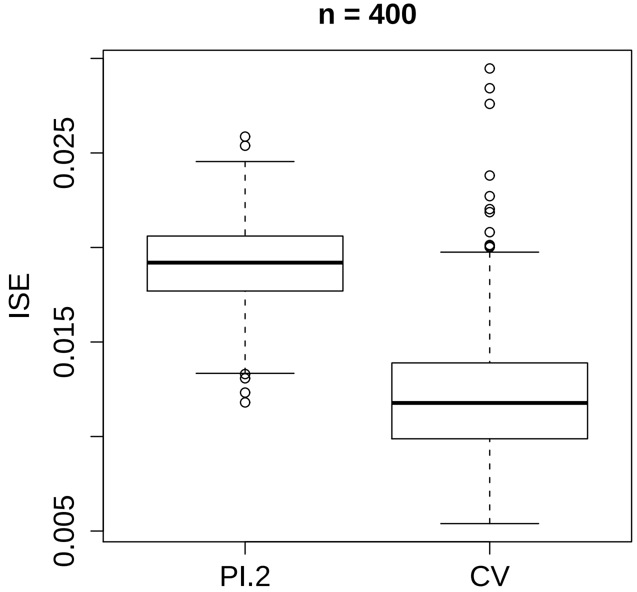
Strongly skewed density:





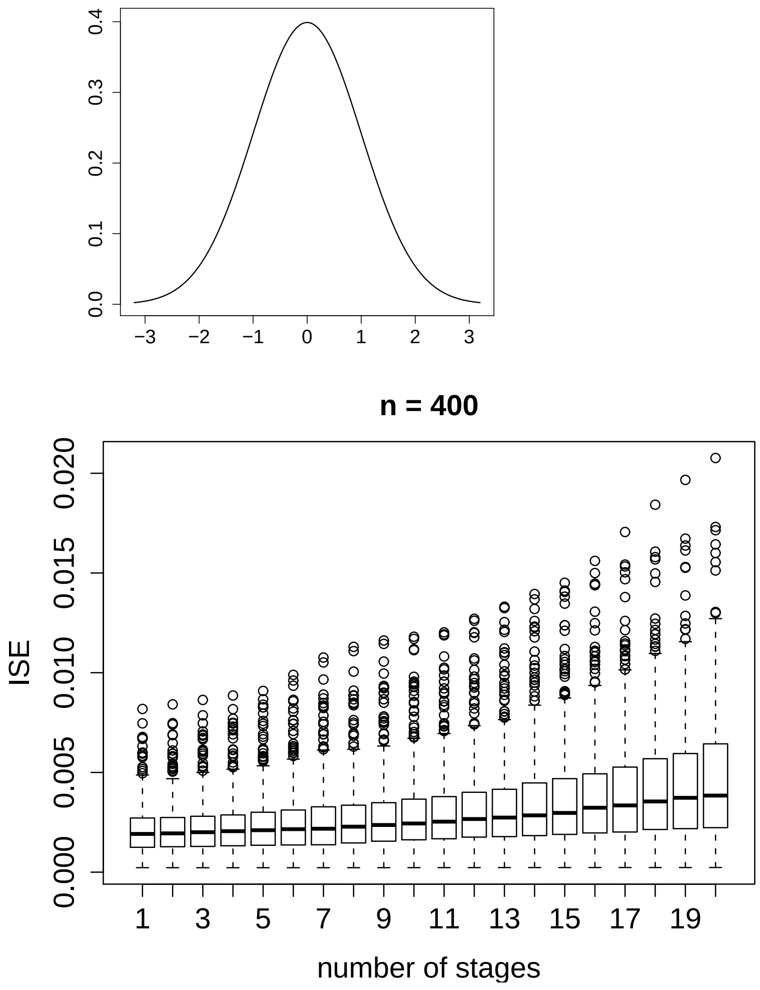
Asymmetric multimodal density:



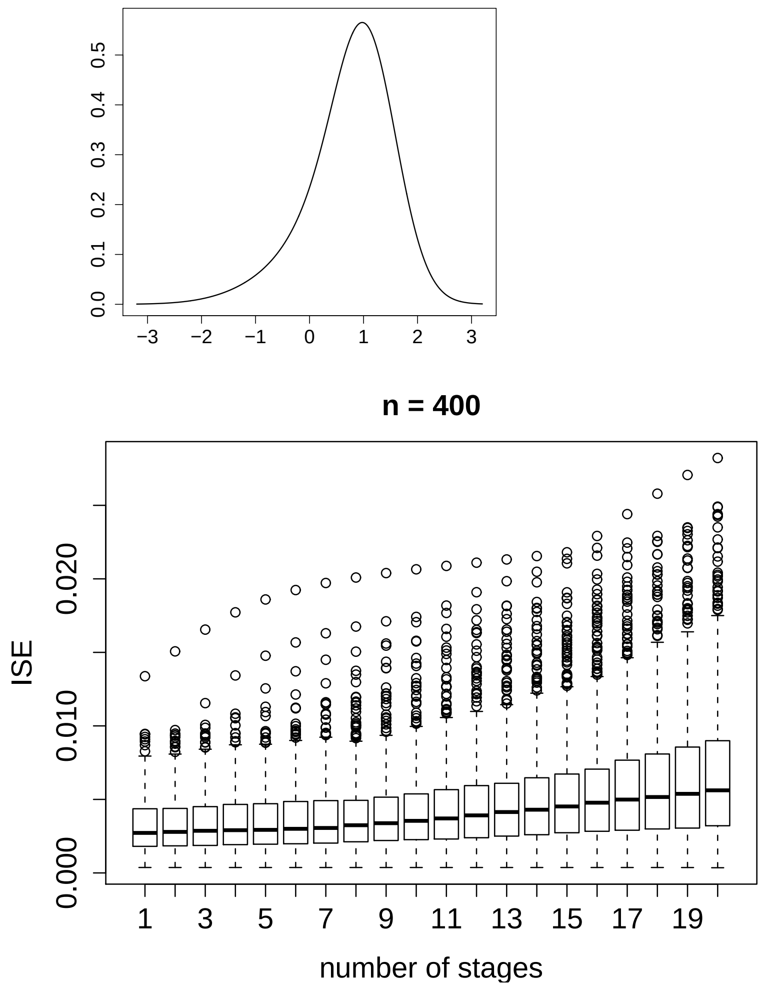


Finite sample behaviour

Standard normal density:

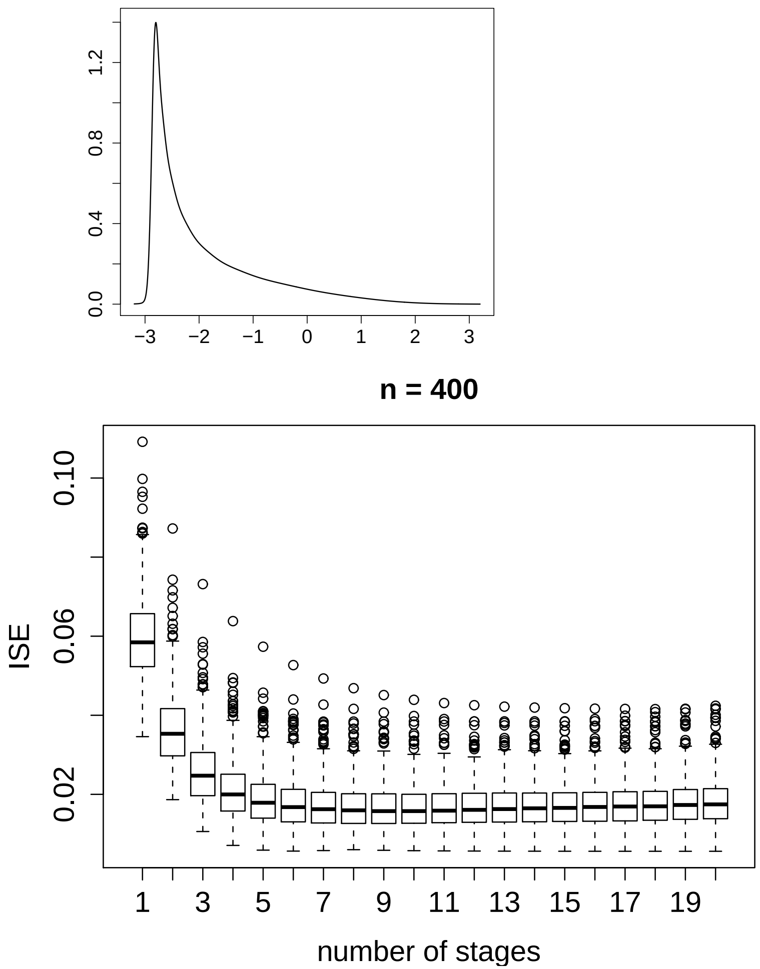


Skewed unimodal density:

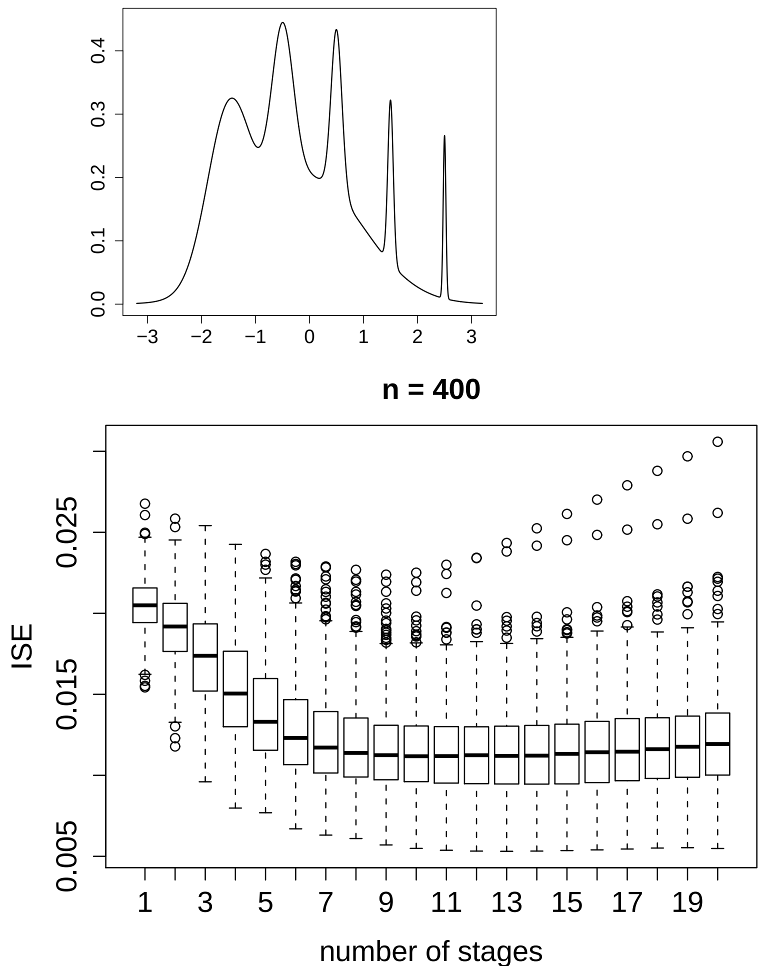


Finite sample behaviour

Strongly skewed density:



Asymmetric multimodal density:



Multistage plug-in bandwidth selector

- From a finite-sample point of view the performance of $\hat{h}_{\mathrm{PI},\ell}$ 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 ℓ?

  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.

- We started by fixing minimum and a maximum number of pilot stages

$$\underline{\mathrm{L}} \quad \text{and} \quad \overline{\mathrm{L}}$$

and by choosing a stage ℓ among the set of possible pilot stages

$$\mathscr{L} = \{\underline{\mathrm{L}}, \underline{\mathrm{L}} + 1, \ldots, \overline{\mathrm{L}}\}$$

- This is equivalent to select one of the bandwidths

$$\hat{h}_{\mathrm{PI},\ell} = c_K \hat{\psi}_{4,\ell}^{-1/5} n^{-1/5}, \ell \in \mathscr{L}.$$

- Recall that each one of these bandwidths has good asymptotic properties.

Combining plug-in and cross-validation procedures

- 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

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

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

- Finally, we take the bandwidth

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

  where

$$\hat{\ell} = \mathrm{argmin}_{\ell \in \mathscr{L}} \mathrm{CV}_\gamma(\hat{h}_{\mathrm{PI},\ell}).$$

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

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

then

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

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

(Chacón & CT, 2008)

- Condition (1) is not very restrictive due to the smoothness of the normal distribution.
- This result justifies the recommendation of using $\underline{L} = 2$.

Proof:

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

$$\mathrm{P}(\Omega_{\underline{L},\overline{L}}) \to 1$$

where

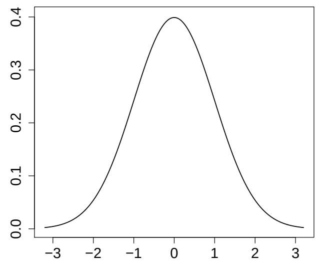
$$\Omega_{\underline{L},\overline{L}} = \left\{ \hat{h}_{\mathrm{PI},\overline{L}} \leq \hat{h}_{\mathrm{PI},\overline{L}-1} \leq \ldots \leq \hat{h}_{\mathrm{PI},\underline{L}+1} \leq \hat{h}_{\mathrm{PI},\underline{L}} \right\}.$$

□ The conclusion follows easily from the asymptotic behaviour of $\hat{h}_{\mathrm{PI},\overline{L}}$ and $\hat{h}_{\mathrm{PI},\underline{L}}$, since for a sample in $\Omega_{\underline{L},\overline{L}}$ we have

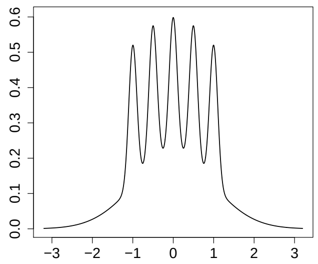
$$\frac{\hat{h}_{\mathrm{PI},\overline{L}}}{h_{\mathrm{MISE}}} - 1 \leq \frac{\hat{h}_{\mathrm{PI},\hat{\ell}(L)}}{h_{\mathrm{MISE}}} - 1 \leq \frac{\hat{h}_{\mathrm{PI},\underline{L}}}{h_{\mathrm{MISE}}} - 1.$$

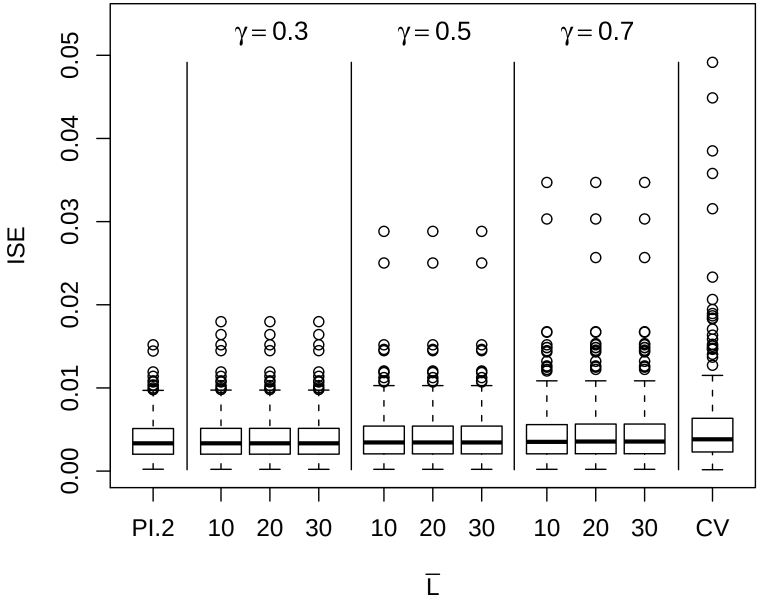
On the role played by $\overline{\mathrm{L}}$ and γ

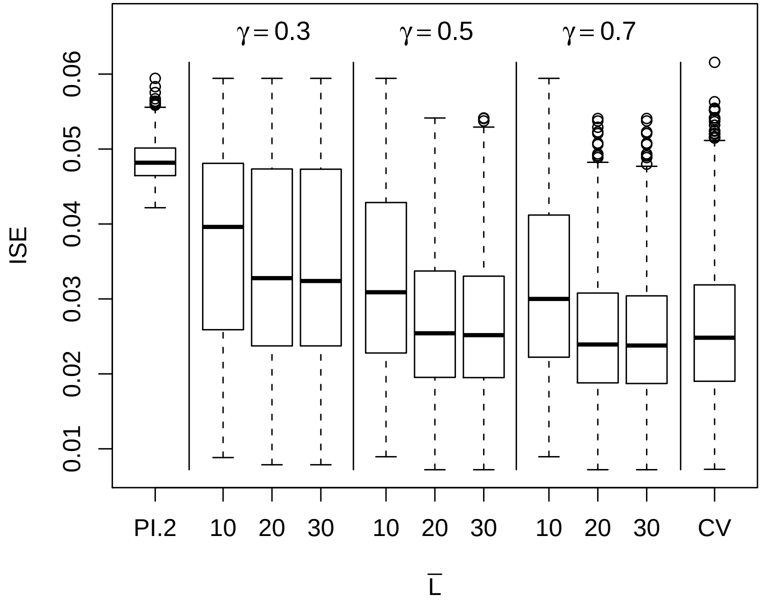
Standard normal density:



Claw density:







Distribution of $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\hat{\ell}})$ as a function of $\overline{\mathrm{L}}$ and γ $(n = 200)$

Choosing $\overline{\mathrm{L}}$ and γ in practice

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

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

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

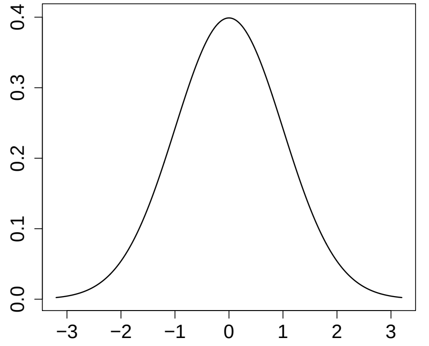
- Regarding the choice of γ, small values of γ are more appropriate for easy-to-estimate densities, whereas large values of γ are more appropriate for hard-to-estimate densities.

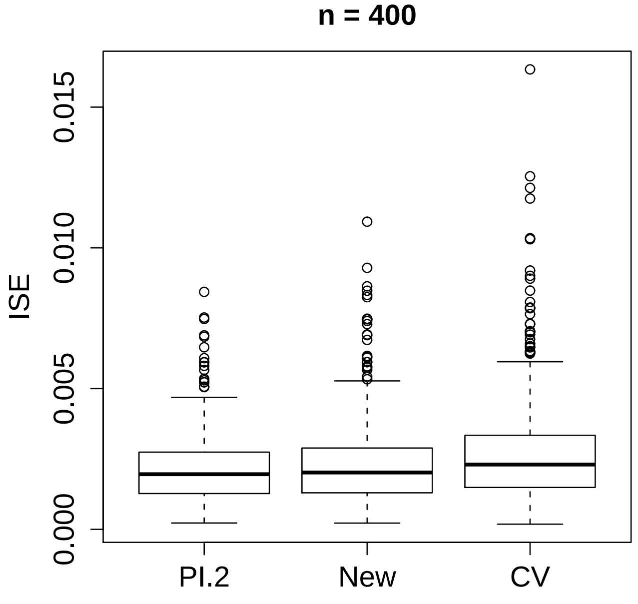
  In order to find a compromise between these two situations we decide to take $\gamma = 0.6$.

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

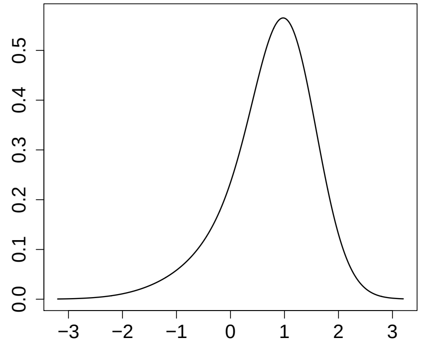
Finite sample behaviour

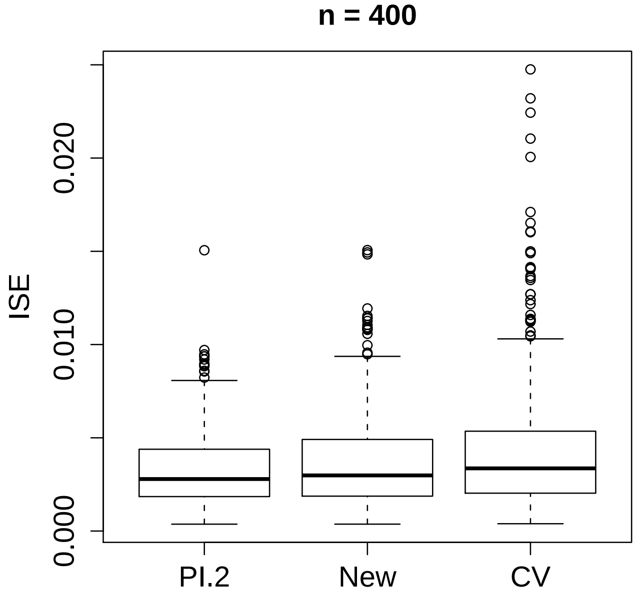
Standard normal density:





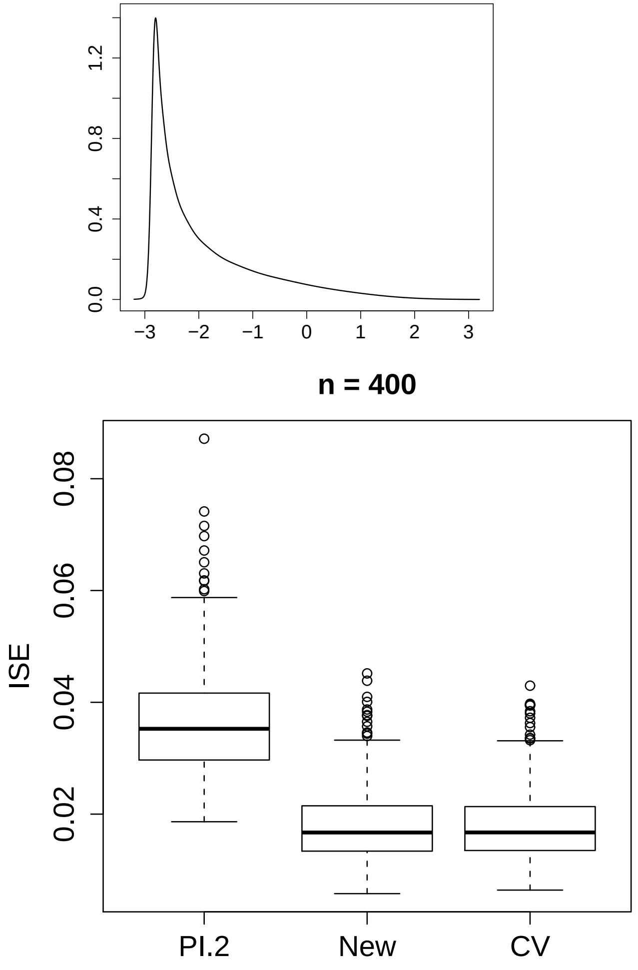
Skewed unimodal density:



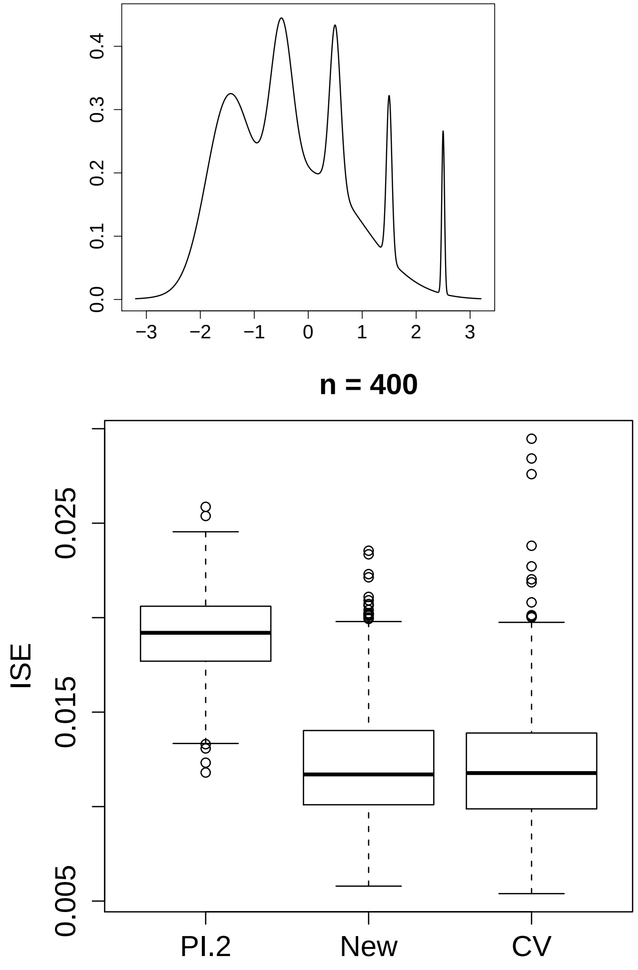


Finite sample behaviour

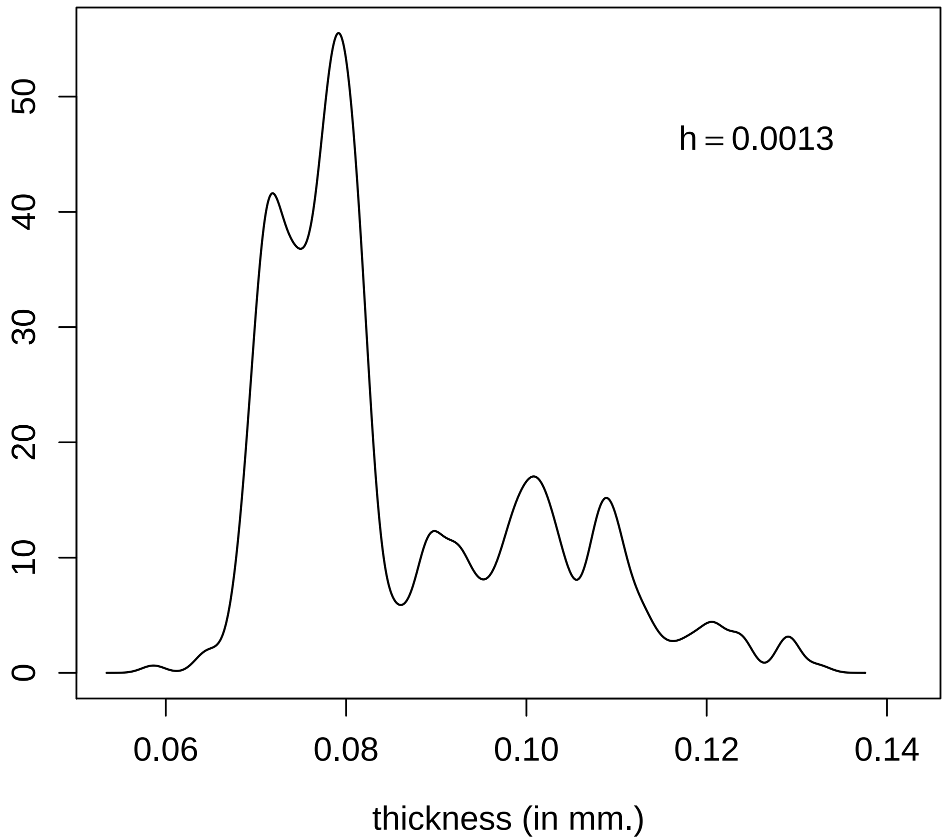
Strongly skewed density:



Asymmetric multimodal density:



Kernel density estimate for the Hidalgo Stamp Data



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

References

Izenman, A.J., Sommer, C.J. (1988).
Philatelic mixtures and multimodal densities.
J. Amer. Statist. Assoc. 83, 941–953.

Chacón, J.E., Tenreiro, C. (2008).
A data-based method for choosing the number of pilot stages for plug-in bandwidth selection.
Pré-publicação 08-59 DMUC (revised, January 2011).

Wand, M.P., Jones, M.C. (1995).
Kernel Smoothing.
Chapman & Hall